Time Series Forecasting

Class Notes

Aradhya Chakrabarti, CSE-36

October 24, 2025

Contents

1 Lesson Plan

1. **Introduction to Time Series Forecasting**
   - 1.a. Examples of Time Series
   - 1.b. Stationary Models and Auto-Correlation Function
   - 1.c. Estimation and Elimination of Trend and Seasonal Components
2. **Stationary Process and ARMA Models**
   - 2.a. Basic Properties and Linear Processes
   - 2.b. Introduction to ARMA Models
   - 2.c. Properties of Sample Mean and Autocorrelation Function
   - 2.d. Forecasting Stationary Time Series
   - 2.e. ARMA (p, q) Processes
   - 2.f. ACF and PACF
   - 2.g. Forecasting of ARMA Processes
3. **Spectral Analysis**
   - 3.a. Spectral Densities
   - 3.b. Time-Invariant Linear Filters
   - 3.c. Spectral Density of an ARMA Process
4. **Modeling and Forecasting with ARMA Processes**
   - 4.a. Preliminary Estimation
   - 4.b. Maximum Likelihood Estimation
   - 4.c. Diagnostics
   - 4.d. Forecasting
   - 4.e. Order Selection
5. **Non-stationary and Seasonal Time Series Models**
   - 5.a. ARIMA Models
   - 5.b. Identification Techniques
   - 5.c. Unit Roots in Time Series
   - 5.d. Forecasting in ARIMA Models
   - 5.e. Seasonal ARIMA Models
   - 5.f. Regression with ARMA Errors
6. **Multivariate Time Series**
   - 6.a. Second-Order Properties of Multivariate Time Series

6.b. Estimation of the Mean and Covariance

6.c. Multivariate ARMA Processes

6.d. Best Linear Predictors of Second Order Random Vectors

6.e. Modeling and Forecasting

7. **State-Space Models**

7.a. State-Space Representations

7.b. Basic Structure Model

7.c. State-Space Representation of ARIMA Models

7.d. Kalman Recursions

7.e. Estimation for State-Space Models

8. **Forecasting Techniques**

8.a. ARAR Algorithm

8.b. Holt-Winter Algorithm

8.c. Holt-Winter Seasonal Algorithm

2 Introduction to Time Series Forecasting

2.1 Introduction

A time series is a sequence of data points collected or recorded at specific time intervals, usually equally spaced (e.g., daily, monthly, yearly).

2.1.1 Key Characteristics

1. **Time-ordered:** Order matters because each value depends on time.
2. **Regular intervals:** Collected at consistent intervals (e.g., every day, every month).
3. **Continuous or discrete:** Can be measured continuously (temperature) or at intervals (monthly sales).

2.1.2 Notation for time series

A time series is often written as:

$$\{y_t\}, \quad t = 1, 2, 3, \ldots, n$$

Where:

- y_t = value of the variable at time t
- t = time index

2.1.3 Objectives of Time Series Analysis

The primary objective of time series analysis is to develop appropriate models that represent the underlying patterns and mechanisms generating a sequence of data points recorded over time.

1. **Modeling the Data-Generating Process:** To represent the underlying patterns (trend, seasonality, randomness) in the data.
2. **Forecasting Future Values:** To predict future observations based on past data.
3. **Seasonal Adjustment:** To identify and remove seasonal effects for clearer trend analysis.
4. **Hypothesis Testing:** To test assumptions or theories using time-based data (e.g., climate change).
5. **Simulation and Risk Assessment:** To simulate future scenarios and estimate probabilities of outcomes in systems.

2.1.4 Types of Analytics

Analytics is the systematic analysis of data to discover patterns, gain insights, and support decision-making. Analytics are of three types:

1. Descriptive
2. Predictive
3. Prescriptive

Descriptive, Predictive, and Prescriptive are three types of analytics that work together in a data-driven decision-making pipeline: first understand the past, then predicting the future and finally recommending optimal actions.

2.2 Statistical Basics

2.2.1 Formal Definition: Mean Function

Let $\{X_t\}$ be a time series with $E(X_t^2) < \infty$. The **mean function** of $\{X_t\}$ is defined as:

$$\mu_X(t) = E(X_t)$$

2.2.2 Formal Definition: Covariance Function

Let $\{X_t\}$ be a time series with $E(X_t^2) < \infty$. The **covariance function** of $\{X_t\}$ is defined as:

$$\gamma_X(r, s) = \text{Cov}(X_r, X_s) = E[(X_r - \mu_X(r))(X_s - \mu_X(s))]$$

for all integers r and s.

2.2.3 Formal Definitions: ACVF and ACF for Stationary Series

Let $\{X_t\}$ be a (weakly) stationary time series.

- The **autocovariance function (ACVF)** of $\{X_t\}$ at lag h is:

  $$\gamma_X(h) = \text{Cov}(X_{t+h}, X_t)$$

  Note that for a stationary series, $\gamma_X(t + h, t)$ depends only on the lag h.

- The **autocorrelation function (ACF)** of $\{X_t\}$ at lag h is:

  $$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = \text{Corr}(X_{t+h}, X_t)$$

2.2.4 Mean (Expected Value)

Mean temperature over time:

$$\mu = E[X_t] = \frac{1}{n} \sum_{t=1}^{n} X_t$$

Example: If $X = [30, 32, 31, 33, 34]$, then

$$\mu = \frac{30 + 32 + 31 + 33 + 34}{5} = 32$$

2.2.5 Variance

How much the temperature varies from the mean:

$$\text{Var}(X_t) = \frac{1}{n} \sum_{t=1}^{n} (X_t - \mu)^2$$

Higher variance → more fluctuation in daily temperatures.

2.2.6 Covariance

Measures how two variables move together. Let X_t be temperature and Y_t be humidity on day t.

$$\text{Cov}(X_t, Y_t) = \frac{1}{n}\sum_{t=1}^{n}(X_t - \overline{X})(Y_t - \overline{Y})$$

2.2.7 Data Distribution

Data distribution refers to the way data values are spread out or arranged in a dataset. It shows how frequently different values or ranges of values occur.

1. It provides a summary of how data behaves.
2. It helps in identifying patterns, outliers, skewness, and the central tendency (like mean and median).
3. Data distribution can be visualized using graphs like: Histograms, Box plots, Probability density functions (for continuous variables).

Understanding the data distribution is a fundamental step in exploratory data analysis (EDA) and model building.

2.2.8 Types of Distributions (Examples)

Distribution Type	**Shape**	**Real-life Example**
Normal (Gaussian)	Bell-shaped, symmetric	Heights of people, test scores
Uniform	Flat	Rolling a fair die
Skewed Right	Long tail to the right	Income distribution
Skewed Left	Long tail to the left	Age at retirement
Bimodal	Two peaks	Exam scores with two groups
Exponential	Rapid decline	Time between arrivals in a queue

2.2.9 Probability Density Function (PDF)

A Probability Density Function (PDF) is a function that describes the likelihood of a continuous random variable taking on a particular value or falling within a specific range.

- A PDF does not give the probability at a single point (since for continuous data, that is zero), but rather the probability over an interval.
- The area under the curve of the PDF over an interval represents the probability of the variable falling within that interval.

Properties of a PDF:

1. The PDF is always non-negative:

$$f(x) \geq 0 \quad \text{for all } x$$

2. The total area under the curve equals 1:

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

3. The probability that a value lies between a and b is:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

2.2.10 Independent and Identically Distributed (i.i.d.)

The term i.i.d. stands for Independent and Identically Distributed. It is a key assumption in many statistical and machine learning methods.

1. **Independent**
   A sequence of random variables is independent if the occurrence of one variable does not influence the occurrence of another. Mathematically:

$$P(X_1, X_2, \ldots, X_n) = P(X_1) \cdot P(X_2) \cdot \cdots \cdot P(X_n)$$

   Example: Rolling a fair die multiple times. Each roll is independent of the others.

2. **Identically Distributed**
   Random variables are said to be identically distributed if they all follow the same probability distribution. Example: Each roll of a fair 6-sided die follows the same uniform distribution:

$$P(X = k) = \frac{1}{6}; \quad \text{for } k = 1, 2, \ldots, 6$$

Definition of i.i.d.
A sequence of random variables $X_1, X_2, \ldots, X_n$ is said to be i.i.d. if:

- Each X_i is independent of the others
- Each X_i is drawn from the same probability distribution

Most time series data are not i.i.d. because:

- **Observations are not independent:** Temperature today is often similar to yesterday's.
- **Observations may not be identically distributed:** Seasonal trends cause changes in mean/variance over time.

In time series analysis, we usually assume dependence and try to model it (using ARIMA, etc.).

2.3 Examples of Time Series

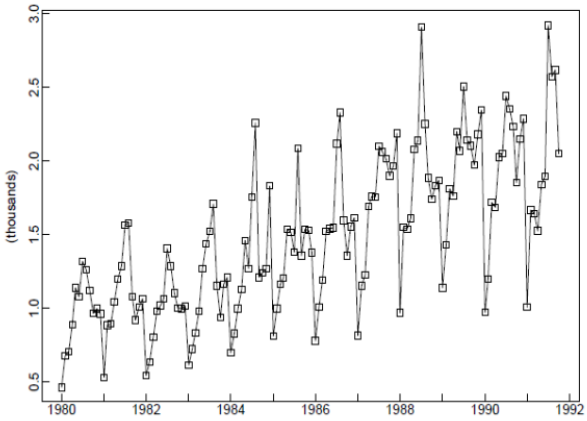

Figure 1: Monthly sales (in kiloliters) of red wine by Australian winemakers from January 1980 through October 1991.

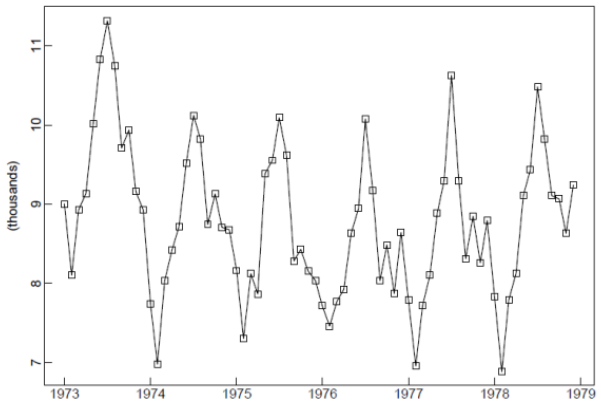

Figure 2: Monthly accidental deaths, U.S.A., 1973-1978

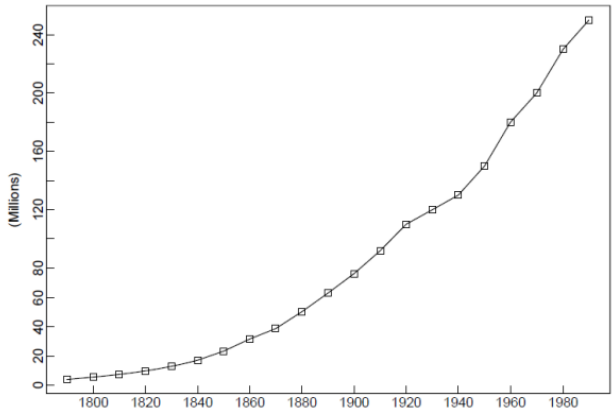

Figure 3: Population of the U.S.A., measured at ten-year intervals

2.4 Stationary Models and Auto-Correlation Function

2.4.1 Stationary Models

A time series is **(weakly) stationary** if:

- Its mean is constant over time.
- Its variance is constant over time.
- Its covariance depends only on the time difference (lag) between observations, not on the specific time points.

$\{X_t\}$ is (weakly) stationary if

(i) $\mu_X(t)$ is independent of t, and

(ii) $\gamma_X(t+h,t)$ is independent of t for each h.

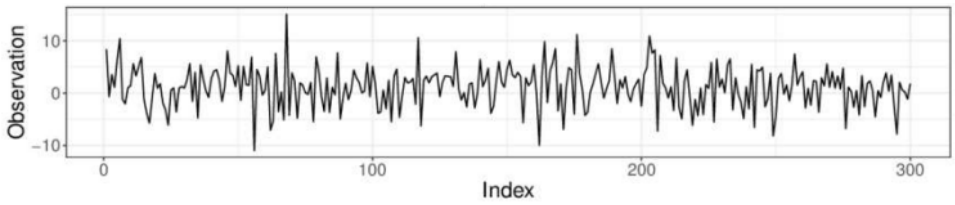

Figure 4: Stationary Time Series

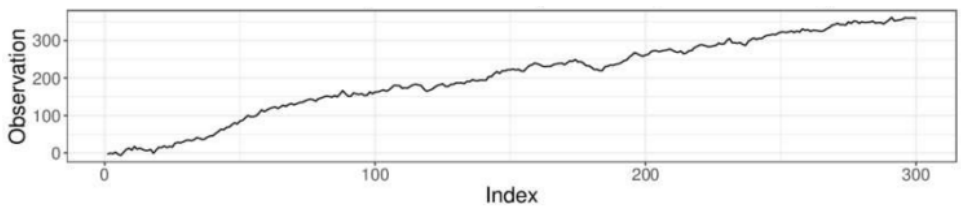

Figure 5: Non-Stationary Time Series (Time-Dependent Mean)

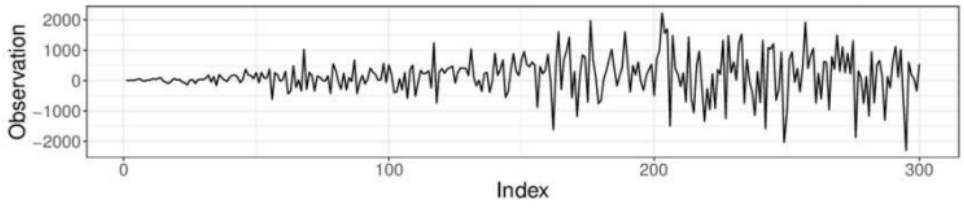

Figure 6: Non-Stationary Time Series (Time-Dependent Variance)

2.4.2 Example: iid noise

If $\{X_t\}$ is iid noise and $E(X_t^2) = \sigma^2 < \infty$, then the first requirement is obviously satisfied, since $E(X_t) = 0$ for all t. By the assumed independence,

$$\gamma_X(t+h,t) = \begin{cases} \sigma^2, & \text{if } h = 0, \\ 0, & \text{if } h \neq 0, \end{cases}$$

which does not depend on t. Hence iid noise with finite second moment is stationary. We shall use the notation

$$\{X_t\} \sim \Pi D(0, \sigma^2)$$

to indicate that the random variables X_t are independent and identically distributed random variables, each with mean 0 and variance σ^2.

2.4.3 Autocovariance Function (ACVF)

Autocovariance measures how much two values of the same time series at different time points change together - that is, how one value at time t is linearly related to a value at time $t + h$, where h is the lag.

This shows how current temperature is related to past values:

$$\gamma(k) = \frac{1}{n-k} \sum_{t=1}^{n-k} (X_t - \overline{X})(X_{t+k} - \overline{X})$$

- $k = 1$: how today's temperature relates to tomorrow's
- If there's a seasonal pattern, autocovariance will be high at certain lags

For stationary time series, $\gamma_X(h)$ depends only on the lag h, not on t. The ACVF is defined as:

$$\gamma_X(h) = \text{Cov}(X_{t+h}, X_t)$$

- $\gamma_X(h)$ is the autocovariance at lag h.
- It measures how the values of the time series at two time points t and $t + h$ co-vary.

2.4.4 Autocorrelation Function (ACF)

For a stationary time series Y_t, the autocorrelation function (ACF) at lag h is defined as:

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

Where:

- $\gamma(h)$ is the autocovariance at lag h.
- $\gamma(0)$ is the variance of Y_t.

Mathematically, the ACF measures the correlation between a time series observation at time t and its past value at time $t - h$ for a given lag h.

$$\rho(h) = \frac{\text{Cov}(Y_t, Y_{t-h})}{\text{Var}(Y_t)} = \frac{\gamma(h)}{\gamma(0)}$$

The ACF captures the **total** linear relationship between the current and past observations in the series.

Autocorrelation is the normalized version of autocovariance:

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)} \in [-1, 1]$$

- $\rho(1)$ = correlation between today and tomorrow
- $\rho(7)$ = correlation 7 days apart (weekly pattern?)

Useful for identifying seasonality or trend.

For a stationary time series, the ACF is defined as:

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = \text{Corr}(X_{t+h}, X_t)$$

- $\rho_X(h)$ is the autocorrelation coefficient at lag h.
- $\gamma_X(0) = \text{Var}(X_t)$ is the variance of the series.
- $\rho_X(h)$ is the normalized version of autocovariance so that:

$$-1 \leq \rho_X(h) \leq 1$$

3 Stationary Process and ARMA Models

3.1 Basic Properties and Linear Processes

For an ARMA(p, q) model to be valid and useful in practice, it must satisfy two essential conditions: stationarity and invertibility.

3.1.1 Condition 1: Stationarity (for AR part)

- **What it means:** The series doesn't "drift" or "explode" over time. Its mean, variance, and autocovariances stay constant over time. The influence of past values $Y_{t-1}, Y_{t-2}, \ldots$ fades away over time.
- **Technical condition:** The roots of the characteristic equation of the AR polynomial must lie outside the unit circle in the complex plane.
- **AR polynomial:** $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$
- We solve for the roots of:

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0$$

  If all roots z satisfy $|z| > 1$, the AR part is stationary.

3.1.2 Condition 2: Invertibility (for MA part)

- **What it means:** We can write the process as an infinite AR model if needed (i.e., express errors in terms of past values). This helps with parameter estimation and uniqueness of model.
- **Technical condition:** The roots of the MA polynomial must also lie outside the unit circle.
- **MA polynomial:** $\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$
- We solve:

$$1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q = 0$$

  If all roots z satisfy $|z| > 1$ the MA part is invertible.

3.1.3 Example: ARMA(1,1) Stationarity and Invertibility Check

Let an $ARMA(1, 1)$ model have the following polynomials:

- $\phi(B) = 1 - 0.6B$
- $\theta(B) = 1 + 0.4B$

Stationarity check (AR part):

$$1 - 0.6z = 0 \Rightarrow z = \frac{1}{0.6} = 1.67$$

Since $|z| = 1.67 > 1$, the process is **Stationary**.

Invertibility check (MA part):

$$1 + 0.4z = 0 \Rightarrow z = -\frac{1}{0.4} = -2.5$$

Since $|z| = 2.5 > 1$, the process is **Invertible**.

3.2 Introduction to ARMA Models

The Autoregressive Moving Average Process, $ARMA(p, q)$, is a combination of an $AR(p)$ and an $MA(q)$ model.

3.3 ARMA (p, q) Processes

3.3.1 Definition: Autoregressive Moving Average Process $ARMA(p, q)$

The general form of an $ARMA(p, q)$ process is:

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + e_t + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q}$$

3.3.2 Examples of ARMA Models

1. **ARMA(1,1):**
$$Y_t = \phi Y_{t-1} + e_t + \theta e_{t-1}$$

2. **ARMA(2,1):**
$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t + \theta_1 e_{t-1}$$

3. **AR(1) (i.e., ARMA(1,0)):**
$$Y_t = \phi Y_{t-1} + e_t$$

4. **MA(1) (i.e., ARMA(0,1)):**
$$Y_t = e_t + \theta e_{t-1}$$

3.4 ACF and PACF

3.4.1 Partial Autocorrelation Function (PACF)

The Partial Autocorrelation Function (PACF) at lag h measures the **direct** correlation between Y_t and Y_{t-h} after removing the effects of the intermediate lags $(Y_{t-1}, Y_{t-2}, \ldots, Y_{t-h+1})$. It helps isolate the contribution of a specific lag.

3.4.2 ACF Behavior for Specific Models

1. ACF of an MA(q) Process For an $MA(q)$ process:

$$Y_t = e_t + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q}$$

The ACF **cuts off** after lag q:

$$\rho(h) = 0 \quad \text{for } h > q$$

The ACF is completely determined by the θ-coefficients and is nonzero only up to lag q.

2. ACF of an AR(p) Process For an $AR(p)$ process:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t$$

The ACF does **not** cut off, but **decays gradually**. It satisfies a linear difference equation based on the AR coefficients (the Yule-Walker equations):

$$\rho(h) = \phi_1 \rho(h-1) + \phi_2 \rho(h-2) + \cdots + \phi_p \rho(h-p) \quad \text{for } h > p$$

So, for $h > p$, the autocorrelations are determined recursively.

3. ACF of an ARMA(p,q) Process An $ARMA(p, q)$ process exhibits two key behaviors in its ACF:

1. **No Cutoff:** Unlike pure $MA(q)$ models, the ACF of an $ARMA(p, q)$ process does **not** cut off after lag q. This is because the AR component introduces infinite memory into the process.
2. **Tails Off Gradually:** After a certain lag (usually after $\max(p, q)$), the autocorrelations **tail off** either exponentially or in a damped sinusoidal fashion (especially when complex roots are involved). This behavior is similar to the ACF of a pure AR process but is modified by the initial impact of the MA terms.

4 Spectral Analysis (or Frequency Domain Analysis)

4.1 Introduction to Spectral Analysis

What is Spectral Analysis?

- Decomposing a time series into sine and cosine functions of different frequencies.
- Cycle dynamics is predominant in many applications such as:
  - Infectious disease data - environmental factors, periodicity of immunity.
  - Business cycles - magnitude and periodicity of cycles.

Core Idea

- Decompose a stationary time series $\{Y_t\}$ into a combination of sinusoids, with random (and uncorrelated) coefficients.
- Identify those frequencies that appear particularly strong or important.
- This can be viewed as a linear multiple regression problem. The dependent variable is Y_t and the independent variables are the sine/cosine functions of all possible (discrete) frequencies.
- The time domain approach considers regression on past values of the time series and shocks. The frequency domain approach considers regression on sinusoids.
- The spectral density function is required for studying the frequency properties of a time series. The inferences regarding the spectral density are called the analysis in the frequency domain.

Spectral Representation Theorem

- The spectral representation theorem states that any stationary time series can be expressed as a combination of sinusoidal functions of different frequencies, each with its own amplitude and phase.
- y_t: stationary time series
- ω: angular frequency

$$y_t = \int_{-\infty}^{\infty} e^{i\omega t} Z(d\omega)$$

- $Z(d\omega)$ is a complex-valued stochastic process which determines the contribution of frequency ω to y_t.

4.2 Spectral Densities

The Spectral Density Function (SDF)

- The spectral density function $S(\omega)$ describes how the variance of a time series is distributed across different frequencies.
- The SDF provides insight into periodic components and is widely used in signal processing, finance, and other fields.

Example

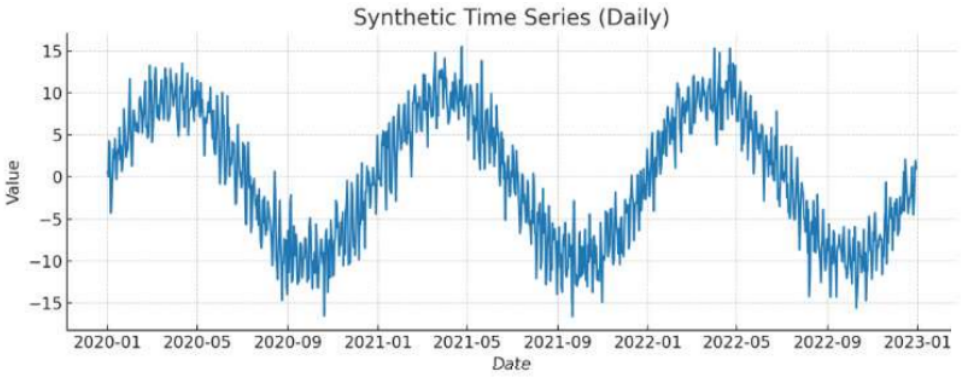

Figure 7: Synthetic Time Series (Daily)

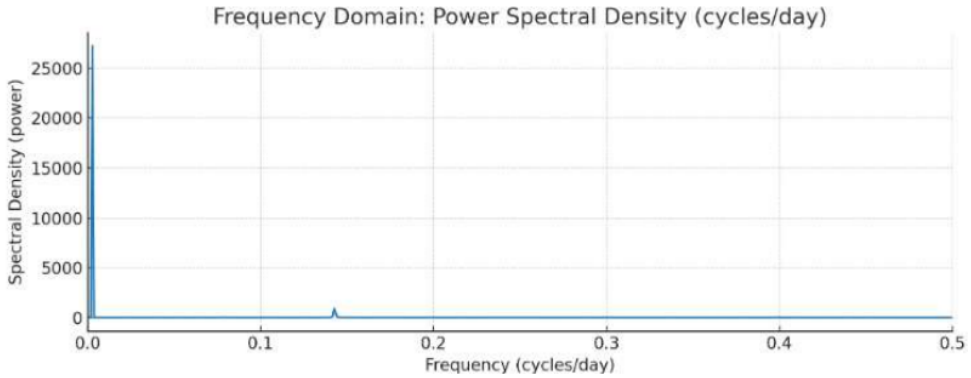

Figure 8: Frequency Domain: Power Spectral Density (cycles/day)

The time plot shows slow, big swings (a strong yearly cycle) plus a smaller, faster wiggle (a weaker weekly cycle). The frequency plot (PSD) shows a large peak near $1/365 \approx 0.0027$ cycles/day (strong long-period component) and a smaller peak near $1/7 \approx 0.1429$ cycles/day (weaker short-period component). The peak heights reflect how much variance each cycle contributes.

Definition of Spectral Density

- For a weakly stationary time series y_t, the spectral density function is the Fourier transform of the auto-covariance function $\gamma(h)$.
- ω: angular frequency
- $\gamma(h)$: auto-covariance function at lag h. $\gamma(h) = E(y_t y_{t+h})$

$$S(\omega) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma(h) e^{-i\omega h}$$

Properties of the SDF

- **Symmetry:** For a real-valued time series, $S(\omega)$ is symmetric. $S(\omega) = S(-\omega)$.

- **Non-negativity:** $S(\omega) \geq 0$ for all ω.
- **Total variance:** The total variance of the time series is the integral of the spectral density.
$$\text{Var}(y_t) = \gamma(0) = \int_{-\pi}^{\pi} S(\omega) d\omega$$
Note: The slide incorrectly shows $\sum \gamma(h)$ for variance. Variance is $\gamma(0)$.
- **Periodicity:** For discrete time series, $S(\omega)$ is periodic with period 2π.
- **Inverse relationship:** There exists an inverse relationship between the auto-covariance function and $S(\omega)$:
$$\gamma(h) = \int_{-\pi}^{\pi} S(\omega) e^{i\omega h} d\omega$$

Interpretation of the SDF

- **Low frequencies** ($\omega \approx 0$)**:** Correspond to long-term trends or slow-moving components in the time series.
- **High frequencies** ($\omega \approx \pi$)**:** Represent rapid fluctuations or noise.
- **Peaks in** $S(\omega)$**:** Indicate dominant periodic components at specific frequencies.

4.3 Spectral Density of an ARMA Process (and related models)

Example 1: White Noise

- For a white noise process, $y_t \sim N(0, \sigma^2)$.
- Auto-covariance: $\gamma(h) = \sigma^2$ for $h = 0$, and 0 otherwise.
- Spectral density: $S(\omega) = \frac{\sigma^2}{2\pi}$. This is constant for all ω.
- This reflects that white noise has equal power across all frequencies.

Example 2: AR(1) Process

- For an autoregressive process of order 1, $y_t = \phi y_{t-1} + \epsilon_t$, with $|\phi| < 1$ and $\epsilon_t \sim N(0, \sigma^2)$.
- Auto-covariance: $\gamma(h) = \frac{\sigma^2}{1-\phi^2} \phi^{|h|}$
- Spectral density: $S(\omega) = \frac{\sigma^2}{2\pi} \frac{1}{|1-\phi e^{-i\omega}|^2}$
- Low frequencies dominate for $|\phi|$ close to 1, indicating slow-moving behavior.

Example 3: MA(q) Process

- For a moving average process of order q, $y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$, with $\epsilon_t \sim N(0, \sigma^2)$.
- Spectral density: $S(\omega) = \frac{\sigma^2}{2\pi} |1 + \theta_1 e^{-i\omega} + \theta_2 e^{-2i\omega} + \cdots + \theta_q e^{-qi\omega}|^2$
- Limited to frequencies below a certain level, determined by q.
- Smoother than an AR process of similar order since it's based on past shocks.

Example 4: Random Walk (Non-stationary)

- For a random walk, $y_t = y_{t-1} + \epsilon_t$, with $\epsilon_t \sim N(0, \sigma^2)$.

- Spectral density: $S(\omega) = \frac{\sigma^2}{2pi} \frac{1}{|1-e^{-i\omega}|^2}$
- The spectral density diverges as $\omega \to 0$, indicating a non-stationary process dominated by low frequencies.

4.4 Practical Applications of Spectral Analysis

- **Signal Processing:** Fundamental for understanding and manipulating signals in the frequency domain.
  - **Noise Filtering:** Identify and remove noise from a signal by filtering out high-frequency components using the SDF. (e.g., cleaning audio, improving image clarity).
- **Telecommunications:**
  - **Modulation and demodulation:** Spectral techniques analyze and design communication signals (AM/FM, digital). (e.g., ensuring signals occupy specific frequency bands).
- **Finance and Economics:** Detect periodic patterns, trends, and cycles.
  - **Market cycle detection:** Identify cycles in stock prices, interest rates, or GDP growth. (e.g., using SDF to uncover business cycles).
- **Medicine and Biology:** Analyzing biomedical signals.
  - **Electrocardiography (ECG):** Detect abnormalities in heart rhythms by identifying irregular frequencies (e.g., diagnosing arrhythmia).

5 Types of Analytics

5.1 Descriptive Analysis

Descriptive analysis summarizes and describes historical data to understand what has happened.

Purpose: To identify patterns, trends, and insights from past events.

Techniques:

- Aggregation (e.g., sum, averages, counts)
- Data visualization (e.g., bar charts, line graphs)
- Statistical Summaries (e.g., mean, variance)
- Time Series decomposition (i.e., trend, seasonality)

Examples:

- A monthly sales report summarizing total revenue and units sold.
- A website traffic overview showing daily visitors and page views.
- A report on customer churn rate over the last quarter.

5.2 Predictive Analytics

Predictive analytics uses historical data to make predictions about future events or outcomes. It aims to model future behavior or trends using statistical and machine learning models. It answers the question: *"What is likely to happen?"*

Techniques:

- Regression analysis
- Time Series forecasting
- Classification models
- Neural Networks

Example: Forecasting next month's sales based on historical sales data.

5.3 Prescriptive Analytics

Prescriptive analytics recommends actions based on predictive insights to achieve desired outcomes. It helps in decision-making by suggesting optimal strategies or solutions. It answers the question: *"What should we do?"*

Techniques:

- Optimization algorithms (e.g., Linear Programming)
- Simulation and modeling
- Reinforcement learning
- Decision analysis

Examples:

- Recommending optimal inventory restocking levels to minimize costs.
- Optimizing delivery routes for a logistics company to save time and fuel.

5.4 Analytics Summary

Descriptive, predictive, and prescriptive analytics form a data-driven decision-making pipeline. It starts with understanding the past (Descriptive), predicting the future (Predictive), and finally recommending optimal actions (Prescriptive).

Table 1: Summary of Analytics Types

Type	**Focus**	**Question Answered**	**Common Techniques**	**Example**
Descriptive	Past data	What happened?	Aggregation, Visualization, Basic Statistics	Sales report for the last quarter
Predictive	Future outcomes	What is likely to happen?	Regression, Forecasting, ML Models	Forecasting next month's revenue
Prescriptive	Decision support	What should we do?	Optimization, Simulation, Reinforcement Learning	Recommending inventory restocking levels

6 Fundamentals of Time Series

6.1 Time Series Data

A time series is a collection of observations recorded in a sequential manner at pre-determined, equally-spaced time intervals (e.g., yearly, monthly, quarterly, or hourly). The ordering of data points is critical and must be preserved.

6.2 Seasonal Adjustment

Seasonal adjustment is the process of removing or correcting for the seasonal component in time series data. This is done so that underlying trends and other non-seasonal patterns can be analyzed more clearly.

6.2.1 Seasonality

Seasonality refers to repeating patterns or fluctuations that occur at regular intervals, such as daily, weekly, monthly, or yearly.

Example: A consistent increase in retail sales every December due to holiday shopping is a seasonal effect.

6.3 Basic Statistical Measures

6.3.1 Mean (Average)

The mean represents the central value of a dataset.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

6.3.2 Variance (Measure of Spread)

The variance measures how much the data points spread out from the mean.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

6.3.3 Standard Deviation

The standard deviation is the square root of the variance and provides a measure of spread in the original units of the data.

$$\sigma = \sqrt{\sigma^2}$$

6.3.4 Covariance (Measure of Joint Variability)

Covariance indicates how two variables change together.

$$\text{Cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- **Positive Covariance:** Both variables tend to increase or decrease together.
- **Negative Covariance:** One variable tends to increase when the other decreases.

6.3.5 Correlation (Standardized Covariance)

Correlation is a scaled version of covariance, providing a standardized measure of the linear relationship between two variables. It always ranges from -1 to +1.

$$r = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y}$$

- **+1:** A perfect positive linear relationship.
- **-1:** A perfect negative linear relationship.
- **0:** No linear relationship.

6.4 Worked Example: Statistical Calculations

Consider the following time series data:

Table 2: Sample Time Series Data

Time	X	Y
1	2	3
2	4	7
3	6	11
4	8	15
5	10	19

6.4.1 Mean of X and Y

$$\bar{x} = \frac{2+4+6+8+10}{5} = 6$$

$$\bar{y} = \frac{3+7+11+15+19}{5} = \frac{55}{5} = 11$$

6.4.2 Variance of X and Y

$$\text{Var}(X) = \frac{1}{5}\sum_{i=1}^{5}(x_i - \bar{x})^2 = \frac{1}{5}[(2-6)^2 + (4-6)^2 + (6-6)^2 + (8-6)^2 + (10-6)^2]$$

$$= \frac{1}{5}[16+4+0+4+16] = \frac{40}{5} = 8$$

Similarly, the variance for Y is calculated:

$$\text{Var}(Y) = \frac{1}{5}[(3-11)^2+(7-11)^2+(11-11)^2+(15-11)^2+(19-11)^2] = \frac{1}{5}[64+16+0+16+64] = \frac{160}{5} = 32$$

6.4.3 Covariance of X and Y

$$\text{Cov}(X,Y) = \frac{1}{5}\sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{1}{5}[(2-6)(3-11) + (4-6)(7-11) + (6-6)(11-11) + (8-6)(15-11) + (10-6)(19-11)]$$

$$= \frac{1}{5}[(-4)(-8) + (-2)(-4) + (0)(0) + (2)(4) + (4)(8)] = \frac{1}{5}[32+8+0+8+32] = \frac{80}{5} = 16$$

6.4.4 Correlation of X and Y

$$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{16}{\sqrt{8 \times 32}} = \frac{16}{\sqrt{256}} = \frac{16}{16} = 1$$

A correlation of 1 indicates a perfect positive linear relationship between X and Y.

7 Probability Concepts for Time Series

7.1 Continuous Random Variables

A random variable that can take any value within a given range or interval is known as a continuous random variable.

Examples:

- The price of a house.
- The weight of an individual.

7.2 Probability Density Function (PDF)

A function $f(x)$ is a Probability Density Function (PDF) if it satisfies the following conditions:

- $f(x) \geq 0$ for all x from $-\infty$ to ∞.
- The total area under the curve is 1: $\int_{-\infty}^{\infty} f(x)dx = 1$.

The PDF is also known as a density function. The probability that a continuous random variable falls within a specific interval is the area under the curve of the PDF over that interval. For any exact point, the probability is zero, i.e., $P(X = x) = 0$.

Example: Let X be a continuous random variable with the following PDF:

$$f(x) = \begin{cases} \lambda(2x - x^2) & 0 < x < 2 \\ 0 & \text{Otherwise} \end{cases}$$

We can find the value of λ by using the property that the total area under the curve is 1. We can also find the probability of an event, such as $P(x > 1)$.

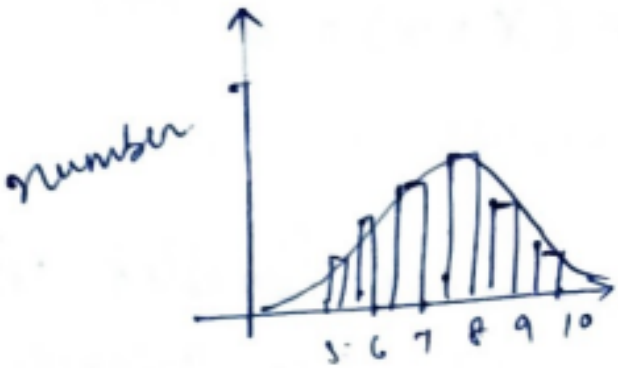

Figure 9: Graph of a Probability Density Function.

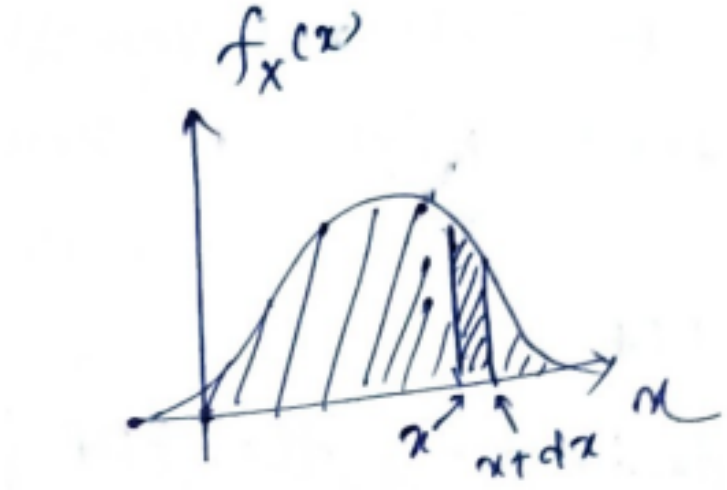

Figure 10: A histogram can approximate a PDF.

7.3 Cumulative Distribution Function (CDF)

The Cumulative Distribution Function (CDF) gives the probability that a random variable X takes on a value less than or equal to a specific value x.

$$F(x) = P(X \leq x)$$

It represents the accumulated probability up to the value x.

