Chapter 5

Stochastic Setup

Niels Bohr, 1986 - as reply to a visitor to his home in Tisvilde who asked him if he really believed a horseshoe above his door brought him luck: "Of course not ... but I am told it works even if you don't believe in it."

The framework of stochastic models is often useful for implementing the following two philosophies:

(Analysis): The primary use of a stochastic framework is to assume that the experiments involved in a certain estimation task follow a proper stochastic rule set. In this way one can abstract away much of the technical irregularities while making life much easier for the analysis of the techniques. The price one has to pay in general for this convenience is that the results 'only' hold 'almost surely', that is, there is an extremely small chance that results go bogus. (Computer scientists like to use the phrase 'with overwhelming probability').

(Constructive): Recent work has shown that the device of randomisation is useful in the design of algorithms. It turns out that this way one can push the boundaries of feasible computation tasks much further theoretically (w.r.t. computational complexity) as well as practically (w.r.t. large-scale computation tasks).

The predominant setup in the analysis of estimation, identification or filtering techniques is that where the involved signals are considered (partially) stochastic. Intuitively this means that the signals itself can be unspecified (to a certain degree), but that the *mechanism generating the signals* is fixed. In practice, stochastic properties manifest themselves as follows: when performing a *stochastic* experiment twice under exactly the same conditions, results could possibly differ. If performing the same experiment twice and results would always be equal, we say that the experiment were *deterministic*.

While I assume that the reader experienced already an introductory class in probability theory or statistics, we will spend some effort in reviewing the basics once more. Not only for the sake of expressing results unambiguously, but also in order to pinpoint the power and limitations of the surveyed results later.

5.1 Getting the Basics Right

5.1.1 Events, Random variables and Derived Concepts

The following definitions establish a proper setup, as introduced by N. Kolmogorov in the 30s. Abstracting from applications, probability theory studies an experiment with a number of possible outcomes $\{\omega\}$. The totality of such outcomes is the sample space $\Omega = \{\omega\}$. An *event* - say A is a subset of the sample space. A *probability measure* P is a function from an event to a number between 0 and 1, or $P : \{\Omega\} \to [0, 1]$, with properties:

1. $0 \leq P(A) \leq 1$
2. $P(\Omega) = 1$
3. Let $\{A_i\}_i$ be any (countable many) set of disjunct events, then $\sum_i P(A_i) = P(\cup_i A_i)$.

Not all possible subsets of Ω need to be events, but the universe of events must form a sigma-field: 'if A is an event, so is $\Omega \backslash A$' and 'the union of any countable number of events must be an event', and 'Ω is an event'. Let's give some examples.

Example 32

- *Sample ω = images on web. A corresponding sample space Ω contains all images present on the web. An event A is e.g. 'all the images in Ω which are black and white' (informally, an image $\omega \in \Omega$ is black-and-white iff $\omega \in A$.)*
- *Sample ω = speech signals. A corresponding sample space Ω is the collection of all possible speech signals. An event A is e.g. the subset of speech signals only containing background noise. (informally, a speech signal ω contains only background noise iff $\omega \in A$.)*
- *Sample ω = weather in Uppsala. A corresponding sample space Ω is the collection of all possible weather regimes in Uppsala. An event A here is e.g. those cases where the weather is called sunny. (informally, a weather regime ω is called sunny iff $\omega \in A$.)*
- *Sample ω = external force on a petrochemical plant. A corresponding sample space Ω is the collection of all possible external forces which could act on the studied plant. An event A here is e.g. the collections of all those external forces which may drive the plant to an unstable working. (informally, an external force ω results in unstable working iff $\omega \in A$.)*

There are a number of derived concepts which we merely summarise:

(Joint): Let $A, B \subset \Omega$ be two events, then the joint probability is defined as

$$P(A, B) \triangleq P(A \cup B). \tag{5.1}$$

(Independence): Let $A, B \subset \Omega$ be two events, then they are called mutually independent if

$$P(A, B) \triangleq P(A)P(B). \tag{5.2}$$

(Conditional): Let $A, B \subset \Omega$ be two events where $B \neq \{\}$, then the conditional probability is defined as

$$P(A|B) \triangleq \frac{P(A, B)}{P(B)}. \tag{5.3}$$

(Bayes): Let $A, B \subset \Omega$ be two events, then Bayes' law says that

$$P(A|B)P(B) = P(B|A)P(A) = P(A, B). \tag{5.4}$$

Often, we are interested in quantities associated with the outcome of an experiment. Such quantity is denoted as a *random variable*. Formally, a random variable is a function defined for any possible $\omega \in \Omega$. If the random variable is evaluated at the sample ω which actually occurred (the observation), we refer to it as a *realisation* of this random variable. This quantity is what we intend with a *value* of a random variable. Following the convention in statistical literature, we denote a random variable as a capital letter. This notational convention makes it easier to discriminate between random variables and deterministic quantities (denoted using lower case letter). This motivates the use of the following notational convention:

$$P(X = x) \triangleq P\left(\{\omega | X(\omega) = x\}\right). \tag{5.5}$$

where $\{\omega|X(\omega) = x\}$ is the set of all samples ω which has a random value $X(\omega)$ equal to x. We have as before that $P : \{\Omega\} \to [0, 1]$, and as such $P\left(\{\omega|X(\omega) = x\}\right)$ gives a number between 0 and 1. Likewise, $P(X > 0)$ means that $P\left(\{\omega|X(\omega) > 0\}\right)$ etc. If X denotes a random variable defined over the outcome space Ω, then $X(\omega)$ denotes a realization measured when ω is sampled from Ω. Sometimes, X can only take a finite number of values, and X is as such called discrete. If not so, X is called a continuous random variable.

Example 33 *The following example illustrates ideas using a simple urn model.*

1. *Consider an urn containing $m = 10$ balls, one ball labeled '2', three balls labeled '1', and 6 of them labeled '0'. The set of all 10 balls is called the 'sampling space' Ω.*
2. *Randomness samples a ball in Ω denoted as ω. This sampling is essentially uniform, any sample comes up equally probable.*
3. *'The subset of balls with label 0' or informally 'A ball with label '0' is drawn', is an event.*
4. *Then the label of this 'random' ball - denoted as the function Z - is a random variable. The actual value $Z(\omega)$ is called a realization of this random variable.*
5. Before *the actual sampling, one could expect a value Z of $\frac{1}{10}(6*0+3*1+1*2) = 0.5$ denoted as $\mathbb{E}[Z] = 0.5$.*
6. *If repeating the experiment $n \to \infty$ times independently, one would end up with the ball labeled '2' in a fraction of $\frac{1}{10}$ of the times. This is captured by the law of large numbers.*

At this elementary level, we make already important conceptual steps:

- The sample space describes the physical reality.
- A random variable is a *mapping* of a sample to its corresponding label.
- 'Randomness' picks any sample with equal probability, while the probability of the corresponding labels is governed by the frequency of the samples with identical labels. This means that the law of probability corresponding to Z is implied by the definition of the random variable, not in the way randomness were implemented!

- Expectations are evaluated *before* the actual experiment is carried out. When doing the calculations when knowledge exists on which ω actually occurred in reality (the observation), the notion of probability is contaminated! In general, a statisticians job is finished right before the actual experiment is implemented (except for the consultancy part).

