

ELEC60002 Statistical Signal Processing and Inference

Timothy Chung

Spring 2024

Contents

0	Foreword	2
1	Introduction	3
1.1	Applications of Statistical Signal Processing	3
2	Linear Stochastic Models	4
2.1	Objectives	4
2.2	WSS Process	4
2.3	ARMA Models	5
2.4	AR Processes	5
2.5	MA Processes	6
2.6	Duality of AR and MA processes	7
2.7	Yule-Walker and the ACF: Motivations	7
2.8	Yule-Walker and the PACF: Motivations	8
2.9	Examples of Modelling with ARMA	8
2.10	Summary: AR and MA Processes	8
2.11	Summary: Wold's Decomposition Theorem and ARMA	9
3	Intro to Estimation Theory	10
3.1	Introduction	10
3.2	Statistical Estimation: Problem Statement	10
3.3	Goodness of an Estimate	11
3.4	Minimum Variance Unbiased (MVU) Estimation	12

Chapter 0

Foreword

The notes are intended to be a gentle introductory supplement to the well-documented course notes, and not a full replacement with all the bells and whistles (e.g. real life use-cases, mathematical examples). The full notes are available on Professor Danilo Mandic's website [here](#).

Thus, a better title for this document would be 'Elements of Statistical Signal Processing and Inference'.

Chapter 1

Introduction

1.1 Applications of Statistical Signal Processing

Uses of Statistical Signal Processing

- Forecasting financial data
- Supply-demand modelling
- Pandemic spread modelling

Main theme: making sense from data based on historical observations.

There are two kinds of statistics: descriptive statistics used to describe data, and inferential statistics that try to gather generalisations from random observations.

Uses of Statistical Inference

- Adaptive learning algorithms like noise-cancelling headphones
- Neural networks for classification, prediction, denoising
- Financial modelling

Prerequisites:

- Basic statistics knowledge, refer to slide 5 of Lecture 1 notes.
- Signals and systems knowledge from 2nd year.

Learning goals:

- Random signals (real-world discrete-time), their properties and statistical descriptors
- Linear stochastic models
- Grounding in linear estimation theory
- Parametric (model-based) and nonparametric (data-driven) modelling
- Optimal estimators for random signals with performance bounds
- Least squares methods
- Adaptive estimation for nonstationary data

Chapter 2

Linear Stochastic Models

2.1 Objectives

- Introduce linear stochastic models for real world data
- Understand stochastic processes for noise
- ARMA models, partial correlations, and optimal model order selection criteria (MDL, AIC,...)

Wold's Decomposition Theorem

Theorem 2.1.1

Any covariance-stationary time series can be decomposed into two different parts: **deterministic** and **stochastic**.

Let w denote a white process, $x_r[n]$ a regular random process, and $x_p[n]$ a predictable process such that $x_r[n] \perp x_p[n]$.

$$x[n] = x_p[n] + x_r[n] \quad (2.1)$$

$$= x_p[n] + \sum_{j=1}^q b_j w[n-j] \quad (2.2)$$

The condition requires $E[x_r[m]x_p[n]] = 0$.

Deterministic vs Stochastic

Definition 2.1.1

- Deterministic means a component can be precisely described by an equation without any randomness, following a predictable pattern that can be determined exactly for any point in time, e.g. a sine wave.
- Stochastic means a component described by a random process. It cannot be described by a simple equation but can be represented using a probability distribution, such as WGN (Filtered White Gaussian) noise.

2.2 WSS Process

A WSS process has two conditions:

1. The mean of the process is constant for all time, i.e., $E[X_t] = \mu$ for all t , where μ is a constant.
2. The autocovariance function $R_x(t, t + \tau) = \text{Cov}(X_t, X_{t+\tau})$ depends only on the time lag τ and not on time t itself, i.e., $R_x(t, t + \tau) = R_x(\tau)$.

