Time Series Analysis

STA457

Course Notes

Compiled by Amir Koutahi
Instructor: Vedant Choudhary

Summer 2026

Contents

Chapter 1

Lecture 1

1.1 Introduction

What is a time series?

Definition 1.1: Time Series

A *time series* is a set of observations x_t, each one being recorded at a specific time t. Equivalently, it is a collection of data $\{x_t\}$ observed sequentially in time.

The corresponding random variables are denoted X_t. The time index t belongs to some discrete index set $\mathcal{T}$, typically equally spaced:

$$\mathcal{T} = \{1, 2, \ldots, n\}.$$

Unequally spaced indices also arise in practice (for example, stock-market data has no observations on weekends and holidays). It is useful to think of $\{x_t\}$ as a **realization** of some stochastic (random) process $\{X_t\}$ at distinct points in time.

Time series vs. stochastic process

Definition 1.2: Discrete Time Series

A *discrete time series* is one in which the set $\mathcal{T}$ of times at which observations are made is a **discrete** set.

Time series data are almost always measured in discrete time, even when the underlying process is continuous.

Definition 1.3: Stochastic Process

A *stochastic process* is a collection of random variables, denoted $\{X_t : t \in \mathcal{T}\}$, that models the evolution of a system over time. (not one RV collection of RVs)

The index set for a stochastic process can be discrete or continuous (e.g. $\mathcal{T} = (0, \infty)$). A continuous-time stochastic process $\{X_t\}$ can generate a discrete time series by:

- **Sampling**: observe X_t at $t = 1, 2, \ldots, n$.
- **Aggregating**: aggregate over intervals $x_t = \int_{t-1}^{t} X_s \, ds$.

In the case of stock markets the data gets updated in milliseconds so extremely noisy use of aggregation is preferred over sampling for example: **VWAP (Volume Weighted Average Price)** is a trading benchmark giving the average price weighted by volume, calculated as VWAP = $\frac{\sum_{i=1}^{n} P_i \cdot V_i}{\sum_{i=1}^{n} V_i}$, where P_i is the typical price $\frac{H+L+C}{3}$ and V_i is the volume of trade i. Prices above VWAP signal bullish sentiment, below signal bearish; institutions use it as an execution benchmark. However for longer periods of stock market analysis we use smapling.

Time series vs. classical statistics

Classical Statistics considers a simple random sample (i.i.d.) $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$. Reordering the samples has no impact.

Time Series Analysis considers $X_1, \ldots, X_T$, where X_t denotes the observation at time t — the observations are *ordered* in time and cannot be reordered.

The main differences are:

1. **Dependence vs. independence:** successive observations are correlated (autocorrelation), not independent.
2. **Single realization vs. replication:** we observe only one realization of $\{X_t\}$ (TSA), not repeated independent copies (LR).
3. **Purposes of models:** description, forecasting, control — not just inference about a population parameter.

Goals of time series analysis

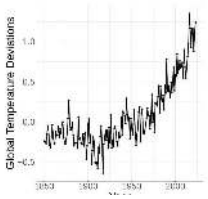
The basic idea of time series modeling is to find a stochastic process whose characteristics are similar to the observed time series.

1. **Description.** Use data to understand the structure of the underlying process. Does not necessarily require a formal model.
2. **Prediction / Forecasting.** Use past and present data to predict future values. Easier said than done — requires careful model selection.
3. **Control.** Adjust inputs to steer future values of the series toward desired targets.
    - Example: rebalance the weight of stocks in an investment portfolio.
    - Example: adjust water level behind a dam to regulate streamflow.

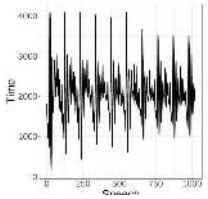
Characteristics of time series

The primary objective of time series analysis is to develop mathematical models that provide plausible descriptions for sample data. When considering time series, one needs to deal with:

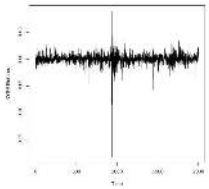
1. **Heterogeneity in data**
    - Time trends.
    - Seasonality.
    - Heteroskedasticity.
2. **Serial dependence / serial correlation.** Observations that are temporally close are most likely dependent in some way.



(a) Time increasing quadratic trend



(b) Seasonality



(c) Heteroskedasticity (non-constant variance

Figure 1.1: Real-world examples of heterogeneity in data

Application: modeling financial data

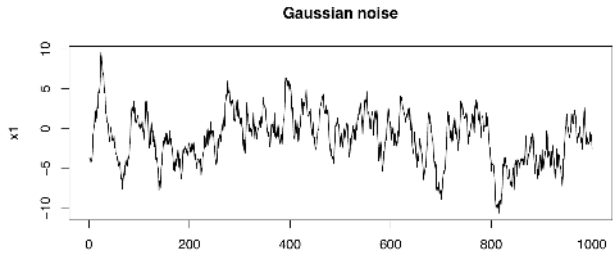
A prominent application of time series is modeling financial data, e.g. daily stock price, interest rate, exchange rate. The *good news*: data are abundant and at very high frequency. The *challenge*: modeling can be very difficult.

The classical time series model is

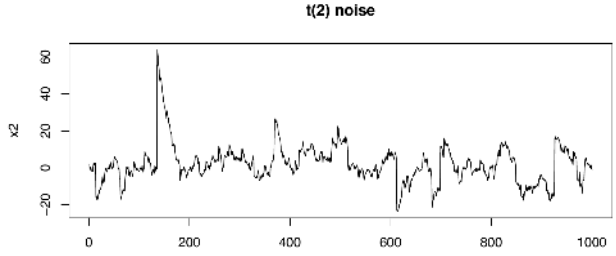
$$x_t = g_\theta\big(\underbrace{x_{t-1}, x_{t-2}, \dots}_{\text{past}}\big) + \varepsilon_t, \qquad \varepsilon_t \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

Two implications of this model often *fail* for financial data:

- Gaussian ε_t implies **light tails** (compared to t distribution) — few extreme values.
- If g_θ is not too non-linear, then $\{x_t\}$ looks the same even if the sequence of observations is reversed. $(x_{t-1}, x_{t-2}, \dots, x_{t-k})$ vs. $(x_{t-k}, x_{t-k+1}, \dots, x_{t-1})$



(a) AR(1) process $x_t = 0.95\, x_{t-1} + \varepsilon_t$ with $\varepsilon_t \overset{\text{iid}}{\sim} N(0, 1)$. Extreme values are rare.



(b) AR(1) process $x_t = 0.95\, x_{t-1} + \varepsilon_t$ with $\varepsilon_t \overset{\text{iid}}{\sim} t(2)$ (Student's t). Occasional extreme spikes are clearly visible — a feature common in financial returns.

Figure 1.2: heavy-tailed distributions (Student's t) allow for extreme values that are virtually absent under Gaussian innovations.

Challenges in modeling financial data

Real financial data typically exhibit:

- **Heavy tails:** far more extreme values than one would expect under a Gaussian model.
- **Conditional heteroscedasticity:** the variance of the noise ε_t depends on past values of x_t — volatility clustering (periods of high volatility followed by periods of low volatility).
- **Economic interventions:** structural breaks, regime changes, crises (e.g. financial crisis of 2008, COVID crash in 2020).

Specialized models that address these issues include:

- ARCH / GARCH models for conditional heteroscedasticity.
- Heavy-tailed innovations (Student's t, stable distributions).
- Regime-switching and jump-diffusion models.

Modeling approaches

Time series analysis is usually compartmentalized into two approaches:

1. **Time Domain.** Analyze time series data with respect to time. Models express explicitly how $\{x_t\}$ varies over time. Current and future values of the series are functions of the past values:

$$x_t = f(x_{t-1}, \ldots, x_{t-p}) + \varepsilon_t,$$

   e.g. autoregressive models.

2. **Frequency (Spectral) Domain.** Analyze time series data with respect to frequency (i.e. cycles in the data). Models for $\{x_t\}$ represent its periodic behavior using a series of random sinusoids:

$$x_t = \sum_j \big\{A_j \cos(2\pi\omega_j t) + B_j \sin(2\pi\omega_j t)\big\},$$

   where $\{\omega_j\}$ are frequencies and $\{A_j\}, \{B_j\}$ are random variables. If $t \in \mathcal{T} = \{0, \Delta, 2\Delta, \ldots\}$, then $\omega_j \in [0, 1/(2\Delta)]$. **Note:** No error terms errors are inside A_j and B_j.

Time domain vs. frequency domain

Which approach is better? Both give different insights, and both can be very useful.

- Time-domain approaches allow us to determine how the future "depends" on the past. More common in practice for forecasting (e.g. predicting stock prices).
- Frequency-domain approaches are useful for **identifying hidden cycles** in data. More common for analyzing historical data for cycles (e.g. recessions in economic data).

Generally, for any frequency-domain method/concept there is an equivalent time-domain method/-concept (and vice versa):

$$\underbrace{\text{autocovariance function}}_{\text{time domain}} \iff \underbrace{\text{spectral density function}}_{\text{frequency domain}}.$$

This course puts more emphasis on the time-domain approach.

1.2 Patterns and Decomposition

Time series patterns

When analyzing a time series plot, we look for four fundamental patterns:

1. **Trend** (m_t): a long-term increase or decrease in the level of the series.
2. **Seasonal** (S_t): a pattern affected by seasonal factors such as time of year, day, or week. Typically a fixed and known frequency.
3. **Cyclic** (C_t): rises and falls of *non-fixed* frequency, typically driven by economic or business cycles. Typically span at least two years. Often combined with trend.
4. **Random** (R_t): the residual variation after accounting for the other three components.

Key distinction: seasonal patterns have a fixed, predictable period; cyclic patterns do not.

Pattern: trend

A **trend** exists when the data show a gradual, long-run shift to higher or lower values. It can be linear, quadratic, or more general (for example, the global mean temperature increasing over 150 years).

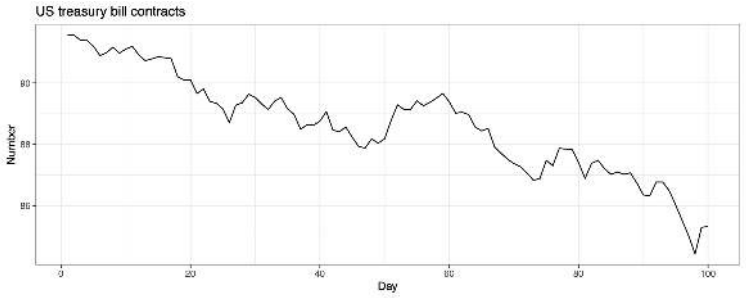


Figure 1.3: US treasury bill contracts. Trend: downward. Seasonality: none apparent. Cyclic: none (although it might exist at a longer horizon).

Pattern: seasonality

A **seasonal pattern** exists when the series exhibits a regular, repeating pattern at a *fixed and known* period s. Common periods are $s = 12$ (monthly data) and $s = 4$ (quarterly data). Examples include peaks in winter electricity demand.

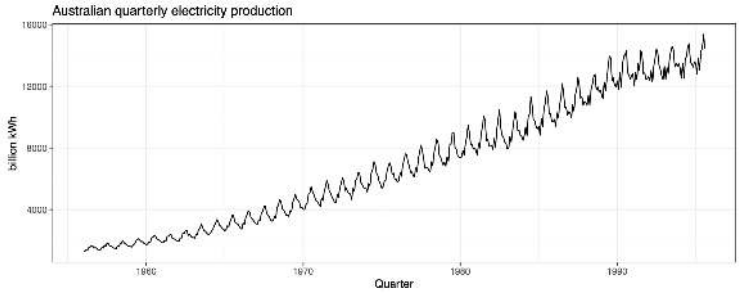


Figure 1.4: Australian quarterly electricity production. Trend: clear upward. Seasonality: strong. Cyclic: none apparent.

Pattern: cyclic variation

A **cyclic pattern** exists when the series exhibits rises and falls that are *not* of a fixed frequency, typically driven by economic or business cycles. Duration varies, usually at least two years but can be much longer. Cycles are harder to model because the period is unknown. Examples include housing prices and GDP growth cycles.

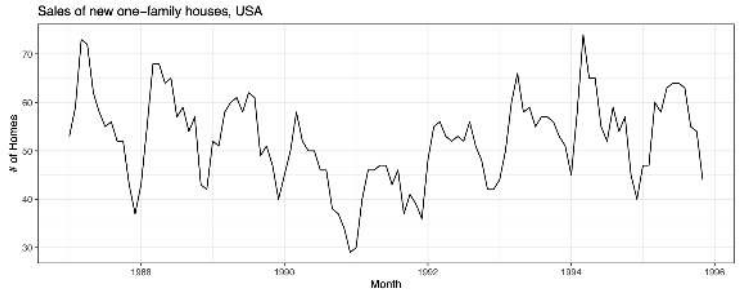


Figure 1.5: We see overall cyclic pattern with seasonality in smaller increments than cycles

Pattern: random component

The **random component** is the residual variation after accounting for the other three components. It has no predetermined pattern or shape.

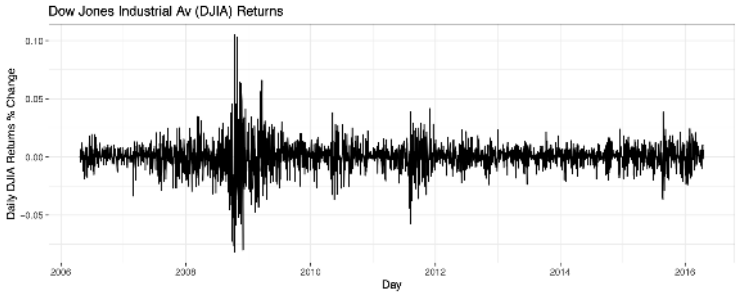


Figure 1.6: Dow Jones Industrial Average (DJIA) returns. Trend: none. Seasonality: none. Cyclic: none. However, it exhibits **volatility clustering**.

Combined patterns

Many real series exhibit *multiple* patterns simultaneously. For example, monthly international air-passenger numbers (1949–1960) show both an upward trend and strong quarterly seasonality. The seasonal amplitude grows with the level — a sign that a *multiplicative* decomposition may be more appropriate than an additive one.

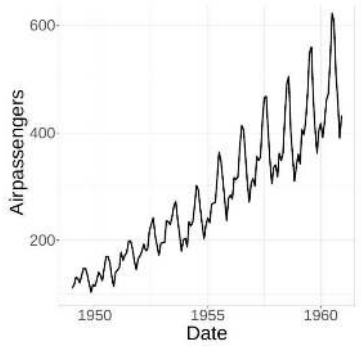


Figure 1.7: Trend + seasonality. Also notice that the amplitude of seasonality changes → multiplicative model (as $t \uparrow$, $m_t \uparrow \Rightarrow m_t \cdot S_t \uparrow$).

Classical decomposition: additive model

The classical approach decomposes a time series into four components:

$$\boxed{x_t = m_t + S_t + C_t + R_t}$$

- m_t — **Trend**: long-term level or direction.
- S_t — **Seasonal**: fixed-period repeating fluctuation.
- C_t — **Cyclic**: medium-term irregular oscillation.
- R_t — **Random/Irregular**: unpredictable residual component.

The first three components are **deterministic**, whereas the last is not.

A **multiplicative** form

$$x_t = m_t \cdot S_t \cdot C_t \cdot R_t$$

is used when the amplitude of fluctuations grows proportionally with the level (a log-transform converts it back to additive). ($m_t \cdot S_t$ increases the amplitude which isn't captured in additive model)

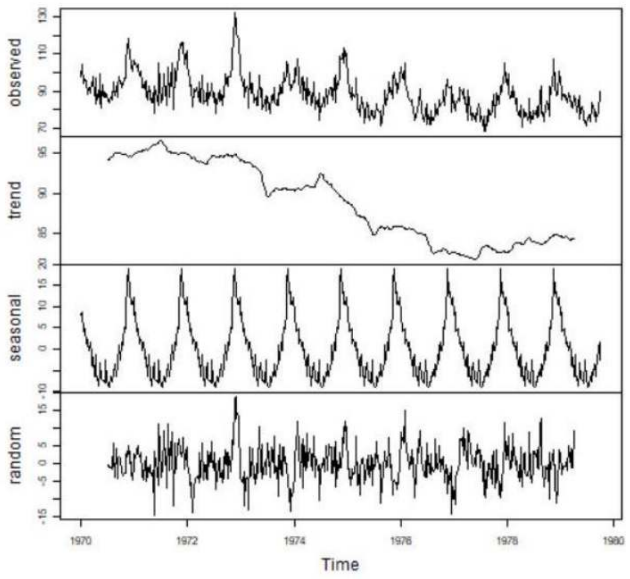


Figure 1.8: Decomposition: 1) observed, 2) trend, 3) seasonality (no error deterministic), 4) Stochastic error

1.3 Classical Regression

Simple forecasting approaches

The simplest forecasting approaches are called **naive methods**:

- **Last value:** $\hat{x}_{T+h} = x_T$ for all horizons h.
- **Average:** $\hat{x}_{T+h} = \bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t$.

These serve as benchmarks — a "good" model should beat them.

Classical regression

We may extend the idea of linear regression to the time series context by assuming

- some output or dependent time series, say x_t, for $t = 1, \ldots, n$,
- is being influenced by a collection of possible inputs or independent series, say $z_{t1}, z_{t2}, \ldots, z_{tq}$, (First discrepancy in MLR the predictors should be independent and not correlated but time series violates it)

where we first regard the inputs as fixed and known. We express this relation through the linear regression model

$$x_t = \beta_0 + \beta_1 z_{t1} + \beta_2 z_{t2} + \cdots + \beta_q z_{tq} + w_t,$$

where $w_t \overset{\text{iid}}{\sim} (0, \sigma_w^2)$ is the random error. (Second discrepancy is that we dont have constant variance in time series)

Capturing patterns via regression

- A **linear trend** is captured by setting $z_{t1} = t$:

$$x_t = \beta_0 + \beta_1 t + w_t.$$

- A **quadratic trend** is captured by setting $z_{t1} = t$, $z_{t2} = t^2$:

$$x_t = \beta_0 + \beta_1 t + \beta_2 t^2 + w_t.$$

- **Seasonality** is captured by adding $d-1$ quarterly/monthly dummy variables (if period of seasonality $= d$), $z_{t,k} = I_{\{t=\text{period } k\}}$:

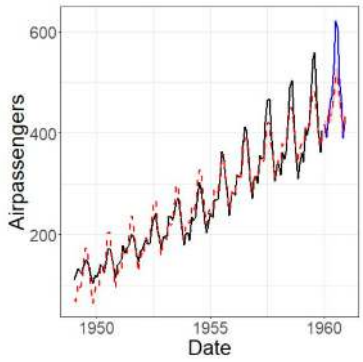
$$x_t = \beta_0 + \delta t + \beta_1 z_{t,1} + \cdots + \beta_{d-1} z_{t,d-1} + w_t.$$

  For example in Figure 1.9 we have 12 months for seasonality $\Rightarrow$ 11 dummy variables $z_{tj}, \quad j \in \{0, 1, \ldots, 11\}$. January will have $z_{tj} = 0$ for all values of j.
- **Cyclicity** can be approximated as a sum of sinusoids at different frequencies:

$$x_t = \beta_0 + \sum_j \big\{\beta_j \cos(2\pi\omega_j t) + \gamma_j \sin(2\pi\omega_j t)\big\} + w_t.$$

Question: How do we formally test for the presence of such behavior? Hypothesis testing for significance of coefficients.

Air passengers: raw model



(a) Regression fit using linear trend and monthly dummy variables (original scale). Red dashed: fitted values; black: observed past; blue: observed future.

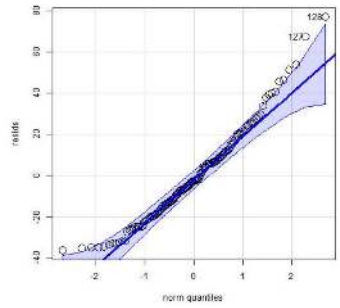


Figure: (a) Residual QQ plot

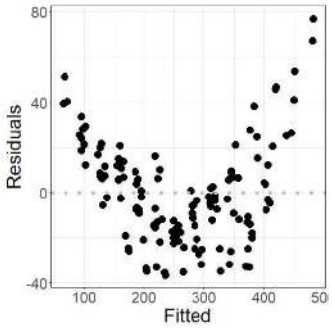
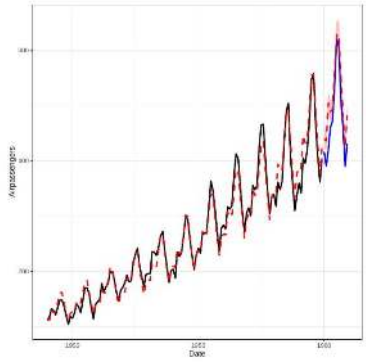


Figure: (b) Residuals versus fitted

(b) Diagnostic plots: residual QQ plot (left) and residuals versus fitted (right). Residuals show clear heteroscedasticity — variance grows with level.

Figure 1.9: Not perfect model and we see non-zero variances in residual plot +deviation from linearity in Q-Q plot

Log-transformed model



(a) Regression fit with linear trend and monthly dummies on log-transformed data. The seasonal amplitude is now stable.

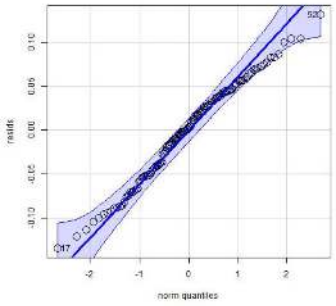


Figure: (a) Residual QQ plot

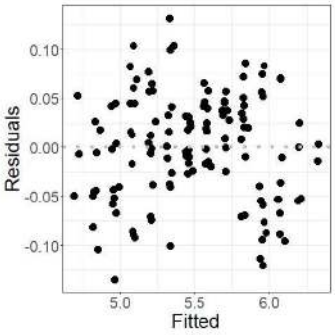


Figure: (b) Residuals versus fitted

(b) Diagnostics for the log-transformed model. Residuals are much better behaved: near-normal QQ plot and no obvious heteroscedasticity.

Figure 1.10: The residuals improved but we see a bit of clustering → left most side only negative residuals and just to the right of it only positive residuals

Box–Cox transformations

When variance grows with the level, a **variance-stabilizing transformation** is needed. The Box–Cox family is

$$y_t = \begin{cases} \log(x_t) & \text{if } \lambda = 0, \\ \dfrac{\operatorname{sign}(x_t)\,|x_t|^{\lambda} - 1}{\lambda} & \text{otherwise.} \end{cases}$$

- $\lambda = 1$: no transformation (identity).

- $\lambda = 0$: natural log (most common for positive data).
- $\lambda = 0.5$: square-root transformation.
- λ is typically estimated from the data by maximum likelihood. (profile log-likelihood)

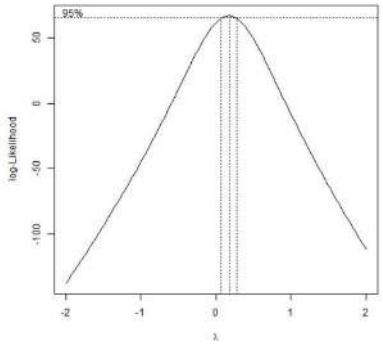
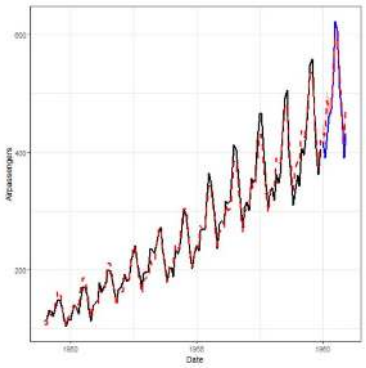


Figure 1.11: Profile log-likelihood for λ on the air-passenger data. The MLE is close to $\lambda = 0$ (log transformation), with a 95% confidence interval shown. So we shouldn't expect much difference between Box-cox and log transformation for this scenario



(a) Regression fit with trend and monthly dummies on Box–Cox transformed data ($\hat{\lambda} \approx 0$).

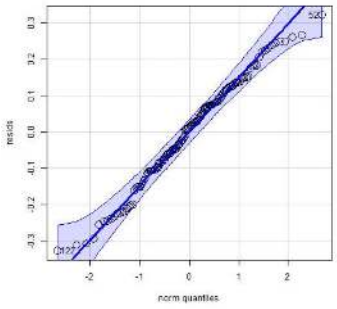


Figure: (a) Residual QQ plot

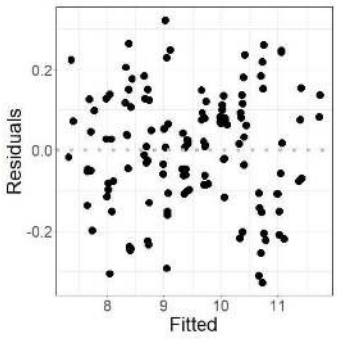


Figure: (b) Residuals versus fitted

(b) Diagnostics for the Box–Cox transformed model the most improved non correlated zero variance residual plot.

Figure 1.12: Box–Cox transformed has similar results to the log model, confirming $\lambda \approx 0$ is appropriate.

Linear model considerations

Typical issues when fitting linear regression to time series:

- **Multicollinearity:** trend and seasonal dummies can be highly correlated; use orthogonal coding or regularization. Leftover autocorrelation means we have not captured all structure
- **Model selection:** choose among competing trend/seasonal specifications using AIC, BIC, or cross-validation.
- **Model evaluation:** residuals should be
  - uncorrelated (check with the ACF plot), and
  - mean zero (no systematic pattern).

Remark 1.1: Residual Plot

uncorrelated (check with the ACF plot), and mean zero (no systematic pattern).

Chapter 2

Lecture 2

2.1 Detrending and Differencing

Signal + noise decomposition

The general logic underlying classical models: suppose x_t is generated by

$$x_t = s_t + \varepsilon_t,$$

where

1. s_t is a **deterministic** signal/trend (estimated by regression),
2. ε_t is **stochastic** noise/error with $\mathbb{E}[\varepsilon_t] = 0$, (not white noise, white noise is the ideal case)
3. $\varepsilon_t = g(\epsilon_t, \epsilon_{t-1}, \ldots)$ where $\{\epsilon_i\}$ are i.i.d. random variables (also called innovations or shocks). Function of previous shocks

If we can estimate s_t and model ε_t, we can:

- Perform prediction and forecasting.
- Quantify uncertainty around forecasts.
- Conduct inference (hypothesis tests, confidence intervals).

Detrending

Definition 2.1: Detrending

Detrending a time series constitutes computing the residuals after estimating the signal/trend component. A *detrended time series* is the series of such residuals.

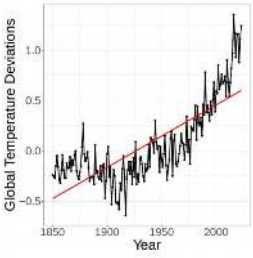
Components that may constitute the "trend" s_t:

- **Trend**: long-term direction (linear, polynomial, or nonparametric).
- **Seasonal**: fixed-period repeating fluctuation.
- **Cyclic**: irregular medium-term oscillation.

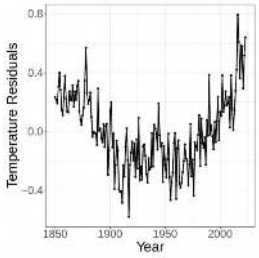
Do not confuse the seasonal component S_t with the trend s_t above.

$$x_t = S_t + \varepsilon_t$$

Estimate S_t: $\hat{\varepsilon}_t = x_t - \hat{S}_t$ (estimated residual after getting rid of the trend component)



(a) OLS estimate of a linear trend overlaid on global temperature deviations.



(b) Residuals after removing the linear OLS trend

Figure 2.1: a "detrended" time series. Structure remains — residuals are not white noise.

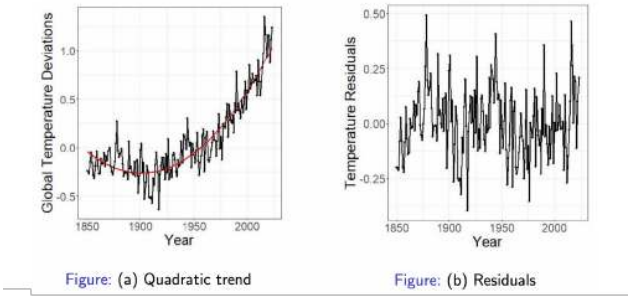


Figure 2.2: Left: quadratic trend fit to global temperature data. Right: corresponding residuals — more symmetric and closer to white noise than with the linear fit.

Other classical approaches

Beyond regression-based detrending, two other common approaches:

- **Detrending**: fit a parametric trend $\hat{s}_t$ (polynomial, spline, etc.) and work with residuals $x_t - \hat{s}_t$.
- **Differencing**: instead of estimating a trend, *remove* it by taking successive differences.

Both are typically combined with regression for seasonal components, though seasonality can also be eliminated by differencing.

2.2 Example Models

Theoretical example: white noise

The simplest series is a collection of uncorrelated random variables w_t ($\text{Cov}(w_i, w_j) = 0 \quad \forall\, i \neq j$) with mean 0 and finite variance σ_w^2.

Definition 2.2: White Noise

The time series $\{w_t\}$ is called *white noise* if the w_t are uncorrelated random variables with mean 0 and finite variance σ_w^2.

- We denote it $w_t \sim \text{wn}(0, \sigma_w^2)$, or simply $w_t \overset{\text{iid}}{\sim} (0, \sigma_w^2)$. (can be from any ditribution)
- If additionally $w_t \sim N(0, \sigma_w^2)$, we call w_t **Gaussian white noise**: $w_t \overset{\text{iid}}{\sim} N(0, \sigma_w^2)$.
- White noise is the "ideal" residual — what we *want* to be left with after fitting a model. (But that is not always. In the definition of time series the only constraint on error is mean zero) (we say that white noise is ideal because there is no more trend that can be exploited from it)

Theoretical example: random walk

Another simple example is the series in which each value x_t equals the previous one x_{t-1} plus some random shock.

Definition 2.3: Random Walk

The *random walk* model is

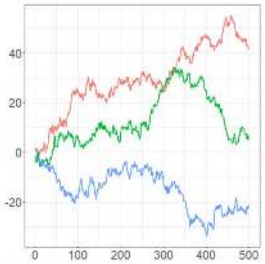
$$x_t = x_{t-1} + w_t \qquad \text{for } t = 1, 2, \ldots$$

with initial condition $x_0 = 0$ and w_t white noise.

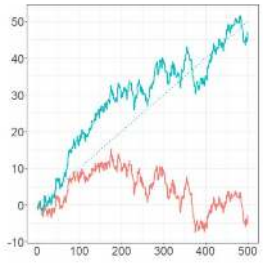
The **random walk with drift** generalizes this:

$$x_t = \delta + x_{t-1} + w_t, \qquad \text{so} \qquad x_t = \delta t + x_0 + \sum_{i=1}^{t} w_i,$$

where δ is the drift. The cumulative sum makes x_t very persistent. According to the random-walk hypothesis, stock prices fluctuate randomly and it is impossible to predict future prices based on past trends — the "efficient market" hypothesis. ($s_t = s_{t-1} + w_t$) Evidently the error term in this model is itself a random walk, and hence is somewhat erratic.



(a) Three independent random-walk realizations. Despite sharing no common trend, they can look deceptively "trendy". (All three don't have drifts)



(b) Random walk with drift (teal, $\delta > 0$) vs. the undrifted random walk (orange). The dotted line shows the deterministic drift δt.

Figure 2.3: Random walks: Can be deceptive

Random Walk with Drift: $x_t = \delta + x_{t-1} + w_t$

$$
\begin{aligned}
&t = 1 : x_1 = \delta + w_1 \\
&t = 2 : x_2 = \delta + x_1 + w_2 = 2\delta + w_1 + w_2 \\
&\qquad \vdots \\
&t = n : x_n = \delta + x_{n-1} + w_n = n\delta + w_1 + w_2 + \cdots + w_n
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}(x_n) &= \mathrm{Var}(n\delta + w_1 + w_2 + \cdots + w_n) \\
&= \mathrm{Var}(w_1) + \mathrm{Var}(w_2) + \cdots + \mathrm{Var}(w_n) \\
&= n\sigma_w^2
\end{aligned}
$$

Note: The drift term $n\delta$ is deterministic, so it contributes nothing to the variance. So the variance is same if there was no drift. The cross-covariance terms vanish because w_i, w_j are uncorrelated for $i \neq j$.

Random walk with drift: differencing

- Detrending a random walk with drift by regression does not work well. Why?
- Because the error term is itself a random walk (non-stationary –¿ non constant mean and non constant variance).
- doesn't have constant variance
- residuals have non-zero covariances
- We may turn to another way.

Differencing works! Under the random walk model,

$$\nabla x_t = x_t - x_{t-1} = \delta + \varepsilon_t,$$

so the sequence of first differences should look like a mean-shifted, white-noise sequence. (shifted by δ)

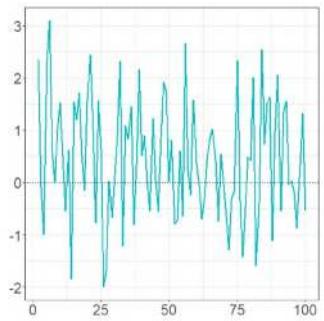


Figure 2.4: First difference of a random-walk-with-drift series: the differenced series fluctuates around the constant δ.

Differencing

Definition 2.4: Lag (Backshift) Operator B

The lag operator B gives the lagged values of the time series:

$$By_t = y_{t-1}, \qquad B^2 y_t = y_{t-2}, \qquad B^k y_t = y_{t-k}.$$

Definition 2.5: Differencing

Differencing a time series constitutes computing the difference between successive terms. The *first-difference operator* ∇ is defined by

$$\nabla x_t = x_t - x_{t-1} = (1 - B)\, x_t.$$

The resulting series $x_2 - x_1, x_3 - x_2, \ldots, x_T - x_{T-1}$ has length $T - 1$. Higher-order differences are defined recursively:

$$\nabla^d x_t = \nabla^{d-1}(\nabla x_t) = (1 - B)^d x_t.$$

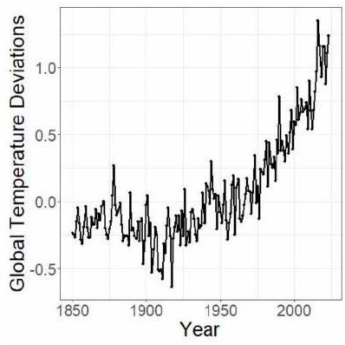


Figure: (a) Quadratic trend

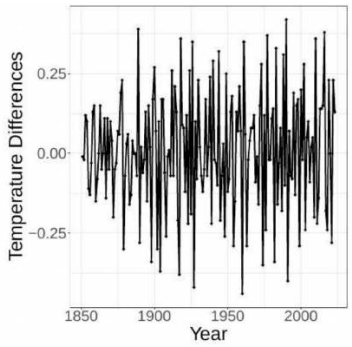


Figure: (b) Residuals

Figure 2.5: Left: global temperature deviations. Right: first differences ∇x_t. The differenced series shows no apparent trend and fluctuates around zero.

Example of higher order differencing:

$$\begin{aligned}
\nabla x_t &= x_t - x_{t-1} \\
\nabla^2 x_t &= \nabla(\nabla x_t) \\
&= \nabla(x_t - x_{t-1}) \\
&= (x_t - x_{t-1}) - (x_{t-1} - x_{t-2}) \\
&= x_t - 2x_{t-1} + x_{t-2}
\end{aligned}$$

Why differencing eliminates polynomial trends

- If the trend is **linear**: $m_t = a + bt$, then
$$\nabla m_t = m_t - m_{t-1} = b \quad \text{(constant — trend gone)}.$$
- If the trend is **quadratic**: $m_t = a + bt + ct^2$, then $\nabla^2 m_t = 2c$.
- **General rule**: if $m_t = a_0 + a_1 t + \cdots + a_k t^k$ is a polynomial of degree k, then $\nabla^k m_t = k!\, a_k$ (a constant), so k differences remove the trend entirely.

For $x_t = m_t + \varepsilon_t$ with polynomial trend of degree k:
$$Z_t = \nabla^k x_t = k!\, a_k + \nabla^k \varepsilon_t,$$
so $\mathbb{E}(Z_t) = k!\, a_k$ is a **constant** — no trend in Z_t.

Caveat: each differencing step **increases** $\mathrm{Var}(Z_t)$, so we test for remaining trend after each step to prevent unnecessary differencing.

Variance under repeated differencing

Let $\mathbb{E}(\varepsilon_t) = 0$, $\mathrm{Var}(\varepsilon_t) = \sigma^2$, and $\mathrm{Cov}(\varepsilon_t, \varepsilon_s) = 0$ for $t \neq s$.

- $\nabla \varepsilon_t = \varepsilon_t - \varepsilon_{t-1} \;\Rightarrow\; \mathrm{Var}(\nabla \varepsilon_t) = 2\sigma^2$.
- $\nabla^2 \varepsilon_t = \varepsilon_t - 2\varepsilon_{t-1} + \varepsilon_{t-2} \;\Rightarrow\; \mathrm{Var}(\nabla^2 \varepsilon_t) = 6\sigma^2$.
- For $k = 2$: $\mathrm{Var}(Z_t) = \sigma^2 + 4\sigma^2 + \sigma^2 = 6\sigma^2$.

Remark 2.1: Key Point

Each differencing step **inflates the variance**. After differencing k times, we test whether any trend remains before differencing again.

2.3 Seasonality Elimination

Seasonal component

Definition 2.6: Seasonal Component

$\{s_t\}$ is a *seasonal component* of period d if it is periodic with period d and sums to zero over each complete cycle:
$$\cdots = s_{t-2d} = s_{t-d} = s_t = s_{t+d} = \cdots \qquad \text{and} \qquad \sum_{j=1}^{d} s_j = 0.$$

- **Why $\sum_{j=1}^{d} s_j = 0$?** The seasonal effects must average to zero over a full cycle so that the trend m_t represents the true long-run mean level of x_t. Without this constraint, any non-zero average of s_t would be absorbed into m_t, making the decomposition non-unique (**identifiability constraint**).
- For **monthly data** with $d = 12$, re-index time as $t = 12(j-1) + k$:
  - $j = 1, 2, \ldots, J$ — year index,
  - $k = 1, 2, \ldots, 12$ — month index within year,

  so $y_t \equiv y_{j,k}$ is the observation in month k of year j.

Eliminating trend and seasonality by differencing

The classical additive model decomposes the series as

$$x_t = m_t + s_t + \varepsilon_t, \qquad \mathbb{E}(\varepsilon_t) = 0, \quad \mathrm{Var}(\varepsilon_t) = \sigma^2 < \infty.$$

The seasonal difference operator. Define

$$\nabla_d x_t = x_t - x_{t-d}.$$

Applied to the additive model, the seasonal component vanishes by the identifiability constraint:

$$\nabla_d x_t \;=\; \underbrace{(m_t - m_{t-d})}_{\text{reduced trend}} + \underbrace{(s_t - s_{t-d})}_{=0} + (\varepsilon_t - \varepsilon_{t-d}).$$

So ∇_d removes seasonality. Whether it also removes the trend depends on the shape of m_t.

Effect on a polynomial trend. If $m_t = a + bt$ is linear, a single seasonal difference kills the trend entirely:

$$\nabla_d m_t = (a + bt) - \big(a + b(t-d)\big) = bd,$$

a constant. But if the trend is quadratic, $m_t = a + bt + ct^2$, then

$$\nabla_d m_t = (bd - cd^2) + 2cd\,t,$$

which still contains a linear-in-t term, so a trend remains.

Remark 2.2: Note

After applying ∇_d, the result $Z_t = \nabla_d x_t$ may or may not still depend on t — it depends entirely on whether the reduced trend $(m_t - m_{t-d})$ is constant.

Remedies for higher-order trends. To fully remove a quadratic trend we can apply any one of the following (each uses two differencing operations in total):

- one ordinary second difference, $\nabla^2 x_t$;
- one seasonal difference followed by one ordinary difference, $\nabla\nabla_d x_t$;
- one second-order seasonal difference, $\nabla_d^2 x_t = x_t - 2x_{t-d} + x_{t-2d}$.

General rule. Each application of ∇ or ∇_d reduces the polynomial degree of the trend by one. A degree-p polynomial trend therefore requires p differencing operations in total, in any combination of ordinary and seasonal differences.

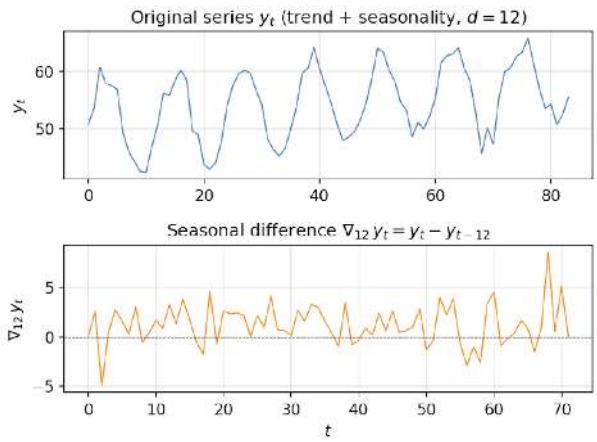


Figure 2.6: Seasonal differencing ($d = 12$) removes annual seasonality, leaving only residual trend (slightly positive if any) and random component.

Differencing vs. Detrending

Detrending and differencing are two ways of reducing a time series to a stationary process (we will see next week).

Pros of differencing over detrending:

- No parameters to estimate; works even for stochastic trends (random walks).
- Higher-order differencing can reduce even very "trendy" series to look like white noise.

Cons of differencing:

- "Washes away" useful features and can introduce more complex autocorrelation / temporal structure in the differenced series.
- The trend component of a series may be of practical interest, and if well estimated may lead to better long-horizon prediction of the series. Discarding it via differencing loses that information.

In practice: unit-root tests (Augmented Dickey–Fuller, KPSS) help determine whether the time series is stationary. If not, detrending or differencing may be appropriate.

2.4 Overview: Time Series Decomposition

Decomposition of time series

Recall that we may consider a time series x_t as generated by

$$x_t = s_t + \varepsilon_t,$$

where

- s_t is a **deterministic** signal/trend,
- ε_t is **stochastic** noise or error, with $\mathbb{E}[\varepsilon_t] = 0$.

This motivates two broad approaches:

1. Applying **classical regression methods** to time series — specify a parametric form for s_t.
2. The **smoothing approach** — estimate s_t non-parametrically.

The process of decomposition and smoothing

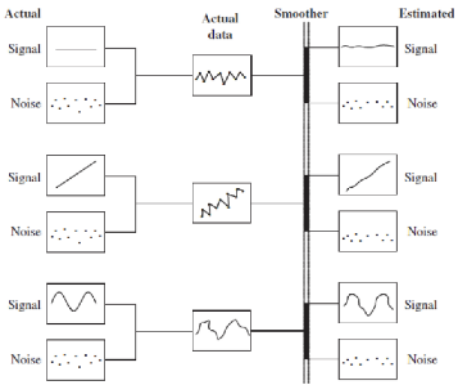


Figure 2.7: A smoother applied to the combined signal+noise recovers the underlying signal.

- **More smoothing**
  - removes noise
  - gives a cleaner trend
  - but loses local structure/details
- **Less smoothing**
  - preserves more information and short-term patterns
  - but may also preserve noise

2.5 Smoothing Approaches

Moving average trend estimation / smoothing

Let $q \geq 0$ be any non-negative integer. Define the $(2q+1)$-point MA:

$$\hat{m}_t = \frac{1}{2q+1} \sum_{j=-q}^{q} x_{t+j}, \qquad q+1 \leq t \leq n-q.$$

Each of the $2q+1$ observations in the window receives weight $\frac{1}{2q+1}$.

Remark 2.3: Remarks

(i) This is an equal-weighted, two-sided MA estimate of trend.
(ii) No trend estimates for the first and last q time points.
(iii) End-point estimates can be recovered using *symmetric padding*.
(iv) MA is a **low-pass filter**: $\text{Var}(\hat{m}_t) < \text{Var}(x_t)$.
(v) This is a **non-causal filter** as it peeks into the future.
(vi) A **causal filter** can be defined by looking at just the q past values.