7.4 Joint Distributions in Time Series

For a time series, the observations at different points in time, such as Y_t and Y_s, are generally dependent. This means their joint PDF is not simply the product of their marginal PDFs:

$$f_{Y_t,Y_s}(y_t, y_s) \neq f_{Y_t}(y_t) \times f_{Y_s}(y_s)$$

Handling these complex joint distributions is challenging because we typically only have a single observation for each random variable Y_t. This difficulty necessitates a simplifying assumption, which leads to the concept of stationarity.

8 Stationarity in Time Series

Stationarity is a fundamental assumption in time series analysis. It implies that the statistical properties of the process do not change over time, meaning the process is in a state of statistical equilibrium.

There are two main types of stationarity:

1. **Strong Stationarity:** The joint distribution of any set of observations is invariant under time shifts.
2. **Weak Stationarity:** Only the first and second moments (mean, variance, and autocovariance) are invariant under time shifts.

8.1 Strong Stationarity

A process is N-th order stationary in distribution if for any time points $t_1, \ldots, t_n$ and any time shift k, the joint distribution remains the same:

$$F_{Y_{t_1},\ldots,Y_{t_n}}(y_1, \ldots, y_n) = F_{Y_{t_1+k},\ldots,Y_{t_n+k}}(y_1, \ldots, y_n)$$

8.2 Weak (Covariance) Stationarity

A time series is considered weakly stationary if it meets the following conditions:

- The mean is constant for all time points: $E(Y_t) = \mu, \forall t$.
- The variance is finite and constant for all time points: $Var(Y_t) = \sigma^2 < \infty, \forall t$.
- The covariance between any two observations depends only on the time lag between them: $Cov(Y_t, Y_{t+k}) = \gamma_k, \forall t$.

This is the most commonly used form of stationarity in practice as it does not make assumptions about the specific distribution of the data.

8.3 Non-Stationarity

A process that is not stationary is called non-stationary. In practice, most time series data exhibit non-stationarity. The main sources of non-stationarity are:

- **Trend:** A long-term increase or decrease in the data.
- **Seasonality:** Predictable, repeating patterns or cycles.
- **Heteroscedasticity:** The variance of the data changes over time.

9 Time Series Decomposition

Time series decomposition is a technique used to break down a time series into its underlying components. This helps in understanding the behavior of the series. The four main components are:

- **Secular Trend (T_t):** The long-term movement of the series.
- **Seasonal Component (S_t):** Regular, periodic variations with a cycle of one year or less.
- **Cyclical Component (C_t):** Long-term, wave-like movements with a duration of more than one year.
- **Irregular Component (I_t):** The random, unpredictable "noise" in the data.

The trend, seasonal, and cyclical components are systematic and predictable, while the irregular component is random.

9.1 Trend Component (T_t)

The trend represents the long-term direction of the time series, such as a consistent upward or downward movement, independent of seasonal or random effects.

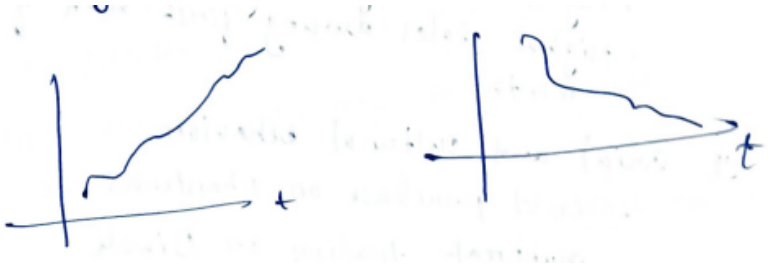

Figure 11: Example of a long-term upward trend.

9.2 Seasonal Component (S_t)

Seasonality refers to systematic, calendar-related effects that repeat at regular intervals.

Examples of causes for seasonality:

- **Natural Conditions:** Higher water consumption in summer.
- **Business Procedures:** Increased travel during holiday seasons.
- **Social and Cultural Behavior:** Higher gold purchases during festivals.

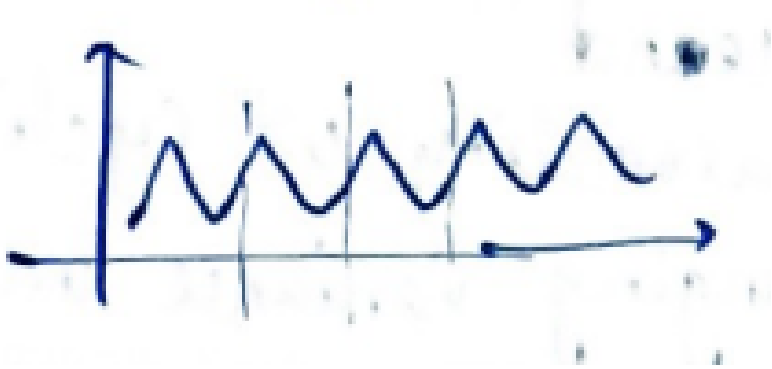

Figure 12: An illustration of seasonal variations.

9.3 Cyclical Component (C_t)

The cyclical component consists of long-term variations that repeat in a systematic, wave-like pattern of expansions and contractions. The duration of these fluctuations is typically at least two years.

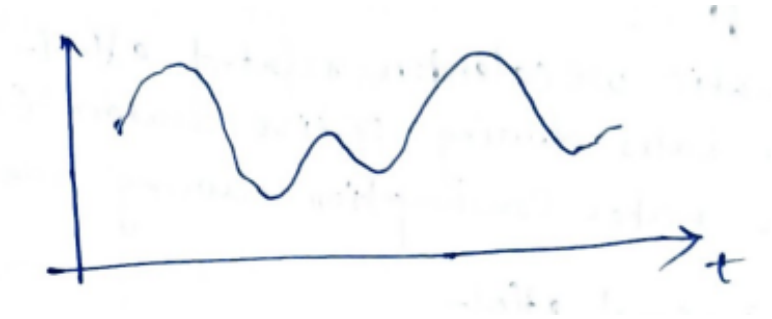

Figure 13: An example of cyclical variations over time.

9.4 Irregular Component (I_t)

Also known as the random or residual component, this is what remains after the trend, seasonal, and cyclical components have been removed. It results from short-term, unpredictable fluctuations in the data. In any given time series, any of the four components can be the dominating factor.

10 Decomposition Types

Two types: Additive Decomposition and Multiplicative Decomposition.

10.1 Additive decomposition

$$Y_t = T_t + S_t + C_t + I_t$$

Represents an absolute amount. Eg. If a manufacturer produces 10,000 more machine parts in November than in October.

10.2 Multiplicative Decomposition

$$Y_t = T_t \times S_t \times C_t \times I_t$$

Represents a relative amount. Eg. If a manufacturer produces 20% more machine parts in November as compared to October.

11 Autocovariance and Autocorrelation

11.1 Example of data autocovariance

Given $X = [4, 6, 8, 9, 7, 10, 8]$. Find Autocovariance at Lag 3. From the data, $n = 7$ and the notes state $\overline{x} = 7$. The formula for autocovariance is:

$$\gamma(k) = \frac{1}{n-k} \sum_{t=1}^{n-k} (x_t - \overline{x})(x_{t+k} - \overline{x})$$

For $k = 3$:

Table 3: Autocovariance Calculation at Lag 3 ($k = 3, \overline{x} = 7$)

t	x_t	x_{t+3}	$x_t - \overline{x}$	$x_{t+3} - \overline{x}$	Product
1	4	9	-3	2	-6
2	6	7	-1	0	0
3	8	10	1	3	3
4	9	8	2	1	2
			Sum (from notes):		-10

$$\gamma(3) = \frac{\sum \text{Product}}{n-k} = \frac{-10}{7-3} = \frac{-10}{4} = -2.5$$

11.2 Auto-correlation formula

Find auto-correlation for the above example at lag 3.

$$\rho_k = \frac{\gamma(k)}{\gamma(0)}$$

First, find $\gamma(0)$ (the variance):

$$\gamma(0) = \frac{1}{N}\sum_{t=1}^{N}(x_t - \overline{x})^2$$

Using $X = [4, 6, 8, 9, 7, 10, 8]$ and $\overline{x} = 7$:

$$\begin{aligned}\gamma(0) &= \frac{1}{7}[(4-7)^2 + (6-7)^2 + (8-7)^2 + (9-7)^2 + (7-7)^2 + (10-7)^2 + (8-7)^2] \\ &= \frac{1}{7}[9 + 1 + 4 + 4 + 0 + 9 + 1] \\ &= \frac{28}{7} = 4.0\end{aligned}$$

Now, calculate the auto-correlation at lag 3:

$$\rho(3) = \frac{\gamma(3)}{\gamma(0)} = \frac{-2.5}{4.0} = -0.625$$

11.3 Note on ACF

- If a dataset / random variable is having **trend**, then the auto-correlation / auto-covariance decreases slowly as we increase the lag.
- If a dataset is having **seasonality**, then auto-correlation is having higher values at regular intervals.

12 Check for Stationarity

12.1 Example 1: $Y_t = e_t$ (White Noise)

Consider a time series $\{Y_t\}$ where $Y_t = e_t$ and $e_t \sim i.i.d(0, \sigma^2)$. Is the process stationary?

- **Mean:** $E(Y_t) = E(e_t) = 0$. (Independent of t)
- **Variance:** $Var(Y_t) = Var(e_t) = \sigma^2 < \infty$. (Independent of t)
- **Covariance:** $Cov(Y_t, Y_{t+k}) = Cov(e_t, e_{t+k}) = 0$ for $k \neq 0$ (since $i.i.d$). (Independent of t, only depends on lag k)

Yes, $Y_t = e_t$ is stationary.

12.2 Example 2: $Y_t = e_t + 0.5e_{t-1}$

Consider $Y_t = e_t + 0.5e_{t-1}$ and $e_t \sim i.i.d(0, \sigma^2)$. Is the process stationary?

- **Mean:** $E(Y_t) = E(e_t + 0.5e_{t-1}) = E(e_t) + 0.5E(e_{t-1}) = 0 + 0.5 \times 0 = 0$. (Independent of t)

- **Variance:**

$$\begin{aligned} Var(Y_t) &= Var(e_t + 0.5e_{t-1}) \\ &= Var(e_t) + 0.5^2 Var(e_{t-1}) + 2 \times 0.5 \times Cov(e_t, e_{t-1}) \\ &= \sigma^2 + 0.25\sigma^2 + 0 \quad (\text{since } Cov(e_t, e_{t-1}) = 0) \\ &= 1.25\sigma^2 \end{aligned}$$

  (Independent of t)

- **Autocovariance (k=1):**

$$\begin{aligned} \gamma_1 &= Cov(Y_t, Y_{t-1}) = Cov(e_t + 0.5e_{t-1}, e_{t-1} + 0.5e_{t-2}) \\ &= Cov(e_t, e_{t-1}) + 0.5Cov(e_t, e_{t-2}) + 0.5Cov(e_{t-1}, e_{t-1}) + (0.5)^2 Cov(e_{t-1}, e_{t-2}) \\ &= 0 + 0 + 0.5Var(e_{t-1}) + 0 \\ &= 0.5\sigma^2 \end{aligned}$$

  (Independent of t, only depends on lag)

Since mean, variance, and autocovariance are independent of time, the time series is stationary.

12.3 Example 3: $Y_t = \sum_{i=1}^{t} e_i$ (Random Walk)

Consider $Y_t = e_1 + e_2 + \cdots + e_t = \sum_{i=1}^{t} e_i$, where $e_t \sim i.i.d(0, \sigma^2)$. Is the process stationary?

- **Mean:** $E(Y_t) = E(\sum_{i=1}^{t} e_i) = \sum_{i=1}^{t} E(e_i) = 0$. (Independent of t)
- **Variance:**

$$\begin{aligned} Var(Y_t) &= Var(e_1 + e_2 + \cdots + e_t) \\ &= \sum_{i=1}^{t} Var(e_i) \quad (\text{due to independence}) \\ &= \sum_{i=1}^{t} \sigma^2 = t\sigma^2 \end{aligned}$$

Variance increases with time t. As variance depends on time, the time series is non-stationary.

12.4 Example 4: $Y_t = a + bt + e_t$

Suppose a time series has the form $Y_t = a + bt + e_t$, where a and b are constants and $e_t \sim i.i.d(0, \sigma^2)$. Is Y_t stationary?

- **Mean:**

$$\begin{aligned} E(Y_t) &= E(a + bt + e_t) \\ &= a + bt + E(e_t) \\ &= a + bt \end{aligned}$$

Here the mean is $a + bt$, which depends on time t. So, the time series is non-stationary.

12.5 Example: Autocovariance Function for MA(1)

Consider the MA(1) process defined by:

$$X_t = Z_t + \theta Z_{t-1}, \quad \{Z_t\} \sim WN(0, \sigma^2)$$

where θ is a constant. We calculate the ACVF, $\gamma_X(h) = \text{Cov}(X_{t+h}, X_t)$.

- **Lag 0:** The variance is calculated as:

$$\begin{aligned}\gamma_X(0) &= \text{Cov}(X_t, X_t) = \text{Var}(X_t) \\ &= \text{Var}(Z_t + \theta Z_{t-1}) \\ &= \text{Var}(Z_t) + \theta^2 \text{Var}(Z_{t-1}) \quad \text{(since } Z_t, Z_{t-1} \text{ are uncorrelated)} \\ &= \sigma^2 + \theta^2\sigma^2 = \sigma^2(1+\theta^2)\end{aligned}$$

- **Lag 1:** The covariance at lag 1 is:

$$\begin{aligned}\gamma_X(1) &= \text{Cov}(X_{t+1}, X_t) = \text{Cov}(Z_{t+1} + \theta Z_t, Z_t + \theta Z_{t-1}) \\ &= \text{Cov}(Z_{t+1}, Z_t) + \theta\text{Cov}(Z_{t+1}, Z_{t-1}) + \theta\text{Cov}(Z_t, Z_t) + \theta^2\text{Cov}(Z_t, Z_{t-1}) \\ &= 0 + 0 + \theta\text{Var}(Z_t) + 0 \\ &= \theta\sigma^2\end{aligned}$$

- **Lag -1:** Since the ACVF is an even function, $\gamma_X(h) = \gamma_X(-h)$, we have $\gamma_X(-1) = \gamma_X(1) = \theta\sigma^2$.
- **Lag $|\mathbf{h}| > \mathbf{1}$:** For $h \geq 2$:

$$\gamma_X(h) = \text{Cov}(X_{t+h}, X_t) = \text{Cov}(Z_{t+h} + \theta Z_{t+h-1}, Z_t + \theta Z_{t-1})$$

  Since the indices $t+h$ and $t+h-1$ are both strictly greater than t and $t-1$ (because $h \geq 2$), all the Z terms involved are uncorrelated. Therefore, the covariance is zero.

$$\gamma_X(h) = 0 \quad \text{for } |h| > 1$$

In summary:

$$\gamma_X(h) = \begin{cases} \sigma^2(1+\theta^2) & \text{if } h = 0 \\ \sigma^2\theta & \text{if } h = \pm 1 \\ 0 & \text{if } |h| > 1 \end{cases}$$

The ACF is obtained by dividing by $\gamma_X(0)$:

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = \begin{cases} 1 & \text{if } h = 0 \\ \frac{\theta}{1+\theta^2} & \text{if } h = \pm 1 \\ 0 & \text{if } |h| > 1 \end{cases}$$

This shows the characteristic cut-off of the MA(1) ACF after lag 1.

12.6 Strong Stationarity

A time series is said to be strongly stationary if for any collection of time points $t_1, t_2, \ldots, t_k$ and any shift h, the joint distribution of $(X_{t_1}, X_{t_2}, \ldots, X_{t_k})$ is identical to the joint distribution of $(X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_k+h})$. This must hold for all k, for any time points, and any integer shift h.

A joint distribution describes the probability of two or more random variables occurring together. It tells how multiple random variables are related and how their probabilities interact.

13 Basic Time Series Models (Recap)

13.1 Time Series Decomposition

Any time series can be decomposed into:

- Secular Trend (T_t)
- Seasonal component (S_t)
- Cyclical Component (C_t)
- Random / irregular Component (I_t)

T_t, S_t, C_t are systematic components. I_t is the irregular component / completely random.

- Additive decomposition (absolute)
- Multiplicative decomposition (Relative)

13.2 White Noise

$$Y_t = e_t$$

Where $e_t \sim i.i.d(0, \sigma^2)$.

13.3 Moving Average

In this model, the current value of the series is expressed as a linear function of past error terms. Purpose: modelling past errors.

13.4 Random Walk

Given random errors e_t with mean 0 and variance σ_e^2. A time series Y_t is defined as: $Y_1 = e_1$
$Y_2 = e_1 + e_2 = Y_1 + e_2$
$Y_3 = e_1 + e_2 + e_3 = Y_2 + e_3$
...

$$Y_t = e_1 + \cdots + e_t$$

$$Y_t = Y_{t-1} + e_t$$

Y_t depends on Y_{t-1}.

13.5 Moving Average (Smoothing)

Moving average is a popular method to estimate the trend component of a time series by smoothing out short-term fluctuations and highlighting long-term patterns. A moving average is the average of a fixed number of consecutive values from a time series. As you move through the series, the average is recalculated, giving a smoother version of the data.

14 Trend Estimation

14.1 Trend Estimation with Moving Average

Formula for a k-point moving average:

$$MA_t = \frac{Y_t + Y_{t+1} + \cdots + Y_{t+(k-1)}}{k} = \frac{1}{k}\sum_{i=0}^{k-1} Y_{t+i}$$

(Note: A centered moving average is also common, but this is the form in the notes.)

14.1.1 Example: 3-month moving average

Estimate trend using 3-month moving average.

Table 4: 3-Month Moving Average Example (Values from notes)

Month	Sales (Y_t)	3-Month MA (Centered on 2nd month)
Jan	120	–
Feb	130	125.6
Mar	125	131.7
Apr	140	138.3
May	150	150.0
Jun	160	153.0
Jul	155	–

Moving average smooths the short-term fluctuations. The trend hence is positive. Conclusion: There is an upward trend. While correlation with time using the original data can show a trend, moving average helps to filter out short-term fluctuations (noise) and better reveal the underlying long-term trend.

14.2 Trend Estimation with Exponential Smoothing

Exponential Smoothing is a time series forecasting and trend elimination method that gives more weight to recent observations while still considering past data. It's useful for smoothing a time series and detecting trends, especially when data has random fluctuations.

$$S_t = \alpha Y_t + (1-\alpha)S_{t-1}$$

Where:

- S_t: Smoothed value at time t
- Y_t: Actual value at time t
- α: Smoothing constant $(0 < \alpha \leq 1)$
- S_{t-1}: Smoothed value at time $t-1$

14.2.1 Example

Time (t)	Sales (Yt)
1	100
2	110
3	105
4	120
5	130

Assume: $\alpha = 0.5$
$S_1 = Y_1 = 100$ (Initial value)

Now Compute:

$$S_t = \alpha Y_t + (1-\alpha)S_{t-1}$$
$$S_2 = 0.5 * 110 + 0.5 * 100 = 105.0$$
$$S_3 = 0.5 * 105 + 0.5 * 105 = 105.0$$
$$S_4 = 0.5 * 120 + 0.5 * 105 = 112.5$$
$$S_5 = 0.5 * 130 + 0.5 * 112.5 = 121.25$$

It is called exponential smoothing because the weight assigned to past data decreases exponentially over time.

$$S_t = \alpha Y_t + (1-\alpha)S_{t-1}$$
$$[\because S_{t-1} = \alpha Y_{t-1} + (1-\alpha)S_{t-2}]$$
$$S_t = \alpha Y_t + (1-\alpha)[\alpha Y_{t-1} + (1-\alpha)S_{t-2}]$$
$$S_t = \alpha Y_t + \alpha(1-\alpha)Y_{t-1} + (1-\alpha)^2 S_{t-2}$$

Time	Actual Sales (Yt)	Smoothed Sales (St)
1	100	100
2	110	105.0
3	105	105.0
4	120	112.5
5	130	121.25

Smoothed value shows gradual upward trend. Fluctuations are dampened. Trend becomes visible while noise is reduced.

14.3 Trend Estimation with Fast Fourier Transform (FFT)

To estimate the trend, we want to remove high-frequency variations (noise) and keep low-frequency components (slow-moving changes).

FFT helps us decompose a time series into a combination of frequencies.

- Low frequency components represents trends.
- High frequency components represents noise and short-term fluctuations.

You can apply FFT on the data, remove high frequency by keeping low frequency and apply inverse FFT to get smoothed trend present in the data.

$$Y_k = \sum_{t=0}^{N-1} Y_t \cdot e^{-\frac{2\pi i k t}{N}} \quad \text{for } k = 0, 1, \ldots, N-1$$

$$Y_t = \frac{1}{N} \sum_{k=0}^{N-1} Y_k \cdot e^{\frac{2\pi i k t}{N}}$$

- Y_t - the sample at time t
- Y_k - k^{th} sample at frequency domain
- i - complex unit ($i^2 = -1$)

14.4 Trend Estimation by Polynomial Fitting

Polynomial fitting is a curve fitting technique where a polynomial function is used to model the trend of a time series.

It works well when the underlying trend is non-linear.

$$Y_t = a_0 + a_1 t + a_2 t^2 + \cdots + a_d t^d$$

Where:

- Y_t = value of time series at time t
- d = degree of polynomial
- $a_0 \ldots a_d$ = polynomial coefficients

We want to estimate the trend using a second degree polynomial (quadratic).

$$T(t) = a_0 + a_1 t + a_2 t^2$$

Fit the model using least square regression. Minimize:

$$\sum_{t=1}^{N} (Y_t - T(t))^2$$

15 Trend Elimination

Trend elimination is a key concept in time-series analysis used to make a non-stationary time series stationary by removing its trend component.

15.1 Trend elimination by differencing

Differencing - it is a transformation where we subtract the current observation from the previous one.

$$Y_t' = Y_t - Y_{t-1}$$

Where:

- Y_t: Original time series
- Y_t': is the differenced time series

Differencing eliminates trend and makes the series stationary.

15.2 Types of differencing

- **First Order differencing:**

  $$Y_t' = Y_t - Y_{t-1}$$

  removes linear trend

- **Second order difference:**

  $$\begin{aligned} Y_t'' &= Y_t' - Y_{t-1}' \\ &= (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) \\ &= Y_t - 2Y_{t-1} + Y_{t-2} \end{aligned}$$

  removes quadratic trend

- **3. Seasonal differencing:**

  $$Y_t' = Y_t - Y_{t-s}$$

  removes seasonality (where s is the seasonality period)

Yt	Yt'
10	-
12	2
14	2
16	2

15.3 Why we need trend elimination

1. To achieve Stationarity
2. To make prediction more accurate
3. To identify underlying patterns
4. To avoid spurious relationships
5. To fit time series model properly.

15.4 Estimation and Elimination of Both Trend and Seasonality

$$Y_t = T_t + S_t + I_t$$

- Y_t: Observed time series
- T_t: trend component
- S_t: seasonal component
- I_t: Irregular/noise residual

Step 1: Estimate the trend component T_t using k-point moving average.

Step 2: Remove the trend to de-trend the series.

$$D_t = Y_t - T_t$$

D_t contains Seasonality and noise component.

Step 3: Estimate the Seasonal Component

- Assumption: Seasonality repeats every d time units.
- 1. group detrended value D_t by seasonal position. (for monthly data, group all Januaries together, all Februaries together etc)
- 2. Compute average of each group:

$$\hat{S}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} D_{j+(i-1)d}$$

  Where $\hat{S}_j$ is the seasonal index for month j.
  N_j is the number of years observed.
  d: represents time unit for which seasonality repeats.

- 3. Normalize seasonal indices:
- for additive model: Ensures they sum to 0.

$$\hat{S}_j^{norm} = \hat{S}_j - \frac{1}{d} \sum_{j=1}^{d} \hat{S}_j$$

Step 4: Estimate the Residuals (Irregular component)

$$\hat{I}_t = Y_t - T_t - S_t$$

15.5 Elimination of Trend and Seasonal components by Differencing

Step 1: Remove the trend by first Order difference.

$$\nabla Y_t = Y_t - Y_{t-1}$$

This removes a linear trend.

Step 2: Seasonal differencing (for seasonality of period d)

$$\nabla_d Y_t = Y_t - Y_{t-d}$$

Step 3: Combined differencing

$$\nabla\nabla_d Y_t = (Y_t - Y_{t-d}) - (Y_{t-1} - Y_{t-1-d})$$

Step 4: The residuals are resulted in $\nabla\nabla_d Y_t$.

Using backshift Operator B:

$$BY_t = Y_{t-1}$$

Trend removal using first difference:

$$\nabla Y_t = (1 - B)Y_t = Y_t - Y_{t-1}$$

Seasonal difference to remove seasonality of period d:

$$\nabla_d Y_t = (1 - B^d)Y_t = Y_t - Y_{t-d}$$

Combined difference:

$$\nabla\nabla_d Y_t = (1 - B)(1 - B^d)Y_t$$

This removes both linear trend and seasonal periodicity of period d.

16 Testing the estimated noise Sequence

Once you have removed trend and seasonality from time series (either by decomposition or differencing), you are left with a residual or noise sequence $\hat{I}_t$.

This sequence should ideally behave like white noise (i.e. no autocorrelation, constant mean and variance, independent values).

We need to verify whether the residuals behave like white noise.

16.1 Visual Inspection

- Plot the residuals, they should fluctuate randomly around 0.
- **ACF plot:**
- Use sample autocorrelations r_k.
- For white noise, r_k should be approximately 0 for all $k \neq 0$.
- ACF bars should lie within 95% confidence bounds: $\pm\frac{1.96}{\sqrt{n}}$ where n = sample size.

16.2 Portmanteau Test

- **Box-Pierce test:**

$$Q = n \sum_{k=1}^{h} r_k^2$$

- **Ljung-Box test (preferred due to better finite sample properties):**

$$Q^* = n(n+2) \sum_{k=1}^{h} \frac{r_k^2}{n-k}$$

- Null hypothesis (H_0): residuals are white noise.
- If Q^* is large, reject $H_0 \rightarrow$ residuals are not white noise.
- If Q^* is small $\rightarrow$ residuals are white noise.

16.3 Example

We are given a moving average process of order 2 (MA(2)) defined as:

$$X_t = Z_t + \theta Z_{t-2}$$

where $\{Z_t\}$ is white noise with mean 0 and variance 1: $Z_t \sim WN(0,1)$.

1. Find the autocovariance and autocorrelation functions for the process when $\theta = 0.8$.

$X_t = Z_t + \theta Z_{t-2}$ with $Z_t \sim WN(0,1)$

Let's compute the autocovariance function:

$$\begin{aligned}
\gamma(0) &= Var(X_t) = Var(Z_t + \theta Z_{t-2}) \\
&= Var(Z_t) + \theta^2 Var(Z_{t-2}) + 2 \cdot \theta \cdot Cov(Z_t, Z_{t-2}) \\
&= Var(Z_t) + \theta^2 Var(Z_{t-2}) \quad (\because \text{Cov}(Z_t, Z_{t-2}) = 0 \text{ for } t \neq t-2) \\
&= 1 + \theta^2 = 1 + (0.8)^2 = 1.64
\end{aligned}$$

$$\begin{aligned}
\gamma(1) &= Cov(X_t, X_{t+1}) \\
&= Cov(Z_t + \theta Z_{t-2}, Z_{t+1} + \theta Z_{t-1}) \\
&= Cov(Z_t, Z_{t+1}) + \theta Cov(Z_t, Z_{t-1}) + \theta Cov(Z_{t-2}, Z_{t+1}) + \theta^2 Cov(Z_{t-2}, Z_{t-1}) \\
&= 0 + 0 + 0 + 0 = 0
\end{aligned}$$

$$\begin{aligned}
\gamma(2) &= Cov(X_t, X_{t+2}) \\
&= Cov(Z_t + \theta Z_{t-2}, Z_{t+2} + \theta Z_t) \\
&= Cov(Z_t, Z_{t+2}) + \theta Cov(Z_t, Z_t) + \theta Cov(Z_{t-2}, Z_{t+2}) + \theta^2 Cov(Z_{t-2}, Z_t) \\
&= 0 + \theta Var(Z_t) + 0 + 0 = \theta = 0.8
\end{aligned}$$

$$\begin{aligned}\gamma(3) &= Cov(X_t, X_{t+3}) \\ &= Cov(Z_t + \theta Z_{t-2}, Z_{t+3} + \theta Z_{t+1}) = 0\end{aligned}$$

Auto-correlation function $\rho(h) = \frac{\gamma(h)}{\gamma(0)}$

$$\begin{aligned}\rho(1) &= \frac{\gamma(1)}{\gamma(0)} = \frac{0}{1.64} = 0 \\ \rho(2) &= \frac{\gamma(2)}{\gamma(0)} = \frac{0.8}{1.64} \approx 0.4878 \\ \rho(h) &= 0 \quad \text{for } h > 2\end{aligned}$$

2. Compute the variance of the sample mean $\overline{X} = \frac{1}{4}(X_1 + X_2 + X_3 + X_4)$

$$\begin{aligned}Var(\overline{X}) &= Var\left(\frac{1}{4}(X_1 + X_2 + X_3 + X_4)\right) \\ &= \frac{1}{16}Var(X_1 + X_2 + X_3 + X_4) \\ &= \frac{1}{16}\left[\sum_{i=1}^{4} Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j)\right] \\ &= \frac{1}{16}[4 \cdot \gamma(0) + 2 \cdot Cov(X_1, X_2) + 2 \cdot Cov(X_1, X_3) + 2 \cdot Cov(X_1, X_4) \\ &\quad + 2 \cdot Cov(X_2, X_3) + 2 \cdot Cov(X_2, X_4) + 2 \cdot Cov(X_3, X_4)] \\ &= \frac{1}{16}[4 \cdot \gamma(0) + 2 \cdot \gamma(1) + 2 \cdot \gamma(2) + 2 \cdot \gamma(3) \\ &\quad + 2 \cdot \gamma(1) + 2 \cdot \gamma(2) + 2 \cdot \gamma(1)] \\ &= \frac{1}{16}[4 \cdot \gamma(0) + 6 \cdot \gamma(1) + 4 \cdot \gamma(2) + 2 \cdot \gamma(3)] \\ &= \frac{1}{16}[4(1.64) + 6(0) + 4(0.8) + 2(0)] \\ &= \frac{1}{16}[6.56 + 0 + 3.2 + 0] \\ &= \frac{9.76}{16} = 0.61\end{aligned}$$

Hence,

$$\begin{aligned}\gamma(3) &= \text{cov}(X_t, X_{t+3}) \\ &= \text{cov}(Z_t + \theta Z_{t-2}, Z_{t+3} + \theta Z_{t+1}) \\ &= 0\end{aligned}$$

16.4 Example

If $\{X_t\}$ and $\{Y_t\}$ are uncorrelated stationary sequences, i.e., X_t and Y_s are uncorrelated for every r, s. Show that $\{X_t + Y_t\}$ is stationary with autocovariance function equal to the sum of the autocovariance functions of $\{X_t\}$ and $\{Y_t\}$.

Given:

- Two time series $\{X_t\}$ and $\{Y_t\}$ are stationary.
- They are uncorrelated, i.e., $\text{cov}(X_r, Y_s) = 0$ for all r, s.

We define a new time series: $Z_t = X_t + Y_t$

Show:

1. $\{Z_t\}$ is stationary.
2. The autocovariance function of $\{Z_t\}$, denoted $\gamma_Z(h)$, is: $\gamma_Z(h) = \gamma_X(h) + \gamma_Y(h)$ where $\gamma_X(h)$ and $\gamma_Y(h)$ are the autocovariance functions of $\{X_t\}$ and $\{Y_t\}$ respectively.

16.5 Step 1: Stationarity of $Z_t = X_t + Y_t$

Prove Mean:

$$\begin{aligned} E[Z_t] &= E[X_t + Y_t] \\ &= E[X_t] + E[Y_t] \\ &= \mu_X + \mu_Y = \text{Constant} \end{aligned}$$

Prove Autocovariance:

$$\begin{aligned} \gamma_Z(h) &= \text{cov}(Z_t, Z_{t+h}) \\ &= \text{cov}(X_t + Y_t, X_{t+h} + Y_{t+h}) \\ &= \text{cov}(X_t, X_{t+h}) + \text{cov}(X_t, Y_{t+h}) + \text{cov}(Y_t, X_{t+h}) + \text{cov}(Y_t, Y_{t+h}) \\ &\text{(Since } \text{cov}(X_t, Y_{t+h}) = 0 \text{ and } \text{cov}(Y_t, X_{t+h}) = 0\text{)} \\ &= \text{cov}(X_t, X_{t+h}) + \text{cov}(Y_t, Y_{t+h}) \\ &= \gamma_X(h) + \gamma_Y(h) \end{aligned}$$

This is independent of time t, and only dependent on lag h.

$\{Z_t\} = \{X_t + Y_t\}$ is stationary. The autocovariance function is $\gamma_Z(h) = \gamma_X(h) + \gamma_Y(h)$.

16.6 Why not all White Noise is i.i.d?

White Noise:

- Zero Mean: $E[X_t] = 0$
- Constant Variance: $\text{var}(X_t) = \sigma^2 < \infty$
- No Autocorrelation: $\text{corr}(X_t, X_{t+h}) = 0$ for all $h \neq 0$

i.i.d (Independent and Identically Distributed): A sequence $\{X_t\}$ is i.i.d if:

- Every X_t is independent from every other X_s.
- All X_t follow the same probability distribution.

White noise only requires uncorrelatedness, not independence. If two variables are independent, they are always uncorrelated. But if two variables are uncorrelated, they are not necessarily independent.

17 Properties of Autocovariance and Autocorrelation Function

1. $\gamma(0) \geq |\gamma(h)|$ for all h (Cauchy-Schwarz inequality)
2. $\gamma(h) = \gamma(-h)$ (Autocovariance function is symmetric)
3. The autocorrelation at lag 0 is always 1: $\rho(0) = 1$

18 Model Identification

The time series models are defined by their structure and the number of parameters (e.g., p, q). To apply these models, we must first estimate their order (e.g., the value of p, q). This process is called model identification and is based on examining the ACF and PACF plots.

ACF and PACF patterns

Model	ACF Behavior	PACF Behavior
AR(p)	Tails off gradually	Cuts off after lag p
MA(q)	Cuts off after lag q	Tails off gradually
ARMA(p,q)	Tails off	Tails off

18.1 Asymptotic Distribution of Sample ACF

For large sample sizes n, the sample autocorrelations $\hat{\rho}_k = (\hat{\rho}(1), \ldots, \hat{\rho}(k))'$ of many stationary time series (including ARMA models) are approximately normally distributed:

$$\hat{\rho}_k \approx N(\rho_k, n^{-1}W)$$

where $\rho_k = (\rho(1), \ldots, \rho(k))'$ is the vector of true autocorrelations, and W is the covariance matrix whose (i, j) element is given by **Bartlett's formula**:

$$\begin{aligned} w_{ij} = \sum_{k=-\infty}^{\infty} \{&\rho(k+i)\rho(k+j) + \rho(k-i)\rho(k+j) \\ &+ 2\rho(i)\rho(j)\rho^2(k) \\ &- 2\rho(i)\rho(k)\rho(k+j) \\ &- 2\rho(j)\rho(k)\rho(k+i)\} \end{aligned}$$

A computationally more convenient form is:

$$\begin{aligned} w_{ij} = \sum_{k=1}^{\infty} &\{\rho(k+i) + \rho(k-i) - 2\rho(i)\rho(k)\} \\ &\times \{\rho(k+j) + \rho(k-j) - 2\rho(j)\rho(k)\} \end{aligned}$$

Special Case: iid Noise: If the underlying process is iid noise, then $\rho(h) = 0$ for $h \neq 0$. Bartlett's formula simplifies to $w_{ii} = 1$ and $w_{ij} = 0$ for $i \neq j$. Thus, for large n, $\hat{\rho}(h) \approx N(0, 1/n)$ for $h \neq 0$, and the sample autocorrelations are approximately uncorrelated. This justifies the $\pm 1.96/\sqrt{n}$ bounds for testing if a series is white noise.

18.2 How to use ACF and PACF to identify model

1. Plot the ACF and PACF of the data (after ensuring stationarity).
2. Look for cut-off points in either ACF or PACF.
    - PACF cuts off after lag p: $\rightarrow$ AR(p)
    - ACF cuts off after lag q: $\rightarrow$ MA(q)
    - Both ACF and PACF tails off: $\rightarrow$ ARMA(p,q)

"Cuts off" means: Value drops to near zero (within confidence bounds) after a certain lag.

Example: Suppose you have:

- ACF: Significant at lag 1 and 2, then dies off.
- PACF: Significant at lag 1, drops afterward.

$\rightarrow$ Likely an MA(2) Model, since ACF cuts off.

18.3 Best Linear Predictor (BLP) Definition

In time series forecasting, we aim to predict future values of a process Y_t using past observations like $Y_t, Y_{t-1}, ...$ The Best Linear Predictor (BLP) is the linear combination of past values that minimizes the Mean Squared Error (MSE) between the actual future value and its forecast. The best linear predictor (BLP) is a fundamental concept used to estimate future values of a time series based on past observations, using a linear function. It aims to minimize the Mean Squared Error (MSE) among all possible linear predictors.