The general form for the power spectrum of a WSS process is

$$P_x(e^{j\omega}) = \sum_{k=1}^N \alpha_k \delta(\omega - \omega_k) + P_{x_r}(e^{j\omega}) \quad (2.3)$$

A covariance-stationary process is called linearly deterministic if $p(x[n] | x[n-1], x[n-2], \dots) = x[n]$. It can be predicted correctly with zero error if we know its entire past data $x[n-1], x[n-2], \dots$

2.3 ARMA Models

ARMA models are a linear stochastic model.

Autoregressive (AR) filters use an all-pole system and Moving Average (MA) filters use an all-zero system. An Autoregressive Moving Average (ARMA) filter utilises both poles and zeroes.

In ARMA modelling we filter white noise $w[n]$ with a causal linear shift-invariant filter (transfer function $H[z]$ that has p poles and q zeroes.

$$X(z) = H(z)W(z) \Rightarrow H(z) = \frac{B_q(z)}{A_p(z)} = \frac{\sum_{k=0}^q b_k z^{-k}}{1 + \sum_{k=1}^p a_k z^{-k}} \quad (2.4)$$

If our filter is WSS, then $x[n]$ is WSS as well. To show this, multiply both sides of Equation 2.4 by $x[n-k]$ and calculate expectation, we have

$$r_{xx}(k) = \underbrace{\sum_{l=1}^p a_l r_{xx}(k-l)}_{\text{easy to calculate}} + \underbrace{\sum_{l=0}^q b_l r_{xw}(k-l)}_{\text{can be complicated}} \quad (2.5)$$

Note that since $X(z) = H(z)W(z)$ the random processes $x[n]$ and $w[n]$ are related by a linear difference equation with constant coefficients. This is

$$\begin{aligned} \text{ARMA}(p, q) = H(z) &= \frac{B(z)}{A(z)} = \frac{\sum_{k=0}^q b_k z^{-k}}{1 + \sum_{k=1}^p a_k z^{-k}} \\ x[n] &= \underbrace{\sum_{l=1}^p a_l x[n-l]}_{\text{autoregressive}} + \underbrace{\sum_{l=0}^q b_l w[n-l]}_{\text{moving average}} \end{aligned} \quad (2.6)$$

Note that for $H(z)$, coefficients must be absolutely summable. For the process to be stationary, $\sum_{j=0}^{\infty} |b_j| < \infty$, and for it to be invertible, $\sum_{j=0}^{\infty} |a_j| < \infty$.

2.4 AR Processes

An autoregressive process of order p , denoted by $AR(p)$ can be described by:

$$x[n] = a_1 x[n-1] + a_2 x[n-2] + \dots + a_p x[n-p] + w[n] \quad (2.7)$$

$$= \sum_{i=1}^p a_i x[n-i] + w[n] \quad (2.8)$$

$$= \mathbf{a}^T \mathbf{x}[n] + w[n] \quad (2.9)$$

2.4.1 ACF of an AR Process

We start by finding $x[n-k]x[n]$:

$$x[n-k]x[n] = a_1 x[n-k]x[n-1] + a_2 x[n-k]x[n-2] + \dots + a_p x[n-k]x[n-p] + x[n-k]w[n] \quad (2.10)$$

When $k > 0$, $E\{x[n-k]w[n]\} = 0$ due to both processes being orthogonal to each other. We are left with:

$$r_{xx}(k) = \begin{cases} a_1 r_{xx}(1) + a_2 r_{xx}(2) + \dots + a_p r_{xx}(p) + \sigma_w^2, & \text{for } k = 0 \\ a_1 r_{xx}(k-1) + a_2 r_{xx}(k-2) + \dots + a_p r_{xx}(k-p), & \text{for } k > 0 \end{cases} \quad (2.11)$$

2.4.2 Normalised ACF of an AR Process

We can normalise by dividing by $r_{xx}(0)$ to get $\rho(k) = r_{xx}(k)/r_{xx}(0)$.