MA as a low-pass filter

In general, the MA estimate can be written as a filtered series. Clearly, the variance in the smoothed series is less than the observed series.

$$\hat{m}_t = \sum_{j=-\infty}^{\infty} a_j \, x_{t+j}, \qquad a_j = \begin{cases} \dfrac{1}{2q+1} & |j| \le q, \\ 0 & \text{otherwise.} \end{cases}$$

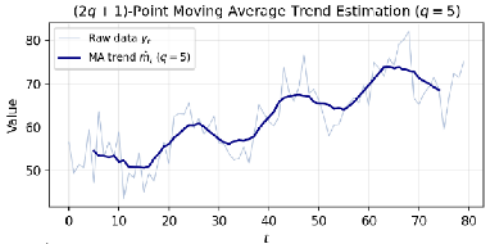


Figure 2.8: $(2q+1)$-point MA trend estimate smooths out high-frequency noise.

Disadvantages of equal-weight MA

1. **Missing endpoints:** no trend estimates for the first and last q time points without padding.
2. **Equal weights:** all $2q+1$ observations in the window receive the same weight — no preference for more recent data.

This motivates the **Exponentially Weighted MA (EWMA) / Simple Exponential Smoothing (SES)**, which addresses the second issue.

Naive forecasts and their interpretation

Introduction: exponential smoothing can be understood as lying between two extreme forecasting rules (naive methods).

For a time series $X_1, \ldots, X_t$, consider:

- **Last value forecast**

$$\hat{x}_{t+1} = x_t.$$

  - Uses only the most recent observation.
  - Corresponds to the **random walk model**: $X_t = X_{t-1} + \varepsilon_t$, $\varepsilon_t \sim \mathrm{WN}(0, \sigma^2)$.
  - Optimal forecast: $\mathbb{E}[X_{t+1} \mid \mathcal{F}_t] = X_t$.
  - Equivalent weights: $\hat{x}_{t+1} = 1 \cdot X_t + 0 \cdot X_{t-1} + \cdots + 0 \cdot X_1$.

- **Sample mean forecast**

$$\hat{x}_{t+1} = \bar{x} = \frac{1}{t} \sum_{j=1}^{t} x_j.$$

  - Uses all past observations equally.
  - Corresponds to the **white noise model**: $X_t \sim \mathrm{WN}(\mu, \sigma^2)$.
  - Optimal forecast: $\mathbb{E}[X_{t+1} \mid \mathcal{F}_t] = \mu \approx \bar{X}$.
  - Equivalent weights: $\hat{X}_{t+1} = \frac{1}{t} X_t + \frac{1}{t} X_{t-1} + \cdots + \frac{1}{t} X_1$.

Remark 2.4: Key Insight

They represent two extremes:

- Random walk → full persistence (recent data dominates).
- White noise → no dependence (all data equally informative).

Exponential smoothing (ES) — the intuition

ES finds a middle ground between two extremes.

- The random walk says "the future is just like the most recent past".
- The mean prediction says "the future is just like the long-run average".

Neither is right in general. ES instead says: "the future is a **weighted average** of everything we have seen, but recent data matters more than old data." The fitted value is a weighted sum of past observations:

$$\hat{x}_{t+1} = \alpha_t x_t + \alpha_{t-1} x_{t-1} + \cdots + \alpha_0 x_0.$$

Desired properties of weights:

- Newer observations should receive *higher* weight.
- Older observations should receive *lower* weight.

Simple exponential smoothing (SES)

Definition 2.7: Simple (First-Order) Exponential Smoothing

$$\hat{x}_{t+1} = \alpha x_t + \alpha(1-\alpha)x_{t-1} + \alpha(1-\alpha)^2 x_{t-2} + \cdots, \qquad 0 < \alpha < 1.$$

The smoothing parameter α is the dial between the two extremes.

1. A high α makes the model reactive and responsive to recent changes, at the cost of being noisy.
2. A low α makes the model stable and smooth, at the cost of being slow to pick up genuine shifts in the series.

The optimal α is learned from the data itself by minimising forecast errors.

Equivalently, the smoothed series is characterised by the **level** ℓ_t:

Definition 2.8: Simple Exponential Smoothing (State-Space Form)

$$\begin{aligned} \text{Forecast equation:} \quad & \hat{x}_{t+1} = \ell_t \\ \text{Smoothing equation:} \quad & \ell_t = \alpha x_t + (1-\alpha)\ell_{t-1} \\ \text{Initial condition:} \quad & \ell_0 \text{ initialised} \end{aligned}$$

The parameters (α, ℓ_0) are estimated by minimising the **one-step-ahead forecast errors**:

$$(\hat{\alpha}, \hat{\ell}_0) = \underset{0\le\alpha\le 1,\, \ell_0}{\arg\min} \sum_{t=1}^{n} \big(x_t - \ell_{t-1}(\alpha, \ell_0)\big)^2.$$

- This is **not** linear regression: the objective is **non-linear in** α because ℓ_{t-1} itself depends on α through the recursion.
- No closed-form solution exists; use **numerical optimisation** (e.g. grid search: fix α, compute ℓ_t recursively, evaluate the SSE, repeat).

Exponentially Weighted Moving Average (EWMA) vs. equal-weight MA

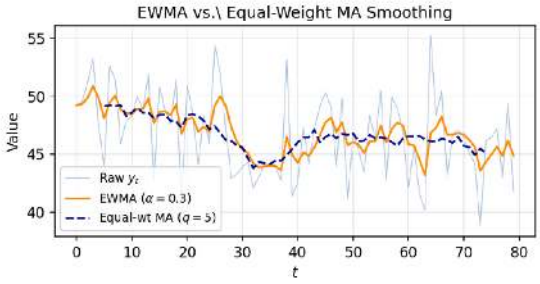


Figure 2.9: EWMA adapts more quickly to local changes; equal-weight MA is much smoother but doesn't capture changes and lags at endpoints.

Shortcomings of SES / EWMA

- The key limitation is that SES only tracks the **current level** of the series.
- If the data is consistently rising or falling (trend), SES will always lag behind.
- It has **no memory of direction / trend, only of position**.

This is what double exponential smoothing fixes. It captures the **trend**.

Double exponential smoothing — the intuition I

Think of double smoothing like describing a moving car:

- To predict where the car will be in five minutes, you need to know not just *where it is now* (the **level**), but also *how fast it is going and in which direction* (the **slope**). SES only tracks position; Holt's method tracks both position and velocity.
- The **level equation** blends the raw new observation with your prior prediction of where the series *should* be right now — that is, the last level plus the last slope.
- The **slope equation** blends what the slope actually turned out to be this period (the change in level) with your prior belief about the slope.
- Both components learn continuously from new data, each at their own learning rate: α for the level and β for the slope.

Double exponential smoothing (Holt's method)

Extends SES to capture a **linear trend** by tracking both a **level** ℓ_t and a **slope** b_t:

Definition 2.9: Holt's Method

$$\begin{aligned}
\text{Forecast equation:} \quad & \hat{x}_{t+h} = \ell_t + h\, b_t \\
& \hat{x}_{t+1} = \ell_t + b_t \\
\text{Level equation:} \quad & \ell_t = \alpha x_t + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\
\text{Trend / Slope equation:} \quad & b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta) b_{t-1}
\end{aligned}$$

- $\alpha \in [0,1]$ controls how quickly the **level** adapts to new observations; $\beta \in [0,1]$ controls how quickly the **slope** adapts.

Double exponential smoothing — the intuition II

- **Forecasting** is then simply extrapolation: start at the current level and project forward in the direction of the current slope.
- The further ahead you forecast, the more the slope matters — and the more uncertainty accumulates. This is why prediction intervals widen as the forecast horizon grows.

Holt's method — multi-step forecasting

The parameters (α, ℓ_0) are estimated by minimising the **one-step-ahead forecast errors**.

- b_0 is the initial estimate of the slope.
- ℓ_0 is the initial estimate of the level.

Remark 2.5: h-step ahead forecast

Forecast equation:

$$\hat{x}_{t+h} = \ell_t + h\, b_t, \qquad \hat{x}_{t+1} = \ell_t + b_t,$$

where ℓ_t is the last smoothed level and b_t is the most updated estimate of the slope.

Also known as **second-order** exponential smoothing or **linear trend** smoothing.

Holt–Winters — the intuition

Holt–Winters adds a **third layer of memory** to Holt's method: the seasonal fingerprint of the series.

- The model maintains a separate **seasonal factor** for each position in the cycle — twelve factors for monthly data with yearly seasonality, seven for daily data with weekly seasonality, and so on. These factors encode the typical deviation from the trend at each point in the cycle, and they are updated each time that position in the cycle comes around again.
- The **level equation** first strips out the current seasonal effect before updating the level, so that seasonal swings do not contaminate the estimate of the underlying trend.
- The **trend equation** stays the same.
- The **seasonal equation** measures the "seasonal surprise" — how much of the current observation is left unexplained after removing the level and trend — and uses this residual to keep the seasonal factors current.

$$\hat{x}_{t+1} = \ell_t + b_t + S_{t+1-p}$$

Triple exponential smoothing (Holt–Winters)

Extends Holt's method to capture both a **linear trend** and **seasonality** of period p, by additionally tracking a **seasonal component** s_t:

Definition 2.10: Holt–Winters Method

$$\begin{aligned}
\text{Forecast equation:} \quad & \hat{x}_{t+1} = \ell_t + b_t + s_{t+1-p} \\
\text{Level equation:} \quad & \ell_t = \alpha(x_t - s_{t-p}) + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\
\text{Trend equation:} \quad & b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta)b_{t-1} \\
\text{Seasonal equation:} \quad & s_t = \gamma(x_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-p}
\end{aligned}$$

- $\alpha \in [0,1]$ governs level adaptation, $\beta \in [0,1]$ governs slope adaptation, and $\gamma \in [0,1]$ governs seasonal adaptation.
- The parameters $(\alpha, \beta, \gamma, \ell_0, b_0, s_0, s_{-1}, \ldots, s_{-p+1})$ are estimated jointly by minimising the sum of squared one-step-ahead forecast errors.

Holt–Winters — multi-step forecasting

For a forecast h steps ahead from the end of the observed series at time t, define

$$k = \left\lfloor \frac{h-1}{p} \right\rfloor,$$

which gives the number of **complete seasonal cycles** that fit between the current time and the forecast horizon. The multi-step forecast is then:

Remark 2.6: h-Step-Ahead Forecast Equation

$$\hat{x}_{t+h} = \ell_t + h\, b_t + s_{t+h-p(k+1)}.$$

- $\ell_t + h\, b_t$ extrapolates the current level forward along the estimated linear trend, exactly as in Holt's method.
- $s_{t+h-p(k+1)}$ looks up the stored seasonal factor for the corresponding position in the cycle at the forecast horizon. The index $t + h - p(k+1)$ always falls within the last observed complete cycle, ensuring we reuse the most recently estimated seasonal factor for that season.

Holt–Winters — the intuition II

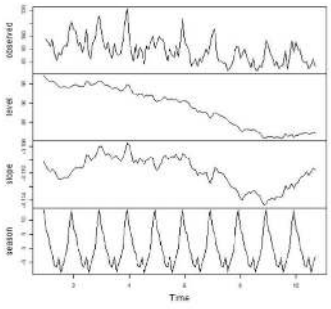
- **Forecasting** projects the trend forward as in Holt's method, then adds back the appropriate stored seasonal factor for the target future time of year.
- This is why Holt–Winters produces forecasts with the correct seasonal shape: it **reads off the learned seasonal pattern from memory** rather than guessing.
- Also known as **third-order** or **trend-plus-seasonality** smoothing.

To forecast at time $t+1$:

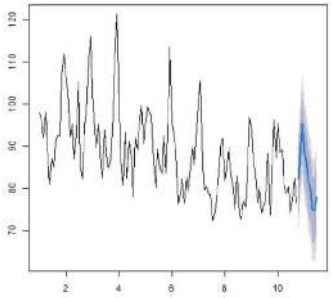
(i) Use the latest available estimate of trend & level, i.e. ℓ_t and b_t.

(ii) Use the latest available estimate of that position/period in the cycle.

- We need to use the position/period corresponding to $t+1$ (i.e. $t+1$ or $t+1-p$ or $t+1-2p, \ldots$).
- We cannot use S_t (although that's the most updated among all seasonal components) because it corresponds to a different position in the cycle.
- We can't use S_{t+1} because we have not yet estimated it based on data up to time t.
- So we use the latest available estimate of that position, which happens to be S_{t+1-p} — this was updated upon observing x_{t+1-p}.



(a) Trend and seasonal decomposition of monthly mortality data.



(b) 10-month ahead forecast with 95% prediction interval.

Figure 2.10: The CI widens as h increases

Another view of exponential smoothing

Two complementary interpretations:

1. **Weighted average view:** In exponential smoothing, forecasts are weighted averages of past observations, weights decaying exponentially as observations get older.
2. **State-space view:** Alternatively, we can view the observed data as combinations of **hidden states** (level, trend, seasonal). The model provides equations for how each unobserved component / state evolves over time.

2.6 ETS State Space Models

Following the state-space view, exponential smoothing is a special case of the **ETS** framework, which considers equations for the following factors / states: **E**rror **T**rend **S**easonality

- The ETS state space models provide a broader framework, and the exponential smoothing model is one example within the framework.
- For the general ETS models, each component can take one of three forms:
  - **N** — None (component absent).
  - **A** — Additive.
  - **M** — Multiplicative.

Remark 2.7: Key Special Cases

ETS(A,N,N)	Simple exponential smoothing
ETS(A,A,N)	Holt's double smoothing
ETS(A,A,A)	Holt–Winters (all additive)
ETS(A,A,M)	Holt–Winters (multiplicative seasonality)

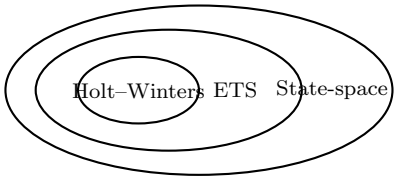


Figure 2.11: Holt–Winters ⊂ ETS ⊂ state-space models.

ETS models — the intuition

The ETS framework asks a more fundamental question than "which smoothing method should I use?": *what is the structural form of the error, trend, and seasonality in this series?* Answering that question determines the model completely.

- **Additive vs. multiplicative seasonal effects.** If monthly sales have a December bump of roughly +500 units every year regardless of how large the business has grown, the seasonal effect is additive — it contributes a fixed amount. But if the December bump is always around 20% of sales, the absolute bump grows with the business, making the seasonal effect multiplicative. Forcing an additive model onto a multiplicative series produces systematically wrong forecasts.
- **Additive vs. multiplicative errors.** Financial returns and prices typically have multiplicative errors — the noise is proportional to the level. Temperature readings or physical measurements often have additive noise of roughly constant absolute size. Choosing the wrong error structure leads to mis-calibrated prediction intervals.

Why use the ETS framework?

1. **Systematic model selection:** compare all plausible ETS combinations via AIC/BIC; pick the best automatically.
2. **Rigorous prediction intervals:** the explicit probability model lets us derive uncertainty bounds formally, not heuristically.
3. **Additive vs. multiplicative:**
   - If seasonal swings are roughly constant in size ⇒ **additive**.
   - If swings grow proportionally with the series level ⇒ **multiplicative**.

Chapter 3

Lecture 3

3.1 Probabilistic Foundations of Time Series

What is a time series?

We have already met a time series informally as a sequence of measurements indexed in time. Before we can do any modeling, we need a more careful probabilistic definition.

Definition 3.1: Univariate Time Series

A *univariate time series* is a sequence of measurements of the **same variable** collected over time. Most often, measurements are made at regular time intervals.

Key distinctions from classical statistics:

- Data are *not* necessarily independent and *not* necessarily identically distributed. (so the i.i.d. assumption from STA257/261 is out the window)
- **Dependence is important** — the whole point of time series analysis is to model and exploit this dependence.
- **The ordering matters** — shuffling the observations destroys the information.

- Ω: sample space.
- $\mathbb{P}$: probability measure, $\mathbb{P} : \Omega \to [0, 1]$, $\omega \mapsto P(\omega)$.
- $\mathcal{F}$: sigma-algebra (a collection of subsets of Ω).

A time series is a sequence

$$\{X_1, X_2, X_3, \ldots\}$$

where each X_t is a random variable. At each time t, X_t has a certain distribution, and the distributions at different times may or may not be the same.

Formal probabilistic definition

The informal definition above is fine for plots, but to do probability we need to recognize that each observation is the realization of a random variable on an underlying probability space.

Definition 3.2: Real-Valued Time Series

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\mathcal{T}$ an index set. A *real-valued time series* is a real-valued function $X(t, \omega)$ defined on $\mathcal{T} \times \Omega$ such that for each fixed t, $X(t, \omega) \equiv X_t$ is a random variable on $(\Omega, \mathcal{F}, P)$.

Note: A time series $\{X_t : t \in \mathcal{T}\}$ is therefore a *collection of random variables*, one for each time point. The single observed sequence $\{x_t\}$ is one realization (one "draw") from this collection.

The L^2 space

Most of the theory in this course — variances, covariances, mean-square convergence — only makes sense when second moments exist. We restrict attention to the space of random variables with finite variance.

Definition 3.3: L^2 Space

A real-valued time series $\{X_t, t \in \mathbb{Z}\}$ is said to be in L^2 if $\mathbb{E}[X_t^2] < \infty$ for each $t \in \mathbb{Z}$. Equivalently,

$$X \in L^2(\Omega, \mathcal{F}, P) \iff \mathbb{E}(X^2) < \infty.$$

- $X \in L^2$ implies $\mathrm{Var}(X_t) < \infty$ for each t. (finite second moment $\Rightarrow$ finite variance)
- In general, we **assume** time series are in L^2 throughout the course.

The joint distribution

A complete probabilistic description of a time series is given by its joint distribution function across all finite collections of time points.

Definition 3.4: Joint Distribution of a Time Series

The joint distribution of a collection of n random variables at arbitrary time points $t_1, t_2, \ldots, t_n$, for any positive integer n, is provided by the joint distribution function evaluated as the probability that the values of the series are jointly less than the n constants $x_1, x_2, \ldots, x_n$; i.e.,

$$F_{t_1,t_2,\ldots,t_n}(x_1, x_2, \ldots, x_n) = P\big(X_{t_1} \leq x_1,\, X_{t_2} \leq x_2,\, \ldots,\, X_{t_n} \leq x_n\big).$$

Unfortunately, these multidimensional distributions cannot be investigated easily **unless** the random variables are *jointly Gaussian*. This motivates working with **summary functions** — mean, autocovariance (ACVF), autocorrelation (ACF) — rather than the full joint distribution. (In Gaussian land, mean and covariance are sufficient statistics, so the summary functions characterize everything anyway.)

Marginal distribution

Looking at one time point at a time gives us the marginal distribution.

Definition 3.5: Marginal Distribution

Let $F_t(x)$ be the marginal distribution function of X_t, defined by

$$F_t(x) = P\{X_t \leq x\}.$$

The corresponding marginal density function is then

$$f_t(x) = \frac{\partial F_t(x)}{\partial x},$$

provided that the density exists.

Mean function

Definition 3.6: Mean Function

The *mean function* of a time series is defined as

$$\mu_{x_t} = \mathbb{E}(X_t) = \int_{-\infty}^{\infty} x\, f_t(x)\, dx,$$

provided it exists, where $\mathbb{E}$ denotes the usual expected value operator.

Note: When no confusion exists about which time series we are referring to, we drop the subscript and write μ_{x_t} as μ_t. The mean function is, in general, **a function of** t — it can drift with time (e.g. the random walk with drift).

3.2 Fundamental Time Series Models

Four fundamental building-block models

Before getting tangled in ARMA, ARIMA, GARCH, etc., it pays to understand the four atomic models from which everything else is built:

1. **White Noise** — pure randomness, no serial dependence.
2. **Moving Averages** — local weighted sums of white noise.
3. **Autoregressions** — present value depends on past values.
4. **Random Walk** — cumulative sum of white noise increments.

These are the atoms from which more complex models (ARMA, ARIMA) are built. Understanding their mean and autocovariance is the foundation of the rest of the course.

White noise

Definition 3.7: White Noise

Suppose $\{w_t,\, t \in \mathbb{Z}\}$ is a sequence of i.i.d. random variables with mean zero and variance $\mathbb{E}[w_t^2] = \sigma_w^2 < \infty$. Then the series $\{w_t\}$ is called a *white noise series*, written $w_t \sim WN(0, \sigma_w^2)$.

Note: w_t being **uncorrelated** (instead of i.i.d.) is sufficient for white noise. The i.i.d. requirement is stronger than needed — uncorrelatedness + zero mean + finite variance is the bare minimum.

Definition 3.8: Gaussian White Noise

If additionally $w_t \sim N(0, \sigma_w^2)$, then w_t is a *Gaussian white noise*, written $w_t \overset{\text{iid}}{\sim} N(0, \sigma_w^2)$.

White noise is the benchmark of "no structure". Any model fitted to a series should reduce its residuals to (approximately) white noise — if structure remains in the residuals, the model has missed something.

Random walk

Definition 3.9: Random Walk with Drift

The series

$$x_t = \delta + x_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim WN(0, \sigma^2)$$

is called a *random walk with drift series*, where δ is the drift.

Unrolling the recursion from x_0:

$$x_t = \delta t + x_0 + \sum_{i=1}^{t} \varepsilon_i.$$

- The process is a **deterministic linear trend** δt plus an *accumulation* of all past shocks.
- Because each shock is **remembered forever**, the variance *grows with time* — the random walk is inherently **non-stationary**. ($\text{Var}(x_t) = t\sigma^2$, which depends on t)

Moving average

A moving average is a tool to smooth the series, and a moving average series is a smoothed series derived from local averaging of the parent series. A moving average of order 3 for the white noise process w_t can be written as

$$v_t = \frac{1}{3}\big(w_{t-1} + w_t + w_{t+1}\big).$$

A more rigorous definition will be given later when we introduce MA(q) models in full generality.

Autoregressive model

An autoregressive series x_t is a series where the present value depends on past values of the series. An example of the autoregressive model (AR of order 2) is

$$x_t = x_{t-1} - 0.9\, x_{t-2} + w_t, \qquad t = 1, 2, \ldots, 500.$$

A more rigorous definition will be given later.

Mean of key examples

Let's flex the mean-function machinery on two of the building blocks.

Example 1 — MA(3):

$$v_t = \tfrac{1}{3}(w_{t-1} + w_t + w_{t+1}), \qquad w_t \overset{\text{iid}}{\sim} N(0, \sigma_w^2).$$

$$\mu_{v_t} = \tfrac{1}{3}\big[\mathbb{E}(w_{t-1}) + \mathbb{E}(w_t) + \mathbb{E}(w_{t+1})\big] = 0.$$

Example 2 — Random Walk (no drift, $x_0 = 0$):

$$x_t = x_{t-1} + w_t, \qquad w_t \overset{\text{iid}}{\sim} N(0, \sigma_w^2).$$

$$\mu_{x_t} = \mathbb{E}(x_{t-1}) + \mathbb{E}(w_t) = \mathbb{E}(x_0) + \sum_{i=1}^{t} \mathbb{E}(w_i) = 0.$$

Remark 3.1: Same mean, different worlds

Both processes have **mean zero**, yet they behave very differently — the MA(3) is bounded and stationary, the random walk drifts arbitrarily far from zero. The distinction lies entirely in their **autocovariance structure**, which is what we look at next.

3.3 Autocovariance and Autocorrelation

Autocovariance function

The mean function is a first-moment summary — it tells us where the series sits on average at each time. The autocovariance is the second-moment summary that captures how the series at one time relates to itself at another time.

Definition 3.10: Autocovariance Function

For a time series x_t, the *autocovariance function* is defined as the second moment product

$$\gamma_x(s,t) = \mathrm{Cov}(x_s, x_t) = \mathbb{E}\big[(x_s - \mu_s)(x_t - \mu_t)\big], \qquad \text{for all } s, t.$$

- $\gamma_x(s,t) = \gamma_x(t,s)$ — **symmetric** in s and t.
- At $s = t$: $\gamma_x(t,t) = \mathrm{Var}(x_t)$. (the autocovariance "at lag 0" is just the variance)

Autocovariance: properties

- The autocovariance measures the **linear dependence** between two points on the same series observed at different times.
- **Smooth series** have autocovariance functions that stay large even when t and s are far apart. **Choppy series** tend to have autocovariance functions that are nearly zero for large separations.
- If $\gamma_x(s,t) = 0$, then x_s and x_t are not *linearly* related, but there may still be some dependence structure between them. If, however, x_s and x_t are **bivariate normal**, then $\gamma_x(s,t) = 0$ ensures their *independence*. (this is the special property of Gaussianity — uncorrelated $\Leftrightarrow$ independent only in the joint Gaussian case)

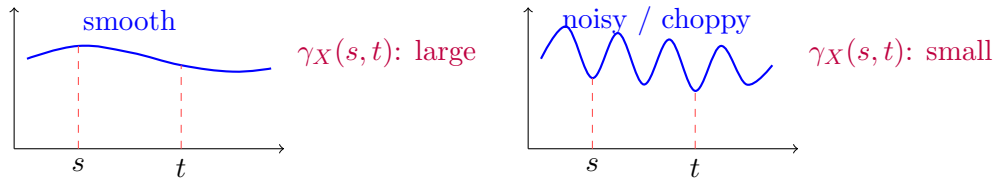


Figure 3.1: Smooth series have large autocovariance between nearby points; choppy series have small autocovariance.

Variance of a time series

Definition 3.11: Variance Function

The *variance function* is defined as

$$\mathrm{Var}(X_t) = \mathbb{E}\big[(X_t - \mu_t)^2\big], \qquad \text{for all } t.$$

Note: For $s = t$, the autocovariance reduces to the variance, because

$$\gamma_x(t,t) = \mathbb{E}\big[(x_t - \mu_t)^2\big] = \mathrm{Var}(x_t).$$

So variance is the $h = 0$ special case of the autocovariance. (one function to rule them all)

Example: Compute the mean and autocovariance function of:

1. White noise sequence w_t
2. The moving average sequence $x_t = w_t + \theta w_{t-1}$
3. The time series $x_t = \cos(2\pi t) + \sin(2\pi t)\, w_t$ (← Exercise)

Note: $w_t \overset{\text{i.i.d.}}{\sim} N(0, \sigma_w^2)$, then $\sin(2\pi t)\, w_t \sim N\big(0, \sin^2(2\pi t)\sigma_w^2\big)$.

1. WN:

$$\begin{aligned} \mathbb{E}[w_t] &= 0 \quad \forall t \\ V(w_t) &= \sigma_w^2 \quad \forall t \\ \gamma_w(t,s) &= \mathrm{Cov}(w_t, w_s) = 0 \quad \forall\, t \neq s \end{aligned}$$

2. MA: $x_t = w_t + \theta w_{t-1}$

$$\mathbb{E}[x_t] = \mathbb{E}[w_t + \theta w_{t-1}] = \mathbb{E}[w_t] + \theta\, \mathbb{E}[w_{t-1}] = 0 + \theta \cdot 0 = 0$$

$$\begin{aligned} V(x_t) &= V(w_t + \theta w_{t-1}) \\ &= V(w_t) + \theta^2 V(w_{t-1}) + 2\,\mathrm{Cov}(w_t, \theta w_{t-1}) \\ &= \sigma_w^2 + \theta^2 \sigma_w^2 + 2\theta\,\mathrm{Cov}(w_t, w_{t-1}) \\ &= \sigma_w^2 + \theta^2 \sigma_w^2 + 2\theta \cdot 0 \\ &= \sigma_w^2 (1 + \theta^2) \end{aligned}$$

$$\begin{aligned}\operatorname{Cov}(x_t, x_{t-1}) &= \operatorname{Cov}(w_t + \theta w_{t-1},\ w_{t-1} + \theta w_{t-2}) \\ &= \operatorname{Cov}(w_t, w_{t-1}) + \operatorname{Cov}(w_t, \theta w_{t-2}) \\ &\quad + \operatorname{Cov}(\theta w_{t-1}, w_{t-1}) + \operatorname{Cov}(\theta w_{t-1}, \theta w_{t-2}) \\ &= 0 + \theta \operatorname{Cov}(w_t, w_{t-2}) + \theta \operatorname{Cov}(w_{t-1}, w_{t-1}) + \theta^2 \operatorname{Cov}(w_{t-1}, w_{t-2}) \\ &= 0 + 0 + \theta \operatorname{Var}(w_{t-1}) + 0 \\ &= \theta \sigma_w^2\end{aligned}$$

$$\operatorname{Cov}(x_t, x_{t-2}) = \operatorname{Cov}(w_t + \theta w_{t-1},\ w_{t-2} + \theta w_{t-3}) = 0$$
since $\operatorname{Cov}(w_s, w_j) = 0\ \ \forall\, s \neq j$, and here all time indices are distinct.

3. Exercise: $x_t = \cos(2\pi t) + \sin(2\pi t)\, w_t$

Mean:

$$\begin{aligned}\mu_t = \mathbb{E}[x_t] &= \mathbb{E}\big[\cos(2\pi t) + \sin(2\pi t)\, w_t\big] \\ &= \cos(2\pi t) + \sin(2\pi t)\, \mathbb{E}[w_t] \\ &= \cos(2\pi t) + \sin(2\pi t) \cdot 0 \\ &= \cos(2\pi t)\end{aligned}$$

Autocovariance: note that $x_t - \mu_t = \sin(2\pi t)\, w_t$.

$$\begin{aligned}\gamma(s,t) = \operatorname{Cov}(x_s, x_t) &= \mathbb{E}\big[(x_s - \mu_s)(x_t - \mu_t)\big] \\ &= \mathbb{E}\big[\sin(2\pi s)\, w_s \cdot \sin(2\pi t)\, w_t\big] \\ &= \sin(2\pi s)\sin(2\pi t)\ \mathbb{E}[w_s w_t]\end{aligned}$$

Case $s = t$: $\mathbb{E}[w_t^2] = \sigma_w^2$, so $\gamma(t,t) = \sin^2(2\pi t)\, \sigma_w^2$.
Case $s \neq t$: $\mathbb{E}[w_s w_t] = 0$, so $\gamma(s,t) = 0$.

$$\gamma(s,t) = \begin{cases} \sin^2(2\pi t)\, \sigma_w^2 & s = t \\ 0 & s \neq t \end{cases}$$

Definition 3.12: Autocorrelation Function (ACF)

The *autocorrelation function* is defined as

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\,\gamma(t,t)}}.$$

Note: The ACF measures the **linear predictability** of the series at time t, say x_t, using only the value x_s.

Properties:

- $-1 \leq \rho(s,t) \leq 1$ always. (it's a correlation — Cauchy–Schwarz guarantees the bound)
- If $x_t = \beta_0 + \beta_1 x_s$ (perfect linear relationship):
  - $\rho(s,t) = +1$ when $\beta_1 > 0$,
  - $\rho(s,t) = -1$ when $\beta_1 < 0$.

Cross-covariance and cross-correlation

So far we have only considered one series at a time. When we have two series and want to measure their joint structure — including possible *leading/lagging* effects — we generalize to the cross-covariance.

Definition 3.13: Cross-Covariance

The *cross-covariance function* between two series x_t and y_t is

$$\gamma_{xy}(s,t) = \mathrm{Cov}(X_s, Y_t) = \mathbb{E}\big[(x_s - \mu_{x_s})(y_t - \mu_{y_t})\big].$$

Definition 3.14: Cross-Correlation Function (CCF)

The scaled version is the *cross-correlation function*:

$$\rho_{xy}(s,t) = \frac{\gamma_{xy}(s,t)}{\sqrt{\gamma_x(s,s)\,\gamma_y(t,t)}}.$$

The CCF measures the linear relationship between x_s and y_t at potentially *different* times — useful for detecting **leading/lagging** relationships between two series. (e.g. does an interest-rate change today predict GDP movements three quarters from now?)

3.4 Stationarity

The concept of stationarity

Intuitively, a process $\{X_t\}$ is **stationary** when:

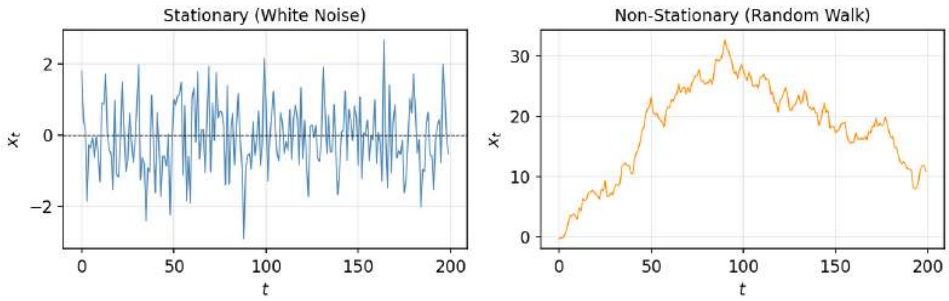
- Its statistical properties **do not change over time**.
- The process is in a state of **statistical equilibrium**.

There are several formal versions of this idea, each making different demands on how much of the distribution is required to be time-invariant. We will cover:

1. Strict (complete) stationarity.
2. Stationarity up to order m.
3. **Weak (covariance / second-order) stationarity** — the main one used in practice in this course.
4. Trend stationarity.

Stationary vs. non-stationary: visual comparison

Before the formal definitions, it helps to know what stationary and non-stationary processes look like in the wild.



re 3: Left: white noise (stationary — constant mean and variance

Figure 3.2: Left: white noise (stationary — constant mean and variance throughout). Right: random walk (non-stationary — variance grows linearly with t).

Strong stationarity

The strongest version of stationarity demands that the entire joint distribution be invariant to time shifts.

Definition 3.15: Strict Stationarity

We say that the time series X_t is *strictly stationary* or *strongly stationary* if its finite-dimensional distributions are shift-invariant. Namely, the probability distribution of

$$\{X_{t_1}, X_{t_2}, \ldots, X_{t_k}\}$$

is identical to that of the time-shifted set

$$\{X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_k+h}\}.$$

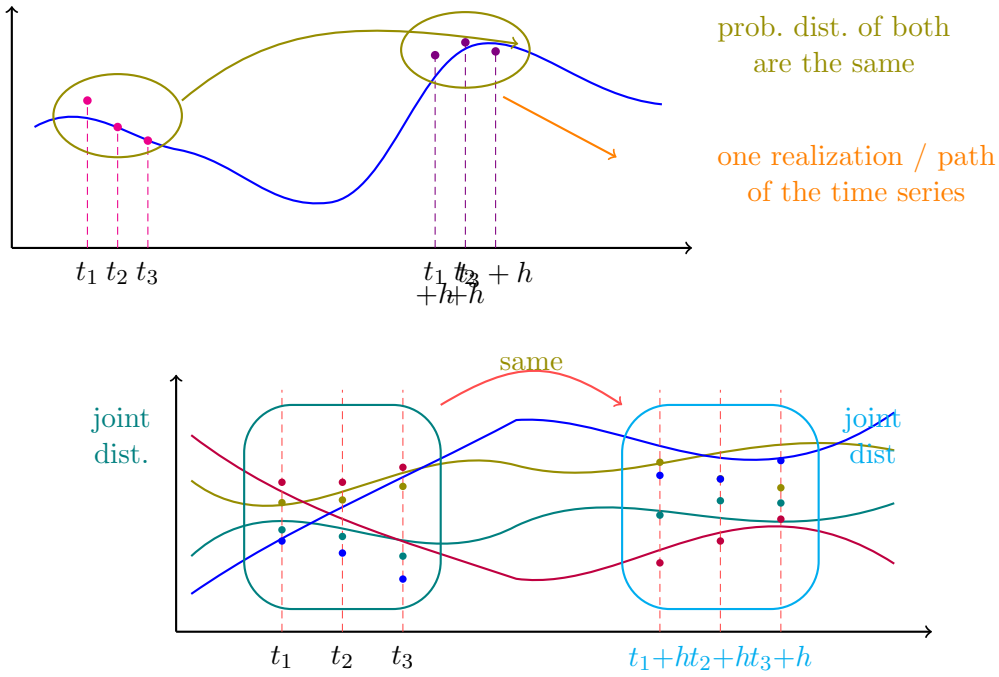
That is,

$$P\{X_{t_1} \leq c_1, \ldots, X_{t_k} \leq c_k\} = P\{X_{t_1+h} \leq c_1, \ldots, X_{t_k+h} \leq c_k\}$$

for all $k = 1, 2, \ldots$, all time points $t_1, t_2, \ldots, t_k$, all numbers $c_1, c_2, \ldots, c_k$, and all time shifts $h = 0, \pm 1, \pm 2, \ldots$.

- This says the distribution of any finite collection is *identical* to the time-shifted collection — complete **time invariance**.
- The marginal distribution of X_{t_i} and X_{t_i+k} are the same for all k — implies the properties such as mean, variance, ACVF, ACF do not change by time shifts.

Reasoning. The joint distribution over t_1, t_2, t_3 equals the joint distribution over t_1+h, t_2+h, t_3+h,



which implies the marginal distribution of X_t is the same for all t. Choosing $k = 1$,

$$P(X_t \leq x) = P(X_{t+h} \leq x) \qquad \forall\, h, t, x$$

$$\Rightarrow \quad X_t \text{ and } X_{t+h} \text{ have the same distribution.}$$

Remark. In practice, we do not observe all realizations of a time series and hence cannot compute the joint distribution. This makes strict stationarity more of a theoretical construct than a practical one.

Chapter 4

Lecture 4

4.1 Stationarity up to Order m

Strict stationarity controls the *entire* joint distribution. A weaker requirement is to control only joint *moments* up to some fixed order.

Definition 4.1: Order-m Stationarity

$\{X_t\}$ is *stationary up to order* m if, for all admissible time points $t_1, \ldots, t_k$, all integers k, and all integers h, every joint moment up to order m exists and is shift-invariant:

$$\mathbb{E}\big(X_{t_1}^{m_1} \cdots X_{t_k}^{m_k}\big) = \mathbb{E}\big(X_{t_1+h}^{m_1} \cdots X_{t_k+h}^{m_k}\big),$$

for all exponents $m_i \geq 0$ with $\sum_{i=1}^{k} m_i \leq m$.

The constraint $\sum m_i \leq m$ controls every mixed moment of total degree at most m. For example, with $m = 3$ the admissible exponent sums are $\sum m_i = 1, 2, 3$.

Implication. Setting $h = -t_1$,

$$\mathbb{E}\big(X_{t_1}^{m_1} X_{t_2}^{m_2}\big) = \mathbb{E}\big(X_{t_1+h}^{m_1} X_{t_2+h}^{m_2}\big) = \mathbb{E}\big(X_0^{m_1} X_{t_2-t_1}^{m_2}\big),$$

so the moment depends only on the **difference** $(t_2 - t_1)$, not on t_1 and t_2 individually. Order-m stationarity also implies order-k stationarity for every $k \leq m$: controlling moments up to degree m automatically controls all lower-degree moments.

Special cases: order 1 and order 2

(i) Mean stationarity (order 1). $\{X_t\}$ is order-1 stationary $\iff$ $\mathbb{E}(X_t)$ exists and $\mathbb{E}(X_t) = \mu$, independent of t.

(ii) Covariance stationarity (order 2). All joint moments up to order 2 exist and:

(a) $\mathbb{E}(X_t) = \mu$, independent of t.
(b) $\mathbb{E}(X_t^2) = \mu_2$, independent of t.
(c) $\mathrm{Var}(X_t) = \mu_2 - \mu^2$, also independent of t.
(d) $\mathrm{Cov}(X_t, X_s)$ is a function of $(s - t)$ only.

Condition (d) follows because $\mathrm{Cov}(X_t, X_s) = \mathbb{E}[X_t X_s] - \mathbb{E}[X_t]\mathbb{E}[X_s] = \mathbb{E}[X_0 X_{s-t}] - \mu^2$, which depends only on $s - t$. Order-2 stationarity is also called **covariance stationarity**, **weak stationarity**, or *stationarity in the wide sense*. The two conditions are **nested**: second-order implies first-order, but not vice versa.

Weak (second-order) stationarity

Definition 4.2: Weak Stationarity

A time series X_t is *weakly stationary* if it is a finite-variance process such that:

(i) the mean $\mathbb{E}(X_t) = \mu_t = \mu$ is **constant**, independent of time t;

(ii) the variance $\mathrm{Var}(X_t) = \sigma^2 < \infty$ is **finite and constant**, independent of t;

(iii) the autocovariance function $\gamma(s, t)$ depends on s and t **only through their difference** $|s - t|$ — and not on their individual values.

Throughout this course, "stationary" means *weakly stationary* unless stated otherwise; this is equivalent to second-order stationarity.

Notation for a stationary series

For a stationary process the mean function is written $\mu_t = \mu$. Since $\gamma(s, t)$ depends only on $|s - t|$, put $s = t + h$, where h is the time shift or **lag**:

$$\gamma(t + h, t) = \mathrm{Cov}(X_{t+h}, X_t) = \mathrm{Cov}(X_h, X_0) = \gamma(h, 0) = \gamma(h).$$

Definition 4.3: Autocovariance Function of a Stationary Series

$$\gamma(h) = \mathrm{Cov}(x_{t+h}, x_t) = \mathbb{E}\big[(x_{t+h} - \mu)(x_t - \mu)\big].$$

Definition 4.4: Autocorrelation Function (ACF) of a Stationary Series

$$\rho(h) = \frac{\gamma(t + h, t)}{\sqrt{\gamma(t + h, t + h)\,\gamma(t, t)}} = \frac{\gamma(h)}{\gamma(0)}.$$

The general normalization is $\rho = \gamma(s, t) \big/ \sqrt{\gamma(t, t)\,\gamma(s, s)}$; under stationarity both terms in the denominator equal $\gamma(0)$.

Properties of the stationary autocovariance

For a stationary series γ satisfies:

1. **Non-negative definite:** for any constants $a_1, \ldots, a_n$,

$$0 \le \mathrm{Var}\left(\sum_{j=1}^{n} a_j X_j\right) = \sum_{j=1}^{n}\sum_{k=1}^{n} a_j a_k\, \gamma(j - k).$$

This is just the variance of a linear combination expanded out: e.g. $\mathrm{Var}(a_1 X_1 + a_2 X_2) = a_1^2 \mathrm{Var}(X_1) + a_2^2 \mathrm{Var}(X_2) + 2a_1 a_2 \mathrm{Cov}(X_1, X_2)$, where each $\gamma(j - k) = \mathrm{Cov}(X_j, X_k)$. The left-hand inequality holds because a variance is $\mathbb{E}[(Y - \mathbb{E}Y)^2]$, the expectation of a square, which

is never negative. So the content of the property is that the matrix $\Gamma_n = [\gamma(j-k)]_{j,k=1}^n$ is a valid covariance matrix: $a^\top \Gamma_n a \geq 0$ for every a.

2. **Bounded by variance:** $|\gamma(h)| \leq \gamma(0) = \mathrm{Var}(x_t)$.
3. **Symmetric (even):** $\gamma(h) = \gamma(-h)$.

Let $\Sigma_{n\times n}$ be a matrix. Σ is n.n.d. if $\forall\, \underline{a} \in \mathbb{R}^n$, we have

$$\underline{a}^\top \Sigma\, \underline{a} \geq 0.$$

Let Σ be the matrix of ACVF:

$$\Sigma_{2\times 2} = \begin{pmatrix} \gamma(0) & \gamma(1) \\ \gamma(1) & \gamma(0) \end{pmatrix}, \qquad \Sigma_{3\times 3} = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \end{array} \begin{array}{c} \begin{array}{ccc} X_1 & X_2 & X_3 \end{array} \\ \begin{pmatrix} \gamma(0) & \gamma(1) & \gamma(2) \\ \gamma(1) & \gamma(0) & \gamma(1) \\ \gamma(2) & \gamma(1) & \gamma(0) \end{pmatrix} \end{array}.$$

As a worked example, take $n = 2$:

$$\begin{aligned} \mathrm{Var}(a_1 X_1 + a_2 X_2) &= \sum_{j=1}^{2}\sum_{k=1}^{2} a_j a_k\, \gamma(j-k) \\ &= \underbrace{a_1 a_1\, \gamma(1-1) + a_1 a_2\, \gamma(1-2)}_{j=1} + \underbrace{a_2 a_1\, \gamma(2-1) + a_2 a_2\, \gamma(2-2)}_{j=2} \\ &= a_1^2\, \gamma(0) + a_2^2\, \gamma(0) + 2 a_1 a_2\, \gamma(1), \qquad \big(\gamma(1) = \gamma(-1)\big). \end{aligned}$$

4.2 Examples and Types of Stationarity

Example 4.1: Stationarity of Standard Models

(1) White noise w_t — *stationary.* $\mathbb{E}(w_t) = 0$ (constant), and $\mathrm{Cov}(w_t, w_s) = 0$ for $t \neq s$, so $\gamma(h) = \sigma_w^2\, \mathbf{1}_{h=0}$: that is, $\gamma(0) = \sigma_w^2$ and $\gamma(h) = 0\ \forall\, h \neq 0$ — a function of lag only.