5.1.2 Continuous Random Variables

However, the above setup does not entirely characterise the intuitive concepts that we were after: a stochastic setup is adopted in order to characterise the mechanism generating the data. This probability function P is however not suited to explain the likelihood of a single sample, but focusses on sets and subsets of events. This subtle difference leads easily to a paradox, as seen in the following example. Consider an event-space such that an infinite number of events may occur. For example, consider the events of all possible 'weathers' in Uppsala: an infinite number of variations can occur, and assume (for the sake of the argument) that any 'weather' is equally probably to occur at an instance. Lets represent the weather which actually occurred as ω. Then $P(\omega) = 0$ necessarily, and the probability of this event equals zero. So it seems that this precise sample (the observation) was not possible to occur after all! This paradox arises as working with infinite sample spaces is not as straightforward as in the discrete case, and a proper notion of 'the probability of a single event' needs an additional apparatus as shown in the following subsection.

In case the sample space Ω contains an (uncountable) infinite number of elements, the above framework needs to be extended slightly in order to deal properly with measurability issues. Let us first look towards the case where a random value X defined over such a sampling space takes values in $\mathbb{R}$.

Definition 14 (CDF and PDF) *The laws of probability associated to a continuous, univariate random variable go as follows:*

(CDF): The Cumulative Distribution Function $F : \mathbb{R} \to [0, 1]$ *(CDF) of a univariate random variable* $X : \Omega \to \mathbb{R}$ *is defined as*

$$F(x) \triangleq P(X \leq x) \triangleq P\left(\{\omega | X(\omega) \leq x\}\right). \tag{5.6}$$

Consequently, one has that $F(-\infty) = 0$, $F(\infty) = 1$ *and the function* F *is monotonically increasing. An example is given in Fig. (5.1.a)*

(PDF): The Probability Density Function $f : \mathbb{R} \to \mathbb{R}_+$ *(PDF) of a univariate random variable* $X : \Omega \to \mathbb{R}$ *with a differential CDF* F *is defined as*

$$f(x) \triangleq \frac{\partial P(X \geq x)}{\partial x} = \frac{\partial F(x)}{\partial x}. \tag{5.7}$$

Those definitions are not mere academical, but clarify for example that a density function does not equal a probability law. Both notions lead also to different tools to estimate the probability laws underlying data.

(HIST): Given a sample of n samples taking values in $\mathbb{R}$, or $\{y_i\}_{i=1}^n \subset \mathbb{R}$, the histogram counts the frequency (normalized number) of samples occurring in a given interval (bin) of $\mathbb{R}$. For example, if we have 5 samples $\{1, 2, 3, 4, 5\}$, and two intervals (bins) $(-\infty, 3]$ and $(3, \infty)$, then

the histogram would say $(3/5, 2/5)$. This is then an estimate of the PDF. A graphical example is given in Fig. (5.2).a of a histogram with 20 bins, and using a sample of $n = 100$. The bins are usually chosen to make the picture look 'pleasing' (ad hoc).

(ECDF): Given a sample of n samples taking values in $\mathbb{R}$, or $\{y_i\}_{i=1}^n \subset \mathbb{R}$, then the *Empirical Cumulative Distribution Function* (ECDF) is a function $\hat{F}_n : \mathbb{R} \to [0, 1]$ which is defined as

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(y_i \leq x), \tag{5.8}$$

where $I(z)$ equals one if z holds true, and zero otherwise. Note that in order to set up this function, one does not need to make choices as the location or size of the bins. This estimator is far more efficient than the histogram, albeit the latter is more often used as it is visually more appealing. A graphical example is given in Fig. (5.2).b of the ECDF using a sample of $n = 100$.

5.1.3 Normal or Gaussian Distribution

Of special (practical as well as theoretical) interest is the Gaussian or Normal distribution with mean μ and standard deviation $\sigma > 0$. Those quantities are also referred to as the first two *moments* of the distribution. The PDF is given for any x as

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{5.9}$$

The quantity σ^2 is also known as the *variance* and characterizes the *spread* of the PDF (see Fig. (5.1).a) This specific distribution is of practical as well as theoretical interest for many reasons, perhaps the most important ones being:

(CLT): (the Central Limit Theorem): This classical result states that the average of a large number n of random variables arising from independently samples tends to a normal distribution with standard deviation $O(\sqrt{n})$. This theorem has a long history, but is now often connected to J.W. Lindenberg.

(Closed): The Gaussian distribution is remarkably stable, meaning that a convolution of two Gaussians is still Gaussian. Often, when performing calculations with Gaussian distributions one can easily derive that the resulting distribution is Gaussian as well. Since the Gaussian is characterized by their first two moments only, one consequently needs only to calculate with those and sidestep working with the functional form for the rest.

(Convenience): A third reason one has for using the Gaussian distribution is its convenience. For example, from a practical point of view many related tools are available in statistical software environments. From a more pen-and-pencil perspective it is plain that it is more easy to work with the two first moments than to work with the full functional form of a distribution.

The first reason also implies that the Gaussian distribution will often turn up as a limit distribution of an estimator.

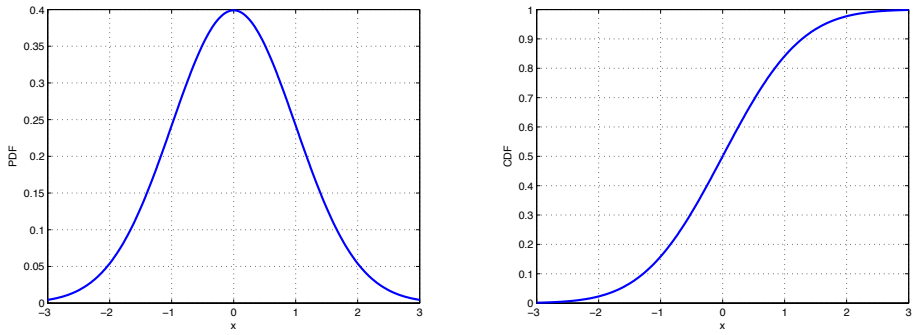


Figure 5.1: (a) PDF of the normal distribution with mean 0 and unit variance. (b) CDF of the normal distribution with mean 0 and unit variance.

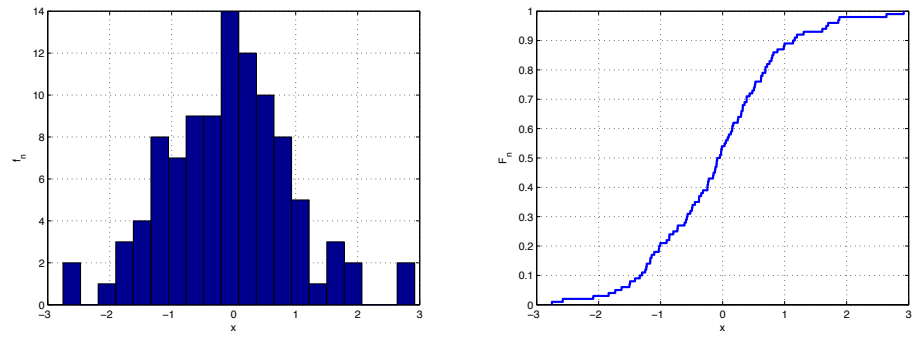


Figure 5.2: Illustration of difference of CDF versus PDF based on a sample of $n = 100$ standard Gaussian distributed values. The histogram - displaying the relative frequency of samples falling within each bin - is the better-known estimate of the pdf. The empirical CDF - defined for each $x \in \mathbb{R}$ as the relative frequency of samples smaller than x - is however much more accurate and fool-proof, but is perhaps less intuitive.