Given a time series $\{X_t\}$, the BLP of X_{t+h} (the value at time $t+h$) based on the observations up to time t is the linear function of past observations:

$$\hat{X}_{t+h|t} = a_0 + \sum_{i=1}^{P} a_i X_{t+1-i}$$

that minimizes the MSE:

$$E[(X_{t+h} - \hat{X}_{t+h|t})^2]$$

Forecasting Goal: We want to forecast Y_{t+h} based on available data up to time t:

$$H_t = \{Y_t, Y_{t-1}, ...\}$$

General form of Best Linear Predictor: The BLP of Y_{t+h} based on H_t is:

$$\hat{Y}_{t+h|t} = a_0 + a_1 Y_t + a_2 Y_{t-1} + \cdots + a_k Y_{t-k+1}$$

The coefficients $a_0, a_1, \ldots, a_k$ are chosen to minimize the mean squared forecast error:

$$E[(Y_{t+h} - \hat{Y}_{t+h|t})^2]$$

18.4 One-Step Ahead & Multi-Step Ahead Forecast

One-step ahead forecast is a prediction for X_{n+1} using only information available up to time n. In ARMA models, it's calculated by:

- plugging in the most recent actual data.
- using the estimated past residuals.

$$P_n X_{n+1} = \phi_1 X_n + \cdots + \theta_1 \hat{e}_n + \dots$$

Note: Since X_n, X_{n-1} etc. are known, they're used directly. The residuals $\hat{e}_t$ are computed as:

$$\hat{e}_t = X_t - P_{t-1} X_t \quad (\text{actual} - \text{predicted})$$

Multi-step ahead forecast (like $X_{n+2}, X_{n+3}, \dots$) A multi-step ahead forecast predicts multiple future values beyond just the next one. It is a recursive process:

$$P_n X_{n+2} = \phi_1 P_n X_{n+1} + \phi_2 X_n + \cdots + \theta_1 \hat{e}_{n+1} + \dots$$

We can't know future random errors $e_{n+1}, e_{n+2}, \dots$ (these are white noise). The best forecast assumes their expected value $= 0$.

18.5 Example

If: $X_t = 0.5X_{t-1} + e_t$ Then:

$$P_n X_{n+1} = 0.5 X_n$$

$$P_n X_{n+2} = 0.5 P_n X_{n+1} = 0.5(0.5 X_n)$$

$$P_n X_{n+2} = (0.5)^2 X_n$$

18.6 Example: AR(1) Fitting

$X = [2.0, 2.5, 3.0, 2.7, 3.2]$. Assume $X_6 = 3.8$. Fit an AR(1) model. Predict X_6. Compute prediction and error.

Model Setup (AR(1)) (Least Squares Estimation) $X_t = \phi X_{t-1} + e_t$ for $t = 2, \dots, n$ Goal is to estimate ϕ given observations $X_1, \dots, X_n$ such that the residuals e_t are minimized in the least squares sense.

$$e_t = X_t - \phi X_{t-1}$$

$$S(\phi) = \sum_{t=2}^{n} (X_t - \phi X_{t-1})^2$$

$$\frac{dS}{d\phi} = \sum_{t=2}^{n} 2(X_t - \phi X_{t-1})(-X_{t-1}) = 0$$

$$\sum_{t=2}^{n} (X_t X_{t-1} - \phi X_{t-1}^2) = 0$$

$$\sum_{t=2}^{n} X_t X_{t-1} = \phi \sum_{t=2}^{n} X_{t-1}^2$$

$$\hat{\phi} = \frac{\sum_{t=2}^{n} X_t X_{t-1}}{\sum_{t=2}^{n} X_{t-1}^2}$$

Compute AR(1) Coefficient $\hat{\phi}$ $X = [2.0, 2.5, 3.0, 2.7, 3.2]$ **Numerator:** $\sum_{t=2}^{5} X_t X_{t-1}$

$$\begin{aligned} &= (2.5)(2.0) + (3.0)(2.5) + (2.7)(3.0) + (3.2)(2.7) \\ &= 5.0 + 7.5 + 8.1 + 8.64 = 29.24 \end{aligned}$$

Denominator: $\sum_{t=2}^{5} X_{t-1}^2$

$$\begin{aligned} &= (2.0)^2 + (2.5)^2 + (3.0)^2 + (2.7)^2 \\ &= 4.0 + 6.25 + 9.0 + 7.29 = 26.54 \end{aligned}$$

$$\hat{\phi} = \frac{29.24}{26.54} \approx 1.102$$

Predict X_6 **(Forecast)** $(X_5 = 3.2)$

$$P_5 X_6 = \hat{\phi} X_5 = 1.102 \times 3.2 = 3.5264 \approx 3.526$$

Forecast Error (Given $X_6 = 3.8$)

$$\text{Error} = X_6 - P_5 X_6 = 3.8 - 3.526 = 0.274$$

Squared Error $= (0.274)^2 \approx 0.075$

19 ARMA(1,1) Model

$$X_t = \phi X_{t-1} + e_t + \theta e_{t-1}$$

where $e_t \sim WN(0, \sigma^2)$ is white noise.

- ϕ is the AR(1) coefficient
- θ is the MA(1) coefficient

Estimate parameters using least squares: Minimize the sum of squared residuals:

$$S(\phi, \theta) = \sum_{t=2}^{n} (X_t - \phi X_{t-1} - \theta \hat{e}_{t-1})^2$$

But $\hat{e}_{t-1}$ must be approximated recursively (e.g., set $\hat{e}_1 = 0$).

19.1 Example: ARMA(1,1) Forecast

$X = [2.0, 2.5, 3.0, 2.7, 3.2]$. Given $X_6 = 3.8, X_7 = 4.4$. Model: $X_t = 0.6X_{t-1} + e_t + 0.4e_{t-1}$, with $e_t \sim WN(0, 1)$. Assume: $e_5 = 0.1$. We want to compute: $P_5 X_6$ and $P_5 X_7$. $(\phi = 0.6, \theta = 0.4)$

Step 1: One-step ahead forecast $(P_5 X_6)$

$$\begin{aligned} P_5 X_6 &= \phi X_5 + \theta e_5 \\ &= (0.6)(3.2) + (0.4)(0.1) \\ &= 1.92 + 0.04 \\ &= 1.96 \end{aligned}$$

Step 2: Forecast Error (e_6)

$$\begin{aligned} e_6 &= X_6 - P_5X_6 \\ &= 3.8 - 1.96 \\ &= 1.84 \end{aligned}$$

Step 3: Two-step ahead forecast (P_5X_7) The formula is $P_5X_7 = \phi P_5X_6 + \theta e_6$. Also e_6 is unknown at time $t = 5 \Rightarrow$ we assume $E[e_6] = 0$.

$$\begin{aligned} P_5X_7 &= \phi P_5X_6 + \theta \cdot 0 \\ &= 0.6 \times 1.96 \\ &= 1.176 \end{aligned}$$

Forecast Error (e_7)

$$e_7 = X_7 - P_5X_7 = 4.4 - 1.176 = 3.224$$

Step 3: Two step ahead forecast (P_5X_7) We have the one-step-ahead forecast:

$$P_5X_6 = 1.96$$

The two-step-ahead forecast is:

$$P_5X_7 = \phi P_5X_6 + \theta\epsilon_6$$

Here, ϵ_6 is unknown, so we assume $\epsilon_6 = 0$.

$$P_5X_7 = \phi P_5X_6 + \theta \cdot 0$$

$$P_5X_7 = 0.6 \times 1.96 = 1.176$$

Forecast Error (ϵ_7)

$$\epsilon_7 = X_7 - P_5X_7$$

$$\epsilon_7 = 4.4 - 1.176 = 3.224$$

19.2 BLP (Best Linear Predictor) Summary

- Predictor that minimizes MSE using past observations.
- One-step forecast includes AR terms and known residuals.
- Multi-step forecast: Recursive, ignores future residuals.
- Error grows as forecast horizon increases.
- Recursive method: uses earlier forecasts to compute later ones.

20 Durbin-Levinson Algorithm

This algorithm is used to efficiently compute the Yule-Walker estimates for an autoregressive (AR) process and to recursively calculate best linear predictors for a stationary time series.

20.1 Yule-Walker Equations

The Yule-Walker equations are a set of equations used primarily in time series analysis to estimate the parameters of an AutoRegressive (AR) model. These estimates are known as Yule-Walker estimates.

For an AR(p) model:

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

Where:

- X_t: Time Series
- $\phi_1, \ldots, \phi_p$: autoregressive coefficients
- ϵ_t: white noise error (mean 0, variance σ^2)

Yule-Walker equations relate the autocovariance (or autocorrelation) function of the time series to the AR model parameters.

Let γ_k be the autocovariance at lag k, then the equations are:

$$\gamma_k = \sum_{j=1}^{P} \phi_j \gamma_{k-j}, \quad \text{for } k = 1, 2, \ldots, p$$

20.1.1 Matrix Form

$$\begin{pmatrix} \gamma(0) & \gamma(1) & \ldots & \gamma(p-1) \\ \gamma(1) & \gamma(0) & \ldots & \gamma(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(p-1) & \gamma(p-2) & \ldots & \gamma(0) \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix} = \begin{pmatrix} \gamma(1) \\ \gamma(2) \\ \vdots \\ \gamma(p) \end{pmatrix}$$

This is a Toeplitz System, which can be solved efficiently.

20.2 Goal of Durbin-Levinson Algorithm

The goal is to find the best linear predictor of X_{n+1} based on $X_1, \ldots, X_n$.

$$\hat{X}_{n+1} = \phi_{n1} X_n + \phi_{n2} X_{n-1} + \cdots + \phi_{nn} X_1$$

The coefficients ϕ_{nj} minimize the MSE of prediction.

- ϕ_{nj}: the coefficient of X_{n+1-j} in the best linear predictor of X_{n+1}.
- v_n: prediction error variance at step n.

20.3 Algorithm Steps

20.3.1 Initialization (n=1)

$$\phi_{11} = \frac{\gamma(1)}{\gamma(0)}$$

$$v_1 = \gamma(0)(1 - \phi_{11}^2)$$

20.3.2 Recursion (for k = 2 to n)

Repeat the following:

1. **Compute the new highest-lag coefficient $\phi_{k,k}$:**

$$\phi_{k,k} = \frac{\gamma(k) - \sum_{j=1}^{k-1} \phi_{k-1,j}\gamma(k-j)}{v_{k-1}}$$

2. **Update other coefficients ($j = 1, 2, \ldots, k-1$):**

$$\phi_{k,j} = \phi_{k-1,j} - \phi_{k,k}\phi_{k-1,k-j}$$

3. **Update the prediction error variance v_k:**

$$v_k = v_{k-1}(1 - \phi_{k,k}^2)$$

This gives new prediction error variances v_k at each step. At the end, you will have all coefficients ϕ_{kj} for $k = 1$ to n, and all prediction error variances v_k for $k = 1$ to n.

You get the best linear predictor at stage k:

$$\hat{X}_{k+1} = \sum_{j=1}^{k} \phi_{kj} X_{k+1-j}$$

20.3.3 Predictors by Stage

- **Stage 1:** Coefficient ϕ_{11}.
  Predictor: $\hat{X}_2 = \phi_{11}X_1$.
- **Stage 2:** Coefficients ϕ_{21}, ϕ_{22}.
  Predictor: $\hat{X}_3 = \phi_{21}X_2 + \phi_{22}X_1$.
- **Stage 3:** Coefficients $\phi_{31}, \phi_{32}, \phi_{33}$.
  Predictor: $\hat{X}_4 = \phi_{31}X_3 + \phi_{32}X_2 + \phi_{33}X_1$.

20.4 Example: Durbin-Levinson

Given $X = \{2, 1, 0, -1, -2\}$

20.4.1 Step 1: Compute Auto covariances

(Assuming these are given from the data)

- $\gamma(0) = \text{Variance} = 2$
- $\gamma(1) = \text{cov}(X_t, X_{t+1}) = 0.8$
- $\gamma(2) = \text{cov}(X_t, X_{t+2}) = -0.2$

20.4.2 Step 2: Durbin-Levinson Algorithm

k=1 (Predict X_2 from X_1)

$$\hat{X}_2 = \phi_{11} X_1$$

$$\phi_{11} = \frac{\gamma(1)}{\gamma(0)} = \frac{0.8}{2} = 0.4$$

$$v_1 = \gamma(0)(1 - \phi_{11}^2) = 2(1 - 0.4^2) = 2(1 - 0.16) = 1.68$$

k=2 (Predict X_3 from X_2, X_1)

$$\phi_{22} = \frac{\gamma(2) - \phi_{11}\gamma(1)}{v_1} = \frac{-0.2 - (0.4)(0.8)}{1.68}$$

$$\phi_{22} = \frac{-0.2 - 0.32}{1.68} = \frac{-0.52}{1.68} = -0.3095$$

$$\phi_{21} = \phi_{11} - \phi_{22}\phi_{11} = 0.4 - (-0.3095)(0.4)$$

$$\phi_{21} = 0.4 + 0.1238 = 0.5238$$

$$v_2 = v_1(1 - \phi_{22}^2) = 1.68(1 - (-0.3095)^2)$$

$$v_2 = 1.68(1 - 0.0958) = 1.519$$

20.4.3 Final Predictors

Predictor for X_2:

$$\hat{X}_2 = \phi_{11} X_1 = 0.4 X_1$$

For $X_1 = 2 \implies \hat{X}_2 = 0.4 \times 2 = 0.8$ (Actual = 1)

Predictor for X_3:

$$\hat{X}_3 = \phi_{21} X_2 + \phi_{22} X_1 = 0.5238 X_2 - 0.3095 X_1$$

For $X_2 = 1, X_1 = 2$:

$$\hat{X}_3 = 0.5238(1) - 0.3095(2)$$

$$\hat{X}_3 = 0.5238 - 0.619 = -0.0952 \text{ (Actual = 0)}$$

21 Innovations Algorithm

The innovations algorithm is a recursive method used in time series analysis to compute:

1. The best linear predictions of future values.
2. Based on past innovations (past prediction errors).
3. Along with their prediction error variance.

21.1 Steps of Innovations Algorithm

Suppose we know the autocovariances $\gamma(0), \gamma(1), \ldots, \gamma(n)$ of a zero-mean stationary time series.

21.1.1 Step 1: Initialization (t=1)

$$e_1 = X_1$$
$$v_1 = \gamma(0)$$

21.1.2 Step 2: Recursion (for t = 2 to n)

For $t = 2, \ldots, n$:

1. **Compute coefficients θ_{tj} for $j = 1$ to $t - 1$:**

$$\theta_{tj} = \frac{1}{v_j}\left(\gamma(t-j) - \sum_{k=1}^{j-1} \theta_{jk}\theta_{tk}v_k\right)$$

   (Note: The sum is 0 if $j = 1$)

2. **Compute the innovation variance:**

$$v_t = \gamma(0) - \sum_{j=1}^{t-1} \theta_{tj}^2 v_j$$

3. **Compute the predictor $\hat{X}_t$ and innovation e_t:**

$$\hat{X}_t = \sum_{j=1}^{t-1} \theta_{tj} e_j$$

$$e_t = X_t - \hat{X}_t$$

21.2 Example: Innovations Algorithm

Given $X = [2, 1, 0, -1, -2]$ Autocovariance: $\gamma(0) = 2$, $\gamma(1) = 0.8$, $\gamma(2) = -0.2$

21.2.1 Step 1 (t=1)

$$e_1 = X_1 = 2$$
$$v_1 = \gamma(0) = 2$$

21.2.2 Step 2 (t=2)

$$\theta_{2,1} = \frac{\gamma(1)}{v_1} = \frac{0.8}{2} = 0.4$$

$$v_2 = \gamma(0) - \theta_{2,1}^2 v_1 = 2 - (0.4^2 \times 2)$$
$$v_2 = 2 - (0.16 \times 2) = 2 - 0.32 = 1.68$$

$$\hat{X}_2 = \theta_{2,1} e_1 = 0.4 \times 2 = 0.8$$

$$e_2 = X_2 - \hat{X}_2 = 1 - 0.8 = 0.2$$

21.2.3 Step 3 (t=3)

Compute $\theta_{3,1}$ (for j=1):

$$\theta_{3,1} = \frac{1}{v_1}(\gamma(3-1) - 0) = \frac{\gamma(2)}{v_1} = \frac{-0.2}{2} = -0.1$$

Compute $\theta_{3,2}$ (for j=2):

$$\theta_{3,2} = \frac{1}{v_2}\left(\gamma(3-2) - \sum_{k=1}^{1} \theta_{2,k}\theta_{3,k}v_k\right)$$

$$\theta_{3,2} = \frac{1}{v_2}(\gamma(1) - \theta_{2,1}\theta_{3,1}v_1)$$

$$\theta_{3,2} = \frac{1}{1.68}(0.8 - (0.4)(-0.1)(2))$$

$$\theta_{3,2} = \frac{1}{1.68}(0.8 + 0.08) = 0.5238$$

Compute v_3:

$$v_3 = \gamma(0) - (\theta_{3,1}^2 v_1 + \theta_{3,2}^2 v_2)$$

$$v_3 = 2 - ((-0.1)^2 \times 2 + (0.5238)^2 \times 1.68)$$

$$v_3 = 2 - ((0.01 \times 2) + (0.2744 \times 1.68))$$

$$v_3 = 2 - (0.02 + 0.4610) = 2 - 0.4810 = 1.519$$

Predict $\hat{X}_3$:

$$\hat{X}_3 = \theta_{3,1}e_1 + \theta_{3,2}e_2$$

$$\hat{X}_3 = (-0.1)(2) + (0.5238)(0.2)$$

$$\hat{X}_3 = -0.2 + 0.1048 = -0.0952$$

Innovation e_3:

$$e_3 = X_3 - \hat{X}_3 = 0 - (-0.0952) = 0.0952$$

22 Spectral Analysis

Spectral analysis in time series is the study of how the variation (or "energy") in the data is distributed across different frequencies, rather than in the time domain.

- **In the time domain,** we ask: "How does the value change over time?"
- **In the frequency domain,** we ask: "Which repeating cycles or periodic patterns are present and how strong are they?"

22.1 Fast Fourier Transform (FFT)

$$X_k = \sum_{t=0}^{N-1} x_t e^{-i2\pi kt/N}$$

$$e^{-i\omega t} = \cos(\omega t) - i\sin(\omega t)$$

22.2 Spectral Density

Spectral density in time series is basically answering this question: "How is the variance of my time series distributed across different frequencies?"

It is a variance-frequency map, telling you which cycles (slow trends, seasonal patterns, rapid oscillations) dominate the data.

22.3 Spectral Density for AR(p) Model

Model: $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \epsilon_t$, where $\epsilon_t \sim WN(0, \sigma^2)$.

Using the lag operator B:

$$\phi(B) X_t = \epsilon_t$$

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

Spectral density formula AR(p):

$$f_x(\omega) = \frac{\sigma^2}{2\pi} \frac{1}{|\phi(e^{-i\omega})|^2}, \quad -\pi \leq \omega \leq \pi$$

22.3.1 AR(1)

$X_t = \phi X_{t-1} + \epsilon_t \implies \phi(B) = 1 - \phi B$

$$f_x(\omega) = \frac{\sigma^2}{2\pi} \frac{1}{|1 - \phi e^{-i\omega}|^2}$$

22.3.2 AR(2)

$$f(\omega) = \frac{\sigma^2}{2\pi} \frac{1}{|1 - \phi_1 e^{-i\omega} - \phi_2 e^{-i2\omega}|^2}$$

22.4 Spectral Density for MA(q) Model

Model: $X_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$, where $\epsilon_t \sim WN(0, \sigma^2)$.

Define the Lag polynomial:

$$X_t = \theta(B) \epsilon_t$$

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

Spectral density formula:

$$f(\omega) = \frac{\sigma^2}{2\pi} |\theta(e^{-i\omega})|^2$$

22.4.1 MA(1)

$$f(\omega) = \frac{\sigma^2}{2\pi} |1 + \theta_1 e^{-i\omega}|^2$$

22.4.2 MA(2)

$$f(\omega) = \frac{\sigma^2}{2\pi} |1 + \theta_1 e^{-i\omega} + \theta_2 e^{-i2\omega}|^2$$

23 Spectral Density for ARMA(p, q)

The ARMA(p, q) model is defined as:

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$$

where $\epsilon_t \sim WN(0, \sigma^2)$.

23.1 Polynomial Form

AR polynomial:

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

MA polynomial:

$$\Theta(B) = 1 + \Theta_1 B + \Theta_2 B^2 + \cdots + \Theta_q B^q$$

The model form using the backshift operator B is:

$$\phi(B) X_t = \Theta(B) \epsilon_t$$

The spectral density for the white noise process ϵ_t is:

$$f_\epsilon(\omega) = \frac{\sigma^2}{2\pi}$$

The spectral density for the ARMA(p, q) process X_t is given by:

$$f_x(\omega) = \left| \frac{\Theta(e^{-i\omega})}{\phi(e^{-i\omega})} \right|^2 f_\epsilon(\omega)$$

Substituting the spectral density for white noise, we get:

$$f(\omega) = \frac{\sigma^2}{2\pi} \frac{|\Theta(e^{-i\omega})|^2}{|\phi(e^{-i\omega})|^2}, \quad -\pi \leq \omega \leq \pi$$

23.2 Examples

23.2.1 ARMA(1, 1)

Formulation of ARMA(1, 1): $Y_t = \phi Y_{t-1} + e_t + \theta e_{t-1}$

Using Backshift operator B: ($Y_{t-1} = BY_t$, $e_{t-1} = Be_t$)

$\Phi(B) = 1 - \phi B$ $\Theta(B) = 1 + \theta B$

$(1 - \phi B) Y_t = (1 + \theta B) e_t$

The model is:

$$X_t = \phi X_{t-1} + \epsilon_t + \theta \epsilon_{t-1}$$

The spectral density is:

$$f(\omega) = \frac{\sigma^2}{2\pi} \frac{|1 + \theta e^{-i\omega}|^2}{|1 - \phi e^{-i\omega}|^2}$$

23.2.2 ARMA(2, 1)

(No formula provided in notes)

24 Order Selection of ARMA Models

There are several statistical criteria for order selection of ARMA models:

24.1 Final Prediction Error (FPE)

It measures how well a model predicts future observations, not just how well it fits the current sample.

$$FPE(p) = \hat{\sigma}_a^2 \frac{n+p}{n-p}$$

Where:

- $\hat{\sigma}_a^2$ is the estimated residual variance:

$$\hat{\sigma}_a^2 = \frac{1}{n} \sum_{t=1}^{n} \hat{\epsilon}_t^2$$

- n = sample size
- p = number of AR parameters (in a pure AR(p) model context)

The residuals are $\epsilon_t \sim WN(0, \sigma_a^2)$.

24.1.1 Interpretation

- If you choose a small value of p, the model will underfit. As a result, you get a high prediction error.
- If you choose a large p, you get a good in-sample-fit, but the denominator $(n-p)$ shrinks, which inflates the FPE (this penalizes complexity).

You should choose the p that minimizes the FPE.

24.2 Akaike Information Criterion (AIC)

It balances goodness-of-fit with model complexity, motivated by information theory.

$$AIC(p,q) = \ln(\hat{\sigma}_a^2) + \frac{2(p+q)}{n}$$

A smaller AIC suggests a better trade-off between fit and complexity.

24.3 Corrected AIC (AICC)

AIC can be biased when the sample size n is small relative to the number of parameters $(p+q)$. So, AICC adds a correction:

$$AICC(p,q) = \ln(\hat{\sigma}_a^2) + \frac{n+p+q}{n-p-q-2}$$

A smaller AICC indicates a better fit.

- This approach reduces the tendency of AIC to overfit in small samples.
- As $n \to \infty$, $AICC \to AIC$.
- This measure is preferred for small to moderate datasets.

24.4 Schwarz Bayesian Criterion (SBC or BIC)

This is derived from Bayesian theory and imposes a stronger penalty for model complexity.

$$BIC(p,q) = \ln(\hat{\sigma}_a^2) + \frac{(p+q)\ln(n)}{n}$$

- A model that is too simple will have a large variance ($\hat{\sigma}_a^2$), so the first term will be high, and the BIC will be large.
- A model that is too complex will have a large penalty term (the second term), so the BIC will be large.

The sweet spot is the model with the minimum BIC.

25 ARIMA(p, d, q)

The general form of an ARIMA(p, d, q) model is:

$$\phi(B)(1-B)^d X_t = \theta(B) Z_t$$

Where $(1-B)^d$ is the differencing operator.

- **First-order difference ($d = 1$):**

$$\nabla X_t = (1-B)X_t = X_t - X_{t-1}$$

- **Second-order difference ($d = 2$):**

$$(1-B)^2 X_t = (1-B)(X_t - X_{t-1}) = (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) = X_t - 2X_{t-1} + X_{t-2}$$

The process involves applying an ARMA(p, q) model to the differenced series $Y_t = (1-B)^d X_t$.

$$\phi(B) Y_t = \theta(B) Z_t$$

Substituting Y_t back gives the full ARIMA model:

$$\phi(B)(1-B)^d X_t = \theta(B) Z_t$$

26 Sample Questions - Time Series Forecasting - MID SEM

26.1 Define Time Series Analysis. Give some examples.

Answer:
Time series analysis involves analyzing time-ordered data points (observations recorded at regular intervals over time) to extract meaningful statistics and other characteristics of the data.

The primary objectives are to understand the underlying structure and patterns (like trend, seasonality, cycles, and randomness), model the data-generating process, and forecast future values. *(Ref: Notes Sec 2.1.3, Book Sec 1.2)*

Examples include:

- Monthly sales data (e.g., Australian red wine sales - Notes Fig 1, Book Fig 1-1)
- Monthly accidental deaths (Notes Fig 2, Book Fig 1-3)
- Population measured over time (e.g., US Population - Notes Fig 3, Book Fig 1-5)
- Daily stock prices
- Annual number of strikes (Book Fig 1-6)
- Signal processing data (Book Fig 1-4)

(Ref: Notes Sec 2.3, Book Sec 1.1)

26.2 Find autocovariance of the given time series at lag 1: $X = \{2, 5, 7, 9, 13, 14\}$.

Answer:
Given the time series $X = \{x_1, x_2, x_3, x_4, x_5, x_6\} = \{2, 5, 7, 9, 13, 14\}$. The number of observations is $n = 6$. First, calculate the sample mean $\bar{x}$:

$$\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t = \frac{2+5+7+9+13+14}{6} = \frac{50}{6} = \frac{25}{3} \approx 8.333$$

The sample autocovariance function at lag h is defined as:

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

For lag $h = 1$:

$$\begin{aligned}
\hat{\gamma}(1) &= \frac{1}{6}\sum_{t=1}^{5}(x_{t+1} - \bar{x})(x_t - \bar{x}) \\
&= \frac{1}{6}\Bigg[(5 - \frac{25}{3})(2 - \frac{25}{3}) + (7 - \frac{25}{3})(5 - \frac{25}{3}) + (9 - \frac{25}{3})(7 - \frac{25}{3}) \\
&\quad + (13 - \frac{25}{3})(9 - \frac{25}{3}) + (14 - \frac{25}{3})(13 - \frac{25}{3})\Bigg] \\
&= \frac{1}{6}\Bigg[(\frac{15-25}{3})(\frac{6-25}{3}) + (\frac{21-25}{3})(\frac{15-25}{3}) + (\frac{27-25}{3})(\frac{21-25}{3}) \\
&\quad + (\frac{39-25}{3})(\frac{27-25}{3}) + (\frac{42-25}{3})(\frac{39-25}{3})\Bigg] \\
&= \frac{1}{6}\Bigg[(\frac{-10}{3})(\frac{-19}{3}) + (\frac{-4}{3})(\frac{-10}{3}) + (\frac{2}{3})(\frac{-4}{3}) + (\frac{14}{3})(\frac{2}{3}) + (\frac{17}{3})(\frac{14}{3})\Bigg] \\
&= \frac{1}{6 \times 9}[190 + 40 - 8 + 28 + 238] \\
&= \frac{1}{54}[488] = \frac{244}{27} \approx 9.037
\end{aligned}$$

The sample autocovariance at lag 1 is $\hat{\gamma}(1) \approx 9.037$. *(Ref: Notes Sec 11.1, Book Def 1.4.4)*

26.3 What do you mean by Stationary time-series data?

Answer:
A time series $\{X_t\}$ is called (weakly) stationary if its statistical properties do not change over time. Specifically, it must satisfy three conditions:

1. The mean function $\mu_X(t) = E[X_t]$ is constant and does not depend on time t. Let $E[X_t] = \mu$.
2. The variance $Var(X_t) = E[(X_t - \mu)^2]$ is constant and finite for all t. Let $Var(X_t) = \sigma^2 < \infty$.
3. The autocovariance function $\gamma_X(r, s) = Cov(X_r, X_s)$ depends only on the lag (the time difference $r - s$) and not on the specific times r and s. That is, $Cov(X_t, X_{t+h}) = \gamma_X(h)$ for all t and any lag h.

In essence, a stationary series fluctuates around a constant mean with constant variance, and the correlation between values depends only on how far apart they are in time, not on when they occurred. Strong stationarity is a stricter condition requiring the entire joint probability distribution of $(X_{t_1}, \ldots, X_{t_k})$ to be the same as that of $(X_{t_1+h}, \ldots, X_{t_k+h})$ for any set of times $t_1, \ldots, t_k$ and any lag h. Weak stationarity is more commonly used in practice. *(Ref: Notes Sec 2.4.1, 8, Book Def 1.4.2)*

26.4 Differentiate between Mean, Variance, Autocovariance, and Autocorrelation.

Answer:
These are fundamental statistical measures used to describe a time series $\{X_t\}$:

- **Mean (μ or $E[X_t]$):** Describes the average value or the central level of the series. For a stationary series, the mean is constant over time, $\mu_X(t) = \mu$. Sample mean: $\bar{x} = \frac{1}{n}\sum x_t$.
- **Variance (σ^2 or $Var(X_t)$):** Measures the spread or variability of the observations around the mean. For a stationary series, the variance is constant over time, $Var(X_t) = \gamma(0) = \sigma^2$. Sample variance: $\hat{\gamma}(0) = \frac{1}{n}\sum(x_t - \bar{x})^2$.
- **Autocovariance Function (ACVF - $\gamma_X(h)$):** Measures the linear dependence between observations at different points in time, separated by a lag h. For a stationary series, $\gamma_X(h) = Cov(X_t, X_{t+h}) = E[(X_t - \mu)(X_{t+h} - \mu)]$. It depends only on the lag h. Sample ACVF: $\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-|h|}(x_{t+|h|} - \bar{x})(x_t - \bar{x})$. Note: $\gamma_X(0) = Var(X_t)$.
- **Autocorrelation Function (ACF - $\rho_X(h)$):** It is the normalized version of the autocovariance, measuring the linear relationship between X_t and X_{t+h} on a scale from -1 to +1. It's defined as $\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)}$. Sample ACF: $\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$. Note: $\rho_X(0) = 1$.

In summary: Mean and Variance describe properties at single time points (level and spread). Autocovariance and Autocorrelation describe the relationship *between* observations at different time points. ACF standardizes ACVF. *(Ref: Notes Sec 2.2, 2.4.3, 2.4.4, 6.3, Book Def 1.4.1, 1.4.3, 1.4.4)*

26.5 Mention the difference between Autocorrelation function (ACF) and Partial Autocorrelation function (PACF).

Answer:
Both ACF and PACF measure the correlation between observations in a time series at different lags, but they do so in different ways:

- **ACF ($\rho(h)$):** Measures the *total* correlation between X_t and X_{t-h}. This includes direct correlation as well as indirect correlation transmitted through the intermediate observations $X_{t-1}, X_{t-2}, \ldots, X_{t-h+1}$.
- **PACF ($\alpha(h)$ or ϕ_{hh}):** Measures the *direct* correlation between X_t and X_{t-h} after removing the linear effects of the intermediate observations $X_{t-1}, X_{t-2}, \ldots, X_{t-h+1}$. It can be thought of as the correlation between X_t and X_{t-h} conditional on the values in between.

The key difference lies in whether the influence of the intermediate lags is accounted for (PACF) or not (ACF). They are used together to identify the orders of ARMA models. *(Ref: Notes Sec 3.4.1, 18.1, Book Sec 3.2.3)*

26.6 What are the main components of a time series?

Answer:
A time series is often considered to be composed of several underlying components that represent different patterns:

1. **Secular Trend (T_t):** The long-term direction or general movement of the series (e.g., upward, downward, or stable).
2. **Seasonal Component (S_t):** Patterns that repeat over a fixed period, typically related to calendar effects (e.g., yearly seasonality in sales, monthly, weekly, or daily cycles). The period d is known ($s_{t+d} = s_t$).
3. **Cyclical Component (C_t):** Longer-term fluctuations or wave-like patterns that are not of a fixed period, often related to broader economic or business cycles. The duration is usually more than a year.
4. **Irregular Component (I_t or Y_t):** The random, unpredictable noise or residual variation left over after accounting for the other components. It's assumed to be stationary (often white noise).

These components can be combined, typically in an additive ($X_t = T_t + S_t + C_t + I_t$) or multiplicative ($X_t = T_t \times S_t \times C_t \times I_t$) model. *(Ref: Notes Sec 9, Book Sec 1.5 - Model 1.5.1)*

26.7 Mention two methods to estimate trend in time series.

Answer:
Two common methods for estimating the trend component (m_t) in a time series are:

1. **Moving Averages:** This method smooths the data by averaging observations over a specific window (e.g., a k-point moving average). A centered moving average $\hat{m}_t = (2q+1)^{-1} \sum_{j=-q}^{q} X_{t-j}$ estimates the trend at time t by averaging values around it. This helps filter out short-term fluctuations (noise) and potentially seasonality if the window size is chosen appropriately.

2. **Polynomial Fitting (Least Squares Regression):** Assumes the trend can be approximated by a polynomial function of time, $m_t = a_0 + a_1 t + a_2 t^2 + \cdots + a_d t^d$. The coefficients $a_0, \ldots, a_d$ are estimated by minimizing the sum of squared differences between the observed data x_t and the polynomial trend $\sum_{t=1}^{n}(x_t - m_t)^2$.

Other methods include exponential smoothing and spectral smoothing (FFT-based). *(Ref: Notes Sec 14, Book Sec 1.5.1)*

26.8 What is the role of the PACF in model identification?

Answer:
The Partial Autocorrelation Function (PACF) plays a crucial role, alongside the ACF, in identifying the order (p) of the autoregressive (AR) part of ARMA models. Its defining characteristic is:

- For a causal **AR(p)** process, the theoretical PACF $\alpha(h)$ will be non-zero for lags $h = 1, \ldots, p$ and will be exactly **zero** for all lags $h > p$.

Therefore, when examining the *sample* PACF ($\hat{\alpha}(h)$) of a stationary time series:

- If the sample PACF shows significant values up to lag p and then abruptly cuts off (i.e., values become statistically insignificant, typically falling within the $\pm 1.96/\sqrt{n}$ bounds) for lags greater than p, this suggests an AR(p) model is appropriate for the data.
- If both the ACF and PACF tail off gradually, it suggests an ARMA(p,q) model with $p > 0$ and $q > 0$.

(Ref: Notes Sec 18.1, Book Sec 3.2.3, Example 3.2.6)

26.9 Write the general form of an ARMA(3,2) process. Also write ARMA(3,2) using backshift operator (B).

Answer:
The general form of an Autoregressive Moving Average process of order (3,2), denoted ARMA(3,2), for a stationary time series $\{X_t\}$ is:

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \phi_3 X_{t-3} = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2}$$

where $\{Z_t\} \sim WN(0, \sigma^2)$ is white noise, and ϕ_1, ϕ_2, ϕ_3 are the autoregressive parameters, and θ_1, θ_2 are the moving average parameters.

Using the backshift operator B, where $B^k X_t = X_{t-k}$ and $B^k Z_t = Z_{t-k}$, the equation can be written more compactly as:

$$(1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3) X_t = (1 + \theta_1 B + \theta_2 B^2) Z_t$$

Or simply $\phi(B) X_t = \theta(B) Z_t$, where $\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \phi_3 z^3$ and $\theta(z) = 1 + \theta_1 z + \theta_2 z^2$. *(Ref: Notes Sec 3.3.1, Book Def 3.1.1)*

26.10 "White noise is not i.i.d". Justify the statement.

Answer:
The statement is generally true, although i.i.d. noise is a specific type of white noise.

- **White Noise (WN):** A sequence $\{X_t\}$ is white noise if it has zero mean ($E[X_t] = 0$), constant variance ($Var(X_t) = \sigma^2$), and is uncorrelated over time ($Cov(X_t, X_s) = 0$ for $t \neq s$). This focuses only on the first and second moments.
- **Independent and Identically Distributed (i.i.d.) Noise:** A sequence $\{X_t\}$ is i.i.d. if all X_t are mutually independent and follow the same probability distribution.

Justification: Independence is a stronger condition than uncorrelatedness. If a sequence is i.i.d. with finite variance, it is automatically white noise (since independence implies zero covariance). However, a sequence can be uncorrelated (satisfying the WN condition) without being independent. For non-Gaussian processes, zero correlation does not guarantee independence.

Example: (Similar to Book Problem 1.8) Let $\{Z_t\}$ be IID N(0,1). Define $X_t = Z_t$ if t is even, and $X_t = (Z_{t-1}^2 - 1)/\sqrt{2}$ if t is odd.

- $E[X_t] = 0$ for all t.
- $Var(X_t) = 1$ for all t.
- $Cov(X_t, X_s) = 0$ for $t \neq s$. (This requires checking cases for t, s even/odd). So, $\{X_t\}$ is WN(0,1).
- However, X_t and X_{t-1} are clearly dependent when t is odd (since X_t is a function of Z_{t-1} and $X_{t-1} = Z_{t-1}$). Thus, $\{X_t\}$ is not independent, and therefore not i.i.d.

(Ref: Notes Sec 16.6, Book Example 1.4.1 vs 1.4.2, Problem 1.8)

26.11 Why is model identification important before forecasting?

Answer:
Model identification is crucial before forecasting because the accuracy and reliability of forecasts depend heavily on using a model that adequately captures the underlying patterns and structure of the time series data.

- **Capturing Dynamics:** Different models (AR, MA, ARMA, ARIMA, etc.) represent different types of temporal dependencies, trends, and seasonalities. Identification helps select a model structure (e.g., the orders p, d, q) that matches the observed behavior in the data's ACF, PACF, and plots.
- **Forecasting Equations:** The specific equations used to generate forecasts are determined by the chosen model type and its estimated parameters. Using an inappropriate model will lead to suboptimal or incorrect forecasting equations.
- **Uncertainty Estimation:** Prediction intervals or measures of forecast uncertainty also depend on the model structure and parameter estimates. An incorrect model yields misleading estimates of future uncertainty.
- **Parsimony:** Identification aims to find the simplest model that adequately represents the data, avoiding overfitting (which leads to poor out-of-sample forecasts) and underfitting (which fails to capture important patterns).