(2) MA(1) $x_t = w_t + \theta w_{t-1}$, $\theta \in \mathbb{R}$, $w_t \sim \mathrm{WN}(0, \sigma^2)$ — *stationary.* The mean is zero and

$$\mathbb{E}(X_t) = 0, \qquad \mathrm{Cov}(X_{t+h}, X_t) = \begin{cases} \sigma^2(1+\theta^2) & h = 0, \\ \theta\sigma^2 & |h| = 1, \\ 0 & |h| > 1. \end{cases}$$

A function of h only $\Rightarrow$ covariance stationary.

(3) Random walk — *not stationary.* Zero-drift $x_t = x_0 + \sum_{i=1}^t w_i$: first-order stationary (constant mean 0) but *not* second-order, since $\mathrm{Var}(X_t) = t\sigma^2$ grows with t. With drift $x_t = x_0 + \sum_{i=1}^t w_i + \delta t$: then $\mathbb{E}(X_t) = x_0 + \delta t$, so it is neither first- nor second-order stationary.

$$x_t = x_0 + \sum_{i=1}^{t} W_i$$

$$\begin{aligned}\mathrm{Cov}(x_{t+h},\, x_t) &= \mathrm{Cov}\Big(\underbrace{W_1 + W_2 + \cdots + W_t}_{\text{shared}} + \underbrace{W_{t+1} + \cdots + W_{t+h}}_{\text{extra}},\; W_1 + W_2 + \cdots + W_t\Big) \\ &= V(W_1) + V(W_2) + \cdots + V(W_t) + 0 \\ &= t\,\sigma_W^2.\end{aligned}$$

Random walk: non-stationarity illustrated

For the zero-drift random walk $x_t = x_0 + \sum_{i=1}^{t} w_i$ we have $\mathbb{E}(X_t) = 0$, but

$$\mathrm{Cov}(X_{t+h}, X_t) = \min(t+h,\, t)\,\sigma^2 = \begin{cases} t\sigma^2 & h > 0, \\ (t+h)\sigma^2 & h < 0. \end{cases}$$

Either way the covariance depends on t, not just on h $\Rightarrow$ the random walk is **not** covariance stationary.

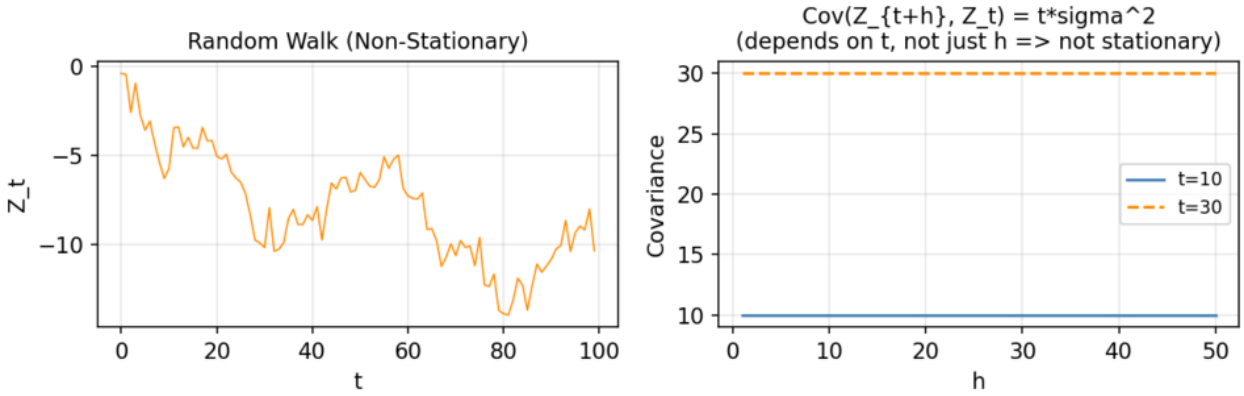


Figure 4.1: Left: a random walk path. Right: $\mathrm{Cov}(Z_{t+h}, Z_t) = t\sigma^2$ depends on t, not just h — confirming non-stationarity.

More examples

(4) Mean-shifted process. Let $\{X_t\}$ be covariance stationary with $\mathbb{E}(X_t) = \mu$. Define

$$Y_t = \begin{cases} X_t, & t \text{ even}, \\ X_t + \alpha, & t \text{ odd}. \end{cases}$$

Then $\mathbb{E}(Y_t)$ alternates between μ and $\mu + \alpha$, a non-constant mean $\Rightarrow$ $\{Y_t\}$ is **not** covariance stationary.

(5) Harmonic process. $X_t = A\cos(\omega t) + B\sin(\omega t)$, where A, B are uncorrelated with $\mathbb{E}(A) = \mathbb{E}(B) = 0$, $\mathrm{Var}(A) = \mathrm{Var}(B) = \sigma^2$, and $\omega \in (-\pi, \pi)$ fixed.

Mean. By linearity,

$$\mathbb{E}(X_t) = \cos(\omega t)\,\mathbb{E}(A) + \sin(\omega t)\,\mathbb{E}(B) = 0.$$

Covariance. Using $\mathrm{Cov}(A, B) = 0$ and $\mathrm{Var}(A) = \mathrm{Var}(B) = \sigma^2$,

$$\begin{aligned}\mathrm{Cov}(X_{t+h}, X_t) &= \mathrm{Cov}\big(A\cos\omega(t+h) + B\sin\omega(t+h),\ A\cos\omega t + B\sin\omega t\big)\\ &= \sigma^2\cos\omega(t+h)\cos\omega t + \sigma^2\sin\omega(t+h)\sin\omega t\\ &= \sigma^2\cos(\omega h),\end{aligned}$$

by the cosine difference identity. This depends only on h, so the process is covariance stationary with $\gamma(h) = \sigma^2\cos(\omega h)$.

Strict vs. weak stationarity

Theorem 4.1: Strict Stationarity Implies Covariance Stationarity

If $\{X_t\}$ is strictly stationary and $\mathbb{E}[X_t^2] < \infty$, then $\{X_t\}$ is also covariance stationary. (Like if $\{X_t\}$ is i.i.d from a distribution but mean isnt finite then $\{X_t\}$ is strict but not covariance)

Proof sketch. For $n = 1$: strict stationarity gives $F_{X_t}(x) = F_{X_{t+k}}(x)$ for all k, so $\mathbb{E}(X_t)$ and $\mathrm{Var}(X_t)$ are time-independent (and $\mathbb{E}[X_t^2] < \infty \Rightarrow \mathrm{Var}(X_t) < \infty$). For $n = 2$: the joint distribution of (X_t, X_s) equals that of (X_{t+h}, X_{s+h}) for all h, so $\mathrm{Cov}(X_t, X_s) = \mathrm{Cov}(X_{t+h}, X_{s+h})$ depends only on the lag $\Rightarrow$ covariance stationary. □

The finite-variance condition is essential. If $X_t \overset{\text{iid}}{\sim} \text{Cauchy}(1)$, then $\{X_t\}$ is strictly stationary, yet $\mathbb{E}[X_t]$ does not exist, so it is not even order-1 (covariance) stationary. The converse of the theorem is also false in general.

Relationships between weak and strong stationarity

1. Strictly stationary $\not\Rightarrow$ weakly stationary (moments may not exist).
2. Strictly stationary + $\mathbb{E}[x_t^2] < \infty$ $\Rightarrow$ weakly stationary.
3. If X_t is weakly stationary *and* a Gaussian process $\Rightarrow$ strictly stationary.
4. If $X_t \notin L^2$ for some t, then not weakly stationary $\Rightarrow$ not strictly stationary (the contrapositive of (2)).

Example 4.2: Weak but Not Strictly Stationary

Let $\{X_t\}$ be a sequence of *independent* random variables with

$$X_t \sim \begin{cases} N(1,1) & t \text{ odd},\\ \mathrm{Exp}(1) & t \text{ even}.\end{cases}$$

Both distributions have mean 1 and variance 1, and $\mathrm{Cov}(X_{t+h}, X_t) = \mathbf{1}_{h=0}$ for all t, so $\{X_t\}$ *is* covariance stationary. But $X_1 \sim N$ and $X_2 \sim \mathrm{Exp}$ have different distributions $\Rightarrow$ $\{X_t\}$ is **not** strictly stationary.

Joint stationarity

Definition 4.5: Joint Stationarity

Two series X_t and Y_t are *jointly stationary* if each is stationary and the cross-covariance is a function of the lag h only:

$$\gamma_{xy}(h) = \mathrm{Cov}(X_{t+h}, Y_t) = \mathbb{E}\big[(X_{t+h} - \mu_x)(Y_t - \mu_y)\big].$$

Definition 4.6: Cross-Correlation Function (CCF)

The CCF of jointly stationary series x_t and y_t is

$$\rho_{xy}(h) = \frac{\gamma_{xy}(h)}{\sqrt{\gamma_x(0)\,\gamma_y(0)}}.$$

Trend stationarity

Consider $X_t = \alpha + \beta t + Y_t$, where Y_t is stationary. The **mean** $\mu_{X_t} = \alpha + \beta t + \mu_y$ depends on t, so X_t is **not stationary** (and not second-order stationary, since $\mathbb{E}[X_t^2] \neq \mathbb{E}[X_{t+h}^2]$). The **autocovariance**, however, is

$$\begin{aligned}\gamma_x(h) = \mathrm{Cov}(X_{t+h}, X_t) &= \mathrm{Cov}\big(\alpha + \beta(t+h) + Y_{t+h},\ \alpha + \beta t + Y_t\big)\\ &= \mathrm{Cov}(Y_{t+h}, Y_t)\\ &= \gamma_y(h).\end{aligned}$$

(constants $\alpha, \beta(t+h), \beta t$ drop out because covariance measures how two random quantities fluctuate together around their means — and constants don't fluctuate.Because covariance measures how two random quantities fluctuate together around their means — and constants don't fluctuate.)a function of the lag h only — the deterministic terms drop out of the covariance. The process has stationary behaviour *around a linear trend*; this is called **trend stationarity**. Removing the trend (by regression or differencing) yields a stationary residual series. (X_t is not second order stat. as $\mathbb{E}[X_t^2] \neq \mathbb{E}[X_{t+h}^2]$)

4.3 Gaussian Processes

Definition 4.7: Multivariate Normal

A p-dimensional random vector $X_{p\times 1} \sim N_p(\mu, \Sigma)$ iff $\alpha' X \sim N_1$ for all $\alpha \in \mathbb{R}^p$, $\alpha \neq 0$. When $\Sigma > 0$ (positive definite),

$$f_X(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\big(-\tfrac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\big)\,.$$

Key properties: any subvector of X is multivariate normal, and every linear combination $\alpha' X$ is univariate normal.

Linear transformations. If $X_{p\times 1} \sim N_p(\mu, \Sigma)$ and $Y_{q\times 1} = A_{q\times p}X + a_{q\times 1}$, then

$$Y \sim N_q\big(a + A\mu,\ A\Sigma A'\big),$$

where Y is $q \times 1$ and $A\Sigma A'$ is $q \times q$; this follows directly from the definition. Every linear combination of *independent* univariate normals is univariate normal, so for independent normals $X_1, \ldots, X_n$ the vector $(X_1, \ldots, X_n)$ is multivariate normal. If the components are *not* independent this can fail: e.g. $X \sim N_1$, $Y = -X$ gives $Y \sim N_1$, yet $X + Y = 0$ is degenerate (not normal with positive variance).

Definition 4.8: Gaussian Time Series

$\{X_t\}$ is a *Gaussian time series* if for all k and all admissible $t_1, \ldots, t_k$ the joint distribution satisfies $(X_{t_1}, \ldots, X_{t_k}) \sim N_k(\mu, \Sigma)$.

Remark 4.1: Gaussian + Covariance Stationary ⇒ Strictly Stationary

If $\{X_t\}$ is Gaussian *and* covariance stationary, then $\{X_t\}$ is also **strictly stationary**. The full probabilistic structure of a Gaussian process is completely determined by its mean function and covariance function — both of which are time-invariant under covariance stationarity.

4.4 Estimation of the ACVF

Why stationarity is critical for estimation

In classical statistics we have many i.i.d. copies of X_t and average across replications. In time series, X_t is a random variable but we observe only **one realisation** — a single realisation cannot by itself recover $\mathbb{E}(X_t)$. Under stationarity, the different time points $x_1, x_2, \ldots, x_n$ all share the same marginal distribution and the same autocovariance structure, so we can **average across time** in place of averaging across replications. Without stationarity each X_t has a different distribution and there is no principled way to pool information across time.

Sample mean

Definition 4.9: Sample Mean

Under the assumption that the series is stationary, the mean function $\mu_t = \mu$ is constant and is estimated by the sample mean

$$\bar{X} = \frac{1}{n} \sum_{t=1}^{n} X_t.$$

$\bar{X}$ is an **unbiased** estimator: $\mathbb{E}(\bar{X}) = \mu$.

Variance of the sample mean

Theorem 4.2: Variance of the Sample Mean

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{t=1}^{n} X_t\right) = \frac{1}{n^2}\mathrm{Cov}\left(\sum_{t=1}^{n} X_t, \sum_{s=1}^{n} X_s\right) = \frac{1}{n}\sum_{h=-n}^{n}\left(1 - \frac{|h|}{n}\right)\gamma_x(h).$$

The variance of $\bar{X}$ is *not* σ^2/n as in the i.i.d. case — there, $\mathrm{Var}\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n^2}\sum \mathrm{Var}(X_i) = \sigma^2/n$. That value emerges here only when $\gamma(h) = 0$ for all $h \neq 0$ (white noise). The triangular weights

$1 - |h|/n$ shrink as $|h|$ increases. Positive serial correlation **inflates** $\mathrm{Var}(\bar{X})$ relative to the i.i.d. case.
$h = 1$: 9 terms $(X_1, X_2), (X_2, X_3), \ldots, (X_9, X_{10})$

Sample ACVF: two estimators

Observe $(X_1, \ldots, X_n)$; the true ACVF is $\gamma_h = \mathrm{Cov}(X_{t+h}, X_t)$.

Case 1 — μ known.

$$\hat{\gamma}^*_\mu(h) = \frac{1}{n-h}\sum_{t=1}^{n-h}(X_t - \mu)(X_{t+h} - \mu) \quad \Rightarrow \quad \text{unbiased},$$

since $\mathbb{E}[\hat{\gamma}^*_\mu(h)] = \frac{1}{n-h}\sum_{t=1}^{n-h}\mathbb{E}[(X_t - \mu)(X_{t+h} - \mu)] = \frac{1}{n-h}(n-h)\gamma_h = \gamma_h$. With divisor n instead,

$$\hat{\gamma}_\mu(h) = \frac{1}{n}\sum_{t=1}^{n-h}(X_t - \mu)(X_{t+h} - \mu) \quad \Rightarrow \quad \mathbb{E}[\hat{\gamma}_\mu(h)] = \frac{n-h}{n}\,\gamma_h,$$

which is asymptotically unbiased since $(n-h)/n \to 1$ as $n \to \infty$.

Case 2 — μ unknown (replace with $\bar{X}_n$).

$$\hat{\gamma}^*(h) = \frac{1}{n-h}\sum_{t=1}^{n-h}(X_t - \bar{X}_n)(X_{t+h} - \bar{X}_n), \qquad \hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(X_t - \bar{X}_n)(X_{t+h} - \bar{X}_n).$$

Both are biased for small samples but consistent.

Why prefer divisor n over $n - h$?

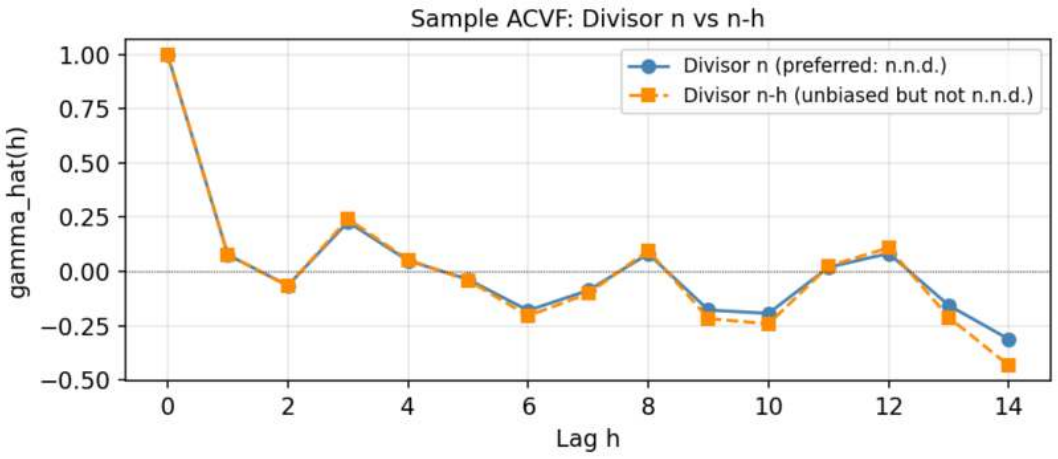


Figure 4.2: Sample ACVF with divisor n vs. divisor $n - h$ for white noise ($n = 50$).

The divisor in $\hat{\gamma}$ is n (not $n - h$) to **guarantee that the resulting autocovariance matrix is non-negative definite**. The divisor $n - h$ does not ensure this, even though it gives an unbiased estimator.

Sample (empirical) autocorrelation

Definition 4.10: Sample Autocovariance

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_t - \bar{x}_n)(x_{t+h} - \bar{x}_n).$$

Definition 4.11: Sample Autocorrelation (ACF)

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}.$$

Both the sample autocovariance and autocorrelation functions are **even**: $\hat{\gamma}(-h) = \hat{\gamma}(h)$ and $\hat{\rho}(-h) = \hat{\rho}(h)$.

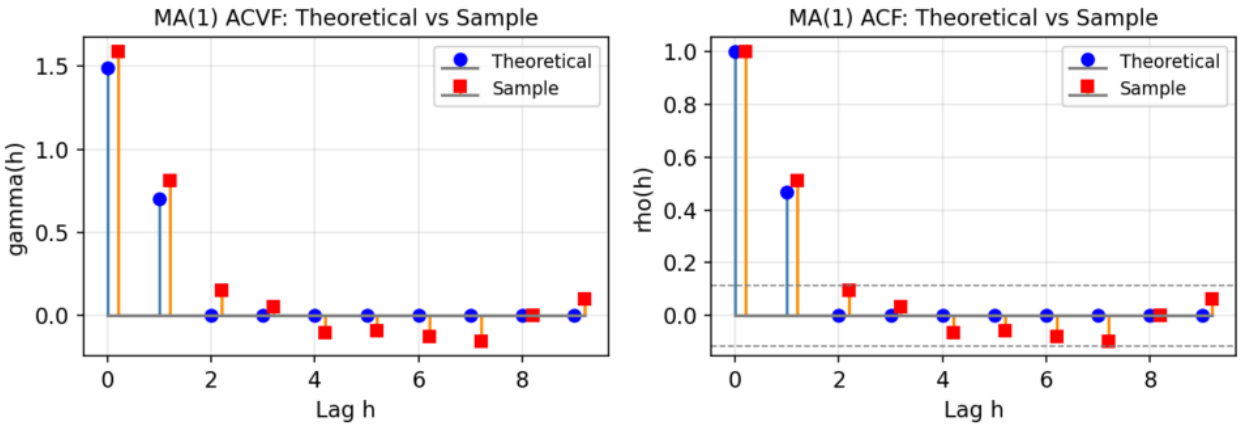


Figure 4.3: MA(1) $Y_t = X_t + 0.7X_{t-1}$ ($n = 300$): theoretical vs. sample ACVF (left) and ACF (right). The ACVF cuts off sharply after lag 1.

The plot of $\hat{\rho}(h)$ against h is called a **correlogram**.

4.5 Large-Sample Properties and ACF Plots

Theorem 4.3: Large-Sample Distribution of $\hat{\rho}(h)$

Under general conditions, if X_t is white noise then for large n the sample ACF $\hat{\rho}_x(h)$, for $h = 1, 2, \ldots, H$ (any fixed H), is approximately normally distributed with zero mean and standard deviation $1/\sqrt{n}$:

$$\hat{\rho}_x(h) \overset{\text{approx}}{\sim} N\left(0, \tfrac{1}{n}\right).$$

Practical significance test. A lag h is deemed statistically significant if

$$|\hat{\rho}(h)| > \frac{2}{\sqrt{n}}.$$

This is the dashed band shown in every correlogram. The "2" is the rounded normal quantile — the exact 95% interval is $\left(-1.96/\sqrt{n},\ 1.96/\sqrt{n}\right)$. For a true white-noise series, approximately 95% of sample ACF values should lie *within* these bounds.

Using the ACF bounds in practice

The significance bounds $\pm 2/\sqrt{n}$ have three main roles:

1. **Diagnostic checking.** Many modelling procedures aim to reduce a time series to white-noise residuals (e.g. fitting ARMA models). After fitting, plot the ACF of the residuals — if the model is adequate, nearly all bars should fall within $\pm 2/\sqrt{n}$.
2. **Model identification.** The pattern of significant lags guides the choice of model order. An MA(q) process has an ACF that cuts off after lag q.
3. **Rough rule.** For white noise of length n, about 95% of the $\hat{\rho}(h)$ values lie within $\pm 2/\sqrt{n}$ by chance alone — so 1 or 2 marginally significant lags out of 20 is not alarming.

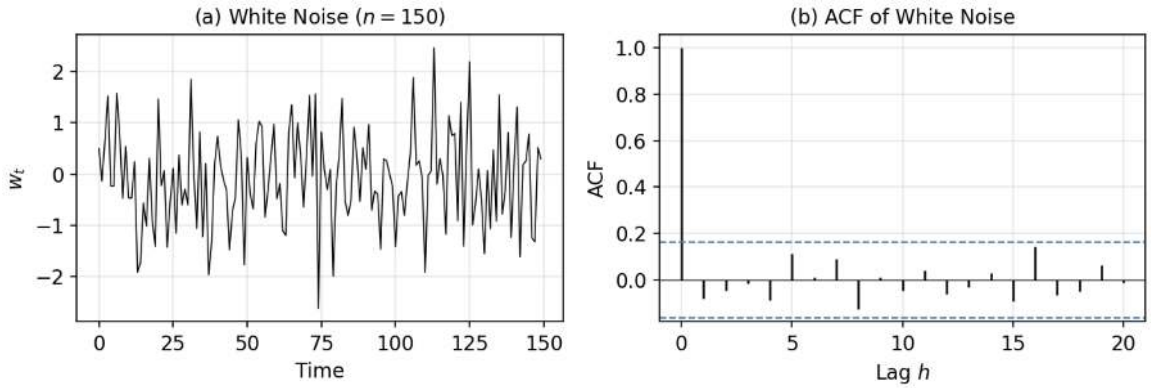


Figure 4.4: White noise series ($n = 150$) and its sample ACF. Dashed lines mark $\pm 2/\sqrt{n}$. All lags fall within the band: no structure.

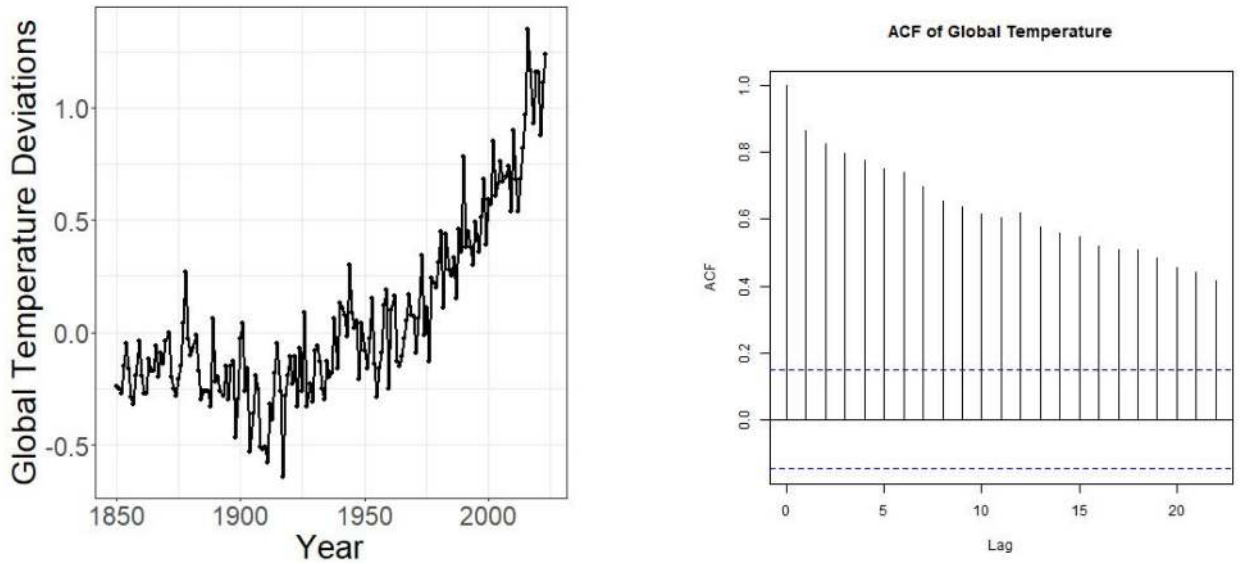


Figure 4.5: Global temperature deviations (trending, non-stationary) and its ACF. A slowly decaying ACF is a hallmark of non-stationarity or trend.

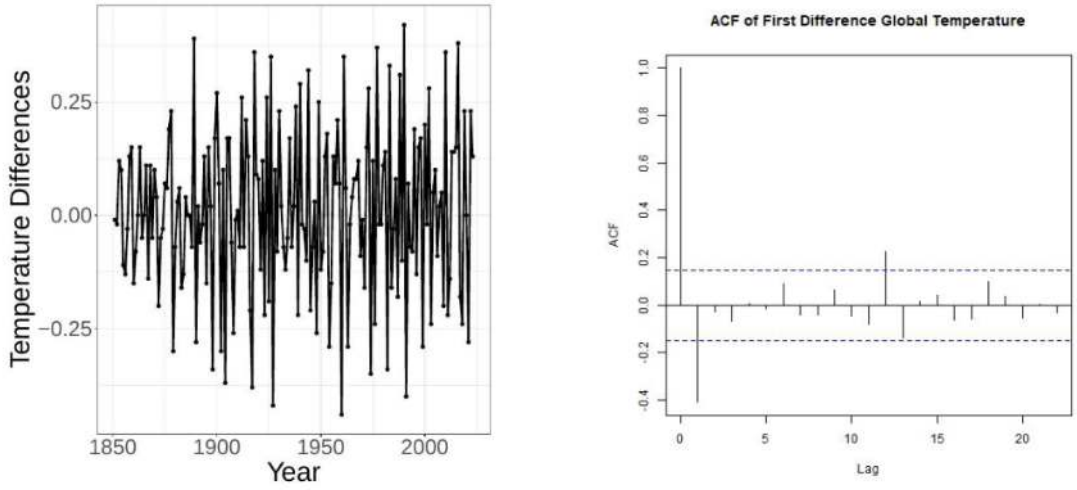


Figure 4.6: First-differenced temperature series and its ACF. After differencing, the ACF drops inside $\pm 2/\sqrt{n}$ beyond the first lag, consistent with a near-white-noise residual.

The lag-1 value $\hat{\rho}(1)$ still sits slightly outside the band, so the differenced series is probably *not* pure white noise — a further difference could be considered.

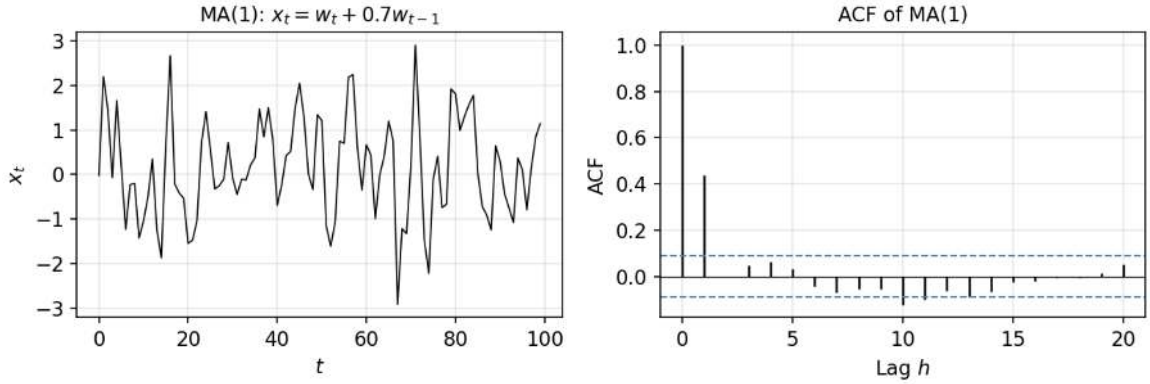


Figure 4.7: MA(1) series $x_t = w_t + 0.7w_{t-1}$ and its sample ACF. The ACF cuts off sharply after lag 1 — a diagnostic signature of a moving-average process.

Chapter 5

Lecture 5

5.1 Autoregressive (AR) Processes

The ARMA family of models

Over the next several lectures we study a hierarchy of models that are built on top of one another:

1. **AR(p)** — autoregressive of order p;
2. **MA(q)** — moving average of order q;
3. **ARMA(p, q)** — autoregressive moving average;
4. **ARIMA(p, d, q)** — autoregressive integrated moving average (covered time permitting).

Each model in this family is a building block for the next, so understanding AR and MA processes thoroughly is essential for everything that follows.

What is an autoregressive model?

An autoregressive model is based on the idea that the current value of a series, x_t, can be explained as a function of its own past p values, where p determines the number of steps into the past needed to forecast the current value:

$$x_t = f(x_{t-1}, x_{t-2}, \ldots, x_{t-p}) + \text{noise}.$$

The intuition mirrors ordinary regression: just as regressing Y on covariates $X_1, X_2, \ldots$ uses *other* variables to explain Y, an autoregression uses the series' *own past* to explain its present.

A simple example is the AR(2) model

$$x_t = x_{t-1} - 0.9\, x_{t-2} + w_t, \qquad w_t \overset{\text{iid}}{\sim} N(0, \sigma_w^2).$$

The current value depends on the two immediately preceding values, plus a fresh shock w_t.

When does autoregressive modelling make sense?

AR models are natural when the current value depends systematically on its own past, leading to persistence, oscillation, or mean-reverting behaviour. Real-world examples include: **economics and finance** (stock returns, interest rates, GDP growth, exchange rates — though AR is not always the best choice here); **epidemiology** (daily infection counts depend on how many were infected in the preceding days); **meteorology** (tomorrow's temperature is best predicted by today's

— autocorrelation is strong); **engineering and signal processing** (speech signals, vibration and electrical measurements all exhibit strong short-term autocorrelation); and **ecology** (animal population dynamics, where this year's population depends on last year's through birth, death, and competition).

Definition 5.1: AR(p) Model

An *autoregressive model of order* p, written AR(p), is

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + w_t,$$

where $w_t \sim \text{WN}(0, \sigma_w^2)$ and $\phi_1, \ldots, \phi_p$ are constants with $\phi_p \neq 0$. We ideally want $\mathbb{E}(x_t) = 0$ for all t, and for x_t to be stationary and causal.

The interpretation of the pieces is as follows. The order p is the **memory** of the model — how many past values directly influence the present. Each coefficient ϕ_j measures the **direct linear effect** of x_{t-j} on x_t, holding all other lags fixed. The shock w_t is a fresh, unpredictable innovation at each time step.

AR(p) with a non-zero mean

If the mean μ of x_t is not zero, set $z_t = x_t - \mu$. Because

$$x_t - \mu = \phi_1(x_{t-1} - \mu) + \phi_2(x_{t-2} - \mu) + \cdots + \phi_p(x_{t-p} - \mu) + w_t,$$

we obtain the **non-zero mean AR(p) model**

$$z_t = \alpha + \phi_1 z_{t-1} + \phi_2 z_{t-2} + \cdots + \phi_p z_{t-p} + w_t, \qquad \alpha = \mu\,(1 - \phi_1 - \phi_2 - \cdots - \phi_p).$$

Working with $z_t = x_t - \mu$ (the zero-mean version) is the standard approach: all theory developed for the zero-mean case applies directly.

Remark 5.1: Where the intercept comes from (annotation)

Suppose we start instead from an AR(p) written *with* a constant term, $x_t = \alpha + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t$. Assuming x_t is stationary with mean μ, take expectations on both sides:

$$\mathbb{E}(x_t) = \mu = \alpha + \phi_1\mu + \cdots + \phi_p\mu + 0 \quad \Longrightarrow \quad \alpha = \mu\,(1 - \phi_1 - \phi_2 - \cdots - \phi_p).$$

Substituting this α back and rearranging recovers $x_t - \mu = \phi_1(x_{t-1} - \mu) + \cdots + \phi_p(x_{t-p} - \mu) + w_t$, so with $z_t = x_t - \mu$ we get the clean zero-mean recursion $z_t = \phi_1 z_{t-1} + \cdots + \phi_p z_{t-p} + w_t$. In particular, if the model genuinely has no intercept ($\alpha = 0$) and $1 - \sum \phi_j \neq 0$, then $\mu = 0$.

The backshift operator

Definition 5.2: Backshift Operator B

The *backshift operator* B is defined by

$$Bx_t = x_{t-1},$$

and is extended to powers by $B^2x_t = B(Bx_t) = Bx_{t-1} = x_{t-2}$, and in general $B^kx_t = x_{t-k}$.

The backshift operator is a formal algebraic tool. We treat polynomials in B as if B were a scalar, which lets the AR model be written compactly and analysed using polynomial algebra.

Remark 5.2: Differencing in backshift notation (annotation)

The same operator expresses the differencing operations from earlier lectures. The d-th order seasonal-style difference applied k times can be written

$$\nabla_d^k = \left(1 - B^d\right)^k,$$

i.e. a polynomial in B. This is the reason differencing and AR analysis share the same algebraic machinery.

AR(p) in backshift operator form

Applying the backshift operator to the AR(p) definition (using $x_{t-1} = Bx_t$, $x_{t-2} = B^2x_t$, …, $x_{t-p} = B^px_t$),

$$x_t = \phi_1 Bx_t + \phi_2 B^2x_t + \cdots + \phi_p B^px_t + w_t,$$

and rearranging gives

$$\underbrace{\left(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p\right)}_{\phi(B)} x_t = w_t, \qquad \text{or simply} \qquad \phi(B)\, x_t = w_t.$$

Definition 5.3: AR Polynomial

The polynomial

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$

is called the *AR polynomial* (or autoregressive operator, or characteristic polynomial).

Does x_t depend on shocks before yesterday?

The AR(p) model is written in terms of the lagged values $x_{t-1}, \ldots, x_{t-p}$ and the latest shock w_t — not in terms of past shocks. So does a shock from two weeks ago still matter today? Yes, and here is why. Take the AR(1) process $x_t = \phi_1 x_{t-1} + w_t$ and substitute repeatedly:

$$x_t = \phi_1 x_{t-1} + w_t = \phi_1(\phi_1 x_{t-2} + w_{t-1}) + w_t = \phi_1^2(\phi_1 x_{t-3} + w_{t-2}) + \phi_1 w_{t-1} + w_t = \cdots$$

Continuing recursively, x_t depends on **every** past shock $w_{t-1}, w_{t-2}, w_{t-3}, \ldots$, not just the most recent one. The weight on shock w_{t-j} is ϕ_1^j, which decays geometrically as j increases (provided

$|\phi_1| < 1$). Recent shocks therefore dominate, but older shocks still contribute — their influence simply fades rather than switching off abruptly.

Remark 5.3: Reading the dependence off the recursion (annotation)

The recursion makes the covariance structure explicit: because x_t contains w_{t-j} with coefficient ϕ_1^j, we have $\mathrm{Cov}(x_t, w_{t-j}) \neq 0$ for every $j \geq 0$. The current value is genuinely informed by the entire history of innovations $w_t, w_{t-1}, w_{t-2}, \ldots, w_1$.

AR versus white noise

This is exactly what distinguishes AR models from white noise: the process has **infinite memory**, but that memory **decays**. The rate of decay is controlled by ϕ — a larger $|\phi|$ means slower forgetting and a more persistent series.

Remark 5.4: Contrast with exponential smoothing (annotation)

This looks superficially similar to exponential smoothing, but the two are not the same. In exponential smoothing the forecast is a weighted combination of lagged values of x_t *itself*,

$$\hat{x}_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \cdots,$$

whereas in the AR linear-process representation the decaying weights multiply the *shocks* w_{t-j}, not the observations.

5.2 Linear Processes

Definition 5.4: Linear Process

A *linear process* x_t is a linear combination of white-noise variates w_t:

$$x_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j \, w_{t-j}, \qquad \text{with} \qquad \sum_{j=-\infty}^{\infty} |\psi_j| < \infty.$$

Its autocovariance function is

$$\gamma_x(h) = \sigma_w^2 \sum_{j=-\infty}^{\infty} \psi_{j+h} \, \psi_j, \qquad h \geq 0.$$

The condition $\sum |\psi_j| < \infty$ is called **absolute summability** of the coefficients. It is exactly what guarantees the infinite sum is well behaved.

Remark 5.5: Why absolute summability gives finite variance (annotation)

We want $\mathrm{Var}(x_t) = \left(\sum_j \psi_j^2\right)\sigma_w^2 < \infty$. We are *given* $\sum_j |\psi_j| < \infty$, which in particular forces

$\sup_j |\psi_j| < \infty$. Let $M := \sup_j |\psi_j| < \infty$. Then

$$\sum_{j=-\infty}^{\infty} \psi_j^2 = \sum_{j=-\infty}^{\infty} |\psi_j| \cdot |\psi_j| \leq \sum_{j=-\infty}^{\infty} |\psi_j| \cdot M = M \sum_{j=-\infty}^{\infty} |\psi_j| < \infty.$$

Hence $\mathrm{Var}(x_t) < \infty$, and it is a constant (it does not depend on t). Absolute summability of the ψ_j is therefore the technical hypothesis that makes the linear process a legitimate, finite-variance stationary series.

Why this matters. The stationary solution of any AR(p) process turns out to be a linear process — it expresses x_t as a (one-sided) infinite weighted sum of past shocks.

Remark 5.6: Why rewrite an AR model as a linear process? (annotation)

Suppose we try to compute $\mathrm{Var}(x_t)$ directly from the AR(1) recursion $x_t = \phi_1 x_{t-1} + w_t$. Using $\mathrm{Cov}(x_{t-1}, w_t) = 0$,

$$\begin{aligned}
\mathrm{Var}(x_t) &= \mathrm{Var}(\phi_1 x_{t-1} + w_t) \\
&= \phi_1^2 \,\mathrm{Var}(x_{t-1}) + \sigma_w^2 \\
&= \phi_1^2 \,\mathrm{Var}(\phi_1 x_{t-2} + w_{t-1}) + \sigma_w^2 \\
&= \phi_1^4 \,\mathrm{Var}(x_{t-2}) + \phi_1^2 \sigma_w^2 + \sigma_w^2 \\
&= \cdots
\end{aligned}$$

The right-hand side never closes — the variance keeps referring back to itself. This makes the recursive form awkward to work with: the variance cannot be read off easily. Rewriting x_t as a linear process $\sum_j \psi_j w_{t-j}$ in *independent* shocks fixes this, since the variance then becomes a simple convergent sum.

Example: the AR(1) process

Example 5.1: AR(1) as a linear process — the problem

Let $x_t = \phi x_{t-1} + w_t$ be an AR(1) process with $w_t \sim \mathrm{WN}(0, \sigma_w^2)$ and $|\phi| < 1$. Show that:

1. $x_t = \sum_{j=0}^{\infty} \phi^j w_{t-j}$ (i.e. x_t is a linear process);
2. the autocovariance function is $\gamma(h) = \dfrac{\sigma_w^2 \, \phi^h}{1 - \phi^2}$ for $h \geq 0$;
3. the autocorrelation function is $\rho(h) = \phi^h$ for $h \geq 0$;
4. x_t is stationary.

Part 1: x_t as a linear process. Starting from $x_t = \phi x_{t-1} + w_t$ and back-substituting,

$$\begin{aligned} x_t &= \phi x_{t-1} + w_t \\ &= \phi(\phi x_{t-2} + w_{t-1}) + w_t \\ &= \phi^2(\phi x_{t-3} + w_{t-2}) + \phi w_{t-1} + w_t \\ &= \phi^3(\phi x_{t-4} + w_{t-3}) + \phi^2 w_{t-2} + \phi w_{t-1} + w_t \\ &\vdots \\ &= \phi^k x_{t-k} + \sum_{j=0}^{k-1} \phi^j w_{t-j}. \end{aligned}$$

As $k \to \infty$: since $|\phi| < 1$, we have $|\phi|^k \to 0$, so the leading term vanishes and

$$x_t = \sum_{j=0}^{\infty} \phi^j w_{t-j}.$$

This is a **causal** linear process with coefficients $\psi_j = \phi^j$ — note that ψ_j is defined for $j \geq 0$ and we set $\psi_j = 0$ for $j < 0$, which is what makes the representation one-sided.

Remark 5.7: Variance from the linear-process form (annotation)

With the representation in hand the variance is now immediate. Because the w_{t-j} are uncorrelated,

$$\mathrm{Var}(x_t) = \sigma_w^2 \sum_{j=0}^{\infty} \phi^{2j} = \frac{\sigma_w^2}{1-\phi^2} < \infty,$$

where the geometric series converges precisely because $|\phi| < 1$. Compare this with the stalled recursion in Remark 5.6 — the linear-process form delivers the answer in one line.

Part 2: autocovariance function. Using the linear-process representation $x_t = \sum_{j=0}^{\infty} \phi^j w_{t-j}$,

$$\gamma(h) = \mathrm{Cov}(x_{t+h}, x_t) = \mathrm{Cov}\left(\sum_{j=0}^{\infty} \phi^j w_{t+h-j},\ \sum_{k=0}^{\infty} \phi^k w_{t-k} \right) = \sigma_w^2 \sum_{j=0}^{\infty} \phi^{j+h}\phi^j = \sigma_w^2\, \phi^h . \frac{1}{1-\phi^2} = \frac{\sigma_w^2\, \phi^h}{1-\phi^2}.$$

The series converges because $|\phi| < 1$, so $\phi^{2j} \to 0$ geometrically.

Remark 5.8: The covariance expansion, term by term (annotation)

It is worth seeing why only the "matched" shock terms survive. Writing both sums out,

$$\mathrm{Cov}\Big(w_{t+h} + \phi w_{t+h-1} + \phi^2 w_{t+h-2} + \cdots,\ w_t + \phi w_{t-1} + \phi^2 w_{t-2} + \cdots \Big).$$

Two shock terms contribute only when their time indices coincide, since $\mathrm{Cov}(w_a, w_b) = 0$ for $a \neq b$ and $\mathrm{Cov}(w_a, w_a) = \sigma_w^2$. Pairing w_{t+h-j} in the first sum with w_{t-k} in the second forces $t + h - j = t - k$, i.e. $j = h + k$. The surviving terms are

$$\phi^h \,\mathrm{Cov}(w_t) + \phi^{h+2} \,\mathrm{Cov}(w_{t-1}) + \phi^{h+4} \,\mathrm{Cov}(w_{t-2}) + \cdots = \sigma_w^2\, \phi^h \big(1 + \phi^2 + \phi^4 + \cdots\big) = \phi^h\, \sigma_w^2\, \frac{1}{1-\phi^2}.$$

Part 3: autocorrelation function. Directly from the definition $\rho(h) = \gamma(h)/\gamma(0)$, with $\gamma(0) = \sigma_w^2/(1-\phi^2)$ and $\gamma(h) = \sigma_w^2\phi^h/(1-\phi^2)$,

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \frac{\sigma_w^2\,\phi^h/(1-\phi^2)}{\sigma_w^2/(1-\phi^2)} = \phi^h.$$

Remark 5.9: Shape of the AR(1) ACF

The ACF of an AR(1) decays *geometrically* with the lag h. For $\phi > 0$ the decay is monotone; for $\phi < 0$ it alternates in sign. Crucially, the ACF never cuts off to exactly zero at any finite lag — this is the key diagnostic difference from MA processes, whose ACF cuts off sharply.