Example 34 *The following examples are instructive. Assume Z is a random variable taking values in $\mathbb{R}^d$, following a Gaussian distribution with the PDF as given in (5.9) for given parameters $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d\times d}$. Then*

$$\mathbb{E}[Z] = \mu, \tag{5.10}$$

and

$$\mathbb{E}\left[(Z-\mu)(Z-\mu)^T\right] = \Sigma. \tag{5.11}$$

and

$$Z \sim \mathcal{N}(\mu, I_d) \Leftrightarrow Z - \mu \sim \mathcal{N}(\mu, I_d). \tag{5.12}$$

Let $z \in \mathbb{R}^d$ be a realization of the random variable Z, then

$$\mathbb{E}[z] = z, \tag{5.13}$$

and

$$\mathbb{E}[z^T Z] = z^T \mu. \tag{5.14}$$

Hence

$$\mathbb{E}\left[(Z-\mu)(z-\mu)^T\right] = 0_d. \tag{5.15}$$

5.1.4 Random Vectors

A random vector is an array of random variables. In general, those random variables are related, and the consequent probability rules governing the sampling of the random vector summarizes both the individual laws as the dependence structure inbetween the different elements. This then leads to the notion of a joint probability distribution functions. Again, we make a difference between the joint Cumulative Distribution Function (joint CDF) and the joint Probability Density Function (joint PDF). Those are also referred to as multivariate distribution functions.

The canonical example is the multivariate Gaussian distribution. The Multivariate Gaussian PDF in d dimensions with mean vector $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ is given for any $\mathbf{x} \in \mathbb{R}^d$ as

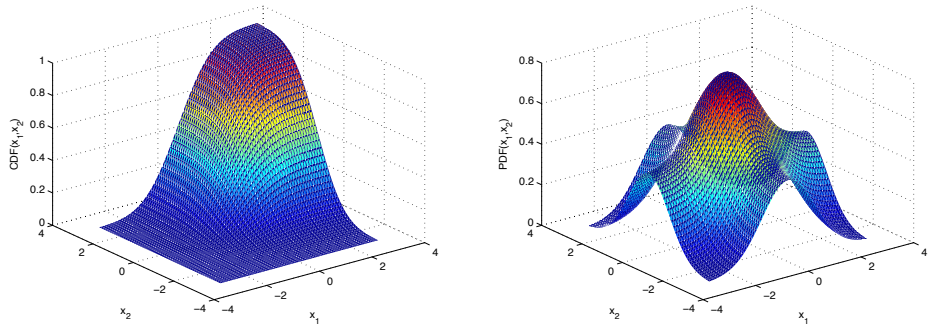
$$f(\mathbf{x};\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right), \tag{5.16}$$

where $|\Sigma|$ denotes the determinant of the matrix Σ, and we assume that Σ has a unique inverse (or the determinant does not equal 0). Figure (5.3) gives an example of the CDF and the PDF of a Multi-Variate Normal (MVN) distribution with mean $\mu = (0,0)^T$ and $\Sigma = I_2$.

5.1.5 Stochastic Processes

In the context of this course we stick to the following definition of a stochastic process.

Definition 15 (Stochastic Process) *A stochastic process Z is a sequence of random variables $Z = \{Z_1, Z_2, \ldots, Z_n\}$ where each Z_t takes values in $\mathbb{R}$. It is entirely defined by its joint probability distribution. A sequence of values $\{z_1, \ldots, z_n\}$ is a* realization *of this process if it is assumed to be sampled from mentioned stochastic process.*



(a) (b)

Figure 5.3: Example of a Multivariate Normal Distribution of two independent random variables. (a) the CDF, and (b) the PDF.

Formally, we consider again an experiment with sample space Ω. Now, a stochastic process is a mapping from a sample ω into a path, i.e. a possibly infinite sequence of numbers. The mathematical description of a path is as a function mapping time instances into its corresponding element in the array of numbers. For example let $z = (z_1, z_2, \dots)$ denote such an array, then there $z(t) = z_t$ for each $t = 1, 2, \dots$. This indicates that there is no formal difference between a function and an indexed array, either concept is a mere notational convention. Since in the context of this course we will primarily be interested in discrete stochastic processes where t could take a finite or countably infinite number of values, we will stick to the indexing notation.

While this looks like a very general definition, it excludes quite some cases which are of interest in different situations. Firstly, we restrict attention to finite sequences of random variables, where the index t ('time') runs from 1 to n. Alternatives are found when the index t can take on continuous values ('Continuous stochastic processes'), or even more complex objects belonging to a well-defined group ('Empirical processes').

The subtlety of such processes goes as follows. A stochastic process is a mapping from an event ω to a corresponding time-series, denoted as a realization of this process. The expected value of a stochastic process is the average of all time-series associated to all possible events. That is, the expected value of a stochastic process is a deterministic timeseries! Let this timeseries be denoted as $m = (\dots, m_0, m_1, m_2, \dots)$. In general, one is interested of a value of one location of this timeseries, say m_t Similarly, one can come up with a definition of the covariance associated to a stochastic process, and the covariance evaluated for certain instances. Often, one makes a simplifying assumption on this series by assuming stationarity:

Definition 16 (Stationary Process) *A stochastic process $\{Z_t\}_t$ is said to be (wide-sense) stationary in case the first two moments do not vary over time, or*

$$\begin{cases} \mathbb{E}[Z_t] = \dots \mathbb{E}[Z_t] = \dots = \mathbb{E}[Z_n] & = m \\ \mathbb{E}[(Z_t - m_t)(Z_{t-\tau} - m_{t-\tau})] & = \mathbb{E}[(Z_{t'} - m_{t'})(Z_{t'-\tau} - m_{t'-\tau})] = r(\tau), \end{cases} \tag{5.17}$$

for all t, t', where one has $|m| < C$ and $|r(\tau)| \leq c$ for some finite constants C, c.

This implies that the covariance structure of a stochastic process has a simple form: namely, that all covariances associated to two different locations are equal. This structural assumption makes stochastic processes behave very similar as the LTIs as studied before (why?). In the context of system identification, one is often working assuming a slightly weaker condition on the involved stochastic processes:

Definition 17 (Quasi-Stationary Process) *A stochastic process $\{Z_t\}_t$ is said to be quasi-stationary in case one has*

$$\begin{cases} \mathbb{E}[Z_t] = m_t \\ \mathbb{E}[Z_t Z_s] = r(t,s) \\ \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} r(t, t-\tau) = r(\tau), \end{cases} \tag{5.18}$$

where for all t one has $|m_t| < C$ and $|r(\tau)| \leq c$ for some finite constants C, c.

That is, we allow the mean of the signal to vary over time, but assume the average covariance be independent over time. The reason that this definition is quite useful is that systems will typically be expressed as stochastic process Y satisfying for all $t = 1, \ldots, n$ that

$$\mathbb{E}[Y_t] = h_t(u_1, \ldots, u_t), \tag{5.19}$$

where h_t is a filter, and $\{u_1, \ldots, u_n\}$ are deterministic. That means that the mean is almost never time-invariant.

An important problem is that in practice we are only given a single realization of a stochastic process. This observation seems to imply that there is nothing much we as a statistician can do. Surely, we must work with expectations of stochastic quantities for which we have only one sample from. And we know that a average of only one sample gives a very poor estimate of the expectation of this sample. Luckily, there is however a way to go ahead. We can shift a bit further in the stochastic process, and uses the so collected samples to build up a proper estimate. If such estimate would indeed converge to the expected value, one says that the process under study is *ergodic*:

Definition 18 (Ergodic Process) *A stochastic process $\{Z_t\}_t$ is said to be ergodic if for any $\tau = 0, 1, \ldots$ one has*

$$\begin{cases} \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} Z_t = \mathbb{E}[Z] \\ \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} Z_t Z_{t-\tau} = \mathbb{E}[Z_t Z_{t-\tau}]. \end{cases} \tag{5.20}$$

This notion turns out to be quite fundamental in the analysis of stochastic processes, but in practice it is often (assumed to be) satisfied.