$$\rho(k) = a_1\rho(k-1) + a_2\rho(k-2) + \dots + a_p\rho(k-p) \quad k > 0 \quad (2.12)$$

2.4.3 Variance of an AR Process

For $k = 0$, the $E\{x[n-k]w[n]\}$ term contributes σ_w^2 to variance, and

$$r_{xx}(0) = a_1r_{xx}(-1) + a_2r_{xx}(-2) + \dots + a_pr_{xx}(-p) + \sigma_w^2 \quad (2.13)$$

Dividing by $r_{xx}(0) = \sigma_x^2$, we get:

$$\sigma_x^2 = \frac{\sigma_w^2}{1 - \rho_1a_1 - \rho_2a_2 - \dots - \rho_pa_p} \quad (2.14)$$

2.4.4 Power Spectrum of an AR Process

Recall the formula for output power of a linear system, $P_{xx} = |H(z)|^2 P_{ww} = H(z)H^*(z)P_{ww}$. We then have

$$P_{xx}(f) = \frac{2\sigma_w^2}{|1 - a_1e^{-j2\pi f} - \dots - a_pe^{-j2\pi pf}|^2} \quad 0 \leq f \leq 1/2 \quad (2.15)$$

2.5 MA Processes

A moving average process of order q , $MA(q)$, is given by:

$$x[n] = w[n] + b_1w[n-1] + \dots + b_qw[n-q] \quad (2.16)$$

$$= w[n] + \sum_{i=1}^q b_iw[n-i] \quad (2.17)$$

$$= \mathbf{b}^T \mathbf{w}[n] + w[n] \quad (2.18)$$

2.5.1 ACF of an MA process

Note that the ACF has a cutoff after lag q .

$$r_{xx}(k) = E[(w[n] + b_1w[n-1] + \dots + b_qw[n-q])(w[n-k])] \quad (2.19)$$

2.5.2 Variance of an MA process

We sub in $k = 0$ into the ACF to obtain the variance:

$$r_{xx}(0) = (1 + b_1^2 + \dots + b_q^2)\sigma_w^2 \quad (2.20)$$

2.5.3 Power Spectrum of an MA process

Since a moving average filter has a transfer function of all zeroes, and no poles except at the origin, which is known as an ARMA process.

$$P(f) = 2\sigma_w^2 |1 + b_1e^{-j2\pi f} + b_2e^{-j4\pi f} + \dots + b_qe^{-j2\pi qf}|^2 \quad (2.21)$$

An MA process has a limited ability to accurately represent time series with spectra that have sharp peaks (high power at specific frequencies). This is because the MA model, being a sum of weighted noise components, tends to produce a smoother spectrum without sharp features. To model time series with sharp spectral peaks (like those that might be seen in signals dominated by sinusoidal components), one would typically need AR or ARMA processes.

2.6 Duality of AR and MA processes

Because of duality between IIR and FIR (infinite and finite impulse response) filters, every AR process has an MA representation. Take for example AR(1):

$$x[n] = a_1 x[n-1] + w[n] \Leftrightarrow \sum_{j=0}^{\infty} b_j w[n-j] \quad (2.22)$$

2.7 Yule-Walker and the ACF: Motivations

Say we want to find the coefficients for an AR(1) process:

$$x[n] = a_1 x[n-1] + w[n] \quad (2.23)$$

The case of $p = 1$ is trivial. We form the over-determined system

$$\underbrace{\begin{pmatrix} x_2 \\ x_3 \\ \vdots \\ x_N \end{pmatrix}}_{\mathbf{b}} = \underbrace{\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{N-1} \end{pmatrix}}_{\mathbf{A}} a_1 \quad (2.24)$$

And solve with the least-squares estimator

$$\hat{a}_1 = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b} = \frac{\sum_{i=1}^{N-1} x_i x_{i+1}}{\sum_{i=1}^{N-1} x_i^2} = \frac{c_1}{c_0} = r_1 \quad (2.25)$$

where c_i, r_i refer to the i -th autocovariance and correlation coefficients respectively.