Part 4: stationarity. We verify the conditions for weak stationarity when $|\phi| < 1$.

- **Constant mean:** $\mathbb{E}(x_t) = \mathbb{E}\big(\sum_{j=0}^{\infty}\phi^j w_{t-j}\big) = \sum_{j=0}^{\infty}\phi^j\mathbb{E}(w_{t-j}) = 0$, which is constant in t.
- **ACVF depends only on the lag:** from Part 2, $\gamma(h) = \sigma_w^2\phi^h/(1-\phi^2)$ depends only on h, not on the absolute time t.
- **Finite variance:** $\gamma(0) = \sigma_w^2/(1-\phi^2) < \infty$ since $|\phi| < 1$.

All three conditions hold, so x_t is weakly stationary.

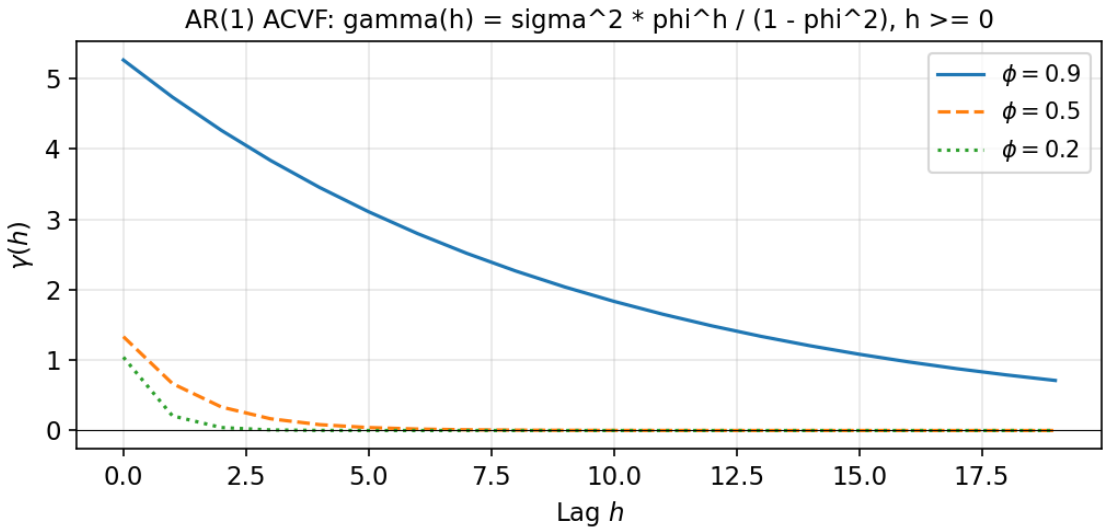


Figure 5.1: AR(1) autocovariance $\gamma(h) = \sigma_w^2\phi^h/(1-\phi^2)$ for three values of ϕ. A larger ϕ means slower decay — the series has longer memory.

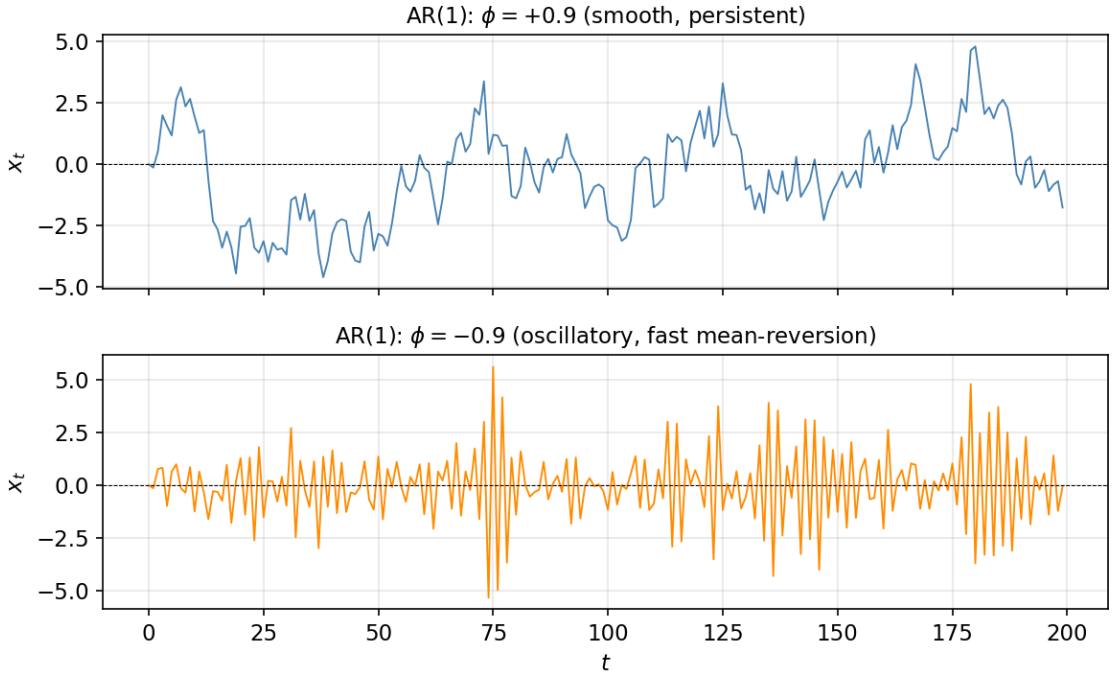


Figure 5.2: Simulated AR(1) trajectories. With $\phi = +0.9$ (top) the series is smooth and persistent with slow mean-reversion; with $\phi = -0.9$ (bottom) it is oscillatory, alternating sign each period, with fast mean-reversion.

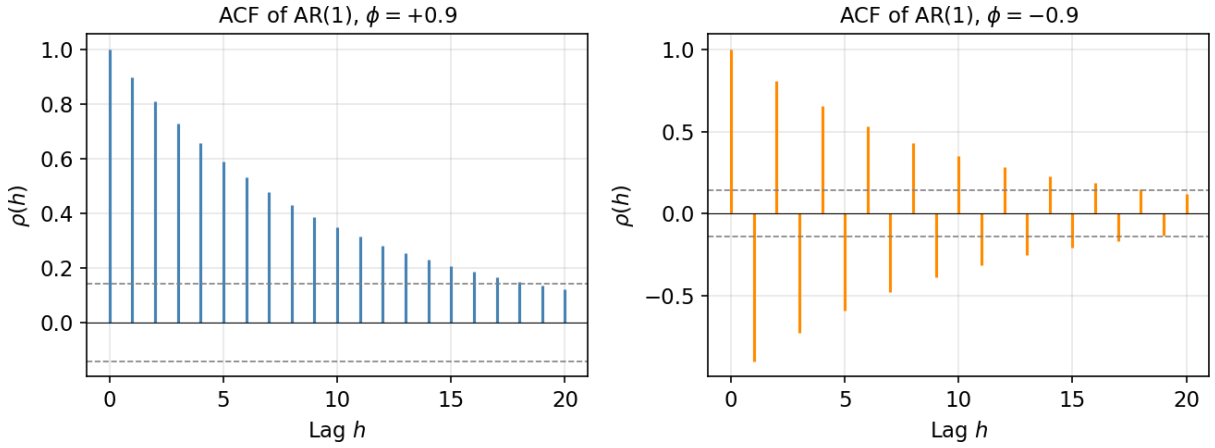


Figure 5.3: Sample ACF of the two AR(1) processes. For $\phi = +0.9$: positive, monotone decay $\rho(h) = 0.9^h$. For $\phi = -0.9$: alternating signs, $\rho(h) = (-0.9)^h$. In both cases the ACF never cuts off abruptly.

5.3 The Stationary Solution

Definition 5.5: Stationary Solution

The *stationary solution* of a time series model is the steady-state (equilibrium) representation of $\{x_t\}$ that

- is consistent with the original stochastic equation governing the process;
- exhibits time-invariant statistical properties (mean, variance, autocovariance);
- is independent of the initial conditions (e.g. x_0).

Intuition. After a sufficiently long time, the influence of the initial state of the series becomes negligible. The series "forgets" where it started and settles into a stable probabilistic regime whose statistics do not change over time.

Finding the stationary solution of AR(p)

The back-substitution trick of Section 2 works neatly for AR(1) but not for a general AR(p). The general approach is **matching coefficients**. Since $\phi(B)x_t = w_t$, we seek a solution of the linear-process form

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j} = \psi(B)\, w_t.$$

Substituting into $\phi(B)x_t = w_t$ gives

$$\phi(B)\,\psi(B)\,w_t = w_t \quad \Longrightarrow \quad \phi(B)\,\psi(B) = 1.$$

Expanding both polynomials and matching the coefficients of $B^0, B^1, B^2, \ldots$ yields the ψ_j recursively. Equivalently,

$$\psi(B) = \phi^{-1}(B),$$

i.e. we treat the backshift operator as a complex number and invert the polynomial.
For the AR(1) case this inversion is the familiar geometric series. Treating B as a complex variable z,

$$\psi(B) = \phi^{-1}(B) = 1 + \phi B + \phi^2 B^2 + \cdots + \phi^j B^j + \cdots, \qquad \phi^{-1}(z) = \frac{1}{1-\phi z} = \sum_{j=0}^{\infty} \phi^j z^j, \quad |z| \le 1,$$

which reproduces $\psi_j = \phi^j$ — exactly the coefficients found by back-substitution.

5.4 Causality and the Characteristic Polynomial

Explosive AR models

Two boundary cases show that stationarity can fail or behave strangely.

The **random walk** ($\phi = 1$): $x_t = x_{t-1} + w_t$. Its autocovariance is

$$\gamma_x(s,t) = \min\{s,t\}\,\sigma_w^2,$$

which depends on t (not just on $|s-t|$), so the random walk is **not stationary**.

The **explosive AR(1)** ($|\phi| > 1$): $x_t = \phi x_{t-1} + w_t$. Each period the previous value is amplified by $|\phi| > 1$ before the shock is added, so the values grow rapidly in magnitude — hence "explosive".

When stationarity is not enough: causality

Surprisingly, an explosive AR(1) *can* be stationary — but only via a representation that depends on *future* shocks. Rewrite the explosive AR(1) by solving forward:

$$x_t = \frac{1}{\phi} x_{t+1} - \frac{1}{\phi} w_{t+1}.$$

Since $|\phi|^{-1} < 1$, iterating this forward gives the stationary but future-dependent solution

$$x_t = -\sum_{j=1}^{\infty} \phi^{-j}\, w_{t+j}.$$

This is stationary, but it uses future innovations — which is useless for forecasting. When a process does *not* depend on the future (as the AR(1) with $|\phi| < 1$ does not), we call it **causal**. The explosive case is stationary but future-dependent, and therefore not causal.

Definition 5.6: Causality

A time series x_t is *causal* if it can be written as a convergent one-sided sum of past and present shocks only:

$$x_t = \sum_{j=0}^{\infty} \psi_j\, w_{t-j} = \psi(B)\, w_t, \qquad \sum_{j=0}^{\infty} |\psi_j| < \infty.$$

Intuition. The current state of the series can be reconstructed entirely from past innovations; no knowledge of future shocks is needed. This is a natural requirement for any model that will be used for forecasting.

Remark 5.10: Stationarity and causality of AR(1) — summary

Causality and stationarity are *distinct* conditions. For the AR(1) process $x_t = \phi x_{t-1} + w_t$:

- $|\phi| < 1$: the series is **stationary and causal**, with stationary solution $x_t = \sum_{j=0}^{\infty} \phi^j w_{t-j}$;
- $|\phi| = 1$: the series is a **random walk** — not stationary;
- $|\phi| > 1$: the series is **stationary but not causal** (its solution is future-dependent), and explosive in the forward direction.

Practical convention. In time series analysis we work exclusively with causal models: stationarity without causality is useless for forecasting. By default, we will treat a series as "stationary" only if it is also causal — and for AR(1) this means $|\phi| < 1$.

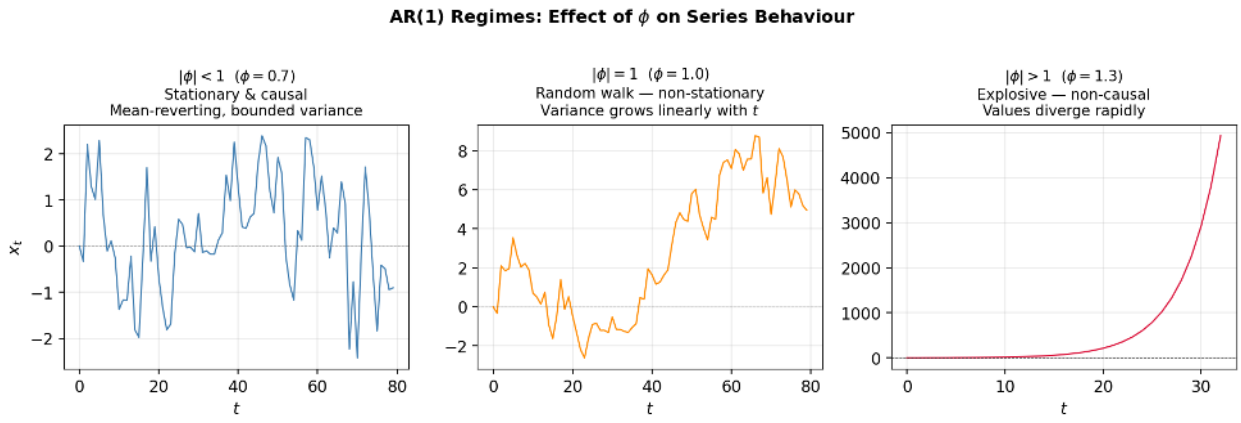


Figure 5.4: The three regimes of AR(1). Left ($|\phi| < 1$): stationary, causal, mean-reverting with bounded variance. Centre ($|\phi| = 1$): random walk, variance grows without bound. Right ($|\phi| > 1$): explosive growth in the forward direction.

The characteristic polynomial

The stationarity and causality conditions for a general AR(p) model are expressed most cleanly through the roots of a polynomial.

Definition 5.7: Characteristic Polynomial

For the AR(p) model $\phi(B)x_t = w_t$, the *characteristic polynomial* is obtained by treating the backshift operator B as a complex variable z:

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p.$$

The *characteristic equation* is $\phi(z) = 0$, and its solutions $z_1, z_2, \ldots, z_p \in \mathbb{C}$ are called the *roots of the AR polynomial*.

Theorem 5.1: Causal Stationarity of AR(p)

The AR(p) process $\phi(B)x_t = w_t$ has a *causal stationary solution* if and only if all roots of the characteristic polynomial $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ lie strictly outside the unit circle:

$$\phi(z) = 0 \implies |z_i| > 1, \qquad i = 1, \ldots, p.$$

In this case the stationary solution is

$$x_t = \sum_{j=0}^{\infty} \psi_j \, w_{t-j}, \qquad \psi(z) = \phi^{-1}(z) = \sum_{j=0}^{\infty} \psi_j z^j, \quad |z| \leq 1.$$

Intuition. The roots encode how the AR polynomial "inverts". If any root lies on or inside the unit circle, $\phi^{-1}(z)$ fails to converge for $|z| \leq 1$, and no causal stationary solution exists.

Stationarity via roots: the AR(1) case

For AR(1), $\phi(z) = 1 - \phi z$, so the characteristic equation $1 - \phi z = 0$ has the single root $z_0 = 1/\phi$. Reading off the three cases:

- $|\phi| < 1$: the root $z_0 = 1/\phi$ satisfies $|z_0| > 1$ — outside the unit circle — so the process is **causal and stationary**;
- $|\phi| = 1$: the root $z_0 = 1$ lies *on* the unit circle — a **unit root**, giving a non-stationary random walk;
- $|\phi| > 1$: the root $z_0 = 1/\phi$ satisfies $|z_0| < 1$ — inside the unit circle — so the process is **not causal** (explosive in the forward direction).

The AR(1) analysis of the earlier sections is thus a special case of the general rule: *all roots of $\phi(z)$ must lie strictly outside the unit circle.*

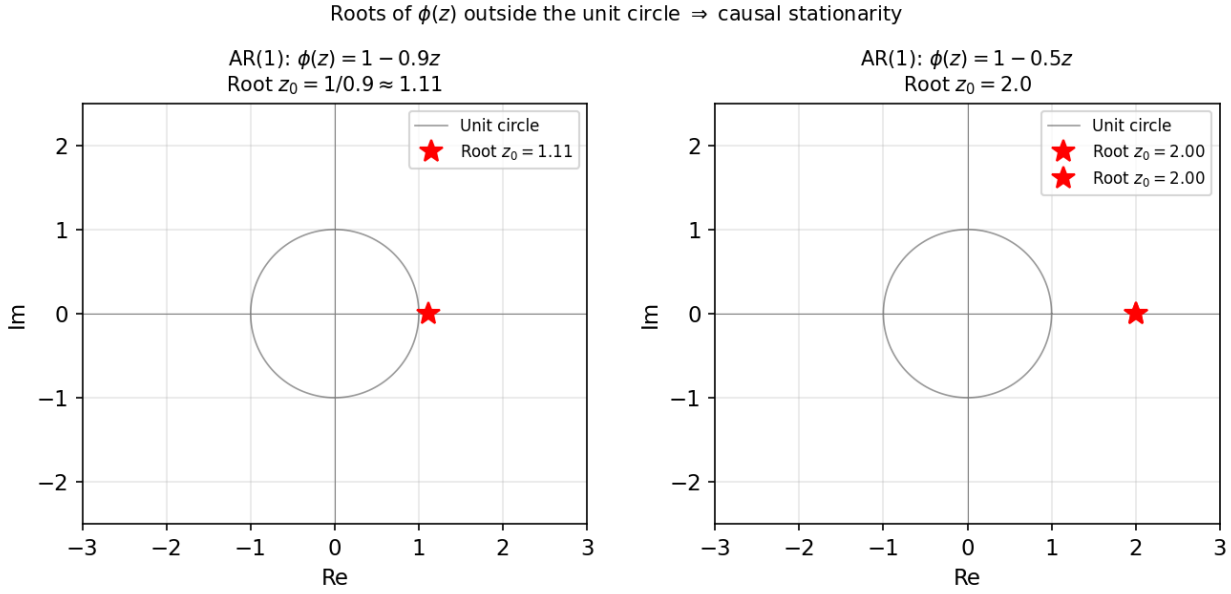


Figure 5.5: Roots of $\phi(z)$ for two AR(1) models. A root lying outside the unit circle indicates causal stationarity.

Worked examples

Example 5.2: Stationarity of AR(p) via roots

Example A. AR(2): $x_t = 1.5x_{t-1} - 0.56x_{t-2} + w_t$. The characteristic polynomial is $\phi(z) = 1 - 1.5z + 0.56z^2$, with roots $z = 1/0.7 \approx 1.43$ and $z = 1/0.8 = 1.25$. Both satisfy $|z_j| > 1$, so the process is **causal and stationary**.

Example B. AR(2): $x_t = x_{t-1} - 0.25x_{t-2} + w_t$. Here $\phi(z) = 1 - z + 0.25z^2 = (1 - 0.5z)^2$, a repeated root at $z = 2$ (outside the unit circle), so the process is **causal and stationary**.

Example C. Random walk: $x_t = x_{t-1} + w_t$. Here $\phi(z) = 1 - z$, with root $z = 1$ lying *on* the unit circle (a unit root), so the process is **not stationary**.

Remark 5.11: Refresher: complex numbers

A complex number $z \in \mathbb{C}$ is written $z = a + bi$ with $a, b \in \mathbb{R}$ and $i = \sqrt{-1}$. Here $a = \mathrm{Re}(z)$ is the real part, $b = \mathrm{Im}(z)$ is the imaginary part, and $\bar{z} = a - bi$ is the complex conjugate. The *modulus* is the distance from the origin in the complex plane,

$$|z| = \sqrt{a^2 + b^2} = \sqrt{z\,\bar{z}}.$$

Key facts: $|z| = 1$ means z lies on the unit circle, $|z| > 1$ means z lies outside it, and the complex roots of a real polynomial always come in conjugate pairs $z, \bar{z}$ with $|z| = |\bar{z}|$.

Example 5.3: AR(2) with complex roots

Consider the model $x_t = x_{t-1} - 0.8\,x_{t-2} + w_t$, i.e. $\phi_1 = 1.0$, $\phi_2 = -0.8$.

Step 1 — characteristic polynomial: $\phi(z) = 1 - z + 0.8z^2$.

Step 2 — solve $\phi(z) = 0$:

$$0.8z^2 - z + 1 = 0 \quad \Longrightarrow \quad z = \frac{1 \pm \sqrt{1 - 3.2}}{1.6} = \frac{1 \pm \sqrt{-2.2}}{1.6} = \frac{1 \pm i\sqrt{2.2}}{1.6}.$$

Step 3 — compute the moduli:

$$|z_{1,2}| = \frac{\sqrt{1^2 + (\sqrt{2.2})^2}}{1.6} = \frac{\sqrt{1 + 2.2}}{1.6} = \frac{\sqrt{3.2}}{1.6} = \frac{1.789}{1.6} \approx 1.118.$$

Step 4 — stationarity check: since $|z_{1,2}| \approx 1.118 > 1$, both roots lie outside the unit circle, so the process is **causal and stationary**.

AR stationarity: summary

For the AR(1) model $x_t = \phi x_{t-1} + w_t$ with root $z_0 = 1/\phi$:

Condition	Root location	Property
$\lvert\phi\rvert < 1$	$\lvert z_0\rvert > 1$: outside circle	Causal & stationary
$\lvert\phi\rvert = 1$	$\lvert z_0\rvert = 1$: on circle	Unit root; non-stationary
$\lvert\phi\rvert > 1$	$\lvert z_0\rvert < 1$: inside circle	Stationary, not causal

For the general AR(p) model $\phi(B)x_t = w_t$ with roots $z_1, \ldots, z_p$:

Root condition	Property
All $\lvert z_j\rvert > 1$ (all outside the unit circle)	Causal & stationary
Any $\lvert z_j\rvert = 1$ (a unit root is present)	Not stationary
Any $\lvert z_j\rvert < 1$ (a root inside the unit circle)	Stationary, not causal

5.5 The AR(1) Autocovariance, Step by Step

In Section 5.2 we obtained the AR(1) autocovariance by first writing x_t as a linear process and then summing. There is a second, more direct route that avoids the linear-process representation

entirely: work straight from the recursion $x_t = \phi x_{t-1} + w_t$, multiplying by lagged values and taking expectations. This is the method that generalises to AR(p) through the *Yule–Walker equations*, so it is worth seeing carefully in the AR(1) case first.

Setting up the computation

Recall the AR(1) model and assume $|\phi| < 1$, so the process is stationary:

$$x_t = \phi x_{t-1} + w_t, \qquad w_t \sim \mathrm{WN}(0, \sigma_w^2), \qquad |\phi| < 1.$$

Remark 5.12: The key assumption on the innovations

The whole computation rests on one fact:

$$\mathrm{Cov}(w_t,\, x_{t-j}) = 0 \qquad \text{for all } j > 0.$$

That is, the current shock is uncorrelated with *all past values* of the series. This holds because x_{t-j} is built only from $w_{t-j}, w_{t-j-1}, \ldots$ — innovations dated strictly before time t — and distinct white-noise terms are uncorrelated.

First, the mean. Taking expectations of both sides and using stationarity ($\mathbb{E}(x_t) = \mathbb{E}(x_{t-1}) = \mu$),

$$\mathbb{E}(x_t) = \phi\, \mathbb{E}(x_{t-1}) + \mathbb{E}(w_t) \quad \Longrightarrow \quad \mu = \phi\mu + 0 \quad \Longrightarrow \quad \mu = 0 \qquad (\text{since } \phi \neq 1).$$

Working with this zero-mean series simplifies the autocovariance to a plain product moment:

$$\gamma(h) = \mathrm{Cov}(x_{t+h}, x_t) = \mathbb{E}[x_{t+h}x_t] - \mathbb{E}[x_{t+h}]\mathbb{E}[x_t] = \mathbb{E}[x_{t+h}x_t].$$

The variance $\gamma(0)$

Square both sides of $x_t = \phi x_{t-1} + w_t$ and take expectations:

$$\mathbb{E}[x_t^2] = \mathbb{E}\big[(\phi x_{t-1} + w_t)^2\big] = \phi^2\, \mathbb{E}[x_{t-1}^2] + 2\phi\, \mathbb{E}[x_{t-1}w_t] + \mathbb{E}[w_t^2].$$

By the key assumption $\mathbb{E}[x_{t-1}w_t] = 0$, and $\mathbb{E}[w_t^2] = \sigma_w^2$. By stationarity $\mathbb{E}[x_t^2] = \mathbb{E}[x_{t-1}^2] = \gamma(0)$, so

$$\gamma(0) = \phi^2\gamma(0) + \sigma_w^2 \quad \Longrightarrow \quad \gamma(0)\,(1-\phi^2) = \sigma_w^2 \quad \Longrightarrow \quad \gamma(0) = \frac{\sigma_w^2}{1-\phi^2}.$$

This is finite and positive because $|\phi| < 1$ gives $1 - \phi^2 > 0$.

The autocovariance $\gamma(h)$ for $h > 0$

Multiply the model at time $t + h$ by x_t and take expectations:

$$\mathrm{Cov}(x_{t+h}, x_t) = \mathbb{E}[x_{t+h}x_t] = \phi\, \mathbb{E}[x_{t+h-1}x_t] + \mathbb{E}[w_{t+h}x_t].$$

Since $h > 0$, the shock w_{t+h} occurs strictly *after* time t, so x_t is already determined before w_{t+h} arrives; hence $\mathbb{E}[w_{t+h}x_t] = 0$. This leaves the **Yule–Walker recursion**

$$\gamma(h) = \phi\, \gamma(h-1), \qquad h > 0.$$

Iterating from the base case $\gamma(0)$,

$$\gamma(1) = \phi\,\gamma(0), \qquad \gamma(2) = \phi\,\gamma(1) = \phi^2\gamma(0), \qquad \ldots, \qquad \gamma(h) = \phi^h\gamma(0).$$

Substituting $\gamma(0)$ and using the symmetry $\gamma(-h) = \gamma(h)$ gives the autocovariance for every lag:

$$\gamma(h) = \frac{\sigma_w^2\,\phi^{|h|}}{1-\phi^2}, \qquad h \in \mathbb{Z}.$$

Remark 5.13: Two routes, one answer

This matches the result of the linear-process calculation in Section 5.2 exactly. The two derivations are complementary: the linear-process route makes the structure $x_t = \sum_j \phi^j w_{t-j}$ explicit, while the recursion route is shorter and — crucially — generalises directly to AR(p).

The autocorrelation function and its shape

Dividing by $\gamma(0)$ gives the autocorrelation function

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \phi^{|h|}.$$

The shape depends on the sign of ϕ. For $\phi > 0$ we have $\rho(h) > 0$ at every lag, and the ACF decays monotonically to zero — positive correlation at all lags. For $\phi < 0$ the ACF $\rho(h) = (-|\phi|)^h$ alternates in sign: even lags are positively correlated, odd lags negatively, and the series oscillates. In both cases $|\rho(h)| \to 0$ as $h \to \infty$ — the AR(1) ACF always *tails off* and never cuts off to exactly zero at a finite lag.

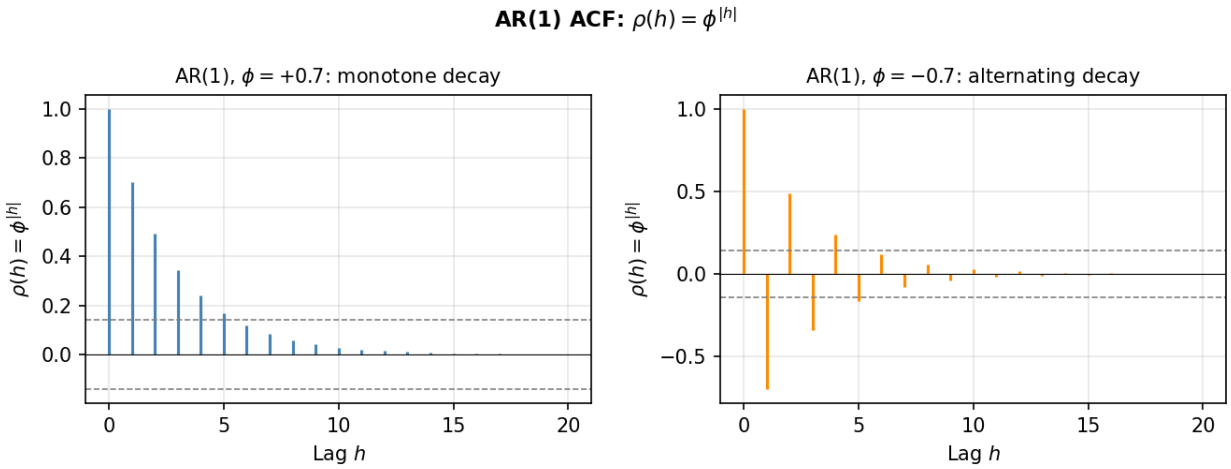


Figure 5.6: AR(1) ACF $\rho(h) = \phi^{|h|}$. Left ($\phi = +0.7$): monotone decay, positive at every lag. Right ($\phi = -0.7$): alternating decay. In neither case does the ACF cut off abruptly.

5.6 Yule–Walker Equations: AR(2) and AR(p)

The recursion route of the previous section is not special to AR(1). Applied to a general AR(p) it produces the *Yule–Walker equations*, a linear system linking the AR coefficients to the autocorrelations.

The AR(2) model and its stationarity region

Definition 5.8: AR(2) Model

The AR(2) model is

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t, \qquad \phi_2 \neq 0, \qquad w_t \sim \mathrm{WN}(0, \sigma_w^2).$$

Stationarity requires all roots of $\phi(z) = 1 - \phi_1 z - \phi_2 z^2$ to lie strictly outside the unit circle. For AR(2) this root condition is equivalent to three simple inequalities on the coefficients:

$$1 - \phi_1 - \phi_2 > 0, \qquad 1 + \phi_1 - \phi_2 > 0, \qquad |\phi_2| < 1.$$

The set of (ϕ_1, ϕ_2) satisfying all three is a triangle — the **AR(2) stationarity region**. Points outside the triangle correspond to non-stationary or explosive processes.

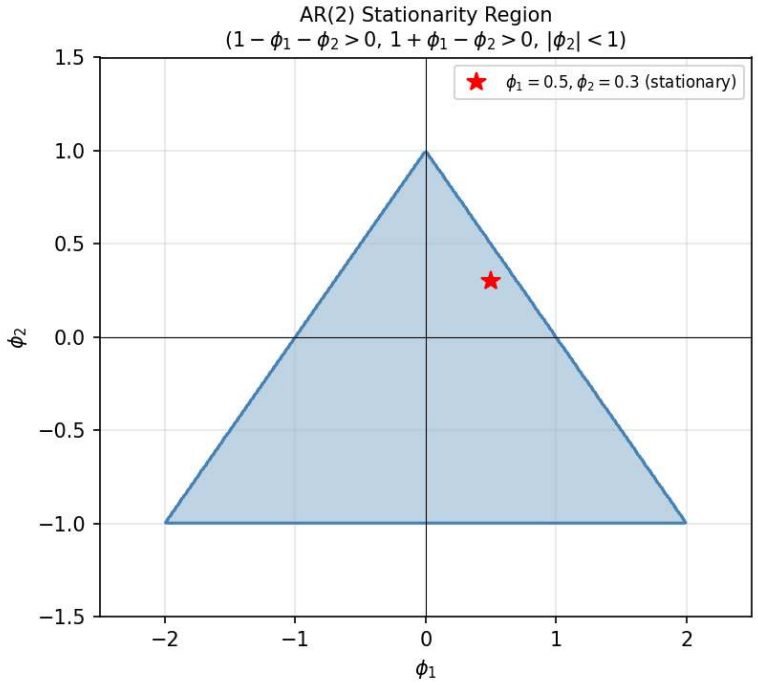


Figure 5.7: The AR(2) stationarity region: the triangle of (ϕ_1, ϕ_2) pairs satisfying $1 - \phi_1 - \phi_2 > 0$, $1 + \phi_1 - \phi_2 > 0$, and $|\phi_2| < 1$. The marked point $(\phi_1, \phi_2) = (0.5, 0.3)$ lies inside, so that process is stationary.

Mean and the Yule–Walker recursion

Taking expectations of the AR(2) equation, $\mu = \phi_1\mu + \phi_2\mu + 0$, so $\mu(1 - \phi_1 - \phi_2) = 0$ and hence $\mu = 0$. With a zero-mean series, the Yule–Walker idea generalises the AR(1) trick: for $h > 0$, multiply both sides of the AR(2) equation (written at time $t + h$) by x_t and take expectations,

$$\mathbb{E}[x_{t+h}x_t] = \phi_1\,\mathbb{E}[x_{t+h-1}x_t] + \phi_2\,\mathbb{E}[x_{t+h-2}x_t] + \underbrace{\mathbb{E}[w_{t+h}x_t]}_{=0\ (h>0)}.$$

This yields the recursion, in autocovariance and (dividing by $\gamma(0)$) in autocorrelation form:

$$\gamma(h) = \phi_1\gamma(h-1) + \phi_2\gamma(h-2), \qquad \rho(h) = \phi_1\rho(h-1) + \phi_2\rho(h-2), \qquad h > 0.$$

Solving for ρ_1 and ρ_2

Evaluate the autocorrelation recursion at $h = 1$, using $\rho(0) = 1$ and the evenness $\rho(-1) = \rho(1)$:

$$\rho(1) = \phi_1\rho(0) + \phi_2\rho(-1) = \phi_1 + \phi_2\rho(1) \quad \Longrightarrow \quad \rho_1(1 - \phi_2) = \phi_1 \quad \Longrightarrow \quad \rho_1 = \frac{\phi_1}{1 - \phi_2}.$$

At $h = 2$,

$$\rho(2) = \phi_1\rho(1) + \phi_2\rho(0) = \phi_1\rho_1 + \phi_2 = \frac{\phi_1^2}{1 - \phi_2} + \phi_2 = \frac{\phi_1^2 + \phi_2(1 - \phi_2)}{1 - \phi_2}.$$

Higher lags follow by applying the recursion $\rho(h) = \phi_1\rho(h - 1) + \phi_2\rho(h - 2)$ iteratively.

The variance $\gamma(0)$ of AR(2)

For $\gamma(0)$, take the $h = 0$ case: multiply $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$ by x_t and take expectations,

$$\gamma(0) = \phi_1\gamma(1) + \phi_2\gamma(2) + \sigma_w^2.$$

Writing $\gamma(1) = \rho_1\gamma(0)$ and $\gamma(2) = \rho_2\gamma(0)$ and solving,

$$\gamma(0)\left(1 - \phi_1\rho_1 - \phi_2\rho_2\right) = \sigma_w^2.$$

Substituting $\rho_1 = \phi_1/(1 - \phi_2)$ and $\rho_2 = \phi_1\rho_1 + \phi_2$ and simplifying gives the closed form

$$\gamma(0) = \frac{\sigma_w^2\,(1 - \phi_2)}{(1 + \phi_2)\left[(1 - \phi_2)^2 - \phi_1^2\right]}.$$

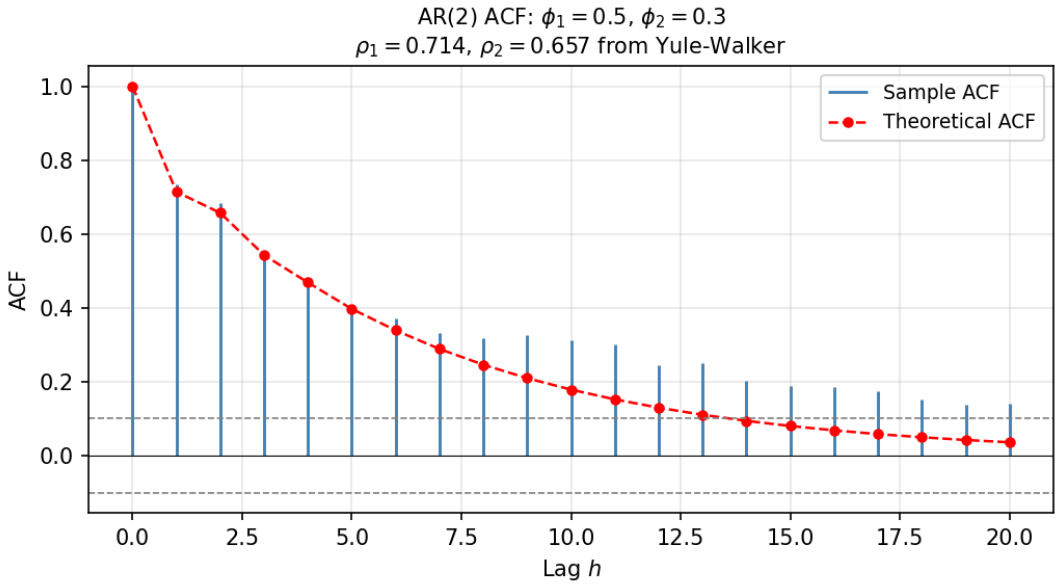


Figure 5.8: AR(2) ACF for $\phi_1 = 0.5$, $\phi_2 = 0.3$. The theoretical ACF (red dashes) computed from the Yule–Walker equations matches the sample ACF closely. As with AR(1), the ACF tails off but never cuts to zero.

The general AR(p) Yule–Walker system

Theorem 5.2: AR(p) Yule–Walker Equations

For the AR(p) model $\phi(B)x_t = w_t$, multiplying by x_t and taking expectations gives, for $h > 0$,

$$\rho(h) = \phi_1\rho(h-1) + \phi_2\rho(h-2) + \cdots + \phi_p\rho(h-p).$$

Evaluating at $h = 1, 2, \ldots, p$ gives the *Yule–Walker system*

$$\underbrace{\begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{p-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{p-2} \\ \vdots & & \ddots & & \vdots \\ \rho_{p-1} & \rho_{p-2} & \cdots & & 1 \end{pmatrix}}_{P(\rho),\ \text{Toeplitz}} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix} = \begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_p \end{pmatrix}.$$

The system can be read in two directions. **Given the coefficients** ϕ_j, it determines the theoretical ACF values ρ_h. **Given a sample ACF** $\hat{\rho}_h$, it can be inverted to estimate the coefficients, $\hat{\boldsymbol{\phi}} = P(\hat{\rho})^{-1}\hat{\rho}$ — this is *Yule–Walker estimation*.

5.7 The MA(∞) Representation: Operator Inversion

We already know that a stationary AR(1) can be written as an infinite moving average, $x_t = \sum_{j=0}^{\infty} \phi^j w_{t-j}$, found in Section 5.2 by back-substitution. Here we re-derive it more formally by *inverting the lag operator*, a method that extends to AR(p).

Operator algebra

The goal is to solve $\phi(B)x_t = w_t$ for x_t, i.e. to find $x_t = \phi(B)^{-1}w_t = \psi(B)w_t$. Define the *truncated inverse*

$$\psi_j^*(B) = 1 + \phi B + \phi^2 B^2 + \cdots + \phi^j B^j.$$

A direct multiplication telescopes:

$$\psi_j^*(B)\,(1 - \phi B) = 1 - \phi^{j+1}B^{j+1}.$$

Applying $\psi_j^*(B)$ to the model $\phi(B)x_t = (1 - \phi B)x_t = w_t$,

$$\psi_j^*(B)\,w_t = \psi_j^*(B)\,\phi(B)\,x_t = x_t - \phi^{j+1}x_{t-j-1}.$$

As $j \to \infty$: since $|\phi| < 1$ and $\mathrm{Var}(x_t) < \infty$, the remainder $\phi^{j+1}x_{t-j-1} \to 0$ in mean square, leaving

$$x_t = \lim_{j\to\infty} \psi_j^*(B)\,w_t = \sum_{j=0}^{\infty} \phi^j w_{t-j}.$$

Coefficient matching

A cleaner algebraic route writes $x_t = \psi(B)w_t$ with $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$ and imposes $\phi(B)\psi(B) = 1$:

$$(1 - \phi B)\big(\psi_0 + \psi_1 B + \psi_2 B^2 + \cdots\big) = 1.$$

Matching the coefficient of each power of B,

$$B^0: \ \psi_0 = 1, \qquad B^1: \ \psi_1 - \phi\psi_0 = 0 \Rightarrow \psi_1 = \phi, \qquad B^2: \ \psi_2 - \phi\psi_1 = 0 \Rightarrow \psi_2 = \phi^2,$$

and in general $\psi_j = \phi\psi_{j-1}$, so $\psi_j = \phi^j$ — the same MA(∞) weights as before.

Remark 5.14: Absolute summability confirms stationarity

The weights $\psi_j = \phi^j$ are absolutely summable:

$$\sum_{j=0}^{\infty} |\psi_j| = \sum_{j=0}^{\infty} |\phi|^j = \frac{1}{1-|\phi|} < \infty \qquad \text{since } |\phi| < 1.$$

This is precisely the condition (Section 5.2) that makes the MA(∞) series a legitimate, finite-variance stationary process.

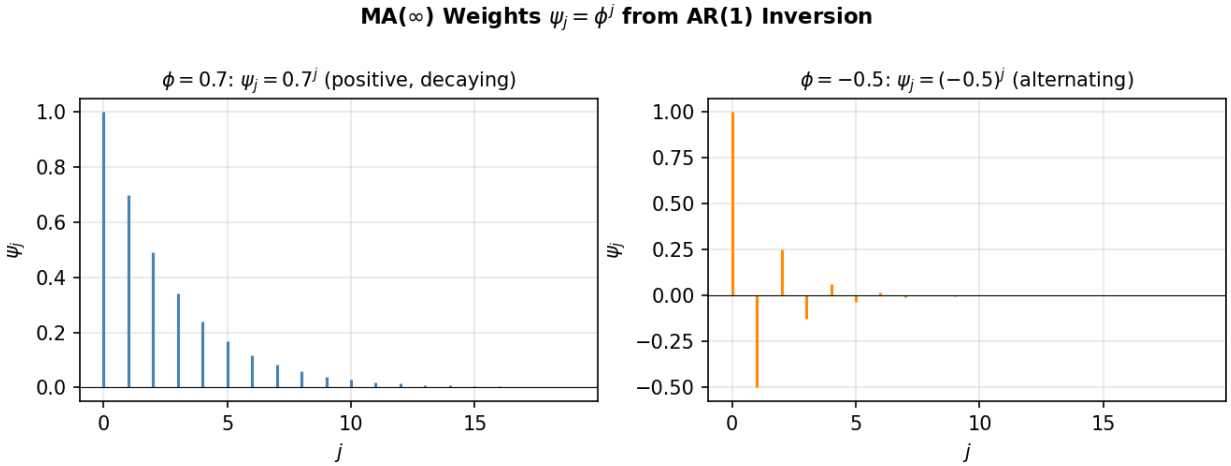


Figure 5.9: MA(∞) weights $\psi_j = \phi^j$ from inverting the AR(1) operator. Left ($\phi = 0.7$): all positive, geometric decay — the distant past still influences x_t, but increasingly less. Right ($\phi = -0.5$): alternating signs, also geometrically damped.

This makes the answer to "does x_t depend on prior shocks?" precise: every past shock w_{t-j} enters x_t with weight ϕ^j. The weights never reach exactly zero, so x_t has *infinite but decaying* memory.