Practically, perhaps the most useful stochastic process is the following.

Definition 19 (Zero Mean White Gaussian Noise) *A stochastic process $Z = \{Z_1, \ldots, Z_n\}$ is called a zero mean white Gaussian noise in case*

(Zero Mean): For each t one has $\mathbb{E}[Z_t] = 0$.

(Gaussian): Each subset of elements is jointly Gaussian distributed with zero mean.

(White): The different elements are uncorrelated.

slightly weaker is

Definition 20 (Zero Mean White Noise) *A stochastic process $Z = \{Z_1, \ldots, Z_n\}$ is called a zero mean white noise in case (i) For each t, $\mathbb{E}[Z_t] = 0$, (ii) the variances of the elements are bounded (i.e. the first two moments exists), and (iii) the different elements are uncorrelated.*

The naming 'white' is historically connected to the related Brownian motion, having a non-vanishing correlation matrix. A popular motivation is that such 'white' noise signal has no 'colouring' due to the fact that all frequencies in its spectrum are equally present.

5.1.6 Interpretations of Probabilities

While notions of probability, random variables and derived concepts were formulated rock-solid (i.e. axiomatic) by A.N. Kolmogorov in the 1930s, there is still ample discussion of what those quantities stand for. This discussion is not only to be fought by philosophers of science, but ones' position here has far-reaching practical impact as well. Rather than surveying the different schools of thought on this matter, let us give the following example by Chernoff suggesting that one should not be guided only by formulas, definitions and formal derivations only in this discussion: statistic is in first instance a practical tool conceived in order to assist decision making in reality. Be critical of its use!

> 'The metallurgist told his friend the statistician how he planned to test the effect of heat on the strength of a metal bar by sawing the bar into six pieces. The first two would go into the hot oven, the next two into the medium oven and the last two into the cool oven. The statistician, horrified, explained how he should randomise in order to avoid the effect of a possible gradient of strength in the metal bar. The method of randomisation was applied, and it turned out that the randomised experiment called for putting the first two into the hot oven, the next two into the medium oven and the last two into the cool oven. "Obviously, we can't do that," said the metallurgist. "On the contrary, you have to do that," said the statistician."'

A point of this example is that one should remain aware that the stochastic framework is - albeit useful - still a framework, that is, it is not absolute. The other thing to understand from this example is that the stochastic framework is most powerful when n is large. If the number of samples is small, paradoxical situations can occur. They become however less likely to occur if the number of samples grow.

5.2 Statistical Inference

Given a statistical setup ('statistical system') associated to an experiment, perhaps encoded as a number of CDFs or PDFs, one can give solutions to many derived problems. For example one can quantify 'what value to expect next', 'how often does a significance test succeed in its purpose', 'when is an observation not 'typical' under this statistical model', and so on. Statistical inference then studies the question how a statistical system can be identified from associated random values. Often such random variables denote the observations which were gathered while performing an experiment of the studied system. We acknowledge at this point that a statistical system is often an highly abstracted description of the actual experiment, and one rather talks about a 'statistical model underlying the observations', however ambiguous that may sound in the context of this book.

5.2.1 In All Likelihood

Definition 21 (Likelihood Function) *Consider a random value, random vector of stochastic process Z_n which takes values in $\mathbb{Z}$, and with associated cdf F and pdf f (assuming it exists). Consider a family of functions $\{f_\theta : \mathbb{Z} \to \mathbb{R}_+\}_\theta$ indexed by $\theta \in \Theta$. The hope is that this family contains an element f_{θ_*} which is in some sense similar to the unknown f. Then the strictly positive likelihood function $L_n : \Theta \to \mathbb{R}_0^+$ is defined as*

$$L_n(\theta) = f_\theta(Z_n). \tag{5.21}$$

The log-Likelihood *of θ on a sample Z_n is defined as $\ell_n(\theta) \triangleq \log L_n(\theta)$*

Note the similarities as well as dissimilarities of the Likelihood function and the pdf $f(Z_n)$ evaluated in the observations. In the special case that there exist a $\theta_* \in \Theta$ such that $f_{\theta_*} = f$, one has obviously that $f(Z_n) = L_n(\theta_*)$.

Definition 22 (The Maximum Likelihood Estimator) *Assume the values $z \in \mathbb{Z}$ observed during an experiment are assumed to follow a random variable Z taking value in $\mathbb{Z}$, obeying a PDF function f which is only known up to some parameters θ. Then the Likelihood function $L_n(\theta)$ can be constructed. A Maximum Likelihood (ML) estimator $\hat{\theta}$ of θ is defined as*

$$\hat{\theta} \in \operatorname*{argmax}_\theta L_n(\theta). \tag{5.22}$$

A prototypical example goes as follows:

Example 35 (Average as an Estimator) *Given n i.i.d. samples from a random variable Z obeying a Gaussian distribution with fixed but unknown mean μ, and a given variance σ^2, or*

$$f_\mu(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right). \tag{5.23}$$

Then the ML estimator for Then given a sample $\{Z_1, \ldots, Z_n\}$ of length n, each one being an independent copy of the Gaussian distribution of (5.23). Then the ML estimate of μ is given as

$$\hat{\mu} = \operatorname*{argmax}_\mu \ell_n(\mu) = \log \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Z_i-\mu)^2}{2\sigma^2}\right). \tag{5.24}$$

Simplifying the expression and neglecting fixed terms gives the equivalent problem

$$\hat{\mu} = \operatorname*{argmax}_\mu \sum_{i=1}^n -(Z_i - \mu)^2. \tag{5.25}$$

which equals the familiar LS estimator, and the closed form formula is given as

$$\bar{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i. \tag{5.26}$$

Note that this equivalence does not hold any longer if σ is unknown too!

This reasoning can easily be generalized to the case where deterministic explanatory vectors $\{\mathbf{x}_i\}_i$ ('inputs') are available as well. At first, let a statistical model be assumed as follows.

$$Y = \mathbf{x}^T\theta_0 + D, \tag{5.27}$$

where $\mathbf{x} \in \mathbb{R}^d$ is a deterministic, fixed vector, $\theta_0 \in \mathbb{R}^d$ is a fixed vector which happened to be unknown. The last term D is a random variable which takes values into $\mathbb{R}$ following certain rules of probabilities. Specifically, we have that it follows a PDF given as $f_D(\cdot;\mu,\sigma)$ with mean $\mu \in \mathbb{R}$ and standard deviation $\sigma \in \mathbb{R}$, defined $\forall z \in \mathbb{R}$ as

$$f(z;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right). \tag{5.28}$$

We assume that e follows a PDF $f(\cdot,0,\sigma)$. This model implies that also Y is a random variable following a PDF with mean $f(\cdot;\mathbf{x}^T\theta,\sigma)$. A last important assumption which is often given is that the samples we observe from this model are independently sampled. That is, the n samples $\{Y_i\}_{i=1}^n$ satisfy the model

$$Y_i = \mathbf{x}_i^T\theta_0 + D_i, \tag{5.29}$$

where $\{D_i\}_{i=1}^n$ are independent, identically distributed (i.i.d.), that is each sample e_i does not contain information about a sample e_j with $i \neq j$, except for their shared PDF function.