In essence, identifying a good model ensures that the forecasts are based on a realistic representation of the process that generated the historical data. *(Ref: Notes Sec 18 Intro, Book Sec 1.2, 1.3.3)*

26.12 How does ACF behave for AR(p) and MA(q) process?

Answer:
The behavior of the Autocorrelation Function (ACF) is a key characteristic used to distinguish between AR and MA processes:

- **MA(q) Process:** The ACF of a Moving Average process of order q **cuts off** after lag q. This means $\rho(h) = 0$ for all lags $h > q$. The ACF is non-zero only for lags $1, \ldots, q$.
- **AR(p) Process:** The ACF of an Autoregressive process of order p **tails off gradually** towards zero as the lag h increases. The decay can be exponential or follow a damped sinusoidal pattern (if the characteristic AR polynomial has complex roots). The ACF values satisfy the Yule-Walker equations for $h > 0$. The ACF does not cut off abruptly.

This contrasting behavior (cut-off vs. tailing off) for ACF (and the opposite behavior for PACF) is fundamental for model identification using sample ACF/PACF plots. *(Ref: Notes Sec 3.4.2, 18 Table, Book Sec 3.2.2, Example 3.2.2)*

26.13 Explain time series decomposition. Mention the types of time series models.

Answer:
Time Series Decomposition: This is a method used to break down a time series into its constituent components, assuming the series can be represented as a combination of these underlying patterns. The goal is to isolate and understand each component. The standard components are:

- Trend (T_t): Long-term movement.
- Seasonality (S_t): Patterns repeating over a fixed period (e.g., year, week).
- Cycle (C_t): Longer-term fluctuations without a fixed period. (Often combined with trend).
- Irregular/Residual (I_t or Y_t): Random noise.

Decomposition can be:

- **Additive:** $X_t = T_t + S_t + C_t + I_t$. Assumes components add up, suitable when seasonal/irregular variation magnitude is constant over time.
- **Multiplicative:** $X_t = T_t \times S_t \times C_t \times I_t$. Assumes components multiply, suitable when variation magnitude changes proportionally to the level of the series. Often converted to additive via logarithms: $\log(X_t) = \log(T_t) + \log(S_t) + \log(C_t) + \log(I_t)$.

(Ref: Notes Sec 9, 10, 13.1, Book Sec 1.5)

Types of Time Series Models: These are mathematical representations used to describe and forecast time series data. Some basic types include:

- **White Noise (WN):** Uncorrelated observations with constant mean (usually 0) and variance. (Notes Sec 13.2, Book Sec 1.4 Ex 1.4.2)
- **i.i.d. Noise:** Stronger than WN; observations are independent and identically distributed. (Notes Sec 2.2.7, Book Sec 1.3.1 Ex 1.3.1)

- **Random Walk:** Current value is previous value plus white noise ($X_t = X_{t-1} + Z_t$). Non-stationary. (Notes Sec 13.4, Book Sec 1.3.1 Ex 1.3.3)
- **Moving Average (MA(q)):** Current value is a linear combination of current and past q white noise terms ($X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$). Stationary. (Notes Sec 13.3, Book Sec 2.1 Prop 2.1.1)
- **Autoregressive (AR(p)):** Current value is a linear combination of past p values plus white noise ($X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t$). Stationary under certain conditions on ϕ's. (Notes Sec 3.3 Ex 3, Book Sec 1.4 Ex 1.4.5)
- **Autoregressive Moving Average (ARMA(p,q)):** Combines AR and MA components ($\phi(B)X_t = \theta(B)Z_t$). A flexible class for modeling stationary series. (Notes Sec 3, Book Sec 3.1)
- **Autoregressive Integrated Moving Average (ARIMA(p,d,q)):** Models non-stationary series by differencing d times to achieve stationarity, then applying an ARMA(p,q) model. (Notes Sec 25, Book Sec 6.1 in later chapters)
- **Seasonal ARIMA (SARIMA):** Extends ARIMA to handle seasonal patterns.
- **State-Space Models:** A general framework involving latent state variables.

(Ref: Notes Sec 13, Book Sec 1.3, 1.4, 2.1, 2.2, 3.1)

26.14 Describe Box-Pierce test and its use.

Answer:
The Box-Pierce test is a *portmanteau test* used for diagnostic checking, specifically to test the hypothesis that a series of residuals (e.g., after fitting an ARMA model) behaves like white noise.

- **Null Hypothesis (H_0):** The residuals are white noise (i.e., uncorrelated).
- **Test Statistic:** The Box-Pierce statistic is calculated as:

$$Q = n \sum_{k=1}^{h} \hat{\rho}_k^2$$

  where n is the number of residuals, $\hat{\rho}_k$ is the sample autocorrelation of the residuals at lag k, and h is a pre-specified maximum lag to consider (e.g., 20 or $\ln(n)$).
- **Distribution under H_0:** If the residuals are truly white noise, Q is approximately distributed as a chi-squared (χ^2) random variable. The degrees of freedom depend on the context. If testing raw data for WN, df $= h$. If testing residuals from an ARMA(p,q) model, df $= h - p - q$.
- **Decision Rule:** Reject H_0 at significance level α if $Q > \chi^2_{1-\alpha}(df)$, the $(1-\alpha)$ quantile of the chi-squared distribution with the appropriate degrees of freedom. A large Q value indicates significant autocorrelation in the residuals, suggesting the model is inadequate.
- **Use:** Primarily used after fitting a time series model (like ARMA) to check if the model has captured the dependence structure adequately. If H_0 is not rejected, the residuals appear consistent with white noise, supporting the model's goodness-of-fit.

- **Refinement:** The Ljung-Box test, with statistic $Q^* = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k}$, is often preferred as its distribution is better approximated by the χ^2 distribution in finite samples.

(Ref: Notes Sec 16.2, Book Sec 1.6(b))

26.15 Check if the given process is Stationary or not: $X_t = e_t + \alpha e_{t-1}$, where $e_t \sim i.i.d(0, \sigma^2)$.

Answer:
This process is a Moving Average process of order 1, MA(1). To check for weak stationarity, we verify the three conditions:

1. **Mean:**
$$E[X_t] = E[e_t + \alpha e_{t-1}] = E[e_t] + \alpha E[e_{t-1}]$$
Since $e_t \sim i.i.d(0, \sigma^2)$, $E[e_t] = 0$ for all t.
$$E[X_t] = 0 + \alpha(0) = 0$$
The mean is constant (0) and independent of t.

2. **Variance:**
$$Var(X_t) = Var(e_t + \alpha e_{t-1})$$
Since e_t and e_{t-1} are independent (because e_t is i.i.d.), their covariance is 0.
$$Var(X_t) = Var(e_t) + Var(\alpha e_{t-1}) = Var(e_t) + \alpha^2 Var(e_{t-1})$$
Since $Var(e_t) = \sigma^2$ for all t.
$$Var(X_t) = \sigma^2 + \alpha^2\sigma^2 = \sigma^2(1+\alpha^2)$$
The variance is constant ($\sigma^2(1+\alpha^2)$) and finite (assuming $\sigma^2 < \infty$), independent of t.

3. **Autocovariance:** We need to check if $Cov(X_t, X_{t+h})$ depends only on the lag h. Let $\gamma(h) = Cov(X_t, X_{t+h}) = E[(X_t - E[X_t])(X_{t+h} - E[X_{t+h}])] = E[X_t X_{t+h}]$.
$$\gamma(h) = E[(e_t + \alpha e_{t-1})(e_{t+h} + \alpha e_{t+h-1})]$$
   - For $h = 0$: $\gamma(0) = E[(e_t + \alpha e_{t-1})^2] = E[e_t^2 + 2\alpha e_t e_{t-1} + \alpha^2 e_{t-1}^2] = \sigma^2 + 0 + \alpha^2\sigma^2 = \sigma^2(1+\alpha^2)$, which matches the variance.
   - For $h = 1$: $\gamma(1) = E[(e_t+\alpha e_{t-1})(e_{t+1}+\alpha e_t)] = E[e_t e_{t+1}+\alpha e_t^2+\alpha e_{t-1}e_{t+1}+\alpha^2 e_{t-1}e_t] = 0 + \alpha\sigma^2 + 0 + 0 = \alpha\sigma^2$.
   - For $h = -1$: $\gamma(-1) = Cov(X_t, X_{t-1}) = Cov(X_{t-1}, X_t) = \gamma(1) = \alpha\sigma^2$.
   - For $|h| > 1$: $\gamma(h) = E[e_t e_{t+h} + \alpha e_t e_{t+h-1} + \alpha e_{t-1}e_{t+h} + \alpha^2 e_{t-1}e_{t+h-1}]$. Since $|h| > 1$, all pairs of indices $(t, t+h)$, $(t, t+h-1)$, $(t-1, t+h)$, and $(t-1, t+h-1)$ are different. Because e_t is i.i.d., the expectation of the product of e at different times is 0. Thus, $\gamma(h) = 0$.

   The autocovariance $\gamma(h)$ depends only on the lag h (it is $\sigma^2(1+\alpha^2)$ for $h = 0$, $\alpha\sigma^2$ for $h = \pm 1$, and 0 otherwise) and not on time t.

Since all three conditions are met, the process $X_t = e_t + \alpha e_{t-1}$ is (weakly) stationary. *(Ref: Notes Sec 12.2, Book Example 1.4.4)*

26.16 Explain the steps of Durbin-Levinson Algorithm.

Answer:
The Durbin-Levinson algorithm is an efficient recursive method for solving the Yule-Walker equations for an autoregressive (AR) process, or more generally, for finding the best linear predictor $P_nX_{n+1} = \sum_{j=1}^{n} \phi_{nj}X_{n+1-j}$ of the next value X_{n+1} based on the past n observations $\{X_n, \ldots, X_1\}$ for a zero-mean stationary time series. It also computes the partial autocorrelation function (PACF).

Let $\gamma(h)$ be the ACVF of the stationary process. Let $v_k = E[(X_{k+1} - P_kX_{k+1})^2]$ be the mean squared error of the one-step prediction based on k past values. The algorithm proceeds as follows:

Initialization (n=1):

1. Set the initial mean squared error $v_0 = \gamma(0)$.
2. Calculate the first coefficient (which is also the first PACF value): $\phi_{11} = \gamma(1)/v_0 = \gamma(1)/\gamma(0) = \rho(1)$.
3. Calculate the mean squared error for the first step: $v_1 = v_0(1 - \phi_{11}^2)$.

Recursion (for k = 2 to n):

1. Calculate the next PACF value (the coefficient of the 'newest' observation X_1 in the predictor P_kX_{k+1}):
$$\phi_{kk} = \frac{\gamma(k) - \sum_{j=1}^{k-1} \phi_{k-1,j}\gamma(k-j)}{v_{k-1}}$$
(ϕ_{kk} is the PACF at lag k, $\alpha(k)$).
2. Update the other coefficients for the k-step predictor using the $(k-1)$-step coefficients and the new PACF value:
$$\phi_{kj} = \phi_{k-1,j} - \phi_{kk}\phi_{k-1,k-j}, \quad \text{for } j = 1, \ldots, k-1$$
3. Update the mean squared error for the k-step prediction:
$$v_k = v_{k-1}(1 - \phi_{kk}^2)$$

Output: After running the recursion up to n, the algorithm yields the coefficients $\phi_{n1}, \ldots, \phi_{nn}$ for the best linear predictor $P_nX_{n+1} = \sum_{j=1}^{n} \phi_{nj}X_{n+1-j}$ and the corresponding minimum mean squared error v_n. It also computes the PACF values $\alpha(k) = \phi_{kk}$ for $k = 1, \ldots, n$. *(Ref: Notes Sec 20.3, Book Sec 2.5.1)*

26.17 Estimate $\hat{X}_3$ and e_3 for $X_t = \{50, 73, 52, 73, 98\}$ using innovations algorithm. Assume autocovariance at lag 0, 1, 2 as 12, 10, 8 respectively ($\gamma(0) = 12, \gamma(1) = 10, \gamma(2) = 8$).

Answer:
We use the Innovations Algorithm (as presented in the Notes, Section 21.1) to find the one-step predictors $\hat{X}_t$ and innovations $e_t = X_t - \hat{X}_t$. The algorithm requires the autocovariances $\gamma(h)$.

We are given $X_1 = 50, X_2 = 73, X_3 = 52$, and $\gamma(0) = 12, \gamma(1) = 10, \gamma(2) = 8$. We assume the process has mean 0.

The algorithm steps: **Step 1: Initialization (t=1)**

- Mean squared prediction error (variance): $v_1 = \gamma(0) = 12$.
- Predictor: $\hat{X}_1 = 0$ (no past data).
- Innovation: $e_1 = X_1 - \hat{X}_1 = 50 - 0 = 50$.

Step 2: Recursion for t=2

- Calculate coefficient $\theta_{2,1}$:

$$\theta_{2,1} = \frac{1}{v_1}\left(\gamma(2-1) - \sum_{k=1}^{1-1}\dots\right) = \frac{\gamma(1)}{v_1} = \frac{10}{12} = \frac{5}{6}$$

- Calculate mean squared error v_2:

$$v_2 = \gamma(0) - \sum_{j=1}^{2-1}\theta_{2,j}^2 v_j = \gamma(0) - \theta_{2,1}^2 v_1 = 12 - \left(\frac{5}{6}\right)^2 \times 12 = 12 - \frac{25}{36}\times 12 = 12 - \frac{25}{3} = \frac{36-25}{3} = \frac{11}{3}$$

- Calculate predictor $\hat{X}_2$:

$$\hat{X}_2 = \sum_{j=1}^{2-1}\theta_{2,j}e_j = \theta_{2,1}e_1 = \frac{5}{6}\times 50 = \frac{250}{6} = \frac{125}{3} \approx 41.67$$

- Calculate innovation e_2:

$$e_2 = X_2 - \hat{X}_2 = 73 - \frac{125}{3} = \frac{219-125}{3} = \frac{94}{3} \approx 31.33$$

Step 3: Recursion for t=3

- Calculate coefficient $\theta_{3,1}$ (for j=1):

$$\theta_{3,1} = \frac{1}{v_1}\left(\gamma(3-1) - \sum_{k=1}^{1-1}\dots\right) = \frac{\gamma(2)}{v_1} = \frac{8}{12} = \frac{2}{3}$$

- Calculate coefficient $\theta_{3,2}$ (for j=2):

$$\theta_{3,2} = \frac{1}{v_2}\left(\gamma(3-2) - \sum_{k=1}^{2-1}\theta_{2,k}\theta_{3,k}v_k\right) = \frac{1}{v_2}(\gamma(1) - \theta_{2,1}\theta_{3,1}v_1)$$

$$\theta_{3,2} = \frac{1}{(11/3)}\left(10 - \left(\frac{5}{6}\right)\left(\frac{2}{3}\right)(12)\right) = \frac{3}{11}\left(10 - \frac{10}{9}\times 12\right) = \frac{3}{11}\left(10 - \frac{40}{3}\right)$$

$$\theta_{3,2} = \frac{3}{11}\left(\frac{30-40}{3}\right) = \frac{3}{11}\left(\frac{-10}{3}\right) = -\frac{10}{11}$$

- Calculate mean squared error v_3:

$$v_3 = \gamma(0) - \sum_{j=1}^{3-1} \theta_{3,j}^2 v_j = \gamma(0) - (\theta_{3,1}^2 v_1 + \theta_{3,2}^2 v_2)$$

$$v_3 = 12 - \left[\left(\frac{2}{3}\right)^2 (12) + \left(-\frac{10}{11}\right)^2 \left(\frac{11}{3}\right) \right] = 12 - \left[\frac{4}{9} \times 12 + \frac{100}{121} \times \frac{11}{3} \right]$$

$$v_3 = 12 - \left[\frac{16}{3} + \frac{100}{11 \times 3} \right] = 12 - \left[\frac{176}{33} + \frac{100}{33} \right] = 12 - \frac{276}{33} = 12 - \frac{92}{11} = \frac{132 - 92}{11} = \frac{40}{11}$$

- Calculate predictor $\hat{X}_3$:

$$\hat{X}_3 = \sum_{j=1}^{3-1} \theta_{3,j} e_j = \theta_{3,1} e_1 + \theta_{3,2} e_2$$

$$\hat{X}_3 = \left(\frac{2}{3}\right)(50) + \left(-\frac{10}{11}\right)\left(\frac{94}{3}\right) = \frac{100}{3} - \frac{940}{33} = \frac{1100 - 940}{33} = \frac{160}{33}$$

- Calculate innovation e_3:

$$e_3 = X_3 - \hat{X}_3 = 52 - \frac{160}{33} = \frac{1716 - 160}{33} = \frac{1556}{33}$$

So, the estimates are $\hat{X}_3 = \frac{160}{33} \approx 4.848$ and $e_3 = \frac{1556}{33} \approx 47.15$. *(Ref: Notes Sec 21.1, 21.2, Book Sec 2.5.2)*

26.18 Estimate trend for $X_t = \{12, 15, 17, 18, 19, 21, 27\}$ using Moving average, Exponential Smoothing ($\alpha = 0.4$), Differencing (1st, 2nd order).

Answer:
Given $X = \{12, 15, 17, 18, 19, 21, 27\}$. $n = 7$.

1. Moving Average (3-point simple MA): We calculate the average of 3 consecutive points. $\hat{m}_t = (X_{t-1} + X_t + X_{t+1})/3$. This gives estimates for $t = 2, \ldots, 6$.

- $\hat{m}_2 = (12 + 15 + 17)/3 = 44/3 \approx 14.67$
- $\hat{m}_3 = (15 + 17 + 18)/3 = 50/3 \approx 16.67$
- $\hat{m}_4 = (17 + 18 + 19)/3 = 54/3 = 18.00$
- $\hat{m}_5 = (18 + 19 + 21)/3 = 58/3 \approx 19.33$
- $\hat{m}_6 = (19 + 21 + 27)/3 = 67/3 \approx 22.33$

The estimated trend series (where calculable) is $\{\_, 14.67, 16.67, 18.00, 19.33, 22.33, \_\}$.

2. Exponential Smoothing ($\alpha = 0.4$): The formula is $S_t = \alpha X_t + (1 - \alpha) S_{t-1}$. Initialize with $S_1 = X_1 = 12$.

- $S_1 = 12$

- $S_2 = 0.4(15) + (1-0.4)(12) = 6 + 0.6(12) = 6 + 7.2 = 13.2$
- $S_3 = 0.4(17) + 0.6(13.2) = 6.8 + 7.92 = 14.72$
- $S_4 = 0.4(18) + 0.6(14.72) = 7.2 + 8.832 = 16.032$
- $S_5 = 0.4(19) + 0.6(16.032) = 7.6 + 9.6192 = 17.2192$
- $S_6 = 0.4(21) + 0.6(17.2192) = 8.4 + 10.33152 = 18.73152$
- $S_7 = 0.4(27) + 0.6(18.73152) = 10.8 + 11.238912 = 22.038912$

The estimated trend (smoothed series) is $\{12, 13.2, 14.72, 16.03, 17.22, 18.73, 22.04\}$.

3. Differencing (Trend Elimination):

- **First Order Differencing** ($\nabla X_t = X_t - X_{t-1}$): This removes a linear trend, leaving a series that should be stationary if the original trend was linear.

$$\nabla X = \{15-12, 17-15, 18-17, 19-18, 21-19, 27-21\} = \{3, 2, 1, 1, 2, 6\}$$

- **Second Order Differencing** ($\nabla^2 X_t = \nabla X_t - \nabla X_{t-1}$): This removes a quadratic trend.

$$\nabla^2 X = \{2-3, 1-2, 1-1, 2-1, 6-2\} = \{-1, -1, 0, 1, 4\}$$

Differencing doesn't estimate the trend directly but removes it to achieve stationarity. *(Ref: Notes Sec 14, 15, Book Sec 1.5)*

26.19 Find autocovariance and autocorrelation function for MA(1) when $\theta = 0.8$, $e_t \sim WN(0, 1)$.

Answer:
The MA(1) process is $X_t = e_t + \theta e_{t-1}$, with $e_t \sim WN(0, \sigma^2)$. Given $\theta = 0.8$ and $\sigma^2 = 1$.

Autocovariance Function (ACVF): $\gamma(h) = Cov(X_t, X_{t+h})$

- Lag 0: $\gamma(0) = Var(X_t) = \sigma^2(1 + \theta^2) = 1(1 + 0.8^2) = 1 + 0.64 = 1.64$.
- Lag 1: $\gamma(1) = Cov(X_t, X_{t+1}) = \sigma^2\theta = 1(0.8) = 0.8$.
- Lag -1: $\gamma(-1) = \gamma(1) = 0.8$.
- Lag $|h| > 1$: $\gamma(h) = 0$.

So, the ACVF is:

$$\gamma(h) = \begin{cases} 1.64 & \text{if } h = 0 \\ 0.8 & \text{if } h = \pm 1 \\ 0 & \text{if } |h| > 1 \end{cases}$$

Autocorrelation Function (ACF): $\rho(h) = \gamma(h)/\gamma(0)$

- Lag 0: $\rho(0) = \gamma(0)/\gamma(0) = 1$.
- Lag 1: $\rho(1) = \gamma(1)/\gamma(0) = 0.8/1.64 = 80/164 = 20/41 \approx 0.4878$.
- Lag -1: $\rho(-1) = \rho(1) \approx 0.4878$.

- Lag $|h| > 1$: $\rho(h) = 0/1.64 = 0$.

So, the ACF is:

$$\rho(h) = \begin{cases} 1 & \text{if } h = 0 \\ 20/41 \approx 0.4878 & \text{if } h = \pm 1 \\ 0 & \text{if } |h| > 1 \end{cases}$$

(Ref: Notes Sec 16.3, Book Example 1.4.4)

26.20 Explain in detail how to choose the order of AR, MA, and ARMA.

Answer:
Choosing the appropriate orders (p, q) for an ARMA(p,q) model (or potentially ARIMA(p,d,q) after differencing d times) involves several steps, primarily using the sample ACF and PACF plots, complemented by information criteria. Assume the series has been made stationary (e.g., by differencing).

1. **Examine Sample ACF and PACF Plots:** Calculate and plot the sample ACF ($\hat{\rho}(h)$) and sample PACF ($\hat{\alpha}(h)$) for the stationary series, including significance bounds (typically $\pm 1.96/\sqrt{n}$). Look for the following theoretical patterns:
   - **AR(p) Signature:** ACF tails off gradually (exponentially or damped sinusoidally). PACF cuts off abruptly after lag p (i.e., $\hat{\alpha}(h)$ is significant for $h \leq p$ and insignificant for $h > p$). $\implies$ Choose AR(p) model.
   - **MA(q) Signature:** ACF cuts off abruptly after lag q (i.e., $\hat{\rho}(h)$ is significant for $h \leq q$ and insignificant for $h > q$). PACF tails off gradually. $\implies$ Choose MA(q) model.
   - **ARMA(p,q) Signature:** Both ACF and PACF tail off gradually. Determining p and q simultaneously is harder. The ACF behavior after lag q is dictated by the AR part, and the PACF behavior after lag p is dictated by the MA part. Look for where the decay pattern seems to stabilize in both plots. $\implies$ Choose ARMA(p,q) model (with $p > 0, q > 0$).
   - **White Noise Signature:** Both ACF and PACF have no significant values beyond lag 0. $\implies$ Model as WN (ARMA(0,0)).
2. **Propose Candidate Models:** Based on the ACF/PACF patterns, propose one or more candidate models (e.g., AR(2), MA(1), ARMA(1,1)). It's common to start with low-order models.
3. **Parameter Estimation:** Estimate the parameters for each candidate model (using methods like Maximum Likelihood Estimation, discussed later).
4. **Diagnostic Checking:** For each fitted model, examine the residuals ($X_t - \hat{X}_t$). The residuals should ideally resemble white noise. Plot the residuals, their ACF, and PACF. Use portmanteau tests (like Ljung-Box, Notes Sec 16.2) to formally check if residual autocorrelations are collectively zero. If residuals are not WN, the model is inadequate.
5. **Information Criteria (Model Comparison):** If multiple models pass diagnostic checks, use information criteria to select the 'best' one. Common criteria include:

- **AIC (Akaike Information Criterion):** $AIC = -2\ln(L)+2k$ or $AIC = n\ln(\hat{\sigma}^2)+2k$
- **AICC (Corrected AIC):** $AICC = AIC + \frac{2k(k+1)}{n-k-1}$ (preferred for small samples)
- **BIC (Bayesian Information Criterion) / SBC:** $BIC = -2\ln(L) + k\ln(n)$ or $BIC = n\ln(\hat{\sigma}^2) + k\ln(n)$

Where L is the maximized likelihood, $\hat{\sigma}^2$ is the estimated residual variance, n is the effective number of observations, and $k = p + q + 1$ (including variance) is the number of estimated parameters. Choose the model with the **minimum** AIC, AICC, or BIC value. BIC tends to penalize complexity more heavily and may select simpler models.

6. **Principle of Parsimony:** Favor the simplest model (fewest parameters) that provides an adequate fit to the data and passes diagnostic checks.

This process is often iterative. If initial models fail diagnostics, re-examine ACF/PACF or try slightly different orders. *(Ref: Notes Sec 18.1, 24, Book Sec 3.2, 5.5)*

26.21 Find ACVF and ACF of AR(1) process with $\phi = 0.6$. Assume the driving noise $Z_t \sim i.i.d(0, \sigma^2)$.

Answer:
The causal AR(1) process is defined by $X_t = \phi X_{t-1} + Z_t$, where $|\phi| < 1$ and $Z_t \sim WN(0, \sigma^2)$ (or i.i.d.(0, σ^2)). We are given $\phi = 0.6$.

Autocovariance Function (ACVF): The variance (ACVF at lag 0) is given by:

$$\gamma(0) = \frac{\sigma^2}{1-\phi^2} = \frac{\sigma^2}{1-(0.6)^2} = \frac{\sigma^2}{1-0.36} = \frac{\sigma^2}{0.64} = \frac{100}{64}\sigma^2 = \frac{25}{16}\sigma^2$$

The ACVF for lag h is given by:

$$\gamma(h) = \phi^{|h|}\gamma(0) = (0.6)^{|h|}\frac{25}{16}\sigma^2$$

Autocorrelation Function (ACF): The ACF is obtained by normalizing the ACVF:

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

So, for $\phi = 0.6$:

$$\rho(h) = (0.6)^{|h|}$$

Specifically:

- $\rho(0) = (0.6)^0 = 1$
- $\rho(1) = (0.6)^1 = 0.6$
- $\rho(2) = (0.6)^2 = 0.36$
- $\rho(3) = (0.6)^3 = 0.216$
- ... and $\rho(-h) = \rho(h)$.

The ACF decays exponentially as the lag $|h|$ increases. *(Ref: Book Example 1.4.5, 2.2.1)*

26.22 What is Spectral density? Why is it used?

Answer:
Spectral Density: The spectral density function (SDF), denoted $f(\lambda)$ or $S(\omega)$, describes how the total variance (or power) of a stationary time series is distributed across different frequencies (λ or ω, typically in the range $[-\pi, \pi]$ or $[0, \pi]$). Mathematically, for a process with absolutely summable ACVF $\gamma(h)$, it is the Fourier transform of the ACVF:

$$f(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma(h) e^{-ih\lambda}$$

Conversely, the ACVF can be recovered from the SDF:

$$\gamma(h) = \int_{-\pi}^{\pi} e^{ih\lambda} f(\lambda) d\lambda$$

The total variance is the integral of the SDF: $\gamma(0) = \int_{-\pi}^{\pi} f(\lambda) d\lambda$.

Why it is used: Spectral analysis (analysis in the frequency domain) provides an alternative perspective to time-domain analysis (based on ACF/PACF). It is particularly useful for:

- **Identifying Periodicity:** Peaks in the spectral density function indicate dominant frequencies, which correspond to periodic components or cycles in the time series (e.g., seasonality, business cycles). The location of the peak gives the frequency, and its height indicates the contribution of that frequency to the total variance.
- **Understanding Dynamics:** Low frequencies ($\lambda \approx 0$) correspond to slow-moving components or trends, while high frequencies ($\lambda \approx \pi$) correspond to rapid fluctuations or noise. The shape of the SDF reveals the nature of the series' fluctuations.
- **Signal Processing:** Used extensively for filtering noise, separating signals based on frequency content, and analyzing the effect of linear filters.
- **Model Characterization:** Different models (AR, MA, ARMA) have characteristic SDF shapes. For example, white noise has a flat spectrum.

(Ref: Notes Sec 4, 22, Book Sec 4.1)

26.23 Discuss properties of Sample Mean and Sample ACF in the context of stationary process. How are they used in time series model building?

Answer:
For a stationary process $\{X_t\}$ with mean μ and ACF $\rho(h)$, observed as $x_1, \ldots, x_n$:

Sample Mean ($\bar{x} = \frac{1}{n} \sum x_t$):

- **Properties:** It is an unbiased estimator of μ ($E[\bar{X}_n] = \mu$). Its variance is given by $Var(\bar{X}_n) = n^{-1} \sum_{h=-(n-1)}^{n-1} (1 - \frac{|h|}{n}) \gamma(h)$. If $\gamma(h) \to 0$ as $h \to \infty$, then $Var(\bar{X}_n) \to 0$ as $n \to \infty$ (mean square consistency). If $\sum |\gamma(h)| < \infty$, then $nVar(\bar{X}_n) \to \sum \gamma(h)$. For many stationary processes (like ARMA), $\bar{X}_n$ is asymptotically normally distributed.

- **Use in Model Building:** Used as the primary estimator for the constant mean μ of the stationary process. Often, the data is mean-centered $(x_t - \bar{x})$ before further analysis like ACF calculation or model fitting.

(Ref: Book Sec 2.4.1)

Sample ACF ($\hat{\rho}(h) = \hat{\gamma}(h)/\hat{\gamma}(0)$):

- **Properties:** It is an estimator of the true ACF $\rho(h)$. It is asymptotically unbiased. For large n, the vector $(\hat{\rho}(1), \ldots, \hat{\rho}(k))$ is approximately multivariate normal with mean $(\rho(1), \ldots, \rho(k))$ and a covariance matrix $n^{-1}W$ where W's elements w_{ij} are given by Bartlett's formula. A key property is that for large n, if the true process is white noise, then $\hat{\rho}(h) \approx N(0, 1/n)$ for $h \neq 0$, and the $\hat{\rho}(h)$ values are approximately uncorrelated. The sample ACVF matrix $[\hat{\gamma}(i-j)]$ (and ACF matrix) is non-negative definite.
- **Use in Model Building:** The sample ACF is a critical tool for:
  - *Assessing Stationarity:* Slow decay in $\hat{\rho}(h)$ suggests non-stationarity (e.g., trend). Periodic behavior in $\hat{\rho}(h)$ suggests seasonality.
  - *Testing for White Noise:* Checking if $\hat{\rho}(h)$ values fall within $\pm 1.96/\sqrt{n}$ bounds for $h \neq 0$.
  - *Model Identification (Order Selection):* The characteristic cut-off pattern for MA(q) models or the tailing-off pattern for AR(p) models in the sample ACF (used in conjunction with sample PACF) helps determine candidate model orders (p, q).
  - *Diagnostic Checking:* After fitting a model, the ACF of the residuals is checked to ensure it resembles white noise, indicating a good fit.

(Ref: Book Sec 1.4.1, 2.4.2, 1.6(a))

26.24 What is Best Linear Predictor (BLP) and discuss its properties.

Answer:
The Best Linear Predictor (BLP) of a future value X_{n+h} (or any random variable Y) based on past observations $X_1, \ldots, X_n$ is the linear function of the predictors $(1, X_1, \ldots, X_n)$ that minimizes the Mean Squared Error (MSE) of prediction.

Let $W = (X_1, \ldots, X_n)'$ (often centered, $W = (X_1 - \mu, \ldots, X_n - \mu)'$). The BLP of X_{n+h} given W is denoted $P(X_{n+h}|W)$ or $P_n X_{n+h}$. If $E[X_t] = \mu$, the BLP has the form:

$$P_n X_{n+h} = \mu + \sum_{i=1}^{n} a_i (X_{n+1-i} - \mu) = \mu + a_n'(X_n - \mu \mathbf{1})$$

where $X_n = (X_n, \ldots, X_1)'$, $\mathbf{1}$ is a vector of ones, and the coefficient vector $a_n = (a_1, \ldots, a_n)'$ is found by solving the linear equations:

$$\Gamma_n a_n = \gamma_n(h)$$

Here, $\Gamma_n = Cov(X_n, X_n) = [\gamma(i-j)]_{i,j=1}^{n}$ is the $n \times n$ covariance matrix of the predictors, and $\gamma_n(h) = Cov(X_{n+h}, X_n) = (\gamma(h), \gamma(h+1), \ldots, \gamma(h+n-1))'$ is the vector of covariances between the target variable and the predictors.

Properties of the BLP ($P_n X_{n+h}$):

1. **Minimizes MSE:** By definition, it yields the smallest $E[(X_{n+h} - L)^2]$ among all linear predictors $L = a_0 + \sum a_i X_{n+1-i}$.
2. **Unbiased Prediction Error:** The expected value of the prediction error is zero: $E[X_{n+h} - P_n X_{n+h}] = 0$.
3. **Orthogonality Principle:** The prediction error $(X_{n+h} - P_n X_{n+h})$ is uncorrelated with all the variables used in the predictor, i.e., uncorrelated with $1, X_1, \ldots, X_n$:

$$E[(X_{n+h} - P_n X_{n+h}) \times 1] = 0$$

$$E[(X_{n+h} - P_n X_{n+h}) X_j] = 0, \quad \text{for } j = 1, \ldots, n$$

   These orthogonality conditions uniquely define the BLP.

4. **Mean Squared Error Formula:** The minimum MSE achieved by the BLP is:

$$MSE = E[(X_{n+h} - P_n X_{n+h})^2] = Var(X_{n+h}) - Cov(P_n X_{n+h}, X_{n+h}) = \gamma(0) - a_n' \gamma_n(h)$$

5. **Linearity:** The prediction operator $P_n(\cdot)$ is linear: $P_n(aU + bV + c) = aP_nU + bP_nV + c$.
6. **Projection:** If Y is already a linear combination of $1, X_1, \ldots, X_n$, then $P_n Y = Y$.
7. **Known Mean:** If the mean μ is known, the predictor is $P_n X_{n+h} = \mu + a_n'(X_n - \mu \mathbf{1})$. If $\mu = 0$, $P_n X_{n+h} = a_n' X_n$.

(Ref: Notes Sec 18.2, 19.2, Book Sec 2.5)

27 Modeling and Forecasting with ARMA Processes

The determination of an appropriate $ARMA(p, q)$ model to represent an observed stationary time series involves choosing p and q (order selection) and estimating the model parameters: the mean, the coefficients $\{\phi_i, i = 1, \ldots, p\}$, $\{\theta_i, i = 1, \ldots, q\}$, and the white noise variance σ^2.

This chapter focuses on estimating the parameters $\phi = (\phi_1, \ldots, \phi_p)'$, $\theta = (\theta_1, \ldots, \theta_q)'$, and σ^2 when p and q are assumed to be known. We will assume the data $x_1, \ldots, x_n$ have been "mean-corrected" by subtracting the sample mean $\overline{y}$, allowing us to fit a zero-mean ARMA model:

$$\phi(B)X_t = \theta(B)Z_t, \quad \{Z_t\} \sim WN(0, \sigma^2)$$

27.1 Preliminary Estimation

Good parameter estimates can be found by maximizing the Gaussian likelihood (Section 27.2). This process is nonlinear and requires a numerical search, which must be initialized with starting parameter values. Preliminary estimation techniques provide these initial values.

We consider four techniques for estimating the parameters of the causal $ARMA(p, q)$ process:

$$\phi(B)X_t = \theta(B)Z_t, \quad \{Z_t\} \sim WN(0, \sigma^2)$$

The Yule-Walker and Burg procedures are for pure autoregressive (AR) models ($q = 0$). The Innovations and Hannan-Rissanen algorithms are used for models with a moving-average component ($q > 0$).

27.1.1 Yule–Walker Estimation

For a pure, causal $AR(p)$ process, $\theta(z) = 1$ and $\psi(z) = 1/\phi(z)$. Multiplying the ARMA equation by X_{t-j} for $j = 0, \ldots, p$ and taking expectations gives the **Yule-Walker equations**:

$$\Gamma_p \phi = \gamma_p \tag{1}$$
$$\sigma^2 = \gamma(0) - \phi' \gamma_p \tag{2}$$

where $\Gamma_p = [\gamma(i - j)]_{i,j=1}^p$ and $\gamma_p = (\gamma(1), \ldots, \gamma(p))'$.

By replacing the theoretical covariances $\gamma(j)$ with their sample estimates $\hat{\gamma}(j)$, we get the **Sample Yule-Walker Equations** for the estimators $\hat{\phi}$ and $\hat{\sigma}^2$:

$$\hat{\Gamma}_p \hat{\phi} = \hat{\gamma}_p \tag{3}$$
$$\hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\phi}' \hat{\gamma}_p \tag{4}$$

where $\hat{\Gamma}_p = [\hat{\gamma}(i - j)]_{i,j=1}^p$ and $\hat{\gamma}_p = (\hat{\gamma}(1), \ldots, \hat{\gamma}(p))'$. If $\hat{\gamma}(0) > 0$, these can be written using the sample autocorrelation function (ACF):

$$\hat{\phi} = (\hat{\phi}_1, \ldots, \hat{\phi}_p)' = \hat{R}_p^{-1} \hat{\rho}_p \tag{5}$$
$$\hat{\sigma}^2 = \hat{\gamma}(0)[1 - \hat{\rho}_p' \hat{R}_p^{-1} \hat{\rho}_p] \tag{6}$$

where $\hat{R}_p = \hat{\Gamma}_p / \hat{\gamma}(0)$ and $\hat{\rho}_p = (\hat{\rho}(1), \ldots, \hat{\rho}(p))'$.

An important property is that the resulting model $1 - \hat{\phi}_1 z - \cdots - \hat{\phi}_p z^p \neq 0$ for $|z| \leq 1$, meaning the **fitted model is causal**. The autocovariances of the fitted model at lags $0, \ldots, p$ are exactly equal to the sample autocovariances $\hat{\gamma}(h)$, $h = 0, \ldots, p$.

Large-Sample Distribution of Yule-Walker Estimators For a large sample from an $AR(p)$ process:

$$\hat{\phi} \approx N(\phi, n^{-1}\sigma^2\Gamma_p^{-1})$$

This result can be used to find confidence regions.