5.8 Partial Fraction Decomposition

For AR(2) and higher orders, inverting $\phi(B)$ by raw coefficient matching still works but gives the ψ_j only recursively. *Partial fractions* give a closed form.

Setting up

Factor the AR(2) polynomial as

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 = (1 - \lambda_1 B)(1 - \lambda_2 B),$$

where $1/\lambda_1, 1/\lambda_2$ are the roots of $\phi(z) = 0$; stationarity requires $|\lambda_i| < 1$. We want $\psi(B) = \phi(B)^{-1}$, i.e. the expansion of

$$\frac{1}{(1-\lambda_1 B)(1-\lambda_2 B)}.$$

Assuming distinct factors ($\lambda_1 \neq \lambda_2$), write the partial fraction decomposition

$$\frac{1}{(1-\lambda_1 B)(1-\lambda_2 B)} = \frac{a}{1-\lambda_1 B} + \frac{b}{1-\lambda_2 B}.$$

Finding the coefficients

Multiplying both sides by $(1-\lambda_1 B)(1-\lambda_2 B)$ gives the identity $1 = a(1-\lambda_2 B) + b(1-\lambda_1 B)$. Evaluating at $B = 1/\lambda_1$ and $B = 1/\lambda_2$ isolates each constant:

$$a = \frac{\lambda_1}{\lambda_1 - \lambda_2}, \qquad b = \frac{\lambda_2}{\lambda_2 - \lambda_1} = -\frac{\lambda_2}{\lambda_1 - \lambda_2}.$$

The MA(∞) weights

Expand each term as a geometric series, valid for $|\lambda_i| < 1$:

$$\frac{1}{1-\lambda_i B} = \sum_{j=0}^{\infty} \lambda_i^j B^j.$$

Combining,

$$\psi(B) = \frac{\lambda_1}{\lambda_1 - \lambda_2} \sum_{j=0}^{\infty} \lambda_1^j B^j - \frac{\lambda_2}{\lambda_1 - \lambda_2} \sum_{j=0}^{\infty} \lambda_2^j B^j = \sum_{j=0}^{\infty} \psi_j B^j, \qquad \boxed{\psi_j = \frac{\lambda_1^{j+1} - \lambda_2^{j+1}}{\lambda_1 - \lambda_2}}$$

As a check: at $j = 0$, $\psi_0 = (\lambda_1 - \lambda_2)/(\lambda_1 - \lambda_2) = 1$; and at $j = 1$, $\psi_1 = (\lambda_1^2 - \lambda_2^2)/(\lambda_1 - \lambda_2) = \lambda_1 + \lambda_2 = \phi_1$, using $\lambda_1 + \lambda_2 = \phi_1$.

The same result by coefficient matching

The partial-fraction formula can also be reached by matching coefficients in $\phi(B)\psi(B) = 1$:

$$(1 - \phi_1 B - \phi_2 B^2)(\psi_0 + \psi_1 B + \psi_2 B^2 + \cdots) = 1,$$

which gives $\psi_0 = 1$, $\psi_1 = \phi_1$, and for $j \geq 2$ the linear recurrence

$$\psi_j = \phi_1 \psi_{j-1} + \phi_2 \psi_{j-2}.$$

This is the *same* recurrence as the Yule–Walker equation for $\rho(h)$. Its general solution is $\psi_j = C_1 \lambda_1^j - C_2 \lambda_2^j$, which matches the partial-fraction result with $C_i = \lambda_i/(\lambda_1 - \lambda_2)$.

Remark 5.15: Generalisation to AR(p)

For an AR(p) process with *distinct* roots, the same decomposition gives

$$\psi_j = C_1 \lambda_1^j + C_2 \lambda_2^j + \cdots + C_p \lambda_p^j,$$

where the constants C_i come from the partial-fraction decomposition of $1/\phi(B)$. The MA(∞) weights are thus a sum of p geometrically decaying modes, one per root.

Remark 5.16: Lecture summary

The threads of this lecture come together as follows.

- **AR(1) ACVF.** $\gamma(h) = \sigma_w^2\,\phi^{|h|}/(1-\phi^2)$ and $\rho(h) = \phi^{|h|}$ — an exponential tail-off that never cuts off.
- **Yule–Walker for AR(p).** $\rho(h) = \phi_1\rho(h-1) + \cdots + \phi_p\rho(h-p)$ for $h > 0$; the matrix system $P(\rho)\,\boldsymbol{\phi} = \rho$ converts between coefficients and ACF values in either direction.
- **MA(∞) representation.** A stationary AR(p) is equivalent to $x_t = \sum_{j=0}^{\infty}\psi_j w_{t-j}$ with $\sum_j |\psi_j| < \infty$; the weights satisfy $\psi_j = \phi_1\psi_{j-1} + \cdots + \phi_p\psi_{j-p}$ and, for distinct roots, have the closed form $\psi_j = \sum_i C_i\lambda_i^j$.

Chapter 6

Lecture 6

6.1 Moving Average (MA) Processes

What is a moving average model?

A **moving average model** is built on the idea that the current value of a series, x_t, is a *finite weighted sum of current and recent white noise shocks*:

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q}.$$

It is instructive to contrast this with the autoregressive model of the previous lecture. In an AR(p) model, x_t is regressed on its *own past values*; in an MA(q) model, x_t is instead a linear combination of *current and past shocks* — no past values of x_t themselves appear on the right-hand side.

This difference has a sharp consequence for memory. The AR model has **infinite memory**: a shock entering the system continues to influence all future values, with an effect that decays geometrically but never vanishes. The MA model, by contrast, has **finite memory**: a shock older than q periods has *exactly zero* influence on the present. The diagnostic signature of this is that the ACF of an MA(q) process is exactly zero beyond lag q — it *cuts off*, rather than tailing off.

When does MA modelling make sense?

MA models are natural whenever the effect of a shock or disturbance lasts for a *limited, known number of periods* and then disappears completely. Put differently, they make sense when a series exhibits **short-memory dependence driven primarily by shocks or noise**, rather than persistent dependence on its own past values. Several settings illustrate the idea.

In **inventory and production**, a supply disruption affects output for a few periods — the time needed to reorder and restock — after which the system returns to normal; an MA(1) or MA(2) captures this well. For **financial returns after news**, a market-moving announcement such as an earnings release or central bank decision causes abnormal returns for one or two trading days, after which prices have fully incorporated the information. In **meteorology**, daily temperature anomalies caused by a passing weather front typically dissipate within a day or two. With **measurement error**, if an instrument averages its readings over the last q periods, the observed values follow an MA(q) model even when the true underlying process is white noise. Finally, **residuals after detrending or differencing** often retain short-range MA structure once trend and seasonality have been removed.

6.2 The MA(q) Model: Definition and Properties

Definition 6.1: MA(q) Model

A *moving average model of order* q, written MA(q), is

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q},$$

where $w_t \sim \mathrm{WN}(0, \sigma_w^2)$ and $\theta_1, \ldots, \theta_q$ are constants with $\theta_q \neq 0$. In backshift operator form,

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

Mean and variance

The first two moments follow directly from the definition. By linearity of expectation, the **mean** is

$$\mathbb{E}(x_t) = \mathbb{E}(w_t) + \theta_1 \mathbb{E}(w_{t-1}) + \cdots + \theta_q \mathbb{E}(w_{t-q}) = 0,$$

so the mean is zero and constant in t for *any* choice of the parameters. For the **variance**, the white noise terms w_{t-j} are uncorrelated, so the cross-terms vanish and only the squared terms survive:

$$\mathrm{Var}(x_t) = \sigma_w^2 \big(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2\big) = \sigma_w^2 \sum_{j=0}^{q} \theta_j^2,$$

where we adopt the convention $\theta_0 = 1$. This is finite for every choice of the θ_j, which already hints that stationarity will be automatic.

Autocovariance function

To find the autocovariance at lag h with $|h| \leq q$, multiply x_{t+h} by x_t and take expectations:

$$\gamma(h) = \mathrm{Cov}(x_{t+h}, x_t) = \mathbb{E}\left[\left(\sum_{j=0}^{q} \theta_j w_{t+h-j} \right) \left(\sum_{k=0}^{q} \theta_k w_{t-k} \right) \right].$$

Since $\mathbb{E}[w_{t+h-j}\, w_{t-k}] = \sigma_w^2\, \mathbf{1}_{h-j=-k} = \sigma_w^2\, \mathbf{1}_{k=j-h}$, only the terms with $k = j - h$ contribute, giving

$$\gamma(h) = \sigma_w^2 \sum_{j=h}^{q} \theta_j\, \theta_{j-h}, \qquad 0 \leq h \leq q,$$

again with $\theta_0 = 1$. For $|h| > q$ the time-index sets of the two sums do not overlap at all, so $\gamma(h) = 0$.

Theorem 6.1: ACF of an MA(q) Process Cuts Off

For an MA(q) process, $\gamma(h) = 0$ for all $|h| > q$. The autocorrelation function is *exactly zero* beyond lag q — it **cuts off**, it does not tail off. This sharp cut-off is the defining diagnostic signature of an MA(q) process.

Stationarity

Every MA(q) process is **always covariance stationary**, regardless of the values of $\theta_1, \ldots, \theta_q$, because x_t is a finite linear combination of white noise terms with constant mean and a covariance structure that depends only on the lag h. There is therefore *no stationarity condition to check* — stationarity is automatic. The substantive issue for MA models is instead **invertibility**, taken up in Section 6.5.

6.3 The MA(1) Process

For the first-order model $x_t = w_t + \theta w_{t-1}$, the general formulae specialise neatly. The **variance** (the case $h = 0$) is

$$\gamma(0) = \mathrm{Var}(x_t) = \sigma_w^2 \, (1 + \theta^2).$$

The **lag-1 covariance** is

$$\gamma(1) = \mathrm{Cov}(x_{t+1}, x_t) = \mathbb{E}\big[(w_{t+1} + \theta w_t)(w_t + \theta w_{t-1})\big] = \theta \sigma_w^2.$$

Expanding the product and using $\mathbb{E}[w_i w_j] = \sigma_w^2 \, \mathbf{1}_{i=j}$, only the cross-term $\mathbb{E}[\theta w_t \cdot w_t] = \theta \sigma_w^2$ survives. For any lag $|h| \geq 2$ the two expressions share no common innovation, so $\gamma(h) = 0$. The autocorrelation function is therefore

$$\rho(h) = \begin{cases} 1, & h = 0, \\ \dfrac{\theta}{1 + \theta^2}, & |h| = 1, \\ 0, & |h| \geq 2. \end{cases}$$

Remark 6.1: The lag-1 ACF of an MA(1) is bounded

The function $\theta/(1+\theta^2)$ satisfies $|\rho(1)| \leq \frac{1}{2}$ for *all* θ, attaining its extreme values $\pm\frac{1}{2}$ at $\theta = \pm 1$. Thus an MA(1) process can never have a lag-1 autocorrelation outside $[-\frac{1}{2}, \frac{1}{2}]$ — a useful sanity check when fitting.

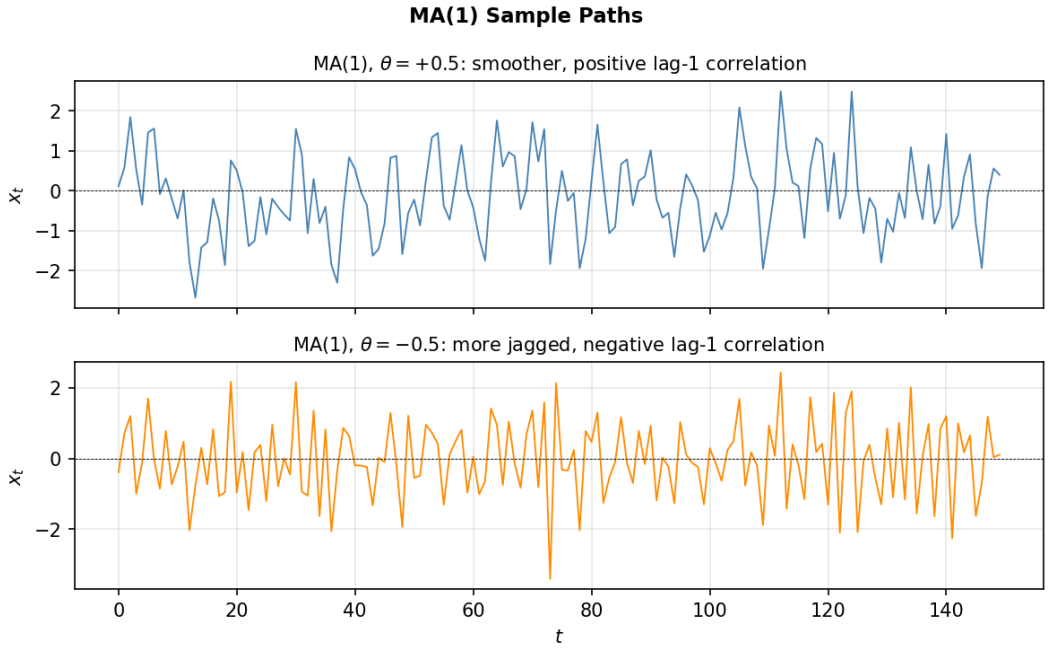


Figure 6.1: MA(1) sample paths. With $\theta = +0.5$ (top), neighbouring values tend to move together, producing a slightly smoother series. With $\theta = -0.5$ (bottom), neighbouring values tend to move in opposite directions, producing a jagged, rapidly oscillating series.

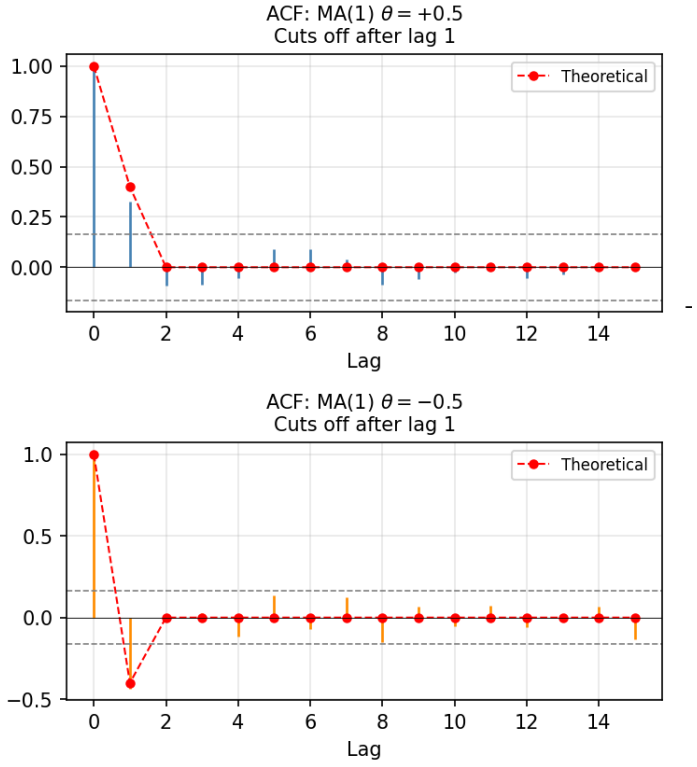


Figure 6.2: MA(1) ACF. The autocorrelation cuts off sharply after lag 1, and the sample values match the theoretical values (red dashes).

6.4 The MA(2) Process

For the second-order model $x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2}$, applying the general autocovariance formula with $\theta_0 = 1$ gives the following. The **variance** is

$$\gamma(0) = \sigma_w^2 \, (1 + \theta_1^2 + \theta_2^2).$$

The **lag-1** autocovariance sums over $j = 1, 2$ with $k = j - 1$:

$$\gamma(1) = \sigma_w^2 \, (\theta_1 \cdot 1 + \theta_2 \cdot \theta_1) = \sigma_w^2 \, \theta_1 (1 + \theta_2).$$

The **lag-2** autocovariance has only the term $j = 2, \; k = 0$:

$$\gamma(2) = \sigma_w^2 \, \theta_2.$$

For $|h| \geq 3$ there is no overlap and $\gamma(h) = 0$. Dividing through by $\gamma(0)$ gives the ACF,

$$\rho(1) = \frac{\theta_1(1 + \theta_2)}{1 + \theta_1^2 + \theta_2^2}, \qquad \rho(2) = \frac{\theta_2}{1 + \theta_1^2 + \theta_2^2}, \qquad \rho(h) = 0 \;\; \text{for } |h| \geq 3.$$

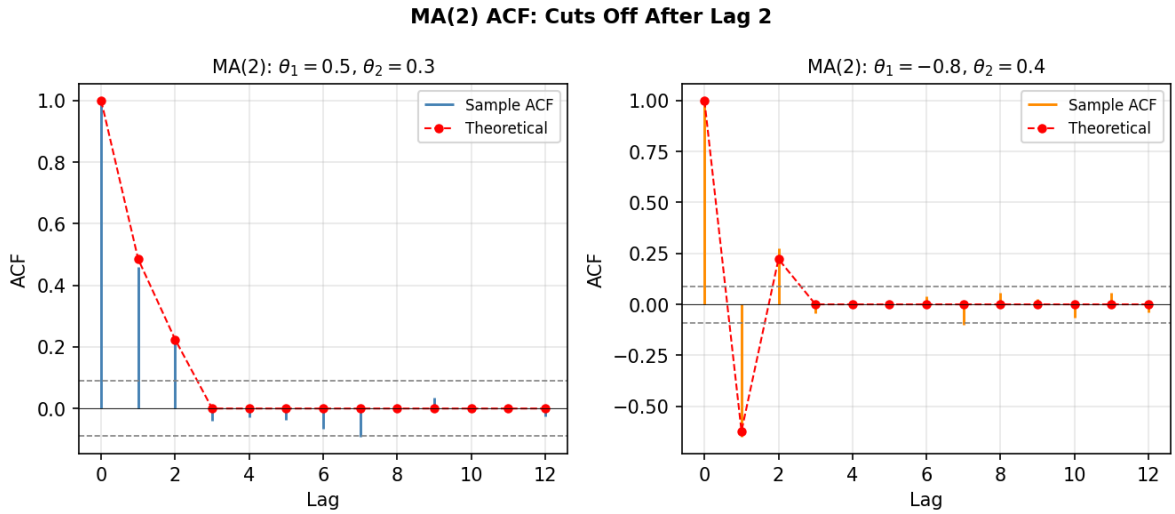


Figure 6.3: MA(2) sample ACF versus theoretical (red dashes) for two parameter sets. In both cases the ACF cuts off cleanly after lag 2. The pattern at lags 1 and 2 encodes the sign and magnitude of θ_1 and θ_2.

6.5 Non-Uniqueness and Invertibility

The non-uniqueness problem

A subtlety of MA models is that two different parameterisations can produce *exactly the same* autocorrelation structure. Consider the two MA(1) models

$$x_t = w_t + \tfrac{1}{5} w_{t-1}, \quad w_t \overset{\text{iid}}{\sim} N(0, 25), \qquad y_t = v_t + 5\, v_{t-1}, \quad v_t \overset{\text{iid}}{\sim} N(0, 1).$$

Computing the lag-1 ACF of each,

$$\rho_x(1) = \frac{1/5}{1 + (1/5)^2} = \frac{1/5}{26/25} = \frac{5}{26}, \qquad \rho_y(1) = \frac{5}{1 + 25} = \frac{5}{26}.$$

The two processes have **identical ACFs** — and, if Gaussian, identical distributions — yet they are parameterised differently. Given only data, we cannot distinguish them on statistical grounds alone. The general pattern is that θ and $1/\theta$, with σ_w rescaled appropriately, always produce the same ACF. This is the **non-uniqueness problem** of MA models.

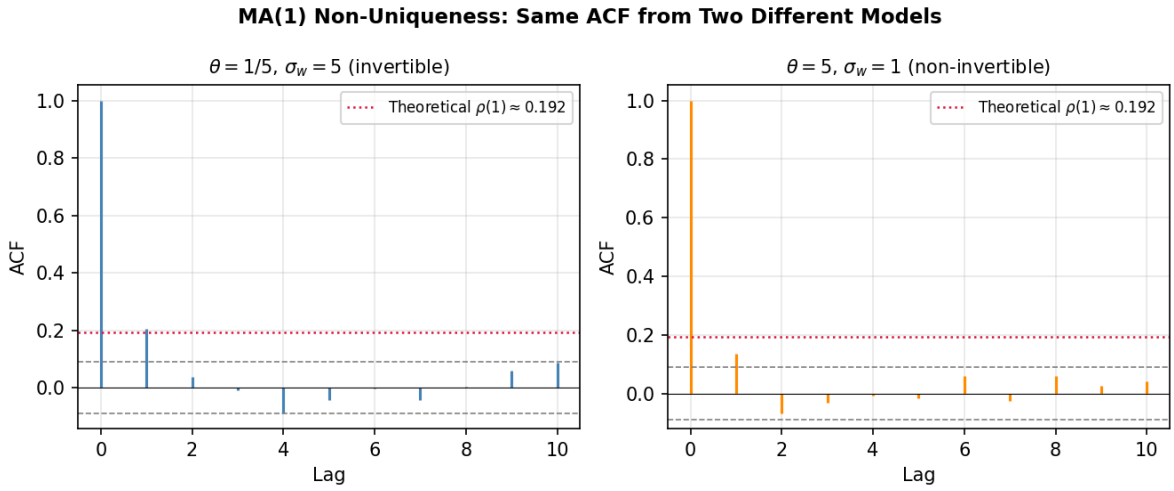


Figure 6.4: Sample ACFs from the $\theta = 1/5$ model (left) and the $\theta = 5$ model (right), both with variance rescaled so that $\rho(1) \approx 5/26$. The theoretical lag-1 value (red dotted line) is identical for both — the data cannot tell the models apart.

Resolving non-uniqueness: invertibility

Since several representations describe the same model, we need a principled rule for choosing one. We take our cue from the AR case. For an AR model we require *causality*: we rewrite the process as an infinite sum of past shocks — an MA(∞) representation — to identify the stationary solution. For the MA model we mimic this criterion in the opposite direction: among the multiple representations of the same model, we choose the one that admits an infinite *autoregressive* representation, AR(∞), expressing w_t in terms of past observations of x_t. Such a process is called **invertible**.

To see how this picks out a unique model, reverse the roles of x_t and w_t. Starting from $x_t = w_t + \theta w_{t-1}$ and rearranging,

$$w_t = -\theta w_{t-1} + x_t.$$

Iterating this recursion backward, if $|\theta| < 1$ the resulting series converges:

$$w_t = \sum_{j=0}^{\infty} (-\theta)^j x_{t-j} = \pi(B)\, x_t.$$

This is a valid AR(∞) representation of the process in terms of its own past, so the process is invertible. If $|\theta| > 1$ the series diverges and no such representation exists.

Theorem 6.2: Invertibility Criterion for MA(1)

The MA(1) process $x_t = w_t + \theta w_{t-1}$ is *invertible* if and only if $|\theta| < 1$.

Theorem 6.3: Invertibility Criterion for MA(q)

Write the MA(q) process in backshift form, $x_t = \theta(B)\, w_t$. The invertible representation is $\pi(B)\, x_t = w_t$ with $\pi(B) = \theta(B)^{-1}$ a formal power series. The MA(q) process is **invertible** if and only if all roots of the MA polynomial

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

lie *strictly outside* the unit circle: $\theta(z) = 0 \implies |z| > 1$.

Just as we always work with the causal (stationary) solution of an AR model, we always choose the **invertible** representation of an MA model. Among all MA models sharing the same ACF, the invertible one is the canonical choice, and it is the representation whose polynomial roots all satisfy $|z_j| > 1$.

The MA(2) invertibility region

For an MA(2) process the polynomial is $\theta(z) = 1 + \theta_1 z + \theta_2 z^2$. Requiring both roots of $\theta(z) = 0$ to lie outside the unit circle is equivalent to the three inequalities

$$|\theta_2| < 1, \qquad 1 + \theta_1 - \theta_2 > 0, \qquad 1 - \theta_1 - \theta_2 > 0.$$

These are exactly the AR(2) stationarity conditions with ϕ_i replaced by θ_i — a direct reflection of the duality between AR and MA models. Parameter pairs inside the resulting triangular region give invertible MA(2) models; pairs outside correspond to non-invertible representations, which we discard in favour of their invertible equivalents.

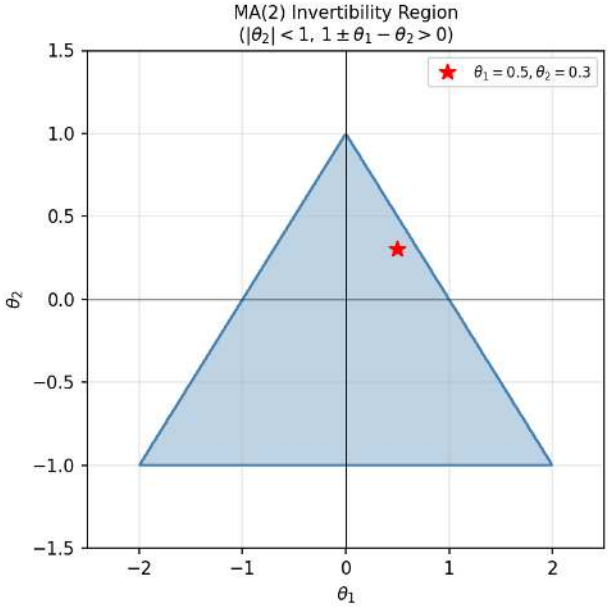


Figure 6.5: The MA(2) invertibility region. Points inside the triangle satisfy $|\theta_2| < 1$ and $1 \pm \theta_1 - \theta_2 > 0$ and give invertible models; points outside do not.

6.6 General MA(q): Summary of Results

Summary of properties

Collecting the results above, the autocovariance and autocorrelation functions of an MA(q) process are

$$\gamma(h) = \begin{cases} \sigma_w^2 \displaystyle\sum_{j=h}^{q} \theta_j \, \theta_{j-h}, & 0 \le h \le q, \\ 0, & h > q, \end{cases} \qquad (\theta_0 = 1),$$

$$\rho(h) = \frac{\sum_{j=h}^{q} \theta_j \, \theta_{j-h}}{\sum_{j=0}^{q} \theta_j^2}, \quad 0 \le h \le q; \qquad \rho(h) = 0, \quad h > q.$$

An MA(q) model is *always* covariance stationary, and it is invertible if and only if all roots of $\theta(z) = 0$ lie strictly outside the unit circle. Among all MA models with a given ACF, the invertible one is the canonical choice.

Duality with AR(p)

The AR and MA families mirror one another point for point. The table below summarises the correspondence.

	AR(p)	**MA(q)**
Definition	$\phi(B)\, x_t = w_t$	$x_t = \theta(B)\, w_t$
Stationarity	Roots of $\phi(z)$ outside unit circle	Always stationary
Key condition	Causality	Invertibility
Causality / invertibility	Roots of $\phi(z)$ outside unit circle	Roots of $\theta(z)$ outside unit circle
MA(∞) representation	$x_t = \psi(B)\, w_t$ (causal)	Always (finite sum)
AR(∞) representation	Always (finite AR)	$\pi(B)\, x_t = w_t$ (invertible)
ACF	Tails off	Cuts off after lag q

Table 6.1: Duality between AR(p) and MA(q) processes.

The MA(∞) process

The MA(q) model can be extended to allow *infinitely many* lagged shocks, giving the MA(∞) process.

Definition 6.2: MA(∞) Process

The *MA(∞) process* is

$$x_t = \sum_{j=0}^{\infty} \theta_j \, w_{t-j} = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots, \qquad w_t \sim \mathrm{WN}(0, \sigma_w^2),$$

where $\theta_0 = 1$ by convention and $\{\theta_j\}$ is a sequence of real coefficients.

The MA(∞) representation matters because it unifies the model families: every causal AR(p) process has an MA(∞) representation, the MA(q) model is the special case in which $\theta_j = 0$ for all $j > q$, and a general stationary linear process is an MA(∞) with suitably summable coefficients.

For the moments, the **mean** is

$$\mathbb{E}(x_t) = \sum_{j=0}^{\infty} \theta_j \, \mathbb{E}(w_{t-j}) = 0,$$

and, since distinct w_{t-j} are uncorrelated, the **variance** is

$$\gamma(0) = \mathrm{Var}(x_t) = \sigma_w^2 \sum_{j=0}^{\infty} \theta_j^2,$$

which is finite if and only if $\sum_{j=0}^{\infty} \theta_j^2 < \infty$. This square-summability condition is what ensures x_t is well defined in L^2; it is weaker than absolute summability.

The **autocovariance** at lag $h \geq 0$ is obtained by pairing terms that share the same innovation:

$$\gamma(h) = \mathrm{Cov}(x_{t+h}, x_t) = \sigma_w^2 \sum_{j=0}^{\infty} \theta_{j+h} \, \theta_j.$$

For this sum to converge absolutely for *all* h, the sufficient condition is **absolute summability**,

$$\sum_{j=0}^{\infty} |\theta_j| < \infty.$$

Absolute summability implies $\sum \theta_j^2 < \infty$, so it is the standard working condition. Under it, $\gamma(h) \to 0$ as $h \to \infty$ and x_t is covariance stationary.

Chapter 7

Lecture 7

7.1 ARMA(p, q) Models

The ARMA family

In the last 2 lectures, we looked at AR(p) — autoregressive of order p — and MA(q) — moving average of order q. In today's lecture, we study the following hierarchy of models:

- **ARMA(p, q)** — autoregressive moving average.
- **ARIMA(p, d, q)** — integrated extension (non-stationary).
- **SARIMA$(p, d, q)(P, D, Q)_s$** — seasonal extension.

Each model is a building block for the next. ARMA combines AR and MA into a single parsimonious framework.

ARMA(p, q): definition

Definition 7.1: ARMA(p, q)

A time series $\{X_t\}$ is ARMA(p, q) if it is stationary and

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

where $\phi_p \neq 0$, $\theta_q \neq 0$, $\varepsilon_t \sim \text{WN}(0, \sigma^2)$. The parameters p and q denote the autoregressive and the moving average orders, respectively.

In backshift operator notation:

$$\phi(B)\, X_t = \theta(B)\, \varepsilon_t,$$

where

$$\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p, \qquad \theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q.$$

Special cases: ARMA$(p,0)$ = AR(p); ARMA$(0,q)$ = MA(q).

ARMA(p, q): non-zero mean

If X_t has a non-zero mean μ, set $\alpha = \mu\,(1 - \phi_1 - \phi_2 - \cdots - \phi_p)$ and write

$$X_t = \alpha + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}.$$

- When $\phi_1 + \cdots + \phi_p \neq 1$ (guaranteed by stationarity), $\mu = \alpha/(1 - \phi_1 - \cdots - \phi_p)$.
- In practice one often works with the mean-corrected series $Z_t = X_t - \mu$, which satisfies a zero-mean ARMA model.

Parameter redundancy

Problem: an ARMA model may not be uniquely identified.

Example 7.1: Parameter Redundancy

Let $X_t = \varepsilon_t$ be a white noise. Multiply both sides by $\eta(B) = 1 - 0.5B$, then the model becomes:

$$(1 - 0.5B)\, X_t = (1 - 0.5B)\, \varepsilon_t \implies X_t = 0.5X_{t-1} + \varepsilon_t - 0.5\varepsilon_{t-1}.$$

This makes X_t look like ARMA(1,1), but X_t is white noise. The root cause of this misconception is parameter redundancy or over-parameterization.

Root of the problem: the AR polynomial $\phi(z) = 1 - 0.5z$ and MA polynomial $\theta(z) = 1 - 0.5z$ share a common factor, creating parameter redundancy.

Solution: require that $\phi(z)$ and $\theta(z)$ have no common factors. This ensures the ARMA representation is unique.

Potential issues

We have seen three problems encountered by the ARMA family of models. Below are solutions to deal with them:

Problem	Source	Solution
Parameter redundancy	ARMA	$\phi(z), \theta(z)$ share no common roots
Model depends on future	AR part	Require causality
Non-unique representation	MA part	Require invertibility

We next look at the formal definition of causality and invertibility for ARMA models, which are similar to the ones for AR and MA models. The AR and MA characteristic polynomials are

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p, \quad \phi_p \neq 0, \qquad \theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q, \quad \theta_q \neq 0,$$

where z is a complex number.

Causal ARMA

Definition 7.2: Causal ARMA

An ARMA(p, q) model is causal if $\{X_t\}$ can be written as a one-sided linear process

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} = \psi(B)\, \varepsilon_t,$$

where

$$\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j, \quad \psi_0 = 1, \quad \text{and} \quad \sum_{j=0}^{\infty} |\psi_j| < \infty.$$

This definition is similar to the definition of a causal AR process, however inverting the ARMA process to a linear process will be slightly more involved due to the presence of the MA component as well.

Theorem 7.1: Causality of ARMA

An ARMA(p, q) model is causal if and only if $\phi(z) \neq 0$ for all $|z| \leq 1$. The coefficients ψ_j of the linear process can be determined by solving

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)}, \qquad |z| \leq 1.$$

In other words, causality requires all roots of $\phi(z)$ to lie strictly outside the unit circle, i.e. $\phi(z) = 0 \Rightarrow |z| > 1$. Causality of ARMA also implies stationarity.

Remark 7.1: Remark

To obtain the coefficients ψ_j, one first inverts the AR characteristic polynomial $\phi(z)$ (using for instance, partial fraction approach) and then factors in the multiplication with the MA characteristic polynomial $\theta(z)$; or directly by coefficient matching.

Invertible ARMA

Definition 7.3: Invertible ARMA

An ARMA(p, q) model is invertible if $\{X_t\}$ can be written as

$$\pi(B)\, X_t = \sum_{j=0}^{\infty} \pi_j X_{t-j} = \varepsilon_t,$$

where

$$\pi(B) = \sum_{j=0}^{\infty} \pi_j B^j, \quad \text{with } \pi_0 = 1, \quad \text{and} \quad \sum_{j=0}^{\infty} |\pi_j| < \infty.$$

Theorem 7.2: Invertibility of ARMA

An ARMA(p, q) model is invertible if and only if $\theta(z) \neq 0$ for all $|z| \leq 1$. The coefficients π_j of the linear process can be determined by solving

$$\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \frac{\phi(z)}{\theta(z)}, \qquad |z| \leq 1.$$

In other words, invertibility requires all roots of $\theta(z)$ to lie strictly outside the unit circle, i.e.

$\theta(z) = 0 \Rightarrow |z| > 1$.

Remark 7.2: Remark

To obtain the coefficients π_j, one first inverts the MA characteristic polynomial $\theta(z)$ (using for instance, partial fraction approach) and then factors in the multiplication with the AR characteristic polynomial $\phi(z)$; or directly by coefficient matching.

Example: ARMA(p, q)

Example 7.2: ARMA(p, q) Identification

Consider the following time series model

$$x_t = 0.4x_{t-1} + 0.45x_{t-2} + w_t + w_{t-1} + 0.25w_{t-2} \tag{1}$$

1. Identify the above model as ARMA(p, q) model (watch out parameter redundancy).
2. Determine whether the model is causal and/or invertible.
3. If the model is causal, write the model as a linear process.

7.2 ACVF of ARMA Models

ARMA(1,1): model and setup

Model:

$$X_t = \phi X_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}, \qquad \varepsilon_t \sim \mathrm{WN}(0, \sigma^2), \quad |\phi| < 1.$$

The MA(∞) representation (from causality) is $X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$ where $\psi_0 = 1$, $\psi_1 = \phi + \theta$, and $\psi_j = \phi^{j-1}(\phi + \theta)$ for $j \geq 1$.

Strategy for ACVF:

1. Compute $\gamma(0)$ using $\mathrm{Var}(X_t)$.
2. Compute $\gamma(1)$ directly from the model.
3. For $h \geq 2$, use the Yule–Walker-type recurrence.

ARMA(1,1): computing $\gamma(0)$ and $\gamma(1)$

Variance $\gamma(0)$:

$$\begin{aligned}\mathrm{Var}(X_t) &= \mathrm{Var}(\phi X_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}) \\ &= \phi^2\sigma_X^2 + \sigma^2 + \theta^2\sigma^2 + 2\phi\theta\mathrm{Cov}(X_{t-1}, \varepsilon_{t-1}) \\ &= \phi^2\sigma_X^2 + (1 + \theta^2)\sigma^2 + 2\phi\theta\sigma^2.\end{aligned}$$

Since $\mathrm{Cov}(X_{t-1}, \varepsilon_{t-1}) = \sigma^2$ (from the MA(∞) rep.), we get

$$\gamma(0)(1 - \phi^2) = \sigma^2(1 + 2\phi\theta + \theta^2) \implies \gamma(0) = \frac{(1 + 2\phi\theta + \theta^2)\sigma^2}{1 - \phi^2}.$$

ACVF at lag 1:

$$\gamma(1) = \mathrm{Cov}(X_{t+1}, X_t) = \mathrm{Cov}(\phi X_t + \varepsilon_{t+1} + \theta\varepsilon_t, X_t) = \phi\,\gamma(0) + \theta\sigma^2.$$

ARMA(1,1): ACVF for $h \geq 2$ and ACF

Recurrence for $h \geq 2$:

$$\begin{aligned}\gamma(h) = \mathrm{Cov}(X_{t+h}, X_t) &= \mathrm{Cov}(\phi X_{t+h-1} + \varepsilon_{t+h} + \theta\varepsilon_{t+h-1}, X_t) \\ &= \phi\,\gamma(h-1) \qquad [\text{since } \varepsilon_{t+h}, \varepsilon_{t+h-1} \perp X_t \text{ for } h \geq 2].\end{aligned}$$

Therefore $\gamma(h) = \phi^{h-1}\gamma(1)$ for $h \geq 1$, and

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \begin{cases} 1 & h = 0, \\ \dfrac{(\phi+\theta)(1+\phi\theta)}{1+2\phi\theta+\theta^2} & h = 1, \\ \phi^{h-1}\rho(1) & h \geq 2. \end{cases}$$

Key feature: the ACF tails off like AR for $h \geq 2$, but the lag-1 value is "shifted" by the MA parameter θ.

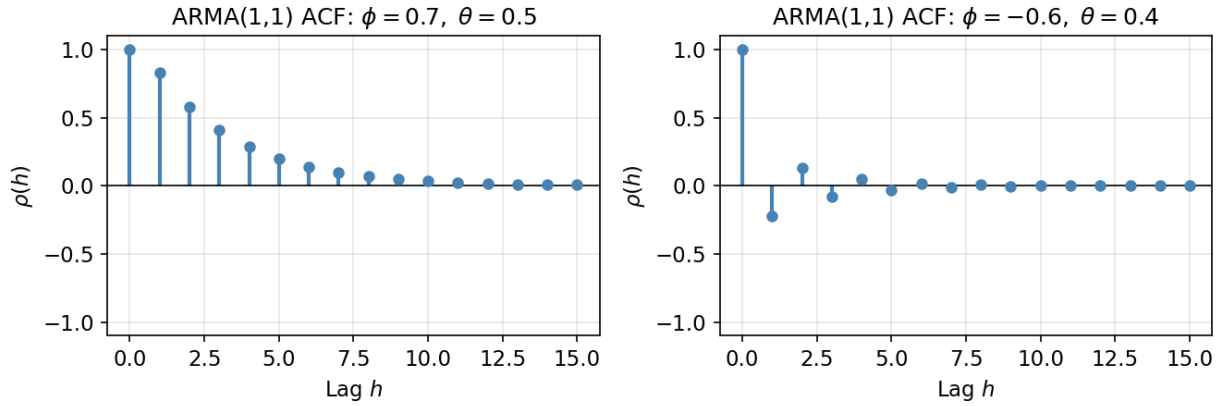


Figure 7.1: ARMA(1,1) ACF. The ACF starts at a lag-1 value determined by both ϕ and θ, then decays geometrically at rate ϕ. Neither cuts off (as MA would) nor decays from lag 1 (as AR does) — it is a mixture of both behaviours.

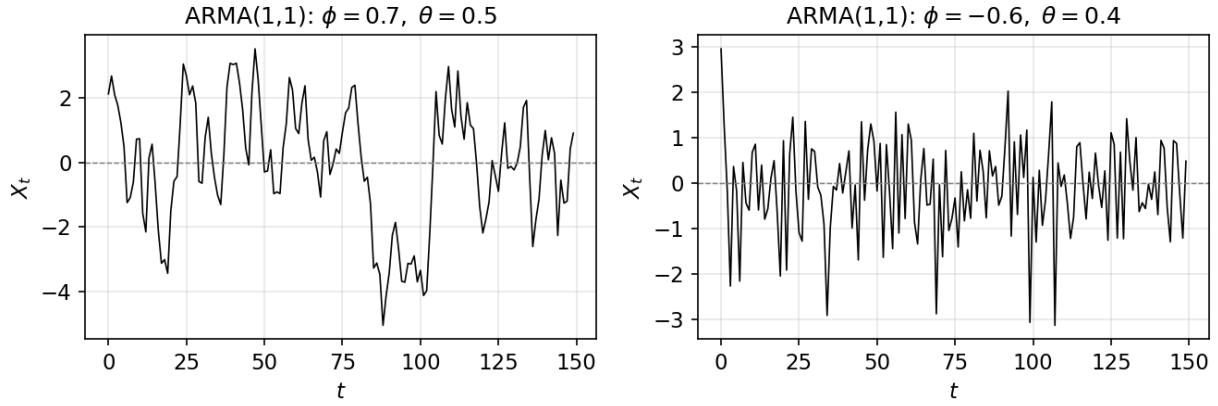


Figure 7.2: ARMA(1,1) sample paths. Both ϕ and θ shape the short-run dynamics: θ amplifies or dampens the immediate effect of each shock, while ϕ governs the persistence / oscillation.

ARMA(2,2): model and setup

Model:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}, \qquad \varepsilon_t \sim \mathrm{WN}(0, \sigma^2).$$

Backshift form: $\phi(B) X_t = \theta(B) \varepsilon_t$.

Strategy:

1. Multiply both sides of the model by X_{t-h} and take expectations.
2. For $h = 0$: obtain $\gamma(0)$ in terms of $\gamma(1)$, $\gamma(2)$, and σ^2-weighted cross-terms involving ε and X.
3. For $h \geq 3$: ACVF satisfies the AR(2) Yule–Walker recurrence $\gamma(h) = \phi_1 \gamma(h-1) + \phi_2 \gamma(h-2)$.

The lags $h = 1$ and $h = 2$ require special treatment because ε_{t-h} is not yet independent of X_{t-h} at those lags.

ARMA(2,2): ACVF at lags 0, 1, 2

For the general ARMA(p, q) case, $\mathrm{Cov}(\varepsilon_{t-k}, X_t) = \sigma^2 \psi_k$ where ψ_k are the MA(∞) coefficients (obtained by coefficient matching from $\phi(B)\psi(B) = \theta(B)$). For ARMA(2,2): $\psi_0 = 1$, $\psi_1 = \phi_1 + \theta_1$, $\psi_2 = \phi_1 \psi_1 + \phi_2 + \theta_2$.

$$\begin{aligned}
\text{Lag 0 (multiply model by } X_t\text{, take } \mathbb{E}): &\quad \gamma(0) = \phi_1 \gamma(1) + \phi_2 \gamma(2) + \sigma^2(\psi_0 + \theta_1 \psi_1 + \theta_2 \psi_2). \\
\text{Lag 1 (multiply by } X_{t-1}): &\quad \gamma(1) = \phi_1 \gamma(0) + \phi_2 \gamma(1) + \sigma^2(\theta_1 \psi_0 + \theta_2 \psi_1). \\
\text{Lag 2 (multiply by } X_{t-2}): &\quad \gamma(2) = \phi_1 \gamma(1) + \phi_2 \gamma(0) + \sigma^2 \theta_2.
\end{aligned}$$

Solve the system $\{\gamma(0), \gamma(1), \gamma(2)\}$ jointly, then use the recurrence for $h \geq 3$.

ARMA(2,2): ACVF for $h \geq 3$ and summary

Recurrence for $h \geq 3$: ($\varepsilon_{t+h-k} \perp X_t$ for $k \leq 2$, $h \geq 3$)

$$\gamma(h) = \phi_1 \gamma(h-1) + \phi_2 \gamma(h-2), \qquad h \geq 3.$$

This is exactly the AR(2) Yule–Walker recurrence, so the ACF tails off as a mixture of geometric / damped oscillatory terms.