Definition 23 (I.I.D.) *A set of random variables $\{D_1,\ldots,D_n\}$ which each take values in $\mathbb{R}$, contains independent random variables iff for all $i \neq j = 1,\ldots,n$ as*

$$\mathbb{E}[D_iD_j] = \mathbb{E}[D_i]\mathbb{E}[D_j]. \tag{5.30}$$

Those random variables are identically distributed iff they share the same probability function, or if D_i has PDF f_i one has

$$f_i(z) = f_j(z), \tag{5.31}$$

for all $i,j = 1,\ldots,n$ and z ranging over the domain $\mathbb{R}$. If both conditions are satisfied, then the set $\{D_1,\ldots,D_n\}$ is denoted as independently and identically distributed, or abbreviated as i.i.d.

This assumption plays a paramount role in most statistical inference techniques. However, it is exactly on those assumptions that time-series analysis, and estimation for dynamical models will deviate. That is, in such context often past errors D_t will say something about the next term D_{t+1}. This cases will be investigated in some details in later chapters.

Now we can combine the different elements. The corresponding Likelihood function of the model of eq. (5.27), the assumed form of the errors as in(5.28), as well as the i.i.d. assumption results in the following Likelihood function expressed in terms of the parameter vector θ:

$$L_n(\theta) = f(Y_1,\ldots,Y_n) = \prod_{i=1} f(Y_i - \mathbf{x}_i^T\theta;0,\sigma). \tag{5.32}$$

Note again, that this function equals the PDF of the n samples in case $\theta = \theta_0$. Now the Maximum Likelihood Estimate is given as

$$\hat{\theta} = \operatorname*{argmax}_\theta L_n(\theta), \tag{5.33}$$

Working out the right-hand side gives

$$\hat{\theta} = \operatorname*{argmax}_{\theta} \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(Y_i - \mathbf{x}_i^T\theta)^2}{2\sigma}\right) \propto -\sum_{i=1}^{n} \left(Y_i - \mathbf{x}_i^T\theta\right)^2. \tag{5.34}$$

In this special case, it is seen that the ML estimator is found by solving the least squares problem $\hat{\theta} = \operatorname{argmin}_\theta \sum_{i=1}^{n}(Y_i - \mathbf{x}_i^T\theta)^2$. That is, in case $\sigma > 0$ is fixed. In case σ needs to be estimated as well, the Likelihood function becomes more intricate.

5.2.2 Power and Limitations of ML

The ML estimator has a number of exquisite properties as discussed in some detail in the abundant literature of statistical inference. Perhaps the most direct property is that the ML method is efficient, that is it performs as well as any estimator under the given assumptions.

Definition 24 (Unbiased) *Given any estimator $\theta_n = g(Y_1, \ldots, Y_n)$ which approximates a parameter (vector) θ_0 based on a sample $Y_1, \ldots, Y_n$. Then this estimator is unbiased iff*

$$\mathbb{E}[\theta_n] = \mathbb{E}[g(Y_1, \ldots, Y_n)] = \theta_0. \tag{5.35}$$

Theorem 3 (EMSE) *Given any estimator $\theta_n = g(Y_1, \ldots, Y_n)$ of θ_0 , then the performance of this estimator can be expressed as the Expected Mean Square Error (EMSE)*

$$V(g) = \mathbb{E}\left\|g(Y_1, \ldots, Y_n) - \theta_0\right\|_2^2 = \mathbb{E}\left\|\theta_n - \theta_0\right\|_2^2. \tag{5.36}$$

Theorem 4 (Covariance of Estimate) *Consider a class of PDFs $\{f_\theta : \theta \in \mathbb{R}^d\}$ where θ_0 is the (unknown) one underlying the data observations, that is f_{θ_0} is the PDF underlying the sample $Y_1, \ldots, Y_n$. Given any estimator $\theta_n = g(Y_1, \ldots, Y_n)$ of a parameter vector $\theta_0 \in \mathbb{R}^d$, then the covariance of this estimator $\mathbf{R}(g) \in \mathbb{R}^d \times d$ can be expressed as*

$$\mathbf{R}(g) = \mathbb{E}\left[(g(Y_1, \ldots, Y_n) - \theta_0)(g(Y_1, \ldots, Y_n) - \theta_0)^T\right] = \mathbb{E}\left[(\theta_n - \theta_0)(\theta_n - \theta_0)^T\right]. \tag{5.37}$$

Theorem 5 (Cramér-Rao Lowerbound) *Given any estimator $\theta_n = g(Y_1, \ldots, Y_n)$ of θ_0 which is unbiased, then*

$$\mathbf{R}(g) \succeq \mathbf{I}_{\theta_0}^{-1}, \tag{5.38}$$

where the so-called Fisher information matrix $\mathbf{I}_{\theta_0}$ is defined as

$$\begin{aligned} \mathbf{I}_{\theta_0} &= \mathbb{E}\left[\frac{d \log f_\theta(Y_1, \ldots, Y_n)}{d\theta} \frac{d^T \log f_\theta(Y_1, \ldots, Y_n)}{d\theta}\Big|_{\theta=\theta_0}\right] \\ &= -\mathbb{E}\left[\frac{d^2 \log f_\theta(Y_1, \ldots, Y_n)}{d\theta^2}\Big|_{\theta=\theta_0}\right]. \end{aligned} \tag{5.39}$$

The general proof can e.g. be found in Ljung's book on System Identification, Section 7.4 and Appendix 7.A. The crucial steps are however present in the following simplified form.

Lemma 6 (Cramér-Rao, simplified) *Consider the case where we have a class of PDFs with a single parameter, say $\{f_\theta : \theta \in \mathbb{R}\}$, such that there is a $\theta_0 \in \mathbb{R}$ such that f_{θ_0} underlies the sample $Y_1, \ldots, Y_n$. Let $\theta_n = g(Y_1, \ldots, Y_n)$ be an unbiased estimator of θ_0, then*

$$\mathbb{E}\left[(\theta_n - \theta_0)^2\right] \geq \frac{1}{m_{\theta_0}}. \tag{5.40}$$

where

$$m_{\theta_0} = \mathbb{E}\left[\frac{df_\theta}{d\theta}\Big|_{\theta=\theta_0}\right]^2. \tag{5.41}$$

5.3 Least Squares Revisited

Let us now turn attention once more to the least squares estimator, and derive some statistical properties on how this works. The analysis is mostly asymptotic, that is properties are derived as if we would have that $n \to \infty$. This is in practice not the case obviously, but those results give nevertheless a good indication of how the estimators behave. We will work under the assumptions that n observations $\{Y_i\}_{i=1}^n$ follow the model

$$Y_i = \mathbf{x}_i^T \theta_0 + D_i, \tag{5.42}$$

where $\theta_0 \in \mathbb{R}^d$ is the *true* parameter which is fixed and deterministic, but which happens to be unknown to us. Here $\{D_1, \ldots, D_n\}$ are i.i.d. and hence is uncorrelated, all have zero mean $\mathbb{E}[D_i] = 0$, but have a fully unspecified PDF except for some regularity conditions. Still the LS estimator has very good properties, although it does not correspond necessarily to a ML estimator.