Order Selection In practice, the true order p is unknown.

- **Using the PACF:** For a true $AR(p)$ process, the partial autocorrelations ϕ_{mm} are zero for $m > p$. The sample PACF values, $\hat{\phi}_{mm}$ for $m > p$, are approximately $N(0, 1/n)$. This suggests choosing p as the smallest m such that $|\hat{\phi}_{kk}| < 1.96n^{-1/2}$ for all $k > m$.
- **Using AICC:** A more systematic approach is to find the p that minimizes the AICC statistic (see Section 27.5).

The Durbin-Levinson algorithm is used to recursively solve the Yule-Walker equations for increasing orders $m = 1, 2, \ldots$.

Confidence Regions for the Coefficients Assuming the order p is correct, an approximate $(1-\alpha)$ confidence region for ϕ_p is:

$$\{\phi \in \mathbb{R}^p : (\hat{\phi}_p - \phi)'\hat{\Gamma}_p(\hat{\phi}_p - \phi) \leq n^{-1}\hat{v}_p\chi^2_{1-\alpha}(p)\} \tag{7}$$

where $\hat{v}_p = \hat{\sigma}^2$ from (5.1.8). An approximate $(1-\alpha)$ confidence interval for a single coefficient ϕ_{pj} is bounded by:

$$\hat{\phi}_{pj} \pm \Phi_{1-\alpha/2}n^{-1/2}\hat{v}_{jj}^{1/2} \tag{8}$$

where $\hat{v}_{jj}$ is the j^{th} diagonal element of $\hat{v}_p\hat{\Gamma}_p^{-1}$.

Example 5.1.1: The Dow Jones Utilities Index (DOWJ.TSM) The sample ACF of this series is slowly decaying, suggesting non-stationarity. We first apply differencing, $Y_t = D_t - D_{t-1}$. We fit an AR process to this new series $\{Y_t\}$, $n = 77$. The sample autocovariances are $\hat{\gamma}(0) = 0.17992$, $\hat{\gamma}(1) = 0.07590$, $\hat{\gamma}(2) = 0.04885$. Applying the Durbin-Levinson algorithm:

$$\begin{aligned}
\hat{\phi}_{11} &= \hat{\rho}(1) = 0.4219 \\
\hat{v}_1 &= \hat{\gamma}(0)[1 - \hat{\rho}^2(1)] = 0.1479 \\
\hat{\phi}_{22} &= [\hat{\gamma}(2) - \hat{\phi}_{11}\hat{\gamma}(1)]/\hat{v}_1 = 0.1138 \\
\hat{\phi}_{21} &= \hat{\phi}_{11} - \hat{\phi}_{11}\hat{\phi}_{22} = 0.3739 \\
\hat{v}_2 &= \hat{v}_1[1 - \hat{\phi}_{22}^2] = 0.1460
\end{aligned}$$

The sample PACF values $\hat{\phi}_{kk}$ for $k > 1$ all lie within the bounds $\pm 1.96/\sqrt{77}$, suggesting an $AR(1)$ model is appropriate. We fit the $AR(1)$ model to the mean-corrected data $X_t = Y_t - 0.1336$. The Yule-Walker $AR(1)$ model for $\{X_t\}$ is:

$$X_t - 0.4219X_{t-1} = Z_t, \quad \{Z_t\} \sim WN(0, 0.1479) \tag{9}$$

The corresponding model for $\{Y_t\}$ is:

$$Y_t - 0.1336 - 0.4219(Y_{t-1} - 0.1336) = Z_t, \quad \{Z_t\} \sim WN(0, 0.1479) \tag{10}$$

Approximate 95% confidence bounds for ϕ_1 are:

$$0.4219 \pm \frac{(1.96)(0.1479)}{(0.17992)\sqrt{77}} = (0.2194, 0.6244)$$

Minimizing the AICC statistic over $AR(p)$ models also selects the $AR(1)$ model with AICC = 74.541.

Yule-Walker Estimation with $q > 0$; Moment Estimators The Yule-Walker equations can be extended to $ARMA(p, q)$ models, but they become nonlinear and difficult to solve. The equations are:

$$\hat{\gamma}(k) - \sum_{i=1}^{p} \phi_i \hat{\gamma}(k-i) = \sigma^2 \sum_{j=k}^{q} \theta_j \psi_{j-k}, \quad 0 \leq k \leq p+q \tag{11}$$

where $\psi(z) = \theta(z)/\phi(z)$. These are rarely used when $q > 0$ as they are computationally complex and less efficient than other methods.

Example 5.1.2: MA(1) Moment Estimators For an $MA(1)$ model, the equations become:

$$\hat{\gamma}(0) = \hat{\sigma}^2(1 + \hat{\theta}_1^2) \tag{12}$$

$$\hat{\rho}(1) = \frac{\hat{\theta}_1}{1 + \hat{\theta}_1^2} \tag{13}$$

If $|\hat{\rho}(1)| > 0.5$, no real solution exists. If $|\hat{\rho}(1)| \leq 0.5$, the invertible solution ($|\hat{\theta}_1| \leq 1$) is:

$$\hat{\theta}_1 = (1 - (1 - 4\hat{\rho}^2(1))^{1/2})/(2\hat{\rho}(1))$$

$$\hat{\sigma}^2 = \hat{\gamma}(0)/(1 + \hat{\theta}_1^2)$$

These moment estimators are known to be much less efficient (have higher variance) than maximum likelihood estimators, especially as $|\theta_1|$ approaches 1.

27.1.2 Burg's Algorithm

Burg's algorithm provides an alternative method for estimating $AR(p)$ parameters. Instead of solving the Yule-Walker equations, it estimates the PACF $\{\phi_{11}, \phi_{22}, \dots\}$ by successively minimizing the sum of squares of **forward and backward one-step prediction errors**.

Let $u_i(t)$ be the forward prediction error (estimating $x_{n+1+i-t}$ from the preceding i observations) and $v_i(t)$ be the backward prediction error (estimating x_{n+1-t} from the subsequent i observations). They satisfy the recursions:

$$u_0(t) = v_0(t) = x_{n+1-t} \tag{14}$$

$$u_i(t) = u_{i-1}(t-1) - \phi_{ii} v_{i-1}(t) \tag{15}$$

$$v_i(t) = v_{i-1}(t) - \phi_{ii} u_{i-1}(t-1) \tag{16}$$

The estimate $\phi_{11}^{(B)}$ is found by minimizing $\sigma_1^2 := \frac{1}{2(n-1)} \sum_{t=2}^{n} [u_1^2(t) + v_1^2(t)]$ with respect to ϕ_{11}. This is repeated to find $\phi_{22}^{(B)}$ by minimizing σ_2^2, and so on up to $\phi_{pp}^{(B)}$. The $AR(p)$ coefficients $\hat{\phi}_{pj}$ are then found by substituting these PACF estimates into the Durbin-Levinson recursions.

Burg's Algorithm Recursions The estimates $\phi_{ii}^{(B)}$ and variance $\sigma_i^{(B)2}$ can be calculated as follows:

$$d(1) = \sum_{t=2}^{n}(u_0^2(t-1) + v_0^2(t))$$
$$\phi_{ii}^{(B)} = \frac{2}{d(i)}\sum_{t=i+1}^{n} v_{i-1}(t)u_{i-1}(t-1)$$
$$d(i+1) = (1-\phi_{ii}^{(B)2})d(i) - v_i^2(i+1) - u_i^2(n)$$
$$\sigma_i^{(B)2} = [(1-\phi_{ii}^{(B)2})d(i)]/[2(n-i)]$$

The large-sample distribution of Burg estimators is the same as for Yule-Walker estimators, $N(\phi, n^{-1}\sigma^2\Gamma_p^{-1})$. For finite samples, Burg's algorithm often yields models with slightly smaller white noise variance estimates and higher likelihoods.

Example 5.1.3: The Dow Jones Utilities Index Applying Burg's algorithm to the same differenced, mean-corrected series as in Example 5.1.1, the minimum AICC model is also $AR(1)$:

$$X_t - 0.4371X_{t-1} = Z_t, \quad \{Z_t\} \sim WN(0, 0.1423) \tag{17}$$

This model has $AICC = 74.492$, which is slightly smaller (better) than the Yule-Walker AICC of 74.541. The 95% confidence bounds are $(0.2354, 0.6388)$.

Example 5.1.4: The lake data This series $\{Y_t, t = 1, \ldots, 98\}$ has a sample PACF that strongly suggests an $AR(2)$ model. Fitting an $AR(2)$ model to the mean-corrected data $X_t = Y_t - 9.0041$:

- **Burg's Method:**

  $$X_t - 1.0449X_{t-1} + 0.2456X_{t-2} = Z_t, \quad \{Z_t\} \sim WN(0, 0.4706)$$

  AICC = 213.55.

- **Yule-Walker Method:**

  $$X_t - 1.0538X_{t-1} + 0.2668X_{t-2} = Z_t, \quad \{Z_t\} \sim WN(0, 0.4920)$$

  AICC = 213.57.

As before, the Burg model has a slightly smaller variance and better AICC value. Minimizing AICC also selects $p = 2$ for both methods.

27.1.3 The Innovations Algorithm

This algorithm (from Section 2.5.2) can be used to fit pure $MA(m)$ models by applying it to the sample ACVF.

Definition 5.1.2: Fitted Innovations MA(m) Model The model is

$$X_t = Z_t + \hat{\theta}_{m1}Z_{t-1} + \cdots + \hat{\theta}_{mm}Z_{t-m}, \quad \{Z_t\} \sim WN(0, \hat{v}_m)$$

where $\hat{\theta}_m = (\hat{\theta}_{m1}, \ldots, \hat{\theta}_{mm})'$ and $\hat{v}_m$ are obtained from the innovations algorithm using the sample ACVF.

Remark 1 (Large-Sample Properties) If $\{X_t\}$ is an invertible $MA(q)$ process, and we fit an $MA(m)$ model where $m \to \infty$ but $n^{-1/3}m \to 0$ as $n \to \infty$, then for a fixed k:

$$n^{1/2}(\hat{\theta}_{m1} - \theta_1, \ldots, \hat{\theta}_{mk} - \theta_k)' \to N(0, A)$$

where $A = [a_{ij}]_{i,j=1}^k$ and

$$a_{ij} = \sum_{r=1}^{\min(i,j)} \theta_{i-r}\theta_{j-r} \quad (\text{with } \theta_0 := 1) \tag{18}$$

Remark 2 (Consistency) Unlike the AR case, the estimator $\hat{\theta}_q = (\hat{\theta}_{q1}, \ldots, \hat{\theta}_{qq})'$ is *not* consistent for $(\theta_1, \ldots, \theta_q)'$. For consistency, one must use $\hat{\theta}_m = (\hat{\theta}_{m1}, \ldots, \hat{\theta}_{mq})'$ where m is large (as in Remark 1). In practice, we increase m until the estimates for the first q coefficients stabilize.

Order Selection

- **Using the ACF:** For a true $MA(q)$ process, $\rho(m) = 0$ for $m > q$. The sample ACF $\hat{\rho}(m)$ for $m > q$ is approximately $N(0, n^{-1}[1 + 2\sum_{j=1}^q \rho^2(j)])$. We can estimate q as the smallest m such that $\hat{\rho}(k)$ is not significant for all $k > m$.
- **Using Coefficients:** We can fit a high-order $MA(m)$ model and inspect the coefficients $\hat{\theta}_{mj}$. We choose q as the largest lag j for which the coefficient is significant (e.g., its ratio to its standard error is > 1).
- **Using AICC:** We can find the q that minimizes the AICC statistic.

Confidence Regions for the Coefficients From Remark 1, approximate 95% confidence bounds for θ_j are:

$$\hat{\theta}_{mj} \pm 1.96 n^{-1/2} \left(\sum_{i=0}^{j-1} \hat{\theta}_{mi}^2 \right)^{1/2} \tag{19}$$

Example 5.1.5: The Dow Jones Utilities Index The sample ACF of the differenced data (Example 5.1.1) suggests an $MA(2)$ model might be a good fit. We fit an $MA(2)$ model using the innovations algorithm (with $m = 17$) to the mean-corrected differenced series X_t. The fitted model is:

$$X_t = Z_t + 0.4269 Z_{t-1} + 0.2704 Z_{t-2}, \quad \{Z_t\} \sim WN(0, 0.1470)$$

The AICC value is 77.467. A check of the $MA(17)$ coefficients shows that only the first two are significant, reinforcing the choice of $q = 2$. This $MA(2)$ model is also the minimum AICC model among all $MA(q)$ models.

Innovations Algorithm Estimates when $p > 0$ and $q > 0$ For a mixed $ARMA(p, q)$ model, we can use the innovations algorithm to get preliminary estimates. A causal ARMA process can be written as $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$. The coefficients ψ_j are estimated by $\hat{\theta}_{mj}$ from a high-order $MA(m)$ fit (where $m \geq p + q$). The parameters ϕ and θ are related to ψ by:

$$\psi_j = \theta_j + \sum_{i=1}^{\min(j,p)} \phi_i \psi_{j-i}, \quad j = 0, 1, \ldots \tag{20}$$

We replace ψ_j with $\hat{\theta}_{mj}$ to get a system of equations. We first solve for $\hat{\phi}$ using the equations for $j = q+1, \ldots, q+p$:

$$\begin{bmatrix} \hat{\theta}_{m,q} & \hat{\theta}_{m,q-1} & \cdots & \hat{\theta}_{m,q+1-p} \\ \hat{\theta}_{m,q+1} & \hat{\theta}_{m,q} & \cdots & \hat{\theta}_{m,q+2-p} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\theta}_{m,q+p-1} & \hat{\theta}_{m,q+p-2} & \cdots & \hat{\theta}_{m,q} \end{bmatrix} \begin{bmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \\ \vdots \\ \hat{\phi}_p \end{bmatrix} = \begin{bmatrix} \hat{\theta}_{m,q+1} \\ \hat{\theta}_{m,q+2} \\ \vdots \\ \hat{\theta}_{m,q+p} \end{bmatrix} \tag{21}$$

Once $\hat{\phi}$ is found, we find $\hat{\theta}$ from:

$$\hat{\theta}_j = \hat{\theta}_{mj} - \sum_{i=1}^{\min(j,p)} \hat{\phi}_i \hat{\theta}_{m,j-i}, \quad j = 1, \ldots, q$$

This method may result in a non-causal model, which cannot be used to initialize the likelihood maximization.

Example 5.1.6: The lake data We fit an $ARMA(1,1)$ model to the mean-corrected lake data $X_t = Y_t - 9.0041$ using the innovations method (with $m = 17$).

$$X_t - 0.7234 X_{t-1} = Z_t + 0.3596 Z_{t-1}, \quad \{Z_t\} \sim WN(0, 0.4757)$$

The AICC value is 212.89, which is smaller (better) than the $AR(2)$ models from Example 5.1.4. This suggests an $ARMA(1,1)$ model may be superior.

27.1.4 The Hannan–Rissanen Algorithm

This algorithm is based on the idea of least squares regression. The $ARMA(p,q)$ equation

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + Z_t + \sum_{j=1}^{q} \theta_j Z_{t-j}$$

looks like a regression of X_t on past X_t and past Z_t. The Z_t terms are unobserved, so the algorithm estimates them.

Step 1. A high-order $AR(m)$ model (with $m > \max(p,q)$) is fitted to the data (e.g., using Yule-Walker). The residuals are computed:

$$\hat{Z}_t = X_t - \hat{\phi}_{m1} X_{t-1} - \cdots - \hat{\phi}_{mm} X_{t-m}, \quad t = m+1, \ldots, n$$

Step 2. The parameters $\beta = (\phi', \theta')'$ are estimated by a least squares regression of X_t onto $(X_{t-1}, \ldots, X_{t-p}, \hat{Z}_{t-1}, \ldots, \hat{Z}_{t-q})$ for $t = m+1+q, \ldots, n$. This involves minimizing:

$$S(\beta) = \sum_{t=m+1+q}^{n} \left(X_t - \sum_{i=1}^{p} \phi_i X_{t-i} - \sum_{j=1}^{q} \theta_j \hat{Z}_{t-j} \right)^2$$

The white noise variance is estimated as $\hat{\sigma}^2_{HR} = S(\hat{\beta})/(n-m-q)$.

Step 3. (Optional) The estimates are refined in a third step to improve efficiency to match the maximum likelihood estimator. (We omit this and use the Step 2 estimates as preliminary values for MLE).

Example 5.1.7: The lake data Fitting an $ARMA(1,1)$ model to the mean-corrected lake data using the Hannan-Rissanen algorithm:

$$X_t - 0.6961X_{t-1} = Z_t + 0.3788Z_{t-1}, \quad \{Z_t\} \sim WN(0, 0.4774)$$

The AICC is 213.18, which is very close to the innovations result (212.89) from Example 5.1.6.

27.2 Maximum Likelihood Estimation

If we assume the time series $\{X_t\}$ is Gaussian with mean zero and ACVF $\kappa(i,j)$, the likelihood of the observation vector $X_n = (X_1, \ldots, X_n)'$ is:

$$L(\Gamma_n) = (2\pi)^{-n/2}(\det \Gamma_n)^{-1/2} \exp\left(-\frac{1}{2}X_n'\Gamma_n^{-1}X_n\right) \tag{22}$$

where $\Gamma_n = E(X_n X_n')$. This calculation can be simplified using the one-step prediction errors $X_j - \hat{X}_j$ and their variances $v_{j-1} = E(X_j - \hat{X}_j)^2$, which are found from the innovations algorithm. The likelihood (5.2.1) is algebraically equivalent to:

$$L(\Gamma_n) = \frac{1}{\sqrt{(2\pi)^n v_0 \ldots v_{n-1}}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{n}\frac{(X_j - \hat{X}_j)^2}{v_{j-1}}\right\} \tag{23}$$

The "maximum likelihood estimators" (MLEs) are the parameter values that maximize this function, even if the process is not truly Gaussian.

The Gaussian Likelihood for an ARMA Process For an $ARMA(p,q)$ process, $v_{j-1} = \sigma^2 r_{j-1}$, where r_{j-1} is computed from the innovations algorithm using the ϕ and θ parameters (with $\sigma^2 = 1$). The likelihood becomes:

$$L(\phi, \theta, \sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)^n r_0 \ldots r_{n-1}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}\frac{(X_j - \hat{X}_j)^2}{r_{j-1}}\right\} \tag{24}$$

Maximum Likelihood Estimators The values $\hat{\phi}$, $\hat{\theta}$, and $\hat{\sigma}^2$ that maximize L are found by first minimizing $l(\phi, \theta)$ with respect to ϕ and θ, and then computing $\hat{\sigma}^2$.

- $\hat{\phi}, \hat{\theta}$ are the values of ϕ, θ that minimize:

$$l(\phi, \theta) = \ln(n^{-1}S(\phi, \theta)) + n^{-1}\sum_{j=1}^{n}\ln r_{j-1} \tag{25}$$

- The MLE for σ^2 is:

$$\hat{\sigma}^2 = n^{-1}S(\hat{\phi}, \hat{\theta}) \tag{26}$$

- where $S(\phi, \theta)$ is the sum of squares:

$$S(\phi, \theta) = \sum_{j=1}^{n}(X_j - \hat{X}_j)^2/r_{j-1} \tag{27}$$

Minimization of $l(\phi, \theta)$ must be done numerically, using the preliminary estimates from Section 27.1 as starting values.

Least Squares Estimation This is an alternative, where one minimizes $S(\phi, \theta)$ directly, rather than $l(\phi, \theta)$. The corresponding variance estimate is $\tilde{\sigma}^2 = S(\tilde{\phi}, \tilde{\theta})/(n - p - q)$.

Large-Sample Distribution of Maximum Likelihood Estimators For a large sample from an $ARMA(p, q)$ process, the estimator $\hat{\beta} = (\hat{\phi}', \hat{\theta}')'$ is approximately normal:

$$\hat{\beta} \approx N(\beta, n^{-1}V(\beta))$$

The matrix $V(\beta)$ can be computed.

Example 5.2.1: $AR(p)$ **model** The asymptotic covariance matrix is $V(\phi) = \sigma^2\Gamma_p^{-1}$ (same as Yule-Walker).

- $AR(1) : V(\phi) = (1 - \phi_1^2)$
- $AR(2) : V(\phi) = \begin{bmatrix} 1 - \phi_2^2 & -\phi_1(1 + \phi_2) \\ -\phi_1(1 + \phi_2) & 1 - \phi_2^2 \end{bmatrix}$

Example 5.2.2: $MA(q)$ **model** Let Γ_q^* be the covariance matrix of an $AR(q)$ process with polynomial $\theta(z)$. Then $V(\theta) = \Gamma_q^{*-1}$.

- $MA(1) : V(\theta) = (1 - \theta_1^2)$
- $MA(2) : V(\theta) = \begin{bmatrix} 1 - \theta_2^2 & \theta_1(1 - \theta_2) \\ \theta_1(1 - \theta_2) & 1 - \theta_2^2 \end{bmatrix}$

Example 5.2.3: $ARMA(1, 1)$ **model**

$$V(\phi, \theta) = \frac{1 + \phi\theta}{(\phi + \theta)^2} \begin{bmatrix} (1 - \phi^2)(1 + \phi\theta) & -(1 - \phi^2)(1 - \theta^2) \\ -(1 - \phi^2)(1 - \theta^2) & (1 - \theta^2)(1 + \phi\theta) \end{bmatrix}$$

Example 5.2.4: The Dow Jones Utilities Index Starting from the Burg $AR(1)$ preliminary estimate (Example 5.1.3), we find the maximum likelihood $AR(1)$ model for the differenced, mean-corrected series:

$$Y_t - 0.4471Y_{t-1} = Z_t, \quad \{Z_t\} \sim WN(0, 0.02117) \tag{28}$$

This model has $-2\ln(L) = 70.321$, slightly better than the preliminary models. The AICC value is 74.483. Using the 'Autofit' procedure confirms this $AR(1)$ model is the minimum-AICC model.

Example 5.2.5: The lake data Using the 'Autofit' procedure on the mean-corrected lake data $X_t = Y_t - 9.0041$, the minimum-AICC model is found to be the $ARMA(1, 1)$ model:

$$X_t - 0.7446X_{t-1} = Z_t + 0.3213Z_{t-1}, \quad \{Z_t\} \sim WN(0, 0.4750) \tag{29}$$

The AICC value is 212.77, which is an improvement on all the preliminary models. The 95% confidence bounds are:

- $\phi : 0.7446 \pm 1.96 \times 0.0773 = (0.5941, 0.8961)$
- $\theta : 0.3208 \pm 1.96 \times 0.1123 = (0.1007, 0.5409)$

27.3 Diagnostic Checking

After fitting a model, we must check its goodness of fit by analyzing the **residuals**. The residuals are defined as:

$$\hat{W}_t = (X_t - \hat{X}_t(\hat{\phi}, \hat{\theta}))/(r_{t-1}(\hat{\phi}, \hat{\theta}))^{1/2}, \quad t = 1, \ldots, n \tag{30}$$

where $\hat{X}_t$ and r_{t-1} are computed from the fitted MLE parameters $\hat{\phi}$ and $\hat{\theta}$. The **rescaled residuals** are $R_t = \hat{W}_t/\hat{\sigma}$, where $\hat{\sigma}$ is the MLE of σ.

If the fitted model is appropriate, the rescaled residuals $\{R_t\}$ should resemble a white noise sequence $WN(0, 1)$. If we assume the underlying process noise $\{Z_t\}$ is $IID(0, \sigma^2)$, then $\{R_t\}$ should resemble an $IID(0, 1)$ sequence.

27.3.1 The Graph of $\{\hat{R}_t, t = 1, \ldots, n\}$

The plot of the rescaled residuals should look like a white noise sequence. We look for any deviations from a zero mean (e.g., trends) or non-constant variance. For the $ARMA(1, 1)$ model fitted to the lake data (Example 5.2.5), the residual graph gives no indication of a nonzero mean or nonconstant variance.

27.3.2 The Sample ACF of the Residuals

If the residuals are consistent with $IID(0, 1)$ noise, their sample autocorrelations $\hat{\rho}(h)$ for $h > 0$ should be small. For large n, the sample ACFs of an IID sequence are approximately $N(0, 1/n)$. We check if the residual ACFs fall within the bounds $\pm 1.96/\sqrt{n}$. If more than a few (e.g., 2 or 3 out of 40) fall outside, or one falls far outside, we question the model's fit. For the $ARMA(1, 1)$ lake data model, the residual ACFs are all well within these bounds, supporting the model.

27.3.3 Tests for Randomness of the Residuals

We apply the statistical tests from Section 1.6 to the residuals. For the $ARMA(1, 1)$ lake data residuals:

- Ljung-Box Portmanteau: $p = 0.964$
- McLeod-Li Portmanteau: $p = 0.788$
- Turning Points: $p = 0.227$
- Difference-Sign: $p = 0.602$
- Rank Test: $p = 0.072$
- Jarque-Bera (Normality): $p = 0.867$

Since all p-values are greater than 0.05, none of these tests give us reason to reject the hypothesis that the residuals are IID. The model passes all diagnostic checks.

27.4 Forecasting

Once a model has been fitted and checked, it can be used for forecasting using the methods from Section 3.3. We treat the fitted maximum likelihood model as the "true" model.

Example 5.4.1: Overshort data In Example 3.2.8, we had overshort data. The maximum likelihood $MA(1)$ model for this data (including the mean) is:

$$X_t + 4.035 = Z_t - 0.818Z_{t-1}, \quad \{Z_t\} \sim WN(0, 2040.75) \tag{31}$$

We have $n = 57$ observations. The h-step-ahead predictors $P_{57}X_{57+h}$ are:

$$P_{57}X_{57+h} = \begin{cases} -4.035 + \theta_{57,1}(X_{57} - \hat{X}_{57}), & \text{if } h = 1 \\ -4.035, & \text{if } h > 1 \end{cases}$$

The mean squared error $E(X_{57+h} - P_{57}X_{57+h})^2$ is:

$$\text{MSE} = \begin{cases} 2040.75 \cdot r_{57}, & \text{if } h = 1 \\ 2040.75 \cdot (1 + (-0.818)^2), & \text{if } h > 1 \end{cases}$$

The forecasts are shown in Table 5.1. Note that for $h > 1$, the forecast is just the estimated mean (-4.035), as an $MA(1)$ process has a memory of only one time step.

Table 5: Forecasts for the overshort data using the $MA(1)$ model.

#	XHAT	SQRT(MSE)	XHAT + MEAN
58	1.0097	45.1753	-3.0254
59	0.0000	58.3602	-4.0351
60	0.0000	58.3602	-4.0351
61	0.0000	58.3602	-4.0351
62	0.0000	58.3602	-4.0351
63	0.0000	58.3602	-4.0351
64	0.0000	58.3602	-4.0351

Note on Forecast MSE The MSE computed above assumes the fitted model parameters are the true parameters. It does not account for the variability in our *estimation* of the parameters. For example, if we fit an $AR(1)$ model $X_t = \phi X_{t-1} + Z_t$ and forecast X_{n+1} using $\hat{\phi}X_n$, the true MSE is:

$$\begin{aligned} E(X_{n+1} - \hat{\phi}X_n)^2 &= E((\phi - \hat{\phi})X_n + Z_{n+1})^2 \\ &= E((\phi - \hat{\phi})X_n)^2 + \sigma^2 \end{aligned} \tag{32}$$

Using large-sample approximations, $E(\phi - \hat{\phi})^2 \approx (1 - \phi^2)/n$. Since $E(X_n^2) = \sigma^2/(1 - \phi^2)$, the MSE is:

$$E(X_{n+1} - \hat{\phi}X_n)^2 \approx \frac{(1 - \phi^2)}{n}\frac{\sigma^2}{(1 - \phi^2)} + \sigma^2 = \frac{\sigma^2}{n} + \sigma^2 = \frac{n+1}{n}\sigma^2$$

The standard MSE calculation gives σ^2. The extra term σ^2/n is the contribution from parameter estimation error. This is negligible for large n, but can be significant for small samples.

27.5 Order Selection

Choosing p and q too large (overfitting) leads to poor forecasts, as the MSE will be inflated by parameter estimation errors. We need a "penalty factor" to discourage overly complex models.

27.5.1 The FPE Criterion

Akaike's Final Prediction Error (FPE) criterion is designed to select the order p of an $AR(p)$ model. The goal is to minimize the one-step MSE when the fitted model is used to predict an *independent* realization $\{Y_t\}$ of the same process. This true MSE is approximately:

$$E(Y_{n+1} - \hat{\phi}_1 Y_n - \cdots - \hat{\phi}_p Y_{n+1-p})^2 \approx \sigma^2 \left(1 + \frac{p}{n}\right) \tag{33}$$

Replacing σ^2 with an unbiased estimator $n\hat{\sigma}^2/(n-p)$ (where $\hat{\sigma}^2$ is the MLE), we get the FPE criterion. We choose p to minimize:

$$FPE_p = \hat{\sigma}^2 \frac{n+p}{n-p} \tag{34}$$

Example 5.5.1: FPE for the lake data We fit $AR(p)$ models of increasing order p to the mean-corrected lake data ($n = 98$) and compute FPE_p.

Table 6: FPE_p for $AR(p)$ models fitted to the lake data.

p	$\hat{\sigma}_p^2$	FPE_p
0	1.7203	1.7203
1	0.5097	0.5202
2	**0.4790**	**0.4989**
3	0.4728	0.5027
4	0.4708	0.5109
5	0.4705	0.5211
...	...	...
10	0.4453	0.5465

The FPE_p value is clearly minimized at $p = 2$, confirming our earlier choice based on the PACF.

27.5.2 The AICC Criterion

A more general criterion is Akaike's Information Criterion (AIC), and its bias-corrected version, AICC. These are based on estimating the Kullback-Leibler discrepancy, a measure of distance between the fitted model and the unknown true model. The AICC criterion is to choose p and q to minimize:

$$AICC := -2 \ln L(\hat{\phi}, \hat{\theta}, \hat{\sigma}^2) + \frac{2(p+q+1)n}{n-p-q-2} \tag{35}$$

where $\hat{\sigma}^2 = S(\hat{\phi}, \hat{\theta})/n$. The first term measures goodness of fit (a smaller $-2 \ln L$ is better), and the second term is a **penalty** for model complexity (it increases with p and q). The AICC has a more extreme penalty for large-order models than the original AIC, which helps prevent overfitting in small samples.

The original **AIC** is:

$$AIC := -2 \ln L(\hat{\phi}, \hat{\theta}, \hat{\sigma}^2) + 2(p+q+1)$$

The **BIC** (Bayesian Information Criterion) is another alternative:

$$BIC = (n-p-q)\ln[n\hat{\sigma}^2/(n-p-q)] + n(1+\ln\sqrt{2\pi}) + (p+q)\ln[(\sum X_t^2 - n\hat{\sigma}^2)/(p+q)] \quad (36)$$

BIC is a *consistent* order selector (i.e., $\hat{p} \to p$ and $\hat{q} \to q$ with probability 1 if a true $ARMA(p,q)$ model exists), whereas AICC/AIC are *asymptotically efficient* (better for prediction). In practice, there is rarely a "true order," and AICC provides a rational criterion for choosing among competing models.

If a model has m coefficients constrained to be zero, the AICC formula is adjusted:

$$AICC := -2\ln L(\hat{\beta}, \hat{\sigma}^2) + \frac{2(m+1)n}{n-m-2} \quad (37)$$

Example 5.5.2: Models for the lake data We compare the best models found for the mean-corrected lake data:

- **ARMA(1,1) model (MLE):** (from Example 5.2.5)
  - $AICC = 212.77$
  - $BIC = 216.86$
- **AR(2) model (MLE):**
  - $AICC = 213.54$
  - $BIC = 217.63$

Both the $ARMA(1,1)$ and $AR(2)$ models pass diagnostic checks. The $ARMA(1,1)$ model is preferred as it minimizes both AICC and BIC. However, the AICC values are very close, so there is no strong reason to discard the $AR(2)$ model.

28 Non-stationary and Seasonal Time Series Models

This chapter addresses modeling time series data $\{x_1, \ldots, x_n\}$ that may not be stationary. If data appear stationary and have a rapidly decaying ACVF, we use ARMA models (Chapter 5). Otherwise, we first seek a transformation (like differencing) to achieve stationarity, leading to ARIMA models.

28.1 ARIMA Models for Nonstationary Time Series

Many nonstationary series can be made stationary by differencing. This leads to the class of Autoregressive Integrated Moving Average (ARIMA) models.

Definition 6.1.1: ARIMA(p, d, q) Process If d is a non-negative integer, then $\{X_t\}$ is an ARIMA(p, d, q) process if $Y_t := (1-B)^d X_t$ is a causal $ARMA(p, q)$ process.

This means $\{X_t\}$ satisfies a difference equation:

$$\phi^*(B)X_t \equiv \phi(B)(1-B)^d X_t = \theta(B)Z_t, \quad \{Z_t\} \sim WN(0, \sigma^2) \tag{38}$$

where $\phi(z)$ and $\theta(z)$ are polynomials of degree p and q respectively, $\phi(z) \neq 0$ for $|z| \leq 1$. The polynomial $\phi^*(z)$ has a zero of order d at $z = 1$.

- The process is stationary if and only if $d = 0$.
- If $d \geq 1$, an arbitrary polynomial trend of degree $(d-1)$ can be added to X_t without violating the equation. ARIMA models are useful for data with trends.
- Even without trends, ARIMA models can be appropriate (e.g., random walks).
- Estimation for ARIMA models is based on the differenced series $Y_t = (1-B)^d X_t$.

Example 6.1.1: ARIMA(1,1,0) Let $(1-\phi B)(1-B)X_t = Z_t$, where $|\phi| < 1$ and $\{Z_t\} \sim WN(0, \sigma^2)$. Then $X_t = X_0 + \sum_{j=1}^{t} Y_j$, where $Y_t = (1-B)X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$ is an $AR(1)$ process.

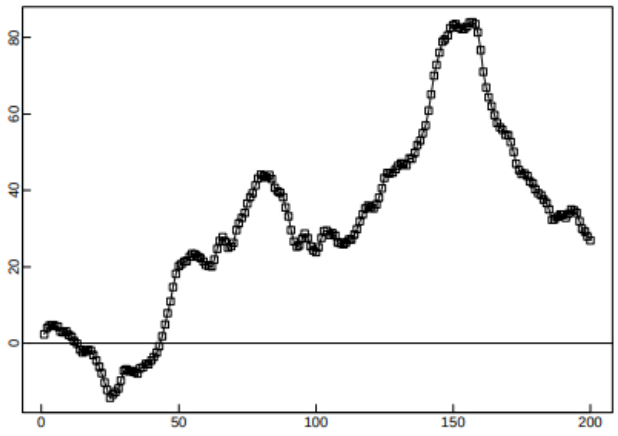

Figure 14: 200 observations of the ARIMA(1,1,0) series X_t with $\phi = 0.8, \sigma^2 = 1$.

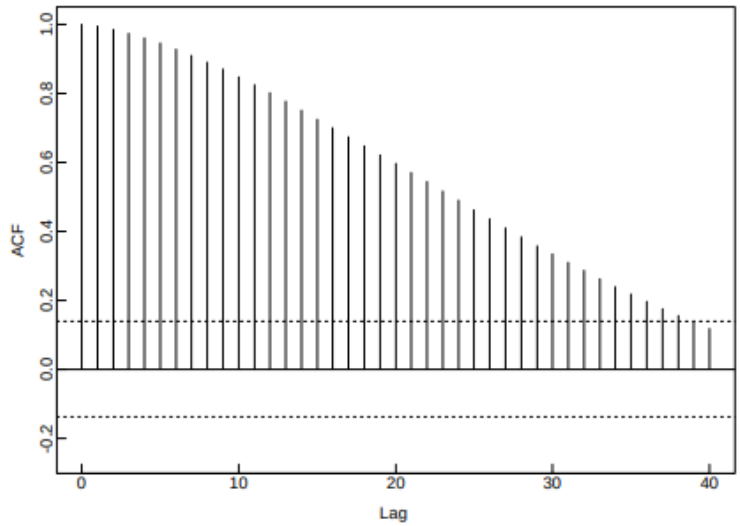

Figure 15: Sample ACF of the data in Figure 14.

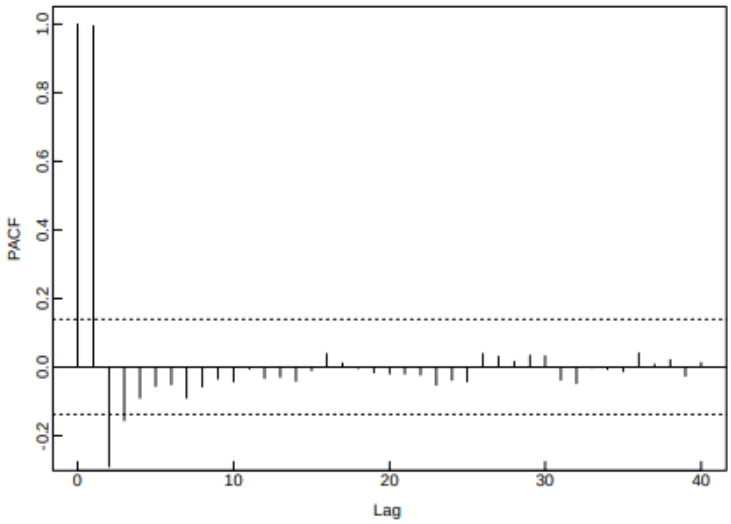

Figure 16: Sample PACF of the data in Figure 14.

The slowly decaying positive sample ACF (Figure 15) suggests non-stationarity and the need for differencing. Applying $\nabla = 1 - B$ once gives the series $Y_t = \nabla X_t$.

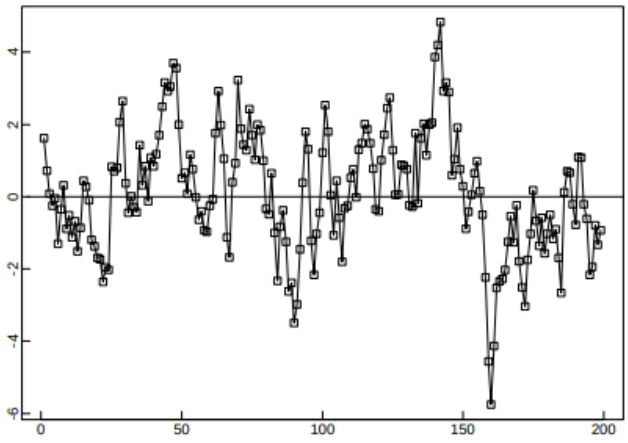

Figure 17: 199 observations of the differenced series $Y_t = \nabla X_t$.