Remark 7.3: General ARMA(p, q) pattern

- For $h > q$: $\gamma(h) = \phi_1 \gamma(h-1) + \cdots + \phi_p \gamma(h-p)$ (pure AR recurrence).
- The first $\min(p, q)$ lags require solving a linear system involving the MA(∞) coefficients ψ_j.
- ACF: always tails off and never cuts off — the only nuance is that the first $\min(p, q)$ lags are "irregular" (shaped by both the AR and MA parameters jointly), and then from lag $\max(p, q)$ onwards it settles into the pure AR(p) decay pattern.

7.3 ARIMA Models

Motivation: non-stationary series

Many real time series are not stationary: they exhibit trends, drifts, or unit roots. The key idea to deal with such cases is: *difference the series until stationarity is achieved, then fit ARMA.*

Random walk example: $X_t = X_{t-1} + \varepsilon_t$, $\varepsilon_t \sim \mathrm{WN}(0, \sigma^2)$.

- $\{X_t\}$ is non-stationary (variance grows with t).
- $\{X_t\}$ has unit root ($\phi(z) = 0 \Rightarrow z = 1$).
- $\nabla X_t = X_t - X_{t-1} = \varepsilon_t \sim \text{WN}(0, \sigma^2) \Rightarrow$ covariance stationary.

Polynomial trend: $X_t = m_t + Y_t$ where m_t is a polynomial trend of degree k and Y_t is covariance stationary.

$$\nabla^k X_t = (1 - B)^d X_t = \nabla^k m_t + \nabla^k Y_t \Rightarrow \text{covariance stationary.}$$

Remark 7.4: Note

Differencing a covariance stationary time series returns a time series that is covariance stationary too.

Integrated of order d

Definition 7.4: Integrated of Order d: $X_t \sim I(d)$

$\{X_t\}$ is said to be *integrated of order d*, i.e. $X_t \sim I(d)$ if $\{\nabla^d X_t\}$ is covariance stationary, where d is the smallest non-negative integer with that property.

Examples:

- Stationary ARMA process: $X_t \sim I(0)$.
- Random walk: $X_t \sim I(1)$ since ∇X_t is white noise.
- Polynomial trend $m_t = \beta_0 + \cdots + \beta_k t^k$: eliminated by ∇^k, so $m_t \sim I(k)$.

Remark 7.5: Note

$X_t \sim I(d)$ with $d > 0 \Rightarrow \{X_t\}$ is not covariance stationary.

Co-integration

Intuition: Cointegration is a statistical property indicating that

- two or more non-stationary time series share a long-term, stable equilibrium;
- individual variables can wander independently over time;
- but a specific linear combination of them remains stationary;
- i.e. their deviations from a baseline relationship are temporary and self-correcting.

Example 7.3: Household Income and Consumption Spending

- Both aggregate disposable income and consumer spending generally trend upward over time due to economic growth and inflation. Both are non-stationary and behave like random walks.
- According to economic theory (the Permanent Income Hypothesis), people base their spending on long-term income expectations. While spending might temporarily spike (e.g., during a holiday shopping season) or drop (e.g., due to a sudden expense), it always adjusts back to income levels.
- The difference between the two—the savings rate—remains stable and fluctuates around a constant mean over the long run.

Definition 7.5: Co-integrated Processes

Let $X_t \sim I(d)$ and $Y_t \sim I(d)$. The processes $\{X_t\}$ and $\{Y_t\}$ are said to be *co-integrated* if there exist constants α, β such that

$$Z_t = \alpha X_t + \beta Y_t \sim I(0).$$

- Existence of (α, β) is not guaranteed in general.
- Economically: two $I(1)$ series may share a long-run equilibrium even though each individually is non-stationary.

Multivariate generalisation: if $X_{1,t}, \ldots, X_{p,t} \sim I(d)$ and $\alpha_1 X_{1,t} + \cdots + \alpha_p X_{p,t} \sim I(0)$, the processes are said to be co-integrated.

ARIMA(p, d, q): definition

Definition 7.6: ARIMA(p, d, q)

A time series $\{X_t\}$ follows an *ARIMA(p, d, q)* model if $Y_t = \nabla^d X_t$ follows a causal ARMA(p, q) process:

$$\phi(B)\, \nabla^d X_t = \theta(B)\, \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{WN}(0, \sigma^2),$$

where $\nabla^d = (1 - B)^d$.

Writing $\Phi^*(B) = \phi(B)(1 - B)^d$, the model is $\Phi^*(B)\, X_t = \theta(B)\, \varepsilon_t$.

- ARIMA$(p,0,q) \equiv$ ARMA(p, q).
- ARIMA$(p, d, q) \equiv$ ARMA$(p + d, q)$ in the extended AR polynomial $\Phi^*(B)$, which has p roots of $\phi(B)$ outside the unit circle and d roots on the unit circle.
- $\{X_t\}$ is non-stationary for $d \geq 1$.

ARIMA(1,1,1): worked example

Example 7.4: ARIMA(1,1,1)

Let $X_t \sim$ ARIMA(1,1,1), so $Y_t = \nabla X_t \sim$ ARMA(1,1):

$$Y_t = \phi Y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}, \qquad |\phi| < 1.$$

Substituting $Y_t = X_t - X_{t-1}$:

$$\begin{aligned} X_t - X_{t-1} &= \phi(X_{t-1} - X_{t-2}) + \varepsilon_t + \theta \varepsilon_{t-1} \\ X_t &= (1 + \phi) X_{t-1} - \phi X_{t-2} + \varepsilon_t + \theta \varepsilon_{t-1}. \end{aligned}$$

So $X_t \sim$ ARMA(2,1) in terms of the extended polynomial $\Phi^*(B) = (1 - \phi B)(1 - B)$, whose roots are $1/\phi$ (outside the unit circle) and 1 (on the unit circle).

$$\Phi^*(B)\, X_t = \theta(B)\, \varepsilon_t, \qquad \text{roots of } \Phi^*(z) = 0 : \ \phi^{-1} \text{ and } 1.$$

ARIMA: differencing in practice

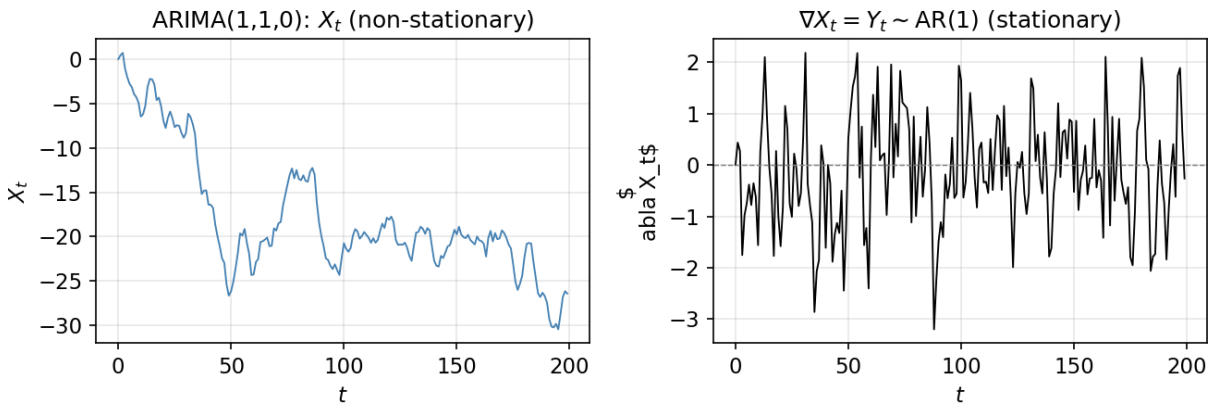


Figure 7.3: ARIMA(1,1,0) process (left) and its first difference (right), which follows AR(1).

Practical rule: difference until the sample ACF decays reasonably quickly. Over-differencing can introduce spurious MA structure.

7.4 Seasonal ARMA and ARIMA Models

Motivation: seasonal time series

Many economic and environmental series have periodic structure at a fixed seasonal period s (e.g. $s = 12$ for monthly, $s = 4$ for quarterly).

AR(1)$_s$ — Seasonal AR(1) is a Restricted AR(s):

$$X_t = \phi_1^{(s)} X_{t-s} + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{WN}(0, \sigma^2).$$

Only the lag s term appears; lags $1, \ldots, s-1$ are absent.

MA(1)$_s$ — Seasonal MA(1) is a Restricted MA(s):

$$X_t = \varepsilon_t + \theta_1^{(s)} \varepsilon_{t-s}.$$

More generally, AR$(P)_s$ and MA$(Q)_s$ use lags at multiples of s:

$$X_t = \phi_1^{(s)} X_{t-s} + \cdots + \phi_P^{(s)} X_{t-Ps} + \varepsilon_t, \qquad X_t = \varepsilon_t + \theta_1^{(s)} \varepsilon_{t-s} + \cdots + \theta_Q^{(s)} \varepsilon_{t-Qs}.$$

Seasonal ARMA$(P, Q)_s$

Definition 7.7: Restricted Seasonal ARMA$(P, Q)_s$

$$X_t = \phi_1^{(s)} X_{t-s} + \cdots + \phi_P^{(s)} X_{t-Ps} + \varepsilon_t + \theta_1^{(s)} \varepsilon_{t-s} + \cdots + \theta_Q^{(s)} \varepsilon_{t-Qs},$$

or equivalently in backshift operator notation:

$$\Phi^{(s)}(B^s)\, X_t = \Theta^{(s)}(B^s)\, \varepsilon_t,$$

where

$$\Phi^{(s)}(B^s) = 1 - \Phi_1^{(s)} B^s - \cdots - \Phi_P^{(s)} B^{Ps}, \qquad \Theta^{(s)}(B^s) = 1 + \Theta_1^{(s)} B^s + \cdots + \Theta_Q^{(s)} B^{Qs}.$$

- $\{X_t\}$ is stationary and causal iff all roots of $\Phi^{(s)}(z^s) = 0$ lie outside the unit circle.
- $\{X_t\}$ is invertible iff all roots of $\Theta^{(s)}(z^s) = 0$ lie outside the unit circle.

Mixed seasonal model: ARMA$(p, q)(P, Q)_s$

In practice one uses a mixed model combining non-seasonal and seasonal components:

$$\phi(B)\,\Phi^{(s)}(B^s)\,X_t = \theta(B)\,\Theta^{(s)}(B^s)\,\varepsilon_t.$$

Example 7.5: ARMA(1,1)(1,1)$_s$

With $\phi(B) = 1 - \phi B$, $\Phi^{(s)}(B^s) = 1 - \Phi B^s$:

$$(1 - \phi B)(1 - \Phi B^s)\,X_t = (1 + \theta B)(1 + \Theta B^s)\,\varepsilon_t.$$

Expanding the left-hand side:

$$X_t = \phi X_{t-1} + \Phi X_{t-s} - \phi\Phi X_{t-s-1} + \varepsilon_t + \theta\varepsilon_{t-1} + \Theta\varepsilon_{t-s} + \theta\Theta\varepsilon_{t-s-1}.$$

ARMA$(p, q)(P, Q)_s$ is a restricted ARMA$(p + Ps, q + Qs)$ with zeros at non-seasonal and non-multiple-of-s lags.

Seasonal differencing

If $\{X_t\}$ is non-stationary due to seasonality, apply the seasonal difference operator:

$$\nabla_s X_t = X_t - X_{t-s} = (1 - B^s)\,X_t.$$

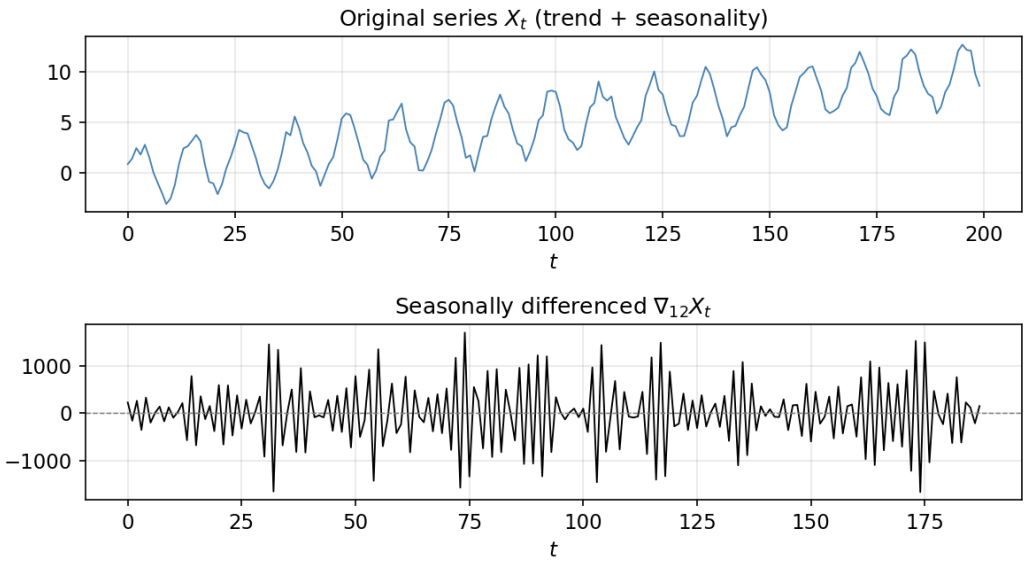


Figure 7.4: Original series with trend and period-12 seasonality (top); after seasonal differencing ∇_{12} (bottom).

SARIMA$(p,d,q)(P,D,Q)_s$: definition

Definition 7.8: SARIMA$(p,d,q)(P,D,Q)_s$

$\{X_t\}$ follows a *SARIMA$(p,d,q)(P,D,Q)_s$* model if $Y_t = \nabla^d \nabla_s^D X_t$ follows a causal, invertible ARMA$(p,q)(P,Q)_s$ process:

$$\phi(B)\,\Phi^{(s)}(B^s)\,(1-B)^d(1-B^s)^D X_t = \theta(B)\,\Theta^{(s)}(B^s)\,\varepsilon_t.$$

- s: period of seasonality.
- D: order of seasonal differencing.
- d: order of non-seasonal differencing.
- This is a non-stationary process; equivalently, $\Phi^*(B)X_t = \Theta^*(B)\varepsilon_t$ where $\Phi^*(B)$ has roots on and outside the unit circle.
- In full generality: SARIMA $\equiv$ ARMA$(p+Ps+d+Ds, q+Qs)$ with the constraint that $d+Ds$ roots lie on the unit circle.

Seasonal differencing: example

Example 7.6: ARIMA(1,1,1)$_{12}$ (common for monthly economic data)

Let $Y_t = \nabla_{12} X_t = X_t - X_{t-12}$. Assume $Y_t \sim$ ARMA(1,1):

$$(1-\phi B)\,Y_t = (1+\theta B)\,\varepsilon_t.$$

Substituting:

$$(1-\phi B)(1-B^{12})\,X_t = (1+\theta B)\,\varepsilon_t.$$

Expanding:

$$X_t = \phi X_{t-1} + X_{t-12} - \phi X_{t-13} + \varepsilon_t + \theta\varepsilon_{t-1}.$$

The roots of $\Phi^*(z) = (1-\phi z)(1-z^{12}) = 0$ are $1/\phi$ (outside the unit circle) and the twelve 12th roots of unity (on the unit circle).

7.5 Summary

Remark 7.6: Lecture 7 Summary

ARMA(p,q): $\phi(B)X_t = \theta(B)\varepsilon_t$; causal stationary iff roots of $\phi(z) = 0$ outside unit circle. Invertible $\Leftrightarrow$ roots of $\theta(z) = 0$ outside unit circle.

ACVF of ARMA: For $h > q$: $\gamma(h) = \phi_1\gamma(h-1) + \cdots + \phi_p\gamma(h-p)$ (AR recurrence). Lags $0, \ldots, q$ require solving a linear system.

ARIMA(p,d,q): $\phi(B)(1-B)^d X_t = \theta(B)\varepsilon_t$. Non-stationary for $d \geq 1$. $X_t \sim I(d)$ means $\nabla^d X_t$ is covariance stationary.

SARIMA$(p,d,q)(P,D,Q)_s$: Adds seasonal AR/MA polynomials in B^s and seasonal differencing $(1-B^s)^D$. Most practical model for seasonal data.

Chapter 8

Lecture 8: Forecasting with BLP and the PACF

8.1 Forecasting: Goals and Setup

Forecasting with time-series models

Goal: given observed data $x_{1:n} = \{x_1, x_2, \ldots, x_n\}$, predict future values

$$x_{n+m}, \qquad m = 1, 2, 3, \ldots$$

- We write x^n_{n+m} for the forecast of x_{n+m} based on $x_{1:n}$.
- Throughout this section we assume $\{x_t\}$ is stationary with known model parameters.
- We use our knowledge of the underlying model to make reliable forecasts.

MMSE predictor

Definition 8.1: Optimal (Minimum Mean Square Error) Predictor

Among all predictors $g(x_{1:n})$, the one that minimizes the mean square error

$$\mathbb{E}[x_{n+m} - g(x_{1:n})]^2$$

is called the *MMSE predictor*. The MMSE predictor is given by the conditional expectation:

$$x^n_{n+m} = \mathbb{E}(x_{n+m} \mid x_{1:n}).$$

In general this requires specifying the full joint distribution (which is unknown in most cases). The above theorem does not explicitly give the form of the predictor g. For practical use we restrict to linear predictors.

8.2 Best Linear Predictors (BLP)

Linear predictors

Definition 8.2: Linear Predictor

A *linear predictor* of x_{n+m} given $x_{1:n}$, x^n_{n+m} has the form

$$x^n_{n+m} = g(x_{1:n}) = \alpha_0 + \sum_{k=1}^{n} \alpha_k x_k,$$

where $\alpha_0, \alpha_1, \ldots, \alpha_n \in \mathbb{R}$.

Remark 8.1: Note

Writing in rigorous notations, the α's should depend on n and m, but for now we drop the dependence from the notation for simplicity if there is no confusion. For example:

- If $n = m = 1$, then x^1_2 is the one-step-ahead linear forecast of x_2 given x_1. That is, $x^1_2 = \alpha_0 + \alpha_1 x_1$.
- But if $n = 2$ and $m = 1$, x^2_3 is the one-step-ahead linear forecast of x_3 given x_1 and x_2. That is, $x^2_3 = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2$.

The values of α_0 and α_1 differ in the two cases.

Best linear predictor

Definition 8.3: Best Linear Predictor (BLP)

The linear predictor that minimises the mean square error

$$\mathbb{E}[x_{n+m} - x^n_{n+m}]^2$$

is called the *best linear predictor (BLP)* and is denoted by $P_{(x_1,\ldots,x_n)}x_{n+m}$ or simply $P_n x_{n+m}$.

Remark 8.2: Remark

If $\{x_t\}$ is a Gaussian time series, the BLP equals the conditional expectation, so no information is lost by restricting to linear predictors.

BLP: prediction equations

The BLP $x^n_{n+m} = \alpha_0 + \sum_{i=1}^n \alpha_i x_i$ is found by solving the prediction equations:

$$\mathbb{E}\big[(x_{n+m} - x^n_{n+m})\, x_i\big] = 0, \qquad i = 0, 1, \ldots, n,$$

where $x_0 = 1$.

Interpretation: the forecast error must be orthogonal to every observed value (and to the constant 1).

Non-zero mean. The $i = 0$ equation gives $\alpha_0 = \big(1 - \sum_{i=1}^{n} \alpha_i\big)\mu$, so

$$x_{n+m}^{n} = \mu + \sum_{i=1}^{n} \alpha_i (x_i - \mu).$$

Without loss of generality we assume $\mu = 0$ going forward. In other words, one works with the mean-centered series.

Derivation

$$\begin{aligned}
\mathbb{E}[X_t] = \mu, \qquad & X_{n+m}^{n} = \alpha_0 + \sum_{i=1}^{n} \alpha_i X_i, \\
\mathbb{E}\big[(X_{n+m} - X_{n+m}^{n}) X_0\big] = 0, \qquad & X_0 = 1, \\
\mathbb{E}[X_{n+m} - X_{n+m}^{n}] &= 0, \\
\mathbb{E}[X_{n+m}] &= \mathbb{E}[X_{n+m}^{n}], \\
\mu &= \mathbb{E}\left[\alpha_0 + \sum_{i=1}^{n} \alpha_i X_i\right] \\
&= \alpha_0 + \sum_{i=1}^{n} \alpha_i \mathbb{E}[X_i] \\
&= \alpha_0 + \mu \sum_{i=1}^{n} \alpha_i .
\end{aligned}$$

$$\begin{aligned}
\alpha_0 &= \mu - \mu \sum_{i=1}^{n} \alpha_i \\
&= \left(1 - \sum_{i=1}^{n} \alpha_i\right) \mu .
\end{aligned}$$

$$\begin{aligned}
X_{n+m}^{n} &= \left(1 - \sum_{i=1}^{n} \alpha_i\right) \mu + \sum_{i=1}^{n} \alpha_i X_i \\
&= \mu - \mu \sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n} \alpha_i X_i \\
&= \mu + \sum_{i=1}^{n} \alpha_i (X_i - \mu).
\end{aligned}$$

BLP: deriving the prediction equations (Step 1)

Setup. Let $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_n)$ and define the mean-square error as a function of α:

$$S(\alpha) = \mathbb{E}\left[\left(x_{n+m} - \alpha_0 - \sum_{i=1}^{n} \alpha_i x_i\right)^2\right].$$

We minimise S over all choices of α. First-order condition w.r.t. α_0:

$$\frac{\partial S}{\partial \alpha_0} = -2\,\mathbb{E}\left[x_{n+m} - \alpha_0 - \sum_{i=1}^{n} \alpha_i x_i\right] = 0.$$

Using $\mathbb{E}(x_t) = \mu$ this gives:

$$\mu - \alpha_0 - \sum_{i=1}^{n} \alpha_i \mu = 0 \implies \alpha_0 = \mu\left(1 - \sum_{i=1}^{n} \alpha_i\right).$$

BLP: deriving the prediction equations (Step 2)

First-order condition w.r.t. α_j, $j = 1, \ldots, n$:

$$\frac{\partial S}{\partial \alpha_j} = -2\,\mathbb{E}\left[\left(x_{n+m} - \alpha_0 - \sum_{i=1}^{n} \alpha_i x_i\right) x_j\right] = 0.$$

Expanding and using $\mathbb{E}(x_t) = \mu$, $\mathbb{E}(x_s x_t) = \gamma(s-t) + \mu^2$:

$$\mathbb{E}[x_{n+m} x_j] - \alpha_0 \mu - \sum_{i=1}^{n} \alpha_i \mathbb{E}[x_i x_j] = 0.$$

Substituting $\alpha_0 = \mu(1 - \sum \alpha_i)$ and rearranging:

$$\gamma_{n+m-j} = \sum_{i=1}^{n} \alpha_i \gamma_{i-j}, \qquad j = 1, \ldots, n.$$

BLP in matrix form

For $\mu = 0$, denote $\alpha_n = (\alpha_1, \ldots, \alpha_n)$. The prediction equations reduce to

$$\Gamma_n \alpha_n = \gamma_n(m),$$

where

$$\Gamma_n = \big[\gamma(i-j)\big]_{i,j=1}^{n} \text{ (Toeplitz matrix)}, \qquad \gamma_n(m) = \begin{pmatrix} \gamma(m+n-1) \\ \gamma(m+n-2) \\ \vdots \\ \gamma(m) \end{pmatrix}.$$

The BLP coefficient vector is

$$\alpha_n^{\text{BLP}} = \Gamma_n^{-1} \gamma_n(m),$$

and the m-step-ahead BLP is $x_{n+m}^n = \alpha_n \cdot (x_1, x_2, \ldots, x_n)$.

BLP: properties and prediction error derivation

For $\mu = 0$ the BLP $P_n x_{n+m} = \alpha_n \cdot x$ where $x = (x_1, x_2, \ldots, x_n)$ satisfies three key properties:

1. $\mathbb{E}[x_{n+m} - P_n x_{n+m}] = 0$,
2. $\mathbb{E}[(x_{n+m} - P_n x_{n+m})\, x_j] = 0$ for $j = 1, \ldots, n$,
3. the mean-square prediction error equals (derived below).

Deriving the MSPE:

$$\begin{aligned}\mathbb{E}[(x_{n+m} - P_n x_{n+m})^2] &= \mathbb{E}[(x_{n+m} - \alpha_n^\top x)^2] \\ &= \gamma_0 + \alpha_n^\top \Gamma_n \alpha_n - 2\alpha_n^\top \gamma_n(m) \\ &= \gamma_0 + \alpha_n^\top \gamma_n(m) - 2\alpha_n^\top \gamma_n(m).\end{aligned}$$

Theorem 8.1: Minimum MSPE

$$P_{n+m}^n = \mathbb{E}[x_{n+m} - P_n x_{n+m}]^2 = \gamma_0 - \alpha_n^\top \gamma_n(m).$$

Prediction error and prediction intervals

Mean-square prediction error:

$$P_{n+m}^n = \mathbb{E}\big(x_{n+m} - x_{n+m}^n\big)^2.$$

For $\mu = 0$, this simplifies to $P_{n+m}^n = \gamma(0) - \alpha_n^\top \gamma_n(m)$.

Prediction interval. A $100(1-\alpha)\%$ prediction interval is

$$x_{n+m}^n \pm C_{\alpha/2}\sqrt{P_{n+m}^n},$$

where $C_{\alpha/2}$ is the appropriate quantile. For Gaussian processes, $C_{0.025} = 1.96$ gives a 95% PI.

Remark 8.3: Note

Prediction uncertainty grows with horizon m.

BLP: prediction equations — summary

Non-zero mean. The α_0 equation gives $\alpha_0 = \big(1 - \sum_{i=1}^n \alpha_i\big)\mu$, so the BLP can be written as:

$$P_n x_{n+m} = \mu + \sum_{i=1}^n \alpha_i (x_i - \mu).$$

Special case: uncorrelated series. If $\{x_t\}$ is uncorrelated ($\gamma_h = 0$ for $h \neq 0$), then $\gamma_n(m) = 0$ and $\alpha_n = 0$, so $P_n x_{n+h} = \mu$ — forecast is always the mean.

Worked example: BLP for AR(2)

Example 8.1: BLP for AR(2)

Setup: $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \varepsilon_t$, covariance stationary.
Predict x_2 from x_1: the BLP $x_2^1 = \alpha x_1$ minimizes $\mathbb{E}[(x_2 - \alpha x_1)]^2$. Setting $\partial/\partial\alpha = 0$:

$$\mathbb{E}(x_2 - \alpha x_1)x_1 = 0 \Rightarrow \gamma(1) = \alpha\gamma(0) \Rightarrow \alpha = \rho(1).$$

Predict x_3 from x_2, x_1: $x_3^2 = \alpha x_2 + \beta x_1$. The prediction equations are

$$\underbrace{\begin{pmatrix}\gamma(0) & \gamma(1)\\ \gamma(1) & \gamma(0)\end{pmatrix}}_{\Gamma_2}\begin{pmatrix}\alpha\\ \beta\end{pmatrix} = \begin{pmatrix}\gamma(1)\\ \gamma(2)\end{pmatrix} \Rightarrow \begin{pmatrix}\alpha\\ \beta\end{pmatrix} = \begin{pmatrix}\phi_1\\ \phi_2\end{pmatrix}.$$

So for $n \geq 2$: $x_{n+1}^n = \phi_1 x_n + \phi_2 x_{n-1}$.

BLP for AR(p): the key theorem

Theorem 8.2: BLP for AR(p)

Let $\{x_t\}$ be a causal AR(p) process with $\varepsilon_t \sim \text{WN}(0, \sigma^2)$. For any $n \geq p$, the BLP of x_{n+1} based on $\{x_n, x_{n-1}, \ldots, x_1\}$ depends only on the most recent p observations:

$$P_{(x_n,\ldots,x_1)}x_{n+1} = \phi_1 x_n + \phi_2 x_{n-1} + \cdots + \phi_p x_{n-p+1}.$$

The older observations $x_{n-p}, x_{n-p-1}, \ldots, x_1$ receive zero weight in the BLP. Moreover, the BLP coefficients equal the AR coefficients exactly.

BLP for AR(p): missing old observations

Why does the BLP have only the p latest observations? The BLP uses all n observations in principle even though the model only looks back p steps. The older lags $x_{n-p}, \ldots, x_1$ receive zero weight in the BLP, not because they are irrelevant to x_{n+1} in general, but because:

- x_{n+1} depends on $x_{n-p}, \ldots, x_1$ only through the intermediaries $x_n, \ldots, x_{n-p+1}$, and
- once those intermediaries are in the conditioning set, the older lags add no further information.

Forecasting AR(p) processes

One-step ahead (for $n \geq p$):

$$x_{n+1}^n = \phi_1 x_n + \phi_2 x_{n-1} + \cdots + \phi_p x_{n-p+1}.$$

The BLP coefficients equal the AR coefficients exactly.

Multi-step ahead: apply the one-step formula recursively, substituting forecasts for unobserved future values:

$$\begin{aligned} x_{n+1}^n &= \phi_1 x_n + \phi_2 x_{n-1} + \cdots + \phi_p x_{n-p+1},\\ x_{n+2}^n &= \phi_1 x_{n+1}^n + \phi_2 x_n + \cdots + \phi_p x_{n-p+2},\\ &\vdots \end{aligned}$$

Prediction variance for AR(1):

$$P_{n+m}^n = \sigma^2 \sum_{k=0}^{m-1} \phi^{2k} = \sigma^2 \frac{1-\phi^{2m}}{1-\phi^2} \xrightarrow{m\to\infty} \frac{\sigma^2}{1-\phi^2} = \gamma(0).$$

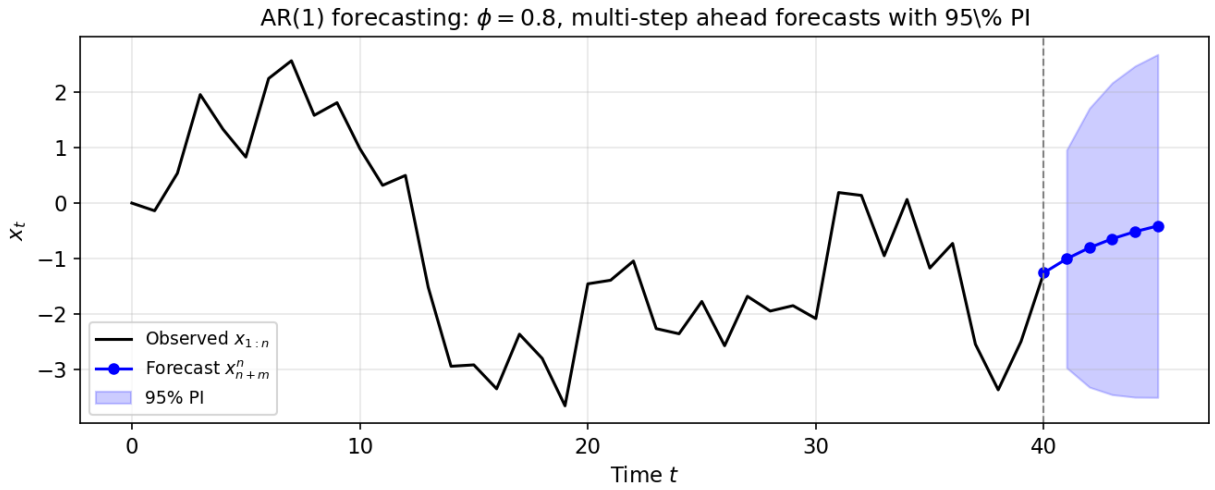


Figure 8.1: AR(1) forecasting illustration. Forecasts revert toward zero (the process mean) geometrically. The 95% PI widens with horizon but is bounded by $\gamma(0)$.

Forecasting MA(q) processes

For $x_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q}$: the key observation is that $\mathbb{E}[x_{n+m} \mid x_{1:n}]$ depends on past innovations $\varepsilon_n, \varepsilon_{n-1}, \ldots$ which can be recovered from observed data.

One-step ahead ($m = 1$):

$$x^n_{n+1} = \theta_1\varepsilon_n + \theta_2\varepsilon_{n-1} + \cdots + \theta_q\varepsilon_{n-q+1}.$$

Multi-step ahead ($m > q$):

$$x^n_{n+m} = 0 \qquad \text{(forecast equals the process mean).}$$

The MA(q) has finite memory: shocks more than q steps ago have no effect on future values.

Remark 8.4: Contrast with AR

AR forecasts never exactly equal the mean at any finite horizon; MA forecasts do so after $m > q$ steps.

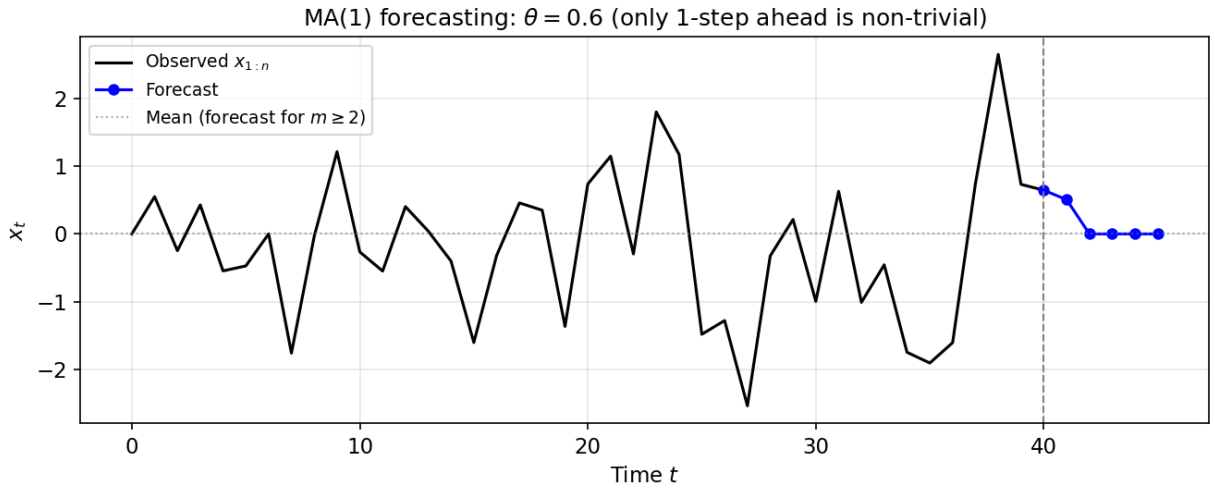


Figure 8.2: MA(1) forecasting illustration. One-step forecast is $\hat{x}_{n+1} = \theta\varepsilon_n$ (non-trivial). For $m \geq 2$: forecast is identically zero. MA processes have no long-range memory.

Applications of the BLP framework

The second-order prediction framework extends beyond simple forecasting.

1. **Missing value estimation.** If x_3 is missing, treat it as the forecast to be made and use available observations $(x_1, x_2, x_4, x_5, \ldots)$ as linear predictors. The BLP gives the best linear estimate of the missing value.
2. **Back-forecasting.** In MA (and ARMA) models, initial innovations $\varepsilon_0, \varepsilon_{-1}, \ldots$ are unobserved. Back-forecasting applies the BLP backwards in time to impute these, enabling conditional likelihood computation.
3. **Defining the PACF.** PACF at lag k involves a forward forecast $P_{(x_2,\ldots,x_k)}x_{k+1}$ and a back-cast $P_{(x_2,\ldots,x_k)}x_1$, both obtained via the BLP. This is the subject of the next section.

Remark 8.5: Remark

The BLP framework requires only finite second moments — no distributional assumptions beyond stationarity.

8.3 Partial Autocorrelation Function (PACF)

Why do we need PACF?

- For MA(q) models, the ACF will be zero for lags greater than q. Moreover, because $\theta_q \neq 0$, the ACF will not be zero at lag q. We can use this property to identify MA models.
- If the process, however, is ARMA or AR, the ACF alone tells us little about the orders of dependence.
- The partial autocorrelation function (PACF) can be used to identify AR models.

Motivation: why ACF is not enough

Consider a causal AR(1) process $x_t = \phi x_{t-1} + \varepsilon_t$, $|\phi| < 1$. We showed:

$$\gamma(2) = \phi^2\gamma(0) \Rightarrow \rho(2) = \phi^2 \neq 0.$$

But x_t depends on x_{t-2} only through x_{t-1}. **Direct verification:**

$$\text{Cov}(x_t - \phi x_{t-1}, x_{t-2} - \phi x_{t-1}) = \text{Cov}(\varepsilon_t, x_{t-2} - \phi x_{t-1}) = 0.$$

Once we remove the effect of x_{t-1}, the remaining correlation is exactly zero.

Remark 8.6: Key idea

The partial autocorrelation at lag h measures the correlation between x_{t+h} and x_t after removing the linear effect of the intermediate observations $x_{t+1}, \ldots, x_{t+h-1}$.

Partial correlation: general definition

Definition 8.4: Partial Correlation

Given random variables X, Y, Z: regress X on Z to get $\hat{X}$; regress Y on Z to get $\hat{Y}$. The *partial correlation* of X and Y given Z is

$$\rho_{XY|Z} = \text{Corr}(X - \hat{X}, Y - \hat{Y}).$$

This removes the linear effect of Z from both X and Y before computing the correlation.

Connection to PACF: setting $X = x_{t+h}$, $Y = x_t$, $Z = (x_{t+1}, \ldots, x_{t+h-1})$:

$$\hat{X} = P_{(x_{t+1},\ldots,x_{t+h-1})}x_{t+h} = \alpha_1 x_{t+h-1} + \cdots + \alpha_{h-1}x_{t+1},$$
$$\hat{Y} = P_{(x_{t+1},\ldots,x_{t+h-1})}x_t = \alpha_1 x_{t+1} + \cdots + \alpha_{h-1}x_{t+h-1}.$$

By stationarity, the coefficients α_j are the same in both regressions.

Regression for PACF

Consider the mean-zero stationary time series x_t and assume $h \geq 2$.

1. Let $\hat{x}_{t+h}$ denote the regression of x_{t+h} on $\{x_{t+h-1}, x_{t+h-2}, \ldots, x_{t+1}\}$. We can write

   $$\hat{x}_{t+h} = \alpha_1 x_{t+h-1} + \alpha_2 x_{t+h-2} + \cdots + \alpha_{h-1}x_{t+1}.$$

   Note: No intercept term is needed because the mean of x_t is zero.
2. Let $\hat{x}_t$ denote the regression of x_t on $\{x_{t+1}, x_{t+2}, \ldots, x_{t+h-1}\}$. Then

   $$\hat{x}_t = \alpha_1 x_{t+1} + \alpha_2 x_{t+2} + \cdots + \alpha_{h-1}x_{t+h-1}.$$

   Note: Because of stationarity, the coefficients $\alpha_1, \alpha_2, \ldots, \alpha_{h-1}$ are the same. The term regression here refers to regression in the population sense. That is, $\hat{x}_{t+h}$ is the linear combination of $x_{t+h-1}, x_{t+h-2}, \ldots, x_{t+1}$ that minimizes the mean squared error $\mathbb{E}[(x_{t+h} - \sum_{j=1}^{h-1} \alpha_j x_{t+j})^2]$.

PACF: formal definition

Definition 8.5: PACF

For a mean-zero stationary process $\{x_t\}$, the *partial autocorrelation function (PACF)* $\alpha(h)$ or ϕ_{hh} at lag h is:

$$\phi_{11} = \text{Corr}(x_{t+1}, x_t) = \rho(1),$$

and for $h \geq 2$:

$$\phi_{hh} = \text{Corr}\big(x_{t+h} - P_{(x_{t+h-1},\ldots,x_{t+1})}x_{t+h},\ x_t - P_{(x_{t+1},\ldots,x_{t+h-1})}x_t\big).$$

Equivalent mechanical definition: ϕ_{hh} is the coefficient of x_1 (i.e. the coefficient on the most distant observation) in the BLP of x_{h+1} based on $x_1, x_2, \ldots, x_h$.

8.4 Computing the PACF: First Principles

PACF via forward and back-forecasting

For lag $h \geq 2$, the PACF is defined as:

$$\alpha(h) = \text{Corr}\big(x_{h+1} - P_{(x_2,\ldots,x_h)}x_{h+1},\ x_1 - P_{(x_2,\ldots,x_h)}x_1\big).$$

- $P_{(x_2,\ldots,x_h)}x_{h+1}$ is the forward forecast of x_{h+1} from $\{x_2, \ldots, x_h\}$.
- $P_{(x_2,\ldots,x_h)}x_1$ is the back-cast of x_1 from $\{x_2, \ldots, x_h\}$.
- Both are obtained through the BLP.

Alternative (mechanical) definition. For $h \geq 2$: $\alpha(h)$ is the coefficient of x_1 in the BLP of x_{h+1} based on $(x_1, x_2, \ldots, x_h)$.

PACF of AR(1): $\alpha(2)$

AR(1): $x_t = \phi x_{t-1} + \varepsilon_t$, $|\phi| < 1$.

Forward forecast: find $P_{x_2}x_3 = \alpha x_2$.

$$S(\alpha) = \mathbb{E}[(x_3 - \alpha x_2)^2], \quad \frac{\partial S}{\partial \alpha} = 0 \Rightarrow \mathbb{E}[(x_3 - \alpha x_2)x_2] = 0 \Rightarrow \alpha = \frac{\gamma_1}{\gamma_0} = \phi.$$

So $P_{x_2}x_3 = \phi x_2$.

Back-cast: find $P_{x_2}x_1 = \beta x_2$.

$$S(\beta) = \mathbb{E}[(x_1 - \beta x_2)^2], \quad \mathbb{E}[(x_1 - \beta x_2)x_2] = 0 \Rightarrow \beta = \frac{\gamma_1}{\gamma_0} = \phi.$$

So $P_{x_2}x_1 = \phi x_2$.

PACF at lag 2:

$$\alpha(2) = \text{Corr}(x_3 - \phi x_2, x_1 - \phi x_2) = \text{Cov}(\varepsilon_3, x_1 - \phi x_2) = 0.$$

PACF of AR(1): $\alpha(3)$

Forward forecast $P_{(x_2,x_3)}x_4 = \alpha_1 x_3 + \alpha_2 x_2$: BLP equations give $\alpha_1 = \phi$, $\alpha_2 = 0$, so $P_{(x_2,x_3)}x_4 = \phi x_3$.

Back-cast $P_{(x_2,x_3)}x_1$: by stationarity, $P_{(x_2,x_3)}x_1 = \alpha_1 x_2 + \alpha_2 x_3$. The BLP equations for the back-cast (with roles reversed) yield $\alpha_1 = \phi$, $\alpha_2 = 0$, so $P_{(x_2,x_3)}x_1 = \phi x_2$.

PACF at lag 3:

$$\alpha(3) = \text{Corr}(x_4 - \phi x_3, x_1 - \phi x_2) = \text{Corr}(\varepsilon_4, x_1 - \phi x_2) = 0,$$

since ε_4 is independent of all past values.

Remark 8.7: General AR(1) result

For all $k \geq 2$: $\alpha(k) = \text{Corr}(\varepsilon_{k+1}, x_1 - \alpha_1 x_2 - \cdots - \alpha_{k-1} x_k) = 0$.

PACF of AR(2): $\alpha(2)$

AR(2): $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \varepsilon_t$.

Forward forecast $P_{x_2} x_3 = \alpha x_2$:

$$\mathbb{E}[(x_3 - \alpha x_2) x_2] = 0 \Rightarrow \gamma_1 = \alpha \gamma_0 \Rightarrow \alpha = \rho_1 = \phi_1.$$

Back-cast $P_{x_2} x_1 = \beta x_2$:

$$\mathbb{E}[(x_1 - \beta x_2) x_2] = 0 \Rightarrow \gamma_1 = \beta \gamma_0 \Rightarrow \beta = \rho_1 = \phi_1.$$

So $P_{x_2} x_3 = \phi_1 x_2$ and $P_{x_2} x_1 = \phi_1 x_2$.