Note at this point the conceptual difference of the deterministic model

$$y_i = \mathbf{x}_i^T \theta + \epsilon_i, \tag{5.43}$$

where $\{\epsilon_1, \ldots, \epsilon_n\}$ are (deterministic) residuals, depending (implicitly) on the choice of θ. For this model, there is no such thing as a true parameter. Moreover, there is no stochastic component, such that e.g. $\mathbb{E}[\epsilon_i] = \epsilon_i$. Note the important differences between the 'true' noise $\{D_i\}_i$ under model (5.42), and the residuals $\{\epsilon_i\}_i$. They only equal each other in the special case that the model (5.42) is assumed to underly the observations $\{y_i\}_i$ (that is if $\{y_i\}_i$ are samples from $\{Y_i\}_i$), and $\theta = \theta_0$ (that is, we have estimated the true parameter *exactly*). Often one makes this assumption that $\{y_i\}_i$ are samples from $\{Y_i\}_i$, but one has merely that $\theta \approx \theta_0$ and the residual terms do not obey the stochastic properties of the noise!

Example 36 (Average, Ct'd) *Consider again the model $Y_i = \theta_0 + D_i$ where $\theta_0 \in \mathbb{R}$ is fixed but unknown, and $\{D_i\}_i$ are i.i.d. random variables with zero mean and standard deviation σ. Then the LS estimator θ_n of θ_0 solves the optimisation problem*

$$V_n(\theta_n) = \min_\theta \sum_{i=1}^n (Y_i - \theta)^2, \tag{5.44}$$

for which the solution is given as $\theta_n = \frac{1}{n}\sum_{i=1}^n Y_i$. How well does θ_n estimate θ_0?

$$\mathbb{E}[\theta_0 - \theta_n]^2 = \mathbb{E}\left[\theta_0 - \frac{1}{n}\sum_{i=1}^n Y_i\right]^2 = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\theta_0 - Y_i)\right]^2 = \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}[D_i^2] = \frac{\sigma^2}{n}. \tag{5.45}$$

Now we answer the question whether the minimal value $V_n(\theta_n)$ says something about the standard deviation σ. Therefore, we elaborate the objective for the optimal θ_n, which gives

$$\begin{aligned}
V_n(\theta_n) &= \mathbb{E}\sum_{i=1}^{n}\left(Y_i - \frac{1}{n}\sum_{i=1}^{n} Y_i\right)^2 \\
&= \mathbb{E}\sum_{i=1}^{n}\left((Y_i-\theta_0) - (\frac{1}{n}\sum_{j=1}^{n} Y_j - \theta_0)\right)^2 \\
&= \mathbb{E}\sum_{i=1}^{n}\left((Y_i-\theta_0)^2 - 2(Y_i-\theta_0)\left(\frac{1}{n}\sum_{j=1}^{n} Y_j - \theta_0\right) + \left(\frac{1}{n}\sum_{j=1}^{n} Y_i - \theta_0\right)^2\right) \\
&= \mathbb{E}\sum_{i=1}^{n}\left(D_i^2 - \frac{2}{n}\sum_{j=1}^{n} D_i D_j + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} D_j D_k\right) \\
&= \sum_{i=1}^{n}\mathbb{E}[D_i^2] - \mathbb{E}\frac{1}{n}\sum_{i=1}^{n} D_i^2 \\
&= (n-1)\,\sigma^2, \qquad (5.46)
\end{aligned}$$

since $\sum_{i=1}^{n}(Y_i - \theta_n) = 0$ by the property of least squares.

Let us now study the covariance and the expected minimal value of the OLS estimate.

Lemma 7 (Statistical Properties of OLS) *Assume the data follows a model $Y_i = \mathbf{x}_i^T\theta + D_i$ with $\{D_1, \ldots, D_n\}$ uncorrelated random variables with mean zero and standard deviation $\sigma > 0$, and $\theta, \mathbf{x}_1, \ldots, \mathbf{x}_n$ are deterministic vectors in $\mathbb{R}^d$. Let the matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ enumerate those such that $\mathbf{X}_i = \mathbf{x}_i^T$ for all $i = 1, \ldots, n$, and assume that $\mathbf{X}$ has full rank such that the inverse $(\mathbf{X}^T\mathbf{X})^{-1}$ is unique. Let θ_n be the LS estimate (as in Chapter 2) solving*

$$V_n(\theta_n) = \min_{\theta} \frac{1}{2}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i^T\theta_n)^2, \qquad (5.47)$$

then

- *The estimate θ_n is unbiased, that is $\mathbb{E}[\theta_n] = \theta_0$.*
- *The covariance of the estimate is given as*

$$\mathbb{E}\left[(\theta_0 - \theta_n)(\theta_0 - \theta_n)^T\right] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}. \qquad (5.48)$$

- *The estimate $V_n(\theta_n)$ implies an unbiased estimate of σ as*

$$\sigma^2 = \frac{2}{n-d}\mathbb{E}[V_n(\theta_n)]. \qquad (5.49)$$

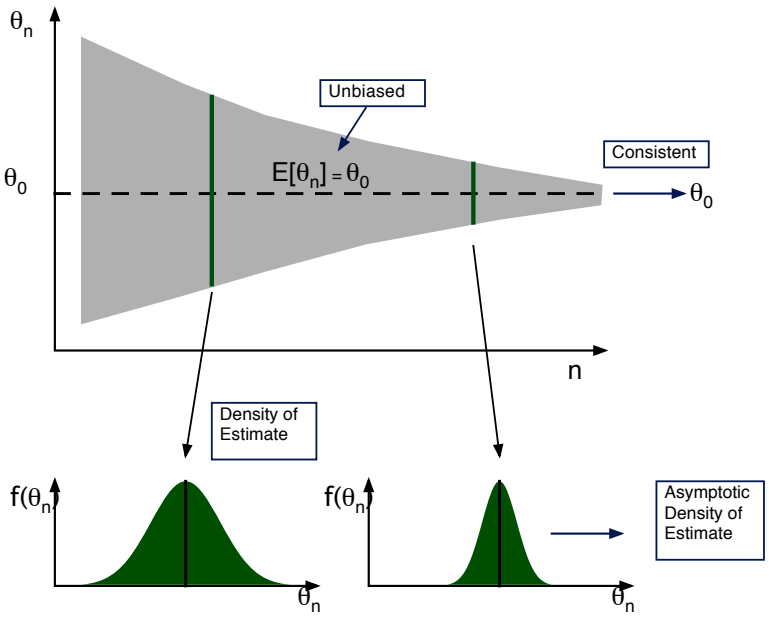


Figure 5.4: Schematical illustration of an *unbiased estimator* θ_n of θ_0. Here $n = 1, 2, \ldots$ denotes the size of the samples $\{Y_1, \ldots, Y_n\}$ on which the estimator $\theta_n = g(Y_1, \ldots, Y_n)$ is based. The estimator is called unbiased if one has for any n that $\mathbb{E}[\theta_n] = \theta_0$. The grey area denoted the possible estimates θ_n for different samples $\{Y_1, \ldots, Y_n\}$. The cross-section of this area for a given n equals the sample distribution, denoted as the 2 bell-shaped curves at the bottom.

Proof: At first, we have the normal equations characterizing θ_n as

$$\theta_n = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\theta_0 + D). \tag{5.50}$$

where $Y = (Y_1, \ldots, Y_n)^T$ and $D = (D_1, \ldots, D_n)^T$ are two random vectors taking values in $\mathbb{R}^n$. Then

$$\theta_n = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\theta_0+D) = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})\theta_0+(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T D = \theta_0+(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T D. \tag{5.51}$$

Taking the expectation of both sides gives

$$\mathbb{E}[\theta_n] = \mathbb{E}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\theta_0 + D)\right] = \theta_0 + \mathbb{E}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T D\right] = \theta_0 + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}\left[D\right] = \theta_0, \tag{5.52}$$

since the vectors $\theta_0, \{\mathbf{x}_i\}$ are deterministic, and hence $\mathbb{E}[\theta_0] = \theta_0, \mathbb{E}[\mathbf{X}] = \mathbf{X}$. This proves unbiasedness of the estimator. Note that the assumption of the vectors $\theta_0, \mathbf{x}_1, \ldots, \mathbf{x}_n$ being deterministic is crucial.