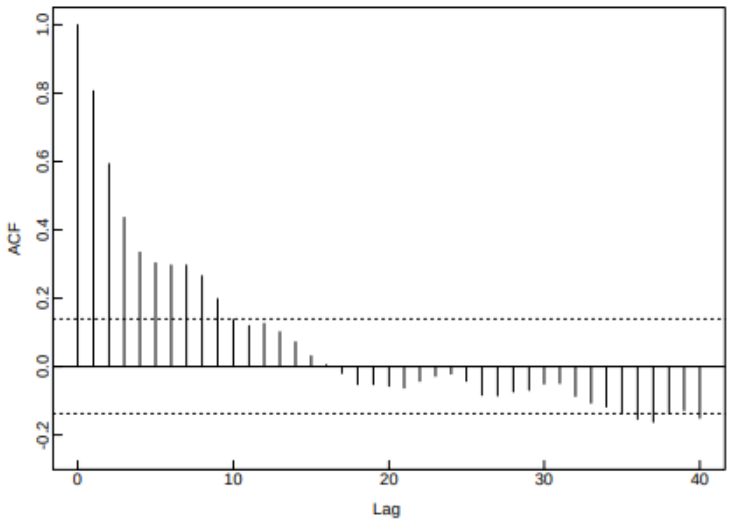

Figure 18: Sample ACF of the differenced series $\{Y_t\}$ in Figure 17.

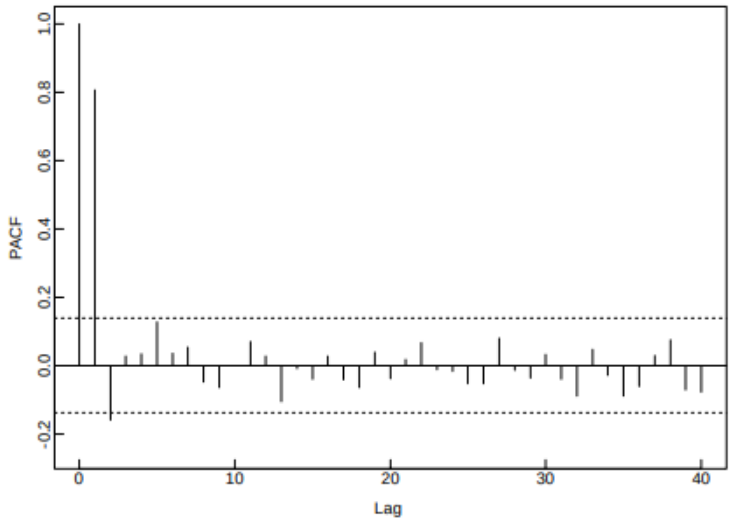

Figure 19: Sample PACF of the differenced series $\{Y_t\}$ in Figure 17.

The ACF and PACF of $\{Y_t\}$ (Figures 18 and 19) suggest an $AR(1)$ model for the differenced series. Maximum likelihood estimation gives:

$$(1 - 0.808B)(1 - B)X_t = Z_t, \quad \{Z_t\} \sim WN(0, 0.978)$$

which is close to the true process $(1 - 0.8B)(1 - B)X_t = Z_t, \{Z_t\} \sim WN(0, 1)$.

Fitting an $AR(2)$ directly to the original data (suggested by Figure 16) gives:

$$(1 - 1.808B + 0.811B^2)X_t = (1 - 0.825B)(1 - 0.983B)X_t = Z_t, \quad \{Z_t\} \sim WN(0, 0.970)$$

This is stationary but has a root very close to 1 (0.983 vs 1.0). Distinguishing between a unit root and a root close to 1 is difficult with finite data. Differencing often leads to a model with roots comfortably outside the unit circle, which is better for estimation.

Other Roots Near the Unit Circle Slowly decaying oscillatory sample ACFs suggest roots near $e^{i\omega}$ for $\omega \neq 0$.

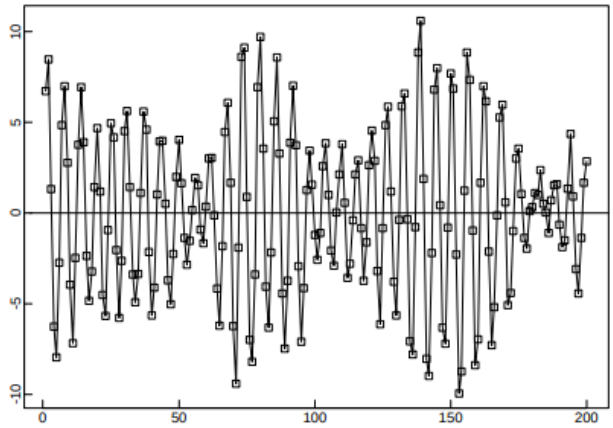

Figure 20: 200 observations of the $AR(2)$ process $X_t - 0.9950X_{t-1} + 0.9901X_{t-2} = Z_t$, with roots near $e^{\pm i\pi/3}$.

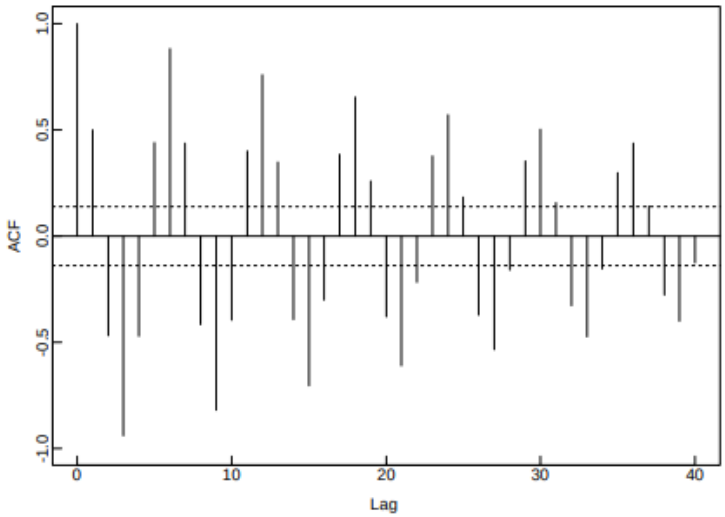

Figure 21: Sample ACF of the data in Figure 20.

The ACF of an $AR(2)$ process $X_t - (2r^{-1}\cos\omega)X_{t-1} + r^{-2}X_{t-2} = Z_t$ is:

$$\rho(h) = r^{-h}\frac{\sin(h\omega + \psi)}{\sin\psi}, \quad h \geq 0 \tag{39}$$

where $\tan\psi = \frac{r^2+1}{r^2-1}\tan\omega$. As $r \downarrow 1$, $\rho(h) \to \cos(h\omega)$. The sample ACF in Figure 21 shows slow sinusoidal decay with period 6 ($\omega = \pi/3$). This suggests applying an operator like $(1 - 2\cos(\pi/3)B + B^2) = (1 - B + B^2)$ or, since the period is close to $s = 6$, the operator $(1 - B^6)$.

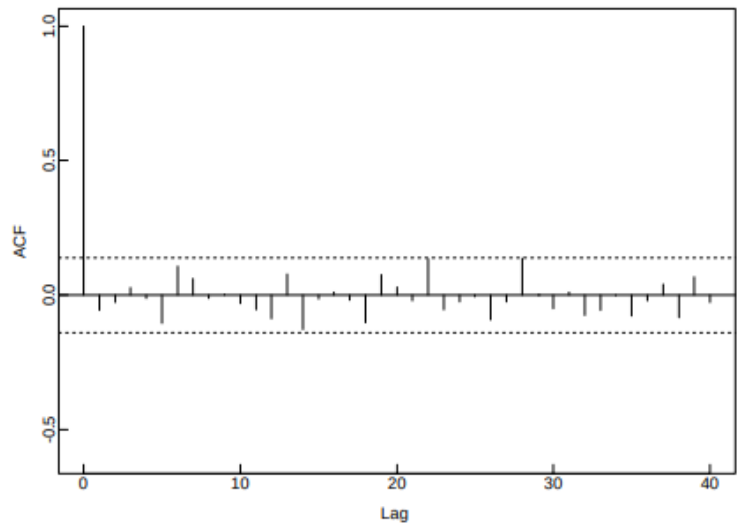

Figure 22: Sample ACF of $(1 - B + B^2)X_t$ with $\{X_t\}$ as in Figure 20.

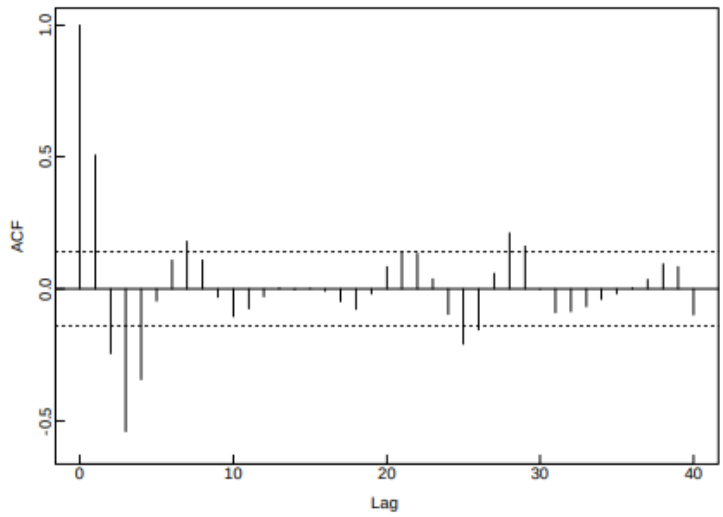

Figure 23: Sample ACF of $(1 - B^6)X_t$ with $\{X_t\}$ as in Figure 20.

Both operators (Figures 22 and 23) produce series whose ACFs decay much faster, allowing an ARMA model to be fitted.

28.2 Identification Techniques

The process of finding an appropriate model involves several steps.

(a) Preliminary Transformations The goal is to transform the data so that fitting a stationary ARMA model is plausible.

- **Variance Stabilization:** If the variability of the series depends on its level (e.g., Figure 1.1), apply a transformation like the logarithm or a Box-Cox transformation $f_\lambda(U)$.

$$f_\lambda(U) = \begin{cases} (U^\lambda - 1)/\lambda, & U_t \geq 0, \lambda > 0 \\ \ln U, & U_t > 0, \lambda = 0 \end{cases}$$

  Often $\lambda = 0$ (log) or $\lambda = 0.5$ (sqrt) is sufficient.

- **Trend and Seasonality Removal:** Detected visually or by slowly decaying/periodic ACFs. Methods include:

  i. *Classical Decomposition:* Estimate and subtract trend (m_t) and seasonal (s_t) components: $Y_t = X_t - m_t - s_t$.

  ii. *Differencing:* Apply operators like $\nabla = (1 - B)$ for trend or $\nabla_s = (1 - B^s)$ for seasonality.

  iii. *Regression:* Fit polynomial trends and/or harmonic terms for seasonality, and model the residuals $W_t = Y_t - x_t'\beta$. (See Section 28.6)

  Applying methods (i) and (ii) to the log-transformed Australian red wine data (Figure 1.17) produces series shown in Figures 6.11 and 6.12.

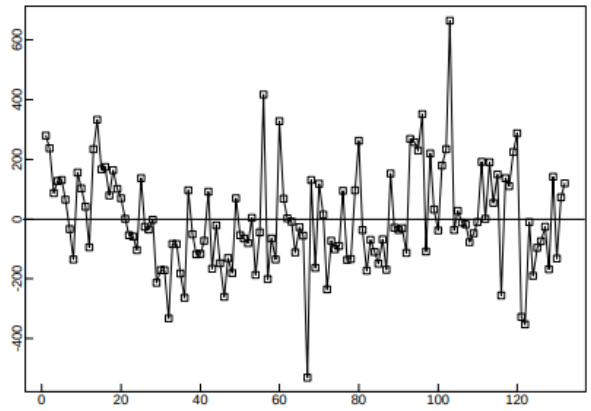

Figure 24: Australian red wine data: log-transformed, then seasonal (period 12) and linear trend components removed via classical decomposition.

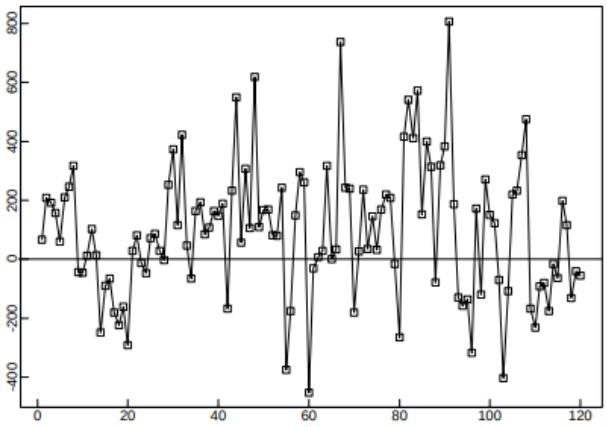

Figure 25: Australian red wine data: log-transformed, then differenced at lag 12, $(1 - B^{12}) \ln(U_t)$.

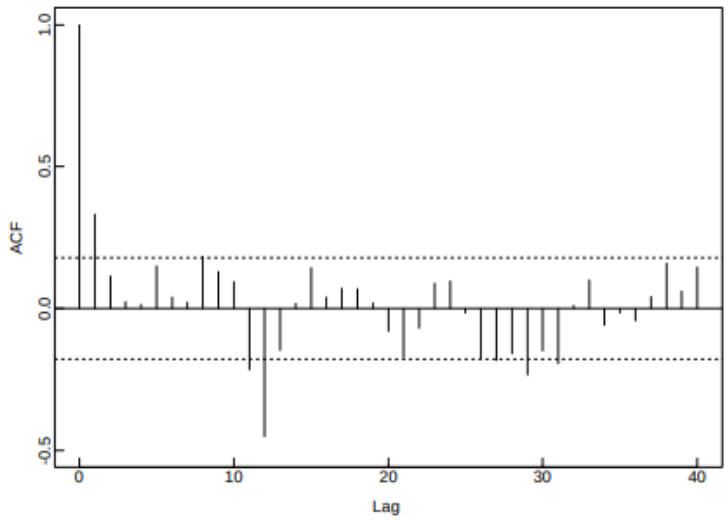

Figure 26: Sample ACF of the data in Figure 25.

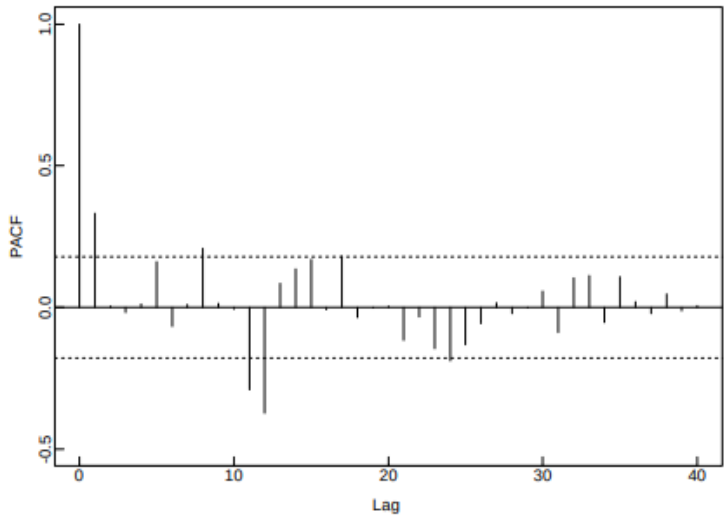

Figure 27: Sample PACF of the data in Figure 25.

Both resulting series appear stationary. Further differencing (e.g., ∇) might be needed if the ACF still decays slowly.

(b) Identification and Estimation Let $\{X_t\}$ be the mean-corrected, transformed series. We need to find the best $ARMA(p, q)$ model.

- **Examine ACF/PACF:** Look for cutoffs (MA or AR models) or decay patterns. Preliminary estimation (Yule-Walker, Burg, Innovations, Hannan-Rissanen) can also guide the choice.
- **Minimize AICC:** Use maximum likelihood estimation for various (p, q) pairs and choose the model that minimizes:

$$AICC(\phi, \theta) = -2 \ln L(\phi, \theta, S(\phi, \theta)/n) + \frac{2(p + q + 1)n}{n - p - q - 2} \tag{40}$$

  The 'Autofit' procedure automates this search.
- **Subset Models:** If some estimated coefficients are not significant (small compared to their standard errors), consider fitting a subset model where these coefficients are constrained to zero. Re-estimate and compare AICC values.
- **Diagnostic Checking:** Check the residuals of the chosen model(s) for whiteness using the tests from Section 5.3. If the minimum AICC model fails, check competing models (those with AICC close to the minimum). Simplicity can be a factor if AICC values are very close.

Example 6.2.1: Australian red wine data Let $\{X_t\}$ be the log-transformed data, differenced at lag 12, and mean-corrected. The sample PACF (Figure 27) suggests an $AR(12)$ model.

- Burg prelim estimation with AICC minimization selects $p = 12$. AICC = -158.77.
- Maximum likelihood $AR(12)$ gives AICC = -158.87.

- Constraining insignificant coefficients (lags 2,3,4,6,7,9,10,11) to zero gives a subset $AR(12)$ model:
$$(1 - 0.270B - 0.224B^5 - 0.149B^8 + 0.099B^{11} + 0.353B^{12})X_t = Z_t$$
with $\hat{\sigma}^2 = 0.0138$ and AICC = -172.49.
- 'Autofit' over $p, q \leq 15$ selects an $ARMA(1, 12)$ model with AICC = -172.74.
- Constraining insignificant MA coefficients (lags 1,3,4,6,7,9,11) gives a subset $ARMA(1, 12)$ model:
$$(1 - 0.286B)X_t = (1 + 0.127B^2 + 0.183B^5 + 0.177B^8 + 0.181B^{10} - 0.554B^{12})Z_t$$
with $\hat{\sigma}^2 = 0.0120$ and AICC = -184.09. This model passes diagnostic checks and is chosen.

Example 6.2.2: The lake data As found in Example 5.2.5, the minimum AICC model for the mean-corrected lake data is $ARMA(1, 1)$, with AICC = 212.77. It passes diagnostic checks.

28.3 Unit Roots in Time Series Models

The presence of a root on or near the unit circle in the AR or MA polynomial is important.

- AR unit root ($\phi(z) = 0$ for $|z| = 1$): Suggests differencing is needed.
- MA unit root ($\theta(z) = 0$ for $|z| = 1$): Suggests overdifferencing.

28.3.1 Unit Roots in Autoregressions

We test if differencing is necessary by testing for a unit root (usually at $z = 1$) in the AR polynomial. The Dickey-Fuller test is commonly used.

AR(1) Case Consider $X_t - \mu = \phi_1(X_{t-1} - \mu) + Z_t$. This can be rewritten as:
$$\nabla X_t = \phi_0^* + \phi_1^* X_{t-1} + Z_t \tag{41}$$
where $\phi_0^* = \mu(1 - \phi_1)$ and $\phi_1^* = \phi_1 - 1$. Testing $H_0 : \phi_1 = 1$ is equivalent to testing $H_0 : \phi_1^* = 0$. We perform an ordinary least squares (OLS) regression of ∇X_t on 1 and X_{t-1} to get $\hat{\phi}_1^*$. The test statistic is the t-ratio:
$$\hat{\tau}_\mu := \hat{\phi}_1^* / \hat{SE}(\hat{\phi}_1^*) \tag{42}$$
Under H_0, the distribution of $\hat{\tau}_\mu$ is *not* the standard t-distribution. Critical values from its specific limit distribution must be used (e.g., -2.86 for $\alpha = 0.05$). H_0 is rejected if $\hat{\tau}_\mu$ is less than the critical value.

AR(p) Case (Augmented Dickey-Fuller Test) The model $X_t - \mu = \sum_{i=1}^p \phi_i(X_{t-i} - \mu) + Z_t$ can be rewritten as:
$$\nabla X_t = \phi_0^* + \phi_1^* X_{t-1} + \sum_{j=2}^{p} \phi_j^* \nabla X_{t-j+1} + Z_t \tag{43}$$
where $\phi_1^* = (\sum_{i=1}^p \phi_i) - 1$. A unit root exists if $\phi(1) = 0$, which is equivalent to $\phi_1^* = 0$. We regress ∇X_t on $1, X_{t-1}, \nabla X_{t-1}, \ldots, \nabla X_{t-p+1}$ via OLS to get $\hat{\phi}_1^*$. The t-ratio:
$$\hat{\tau}_\mu := \hat{\phi}_1^* / \hat{SE}(\hat{\phi}_1^*) \tag{44}$$
has the *same* limiting distribution and critical values as the AR(1) case.

Example 6.3.1: ARIMA(1,1,0) Data For the data in Figure 6.1 (true model has a unit root), we fit an AR(3) model to test for the unit root using the augmented form (6.3.4). The regression gives $\hat{\phi}_1^* = -0.0041$ with $SE = 0.0028$.

$$\hat{\tau}_\mu = -0.0041/0.0028 = -1.464$$

Since $-1.464 > -2.57$ (the 10% critical value), we do not reject H_0. The test correctly suggests the presence of a unit root (or fails to reject it). If the standard t-distribution were used, we might incorrectly reject H_0. Tests excluding the intercept term (ϕ_0^*) use different critical values.

28.3.2 Unit Roots in Moving Averages

Testing for a unit root in $\theta(z)$ can indicate overdifferencing. If $\phi(B)X_t = \theta(B)Z_t$ is the correct model, then ∇X_t follows a model with MA polynomial $\theta(z)(1-z)$, which has a unit root at $z = 1$.

MA(1) Case Consider $X_t = Z_t + \theta Z_{t-1}, \{Z_t\} \sim IID(0, \sigma^2)$. We test $H_0 : \theta = -1$ vs $H_1 : \theta > -1$.

- **Test based on $\hat{\theta}$:** Let $\hat{\theta}$ be the MLE. Under H_0, $n(\hat{\theta}+1)$ converges to a specific distribution. We reject H_0 at level α if $\hat{\theta} > -1 + c_\alpha/n$, where c_α is the $(1-\alpha)$ quantile of the limit distribution ($c_{.05} = 6.80$).
- **Likelihood Ratio Test:** Compute the statistic

  $$\lambda_n := -2\ln\left(\frac{L(-1, S(-1)/n)}{L(\hat{\theta}, \hat{\sigma}^2)}\right)$$

  Reject H_0 if $\lambda_n > c_{LR,\alpha}$, where $c_{LR,\alpha}$ are critical values from the limit distribution ($c_{LR,.05} = 1.94$).

Example 6.3.2: Overshort data The MLE for the mean-corrected overshort data (Example 5.4.1) was $\hat{\theta} = -0.818$ with $n = 57$.

- **Test based on $\hat{\theta}$:** The critical value is $-1 + 6.80/57 = -0.881$. Since $-0.818 > -0.881$, we reject $H_0 : \theta = -1$ at $\alpha = 0.05$.
- **Likelihood Ratio Test:** $-2\ln L(-1, \hat{\sigma}^2(-1)) = 604.584$ and $-2\ln L(\hat{\theta}, \hat{\sigma}^2) = 597.267$. So $\lambda_n = 7.317$. Since $7.317 > 4.41 = c_{LR,.01}$, we reject H_0 at $\alpha = 0.01$.

Both tests suggest $\theta \neq -1$.

28.4 Forecasting ARIMA Models

Forecasting an ARIMA(p,d,q) process X_t, where $Y_t = (1-B)^d X_t$ is a causal $ARMA(p,q)$, requires additional assumptions because the mean and ACVF of X_t are not uniquely determined if $d \geq 1$.

We assume observations $X_{1-d}, \ldots, X_n$ are available, and the initial state $(X_{1-d}, \ldots, X_0)$ is uncorrelated with Y_t for $t > 0$. The model can be written recursively:

$$X_t = Y_t - \sum_{j=1}^{d} \binom{d}{j} (-1)^j X_{t-j}, \quad t = 1, 2, \ldots \tag{45}$$

Let P_n be the best linear predictor based on $\{1, X_{1-d}, \ldots, X_n\}$. Applying P_n to the equation for X_{n+h}:

$$P_n X_{n+h} = P_n Y_{n+h} - \sum_{j=1}^{d} \binom{d}{j} (-1)^j P_n X_{n+h-j} \tag{46}$$

Here, $P_n Y_{n+h}$ is the best linear predictor of Y_{n+h} based on $\{1, Y_1, \ldots, Y_n\}$, calculated as in Section 3.3. The predictors $P_n X_{n+h}$ can be computed recursively starting from $h = 1$, noting $P_n X_{n+h-j} = X_{n+h-j}$ if $n + h - j \leq n$.

Let $\phi^*(z) = \phi(z)(1-z)^d = 1 - \sum_{j=1}^{p+d} \phi_j^* z^j$. The predictor $P_n X_{n+h}$ satisfies:

$$P_n X_{n+h} = \sum_{j=1}^{p+d} \phi_j^* P_n X_{n+h-j} + \sum_{j=h}^{q} \theta_{n+h-1,j} (X_{n+h-j} - \hat{X}_{n+h-j}) \tag{47}$$

where $X_t - \hat{X}_t = Y_t - \hat{Y}_t$ are the one-step prediction errors for Y_t.

The mean squared error is:

$$\sigma_n^2(h) = E(X_{n+h} - P_n X_{n+h})^2 = \sum_{j=0}^{h-1} \left(\sum_{r=0}^{j} \chi_r \theta_{n+h-r-1,j-r} \right)^2 v_{n+h-j-1} \tag{48}$$

where $\chi(z) = (\phi^*(z))^{-1} = \sum \chi_r z^r$, $v_k = E(Y_{k+1} - \hat{Y}_{k+1})^2$. For large n, if $\theta(z) \neq 0$ for $|z| \leq 1$:

$$\sigma_n^2(h) \approx \sigma^2 \sum_{j=0}^{h-1} \psi_j^2 \tag{49}$$

where $\psi(z) = \theta(z)/\phi^*(z) = \sum \psi_j z^j$.

28.4.1 The Forecast Function

For fixed n and $h > q$, the predictor $g(h) := P_n X_{n+h}$ satisfies the homogeneous difference equation:

$$g(h) - \sum_{j=1}^{p+d} \phi_j^* g(h-j) = 0, \quad h > q \tag{50}$$

The general solution depends on the roots of $\phi^*(z) = \phi(z)(1-z)^d = 0$. If the roots are 1 (multiplicity d) and $\xi_1^{-1}, \ldots, \xi_p^{-1}$ (distinct, from $\phi(z)$), the solution is:

$$g(h) = a_0 + a_1 h + \cdots + a_{d-1} h^{d-1} + \sum_{j=1}^{p} b_j \xi_j^{-h}, \quad h > q - p - d \tag{51}$$

The coefficients a_i, b_j are determined by matching this formula with the computed values of $g(h)$ for $q - p - d < h \leq q$. This provides an explicit formula for forecasts at all lead times.

Example 6.4.1: ARIMA(1,1,0) model for Dow Jones The model for the original index D_t is $(1 - 0.4471B)((1-B)D_t - 0.1336) = Z_t$. This rearranges to:

$$(1 - 1.4471B + 0.4471B^2) D_t = 0.1336(1 - 0.4471) + Z_t$$

The forecast function $g(h) = P_{77}D_{77+h}$ satisfies:

$$(1 - 1.4471B + 0.4471B^2)g(h) = 0.1336(1 - 0.4471) = 0.07387, \quad h > 0$$

The roots of $\phi^*(z) = (1 - z)(1 - 0.4471z)$ are 1 and $1/0.4471$. The general solution form is $g(h) = a_0 + a_1 h + b_1(0.4471)^h$. A particular solution is $g(h) = 0.1336h$. So,

$$g(h) = 0.1336h + a + b(0.4471)^h, \quad h > -2$$

Using $g(0) = D_{77} = 121.23$ and $g(-1) = D_{76} = 122$, we solve for a and b to get:

$$g(h) = 0.1336h + 120.50 + 0.7331(0.4471)^h$$

Forecasts: $g(1) = P_{77}D_{78} = 120.97$, $g(2) = P_{77}D_{79} = 120.94$. MSEs: $\sigma^2_{77}(1) = \sigma^2 = 0.1455$, $\sigma^2_{77}(2) = \sigma^2(1 + \psi_1^2) = \sigma^2(1 + (1.4471)^2) = 0.4502$. (Note $\psi_1 = \phi_1^* = 1.4471$ here).

28.5 Seasonal ARIMA Models

Seasonal patterns can be modeled using differencing at the seasonal lag s, $Y_t = (1 - B^s)X_t$, and then fitting an ARMA model to Y_t. This is a special case of the general Seasonal ARIMA (SARIMA) model.

Definition 6.5.1: SARIMA(p, d, q) x (P, D, Q)$_s$ Process If d, D are non-negative integers, $\{X_t\}$ is a SARIMA(p, d, q) x (P, D, Q)$_s$ process with period s if the differenced series $Y_t := (1 - B)^d(1 - B^s)^D X_t$ is a causal ARMA process defined by:

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)Z_t, \quad \{Z_t\} \sim WN(0, \sigma^2) \tag{52}$$

where:

- $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ (non-seasonal AR polynomial)
- $\Phi(z) = 1 - \Phi_1 z - \cdots - \Phi_P z^P$ (seasonal AR polynomial)
- $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$ (non-seasonal MA polynomial)
- $\Theta(z) = 1 + \Theta_1 z + \cdots + \Theta_Q z^Q$ (seasonal MA polynomial)

Causality requires $\phi(z) \neq 0$ and $\Phi(z) \neq 0$ for $|z| \leq 1$. Invertibility requires $\theta(z) \neq 0$ and $\Theta(z) \neq 0$ for $|z| \leq 1$. The model allows for random variation in the seasonal pattern over time.

Identification 1. Choose d and D (usually 0 or 1) by differencing until the series $Y_t = (1 - B)^d(1 - B^s)^D X_t$ appears stationary. 2. Examine the sample ACF and PACF of Y_t. 3. Look at lags $s, 2s, 3s, \ldots$ to identify the seasonal orders P, Q. The pattern should resemble the ACF/PACF of an $ARMA(P, Q)$ process. 4. Look at lags $1, 2, \ldots, s-1$ (within the first season) to identify the non-seasonal orders p, q. The pattern should resemble the ACF/PACF of an $ARMA(p, q)$ process. 5. Use AICC minimization over candidate $(p, d, q) \times (P, D, Q)_s$ models. 6. Perform diagnostic checks on the residuals of the chosen model.

Estimation Parameters $(\phi, \Phi, \theta, \Theta, \sigma^2)$ are estimated by maximizing the likelihood of the differenced series Y_t. This is equivalent to fitting a constrained $ARMA(p + sP, q + sQ)$ model to Y_t.

Example 6.5.4: Monthly accidental deaths Let X_t be the monthly deaths. The series $Y_t = (1-B)(1-B^{12})X_t$ appears stationary. Its sample ACF is shown in Figure 6.17.

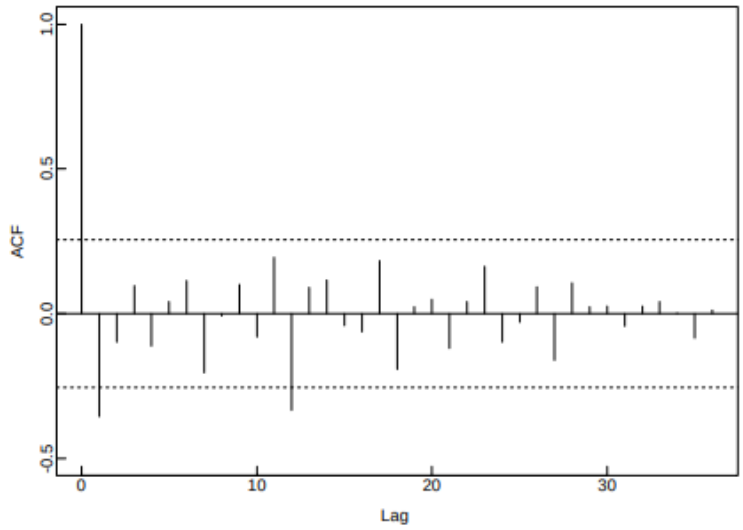

Figure 28: Sample ACF of the differenced accidental deaths $\{\nabla\nabla_{12}X_t\}$.

- Lags $12, 24, 36$: $\hat{\rho}(12) = -0.333, \hat{\rho}(24) = -0.099, \hat{\rho}(36) = 0.013$. Suggests a seasonal MA(1), i.e., $P = 0, Q = 1$.
- Lags $1, \ldots, 11$: Only $\hat{\rho}(1)$ is significant. Suggests a non-seasonal MA(1), i.e., $p = 0, q = 1$.

This leads to a SARIMA(0,1,1) x $(0,1,1)_{12}$ model for X_t, meaning $Y_t = \mu + (1+\theta_1 B)(1+\Theta_1 B^{12})Z_t$. Maximum likelihood estimation gives:

$$\nabla\nabla_{12}X_t = 28.831 + (1-0.478B)(1-0.588B^{12})Z_t, \quad \{Z_t\} \sim WN(0, 94390)$$

AICC = 855.53. An alternative approach fits a subset MA(13) model directly to Y_t. After constraining insignificant coefficients, the model is:

$$\nabla\nabla_{12}X_t = 28.831 + Z_t - 0.596Z_{t-1} - 0.407Z_{t-6} - 0.685Z_{t-12} + 0.460Z_{t-13}$$

with $\{Z_t\} \sim WN(0, 71240)$ and AICC = 855.61. Both models pass diagnostic checks and have similar AICC values.

28.5.1 Forecasting SARIMA Processes

Forecasting follows the same logic as for ARIMA processes. Let $Y_t = (1-B)^d(1-B^s)^D X_t$. Rearranging gives X_t in terms of Y_t and past X_t. Let a_j be coefficients such that $(1-B)^d(1-B^s)^D = 1 - \sum_{j=1}^{d+Ds} a_j B^j$.

$$X_{n+h} = Y_{n+h} + \sum_{j=1}^{d+Ds} a_j X_{n+h-j}$$

Applying P_n gives the recursive forecast equation:

$$P_n X_{n+h} = P_n Y_{n+h} + \sum_{j=1}^{d+Ds} a_j P_n X_{n+h-j} \tag{53}$$

where P_nY_{n+h} is the forecast for the ARMA process Y_t. The MSE is given by:

$$\sigma_n^2(h) = E(X_{n+h} - P_nX_{n+h})^2 = \sum_{j=0}^{h-1} \left(\sum_{r=0}^{j} \chi_r \theta_{n+h-r-1,j-r} \right)^2 v_{n+h-j-1} \tag{54}$$

where $\chi(z) = [\phi(z)\Phi(z^s)(1-z)^d(1-z^s)^D]^{-1} = \sum \chi_r z^r$. For large n, if the model is invertible:

$$\sigma_n^2(h) \approx \sigma^2 \sum_{j=0}^{h-1} \psi_j^2 \tag{55}$$

where $\psi(z) = \frac{\theta(z)\Theta(z^s)}{\phi(z)\Phi(z^s)(1-z)^d(1-z^s)^D} = \sum \psi_j z^j$.

Example 6.5.5: Monthly accidental deaths Forecasting the next 6 values using models (6.5.8) and (6.5.9). The numerical results are in Table 6.1.

28.6 Regression with ARMA Errors

Standard linear regression assumes independent errors. When errors are correlated (common in time series), we use models like:

$$Y_t = x_t'\beta + W_t, \quad t = 1, \ldots, n \quad \text{or} \quad Y = X\beta + W \tag{56}$$

where Y is the observation vector, X is the design matrix (with t^{th} row x_t'), β is the vector of regression coefficients, and W is the error vector, assumed to be from a causal, zero-mean $ARMA(p,q)$ process:

$$\phi(B)W_t = \theta(B)Z_t, \quad \{Z_t\} \sim WN(0, \sigma^2) \tag{57}$$

28.6.1 OLS and GLS Estimation

- **Ordinary Least Squares (OLS):** Minimizes $\sum(Y_t - x_t'\beta)^2$.

  $$\hat{\beta}_{OLS} = (X'X)^{-1}X'Y \tag{58}$$

  OLS is unbiased ($E(\hat{\beta}_{OLS}) = \beta$) but not efficient if errors are correlated. Its covariance is:

  $$Cov(\hat{\beta}_{OLS}) = (X'X)^{-1}X'\Gamma_n X(X'X)^{-1} \tag{59}$$

  where $\Gamma_n = E(WW')$ is the covariance matrix of the ARMA errors. Using the incorrect formula $\sigma^2(X'X)^{-1}$ (which assumes independence) can lead to very inaccurate standard errors.

- **Generalized Least Squares (GLS):** Minimizes $(Y - X\beta)'\Gamma_n^{-1}(Y - X\beta)$.

  $$\hat{\beta}_{GLS} = (X'\Gamma_n^{-1}X)^{-1}X'\Gamma_n^{-1}Y \tag{60}$$

  GLS is the Best Linear Unbiased Estimator (BLUE). Its covariance is:

  $$Cov(\hat{\beta}_{GLS}) = (X'\Gamma_n^{-1}X)^{-1} \tag{61}$$

  GLS requires knowing Γ_n (i.e., the ARMA parameters ϕ, θ, σ^2).

Let $V = \sigma^{-2}\Gamma_n$. Let T be a matrix such that $T'T = V^{-1}$. Multiplying the regression model by T:

$$TY = TX\beta + TW$$

The transformed errors TW are uncorrelated with variance σ^2. Applying OLS to this transformed model yields the GLS estimator:

$$\hat{\beta}_{GLS} = ((TX)'(TX))^{-1}(TX)'(TY) = (X'T'TX)^{-1}X'T'TY$$

The matrix T can be found using the innovations algorithm; multiplying W by T effectively computes the standardized one-step prediction errors (residuals) of the ARMA process W_t. GLS estimation involves: 1. Transform $Y \to Y^* = TY$ (compute ARMA residuals for Y). 2. Transform each column X_j of $X \to X_j^* = TX_j$ (compute ARMA residuals for X_j). 3. Compute $\hat{\beta}_{GLS} = (X^{*'}X^*)^{-1}X^{*'}Y^*$.