For the case of $p = 2$, with process

$$x[n] = a_1 x[n-1] + a_2 x[n-2] + w[n] \quad (2.26)$$

We have

$$\underbrace{\begin{pmatrix} x_3 \\ x_4 \\ \vdots \\ x_N \end{pmatrix}}_{\mathbf{b}} = \underbrace{\begin{pmatrix} x_2 & x_1 \\ x_3 & x_2 \\ \vdots & \vdots \\ x_{N-1} & x_{N-2} \end{pmatrix}}_{\mathbf{A}} \underbrace{\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}}_{\mathbf{a}} \quad (2.27)$$

But notice how that as the order grows, it becomes more computationally intensive to compute the inverse in $\hat{\mathbf{a}} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b}$ as $\mathbf{A}^T \mathbf{A}$ is not guaranteed to be diagonal, so inverting it becomes difficult.

Instead, we can derive a more efficient process for any order p . From Equation 2.12, we note that

$$\rho(k) = a_1 \rho(k-1) + a_2 \rho(k-2) + \cdots + a_p \rho(k-p) \quad k > 0$$

We can list the equations as such, notating $\rho(k)$ as ρ_k :

$$\begin{array}{lllllll} \rho_1 = & a_1 \rho_0 + & a_2 \rho_1 + & a_3 \rho_2 + & \cdots + & a_{p-1} \rho_{p-2} + & a_p \rho_{p-1} \\ \rho_2 = & a_1 \rho_1 + & a_2 \rho_0 + & a_3 \rho_1 + & \cdots + & a_{p-1} \rho_{p-3} + & a_p \rho_{p-2} \\ \rho_{p-1} = & a_1 \rho_{p-2} + & a_2 \rho_{p-3} + & a_3 \rho_{p-4} + & \cdots + & a_{p-1} \rho_0 + & a_p \rho_1 \\ \rho_p = & a_1 \rho_{p-1} + & a_2 \rho_{p-2} + & a_3 \rho_{p-3} + & \cdots + & a_{p-1} \rho_1 + & a_p \rho_0 \end{array}$$

Since $\rho_0 = 1$, we have:

$$\underbrace{\begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_{p-1} \\ \rho_p \end{pmatrix}}_{\mathbf{r}} = \underbrace{\begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{p-2} & \rho_{p-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{p-3} & \rho_{p-2} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \rho_{p-2} & \rho_{p-3} & \rho_{p-4} & \cdots & 1 & \rho_1 \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} & \cdots & \rho_1 & 1 \end{pmatrix}}_{\mathbf{R}} \underbrace{\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{p-1} \\ a_p \end{pmatrix}}_{\mathbf{a}} \quad (2.28)$$

This is the form $\mathbf{R}\mathbf{a} = \mathbf{r}$, where \mathbf{R} is a square coefficient matrix that is full rank and symmetric, and invertability is guaranteed. We can then go on to calculate $\hat{\mathbf{a}} = \mathbf{R}^{-1}\mathbf{r}$.

2.8 Yule-Walker and the PACF: Motivations

The PACF stands for the Partial Autocorrelation function, which shows the relationship between an observation in a time series with observations at prior time steps, with the relationships of intervening observations removed.

It is a vector π defined by

$$\pi(k) = \begin{cases} 1 & \text{if } k = 0 \\ a_{kk} & \text{if } k \geq 1 \end{cases} \quad (2.29)$$

where a_{kk} is the last component of $\mathbf{a}_k = [a_{k1}, a_{k2}, \dots, a_{kk}]^T$ from $\mathbf{a}_k = \mathbf{R}_k^{-1}\mathbf{r}_k$. We denote a_{kj} to be the j th coefficient in an autoregressive representation of order k , at Equation 2.28.