Secondly, the covariance expression can be derived as follows. Here the crucial insight is that we have by assumption of zero mean i.i.d. noise (or white noise) that $\mathbb{E}[DD^T] = \sigma^2 I_n$ where $I_n = \mathrm{diag}(1, \ldots, 1) \in \mathbb{R}^{n\times n}$. Then we have from eq. (5.51) that

$$\begin{aligned}
\mathbb{E}\left[(\theta_0 - \theta_n)(\theta_0 - \theta_n)^T\right] &= \mathbb{E}\left[((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T D)(D^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-T})\right] \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}[DD^T]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-T} \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-T} \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.
\end{aligned} \tag{5.53}$$

Thirdly, the minimal value of the minimization problem is given as

$$\begin{aligned}
V_n(\theta_n) &= \frac{1}{2}(Y - \mathbf{X}\theta_n)^T(Y - \mathbf{X}\theta_n) \\
&= \frac{1}{2}\left(Y^TY - 2Y^T\mathbf{X}\theta_n + \theta_n^T\mathbf{X}^T\mathbf{X}\theta_n\right) \\
&= \frac{1}{2}\left(Y^TY - 2Y^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TY + Y^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TY\right) \\
&= \frac{1}{2}\left(Y^TY - Y^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TY\right).
\end{aligned} \tag{5.54}$$

Hence

$$\begin{aligned}
2V_n(\theta_n) &= Y^T\left(I_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)Y \\
&= (\mathbf{X}\theta_0 + D)^T\left(I_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)(\mathbf{X}\theta_0 + D) \\
&= D^T\left(I_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)D.
\end{aligned} \tag{5.55}$$

Then, using the properties of the trace operator $\mathrm{tr}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = \mathrm{tr}(\mathbf{x}\mathbf{x}^T\mathbf{A})$ and $\mathrm{tr}(\mathbf{A}+\mathbf{B}) = \mathrm{tr}(\mathbf{A}) +$

$\mathrm{tr}(\mathbf{B})$ for $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ and $\mathbf{x} \in \mathbb{R}^n$ gives

$$\begin{aligned} 2\mathbb{E}[V_n(\theta_n)] &= \mathbb{E}\,\mathrm{tr}\left(D^T\left(I_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)D\right) \\ &= \mathbb{E}\,\mathrm{tr}\left(D^T D\left(I_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)\right) \\ &= \mathrm{tr}\left(\mathbb{E}[D^T D]\left(I_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)\right) \\ &= \sigma^2\,\mathrm{tr}(I_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) \\ &= \sigma^2\left(\mathrm{tr}(I_n) - \mathrm{tr}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\right) \\ &= \sigma^2\left(\mathrm{tr}(I_n) - \mathrm{tr}((\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1})\right) \\ &= \sigma^2\left(\mathrm{tr}(I_n) - \mathrm{tr}(I_d)\right) = \sigma^2(n-d). \end{aligned} \tag{5.56}$$

□

This result is slightly generalized as follows.

Theorem 6 (Gauss-Markov Theorem) *Given a model with deterministic values $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$ with fixed but unknown $\theta_0 \in \mathbb{R}^d$ such that*

$$Y_i = \mathbf{x}_i^T\theta_0 + D_i, \tag{5.57}$$

where $\{D_1, \ldots, D_n\}$ are uncorrelated, all have zero mean $\mathbb{E}[D_i] = 0$ and have (finite) equal variances, i.e. $\mathbb{E}[D_i^2] = \cdots = \mathbb{E}[D_n^2] = \sigma^2 < \infty$ (i.e. $\{D_i\}_i$ is homoskedastic). Suppose that we have an estimator $\theta_n = g(Y_1, \ldots, Y_n)$. Then its performance can be measured as the variance

$$V(\theta_n) = \mathbb{E}\left\|\theta_0 - \theta_n\right\|^2. \tag{5.58}$$

Then the linear *estimator achieving the minimal possible variance $V(\theta_n)$ such that it is unbiased $\mathbb{E}[\theta_n] = \theta_0$ is given as*

$$\theta_n = \underset{\theta}{\mathrm{argmin}}\,\frac{1}{2}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i^T\theta)^2. \tag{5.59}$$

This estimator is hence called the Best Linear Unbiased Estimator (BLUE).

An estimator $\theta_n(y_1, \ldots, y_n)$ is *linear* if it satisfies the superposition principle, i.e. if $\hat{\theta} = \theta_n(y_1, \ldots, y_n)$ and $\hat{\theta}' = \theta_n(y_1', \ldots, y_n')$, iff $\hat{\theta} + \hat{\theta}' = \theta_n(y_1 + y_1', \ldots, y_n + y_n')$. In other words, the estimator can be written as a linear combination of the outputs, or

$$\theta_{n,j}(y_1, \ldots, y_n) = \sum_{i=1}^{n} c_{i,j}y_i, \quad \forall j = 1, \ldots, d, \tag{5.60}$$

where $\{c_{i,j}\}$ do not depend on $\{y_i\}$. Indeed the OLS estimate obeys this form as it can be written as

$$\theta_n = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}, \tag{5.61}$$

according to the formulae in chapter 2.

In turn we also have that the least squares estimate needs modification if the coloring of the noise is known to equal a matrix $\mathbf{R}$. The reasoning goes as follows. Assume again that the linear model

$$Y_i = \mathbf{x}_i^T\theta_0 + D_i, \tag{5.62}$$

with $\mathbf{x}_1, \dots, \mathbf{x}_n, \theta \in \mathbb{R}^d$, but where $\{D_1, \dots, D_n\}$ are zero mean random variables with covariance $\mathbf{R}$ such that $\mathbf{R}_{ij} = \mathbb{E}[D_i D_j]$ for all $i, j = 1, \dots, n$. Then the estimator θ_n of θ_0 with minimal expected error is given as

$$\theta_n = \operatorname*{argmin}_{\theta} \frac{1}{2}(Y - \mathbf{X}\theta)^T \mathbf{R}^{-1}(Y - \mathbf{X}\theta). \tag{5.63}$$

This estimator is known as the Best Linear Unbiased Estimate (BLUE). The following simple example illustrates this point:

Example 37 (Heteroskedastic Noise) *Consider again the model*

$$Y_i = \theta_0 + D_i, \tag{5.64}$$

where $\theta_0 \in \mathbb{R}$ and $\{D_i\}_i$ are uncorrelated (white) zero mean stochastic variables, with variances $\mathbb{E}[D_i^2] = \sigma_i^2 > 0$ which are different for all samples, i.e. for all $i = 1, \dots, n$. Then the BLUE estimator becomes

$$\theta_n = \operatorname{argmin} \sum_{i=1}^{n} (Y - 1_n\theta_0)^T \mathbf{M}(Y - 1_n\theta_0), \tag{5.65}$$

where

$$\mathbf{M} = \begin{bmatrix} \sigma_1^{-2} & & 0 \\ & \ddots & \\ 0 & & \sigma_n^{-2} \end{bmatrix}. \tag{5.66}$$

The solution is hence given as

$$1_n^T \mathbf{R} 1_n \theta_n = 1_n^T \mathbf{R} Y, \tag{5.67}$$

where $Y = (Y_1, \dots, Y_n)^T$ takes elements in $\mathbb{R}^n$. Equivalently,

$$\theta_n = \frac{1}{\sum_{i=1}^n \sigma_i^2} \sum_{i=1}^{n} \frac{Y_i}{\sigma_i^2}. \tag{5.68}$$

Note that the influence of a sample Y_i in the total sum is small in case it is inaccurate, or σ_i is large, and vice versa.