28.6.2 ML Estimation

If ARMA parameters (ϕ, θ, σ^2) are unknown, we estimate them simultaneously with β by maximizing the Gaussian likelihood:

$$L(\beta, \phi, \theta, \sigma^2) = (2\pi)^{-n/2}(\det \Gamma_n)^{-1/2} \exp\{-\frac{1}{2}(Y - X\beta)'\Gamma_n^{-1}(Y - X\beta)\}$$

This is equivalent to minimizing:

$$l(\beta, \phi, \theta) = \ln(n^{-1}S(\beta, \phi, \theta)) + n^{-1}\sum_{t=1}^{n} \ln r_{t-1} \tag{62}$$

where $S(\beta, \phi, \theta) = \sum(W_t - \hat{W}_t)^2/r_{t-1}$ is the residual sum of squares for $W_t = Y_t - x_t'\beta$, calculated using the ARMA parameters ϕ, θ. The MLE for σ^2 is $\hat{\sigma}^2 = S(\hat{\beta}, \hat{\phi}, \hat{\theta})/n$.

An iterative scheme (generalizing Cochrane-Orcutt) can be used:

1. Start with $\hat{\beta}^{(0)} = \hat{\beta}_{OLS}$. Compute residuals $\hat{W}^{(0)} = Y - X\hat{\beta}^{(0)}$.
2. Fit an ARMA(p,q) model to $\hat{W}^{(0)}$ via MLE to get $\hat{\phi}^{(1)}, \hat{\theta}^{(1)}$.
3. Using $\hat{\phi}^{(1)}, \hat{\theta}^{(1)}$, compute $\hat{\beta}^{(1)} = \hat{\beta}_{GLS}(\hat{\phi}^{(1)}, \hat{\theta}^{(1)})$.
4. Compute new residuals $\hat{W}^{(1)} = Y - X\hat{\beta}^{(1)}$.
5. Repeat steps 2-4 until estimates converge.

The MLEs $\hat{\beta}, \hat{\phi}, \hat{\theta}$ are asymptotically normal. $\hat{\beta}$ is asymptotically independent of $(\hat{\phi}, \hat{\theta})$. The asymptotic covariance matrix of $\hat{\beta}$ is the same as for GLS.

Example 6.6.1: Overshort data Model $Y_t = \beta + W_t$, where $W_t = Z_t + \theta Z_{t-1}$. OLS gives $\hat{\beta}_{OLS} = \bar{Y} = -4.035$. Variance assuming independence is 59.92. Iterative MLE/GLS:

- Iteration 0: $\hat{\beta}^{(0)} = -4.035$. Residuals lead to $\hat{\theta}^{(1)} = -0.818$.
- Iteration 1: GLS using $\hat{\theta}^{(1)}$ gives $\hat{\beta}^{(1)} = -4.745$. Residuals lead to $\hat{\theta}^{(2)} = -0.848$.
- Iteration 2: GLS using $\hat{\theta}^{(2)}$ gives $\hat{\beta}^{(2)} = -4.780$. Residuals lead to $\hat{\theta}^{(3)} = -0.848$.

Converged estimates: $\hat{\beta} = -4.780$, $\hat{\theta} = -0.848$. The standard error for $\hat{\beta}$ is now $\sqrt{1.408} \approx 1.187$. The 95% CI is $(-7.07, -2.43)$, strongly suggesting $\beta \neq 0$. Accounting for dependence was crucial.

Example 6.6.2: Lake data Model $Y_t = \beta_0 + \beta_1 t + W_t$, where W_t is $AR(2)$. OLS: $\hat{\beta}_{OLS} = (10.202, -0.0242)'$. Covariance assuming independence: $\begin{bmatrix} 0.0730 & -0.0011 \\ -0.0011 & 0.00002 \end{bmatrix}$. Iterative MLE/GLS converges quickly to $\hat{\beta} = (10.09, -0.0216)'$, $\hat{\phi}_1 = 1.005$, $\hat{\phi}_2 = -0.291$. Correct asymptotic covariance for $\hat{\beta}_{GLS}$ (and $\hat{\beta}_{MLE}$) is estimated as $\begin{bmatrix} 0.2139 & -0.0032 \\ -0.0032 & 0.00006 \end{bmatrix}$. The variance estimates are about 3 times larger than those assuming independence. The 95% CI for β_1 is $-0.0216 \pm 1.96\sqrt{0.00006} = -0.0216 \pm 0.0048$, indicating a significant decreasing trend.

Example 6.6.3: Seat-belt legislation Y_t = monthly deaths/injuries. Model $Y_t = a + bf(t) + W_t$, where $f(t) = 0$ before Feb '83 ($t = 98$) and $f(t) = 1$ from Feb '83 ($t = 99$) onwards. We want to test if $b < 0$.

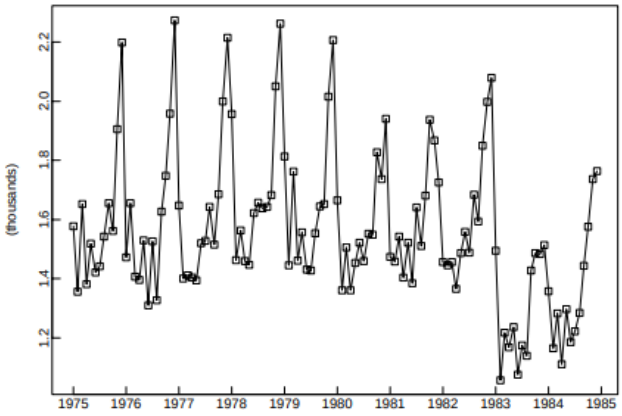

Figure 29: Monthly deaths and serious injuries $\{Y_t\}$ on UK roads, Jan '75 - Dec '84.

OLS residuals show strong seasonality (period 12). We difference at lag 12: $X_t = Y_t - Y_{t-12}$. The model becomes approximately $X_t = bg_t + N_t$, where $g_t = f_t - f_{t-12}$ (1 for $99 \leq t \leq 110$, 0 otherwise) and $N_t = W_t - W_{t-12}$ is stationary. OLS on $X_t = bg_t + N_t$ gives $\hat{b} = -346.92$. Residuals suggest an MA(12) model for N_t. Iterative MLE/GLS converges to:

$$X_t = -328.45 g_t + N_t$$

where N_t is MA(12) with $\hat{\sigma}^2 = 12581$. The significant MA coefficients are at lags 1, 11, and 12 ($\hat{\theta}_1 = 0.219, \hat{\theta}_{11} = 0.183, \hat{\theta}_{12} = -0.627$). The standard error for $\hat{b}$ is 49.41. Since $\hat{b} = -328.45$ is highly significant and negative, the legislation appears effective.

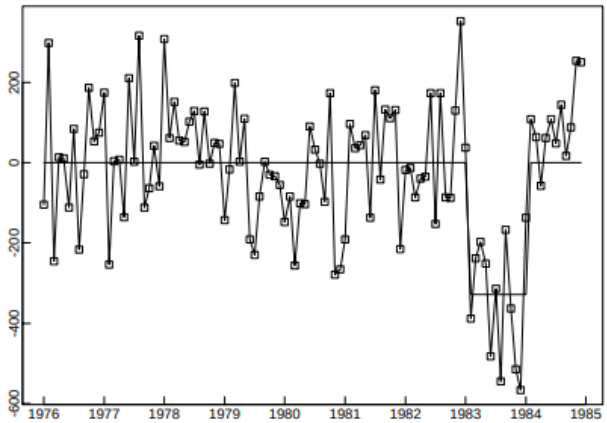

Figure 30: Differenced deaths and serious injuries $\{X_t = Y_t - Y_{t-12}\}$ showing the fitted GLS regression function (step function based on g_t).

29 Multivariate Time Series

Often, multiple time series $\{X_{t1}\}, \{X_{t2}\}, \ldots, \{X_{tm}\}$ are related. We analyze them jointly as a vector time series $X_t = (X_{t1}, \ldots, X_{tm})'$. This allows us to model interdependence between component series.

29.1 Second-Order Properties of Multivariate Time Series

Let $X_t = (X_{t1}, \ldots, X_{tm})'$ be an m-variate time series with $EX_{ti}^2 < \infty$. The second-order properties are defined by:

- **Mean Vector**: $\mu_t := EX_t = (\mu_{t1}, \ldots, \mu_{tm})'$
- **Covariance Matrices**: $\Gamma(t+h, t) := E[(X_{t+h} - \mu_{t+h})(X_t - \mu_t)']$. This is an $m \times m$ matrix where the (i,j)-element is $\gamma_{ij}(t+h, t) := \text{Cov}(X_{t+h,i}, X_{t,j})$.

Definition 7.2.1: Stationary Multivariate Time Series The m-variate series $\{X_t\}$ is (weakly) stationary if

(i) μ_t is independent of t (constant mean vector μ).

(ii) $\Gamma(t+h, t)$ depends only on the lag h, for all t (covariance matrix function $\Gamma(h)$).

For a stationary series:

$$\mu = EX_t$$

$$\Gamma(h) = E[(X_{t+h} - \mu)(X_t - \mu)'] = [\gamma_{ij}(h)]_{i,j=1}^m$$

- The diagonal elements $\gamma_{ii}(h)$ are the autocovariance functions of the component series $\{X_{ti}\}$.
- The off-diagonal elements $\gamma_{ij}(h)$ for $i \neq j$ are the **cross-covariance functions**. Note that $\gamma_{ij}(h) = \text{Cov}(X_{t+h,i}, X_{t,j})$ and $\gamma_{ji}(h) = \text{Cov}(X_{t+h,j}, X_{t,i})$.

The **Correlation Matrix Function** $R(h)$ is defined as:

$$R(h) := [\rho_{ij}(h)]_{i,j=1}^m, \quad \text{where } \rho_{ij}(h) = \frac{\gamma_{ij}(h)}{\sqrt{\gamma_{ii}(0)\gamma_{jj}(0)}} \tag{63}$$

$\rho_{ij}(h)$ is the cross-correlation function.

Example 7.2.1 Let $X_{t1} = Z_t$, $X_{t2} = Z_t + 0.75Z_{t-10}$, where $\{Z_t\} \sim WN(0,1)$. This bivariate series is stationary with $\mu = 0$. The non-zero covariance matrices are:

$$\Gamma(0) = \begin{bmatrix} 1 & 1 \\ 1 & 1.5625 \end{bmatrix}, \quad \Gamma(10) = \begin{bmatrix} 0 & 0 \\ 0.75 & 0.75 \end{bmatrix}, \quad \Gamma(-10) = \Gamma(10)' = \begin{bmatrix} 0 & 0.75 \\ 0 & 0.75 \end{bmatrix}$$

The non-zero correlation matrices are:

$$R(0) = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}, \quad R(10) = \begin{bmatrix} 0 & 0 \\ 0.60 & 0.48 \end{bmatrix}, \quad R(-10) = \begin{bmatrix} 0 & 0.60 \\ 0 & 0.48 \end{bmatrix}$$

Basic Properties of $\Gamma(\cdot)$

1. $\Gamma(h) = \Gamma'(-h)$
2. $|\gamma_{ij}(h)| \leq \sqrt{\gamma_{ii}(0)\gamma_{jj}(0)}$
3. $\gamma_{ii}(\cdot)$ is an autocovariance function for each i.
4. $\sum_{j,k=1}^{n} a_j'\Gamma(j-k)a_k \geq 0$ for all $n \in \{1, 2, \dots\}$ and $a_1, \dots, a_n \in \mathbb{R}^m$ (non-negative definite property).

Note that $|\gamma_{ij}(h)|$ can be greater than $|\gamma_{ij}(0)|$ if $i \neq j$.

Definition 7.2.2: Multivariate White Noise The m-variate series $\{Z_t\}$ is white noise with mean 0 and covariance matrix Σ, written $\{Z_t\} \sim WN(0, \Sigma)$, if $\{Z_t\}$ is stationary with $EZ_t = 0$ and

$$\Gamma_Z(h) = \begin{cases} \Sigma, & \text{if } h = 0 \\ 0, & \text{otherwise} \end{cases}$$

where Σ is a positive semidefinite matrix.

Definition 7.2.3: Multivariate IID Noise The m-variate series $\{Z_t\}$ is IID noise with mean 0 and covariance matrix Σ, written $\{Z_t\} \sim IID(0, \Sigma)$, if the random vectors $\{Z_t\}$ are independent and identically distributed with $EZ_t = 0$ and $E(Z_t Z_t') = \Sigma$.

Definition 7.2.4: Multivariate Linear Process The m-variate series $\{X_t\}$ is a linear process if it has the representation

$$X_t = \sum_{j=-\infty}^{\infty} C_j Z_{t-j}, \quad \{Z_t\} \sim WN(0, \Sigma) \tag{64}$$

where $\{C_j\}$ is a sequence of $m \times m$ matrices whose components are absolutely summable ($\sum_j |(C_j)_{kl}| < \infty$ for all k, l). The linear process is stationary with mean 0 and covariance function:

$$\Gamma_X(h) = \sum_{j=-\infty}^{\infty} C_{j+h}\Sigma C_j' \tag{65}$$

If $C_j = 0$ for $j < 0$, it is an $MA(\infty)$ process.

29.2 Estimation of the Mean and Covariance Function

29.2.1 Estimation of μ

Based on observations $X_1, \dots, X_n$, the mean vector μ is estimated by the sample mean vector:

$$\overline{X}_n = \frac{1}{n}\sum_{t=1}^{n} X_t$$

This is unbiased. Its consistency is stated in Proposition 7.3.1.

Proposition 7.3.1 If $\{X_t\}$ is stationary with mean μ and covariance $\Gamma(\cdot)$:

- $E(\overline{X}_n - \mu)'(\overline{X}_n - \mu) \to 0$ if $\gamma_{ii}(n) \to 0$ for all i.
- $nE(\overline{X}_n - \mu)'(\overline{X}_n - \mu) \to \sum_{i=1}^{m} \sum_{h=-\infty}^{\infty} \gamma_{ii}(h)$ if $\sum_h |\gamma_{ii}(h)| < \infty$ for all i.

Under further conditions, $\overline{X}_n$ is approximately normal for large n. Approximate confidence regions for μ can be constructed based on individual confidence intervals for each component μ_i, using estimates of the spectral densities at zero, $2\pi f_i(0) = \sum_k \gamma_{ii}(k)$.

29.2.2 Estimation of $\Gamma(h)$

The sample covariance matrix function is:

$$\hat{\Gamma}(h) = \begin{cases} n^{-1} \sum_{t=1}^{n-h} (X_{t+h} - \overline{X}_n)(X_t - \overline{X}_n)', & 0 \le h \le n-1 \\ \hat{\Gamma}'(-h), & -n+1 \le h < 0 \end{cases} \tag{66}$$

The sample cross-correlation function is:

$$\hat{\rho}_{ij}(h) = \frac{\hat{\gamma}_{ij}(h)}{\sqrt{\hat{\gamma}_{ii}(0)\hat{\gamma}_{jj}(0)}} \tag{67}$$

Theorem 7.3.1 (Large-sample distribution of $\hat{\rho}_{12}(h)$ under independence) Let $\{X_{t1}\}$ and $\{X_{t2}\}$ be independent linear processes: $X_{t1} = \sum \alpha_k Z_{t-k,1}$, $X_{t2} = \sum \beta_k Z_{t-k,2}$, where $\{Z_{t1}\}$ and $\{Z_{t2}\}$ are independent IID sequences. Then for large n, $n^{1/2}\hat{\rho}_{12}(h)$ and $n^{1/2}\hat{\rho}_{12}(k)$ $(h \neq k)$ are approximately bivariate normal with:

- Means: 0
- Variance: $\sum_{j=-\infty}^{\infty} \rho_{11}(j)\rho_{22}(j)$
- Covariance: $\sum_{j=-\infty}^{\infty} \rho_{11}(j)\rho_{22}(j+k-h)$

Important Consequence: If one series (say $\{X_{t1}\}$) is white noise, then $\rho_{11}(j) = 0$ for $j \neq 0$. The variance becomes $\rho_{11}(0)\rho_{22}(0) = 1$, and the covariance is 0. Thus, if one series is white noise and they are independent, $\hat{\rho}_{12}(h)$ is approx $N(0, 1/n)$ for $h \neq 0$.

If neither series is white noise, the variance of $\hat{\rho}_{12}(h)$ can be much larger than $1/n$, even under independence. A large $\hat{\rho}_{12}(h)$ does not necessarily imply dependence unless the autocorrelations $\rho_{11}(\cdot)$ and $\rho_{22}(\cdot)$ are accounted for.

29.2.3 Testing for Independence of Two Stationary Time Series

Because the distribution of $\hat{\rho}_{12}(h)$ depends on the individual series' ACFs, we use **prewhitening**. 1. Fit appropriate invertible ARMA models to each series $\{X_{t1}\}$ and $\{X_{t2}\}$ individually. 2. Compute the residuals (whitened series) $\{\hat{W}_{t1}\}$ and $\{\hat{W}_{t2}\}$. 3. Compute the sample cross-correlation function $\hat{\rho}_{W,12}(h)$ between the residual series. 4. If the original series were independent, the whitened series should also be independent. By Theorem 7.3.1 (since residuals are approximately white noise), $\hat{\rho}_{W,12}(h)$ should be approximately $N(0, 1/n)$ for large n. 5. Check if the values $\hat{\rho}_{W,12}(h)$ fall within the bounds $\pm 1.96/\sqrt{n}$. If significantly many fall outside, reject the independence hypothesis.

Example 7.3.1 A bivariate series (length 200) shows significant sample cross-correlations. However, both component series are strongly autocorrelated (resembling AR(1) processes). After fitting AR(1) models to each component and obtaining the residuals, the sample cross-correlations between the residual series are computed. Nearly all fall within $\pm 1.96/\sqrt{200} = \pm 0.139$. This suggests the original series were likely independent (which they were by construction in this example).

29.2.4 Bartlett's Formula (Multivariate)

Provides a large-sample approximation for the covariance between sample cross-correlations $\hat{\rho}_{12}(h)$ and $\hat{\rho}_{12}(k)$ for a Gaussian bivariate process $\{X_t\}$ (without assuming independence). The formula is complex:

$$\begin{aligned}\lim_{n\to\infty} n\text{Cov}(\hat{\rho}_{12}(h), \hat{\rho}_{12}(k)) = \sum_{j=-\infty}^{\infty} &[\rho_{11}(j)\rho_{22}(j+k-h) + \rho_{12}(j+k)\rho_{21}(j-h) \\ &- \rho_{12}(h)\{\rho_{11}(j)\rho_{12}(j+k) + \rho_{22}(j)\rho_{21}(j-k)\} \\ &- \rho_{12}(k)\{\rho_{11}(j)\rho_{12}(j+h) + \rho_{22}(j)\rho_{21}(j-h)\} \\ &+ \rho_{12}(h)\rho_{12}(k)\{\tfrac{1}{2}\rho_{11}^2(j) + \rho_{12}^2(j) + \tfrac{1}{2}\rho_{22}^2(j)\}]\end{aligned}$$

Corollary 7.3.1 If $\{X_t\}$ satisfies the conditions for Bartlett's formula, if either $\{X_{t1}\}$ or $\{X_{t2}\}$ is white noise, and if $\rho_{12}(h) = 0$ for $h \notin [a, b]$, then $\lim_{n\to\infty} n\text{Var}(\hat{\rho}_{12}(h)) = 1$ for $h \notin [a, b]$.

Example 7.3.2: Sales with a leading indicator Analyzes differenced sales $\{D_{t2}\}$ and leading indicator $\{D_{t1}\}$. Univariate models are MA(1) and ARMA(1,1) respectively. Prewhitening both series yields residuals $\{\hat{W}_{t1}\}$ and $\{\hat{W}_{t2}\}$. Their sample cross-correlation function shows only one significant value at lag $h = -3$ ($\hat{\rho}_{12}(-3) = 0.969$). This suggests a relationship of the form $\hat{W}_{t2} \approx c\hat{W}_{t-3,1} + N_t$, where c is related to the ratio of variances and N_t is noise. This leads towards a transfer function model relating the original series.

29.3 Multivariate ARMA Processes

Definition 7.4.1: Multivariate ARMA(p, q) Process $\{X_t\}$ is an m-variate ARMA(p, q) process if $\{X_t\}$ is stationary and satisfies:

$$X_t - \Phi_1 X_{t-1} - \cdots - \Phi_p X_{t-p} = Z_t + \Theta_1 Z_{t-1} + \cdots + \Theta_q Z_{t-q} \tag{68}$$

where $\{Z_t\} \sim WN(0, \Sigma)$, and Φ_i, Θ_j are $m \times m$ coefficient matrices. Compact form:

$$\Phi(B)X_t = \Theta(B)Z_t \tag{69}$$

where $\Phi(z) = I - \Phi_1 z - \cdots - \Phi_p z^p$ and $\Theta(z) = I + \Theta_1 z + \cdots + \Theta_q z^q$ are matrix polynomials.

Example 7.4.1: Multivariate AR(1) $X_t = \Phi X_{t-1} + Z_t$. If $det(I - z\Phi) \neq 0$ for $|z| \leq 1$ (i.e., all eigenvalues of Φ are less than 1 in absolute value), then the unique stationary causal solution is:

$$X_t = \sum_{j=0}^{\infty} \Phi^j Z_{t-j} \tag{70}$$

Causality An ARMA(p,q) process is causal if it can be written as $X_t = \sum_{j=0}^{\infty} \Psi_j Z_{t-j}$ where components of Ψ_j are absolutely summable. Condition: $\det \Phi(z) \neq 0$ for all $|z| \leq 1$. The matrices Ψ_j satisfy $\Psi_j = \Theta_j + \sum_{k=1}^{p} \Phi_k \Psi_{j-k}$, $j \geq 0$ (with $\Psi_j = 0$ for $j < 0$, $\Theta_0 = I$, $\Theta_j = 0$ for $j > q$, $\Phi_k = 0$ for $k > p$).

Invertibility An ARMA(p,q) process is invertible if $Z_t = \sum_{j=0}^{\infty} \Pi_j X_{t-j}$ where components of Π_j are absolutely summable. Condition: $\det \Theta(z) \neq 0$ for all $|z| \leq 1$. The matrices Π_j satisfy $\Pi_j = -\Phi_j - \sum_{k=1}^{q} \Theta_k \Pi_{j-k}$, $j \geq 0$ (with $\Pi_j = 0$ for $j < 0$, $\Phi_0 = -I$, $\Phi_j = 0$ for $j > p$, $\Theta_k = 0$ for $k > q$).

Remark 3 (Non-uniqueness) Consider a bivariate AR(1) with $\Phi = \left[\begin{smallmatrix} 0 & 0.5 \\ 0 & 0 \end{smallmatrix}\right]$. Then $\Phi^j = 0$ for $j > 1$. The process is $X_t = Z_t + \Phi Z_{t-1}$, which is also an MA(1). This shows multivariate ARMA models may not have unique orders without further restrictions.

29.3.1 The Covariance Matrix Function of a Causal ARMA Process

For a causal ARMA process $X_t = \sum \Psi_j Z_{t-j}$, the covariance matrix function is:

$$\Gamma(h) = E(X_{t+h} X_t') = \sum_{j=0}^{\infty} \Psi_{j+h} \Sigma \Psi_j', \quad h \geq 0 \tag{71}$$

and $\Gamma(h) = \Gamma'(-h)$ for $h < 0$. Alternatively, $\Gamma(h)$ satisfies the Yule-Walker type equations:

$$\Gamma(j) - \sum_{r=1}^{p} \Phi_r \Gamma(j-r) = \sum_{r=j}^{q} \Theta_r \Sigma \Psi_{r-j}', \quad j = 0, 1, \ldots \tag{72}$$

These can be solved for $\Gamma(0), \ldots, \Gamma(p)$ and then recursively for $h > p$.

29.4 Best Linear Predictors of Second-Order Random Vectors

Let $\{X_t\}$ be an m-variate series with mean μ_t and covariance $K(i,j)$. Let Y be another m-variate random vector with mean μ. The best linear predictor of Y based on $\{1, X_1, \ldots, X_n\}$ is $P_n(Y) = (P_n Y_1, \ldots, P_n Y_m)'$, where $P_n Y_j$ is the best predictor of Y_j using all components of $X_1, \ldots, X_n$. It has the form:

$$P_n(Y) = \mu + \sum_{i=1}^{n} A_i (X_{n+1-i} - \mu_{n+1-i}) \tag{73}$$

and satisfies the orthogonality condition:

$$E[(Y - P_n(Y)) X_{n+1-i}'] = 0 \quad \text{(zero matrix), for } i = 1, \ldots, n \tag{74}$$

The one-step predictor for a zero-mean series is $\hat{X}_{n+1} = P_n(X_{n+1}) = \sum_{j=1}^{n} \Phi_{nj} X_{n+1-j}$. The coefficient matrices satisfy:

$$\sum_{j=1}^{n} \Phi_{nj} K(n+1-j, n+1-i) = K(n+1, n+1-i), \quad i = 1, \ldots, n \tag{75}$$

If $\{X_t\}$ is stationary, $K(i,j) = \Gamma(i-j)$, and the equations become:

$$\sum_{j=1}^{n} \Phi_{nj}\Gamma(i-j) = \Gamma(i), \quad i = 1, \ldots, n \tag{76}$$

These can be solved recursively using Whittle's generalization of the Durbin-Levinson algorithm. The algorithm also yields the prediction error covariance matrices $V_n = E(X_{n+1} - \hat{X}_{n+1})(X_{n+1} - \hat{X}_{n+1})'$.

$$V_n = \Gamma(0) - \sum_{j=1}^{n} \Phi_{nj}\Gamma(-j)$$

Multivariate versions of the innovations algorithm also exist.

29.5 Modeling and Forecasting with Multivariate AR Processes

29.5.1 Estimation for Autoregressive Processes Using Whittle's Algorithm

For a causal m-variate $AR(p)$ process $X_t = \sum_{i=1}^{p} \Phi_i X_{t-i} + Z_t$, $\{Z_t\} \sim WN(0, \Sigma)$, the parameters $\Phi_1, \ldots, \Phi_p, \Sigma$ satisfy the Yule-Walker equations:

$$\Gamma(i) = \sum_{j=1}^{p} \Phi_j \Gamma(i-j), \quad i = 1, \ldots, p \tag{77}$$

$$\Sigma = \Gamma(0) - \sum_{j=1}^{p} \Phi_j \Gamma(-j) \tag{78}$$

The Yule-Walker estimators $\hat{\Phi}_1, \ldots, \hat{\Phi}_p, \hat{\Sigma}$ are obtained by replacing $\Gamma(h)$ with $\hat{\Gamma}(h)$ and solving. This is computationally equivalent to finding the prediction coefficients using Whittle's algorithm. Order selection can be based on minimizing the multivariate AICC:

$$AICC = -2\ln L(\hat{\Phi}_p, \hat{\Sigma}) + \frac{2(pm^2+1)nm}{nm - pm^2 - 2} \tag{79}$$

where L is the Gaussian likelihood based on the fitted $AR(p)$ model.

Example 7.6.1: Dow Jones and All Ordinaries Indices Analyzes the bivariate series $X_t = (X_{t1}, X_{t2})'$ of percentage returns. Using Yule-Walker estimation with AICC minimization selects an $AR(1)$ model:

$$X_t = \begin{bmatrix} 0.0288 \\ 0.00836 \end{bmatrix} + \begin{bmatrix} -0.0148 & 0.0357 \\ 0.6589 & 0.0998 \end{bmatrix} X_{t-1} + Z_t$$

with $Z_t \sim WN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.3653 & 0.0224 \\ 0.0224 & 0.6016 \end{bmatrix}\right)$.

Example 7.6.2: Sales with a leading indicator Analyzes the differenced bivariate series $X_t = \nabla Y_t$. Minimum AICC Yule-Walker selects an $AR(5)$ model. The estimated coefficient matrices $\hat{\Phi}_1, \ldots, \hat{\Phi}_5$ and $\hat{\Sigma}$ are given. Analysis suggests X_{t1} (differenced indicator) influences X_{t2} (differenced sales) at lags 3, 4, 5, leading towards a transfer function model.

29.5.2 Forecasting Multivariate Autoregressive Processes

For a causal $AR(p)$ process $X_t = \phi_0 + \sum_{i=1}^{p} \Phi_i X_{t-i} + Z_t$, the best linear predictor based on $X_1, \ldots, X_n$ (assuming $n \geq p$) is:

- 1-step predictor: $P_n X_{n+1} = \phi_0 + \sum_{i=1}^{p} \Phi_i X_{n+1-i}$
- h-step predictor ($h > 1$): Computed recursively using

$$P_n X_{n+h} = \phi_0 + \sum_{i=1}^{p} \Phi_i P_n X_{n+h-i} \tag{80}$$

  (where $P_n X_k = X_k$ if $k \leq n$).

The h-step prediction error is $X_{n+h} - P_n X_{n+h} = \sum_{j=0}^{h-1} \Psi_j Z_{n+h-j}$, where $\Psi(z) = (\Phi(z))^{-1} = \sum \Psi_j z^j$. The error covariance matrix is:

$$E[(X_{n+h} - P_n X_{n+h})(X_{n+h} - P_n X_{n+h})'] = \sum_{j=0}^{h-1} \Psi_j \Sigma \Psi_j' \tag{81}$$

Example 7.6.3: Dow Jones and All Ordinaries Indices Compares predictors for X_{t2} (All Ordinaries return): 1. Sample mean $\hat{\mu}_2$. 2. Univariate AR(1) model fitted to $\{X_{t2}\}$. 3. Bivariate VAR(1) model fitted to $X_t = (X_{t1}, X_{t2})'$. Using data $t = 1, \ldots, 250$ to fit, and predicting for $t = 251, \ldots, 290$:

- Model MSEs (from fit): $\hat{\gamma}_{22}(0) = 0.771$, AR(1)=0.760, VAR(1)=0.602.
- Actual Avg Sq Errors (on $t = 251..290$): SampleMean=0.471, AR(1)=0.459, VAR(1)=0.396.

The VAR(1) model provides the best forecasts, demonstrating the benefit of using the Dow Jones return (X_{t1}) to predict the All Ordinaries return (X_{t2}).

Forecasting Multivariate ARIMA/SARIMA If Y_t is nonstationary and $U_t = D(B)Y_t$ is a stationary ARMA process $U_t = \mu + X_t$, where $\Phi(B)X_t = \Theta(B)Z_t$. Let $D(z) = 1 - \sum d_j z^j$. Predictors $\tilde{P}_n Y_{n+h}$ are found recursively:

$$\tilde{P}_n Y_{n+h} = P_n U_{n+h} + \sum_{j=1}^{r} d_j \tilde{P}_n Y_{n+h-j} \tag{82}$$

where $P_n U_{n+h}$ is the predictor for the ARMA process U_t. The error covariance matrix (approximate for large n if $q > 0$, exact for AR models if $n \geq p$) is:

$$\approx \sum_{j=0}^{h-1} \Psi_j^* \Sigma \Psi_j^{*'} \tag{83}$$

where $\Psi^*(z) = D(z)^{-1}\Phi(z)^{-1}\Theta(z) = \sum \Psi_j^* z^j$.

29.6 Cointegration

Addresses nonstationary multivariate series where components tend to "move together".

- A univariate series $\{X_t\}$ is **integrated of order d**, $I(d)$, if $\nabla^d X_t$ is stationary but $\nabla^{d-1} X_t$ is nonstationary.
- A k-variate series $\{X_t\}$ is $I(d)$ if $\nabla^d X_t$ is stationary but $\nabla^{d-1} X_t$ is not.
- An $I(d)$ process $\{X_t\}$ is **cointegrated** with cointegration vector α ($\alpha \neq 0$) if the linear combination $\alpha' X_t$ is integrated of order less than d (often $d-1$, typically $I(0)$ when $d = 1$).

Example 7.7.1 Let $X_t = \sum_{j=1}^{t} Z_j$ be a random walk ($I(1)$). Let $Y_t = X_t + W_t$, where W_t is $I(0)$ noise independent of Z_t. Then Y_t is also $I(1)$. The vector process $(X_t, Y_t)'$ is $I(1)$. The linear combination $\alpha'(X_t, Y_t)'$ with $\alpha = (1, -1)'$ is $X_t - Y_t = -W_t$, which is $I(0)$. So $(X_t, Y_t)'$ is cointegrated with vector $(1, -1)'$. This captures long-run equilibrium relationships between nonstationary series.

Example 7.7.2 Differencing the cointegrated process from Example 7.7.1 gives $U_t = \nabla X_t = Z_t$ and $V_t = \nabla Y_t = Z_t + W_t - W_{t-1}$. The vector $(U_t, V_t)'$ is a stationary vector MA(1) process. However, its MA matrix polynomial $\Theta(z)$ has $\det \Theta(1) = 0$, meaning the differenced process is non-invertible. This non-invertibility is characteristic of differenced cointegrated systems and relates to Error Correction Models (ECMs). Estimation requires specialized techniques beyond standard VAR/ARMA methods.

30 State-Space Models

State-space representations offer a powerful and flexible framework for time series analysis, encompassing linear ARIMA models, classical decomposition, and more complex structural models. They originated in control theory. This chapter introduces the general state-space model, structural models, ARIMA representations in state-space form, the crucial Kalman recursions for prediction and estimation, handling missing data, the EM algorithm, and generalized state-space models.

Notation: $\{W_t\} \sim WN(0, \{R_t\})$ means $EW_t = 0$ and $E(W_s W_t') = R_t$ if $s = t$, and 0 otherwise.

30.1 State-Space Representations

A state-space model for a w-dimensional observation sequence $\{Y_t, t = 1, 2, \dots\}$ involves an unobserved v-dimensional state vector X_t.

Observation Equation: Relates the observation Y_t to the current state X_t.

$$Y_t = G_t X_t + W_t, \quad t = 1, 2, \dots \tag{84}$$

where $\{W_t\} \sim WN(0, \{R_t\})$ is the observation noise, and G_t is a $w \times v$ matrix.

State Equation: Describes the evolution of the state vector over time.

$$X_{t+1} = F_t X_t + V_t, \quad t = 1, 2, \dots \tag{85}$$

where $\{V_t\} \sim WN(0, \{Q_t\})$ is the state noise, F_t is a $v \times v$ transition matrix, and $\{V_t\}$ is uncorrelated with $\{W_t\}$. The initial state X_1 is uncorrelated with all noise terms.

Remarks

- More general forms allow V_t, W_t correlation and control terms.
- Often, matrices F_t, G_t, Q_t, R_t are time-invariant (F, G, Q, R).
- X_t depends on X_1 and $V_1, \dots, V_{t-1}$. Y_t depends on $X_1, V_1, \dots, V_{t-1}, W_t$.
- V_t is uncorrelated with X_s, Y_s for $s \le t$. W_t is uncorrelated with X_s for $s \le t$ and Y_s for $s < t$.
- If noise terms are independent, $\{X_t\}$ is a Markov process.

Definition 8.1.1 A time series $\{Y_t\}$ has a state-space representation if such a model exists.

Example 8.1.1: AR(1) Process For $Y_t = \phi Y_{t-1} + Z_t$, $\{Z_t\} \sim WN(0, \sigma^2)$. Let the state be $X_t = Y_t$. State eq: $X_{t+1} = \phi X_t + V_t$, where $V_t = Z_{t+1}$. $(F = \phi, Q = \sigma^2)$ Obs eq: $Y_t = X_t$. $(G = 1, W_t = 0, R = 0)$

Example 8.1.2: ARMA(1,1) Process For $Y_t = \phi Y_{t-1} + Z_t + \theta Z_{t-1}$. Let U_t be the AR(1) process $\phi(B)U_t = Z_t$. Then $Y_t = \theta(B)U_t = U_t + \theta U_{t-1}$. Let the state be $X_t = (U_{t-1}, U_t)'$. State eq: $X_{t+1} = \begin{bmatrix} 0 & 1 \\ 0 & \phi \end{bmatrix} X_t + \begin{bmatrix} 0 \\ 1 \end{bmatrix} Z_{t+1}$. $(F = \begin{bmatrix} 0 & 1 \\ 0 & \phi \end{bmatrix}, V_t = (0, Z_{t+1})', Q = \begin{bmatrix} 0 & 0 \\ 0 & \sigma^2 \end{bmatrix})$ Obs eq: $Y_t = \begin{bmatrix} \theta & 1 \end{bmatrix} X_t$. $(G = [\theta\ 1], W_t = 0, R = 0)$

State-Space Models with $t \in \{0, \pm 1, \dots\}$ For stationary models, it's often natural to assume the equations hold for all integer t:

$$Y_t = GX_t + W_t, \quad \{W_t\} \sim WN(0, R) \tag{86}$$
$$X_{t+1} = FX_t + V_t, \quad \{V_t\} \sim WN(0, Q) \tag{87}$$

where F, G, Q, R are constant matrices, and $E(V_s W_t') = 0$. The state equation is **stable** if all eigenvalues of F are inside the unit circle ($\det(I - Fz) \neq 0$ for $|z| \leq 1$). In the stable case, the unique stationary solution is:

$$X_t = \sum_{j=0}^{\infty} F^j V_{t-j-1} \tag{88}$$

$$Y_t = W_t + \sum_{j=0}^{\infty} GF^j V_{t-j-1} \tag{89}$$

Both $\{X_t\}$ and $\{Y_t\}$ are stationary.

30.2 The Basic Structural Model

Structural models define Y_t directly in terms of unobserved components like trend (M_t), seasonal (S_t), and irregular (W_t), allowing these components to evolve randomly.