$\rho_k = a_1\rho_{k-1} + a_2\rho_{k-2} + \dots + a_p\rho_{k-p}$ $k > 0$ now becomes:

$$\rho_j = a_{kj}\rho_{j-1} + \dots + a_{k(k-1)}\rho_{j-k+1} + a_{kk}\rho_{j-k} \quad j = 1, 2, \dots, k \quad (2.30)$$

$$\underbrace{\begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_{k-1} \\ \rho_k \end{pmatrix}}_{\mathbf{r}} = \underbrace{\begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_{k-2} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \rho_{k-2} & \rho_{k-3} & \rho_{k-4} & \cdots & 1 & \rho_1 \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & 1 \end{pmatrix}}_{\mathbf{R}} \underbrace{\begin{pmatrix} a_{k1} \\ a_{k2} \\ \vdots \\ a_{k(k-1)} \\ a_{kk} \end{pmatrix}}_{\mathbf{a}} \quad (2.31)$$

We can solve for $k = 1, 2 \dots$ manually:

$$a_{11} = \rho_1, \quad a_{22} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2}, \quad a_{33} = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_1 \\ \rho_1 & 1 & \rho_2 \\ \rho_2 & \rho_1 & \rho_3 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{vmatrix}}, \quad \text{etc} \quad (2.32)$$

The partial autocorrelation function at lag k , denoted $\pi(k)$ (or equally the AR coefficient a_{kk}), measures the linear relationship between $x(n)$ and $x(n-k)$, once we have removed the influence of $x_{n-1}, \dots, x_{n-k+1}$, i.e.,

$$a_{kk} = \text{corr}(x(n) - \hat{x}(n), x(n-k) - \hat{x}(n-k)).$$

The PACF is used to determine the order p of an AR Model (when we are trying to model the data with an AR model), as when lag k reaches p , PACF should drop off, indicating no linear relationship beyond that point.

Therefore, for an AR(p) process, the PAC a_{kk} is nonzero for all $k \leq p$ and zero everywhere else. In practice, it is difficult to guarantee this for real world data, so a small threshold for tolerance is needed.

2.9 Examples of Modelling with ARMA

See the slides, page 42 onwards.

2.10 Summary: AR and MA Processes

1. A stationary finite AR(p) process can be represented as an infinite order MA process. A finite MA process can be represented as an infinite AR process.
2. The finite MA(q) process has an Autocorrelation Function (ACF) that is zero beyond q . For an AR process, the ACF is infinite in length and consists of a mixture of damped exponentials and/or damped sine waves.

3. Finite MA processes are always stable, and there is no requirement on the coefficients of MA processes for stationarity. However, for invertibility, the roots of the characteristic equation must lie inside the unit circle.
4. AR processes produce spectra with sharp peaks (two poles of $A(z)$ per peak), whereas MA processes cannot produce peaky spectra.

ARMA modelling is a classic technique which has found a tremendous number of applications.

2.11 Summary: Wold's Decomposition Theorem and ARMA

- Every stationary time series can be represented as a sum of a perfectly predictable process and a feasible moving average process.
- Two time series with the same Wold representations are the same, as the Wold representation is unique.
- Since any MA process also has an ARMA representation, working with ARMA models is not an arbitrary choice but is physically justified.
- The causality and stationarity on ARMA processes depend entirely on the AR parameters and not on the MA parameters.
- An MA process is not uniquely determined by its ACF.
- An $\text{AR}(p)$ process is always invertible, even if it is not stationary.
- An $\text{MA}(q)$ process is always stationary, even if it is non-invertible.

Chapter 3

Intro to Estimation Theory

3.1 Introduction

- Historical data still allows for accurate modelling of complex systems.
- This links to a need for a unifying and rigorous framework that defines a ‘goodness of performance’ measure for any Data Analytic model

Terminology: **Prediction** refers to a pre-built model based on in-sample data to estimate out-of-sample values. **Forecasting** is a form of prediction that implicitly assumes time-series methods, and we use historical data to predict future data, often with probabilistic bounds (confidence intervals).