Lets now give an example where the inputs are stochastic as well, or

Example 38 (Stochastic Inputs) *Assume the observations $\{Y_i\}_i$ are modeled using the random vectors $\{X_i\}_i$ taking values in $\mathbb{R}^d$, $\theta_0 \in \mathbb{R}^d$ is deterministic but unknown*

$$Y_i = X_i^T \theta_0 + D_i, \tag{5.69}$$

where $\{D_i\}_i$ are zero mean i.i.d. and are assumed to be independent from $\{X_i\}$. *This assumption is crucial as we will see later. Then the above derivations still hold more or less. Consider the LS estimate θ_n. It is an unbiased estimate of θ_0 as could be seen by reproducing the above proof. Let $D = (D_1, \dots, D_n)^T$, then*

$$\mathbb{E}[\theta_n] = \mathbb{E}\left[(X^T X)^{-1} X^T (X\theta_0 + D)\right] = \theta_0 + \mathbb{E}\left[(X^T X)^{-1} X^T D\right] = \theta_0 + \mathbb{E}[(X^T X)^{-1} X^T]\mathbb{E}\left[D\right] = \theta_0, \tag{5.70}$$

where X is the random matrix taking elements in $\mathbb{R}^{n\times d}$ such that $\mathbf{e}_i^T X = X_i$ for all $i = 1, \ldots, n$. Here we need the technical condition that $\mathbb{E}[(X^TX)^{-1}]$ exists, or that X^TX is almost surely full rank. This equation implies asymptotic unbiasedness of the estimator θ_n. Similarly, one can proof that the covariance of θ_n is given as

$$\mathbb{E}\left[(\theta_0 - \theta_n)(\theta_0 - \theta_n)^T\right] = \sigma^2 \mathbb{E}\left[(X^TX)^{-1}\right]. \tag{5.71}$$

Note that $\mathbb{E}[(X^TX)^{-1}] \neq (\mathbb{E}[X^TX])^{-1}$ exactly, although such relation holds asymptotically since $\lim_{n\to\infty} \frac{1}{n} X_i X_i^T \approx \mathbb{E}[XX^T]$. Finally, the minimal value $V_n(\theta_n)$ satisfies

$$\sigma^2 = \frac{2}{n-d}\mathbb{E}[V_n(\theta_n)]. \tag{5.72}$$

The key property which causes this to work is the fact that $\mathbb{E}[X_tD_t] = \mathbb{E}[X_t]\mathbb{E}[D_t]$. This condition was trivially satisfied if $\mathbf{x}_t$ were deterministic, leading to the many optimality principles of least squares estimates as stated in the Gauss-Markov theorem.

5.4 Instrumental Variables

Example 39 (Dependent Noise) *Consider again the following model using the definitions as given in the previous example:*

$$Y_t = X_t^T\theta_0 + D_t, \tag{5.73}$$

and D_t is a random variable with bounded variance and zero mean, then the least squares estimate θ_n is given by the solution of

$$\theta_n = \left(\frac{1}{n}\sum_{t=1}^{n} X_tX_t^T\right)^{-1}\left(\frac{1}{n}\sum_{t=1}^{n} X_tY_t\right). \tag{5.74}$$

In case $n \to \infty$, one has by definition that

$$\theta_n = \left(\mathbb{E}\left[X_tX_t^T\right]\right)^{-1}\mathbb{E}\left[X_tY_t\right]. \tag{5.75}$$

Assuming that $\mathbb{E}[X_tX_t^T]$ exists and is invertible, one can write equivalently that

$$\begin{aligned}
\theta_0 - \theta_n &= \theta_0 - \left(\mathbb{E}\left[X_tX_t^T\right]\right)^{-1}\mathbb{E}\left[X_tY_t\right] \\
&= \left(\mathbb{E}\left[X_tX_t^T\right]\right)^{-1}\mathbb{E}\left[X_tX_t^T\right]\theta_0 - \left(\mathbb{E}\left[X_tX_t^T\right]\right)^{-1}\mathbb{E}\left[X_t(X_t^T\theta_0 + D_t)\right] \\
&= \left(\mathbb{E}\left[X_tX_t^T\right]\right)^{-1}\mathbb{E}\left[X_tD_t\right].
\end{aligned} \tag{5.76}$$

And the estimate θ_n is only (asymptotically) unbiased if $\mathbb{E}\left[X_tD_t\right] = 0_d$.

This reasoning implies that we need different parameter estimation procedures in case the noise is dependent on the inputs. Such condition is often referred to as the 'colored noise' case. One way to construct such an estimator, but retaining the convenience of the LS estimator and corresponding normal equations goes as follows.

We place ourselves again in a proper stochastic framework, where the system is assumed to be

$$Y_i = X_i^T\theta_0 + D_i, \tag{5.77}$$

where $X_1, \ldots, X_n, \theta_0 \in \mathbb{R}^d$ are random vectors, and $\{D_i\}$ is zero mean stochastic noise. As in the example this noise can have a substantial coloring, and an ordinary least squares estimator wont give consistent estimates of θ_0 in general. Now let us suppose that we have the random vectors $\{Z_t\}_t$ taking values in $\mathbb{R}^d$ such that

$$\mathbb{E}\left[Z_t D_t\right] = 0_d. \tag{5.78}$$

That is, the instruments are orthogonal to the noise. Then the IV estimator θ_n is given as the solution of $\theta \in \mathbb{R}^d$ to the following system of linear equations

$$\sum_{t=1}^{n} Z_t (Y_t - X_t^T \theta) = 0_d, \tag{5.79}$$

where expectation is replaced by a sample average. That means that we estimate the parameters by imposing the sample form of the assumed independence: that is the estimated model necessarily matches the assumed moments of the involved stochastic quantities. Note that this expression looks similar to the normal equations. If $\sum_{t=1}^{n}(Z_t X_t^T)$ were invertible, then the solution is unique and can be written as

$$\theta_n = \left(\sum_{t=1}^{n} Z_t X_t^T\right)^{-1} \left(\sum_{t=1}^{n} Z_t Y_t^T\right). \tag{5.80}$$

So the objective for us id to design instruments, such that

- The instruments are orthogonal to the noise, or $\mathbb{E}\left[Z_t D_t\right] = 0_d$.
- The matrix $\mathbb{E}[Z_t X_t^T]$ were of full rank, such that also with high probability $\sum_{t=1}^{n}(Z_t X_t^T)$ has a unique inverse.

Example 40 *A common choice in the context of dynamical systems for such instruments goes as follows. Assume that the random vectors X_t consists of delayed elements of the output $Y_{t-\tau}$ of the system which cause the troublesome correlation between D_t and X_t. This is for example typically the case in an ARMAX model. Then a natural choice for the instruments would be to take delayed entries of the input $\{U_t\}_t$ of the system*

$$Z_t = \left(U_{t-1}, \ldots, U_{t-d}\right), \tag{5.81}$$

which takes values in $\mathbb{R}^d$. This is a good choice if the inputs were assumed to be independent of the (colored) noise.