Example 8.2.1: Random Walk Plus Noise Model (Local Level Model) Observation eq: $Y_t = M_t + W_t$, $\{W_t\} \sim WN(0, \sigma_w^2)$. State eq: $M_{t+1} = M_t + V_t$, $\{V_t\} \sim WN(0, \sigma_v^2)$. Here, the state $X_t = M_t$. $F = 1, G = 1, Q = \sigma_v^2, R = \sigma_w^2$. The text shows a realization (Figure 8.1) where the underlying random walk M_t (level) changes over time, and the observations Y_t fluctuate around it. The differenced series $\nabla Y_t = V_{t-1} + W_t - W_{t-1}$ is stationary. It is an MA(1) process:

$$\nabla Y_t = Z_t + \theta Z_{t-1}, \quad \{Z_t\} \sim WN(0, \sigma^2)$$

where θ and σ^2 are determined by σ_v^2, σ_w^2 via:

$$\frac{\theta}{1+\theta^2} = \frac{-\sigma_w^2}{2\sigma_w^2 + \sigma_v^2}, \quad \theta\sigma^2 = -\sigma_w^2 \tag{90}$$

The unique solution with $|\theta| \leq 1$ exists. Thus, the Random Walk Plus Noise model is equivalent to an ARIMA(0,1,1) process. The sample ACF of the differenced data (Figure 8.2) for the realization in Figure 8.1 shows a significant negative spike only at lag 1, consistent with an MA(1). Testing if the level M_t is constant is equivalent to testing $\sigma_v^2 = 0$, which corresponds to $\theta = -1$ (a unit root in the MA polynomial) for the differenced series.

Local Linear Trend Model Extends the local level model by allowing a randomly varying slope B_t. Obs eq: $Y_t = M_t + W_t$, $\{W_t\} \sim WN(0, \sigma_w^2)$. State eqs:

$$M_{t+1} = M_t + B_t + V_t, \quad \{V_t\} \sim WN(0, \sigma_v^2)$$
$$B_{t+1} = B_t + U_t, \quad \{U_t\} \sim WN(0, \sigma_u^2)$$

State vector $X_t = (M_t, B_t)'$. State eq: $X_{t+1} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} X_t + \begin{bmatrix} V_t \\ U_t \end{bmatrix}$. ($F = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, State noise is $(V_t, U_t)'$, $Q = \text{diag}(\sigma_v^2, \sigma_u^2)$ if V_t, U_t uncorrelated). Obs eq: $Y_t = \begin{bmatrix} 1 & 0 \end{bmatrix} X_t + W_t$. ($G = [1\ 0], R = \sigma_w^2$).

Example 8.2.2: Seasonal Series with Noise Generalizes the deterministic seasonal component s_t satisfying $\sum_{j=0}^{d-1} s_{t-j} = 0$ (or $s_{t+1} = -s_t - \cdots - s_{t-d+2}$). Let the random seasonal component Y_t satisfy:

$$Y_{t+1} = -Y_t - \cdots - Y_{t-d+2} + S_t, \quad \{S_t\} \sim WN(0, \sigma_s^2)$$

State vector $X_t = (Y_t, Y_{t-1}, \ldots, Y_{t-d+2})'$, dimension $d-1$. Obs eq: $Y_t = \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix} X_t$. State eq: $X_{t+1} = FX_t + V_t$, where $V_t = (S_t, 0, \ldots, 0)'$ and

$$F = \begin{bmatrix} -1 & -1 & \ldots & -1 \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \end{bmatrix}$$

(F is $(d-1) \times (d-1)$ with 1s on first subdiagonal).

Example 8.2.3: Trend + Seasonality + Noise Combine the local linear trend and random seasonal models by stacking state vectors. Let X_t^1 be the state for trend (dim 2), X_t^2 be the state for seasonality (dim $d-1$). Combined state $X_t = (X_t^{1'}, X_t^{2'})'$. State eq: $X_{t+1} = \begin{bmatrix} F_1 & 0 \\ 0 & F_2 \end{bmatrix} X_t + \begin{bmatrix} V_t^1 \\ V_t^2 \end{bmatrix}$. ($F = \text{blockdiag}(F_1, F_2)$, $V_t = (V_t^{1'}, V_t^{2'})'$) Obs eq: $Y_t = \begin{bmatrix} G_1 & G_2 \end{bmatrix} X_t + W_t$. ($G = [1\ 0\ 1\ 0 \ldots 0]$)

30.3 State-Space Representation of ARIMA Models

Example 8.3.1: Causal AR(p) Process $Y_{t+1} = \sum_{j=1}^{p} \phi_j Y_{t+1-j} + Z_{t+1}$. State vector $X_t = (Y_{t-p+1}, \ldots, Y_t)'$. Obs eq: $Y_t = \begin{bmatrix} 0 & \ldots & 0 & 1 \end{bmatrix} X_t$. State eq: $X_{t+1} = FX_t + V_t$, where $V_t = (0, \ldots, 0, Z_{t+1})'$ and

$$F = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \\ \phi_p & \phi_{p-1} & \phi_{p-2} & \ldots & \phi_1 \end{bmatrix}$$

(This is the companion matrix form). Causality ($\phi(z) \neq 0$ for $|z| \leq 1$) is equivalent to stability of F (eigenvalues inside unit circle).

Example 8.3.2: Causal ARMA(p,q) Process $\phi(B)Y_t = \theta(B)Z_t$. Let $r = \max(p, q+1)$. Define $\phi_j = 0$ for $j > p$, $\theta_j = 0$ for $j > q$, $\theta_0 = 1$. Let U_t be the AR(p) process $\phi(B)U_t = Z_t$. Then $Y_t = \theta(B)U_t = \sum_{j=0}^{q} \theta_j U_{t-j}$. State vector $X_t = (U_{t-r+1}, \ldots, U_t)'$. State eq: $X_{t+1} = FX_t + V_t$, where F is the $r \times r$ companion matrix for $\phi(z)$ (extended with zeros if $r > p$) and $V_t = (0, \ldots, 0, Z_{t+1})'$. Obs eq: $Y_t = \begin{bmatrix} \theta_{r-1} & \theta_{r-2} & \ldots & \theta_1 & \theta_0 \end{bmatrix} X_t$.

Example 8.3.3: ARIMA(p,d,q) Process Let $W_t = \nabla^d Y_t$ be a causal ARMA(p,q). W_t has a state-space representation $W_t = G_W X_t^W$, $X_{t+1}^W = F_W X_t^W + V_t^W$. We need to augment the state to include lagged values of Y_t. Let $Y_{t-1}^{(d)} = (Y_{t-d}, \ldots, Y_{t-1})'$. Using $Y_t = W_t - \sum_{j=1}^{d} \binom{d}{j} (-1)^j Y_{t-j}$, we can write $Y_t = G_W X_t^W + \text{linear combination of } Y_{t-1}^{(d)}$. Define combined state $X_t = (X_t^{W'}, Y_{t-1}^{(d)'})'$. We can derive matrices F, G, Q for the model $X_{t+1} = FX_t + V_t$, $Y_t = GX_t$. The initial state $Y_0^{(d)}$ is assumed uncorrelated with future noise.

30.4 The Kalman Recursions

These recursions solve three key problems for the state-space model (8.1.1)-(8.1.2), assuming known matrices F_t, G_t, Q_t, R_t and initial state X_1 properties $(EX_1, Var(X_1))$. Let $Y^{(t)} = (Y_0, Y_1, \ldots, Y_t)$. We want minimum mean square error linear estimates $P_t(\cdot)$ based on $Y^{(t)}$.

- **Prediction:** Estimate X_t based on $Y^{(t-1)}$. Notation: $\hat{X}_t = P_{t-1}(X_t)$. Error covariance: $\Omega_t = E[(X_t - \hat{X}_t)(X_t - \hat{X}_t)']$.
- **Filtering:** Estimate X_t based on $Y^{(t)}$. Notation: $X_{t|t} = P_t(X_t)$. Error covariance: $\Omega_{t|t} = E[(X_t - X_{t|t})(X_t - X_{t|t})']$.
- **Smoothing:** Estimate X_t based on $Y^{(n)}$ with $n > t$. Notation: $X_{t|n} = P_n(X_t)$. Error covariance: $\Omega_{t|n} = E[(X_t - X_{t|n})(X_t - X_{t|n})']$.

Kalman Prediction Initial conditions: $\hat{X}_1 = P(X_1|Y_0)$, $\Omega_1 = E[(X_1 - \hat{X}_1)(X_1 - \hat{X}_1)']$. Recursions for $t = 1, 2, \ldots$: Define innovation $I_t = Y_t - P_{t-1}Y_t = Y_t - G_t\hat{X}_t$. Innovation covariance $\Delta_t = E(I_t I_t') = G_t\Omega_t G_t' + R_t$. Intermediate term $\Theta_t = E(X_{t+1}I_t') = F_t\Omega_t G_t'$. Update state prediction:

$$\hat{X}_{t+1} = F_t\hat{X}_t + \Theta_t\Delta_t^{-1}I_t = F_t\hat{X}_t + \Theta_t\Delta_t^{-1}(Y_t - G_t\hat{X}_t) \tag{91}$$

Update error covariance prediction:

$$\Omega_{t+1} = F_t\Omega_t F_t' + Q_t - \Theta_t\Delta_t^{-1}\Theta_t' \tag{92}$$

(Δ_t^{-1} is any generalized inverse).

Kalman Filtering Update predicted state to filtered state using current observation Y_t:

$$X_{t|t} = \hat{X}_t + \Omega_t G_t'\Delta_t^{-1}(Y_t - G_t\hat{X}_t) \tag{93}$$

Update predicted error covariance to filtered error covariance:

$$\Omega_{t|t} = \Omega_t - \Omega_t G_t'\Delta_t^{-1}G_t\Omega_t \tag{94}$$

Kalman Fixed-Point Smoothing Updates the estimate of X_t as more future data $Y_{t+1}, \ldots, Y_n$ becomes available. For fixed t, iterate for $n = t, t+1, \ldots$: Initial conditions: $X_{t|t-1} = \hat{X}_t$, $\Omega_{t|t-1} = \Omega_t$. $\Omega_{t,t} = \Omega_t$. Recursions:

$$X_{t|n} = X_{t|n-1} + \Omega_{t,n}G_n'\Delta_n^{-1}(Y_n - G_n\hat{X}_n) \tag{95}$$
$$\Omega_{t,n+1} = \Omega_{t,n}[F_n - \Theta_n\Delta_n^{-1}G_n]' \tag{96}$$
$$\Omega_{t|n} = \Omega_{t|n-1} - \Omega_{t,n}G_n'\Delta_n^{-1}G_n\Omega_{t,n}' \tag{97}$$

where $\Omega_{t,n} = E[(X_t - \hat{X}_t)(X_n - \hat{X}_n)']$.

h-step Prediction of Y_t Predictors P_tY_{t+h} and MSE matrices $\Delta_t^{(h)}$ can be found using the predicted states:

$$P_tX_{t+1} = F_tX_{t|t} \tag{98}$$
$$P_tX_{t+h} = F_{t+h-1}P_tX_{t+h-1} = (F_{t+h-1}\ldots F_{t+1})P_tX_{t+1}, \quad h \geq 2 \tag{99}$$
$$P_tY_{t+h} = G_{t+h}P_tX_{t+h}, \quad h \geq 1 \tag{100}$$

Error covariances $\Omega_t^{(h)} = E[(X_{t+h} - P_t X_{t+h})(X_{t+h} - P_t X_{t+h})']$ satisfy:

$$\Omega_t^{(1)} = \Omega_{t+1} \tag{101}$$

$$\Omega_t^{(h)} = F_{t+h-1}\Omega_t^{(h-1)}F'_{t+h-1} + Q_{t+h-1}, \quad h \geq 2 \tag{102}$$

And $\Delta_t^{(h)} = G_{t+h}\Omega_t^{(h)}G'_{t+h} + R_{t+h}$.

Example 8.4.1: Random Walk Plus Noise and Exponential Smoothing Applying Kalman prediction to $Y_t = X_t + W_t, X_{t+1} = X_t + V_t$ yields:

$$\hat{Y}_{t+1} = P_t Y_{t+1} = \hat{X}_{t+1} = \hat{X}_t + \frac{\Omega_t}{\Omega_t + \sigma_w^2}(Y_t - \hat{X}_t)$$

Let $a_t = \Omega_t/(\Omega_t+\sigma_w^2)$. Then $\hat{X}_{t+1} = (1-a_t)\hat{X}_t + a_t Y_t$. This is exponential smoothing with time-varying coefficient a_t. The error covariance Ω_t converges to a steady-state value Ω satisfying $\Omega = \Omega + \sigma_v^2 - \Omega^2/(\Omega + \sigma_w^2)$. The limiting value $a = \Omega/(\Omega + \sigma_w^2)$ corresponds to the standard exponential smoothing coefficient, and relates to the MA(1) parameter θ of the differenced series ∇Y_t via $a = 1 + \theta$.

30.5 Estimation For State-Space Models

If the model matrices F_t, G_t, Q_t, R_t depend on a parameter vector θ, we estimate θ by maximizing the Gaussian likelihood of the observations $Y_1, \ldots, Y_n$. Using the innovations $I_j = Y_j - P_{j-1}Y_j$ and their covariances Δ_j (computed via Kalman prediction):

$$L(\theta; Y_1, \ldots, Y_n) = (2\pi)^{-nw/2}\left(\prod_{j=1}^{n} \det \Delta_j\right)^{-1/2} \exp\left\{-\frac{1}{2}\sum_{j=1}^{n} I_j'\Delta_j^{-1}I_j\right\} \tag{103}$$

This likelihood can be numerically maximized w.r.t. θ.

Application to Structural Models For models like (8.5.3)-(8.5.4) with unknown initial state mean $\mu = X_1$ and unknown noise variances in $Q = \sigma_w^2 Q^*$ and $R = \sigma_w^2$, a stepwise maximization is possible. Maximize likelihood w.r.t. μ and σ_w^2 for fixed Q^*, then maximize the resulting "reduced likelihood" w.r.t. Q^*. Formulas for $\hat{\mu}(Q^*)$ and $\hat{\sigma}_w^2(Q^*)$ and the reduced likelihood $l(Q^*)$ are given.

Example 8.5.1: Random walk plus noise Fitting the model to the data from Fig 8.1. Estimates $\hat{\mu} = 0.906$, $\hat{\sigma}_v^2 = 5.351$, $\hat{\sigma}_w^2 = 8.233$ are obtained, close to true values (0, 4, 8).

Example 8.5.2: Airline passengers The airline data (Fig 8.3) show trend and seasonality, with increasing variance. A structural model combining local linear trend and random seasonality (Example 8.2.3) is fitted to the raw data. Parameter estimation yields variance estimates for trend level, slope, seasonal component, and observation noise. The estimated slope variance is near zero, suggesting a near-constant slope. Figures 8.4 and 8.5 (summarized in text) show the predicted components and the overall one-step predictions versus actual data, indicating a good fit.

30.6 State-Space Models with Missing Observations

The state-space framework handles missing data naturally.

Likelihood Calculation Suppose we observe $Y_{i_1}, \ldots, Y_{i_r}$ where $1 \leq i_1 < \cdots < i_r \leq n$. Define a modified observation sequence Y_t^* and model:

$$Y_t^* = G_t^* X_t + W_t^*$$
$$X_{t+1} = F_t X_t + V_t$$

where $G_t^* = G_t$ if $t \in \{i_1, \ldots, i_r\}$ and 0 otherwise. $W_t^* = W_t$ if $t \in \{i_1, \ldots, i_r\}$ and independent $N(0, I)$ noise otherwise. Define observed sequence $y_t^* = y_t$ if $t \in \{i_1, \ldots, i_r\}$ and 0 otherwise. The likelihood of the original sparse observations $L_1(\theta; y_{i_1}, \ldots, y_{i_r})$ is related to the likelihood $L_2(\theta; y_1^*, \ldots, y_n^*)$ of the modified full sequence under the modified model:

$$L_1 = (2\pi)^{(n-r)w/2} L_2$$

L_2 can be computed using the Kalman prediction recursions applied to the modified model with the observation sequence $y_1^*, \ldots, y_n^*$. The terms in the likelihood sum corresponding to missing times ($t \notin \{i_1, \ldots, i_r\}$) will have $\Delta_t^* \approx I$ and $I_t^* \approx 0$ (if y_t^* is set to 0), contributing negligibly or constant factors.

Example 8.6.1: AR(1) with missing Y_2 Shows derivation of the likelihood $L(\phi, \sigma^2; y_1, y_3, y_4, y_5)$ using the modified model approach.

Estimation of Missing Values The minimum mean square estimate of a missing Y_t given observations $Y_{obs} = \{Y_{i_1}, \ldots, Y_{i_r}\}$ is $P(Y_t | Y_0, Y_{obs})$. This can be found using the modified model for Y_t^*:

$$P(X_t | Y_0, Y_{obs}) = P(X_t | Y_0^*, Y_1^*, \ldots, Y_n^*)$$

The right side is the smoothed state estimate $X_{t|n}$ obtained by applying Kalman smoothing to the modified model using the sequence $y_1^*, \ldots, y_n^*$. Then, the estimate of Y_t is:

$$P(Y_t | Y_0, Y_{obs}) = G_t P(X_t | Y_0^*, \ldots, Y_n^*) = G_t X_{t|n}$$

Example 8.6.2: AR(1) with missing Y_2 Applies Kalman smoothing to the modified model to find the estimate of X_2 given y_1, y_3, y_4, y_5. Result: $\hat{Y}_2 = \phi(Y_1 + Y_3)/(1 + \phi^2)$.

Direct Estimation for AR Models For Gaussian AR(p) models, maximizing the joint density of observed and missing values w.r.t. missing values yields the conditional expectation (best estimate). This is equivalent to minimizing $\sum_{t=p+1}^{n} Z_t^2 = \sum (Y_t - \sum \phi_j Y_{t-j})^2$. For AR(1) with missing Y_2, minimizing $(Y_2 - \phi Y_1)^2 + (Y_3 - \phi Y_2)^2$ w.r.t Y_2 gives the same result as Example 8.6.2.

30.7 The EM Algorithm

An iterative method for finding MLEs when data is incomplete (e.g., missing observations, or unobserved states in state-space models). Let $W = (X, Y)$ be complete data (X unobserved, Y observed). Let $l(\theta; W) = \ln f(W; \theta)$ be the complete data log-likelihood. Iteration $i + 1$ from estimate $\theta^{(i)}$:

- **E-step (Expectation):** Compute $Q(\theta|\theta^{(i)}) = E_{\theta^{(i)}}[l(\theta; X, Y)|Y]$. This is the expected complete data log-likelihood, averaging over the distribution of missing data X given observed Y, using current parameters $\theta^{(i)}$.
- **M-step (Maximization):** Find $\theta^{(i+1)}$ that maximizes $Q(\theta|\theta^{(i)})$.

The observed data likelihood $L(\theta^{(i)}; Y)$ is non-decreasing. Limit points are typically stationary points of the likelihood. Useful when maximizing Q is much easier than maximizing $L(\theta; Y)$ directly.

Missing Data Application If $W = (X', Y')'$ is Gaussian with mean 0, covariance $\Sigma(\theta)$. X is missing, Y observed. Let $\hat{X} = E_{\theta^{(i)}}(X|Y)$ and $\Sigma_{X|Y}(\theta^{(i)})$ be the conditional mean and covariance. $Q(\theta|\theta^{(i)}) = E_{\theta^{(i)}}[-\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln\det\Sigma(\theta) - \frac{1}{2}W'\Sigma(\theta)^{-1}W|Y]$. A simplified/modified version often used replaces Q with $\tilde{Q}(\theta|\theta^{(i)}) = l(\theta; \hat{W})$, where $\hat{W} = (\hat{X}', Y')'$. E-step: Estimate missing values X using current $\theta^{(i)}$ (e.g., Kalman smoother). M-step: Maximize the complete data likelihood using the observed Y and estimated X (treat $\hat{X}$ as observed).

Example 8.7.1: Lake data with missing values Fit AR(2) to lake data with 10 artificial missing values. Iter 0: $\phi^{(0)} = (0, 0)$. E-step: $\hat{W}_{missing} = 0$. M-step: Fit AR(2) to filled data $\to \phi^{(1)}$. Iter 1: E-step: Estimate missing W using $\phi^{(1)}$. M-step: Fit AR(2) to new filled data $\to \phi^{(2)}$. ... Converges quickly.

30.8 Generalized State-Space Models

Extend the linear Gaussian framework. Two types: parameter-driven and observation-driven.

30.8.1 Parameter-Driven Models

State X_t evolves independently of past observations $Y^{(t-1)}$. Assume conditional independence:

- $Y_t|(X_t, X^{(t-1)}, Y^{(t-1)})$ depends only on X_t, via density $p(y_t|x_t)$.
- $X_{t+1}|(X_t, X^{(t-1)}, Y^{(t)})$ depends only on X_t, via density $p(x_{t+1}|x_t)$ (Markov property for states).

Initial state density $p_1(x_1)$. Joint density: $p(y^{(n)}, x^{(n)}) = \left(\prod_{j=1}^{n} p(y_j|x_j)\right)\left(\prod_{j=2}^{n} p(x_j|x_{j-1})\right)p_1(x_1)$. $Y_1, \ldots, Y_n$ are conditionally independent given $X_1, \ldots, X_n$.

Filtering/Prediction via conditional densities:

$$p(x_t|y^{(t)}) \propto p(y_t|x_t)p(x_t|y^{(t-1)}) \quad \text{(Filtering update)} \tag{104}$$

$$p(x_{t+1}|y^{(t)}) = \int p(x_{t+1}|x_t)p(x_t|y^{(t)})d\mu(x_t) \quad \text{(Prediction update)} \tag{105}$$

Start with $p(x_1|y^{(0)}) = p_1(x_1)$. These update the entire distribution, not just mean/covariance. Estimates are conditional means, e.g., $E(X_t|Y^{(t)})$.

Example 8.8.1: Linear Gaussian model revisited If $p(y_t|x_t)$ and $p(x_{t+1}|x_t)$ are Gaussian, then $p(x_t|y^{(t)})$ and $p(x_t|y^{(t-1)})$ are also Gaussian. The means and variances satisfy the Kalman filter recursions.

Example 8.8.2: Non-Gaussian model State: $X_t = aX_{t-1}$ (deterministic given X_{t-1}). Obs: $Y_t|X_t \sim \text{Poisson}(\pi X_t)$. Initial state $X_1 \sim \text{Gamma}(\alpha, \lambda)$. The recursions show that if $p(x_t|y^{(t)})$ is Gamma, then $p(x_{t+1}|y^{(t)})$ is also Gamma (scaled), and $p(x_{t+1}|y^{(t+1)})$ is Gamma. Explicit formulas for parameters are derived. Prediction density $p(y_{t+1}|y^{(t)})$ is Negative Binomial.

Example 8.8.3: Model for time series of counts (Polio data) Obs: $Y_t|X_t \sim \text{Poisson}(e^{X_t})$. State: $X_t = \beta' u_t + W_t$, where $W_t = \phi W_{t-1} + Z_t$, $\{Z_t\} \sim N(0, \sigma^2)$. u_t are covariates (trend, seasonals). Transition density $p(x_{t+1}|x_t)$ is Gaussian. Likelihood is intractable. Monte Carlo EM (MCEM) used for estimation: E-step: Approximate $Q(\theta|\theta^{(i)}) = E_{\theta^{(i)}}[\ln L(\theta; X^{(n)})|Y^{(n)}]$ using simulations (e.g., Gibbs sampler) of $X^{(n)}$ from $p(x^{(n)}|y^{(n)}; \theta^{(i)})$. M-step: Maximize the approximated Q (similar to fitting regression with AR(1) errors). Analysis of polio data finds significant seasonal components and slight downward trend.

30.8.2 Observation-Driven Models

State evolution depends directly on past observations. Assume conditional independence:

- $Y_t|(X_t, X^{(t-1)}, Y^{(t-1)})$ depends only on X_t, via $p(y_t|x_t)$.
- State X_{t+1} given $Y^{(t)}$ has density $p(x_{t+1}|y^{(t)})$, specified directly.

Filtering is simple: $p(x_t|y^{(t)}) \propto p(y_t|x_t)p(x_t|y^{(t-1)})$. Prediction is given by the state eq. Likelihood $p(y^{(n)}) = \prod p(y_t|y^{(t-1)})$ is easier to compute. Model specification might be ambiguous regarding the underlying state process unless additional assumptions are made (e.g., X_{t+1} depends only on $Y^{(t)}$).

Exponential Family Models If $p(y_t|x_t) = \exp\{y_t x_t - b(x_t) + c(y_t)\}$ (natural parameter x_t), and if the state density $p(x_{t+1}|y^{(t)})$ belongs to the conjugate prior family $f(x; \alpha, \lambda) = \exp\{\alpha x - \lambda b(x) + A(\alpha, \lambda)\}$, then the posterior $p(x_t|y^{(t)})$ is also in this family. Parameters update: $\lambda_t = 1 + \lambda_{t|t-1}$, $\alpha_t = y_t + \alpha_{t|t-1}$. One-step predictor $\hat{Y}_{t+1} = E(Y_{t+1}|y^{(t)}) = \alpha_{t+1|t}/\lambda_{t+1|t}$. If a "power steady model" assumption is made: $\lambda_{t+1|t} = \delta\lambda_t$, $\alpha_{t+1|t} = \delta\alpha_t$ $(0 < \delta < 1)$, then predictors follow exponential smoothing $\hat{Y}_{t+1} = (1-\delta)y_t + \delta\hat{Y}_t$. However, these models can have degenerate long-term behavior.

Example 8.8.7: Goals scored by England vs Scotland Count data (goals per match). Data shows overdispersion (variance > mean) compared to Poisson, and dependence (transition probabilities vary). Fit Poisson observation model with log-gamma state density, using power steady update ($\lambda_{t+1|t} = \delta\lambda_t, \alpha_{t+1|t} = \delta\alpha_t$). Estimate δ by maximizing conditional likelihood (product of negative binomial prediction densities). Result $\hat{\delta} = 0.844$. One-step predictors resemble exponential smoothing.

31 Forecasting Techniques

This chapter discusses three forecasting techniques that place less emphasis on explicit model construction compared to ARIMA modeling. They select an optimal algorithm from a limited class based on specified criteria and are effective for various real data sets.

31.1 The ARAR Algorithm

This algorithm adapts the ARARMA approach. It applies automatically selected "memory-shortening" transformations if the data is deemed "long-memory", and then fits a subset autoregressive (AR) model to the transformed series.

31.1.1 Memory Shortening

Given data $\{Y_t, t = 1, \ldots, n\}$, the algorithm classifies the series memory and applies transformations if needed. The allowed transformations are:

$$\tilde{Y}_t = Y_t - \hat{\phi}(\hat{\tau})Y_{t-\hat{\tau}} \tag{106}$$

$$\tilde{Y}_t = Y_t - \hat{\phi}_1 Y_{t-1} - \hat{\phi}_2 Y_{t-2} \tag{107}$$

The classification algorithm is:

1. For each $\tau = 1, \ldots, 15$, find $\hat{\phi}(\tau)$ minimizing $ERR(\phi, \tau) = \frac{\sum_{t=\tau+1}^{n}[Y_t - \phi Y_{t-\tau}]^2}{\sum_{t=\tau+1}^{n} Y_t^2}$. Define $Err(\tau) = ERR(\hat{\phi}(\tau), \tau)$. Choose $\hat{\tau}$ that minimizes $Err(\tau)$.
2. If $Err(\hat{\tau}) \leq 8/n$, classify as **L** (Long-memory) and use transformation (106).
3. If $\hat{\phi}(\hat{\tau}) \geq 0.93$ and $\hat{\tau} > 2$, classify as **L** and use transformation (106).
4. If $\hat{\phi}(\hat{\tau}) \geq 0.93$ and $\hat{\tau} \in \{1, 2\}$, find $\hat{\phi}_1, \hat{\phi}_2$ minimizing $\sum_{t=3}^{n}[Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2}]^2$. Classify as **M** (Moderately long-memory) and use transformation (107).
5. If $\hat{\phi}(\hat{\tau}) < 0.93$, classify as **S** (Short-memory). No transformation needed at this step.

If L or M is chosen, the transformed series $\{\tilde{Y}_t\}$ is re-checked. This continues until the series is classified as S (at most 3 transformations).

31.1.2 Fitting a Subset Autoregression

Let $\{S_t, t = k + 1, \ldots, n\}$ be the final memory-shortened series (or original series if classified S initially). Let $\overline{S}$ be its sample mean. Fit a subset AR model to the mean-corrected series $X_t = S_t - \overline{S}$. The model form is:

$$X_t = \phi_1 X_{t-1} + \phi_{l_1} X_{t-l_1} + \phi_{l_2} X_{t-l_2} + \phi_{l_3} X_{t-l_3} + Z_t, \quad \{Z_t\} \sim WN(0, \sigma^2) \tag{108}$$

The lags l_1, l_2, l_3 (where $1 < l_1 < l_2 < l_3 \leq m$, with $m = 13$ or 26) and coefficients ϕ_j are chosen to minimize the Yule-Walker estimate of σ^2 (or maximize Gaussian likelihood). The Yule-Walker estimates are found by solving the appropriate system based on sample autocorrelations $\hat{\rho}(\cdot)$ and autocovariances $\hat{\gamma}(\cdot)$ of $\{X_t\}$.

31.1.3 Forecasting

Let the memory-shortening filter be $\psi(B) = 1+\psi_1 B+\cdots+\psi_k B^k$ (where $k = 0$ if no shortening). Let the subset AR filter be $\phi(B) = 1 - \phi_1 B - \phi_{l_1} B^{l_1} - \phi_{l_2} B^{l_2} - \phi_{l_3} B^{l_3}$. The overall model for the original series $\{Y_t\}$ is:

$$\xi(B)Y_t = \phi(1)\overline{S} + Z_t \tag{109}$$

where $\xi(B) = \psi(B)\phi(B) = 1 + \xi_1 B + \cdots + \xi_{k+l_3} B^{k+l_3}$. Forecasts $P_n Y_{n+h}$ for $n > k + l_3$ are computed recursively:

$$P_n Y_{n+h} = -\sum_{j=1}^{k+l_3} \xi_j P_n Y_{n+h-j} + \phi(1)\overline{S}, \quad h \geq 1 \tag{110}$$

with $P_n Y_{n+h} = Y_{n+h}$ for $h \leq 0$. The mean squared error is:

$$E[(Y_{n+h} - P_n Y_{n+h})^2] = \sum_{j=0}^{h-1} \tau_j^2 \sigma^2 \tag{111}$$

where τ_j are coefficients of $1/\xi(z) = \sum_{j=0}^{\infty} \tau_j z^j$, found from $\tau_0 = 1$ and $\sum_{j=0}^{m} \tau_j \xi_{m-j} = 0$ for $m \geq 1$.

Example 9.1.1: Accidental Deaths Data Applying ARAR to DEATHS.TSM (Optimize based on WN variance, max lag 26). Memory-shortening filter: $(1-0.9779B^{12})$. Subset AR fit to shortened, mean-corrected series: Lags (1, 3, 12, 13), Coeffs (0.5915, -0.3822, -0.3022, 0.2970), $\hat{\sigma}^2 = 12314$. The overall filter $\xi(B)$ coefficients are provided. Forecasts for $t = 73, \ldots, 78$ are computed. The text shows a graph (Figure 9.1) of the data and 24 forecasts. Table 9.1 compares the first 6 ARAR forecasts with observed values and forecasts from ARIMA models (6.5.8) and (6.5.9). Observed RMSE for $h = 1..6$: ARAR=253, ARIMA(6.5.8)=583, ARIMA(6.5.9)=501. ARAR performs best here.

31.2 The Holt-Winters Algorithm

(Non-Seasonal) Generalizes exponential smoothing for series with locally linear trend.

31.2.1 The Algorithm

Forecast function:

$$P_n Y_{n+h} = \hat{a}_n + \hat{b}_n h, \quad h = 1, 2, \ldots \tag{112}$$

where $\hat{a}_n$ is the estimated level and $\hat{b}_n$ the estimated slope at time n. Recursions for $\hat{a}_t, \hat{b}_t$:

$$\hat{a}_{n+1} = \alpha Y_{n+1} + (1 - \alpha)(\hat{a}_n + \hat{b}_n) \tag{113}$$

$$\hat{b}_{n+1} = \beta(\hat{a}_{n+1} - \hat{a}_n) + (1 - \beta)\hat{b}_n \tag{114}$$

where $\alpha, \beta \in [0, 1]$ are smoothing parameters. Initial conditions (natural choice):

$$\hat{a}_2 = Y_2 \tag{115}$$

$$\hat{b}_2 = Y_2 - Y_1 \tag{116}$$

Solve recursively for $\hat{a}_3, \hat{b}_3, \ldots, \hat{a}_n, \hat{b}_n$. Parameters α, β can be specified or optimized to minimize $\sum_{i=3}^{n} (Y_i - P_{i-1} Y_i)^2$.

Connection to State-Space The Holt-Winters recursions (113)-(114) correspond to the steady-state Kalman filter update equations for the local linear trend model (Section 8.2.1) $Y_t = M_t + W_t$, $M_{t+1} = M_t + B_t + V_t$, $B_{t+1} = B_t + U_t$. The smoothing parameters α, β are related to the steady-state Kalman gain and noise variances.

Example 9.2.1: Accidental Deaths Data Applying non-seasonal Holt-Winters (Optimize coefficients) to DEATHS.TSM. Figure 9.2 (summarized in text) shows data and 24 forecasts. Optimal α, β are found. Table 9.2 shows forecasts for $t = 73, \ldots, 78$. Observed RMSE for $h = 1..6$: 1143. Much worse than ARAR or ARIMA, as seasonality is ignored.

31.2.2 Holt-Winters and ARIMA Forecasting

Forecasting via exponential smoothing (Eq 1.5.7, $\hat{m}_{n+1} = \alpha Y_{n+1} + (1-\alpha)\hat{m}_n$, $P_n Y_{n+h} = \hat{m}_n$) is equivalent to forecasting using the ARIMA(0,1,1) model:

$$\nabla Y_t = Z_t - (1-\alpha)Z_{t-1} = (1 - (1-\alpha)B)Z_t \tag{117}$$

Forecasting via the non-seasonal Holt-Winters algorithm is equivalent to forecasting using the ARIMA(0,2,2) model:

$$\nabla^2 Y_t = (1-B)^2 Y_t = Z_t - (2 - \alpha - \alpha\beta)Z_{t-1} + (1-\alpha)Z_{t-2} \tag{118}$$

31.3 The Holt-Winters Seasonal Algorithm

Extends Holt-Winters to handle trend and seasonality of period d.

31.3.1 The Algorithm

Forecast function:

$$P_n Y_{n+h} = \hat{a}_n + \hat{b}_n h + \hat{c}_{n+h}, \quad h = 1, 2, \ldots \tag{119}$$

where $\hat{a}_n$ (level), $\hat{b}_n$ (slope), $\hat{c}_n$ (seasonal component) are estimates at time n. The seasonal estimate is projected cyclically:

$$\hat{c}_{n+h} = \hat{c}_{n+h-kd} \quad \text{where } k = \lfloor (h-1)/d \rfloor + 1 \tag{120}$$

Recursions for $\hat{a}_t, \hat{b}_t, \hat{c}_t$:

$$\hat{a}_{n+1} = \alpha(Y_{n+1} - \hat{c}_{n+1-d}) + (1-\alpha)(\hat{a}_n + \hat{b}_n) \tag{121}$$

$$\hat{b}_{n+1} = \beta(\hat{a}_{n+1} - \hat{a}_n) + (1-\beta)\hat{b}_n \tag{122}$$

$$\hat{c}_{n+1} = \gamma(Y_{n+1} - \hat{a}_{n+1}) + (1-\gamma)\hat{c}_{n+1-d} \tag{123}$$

where $\alpha, \beta, \gamma \in [0, 1]$ are smoothing parameters. Initial conditions (natural choice): Need at least $d+1$ observations.

$$\hat{a}_{d+1} = Y_{d+1} \tag{124}$$

$$\hat{b}_{d+1} = (Y_{d+1} - Y_1)/d \tag{125}$$

$$\hat{c}_i = Y_i - (\hat{a}_{d+1} + \hat{b}_{d+1}(i - (d+1))), \quad i = 1, \ldots, d+1 \tag{126}$$

Solve recursively for $t = d+1, \ldots, n$. Parameters α, β, γ can be specified or optimized to minimize $\sum_{i=d+2}^{n} (Y_i - P_{i-1}Y_i)^2$.

Example 9.3.1: Accidental Deaths Data Applying seasonal Holt-Winters (Optimize coefficients, period $d = 12$) to DEATHS.TSM. Figure 9.3 (summarized in text) shows data and 24 forecasts. Optimal α, β, γ are found. Table 9.3 shows forecasts for $t = 73, \ldots, 78$. Observed RMSE for $h = 1..6$: 401. Better than non-seasonal HW and the ARIMA models, but not as good as ARAR for this specific horizon and data split.

31.3.2 Holt-Winters Seasonal and ARIMA Forecasting

The Holt-Winters seasonal algorithm corresponds to forecasting using a specific ARIMA model (additive seasonality):

$$(1 - B)(1 - B^d)Y_t = \theta(B)Z_t \tag{127}$$

where $\theta(B)$ is an MA polynomial of order $d + 1$ whose coefficients are complex functions of α, β, γ. The exact form is given in the text but involves $Z_{t-1}, \ldots, Z_{t-d+1}$, $Z_{t-d} - Z_{t-d-1}$, etc.

31.4 Choosing a Forecasting Algorithm

- Real data rarely perfectly match a simple model like ARIMA.
- Mean squared error isn't always the best error measure. Desired forecast horizon (1-step vs h-step) also matters.
- Heuristic algorithms (ARAR, HW, HWS) are valuable alternatives.
- Choice can be based on historical performance: apply competing algorithms to past data and compare errors (e.g., mean absolute percentage error, MAPE) for the desired forecast horizon.
- Relative performance varies by dataset. For DEATHS.TSM (first 72 points, predict next 6), ARAR was best based on RMSE. For logged WINE.TSM (first 130 points, predict next 12), an ARIMA model was best, followed by HWS, then ARAR.
- Multiplicative versions of HW exist for processes like $Y_t = m_t s_t Z_t$. Alternatively, apply additive HW to $\ln Y_t$.
- ARAR's general memory shortening can handle some multiplicative effects reasonably well even without transformation, as shown by forecasting AIRPASS.TSM directly (RMSE=18.21) compared to fitting ARIMA to logged, differenced data (RMSE=21.67) for predicting the last 12 values. Figure 9.4 (summarized in text) shows the ARAR forecasts for AIRPASS.