PACF at lag 2:

$$\alpha(2) = \text{Corr}(x_3 - \phi_1 x_2, x_1 - \phi_1 x_2) = \frac{\text{Cov}(x_3 - \phi_1 x_2, x_1 - \phi_1 x_2)}{[\text{Var}(x_3 - \phi_1 x_2)\text{Var}(x_1 - \phi_1 x_2)]^{1/2}}.$$

PACF of AR(2): computing $\alpha(2)$ explicitly

Variances (both equal by stationarity):

$$\text{Var}(x_3 - \phi_1 x_2) = \gamma_0 + \phi_1^2 \gamma_0 - 2\phi_1 \gamma_1.$$

Covariance:

$$\begin{aligned} \text{Cov}(x_3 - \phi_1 x_2, x_1 - \phi_1 x_2) &= \gamma_2 - \phi_1 \gamma_1 - \phi_1 \gamma_1 + \phi_1^2 \gamma_0 \\ &= \gamma_2 - 2\phi_1 \gamma_1 + \phi_1^2 \gamma_0. \end{aligned}$$

After simplification:

$$\alpha(2) = \frac{\gamma_2 - 2\phi_1 \gamma_1 + \phi_1^2 \gamma_0}{\gamma_0 - 2\phi_1 \gamma_1 + \phi_1^2 \gamma_0} \neq 0.$$

Remark 8.8: Result

$\alpha(2) \neq 0$. $\alpha(3) = 0$ (same argument as AR(1) case for $k > p$). PACF cuts off after lag 2.

PACF of MA(1): $\alpha(2)$ from first principles

MA(1): $x_t = \varepsilon_t + \theta\varepsilon_{t-1}$, $\theta^* = \rho_1 = \theta/(1+\theta^2)$, $\rho_k = 0$ for $k \geq 2$.

Forward: $P_{x_2}x_3 = \theta^* x_2$ (same as AR(1) calculation above). **Back-cast:** $P_{x_2}x_1 = \theta^* x_2$.

PACF at lag 2:

$$\alpha(2) = \frac{\text{Cov}(x_3 - \theta^* x_2, x_1 - \theta^* x_2)}{[\text{Var}(x_3 - \theta^* x_2)\text{Var}(x_1 - \theta^* x_2)]^{1/2}}.$$

Numerator:

$$\text{Cov}(x_3 - \theta^* x_2, x_1 - \theta^* x_2) = \gamma_2 - \theta^*\gamma_1 - \theta^*\gamma_1 + (\theta^*)^2\gamma_0 = -(\theta^*)^2\gamma_0.$$

(Using $\gamma_2 = 0$ and $\gamma_1 = \theta^*\gamma_0$.) **Denominator:** $\text{Var}(x_3 - \theta^* x_2) = \gamma_0(1 + (\theta^*)^2) - 2(\theta^*)^2\gamma_0 = \gamma_0(1 - (\theta^*)^2)$.

$$\alpha(2) = \frac{-(\theta^*)^2\gamma_0}{\gamma_0(1-(\theta^*)^2)} = \frac{-(\theta^*)^2}{1-(\theta^*)^2} \neq 0.$$

PACF of MA(1): general formula

Continuing the pattern, for general $k \geq 2$:

$$\alpha(k) = \frac{-(-\theta)^k}{1 + \theta^2 + \theta^4 + \cdots + \theta^{2k}} \neq 0.$$

Key features:

- $|\alpha(k)| \to 0$ geometrically as $k \to \infty$ (tails off).
- The sign alternates if $\theta > 0$: $\alpha(2) < 0$, $\alpha(3) > 0$, ...
- This contrasts sharply with AR(p) processes where $\alpha(k) = 0$ exactly for $k > p$.

Remark 8.9: Remark

Computing $\hat{\alpha}(k)$ from data will give a value close to 0 but not exactly 0 for the MA(1) case, since $\alpha(k)$ is a continuous r.v. — $P(\alpha(k) = 0) = 0$ exactly. To determine the order of AR(p) process since PACF eventually becomes 0, we use the asymptotic test $\hat{\alpha}(k) \overset{\text{asym}}{\sim} N(0, 1/n)$ for $k > p$. This test applies only to AR processes.

8.5 ACF/PACF Plots for AR, MA, ARMA

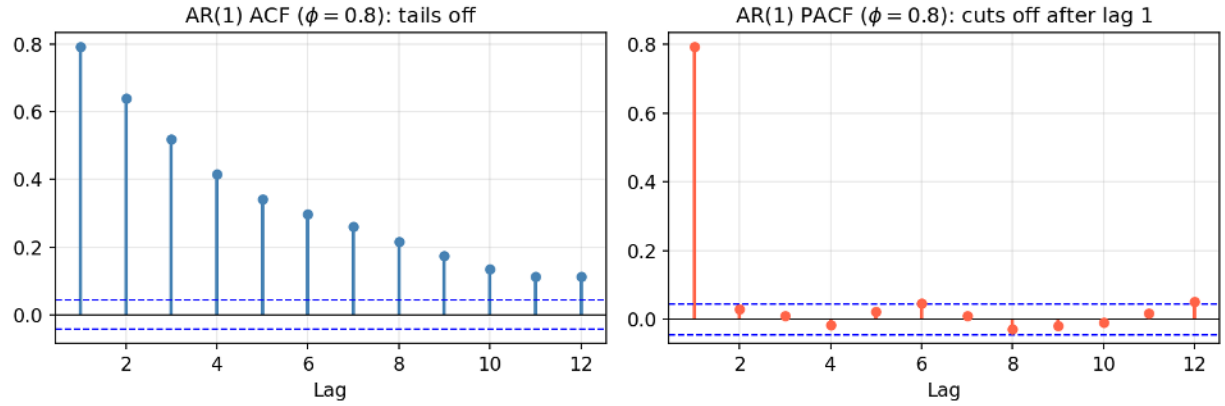


Figure 8.3: AR(1): ACF exponential decay (tails off); PACF single spike at lag 1, zero thereafter (cuts off after lag 1). Dashed lines are $\pm 1.96/\sqrt{n}$ significance bands.

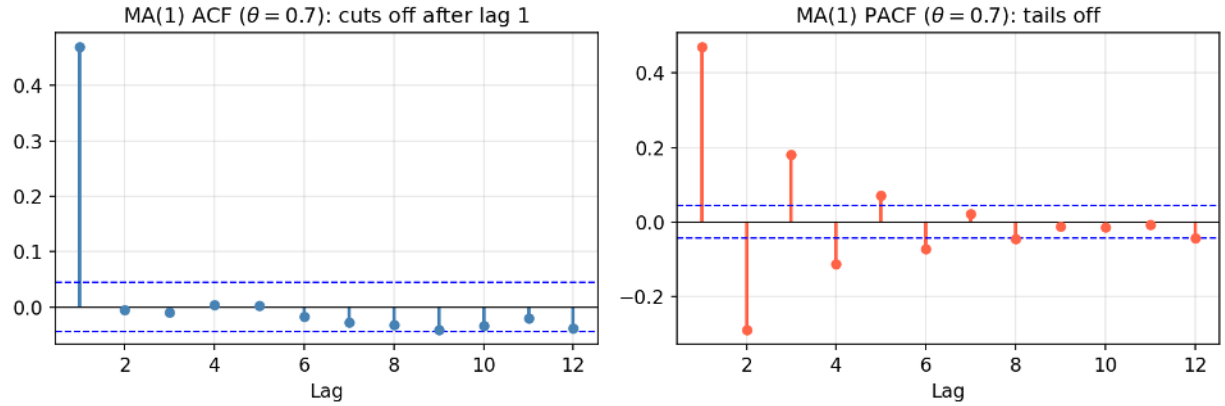


Figure 8.4: MA(1): ACF single non-zero spike at lag 1 (cuts off); PACF geometrically decaying spikes (tails off).

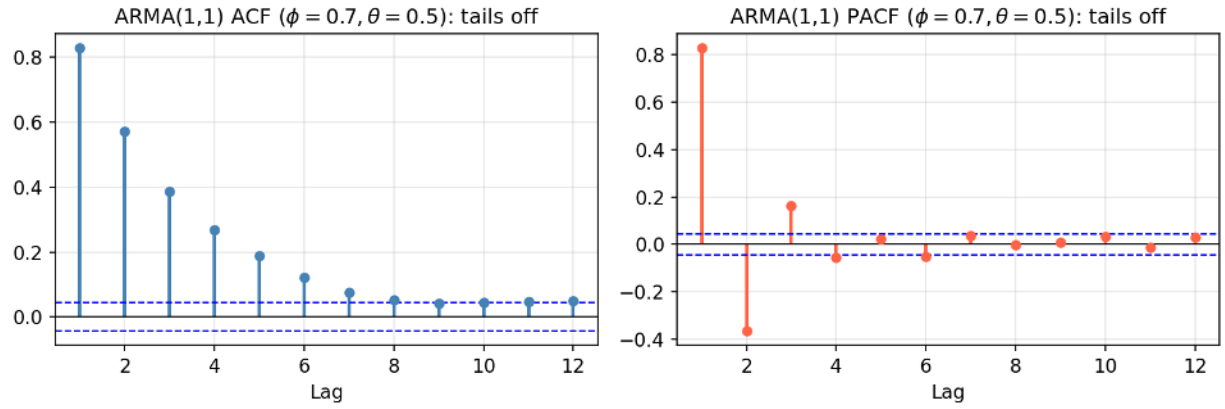


Figure 8.5: ARMA(1,1): ACF decays geometrically (tails off); PACF also tails off. Neither cuts off ⇒ need information criteria (AIC/BIC) to determine order.

8.6 Model Identification Summary

ACF and PACF: model identification table

Model	**ACF** $\rho(h)$	**PACF** $\alpha(h)$
AR(1)	Tails off: $\rho(h) = \phi^h$	Cuts off after lag 1
AR(p)	Tails off: mixture of geometric / damped oscillation	Cuts off after lag p: exactly p non-zero spikes
MA(1)	Cuts off after lag 1	Tails off
MA(q)	Cuts off after lag q: exactly q non-zero spikes	Tails off
ARMA(1,1)	Tails off	Tails off
ARMA(p, q)	Tails off (pure AR recurrence for $h > q$)	Tails off (pure MA recurrence for $h > p$)

How to use the table: step-by-step

1. Plot the sample ACF and PACF. Add significance bands $\pm 1.96/\sqrt{n}$.
2. Does the ACF cut off? If yes at lag q ⇒ candidate: MA(q).
3. Does the PACF cut off? If yes at lag p ⇒ candidate: AR(p).
4. Both tail off? ⇒ candidate: ARMA(p, q) for some $p, q \geq 1$. Use AIC / BIC to select p and q.
5. Estimate and check residuals. Verify that residuals behave like white noise (Ljung–Box test, residual ACF near zero).

Remark 8.10: Caveat

In finite samples, "cuts off" vs "tails off" can be hard to distinguish. Model selection criteria (AIC, BIC) and formal hypothesis tests should supplement visual inspection. When both ACF and PACF tail off (ARMA case), visual methods fail. Use penalised likelihood criteria.

Information criteria for order selection

Definition 8.6: Akaike Information Criterion (AIC)

$$\mathrm{AIC}(p,q) = -2\log L(\hat{\theta}) + 2K,$$

where L is the maximum likelihood and $K = p + q + 1$ is the number of free parameters (including σ^2).

Definition 8.7: Bayesian Information Criterion (BIC)

$$\mathrm{BIC}(p,q) = -2\log L(\hat{\theta}) + K\log n.$$

BIC penalises complexity more heavily and tends to select more parsimonious models.

Procedure: fit ARMA(p,q) for a grid of (p,q) values; select the model with the smallest AIC (or BIC).

Summary

Remark 8.11: Lecture 8 Summary

Forecasting: BLP minimises MSE among all linear predictors; found by solving $\Gamma_n\alpha_n = \gamma_n(m)$. For AR$(p)$: BLP coefficients = model coefficients; apply recursively for multi-step. For MA(q): forecasts equal zero after horizon q.

PACF: $\alpha(h) = \phi_{hh}$ = last coefficient in the order-h Yule–Walker system.

Model Identification: AR(p) — ACF tails off, PACF cuts off after p. MA(q) — ACF cuts off after q, PACF tails off. ARMA(p,q) — both tail off. Use AIC/BIC when both tail off.

Chapter 9

Lecture 9

9.1 Computing the PACF: Yule–Walker

PACF via regression: the setup

There is a natural way to compute the PACF through a sequence of regressions. For each lag k, consider regressing x_{t+k} on the k intervening observations:

$$x_{t+k} = \phi_{k1} x_{t+k-1} + \phi_{k2} x_{t+k-2} + \cdots + \phi_{kk} x_t + \varepsilon_{t+k},$$

where ε_{t+k} has mean zero and is uncorrelated with x_{t+k-j} for $j = 1, \ldots, k$.

- ϕ_{ki} is the i-th regression coefficient at lag k.
- The **last** coefficient ϕ_{kk} — on x_t, the most distant observation — is the PACF at lag k.
- This is exactly the BLP characterisation: ϕ_{kk} is the coefficient of x_t in the BLP of x_{t+k} given $\{x_{t+1}, \ldots, x_{t+k-1}\}$.

Goal: find ϕ_{kk} for $k = 1, 2, 3, \ldots$ without fitting a regression for each k from scratch.

Deriving the Yule–Walker system

Multiply both sides of the regression equation by x_{t+k-j} and take expectations. Using $\mathbb{E}[\varepsilon_{t+k}\, x_{t+k-j}] = 0$:

$$\gamma(j) = \phi_{k1}\gamma(j-1) + \phi_{k2}\gamma(j-2) + \cdots + \phi_{kk}\gamma(j-k).$$

Dividing through by $\gamma(0)$:

$$\rho(j) = \phi_{k1}\rho(j-1) + \phi_{k2}\rho(j-2) + \cdots + \phi_{kk}\rho(j-k).$$

Evaluating at $j = 1, 2, \ldots, k$ gives a system of k equations in k unknowns $(\phi_{k1}, \ldots, \phi_{kk})$:

$$\begin{aligned}
\rho(1) &= \phi_{k1}\rho(0) + \phi_{k2}\rho(1) + \cdots + \phi_{kk}\rho(k-1), \\
\rho(2) &= \phi_{k1}\rho(1) + \phi_{k2}\rho(0) + \cdots + \phi_{kk}\rho(k-2), \\
&\vdots \\
\rho(k) &= \phi_{k1}\rho(k-1) + \phi_{k2}\rho(k-2) + \cdots + \phi_{kk}\rho(0).
\end{aligned}$$

Yule–Walker system: matrix form

Using $\rho(0) = 1$ and writing $\mathbf{R}_k$ for the $k \times k$ autocorrelation matrix, the system becomes

$$\underbrace{\begin{pmatrix} 1 & \rho(1) & \cdots & \rho(k-1) \\ \rho(1) & 1 & \cdots & \rho(k-2) \\ \vdots & & \ddots & \vdots \\ \rho(k-1) & \rho(k-2) & \cdots & 1 \end{pmatrix}}_{\mathbf{R}_k} \begin{pmatrix} \phi_{k1} \\ \phi_{k2} \\ \vdots \\ \phi_{kk} \end{pmatrix} = \begin{pmatrix} \rho(1) \\ \rho(2) \\ \vdots \\ \rho(k) \end{pmatrix}.$$

Definition 9.1: PACF from Cramer's Rule

The last element of the solution, ϕ_{kk}, can be read off using Cramer's rule — replace the last column of $\mathbf{R}_k$ with the right-hand side vector $(\rho(1), \ldots, \rho(k))^\top$ and divide determinants:

$$\alpha(k) = \phi_{kk} = \frac{\det \mathbf{R}_k^*}{\det \mathbf{R}_k}.$$

Cramer's rule: explicit formulas for $k = 1, 2$

$k = 1$: the system is just $\phi_{11} \cdot 1 = \rho(1)$, so

$$\alpha(1) = \phi_{11} = \rho(1).$$

$k = 2$: replace the last column of $\begin{pmatrix} 1 & \rho(1) \\ \rho(1) & 1 \end{pmatrix}$ with $\begin{pmatrix} \rho(1) \\ \rho(2) \end{pmatrix}$:

$$\alpha(2) = \phi_{22} = \frac{\det \begin{pmatrix} 1 & \rho(1) \\ \rho(1) & \rho(2) \end{pmatrix}}{\det \begin{pmatrix} 1 & \rho(1) \\ \rho(1) & 1 \end{pmatrix}} = \frac{\rho(2) - \rho(1)^2}{1 - \rho(1)^2}.$$

Cramer's rule: explicit formula for $k = 3$

$$\alpha(3) = \phi_{33} = \frac{\det \begin{pmatrix} 1 & \rho(1) & \rho(1) \\ \rho(1) & 1 & \rho(2) \\ \rho(2) & \rho(1) & \rho(3) \end{pmatrix}}{\det \begin{pmatrix} 1 & \rho(1) & \rho(2) \\ \rho(1) & 1 & \rho(1) \\ \rho(2) & \rho(1) & 1 \end{pmatrix}}.$$

Relation to one-step-ahead BLP

For the **one-step-ahead** case ($m = 1$), write the BLP as

$$x_{k+1}^k = \phi_{k1} x_k + \phi_{k2} x_{k-1} + \cdots + \phi_{kk} x_1.$$

(We relabel $\alpha_i \to \phi_{k,k-i+1}$.) The prediction equations become the **Yule–Walker system**:

$$\sum_{j=1}^{k} \phi_{kj}\, \gamma(i-j) = \gamma(i), \qquad i = 1, \ldots, n.$$

Remark 9.1: Key fact

The **last coefficient** ϕ_{kk} is exactly the **PACF at lag** k. This gives a recursive way to compute the PACF: solve the Yule–Walker system of increasing order $k = 1, 2, 3, \ldots$ and record the last coefficient each time.

Worked example: PACF of AR(1) via Yule–Walker

AR(1): $\rho(h) = \phi^h$.

$$k = 1: \quad \alpha(1) = \rho(1) = \phi.$$

$$k = 2: \quad \alpha(2) = \frac{\rho(2) - \rho(1)^2}{1 - \rho(1)^2} = \frac{\phi^2 - \phi^2}{1 - \phi^2} = 0.$$

$k = 3$: substituting $\rho(j) = \phi^j$ into the 3×3 system and solving (or using Cramer's rule) gives $\phi_{33} = 0$.

General $k \geq 2$: the numerator determinant always has two columns that are multiple of each other (since $\rho(j) = \phi^j$ satisfies the AR(1) Yule–Walker recurrence exactly), making it zero.

Remark 9.2: Result

$\alpha(k) = \phi$ for $k = 1$; $\alpha(k) = 0$ for $k \geq 2$. **PACF cuts off after lag 1.**

Worked example: PACF of general AR(p) via Yule–Walker

For a general AR(p), the Yule–Walker equations at order $k = p$ give

$$\mathbf{R}_p \left(\phi_1, \ldots, \phi_p\right)^\top = \left(\rho(1), \ldots, \rho(p)\right)^\top,$$

which is satisfied exactly by the model parameters, so $\phi_{pp} = \phi_p \neq 0$. For $k = p + 1$: the extra equation

$$\rho(p+1) = \phi_{k1}\rho(p) + \cdots + \phi_{k,p+1}\rho(0)$$

is *automatically satisfied* with $\phi_{k,p+1} = 0$, because the AR(p) ACF satisfies $\rho(p+1) = \phi_1\rho(p) + \cdots + \phi_p\rho(1)$ exactly (the Yule–Walker recurrence continues for all $h > 0$). This forces the last coefficient to be zero.

Remark 9.3: General result

For AR(p): $\alpha(k) = \phi_{kk} \neq 0$ for $k \leq p$; $\alpha(k) = 0$ for $k > p$. **PACF cuts off after lag p.**

PACF of MA(1): explicit calculation

MA(1): $x_t = \varepsilon_t + \theta\varepsilon_{t-1}$, $\theta^* \equiv \rho(1) = \theta/(1+\theta^2)$, $\rho(h) = 0$ for $h \geq 2$ (via MA ACF formula).

Lag $h = 1$: $\phi_{11} = \rho(1) = \theta^*$.

Lag $h = 2$:

$$\phi_{22} = \frac{\rho(2) - \rho(1)^2}{1 - \rho(1)^2} = \frac{0 - (\theta^*)^2}{1 - (\theta^*)^2} = \frac{-(\theta^*)^2}{1 - (\theta^*)^2} \neq 0.$$

General $h \geq 2$:

$$\alpha(h) = \phi_{hh} = \frac{-(-\theta)^h}{1 + \theta^2 + \cdots + \theta^{2h}} \neq 0.$$

As $h \to \infty$, $\alpha(h) \to 0$ geometrically.

Remark 9.4: Result: MA(1) PACF

$\phi_{hh} \neq 0$ for all h. **PACF tails off.**

PACF of general AR(p): cuts off after lag p

For AR(p) with $h > p$, the BLP of x_{h+1} based on $x_1, \ldots, x_h$ is

$$P_{(x_1,\ldots,x_h)} x_{h+1} = \phi_1 x_h + \phi_2 x_{h-1} + \cdots + \phi_p x_{h-p+1}.$$

The forecast residual is $x_{h+1} - P_{(\cdots)} x_{h+1} = \varepsilon_{h+1}$. For the *back-cast* residual: since x_1 is already explained by $x_2, \ldots, x_h$ through the AR structure (no further link to x_{h+1}),

$$\alpha(h) = \mathrm{Corr}\big(\varepsilon_{h+1},\ x_1 - P_{(\cdots)} x_1\big) = 0.$$

Remark 9.5: General result

For AR(p): $\alpha(h) = 0$ for all $h > p$. The PACF **cuts off after lag** p.

Asymptotic results for the sample PACF

Theorem 9.1: General Result (Quenouille, 1949)

For a linear process satisfying mild regularity conditions, if the true $\alpha(k) = 0$ at some lag k, then

$$\sqrt{n}\,\hat{\alpha}(k) \xrightarrow{d} N(0,1) \quad \text{as } n \to \infty,$$

i.e. $\hat{\alpha}(k) \overset{\text{asym}}{\sim} N(0, 1/n)$ under $H_0 : \alpha(k) = 0$. The 95% significance bands $\pm 1.96/\sqrt{n}$ follow directly.

Remark. $\alpha(k)$ is a fixed population quantity, not a random variable. The sample PACF $\hat{\alpha}(k)$ is random, and $\hat{\alpha}(k) \neq 0$ in finite samples even when $\alpha(k) = 0$, due to sampling variability.

Asymptotic PACF: AR, MA, and ARMA

Model	**True** $\alpha(k)$	**Asymptotic result**
AR(p)	$\alpha(k) = 0$ for $k > p$	$\hat{\alpha}(k) \overset{\text{asym}}{\sim} N(0, 1/n)$ for $k > p$. Bands are a valid test of order.
MA(q)	$\alpha(k) \neq 0$ for all $k \geq 1$	$\hat{\alpha}(k) \xrightarrow{p} \alpha(k) \neq 0$. Bands are **not** valid; PACF tails off but never cuts off.
ARMA(p, q)	$\alpha(k) \neq 0$ for all $k \geq 1$	Same as MA: bands are **not** valid for any finite lag.

Remark 9.6: Practical implication

The significance bands $\pm 1.96/\sqrt{n}$ are formally valid **only** when testing $H_0 : \alpha(k) = 0$ for $k > p$ under an AR(p) model. Applying them to MA or ARMA processes is informal and can be misleading — both the sample ACF and PACF tail off, and model selection should rely on AIC/BIC rather than band-crossing alone.

9.2 Building ARIMA Models

Steps for building ARIMA models

Fitting an ARIMA model to time series data involves the following steps:

1. Plot the data and interpret the plot.
2. Transform the data if necessary (e.g. log, first difference).
3. Identify the dependence orders of the model (p, d, q).
4. Estimate the parameters.
5. Diagnostics of the residuals (autocorrelation, interpretation, normality assumptions).
6. Select the best model among candidates.

Steps 1 & 2: plotting and transforming the data

Step 1 — Plot the data. Always begin with a time plot. Inspect for trends, changing variance, seasonal patterns, or obvious anomalies before any modelling.

Step 2 — Transform the data. If the variability grows with the level of the series, a variance-stabilising transformation is needed. The **Box–Cox** family of power transformations is a standard choice:

$$x_t^{(\lambda)} = \begin{cases} \dfrac{x_t^{\lambda} - 1}{\lambda}, & \lambda \neq 0, \\ \log x_t, & \lambda = 0. \end{cases}$$

Step 3: identifying the orders (p, d, q)

After transforming the data, identify preliminary values of:

- d — the order of differencing,
- p — the autoregressive order,
- q — the moving-average order.

We address d first, then (p, q) jointly from the differenced series.

Step 3.1: identifying the differencing order d

Remark 9.7: Detecting the need for differencing

A **slow decay** in the sample ACF of x_t suggests the series is non-stationary and differencing may be needed. One can also check for non-stationarity of time series / presence of unit roots by **hypothesis testing**: Augmented Dickey–Fuller (ADF) test, KPSS test, Phillips–Perron (PP) test.

Procedure:

1. If differencing is needed, compute ∇x_t and inspect its time plot and ACF. Perform testing to check for unit roots / non-stationarity.
2. If further differencing is needed, compute $\nabla^2 x_t$ and repeat.
3. Stop as soon as the differenced series looks stationary.

Remark 9.8: Warning: over-differencing

Do **not** difference more than necessary. For example, $x_t = w_t$ is white noise, but $\nabla x_t = w_t - w_{t-1}$ is MA(1) — differencing has **introduced dependence** where none existed.

Step 3.2: identifying p and q

With d fixed, examine the sample ACF and PACF of $\nabla^d x_t$ and apply the standard identification rules:

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

Remark 9.9: Important remarks

- The ACF and PACF *cannot both* cut off simultaneously.
- With real data it may not always be clear whether a plot is "tailing off" or "cutting off" — do not over-interpret.
- Two seemingly different models can be very similar in fit; retain a few candidate (p, d, q) triples for the estimation step.

Step 4: estimate model parameters

For all candidate models (different triplets of (p, d, q)), we estimate model parameters using either the **Method of Moments (MOM)** or the **Maximum Likelihood (MLE)** approaches.

Step 5: model diagnostics

This investigation focuses on the analysis of the residuals. The diagnostic results provide base for model selection.

Definition 9.2: Standardized Innovations

The standardized innovations or residuals can be computed by

$$\hat{e}_t = \frac{x_t - \hat{x}_t^{t-1}}{\sqrt{\hat{P}_t^{t-1}}},$$

where $\hat{x}_t^{t-1}$ is the **one-step-ahead prediction** of x_t based on the fitted model and $\hat{P}_t^{t-1}$ is the estimated one-step-ahead error variance.

If the model is correctly specified / fits well, the standardized residuals $\{\hat{e}_t\}$ should behave as an **iid sequence** with mean 0 and variance 1. A normal probability plot or a Q–Q plot can help in identifying departures from normality.

Key diagnostic plots:

- **Time plot** of $\{\hat{e}_t\}$: check for outliers, changing variance, or remaining structure.
- **Normal Q–Q plot**: assess departures from normality.
- **Sample ACF** of $\{\hat{e}_t\}$: plot $\hat{\rho}_e(h)$ vs. h with bounds $\pm 2/\sqrt{n}$; significant spikes suggest model misspecification and residual dependence structure in residuals.
- The **Ljung–Box test** can be used to identify whether $\hat{\rho}_e(h)$ is small in magnitude for a given lag h. This test checks if the correlations of the residuals are 0 or not.

Step 5: the Ljung–Box–Pierce test

Definition 9.3: Ljung–Box–Pierce Test

Hypotheses: H_0: residuals are uncorrelated up to lag H vs. H_1: at least one $\rho_e(h) \neq 0$ for $1 \leq h \leq H$.

Test statistic (Ljung–Box):

$$Q(H) = n(n+2) \sum_{h=1}^{H} \frac{\hat{\rho}_e(h)^2}{n-h}.$$

Under H_0, for large n:

$$Q(H) \xrightarrow{d} \chi^2(H-p-q),$$

where p and q are the AR and MA orders of the fitted model.

Decision rule: Reject H_0 at level α if $Q(H) > \chi^2_\alpha(H-p-q)$, or equivalently if the p-value $< \alpha$. Large p-values across all lags indicate a well-fitting model.

Remark 9.10: Remarks

- The degrees of freedom correction $H-p-q$ accounts for the parameters already estimated from the data.
- The test should be applied at **several values of** H (e.g. $H = 10, 15, 20$) to check residual dependence at multiple horizons.
- Failing to reject H_0 does *not* prove the model is correct — it only indicates no detectable autocorrelation in the residuals up to lag H.

Step 6: model selection

There may be multiple candidate models after Step 3.

1. **Discard** any models with major diagnostic violations (significant residual autocorrelation, non-normality, etc.).
2. Among the remaining models, use an **information-theoretic criterion** to select the best fit — lower values are preferred.

Definition 9.4: Information Criteria

$$\text{AIC} = -2\,\ell(\hat{\theta}) + 2k, \qquad \text{BIC} = -2\,\ell(\hat{\theta}) + k \log n,$$

where $k = p + q + 1$ is the number of estimated parameters and $\ell(\hat{\theta})$ is the maximised log-likelihood. **Prefer the model with the smallest value.**

Remark 9.11: Remarks

BIC penalises model complexity more heavily than AIC and tends to favour **simpler** (more parsimonious) models, especially for large n.

Example: analysis of US GNP data

Example 9.1: US GNP Data

We consider the analysis of quarterly U.S. GNP from 1947(1) to 2002(3), $n = 223$ observations. The data are real U.S. gross national product in billions of chained 1996 dollars and have been seasonally adjusted. The data were obtained from the Federal Reserve Bank of St. Louis (`http://research.stlouisfed.org/`).

Step 1 & 2.

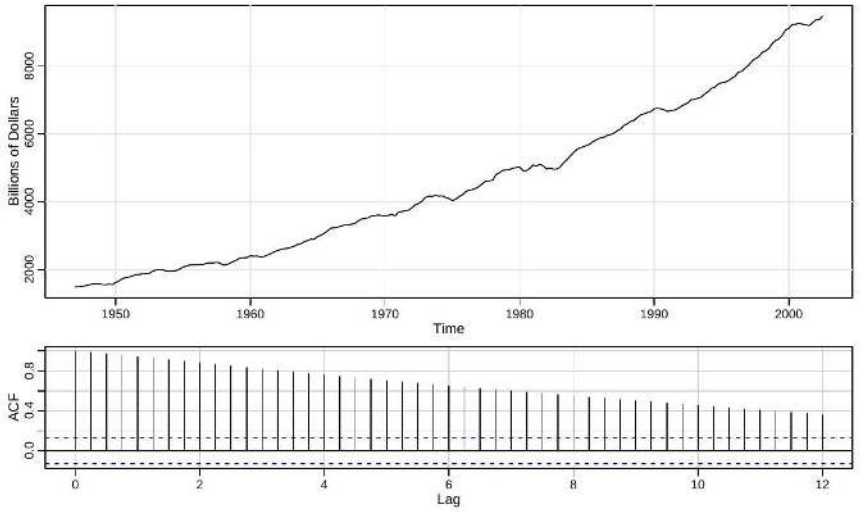


Figure 9.1: Quarterly U.S. GNP from 1947(1) to 2002(3) (top) and its sample ACF (bottom).

There is a clear upward trend in the data. The slowly decaying ACF also indicates non-stationarity.

Step 3. When reports of GNP and similar economic indicators are given, it is often in growth rate (percent change) rather than in actual (or adjusted) values that is of interest. The growth rate $x_t = \nabla \log(y_t) = \log y_t - \log y_{t-1}$ is plotted.

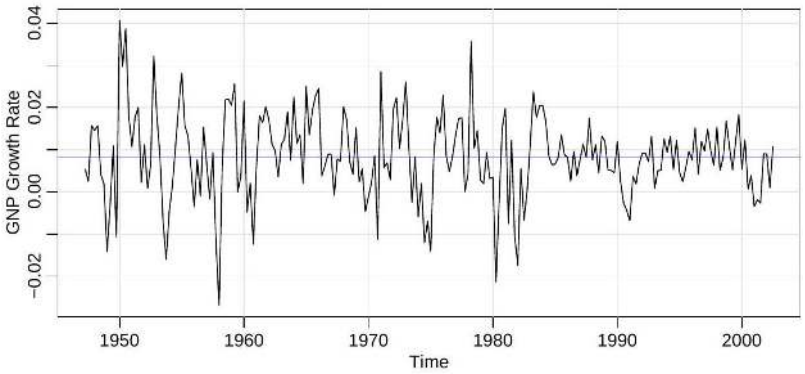


Figure 9.2: U.S. GNP quarterly growth rate.

The growth-rate series appears stable / stationary, so we set $d = 1$ on $\log y_t$.

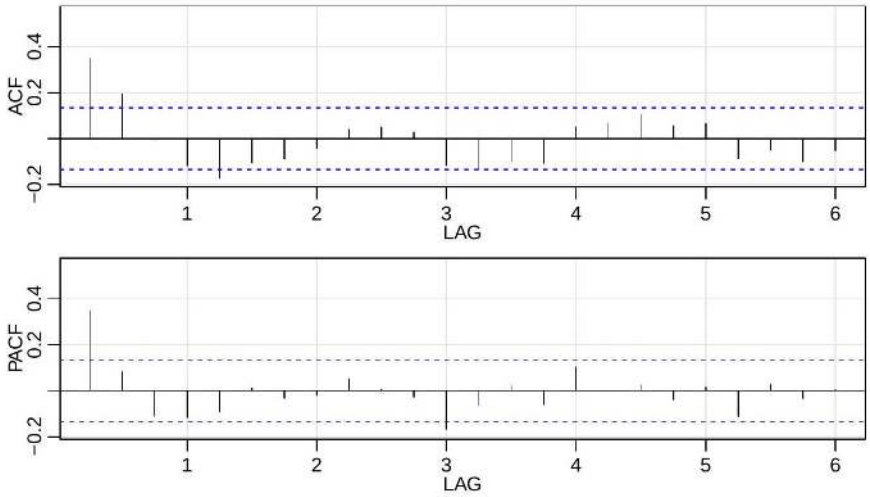


Figure 9.3: Sample ACF and PACF of the GNP quarterly growth rate.

Inspecting the sample ACF and PACF, we might feel that two models are suitable for the data:

1. The ACF is cutting off at lag 2 and the PACF is tailing off. This would suggest the GNP growth rate follows an MA(2) process, or log GNP follows an ARIMA(0,1,2) model.
2. The ACF is tailing off and the PACF is cutting off at lag 1. This suggests an AR(1) model for the growth rate, or ARIMA(1,1,0) for log GNP.

Rather than focus on one model, we will fit both models.

Step 4. Using MLE to fit the models:

$$\text{MA(2):} \quad \hat{x}_t = .008_{(.001)} + .303_{(.065)}\hat{w}_{t-1} + .204_{(.064)}\hat{w}_{t-2} + \hat{w}_t, \qquad \hat{\sigma}_w = .0094.$$
$$\text{AR(1):} \quad \hat{x}_t = .008_{(.001)}(1 - .347) + .347_{(.063)}\hat{x}_{t-1} + \hat{w}_t, \qquad \hat{\sigma}_w = .0095.$$

The values in parentheses are the corresponding estimated standard errors. All of the regression coefficients are significant, including the constant.

Step 5. We take the MA(2) for example.

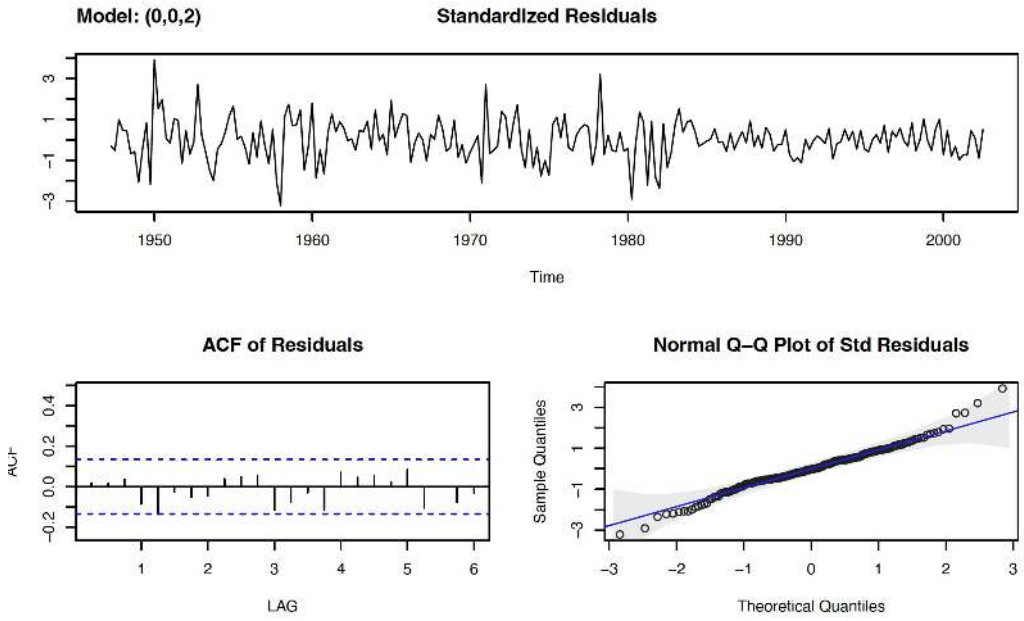


Figure 9.4: Residual plots of the GNP quarterly growth rate for the MA(2) model: standardized residuals (top), ACF of residuals (bottom left), and normal Q–Q plot of standardized residuals (bottom right).

Performing the Ljung–Box test:

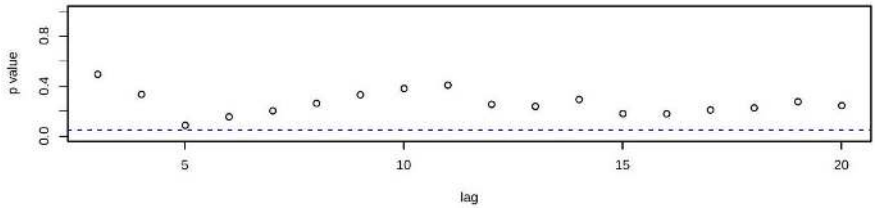


Figure 9.5: p-values associated with the Ljung–Box Q-statistic at lags $H = 3$ through $H = 20$ (with corresponding degrees of freedom $H - 2$).

None of the p-values are below the threshold, so we cannot reject the null hypothesis.

9.3 Forecasting

General setup: second-order prediction

The BLP framework extends beyond time series to any collection of random variables with finite second moments.

Setup: Suppose Y and $W_1, \ldots, W_n$ have finite second moments. The BLP of Y given $\mathbf{W}_n = (W_n, \ldots, W_1)^\top$ is

$$P_{\mathbf{W}_n} Y = \mu_Y + \boldsymbol{\alpha}_n^\top (\mathbf{W}_n - \boldsymbol{\mu}),$$

where $\boldsymbol{\alpha}_n$ solves $\mathbf{T}\boldsymbol{\alpha}_n = \boldsymbol{\gamma}$ with $T_{ij} = \mathrm{Cov}(W_i, W_j)$ and $\gamma_i = \mathrm{Cov}(Y, W_i)$. The three key properties of the BLP:

1. $\mathbb{E}[Y - P_{\mathbf{W}_n} Y] = 0$,

2. $\mathbb{E}[(Y - P_{\mathbf{W}_n} Y)\, W_i] = 0$ for all i,
3. Mean-square error: $\mathbb{E}[Y - P_{\mathbf{W}_n} Y]^2 = \mathrm{Var}(Y) - \boldsymbol{\alpha}_n^\top \boldsymbol{\gamma}$.

Applications: forecasting, missing value imputation, back-forecasting, and defining the PACF.

ARIMA forecasting

Consider a time series x_t that is ARIMA(p, d, q). It should be clear that, since $y_t = \nabla^d x_t$ is ARMA, we can use the previously introduced methods to obtain forecasts of y_t, which in turn lead to forecasts for x_t.

For example, if $d = 1$, given forecasts y_{n+m}^n for $m = 1, 2, \ldots$, we have $y_{n+m}^n = x_{n+m}^n - x_{n+m-1}^n$, so that

$$x_{n+m}^n = y_{n+m}^n + x_{n+m-1}^n$$

with initial condition $x_{n+1}^n = y_{n+1}^n + x_n$.

Example: IMA(1,1) and EWMA model

The ARIMA(0,1,1), or IMA(1,1) model is of interest because many economic time series can be successfully modeled this way. In addition, the model leads to a frequently used forecasting method called **exponentially weighted moving averages (EWMA)**. We will write the model as

$$x_t = x_{t-1} + w_t - \lambda w_{t-1},$$

with $|\lambda| < 1$, for $t = 1, 2, \ldots$, and $x_0 = 0$. We could also include a drift term in the formula.

If we write

$$y_t = w_t - \lambda w_{t-1},$$

we may write the IMA(1,1) as $x_t = x_{t-1} + y_t$. Because $|\lambda| < 1$, y_t has an invertible representation, $y_t = \sum_{j=1}^{\infty} \lambda^j y_{t-j} + w_t$, and substituting $y_t = x_t - x_{t-1}$, we may write

$$x_t = \sum_{j=1}^{\infty} (1 - \lambda)\lambda^{j-1} x_{t-j} + w_t,$$

as an approximation for large t (put $x_t = 0$ for $t \leq 0$), which is the exponentially weighted moving averages (EWMA).

Chapter 10

Lecture 10

10.1 Estimation of ARMA Models: Setup

Three approaches

We study three approaches to estimating the parameters of ARMA models:

1. **Least Squares (LS)**;
2. **Method of Moments (MOM)**;
3. **Maximum Likelihood Estimation**: Conditional MLE (CMLE) and Exact MLE (EMLE).

Consider the causal stationary AR(p) model:

$$x_t = c + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \varepsilon_t, \qquad \varepsilon_t \overset{\text{iid}}{\sim} \text{WN}(0, \sigma^2).$$

- Model parameters: $\boldsymbol{\eta} = (c, \phi_1, \ldots, \phi_p)$.
- Given observations $(x_1, \ldots, x_n)$, the goal is to estimate $\boldsymbol{\eta}$ and σ^2.

Remark 10.1: AR models: estimators coincide

For AR models, CMLE and the OLS estimator coincide. CMLE and EMLE share the same asymptotic distribution.

10.2 Approach I: Least Squares Estimation

Least squares for AR(p)

Define the sum of squared residuals:

$$\Psi(c, \phi_1, \ldots, \phi_p) = \sum_{t=p+1}^{n} \left(x_t - c - \phi_1 x_{t-1} - \cdots - \phi_p x_{t-p}\right)^2.$$

Definition 10.1: Least Squares Estimator

The least squares estimator is

$$\hat{\boldsymbol{\eta}}_{\text{LSE}} = \arg\min_{\boldsymbol{\eta}} \ \Psi(c, \phi_1, \ldots, \phi_p).$$

The LS normal equations are:

$$\sum_{t=p+1}^{n} \big(x_t - c - \phi_1 x_{t-1} - \cdots - \phi_p x_{t-p}\big) = 0,$$

$$\sum_{t=p+1}^{n} \big(x_t - c - \phi_1 x_{t-1} - \cdots - \phi_p x_{t-p}\big)\, x_{t-i} = 0, \qquad i = 1, \ldots, p.$$

Noise variance estimate and remarks

Solving the LS normal equations yields $\hat{\boldsymbol{\eta}}_{\mathrm{LSE}}$, and the noise variance is estimated by:

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{t=p+1}^{n} \big(x_t - \hat{c} - \hat{\phi}_1 x_{t-1} - \cdots - \hat{\phi}_p x_{t-p}\big)^2.$$

Remarks:

- The LS approach does not make any distributional assumptions about the time series.
- LS does not inherently enforce constraints on the parameters to enforce stationarity / invertibility / causality.
- The LS approach does not rely on stationarity properties of the time series to begin with.

10.3 Approach II: Method of Moments (Yule–Walker)

The MOM procedure

The **Method of Moments (MOM)** estimates model parameters by matching theoretical moments to their sample counterparts.

1. **Express moments in terms of parameters.** Write the theoretical autocovariances $\gamma(0), \gamma(1), \ldots, \gamma(h)$ as explicit functions of the model parameters $(\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2)$.
2. **Invert the moment equations.** Solve those equations to re-express the parameters $(\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2)$ as functions of $\gamma(0), \gamma(1), \ldots, \gamma(h)$.
3. **Plug in sample estimates.** Replace each $\gamma(h)$ with its sample estimate $\hat{\gamma}(h)$ (equivalently $\hat{\rho}(h)$) computed from the data. The resulting values are the MOM estimates of the model parameters.