3.2 Statistical Estimation: Problem Statement

Estimators and Estimate	Definition 3.2.1
Given an N-point dataset, $x[0], x[1], \dots, x[N-1]$ that depends on unknown scalar parameter θ . We let $\hat{\theta}$ be an estimator that is defined as a function $g(\cdot)$ of dataset $\{x\}$, that can estimate our unknown parameter θ .	
$\hat{\theta} = g(x[0], x[1], \dots, x[N-1]) \quad (3.1)$	
This is the single parameter or scalar case. We can consider the vector case, determining a set of parameters $\theta = [\theta_1, \dots, \theta_p]^T$ from data samples $\mathbf{x} = [x[0], \dots, x[N-1]]^T$ where such parameters will yield the highest probability of obtaining the observed data.	
$\max_{\text{span } \theta} p(\mathbf{x}; \theta) \quad (3.2)$	
The estimator refers to our rule $g(\mathbf{x})$ that assigns a value to parameter θ . And an estimate of the true value of θ is $\hat{\theta}$.	

We can estimate θ with **classical estimation**, assuming θ is deterministic with no *a priori* information about it (minimum-variance solution, least-squares), or use **Bayesian estimation**, applying prior knowledge to it (Wiener and Kalman filters, adaptive signal processing).

$p(\mathbf{x}; \theta)$ contains all the information we need to find $\hat{\theta}$, however, in practice, the PDF is not given so we need to choose a model that captures the essence of the system we are trying to model – leading to a mathematically tractable estimator.

Gaussian RV	Definition 3.2.2
A Gaussian random variable $X \sim N(\mu, \sigma^2)$ has the pdf	
$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (3.3)$	
μ is the mean, σ is the standard deviation, and is greater than zero. σ^2 is the variance.	

Conditional PDF	Definition 3.2.3
<p>A conditional pdf $p_{y x}(y x)$ can be thought of as a sliced, normalised form of the joint pdf $p(x, y)$.</p> <p>It is formally defined as</p> $p_{Y X}(y x) = \begin{cases} \frac{p_{XY}(x, y)}{p_X(x)} & p_X(x) \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad (3.4)$ <p>General shorthand and notation means we define it as</p> $p(x y) = \begin{cases} \frac{p(x, y)}{p(y)} & p(y) \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad (3.5)$	

Bias	Definition 3.2.4
<p>The bias is the difference between the expected value of the estimate $\hat{\theta}$ and actual value of the parameter θ. It is denoted by B. For N data samples, we have</p> $B = E\{\hat{\theta}_N - \theta\} \quad (3.6)$ <p>It can also be defined as $E[\eta]$, where $\eta = \hat{\theta} - \theta$.</p>	

3.3 Goodness of an Estimate

We assume noise \mathbf{w} is white with i.i.d samples. But whiteness is not realistic, much rather, it is Gaussian. We can also assume it has zero-mean. These assumptions (later on) allow us to find a performance bound for optimal estimators, allowing us to gauge performance of an estimator.

Goodness analysis is usually a function of noise variance σ_w^2 , usually expressed in terms of signal-to-noise ratio, SNR. Usually, we can only assess performance if we know true θ . Typically the goodness of an estimator is captured through the mean and variance of $\hat{\theta} = g(\mathbf{x})$.