Essentially, treat the parameters as unknowns and solve for a system of equations (you need as many sample ACVFs as the number of parameters). For AR(p) processes this procedure yields the **Yule–Walker estimators**.

Yule–Walker equations for AR(p)

For a zero-mean AR(p) process, multiplying $x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \varepsilon_t$ by x_{t-h} and taking expectations gives, for $h > 0$:

$$\gamma_h = \phi_1 \gamma_{h-1} + \cdots + \phi_p \gamma_{h-p}.$$

In matrix form (**Yule–Walker equations**):

$$\boldsymbol{\Gamma}_p\, \boldsymbol{\phi} = \boldsymbol{\gamma}_p, \qquad \sigma^2 = \gamma_0 - \boldsymbol{\phi}^\top \boldsymbol{\gamma}_p,$$

where $\boldsymbol{\Gamma}_p = \{\gamma_{k-j}\}_{j,k=1}^{p}$ and $\boldsymbol{\gamma}_p = (\gamma_1, \ldots, \gamma_p)^\top$.

Definition 10.2: Yule–Walker Estimators

The Yule–Walker estimators replace γ_h with sample autocovariances $\hat{\gamma}_h$:

$$\hat{\boldsymbol{\phi}}_{\mathrm{YW}} = \hat{\boldsymbol{\Gamma}}_p^{-1}\hat{\boldsymbol{\gamma}}_p, \qquad \hat{\sigma}^2_{\mathrm{YW}} = \hat{\gamma}_0 - \hat{\boldsymbol{\gamma}}_p^{\top}\hat{\boldsymbol{\Gamma}}_p^{-1}\hat{\boldsymbol{\gamma}}_p.$$

For AR(1): $h = 1 \Rightarrow \hat{\gamma}_1 = \hat{\phi}_{\mathrm{YW}}\,\hat{\gamma}_0$, so $\hat{\phi}_{\mathrm{YW}} = \hat{\gamma}_1/\hat{\gamma}_0$.

Large-sample results for Yule–Walker estimators

Theorem 10.1: Asymptotics of Yule–Walker Estimators

The asymptotic $n \to \infty$ behavior of the Yule–Walker estimators in the case of causal AR(p) processes is as follows:

$$\sqrt{n}\,(\hat{\boldsymbol{\phi}}_{\mathrm{YW}} - \boldsymbol{\phi}) \xrightarrow{d} N\big(\mathbf{0},\ \sigma^2\,\boldsymbol{\Gamma}_p^{-1}\big), \qquad \hat{\sigma}^2_{\mathrm{YW}} \xrightarrow{P} \sigma^2.$$

Theorem 10.2: Corollary — Large-Sample Distribution of the PACF

For a causal AR(p) process, for $h > p$:

$$\sqrt{n}\,\hat{\phi}_{hh} \xrightarrow{d} N(0,1).$$

This justifies the bounds $\pm 2/\sqrt{n}$ for significance in the sample PACF.

MOM estimation: AR(1) example

Example 10.1: MOM for AR(1)

Consider a zero-mean AR(1): $x_t = \phi x_{t-1} + \varepsilon_t$, $\varepsilon_t \sim \mathrm{WN}(0, \sigma^2)$.

Step 1 — Express ACVF in terms of parameters. Multiply both sides by x_{t-h} and take expectations:

$$\gamma(h) = \phi\,\gamma(h-1), \quad h \geq 1, \qquad \gamma(0) = \frac{\sigma^2}{1-\phi^2}.$$

Step 2 — Invert to express parameters in terms of ACVF.

$$\phi = \frac{\gamma(1)}{\gamma(0)} = \rho(1), \qquad \sigma^2 = \gamma(0)\,(1-\phi^2).$$

Step 3 — Substitute sample estimates.

$$\hat{\phi}_{\mathrm{YW}} = \frac{\hat{\gamma}(1)}{\hat{\gamma}(0)} = \hat{\rho}(1), \qquad \hat{\sigma}^2_{\mathrm{YW}} = \hat{\gamma}(0)\big(1-\hat{\rho}(1)^2\big).$$

MOM estimation: MA(1) example

Example 10.2: MOM for MA(1)

Consider a zero-mean MA(1): $x_t = \varepsilon_t + \theta\varepsilon_{t-1}$, $\varepsilon_t \sim \mathrm{WN}(0, \sigma^2)$.
Step 1 — Express ACVF in terms of parameters.

$$\gamma(0) = \sigma^2(1+\theta^2), \qquad \gamma(1) = \sigma^2\theta, \qquad \gamma(h) = 0 \text{ for } h \geq 2.$$

Step 2 — Invert to express parameters in terms of ACVF. Dividing gives $\rho(1) = \theta/(1+\theta^2)$. Solving for θ:

$$\rho(1)\,\theta^2 - \theta + \rho(1) = 0 \implies \theta = \frac{1 \pm \sqrt{1-4\rho(1)^2}}{2\rho(1)}.$$

Take the root with $|\theta| \leq 1$ (invertibility). Then $\sigma^2 = \gamma(0)/(1+\theta^2)$.
Step 3 — Substitute sample estimates.

$$\hat{\theta}_{\mathrm{MOM}} = \frac{1 \pm \sqrt{1-4\hat{\rho}(1)^2}}{2\hat{\rho}(1)}, \qquad \hat{\sigma}^2_{\mathrm{MOM}} = \frac{\hat{\gamma}(0)}{1+\hat{\theta}^2}.$$

Take the root with $|\hat{\theta}_{\mathrm{MOM}}| \leq 1$ (invertibility) and the corresponding $\hat{\sigma}^2_{\mathrm{MOM}}$.

Remark 10.2: Existence of a real solution

A real solution requires $|\hat{\rho}(1)| \leq 1/2$. If violated, the data are unlikely to have come from an MA(1).

10.4 Approach III: Maximum Likelihood Estimation

General approach

Under the assumption $\varepsilon_t \overset{\text{iid}}{\sim} N(0, \sigma^2)$, MLE finds the parameter values that make the observed data most probable.

1. **Write down the joint likelihood:** use the **conditional structure** of the model to factorise $L(\boldsymbol{\eta}) = \prod_t f(x_t \mid \text{past})$. For Gaussian innovations each factor is a normal density.
2. **Choose exact or conditional likelihood:** the **exact** likelihood includes the marginal density of the first p observations; the **conditional** likelihood treats them as fixed, simplifying computation.
3. **Maximise the log-likelihood:** closed-form solutions exist only in special cases (e.g. AR(1) CMLE = OLS). In general, use iterative numerical methods (Newton–Raphson, Levenberg–Marquardt).
4. **Estimate σ^2:** given $\hat{\boldsymbol{\eta}}$, estimate σ^2 by the average squared residual from the fitted model.

Exact MLE for AR(1)

Suppose $\{x_t\}$ is a Gaussian AR(1): $x_t = c + \phi x_{t-1} + \varepsilon_t$, $|\phi| < 1$, $\varepsilon_t \sim N(0, \sigma^2)$.

The **Markov structure** gives, for $t \geq 2$:

$$x_t \mid x_{t-1}, \ldots, x_1 \overset{d}{=} x_t \mid x_{t-1} \sim N(c + \phi x_{t-1},\ \sigma^2).$$

The stationary marginal is $x_1 \sim N\left(\frac{c}{1-\phi},\ \frac{\sigma^2}{1-\phi^2}\right)$.

Joint pdf of $(x_1, \ldots, x_n)$:

$$f_{x_1,\ldots,x_n} = f_{x_1} \prod_{t=2}^{n} f_{x_t|x_{t-1}}.$$

Likelihood: $L(\boldsymbol{\eta}) = f_{x_1} \prod_{t=2}^{n} f_{x_t|x_{t-1}}$.

Exact MLE for AR(1): log-likelihood

The log-likelihood is:

$$\ell(\boldsymbol{\eta}) = \underbrace{\left(-\tfrac{1}{2}\log 2\pi - \tfrac{1}{2}\log\tfrac{\sigma^2}{1-\phi^2} - \tfrac{1-\phi^2}{2\sigma^2}\left(x_1 - \tfrac{c}{1-\phi}\right)^2\right)}_{\log f_{x_1}}$$
$$+ \underbrace{\left(-(n-1)\log\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}\sum_{t=2}^{n}(x_t - c - \phi x_{t-1})^2\right)}_{\sum_{t=2}^{n}\log f_{x_t|x_{t-1}}},$$

$$\hat{\boldsymbol{\eta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\eta}}\ \ell(\boldsymbol{\eta}).$$

Remark: use iterative optimisation (Newton–Raphson, Levenberg–Marquardt, downhill simplex) to obtain the MLEs.

Conditional MLE for AR(1)

Definition 10.3: Conditional MLE — Difference from Exact MLE

Treat x_1 as **fixed/deterministic** and condition on it. For AR(p), treat $x_1, \ldots, x_p$ as fixed. The conditional likelihood is:

$$L_c(\boldsymbol{\eta}) = \prod_{t=2}^{n} f_{x_t|x_{t-1}}.$$

The conditional log-likelihood is:

$$\ell_c(\boldsymbol{\eta}) = -\frac{n-1}{2}\log 2\pi - \frac{n-1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=2}^{n}(x_t - c - \phi x_{t-1})^2.$$

Setting $\partial\ell_c/\partial c = 0$ and $\partial\ell_c/\partial\phi = 0$ gives:

$$\begin{pmatrix}\hat{c}_{\text{CMLE}} \\ \hat{\phi}_{\text{CMLE}}\end{pmatrix} = \begin{pmatrix} n-1 & \sum x_{t-1} \\ \sum x_{t-1} & \sum x_{t-1}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum x_t \\ \sum x_t x_{t-1} \end{pmatrix},$$

which is the **same as the OLS/LSE**.

The CMLE variance estimator is:

$$\hat{\sigma}^2_{\text{CMLE}} = \frac{1}{n-1}\sum_{t=2}^{n}\left(x_t - \hat{c}_{\text{CMLE}} - \hat{\phi}_{\text{CMLE}}\, x_{t-1}\right)^2.$$

Remark 10.3: CMLE vs. EMLE

CMLE and EMLE have the **same asymptotic distribution**.

Exact MLE for AR(p)

For a stationary Gaussian AR(p) process $\{x_t\}$:

$$\mathbf{x}_p = (x_1, \ldots, x_p)^\top \sim N_p(\boldsymbol{\mu}_p,\ \sigma^2 V_p),$$

where

$$\boldsymbol{\mu}_p = \frac{c}{1-\phi_1-\cdots-\phi_p}\,\mathbf{1}_p, \qquad \sigma^2 V_p = \operatorname{Cov}(\mathbf{x}_p) = \boldsymbol{\Gamma}_p = \{\gamma_{k-j}\}_{j,k=1}^{p}.$$

For $t > p$, the Markov property gives:

$$x_t \mid x_{t-1}, \ldots, x_1 \overset{d}{=} x_t \mid x_{t-1}, \ldots, x_{t-p} \sim N\big(c + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p},\ \sigma^2\big).$$

The joint pdf factors as:

$$f_{x_n,\ldots,x_1} = f_{\mathbf{x}_p}\prod_{t=p+1}^{n} f_{x_t|x_{t-1},\ldots,x_{t-p}}.$$

Exact MLE for AR(p): log-likelihood

The log-likelihood is:

$$\begin{aligned}\ell(\boldsymbol{\eta}) = &\left(-\frac{p}{2}\log 2\pi - \frac{p}{2}\log\sigma^2 + \frac{1}{2}\log|V_p^{-1}| - \frac{1}{2\sigma^2}(\mathbf{x}_p-\boldsymbol{\mu}_p)^\top V_p^{-1}(\mathbf{x}_p-\boldsymbol{\mu}_p)\right)\\ &+\left(-\frac{n-p}{2}\log 2\pi - \frac{n-p}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=p+1}^{n}\left(x_t - c - \phi_1 x_{t-1} - \cdots - \phi_p x_{t-p}\right)^2\right).\end{aligned}$$

The entries of $V_p^{-1} = ((v_p^{ij}))$ are given by **Galbraith's formula**:

$$v_p^{ij} = \sum_{k=0}^{i-1}\phi_k\phi_{k+j-i} - \sum_{k=p+1-i}^{p+i-j}\phi_k\phi_{k+j-i}, \qquad \phi_0 = -1.$$

- To write down the exact likelihood for AR(p), we need V_p^{-1} and $|V_p|$ explicitly, where $\sigma^2 V_p = \boldsymbol{\Gamma}_p$ is the autocovariance matrix of the initial vector $(x_1, \ldots, x_p)^\top$. In general, inverting an arbitrary $p \times p$ Toeplitz matrix requires numerical methods.
- Galbraith's formula provides a **closed-form expression** for the (i,j) entry of V_p^{-1} directly in terms of the AR coefficients $\phi_1, \ldots, \phi_p$, bypassing numerical inversion. This makes the quadratic form $(\mathbf{x}_p-\boldsymbol{\mu}_p)^\top V_p^{-1}(\mathbf{x}_p-\boldsymbol{\mu}_p)$ and the determinant $|V_p|$ in the exact log-likelihood **analytically tractable**.

Conditional MLE for AR(p)

Definition 10.4: Conditional Likelihood for AR(p)

Conditioning on the first p observations:

$$L_c(\boldsymbol{\eta}) = \prod_{t=p+1}^{n} f_{x_t|x_{t-1},\ldots,x_{t-p}} = f_{x_n,\ldots,x_{p+1}|x_p,\ldots,x_1}.$$

Conditional log-likelihood:

$$\ell_c(\boldsymbol{\eta}) = -\frac{n-p}{2}\log 2\pi - \frac{n-p}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=p+1}^{n}\big(x_t - c - \phi_1 x_{t-1} - \cdots - \phi_p x_{t-p}\big)^2.$$

Maximising gives $\hat{\boldsymbol{\eta}}_{\text{CMLE}} = \hat{\boldsymbol{\eta}}_{\text{LSE}}$, with

$$\hat{\sigma}^2_{\text{CMLE}} = \frac{1}{n-p}\sum_{t=p+1}^{n}\big(x_t - \hat{c} - \hat{\phi}_1 x_{t-1} - \cdots - \hat{\phi}_p x_{t-p}\big)^2.$$

Conditional MLE for MA(1)

Consider a Gaussian MA(1): $x_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1}$, $\varepsilon_t \overset{\text{iid}}{\sim} N(0,\sigma^2)$, $\boldsymbol{\eta} = (\mu, \theta, \sigma^2)$.

Conditional MLE: assume $\varepsilon_0 = 0$ (its expected value) and derive the likelihood conditional on $\varepsilon_0 = 0$.

The innovations can be recovered recursively:

$$\begin{aligned}
\varepsilon_1 &= x_1 - \mu - \theta\varepsilon_0 = x_1 - \mu,\\
\varepsilon_2 &= x_2 - \mu - \theta\varepsilon_1 = (x_2 - \mu) - \theta(x_1 - \mu),\\
\varepsilon_3 &= x_3 - \mu - \theta\varepsilon_2 = (x_3 - \mu) - \theta(x_2 - \mu) + \theta^2(x_1 - \mu),\\
&\vdots
\end{aligned}$$

Thus $(\varepsilon_1, \ldots, \varepsilon_n)$ can be expressed in terms of $(x_1, \ldots, x_n)$, μ, and θ.

For $t \geq 1$: $x_t \mid \varepsilon_{t-1} \sim N(\mu + \theta\varepsilon_{t-1},\ \sigma^2)$. The conditional log-likelihood is:

$$\ell_c(\boldsymbol{\eta}) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{n}\varepsilon_t^2, \qquad \hat{\boldsymbol{\eta}}_{\text{CMLE}} = \arg\max_{\boldsymbol{\eta}}\ \ell_c(\boldsymbol{\eta}).$$

Iterative optimisation is required (no closed form in general).

Exact MLE for MA(1): covariance structure

For the **exact** likelihood, treat $\mathbf{x}_n = (x_1, \ldots, x_n)$ as a single multivariate Gaussian observation with **unconditional distribution**:

$$\mathbf{x}_n \sim N_n(\mu\,\mathbf{1}_n,\ \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} = \text{Cov}(\mathbf{x}_n)$ follows from $\gamma(0) = \sigma^2(1+\theta^2)$, $\gamma(1) = \sigma^2\theta$, $\gamma(h) = 0$ for $|h| \geq 2$:

$$\boldsymbol{\Sigma} = \sigma^2\begin{pmatrix}
1+\theta^2 & \theta & 0 & \cdots & 0 & 0\\
\theta & 1+\theta^2 & \theta & \cdots & 0 & 0\\
0 & \theta & 1+\theta^2 & \cdots & 0 & 0\\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\
0 & 0 & 0 & \cdots & 1+\theta^2 & \theta\\
0 & 0 & 0 & \cdots & \theta & 1+\theta^2
\end{pmatrix}.$$

This is a symmetric **tridiagonal Toeplitz** matrix — all diagonal entries equal $\sigma^2(1+\theta^2)$ and all super/sub-diagonal entries equal $\sigma^2\theta$.

Exact MLE for MA(1): Cholesky factorisation

To evaluate

$$L(\boldsymbol{\eta}) = (2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\Big(-\tfrac{1}{2}(\mathbf{x}_n-\mu\mathbf{1}_n)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\mu\mathbf{1}_n)\Big)$$

we need $\boldsymbol{\Sigma}^{-1}$ and $|\boldsymbol{\Sigma}|$. Use the Cholesky factorisation $\boldsymbol{\Sigma} = ADA^\top$ where A is unit lower-triangular and $D = \operatorname{diag}(d_{11},\dots,d_{nn})$ with

$$d_{ii} = \sigma^2\,\frac{1+\theta^2+\cdots+\theta^{2i}}{1+\theta^2+\cdots+\theta^{2(i-1)}}, \qquad i = 1,\dots,n.$$

Setting $\mathbf{y} = A^{-1}(\mathbf{x}_n-\mu\mathbf{1}_n)$, the log-likelihood becomes:

$$\ell(\boldsymbol{\eta}) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{n}\log d_{tt} - \frac{1}{2}\sum_{t=1}^{n}\frac{y_t^2}{d_{tt}}.$$

Initialisation: compute $\hat\rho(1)$ from data, solve $\hat\rho(1) = \hat\theta/(1+\hat\theta^2)$ for $\hat\theta$ (quadratic in $\hat\theta$); set $\mu = \bar x$.

Conditional MLE for MA(q)

For a Gaussian MA(q) process: $x_t = c+\varepsilon_t+\theta_1\varepsilon_{t-1}+\cdots+\theta_q\varepsilon_{t-q}$, $\boldsymbol{\eta} = (c,\theta_1,\dots,\theta_q,\sigma^2)$.

Initialise at expected values: $\varepsilon_0 = \varepsilon_{-1} = \cdots = \varepsilon_{-(q-1)} = 0$.

Recursively compute innovations:

$$\varepsilon_t = x_t - c - \theta_1\varepsilon_{t-1} - \cdots - \theta_q\varepsilon_{t-q}.$$

The conditional log-likelihood is:

$$\ell_c(\boldsymbol{\eta}) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{n}\varepsilon_t^2.$$

Conditional MLE for ARMA(p,q)

For a Gaussian ARMA(p,q): $x_t = c+\sum_{i=1}^{p}\phi_i x_{t-i}+\varepsilon_t+\sum_{i=1}^{q}\theta_i\varepsilon_{t-i}$, $\boldsymbol{\eta} = (c,\phi_1,\dots,\phi_p,\theta_1,\dots,\theta_q,\sigma^2)$.

Initialise at expected values for $t\le 0$:

$$\boldsymbol{\varepsilon}_0 = \mathbf{0}_q, \qquad \mathbf{x}_0 = \frac{c}{1-\phi_1-\cdots-\phi_p}\,\mathbf{1}_p = \mathbb{E}(\mathbf{x}_0).$$

Innovations are computed recursively, and for $t\ge 2$:

$$x_t \mid x_{t-1},\dots,x_1,\mathbf{x}_0,\boldsymbol{\varepsilon}_0 \sim N\left(c+\sum_{i=1}^{p}\phi_i x_{t-i}+\sum_{i=1}^{q}\theta_i\varepsilon_{t-i},\ \sigma^2\right).$$

The conditional log-likelihood can again be written down in a similar fashion using the density of a normal distribution.

Asymptotic distribution of MLE for ARMA(p,q)

Let $\{x_t\}$ be a causal and invertible ARMA(p,q): $\phi(B)x_t = \theta(B)\varepsilon_t$, with $\boldsymbol{\beta} = (\phi_1,\dots,\phi_p,\theta_1,\dots,\theta_q)^\top$.

Theorem 10.3: Asymptotic Distribution of the MLE

$$\sqrt{n}\,(\hat{\boldsymbol{\beta}}_{\text{MLE}} - \boldsymbol{\beta}) \xrightarrow{d} N_{p+q}\big(\mathbf{0},\ \text{Var}(\boldsymbol{\beta})\big),$$

where

$$\text{Var}(\boldsymbol{\beta}) = \sigma^2 \begin{bmatrix} \mathbb{E}(\boldsymbol{U}_t \boldsymbol{U}_t^\top) & \mathbb{E}(\boldsymbol{U}_t \boldsymbol{V}_t^\top) \\ \mathbb{E}(\boldsymbol{V}_t \boldsymbol{U}_t^\top) & \mathbb{E}(\boldsymbol{V}_t \boldsymbol{V}_t^\top) \end{bmatrix}^{-1},$$

$$\boldsymbol{U}_t = (U_t, U_{t-1}, \ldots, U_{t-(p-1)})^\top, \qquad \boldsymbol{V}_t = (V_t, V_{t-1}, \ldots, V_{t-(q-1)})^\top,$$

and $\{U_t\}$, $\{V_t\}$ are AR processes satisfying $\phi(B)U_t = \varepsilon_t$ and $\theta(B)V_t = \varepsilon_t$.

Asymptotic distribution: special cases

AR(p) (i.e. $q = 0$), $\boldsymbol{\beta} = \boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^\top$:

$$\sqrt{n}\,(\hat{\boldsymbol{\phi}}_{\text{MLE}} - \boldsymbol{\phi}) \xrightarrow{d} N_p\big(\mathbf{0},\ \sigma^2 \big[\mathbb{E}(\boldsymbol{U}_t \boldsymbol{U}_t^\top)\big]^{-1}\big).$$

MA(q) (i.e. $p = 0$), $\boldsymbol{\beta} = \boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^\top$:

$$\sqrt{n}\,(\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}) \xrightarrow{d} N_q\big(\mathbf{0},\ \sigma^2 \big[\mathbb{E}(\boldsymbol{V}_t \boldsymbol{V}_t^\top)\big]^{-1}\big).$$

Remark 10.4: Key special cases

- AR(1): $\sqrt{n}\,(\hat{\phi}_{\text{MLE}} - \phi) \xrightarrow{d} N(0,\ 1 - \phi^2)$.
- MA(1): $\sqrt{n}\,(\hat{\theta}_{\text{MLE}} - \theta) \xrightarrow{d} N(0,\ 1 - \theta^2)$.

Summary on MOM and MLE

- The **Method of Moments** may give a **suboptimal** estimator. An improved approach is to use the estimates obtained from the Method of Moments approach as initial values for the Maximum Likelihood approach to get an optimal estimate.
- Under appropriate conditions, for causal and invertible ARMA processes, the maximum likelihood estimator, initialized by the method of moments estimator, provides **optimal** estimators of σ_w^2 and $\boldsymbol{\beta}$, in the sense that $\hat{\sigma}_w^2$ is consistent, and the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ is the best asymptotic normal distribution.

Chapter 11

Lecture 11

11.1 Cyclical Behaviour and Periodic Series

Time domain vs. frequency domain

So far we have analysed time series in the **time domain**: modelling dependence through autocovariances, the ACF, the PACF, and ARMA structure. The **frequency domain** offers a complementary view.

- Some series exhibit strong **periodic** or **cyclical** behaviour.
- The frequency domain decomposes the series into sinusoidal components and asks: *at which frequencies does most of the variation lie?*

The two approaches are equivalent — they contain the same information about a stationary process — but each can be more convenient depending on the question. (Time-domain ACF and frequency-domain spectrum are Fourier transforms of each other.)

Cycle and period

In order to define the rate at which a series oscillates, we first define the *cycle* of a series.

Definition 11.1: The Cycle and Period of a Series

A *cycle* is a single complete repetition of a periodic function (such as a sine or cosine function) over a specific time interval. The duration of the interval taken to complete one cycle is defined as the *period* of the oscillation / series.

For example:

- Yearly data, 1 cycle per 5 years ⇒ period is 5 years.
- Monthly data with an annual cycle, 1 cycle per 12 months ⇒ period is 12 months.

Cyclical behaviour

Consider the periodic process

$$x_t = A\cos(2\pi\omega t + \phi), \qquad t = 0, \pm 1, \pm 2, \ldots$$

We introduce the following concepts:

- ω is the *frequency*: cycles per unit time — it is the inverse of the period;
- A is the *amplitude*: the maximum height of the oscillation / function;
- ϕ is the *phase*: the start point of the cosine function.

Note: We can introduce random variation in this series by allowing the amplitude and phase to vary randomly (i.e. A and ϕ can be random variables).

Using the identity $\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$, we can rewrite x_t as

$$x_t = U_1\cos(2\pi\omega t) + U_2\sin(2\pi\omega t), \qquad U_1 = A\cos\phi, \quad U_2 = -A\sin\phi.$$

Note:

- The amplitude is $A = \sqrt{U_1^2 + U_2^2}$.
- The phase is $\phi = \tan^{-1}(-U_2/U_1)$.
- U_1 and U_2 are often taken to be normally distributed random variables if we want to model randomness. (This trades a fixed amplitude/phase for two random coefficients — which is much easier to work with.)

Stationarity of a cyclic series

Theorem 11.1: Stationarity of a Cyclic Series

If U_1 and U_2 are uncorrelated with mean 0 and variance σ^2, then the cyclic time series

$$x_t = U_1\cos(2\pi\omega t) + U_2\sin(2\pi\omega t)$$

is stationary with

$$\mathbb{E}(x_t) = 0, \qquad \gamma(h) = \sigma^2\cos(2\pi\omega h).$$

Interpretation: The amplitude $A^2 = U_1^2 + U_2^2$ is a random variable. Given observed values $U_1 = a$, $U_2 = b$, a natural estimate of σ^2 is

$$\hat{\sigma}^2 = a^2 + b^2,$$

the sample variance of the two observations (a, b).

Mixtures of periodic series

A general periodic series can be written as a mixture of components at **different frequencies and amplitudes**:

$$x_t = \sum_{k=1}^{q}\left[U_{k1}\cos(2\pi\omega_k t) + U_{k2}\sin(2\pi\omega_k t)\right],$$

where U_{k1}, U_{k2} for $k = 1, \ldots, q$ are uncorrelated zero-mean random variables with variance σ_k^2, and the ω_k are distinct frequencies.

Theorem 11.2: Autocovariance of a Mixture

$$\gamma(h) = \sum_{k=1}^{q}\sigma_k^2\cos(2\pi\omega_k h).$$

Each σ_k^2 measures the **contribution** of frequency ω_k to the total variance of the series. (Setting $h = 0$ gives $\gamma(0) = \mathrm{Var}(x_t) = \sum_k \sigma_k^2$, so the total variance is split across frequencies.)

Example 11.1: Mixture of Three Components

Consider three components with frequencies 6/100, 10/100, 40/100:

$$x_{t1} = 2\cos(2\pi t \cdot 6/100) + 3\sin(2\pi t \cdot 6/100), \qquad A_1^2 = 4 + 9 = 13,$$
$$x_{t2} = 4\cos(2\pi t \cdot 10/100) + 5\sin(2\pi t \cdot 10/100), \qquad A_2^2 = 16 + 25 = 41,$$
$$x_{t3} = 6\cos(2\pi t \cdot 40/100) + 7\sin(2\pi t \cdot 40/100), \qquad A_3^2 = 36 + 49 = 85.$$

Combine these components to form a time series $x_t = x_{t1} + x_{t2} + x_{t3}$ and draw 100 samples from it for $t = 1, \ldots, 100$.

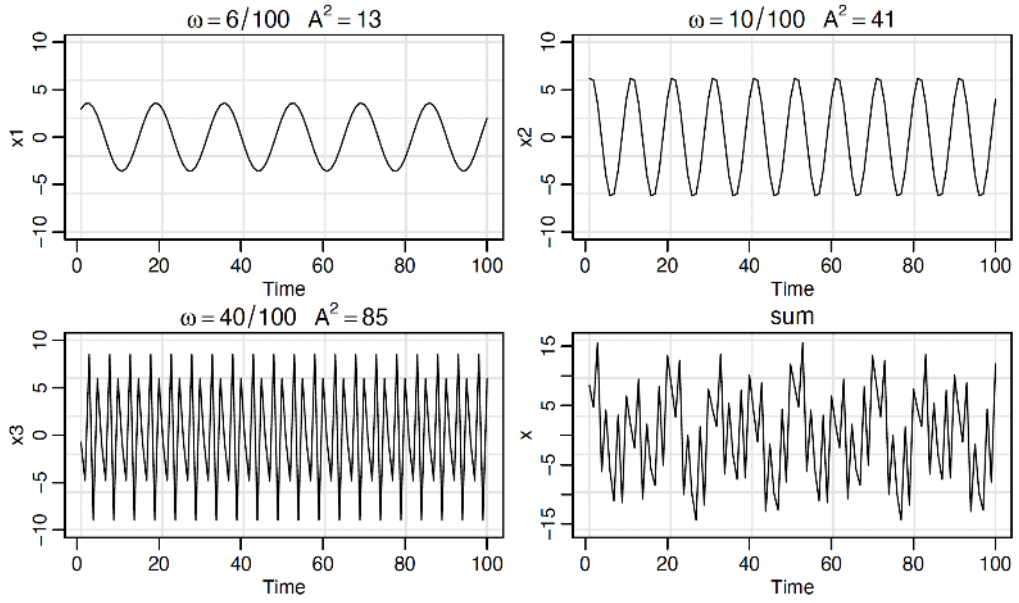


Figure 11.1: Three sinusoidal components and their sum x_t. The higher-frequency component ($\omega = 40/100$) oscillates fastest, and the largest amplitude ($A^2 = 85$) dominates the appearance of the sum.

Fundamental frequency decomposition

Theorem 11.3: Fundamental Frequency Decomposition

Any time series $x_1, \ldots, x_n$ can be written *exactly* as

$$x_t = a_0 + \sum_{j=1}^{\lfloor n/2 \rfloor} \left[a_j \cos(2\pi t j/n) + b_j \sin(2\pi t j/n) \right],$$

with

$$a_0 = \bar{x}, \qquad a_j = \frac{2}{n} \sum_{t=1}^{n} x_t \cos(2\pi t j/n), \qquad b_j = \frac{2}{n} \sum_{t=1}^{n} x_t \sin(2\pi t j/n).$$

- The frequencies $\omega_j = j/n$ are called the **Fourier** or **fundamental frequencies**.
- The decomposition is **exact** (zero residuals), not an approximation. For n observations there are exactly n basis functions — one constant plus cosine/sine pairs at each fundamental frequency — forming a set of **mutually orthogonal** vectors in $\mathbb{R}^n$ that span the entire space.

The scaled periodogram

Under the fundamental frequency decomposition, the *scaled periodogram* is defined by

$$P(j/n) = a_j^2 + b_j^2.$$

It has the following properties:

- $P(j/n)$ is an estimate of σ_j^2, which corresponds to $\omega_j = j/n$ (i.e. the scaled periodogram is the sample variance at each frequency component).
- It indicates which frequency components are large in magnitude and which are small.
- Large values of $P(j/n)$ indicate the frequencies $\omega_j = j/n$ are predominant in the series, whereas small values of $P(j/n)$ may be associated with noise.

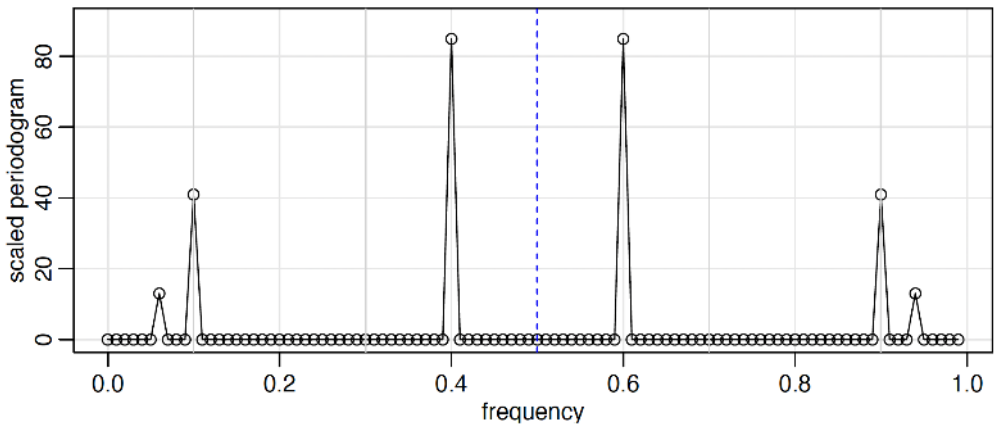


Figure 11.2: Scaled periodogram of x_t. Peaks at $\omega = 0.06, 0.10, 0.40$ recover $A_k^2 = 13, 41, 85$. The dashed line marks $\omega = 0.5$; note the mirror symmetry about 0.5.

Note: The periodogram is symmetric about $\omega = 0.5$, so in practice we only inspect $0 \leq \omega \leq 0.5$.

11.2 Spectral Density Function

Motivation

Background:

- The fundamental frequency decomposition allows us to express a time series as a sum of multiple cyclical components at different frequencies.
- Some frequencies capture a significant portion of the variation in the original series (i.e. they are *predominant*), while others contribute minimally and may resemble white noise.

Key question: Can we quantify how much variation is contributed by each frequency — the *relative importance* of each frequency component in the overall variability of the series? This leads us to the concept of *spectral density analysis*.

Remark 11.1: Why "spectral"?

If we consider the data x_t as a colour (waveform) made up of three primary colours x_{t1}, x_{t2}, x_{t3} at various strengths (amplitudes), then the periodogram decomposition acts like a prism that decomposes the colour x_t into its primary colours (spectrum). Hence the term *spectral analysis*.

Spectral representation of an autocovariance function

We can decompose a stationary time series into periodic components.

Definition 11.2: The Spectral Distribution Function

If $\{x_t\}$ is stationary with autocovariance $\gamma(h) = \text{Cov}(x_{t+h}, x_t)$, then there exists a unique monotonically increasing function $F(\omega)$, called the *spectral distribution function*, with $F(-\infty) = F(-1/2) = 0$ and $F(\infty) = F(1/2) = \gamma(0)$, such that

$$\gamma(h) = \int_{-1/2}^{1/2} e^{2\pi i \omega h} \, dF(\omega).$$

For the spectral distribution function $F(\omega)$, note that:

- $F(\omega)$ behaves like a cumulative distribution function for a discrete distribution, except that $F(\infty) = \sigma^2 = \text{Var}(x_t)$ instead of 1.
- In fact $F(\omega)$ is a cumulative distribution function, not of probabilities, but rather of variances, with $F(\infty)$ being the total variance of the process x_t.
- $F(\omega)$ characterizes how the total variance of the time series x_t is distributed across different frequencies ω.

The spectral density function

Definition 11.3: The Spectral Density Function

If the autocovariance function $\gamma(h)$ of a stationary process satisfies

$$\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty,$$

then it has the representation

$$\gamma(h) = \int_{-1/2}^{1/2} e^{2\pi i \omega h} f(\omega)\, d\omega, \qquad h = 0, \pm 1, \pm 2, \ldots,$$

where $f(\omega)$ is called the *spectral density function*.

We may also define the spectral density as the Fourier inverse transformation:

$$f(\omega) = \sum_{h=-\infty}^{\infty} \gamma(h)\, e^{-2\pi i \omega h}, \qquad -\tfrac{1}{2} \le \omega \le \tfrac{1}{2}.$$

Key properties:

- The spectral density is the analogue of the probability density function. That is, we have (a) $f(\omega) \ge 0$, and (b) $f(\omega) = f(-\omega)$.
- $\gamma(0) = \mathrm{Var}(x_t) = \int_{-1/2}^{1/2} f(\omega)\, d\omega$ (total variance = area under the spectral density).
- We will typically only plot $f(\omega)$ for $0 \le \omega \le 1/2$.

The normalised spectral density

It is often convenient to work with the **normalised** spectral density, which removes the dependence on σ^2:

$$f^*(\omega) = \frac{f(\omega)}{\gamma(0)} = \frac{f(\omega)}{\mathrm{Var}(x_t)},$$

so that $\int_{-1/2}^{1/2} f^*(\omega)\, d\omega = 1$. Equivalently,

$$f^*(\omega) = \sum_{h=-\infty}^{\infty} \rho(h)\, e^{-2\pi i \omega h},$$

i.e. the normalised spectral density is the Fourier transform of the ACF $\rho(h)$.

Example 11.2: Spectral Density of White Noise

Let $w_t \sim \mathrm{WN}(0, \sigma_w^2)$. Since $\gamma(0) = \sigma_w^2$ and $\gamma(h) = 0$ for $h \neq 0$:

$$f(\omega) = \sum_{h=-\infty}^{\infty} \gamma(h)\, e^{-2\pi i \omega h} = \sigma_w^2, \qquad -\tfrac{1}{2} \le \omega \le \tfrac{1}{2}.$$

The spectral density is **flat** — all frequencies contribute equally to the total variance. This is the origin of the name *white noise*, by analogy with white light which contains all colours at

equal intensity.

11.3 Spectral Density of ARMA Models

The linear filter

In order to derive the general spectral density of an ARMA(p, q) process, we need the following tool.

Definition 11.4: Linear Filter

A linear filter uses a set of specified coefficients a_j, for $j = 0, \pm 1, \pm 2, \ldots$, to transform an input series x_t into an output series y_t of the form

$$y_t = \sum_{j=-\infty}^{\infty} a_j x_{t-j}, \qquad \sum_{j=-\infty}^{\infty} |a_j| < \infty.$$

Note:

- The above form is also called a *convolution* in some statistical contexts.
- The coefficients are collectively called the *impulse response function*.

Theorem 11.4: Output Spectrum of a Filtered Stationary Series

For the linear filter $y_t = \sum_j a_j x_{t-j}$, if x_t has spectrum $f_x(\omega)$, then the spectrum of the filtered output y_t, say $f_y(\omega)$, is related to the spectrum of the input x_t by

$$f_y(\omega) = |A(\omega)|^2 f_x(\omega),$$

where the *frequency response function* $A(\omega)$ is defined as

$$A(\omega) = \sum_{j=-\infty}^{\infty} a_j \exp(-2\pi i \omega j).$$

Spectral density of ARMA(p, q)

An ARMA process is a *linear filter* applied to white noise.

Theorem 11.5: Spectral Density of an ARMA Process

If $\{x_t\}$ is ARMA(p, q) with $\phi(B)x_t = \theta(B)w_t$, its spectral density is

$$f_x(\omega) = \sigma_w^2 \frac{|\theta(e^{-2\pi i\omega})|^2}{|\phi(e^{-2\pi i\omega})|^2},$$

where $\phi(z) = 1 - \sum_{k=1}^{p} \phi_k z^k$ and $\theta(z) = 1 + \sum_{k=1}^{q} \theta_k z^k$.

Special cases:

- AR(p): $\theta \equiv 1$, so $f_x(\omega) = \sigma_w^2 / |\phi(e^{-2\pi i\omega})|^2$.

- MA(q): $\phi \equiv 1$, so $f_x(\omega) = \sigma_w^2 \, |\theta(e^{-2\pi i\omega})|^2$.
- White noise: $f_x(\omega) = \sigma_w^2$.

Remark 11.2: Useful Identity

A key identity used when computing $|\phi(e^{-2\pi i\omega})|^2$ for AR models is

$$|1 - \phi e^{-2\pi i\omega}|^2 = (1 - \phi e^{-2\pi i\omega})(1 - \phi e^{2\pi i\omega}) = 1 - 2\phi\cos(2\pi\omega) + \phi^2.$$

This follows from $(a-b)(a-\bar{b}) = |a|^2 - 2\,\mathrm{Re}(a\bar{b}) + |b|^2$ with $a = 1$ and $b = \phi e^{2\pi i\omega}$. Recall also $e^{i\theta} = \cos\theta + i\sin\theta$, hence $e^{i\theta} + e^{-i\theta} = 2\cos\theta$, so $\phi e^{-2\pi i\omega} + \phi e^{2\pi i\omega} = 2\phi\cos(2\pi\omega)$.

Example 11.3: Spectral Density of MA(1)

Let $x_t = w_t + \theta w_{t-1}$, so $\theta(z) = 1 + \theta z$. Then

$$f_x(\omega) = \sigma_w^2 \, |1 + \theta e^{-2\pi i\omega}|^2 = \sigma_w^2 \, (1 + \theta e^{-2\pi i\omega})(1 + \theta e^{2\pi i\omega}).$$

Expanding,

$$f_x(\omega) = \sigma_w^2 \big(1 + \theta^2 + \theta(e^{2\pi i\omega} + e^{-2\pi i\omega})\big) = \sigma_w^2 \big(1 + \theta^2 + 2\theta\cos(2\pi\omega)\big).$$

Check: integrating over $[-1/2, 1/2]$ should give $\gamma(0)$:

$$\int_{-1/2}^{1/2} f_x(\omega)\, d\omega = \sigma_w^2(1 + \theta^2) = \gamma(0). \checkmark$$

Example 11.4: Spectral Density of AR(1)

Let $x_t = \phi x_{t-1} + w_t$, so $\phi(z) = 1 - \phi z$ and $|\phi| < 1$. Then

$$f_x(\omega) = \frac{\sigma_w^2}{|1 - \phi e^{-2\pi i\omega}|^2} = \frac{\sigma_w^2}{1 - 2\phi\cos(2\pi\omega) + \phi^2}.$$

Shape:

- $\phi > 0$: spectral density is large at *low* frequencies (smooth, slowly varying series).
- $\phi < 0$: spectral density is large at *high* frequencies (rapidly oscillating series).
- $\phi = 0$: flat spectrum (white noise).

Note: The ACF of AR(1) is $\rho(h) = \phi^{|h|}$, which you can verify by checking that $\int_{-1/2}^{1/2} f_x(\omega) e^{2\pi i\omega h}\, d\omega$ reproduces $\gamma(h) = \sigma_w^2 \phi^{|h|}/(1 - \phi^2)$.

11.4 Estimating the Spectral Density

The problem and the DFT

Problem: Given observed data $x_1, \ldots, x_n$, how do we estimate the spectral density $f(\omega)$?

- Parametrically (recall the spectral density for ARMA).

- Non-parametrically using the Discrete Fourier Transform (DFT).

Definition 11.5: Discrete Fourier Transform (DFT)

Given $x_1, \ldots, x_n$, the *DFT* at fundamental frequency $\omega_j = j/n$ is

$$d(\omega_j) = n^{-1/2} \sum_{t=1}^{n} x_t \, e^{-2\pi i \omega_j t}, \qquad j = 0, 1, \ldots, n-1.$$

Interpretation:

- $d(\omega_j)$ is a complex-valued summary of how much the data oscillates at frequency ω_j.
- Large $|d(\omega_j)|$ means frequency ω_j is strongly present in the series.

The periodogram

Definition 11.6: The Periodogram

The *periodogram* is the squared modulus of the DFT:

$$I(\omega_j) = |d(\omega_j)|^2 = \sum_{h=-(n-1)}^{n-1} \hat{\gamma}(h) \, e^{-2\pi i \omega_j h}, \qquad j \neq 0.$$

Application and properties:

- $I(\omega_j)$ is the *sample analogue* of $f(\omega_j)$ — we may think of the periodogram as the sample spectral density of x_t.
- It estimates how much variance the series carries at each frequency.
- The scaled periodogram (introduced earlier) can be computed by

$$P(j/n) = \frac{4}{n} |d(\omega_j)|^2.$$