We want $\mu_{\hat{\theta}} = E[\hat{\theta}] = \theta$, and a small variance, $\sigma_{\hat{\theta}} = E\{(\hat{\theta} - E\{\hat{\theta}\})^2\}$

Estimation error η is defined by $\eta = \hat{\theta} - \theta$. Since $\hat{\theta}$ is a random variable and θ is a constant, then η is a random variable too. Also, $\eta = 0$ refers to an unbiased estimator, which exactly matches the parameter θ . We need:

1. $E\{\eta\} = 0$, an unbiased estimator
2. A minimum variance, where $\text{Var}(\eta) = E\{(\eta - E\{\eta\})^2\}$ is small

3.3.1 Bias

- For sufficiently large N observations of $x[n]$, the expected value of an unbiased estimate $\hat{\theta}$ converges towards its true value:

$$E\{\hat{\theta}_N\} = \theta \equiv B = E\{\hat{\theta}_N\} - \theta = 0 \quad (3.7)$$

- Conversely, if $B \neq 0$ then the estimator $\hat{\theta} = g(\mathbf{x})$ is biased asymptotically.
- Therefore it can be described that an estimator's bias is a critical measure of its **accuracy**. An unbiased estimator has an expected value equal to the parameter it estimates, while an asymptotically unbiased estimator approaches this property as the sample size grows infinitely large.

3.3.2 Variance

- The **precision** of an estimator is assessed by how closely it can consistently estimate the parameter. This is formalised by the variance of the estimator approaching zero as the sample size increases indefinitely, called

the **variance criterion**:

$$\lim_{N \rightarrow \infty} \text{var}(\hat{\theta}_N) = 0 \quad (3.8)$$

- This notion of precision can also be expressed in terms of the estimator's mean squared error (MSE) diminishing as the sample size grows:

$$\lim_{N \rightarrow \infty} \text{E} \left[(\hat{\theta}_N - \text{E}[\hat{\theta}_N])^2 \right] = 0 \quad (3.9)$$

The Mean Square Convergence is a stronger form of converge than the aforementioned **variance criterion**.

- When the estimator is unbiased, which means its expected value is the true parameter value, we can apply Chebyshev's inequality to show that the estimator's probability of deviating from the true parameter by more than any positive number ϵ decreases as variance decreases:

$$\Pr \left\{ |\hat{\theta}_N - \theta| \geq \epsilon \right\} \leq \frac{\text{var}(\hat{\theta}_N)}{\epsilon^2} \quad (3.10)$$

- If the **variance of the estimator does indeed tend to zero as the sample size approaches infinity**, and the estimator is **asymptotically unbiased**, the estimator is **consistent**. This means that the probability that $\hat{\theta}$ differs from θ by more than ϵ will go to zero; it converges in probability to the true parameter value:

$$\hat{\theta}_N \xrightarrow{N \rightarrow \infty} \theta \quad (3.11)$$

Tschebycheff (or Chebyshev) inequality

Extra, Not Assessed 3.3.1

Fun trivia: Tschebycheff is one of many transliterations for Chebyshev, more include Tchebichef, Tchebychev, Tchebycheff, Tschebyshev...

Chebyshev's inequality provides a bound on the probability that a random variable deviates from its mean. To prove this, we will use Markov's inequality, which states that for a non-negative random variable X and any $a > 0$,

$$P(X \geq a) \leq \frac{\text{E}[X]}{a}. \quad (3.12)$$

Now let X be a random variable with finite mean μ and finite variance σ^2 . For any $\epsilon > 0$, consider the non-negative random variable $(X - \mu)^2$. Applying Markov's inequality to this random variable, we have:

$$P((X - \mu)^2 \geq \epsilon^2) \leq \frac{\text{E}[(X - \mu)^2]}{\epsilon^2}. \quad (3.13)$$

Since the variance of X is σ^2 , we have $\text{E}[(X - \mu)^2] = \sigma^2$. Therefore, we can rewrite the inequality as:

$$P((X - \mu)^2 \geq \epsilon^2) \leq \frac{\sigma^2}{\epsilon^2}. \quad (3.14)$$

Noticing that $(X - \mu)^2 \geq \epsilon^2$ if and only if $|X - \mu| \geq \epsilon$, we can write:

$$P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}. \quad (3.15)$$

This is the statement of Chebyshev's inequality. ■

3.4 Minimum Variance Unbiased (MVU) Estimation