TOPICS IN ECONOMETRICS

Notes for MSc by

MICHAEL CURRAN

TRINITY COLLEGE DUBLIN DEPT. OF ECONOMICS
2012

Contents

Preface

The goal of this module is to introduce more advanced econometric topics in addition to building upon the first term's work on Econometrics. The first two chapters are background reading on measure theoretic probability and asymptotic theory that underlie most of what we will be covering in the course. They relate to some of the work you covered in maths-camp in September. Most proofs will be omitted as this course is primarily focused on applied econometrics, reflecting the majority of interests of the department and of many past students of the MSc program. While theoretical aspects certainly are touched upon throughout the course to provide some balance, these are not developed in much detail. Some of you may end up working on theoretical issues at central banks or even at academic institutes where you may decide to specialise in theoretical economics and / or theoretical econometrics. The material in the this course, particularly in the first two chapters will provide some elementary foundations for later courses you may choose to take if you go in this direction. At times, you may feel that what you are studying is rather theoretical. To an extent, you are correct, but practically all of the MSc program is oriented towards the practice of applied economics from a theoretical angle rather than on developing your abilities towards the theorist route. Do not misinterpret this, however! Certainly, you will be well trained to pursue theoretical economics after this course and you will also be ready for further study or work in applied fields too. What you take from this course will be extremely beneficial for most areas (at least related to economics) you could choose afterwards, but keep in mind that you will receive the most specific training from 'on-the-job' training, whereas here you will learn a great deal of transferable skills – e.g. the use of Stata for most statistical computing, MATLAB for more creative, advanced programming, presentation experience, project experience, a certain amount of mathematical and problem solving skills and a greater appreciation and understanding of issues in applied economics work in addition to a good introduction to most of the main topics in theoretical and applied micro and macro economists have been thinking about over the last ten to fifty years.[1]

[1]Stata and MATLAB seem to be the optimum combination of computer package skills that form the toolbox of most economists, reflected in their wide use across both top American and European schools. You will be hard pressed to find a problem in economics that neither can be used for, though there are some rare instances where of course, more specialised packages might have certain slight advantages; e.g. Mathematica for computing derivatives

Since what I will be covering in the course will rely on definitions and theorems from the first two chapters, I may neglect to define these theorems and definitions from time to time; rest assured, if they are not explained in the later chapters, they will be included in one of the first two chapters. While the material is not directly examinable, you need to understand it to understand many parts of the course and you could be examined indirectly on it, for instance through applying the law of iterated expectations and law of total probability to put bounds on some quantity of interest in a question relating to chapter 3.

One recurring theme throughout the course will be identification. That is, suppose we look at the entire population and want to know some statistic, for instance the mean. Given the available information, what can we say about the mean? What can we say about it if some observations are missing? What bounds – if any – can we put on the mean so that we can tell it is between these bounds? We will study this topic in detail in chapter 3 as a motivation for the rest of the course.

Once we have dealt with the initial 'housekeeping' chapters, we will be free to concentrate on the central part of this course. While this course is provisionally called 'Topics in Econometrics', it could also be called 'Stationary Univariate Time Series' for the amount of time we will be devoting to this area. We will first cover the basics, some of you who took Econometrics before will have covered this, i.e. auto-regressive, moving average and auto-regressive moving average models, auto-correlation and partial auto-correlation functions, and identifying, estimating, testing and forecasting with these models. These topics will be the concern of chapter 4.

Chapter 5 is a further development of chapter 4 in relation to forecasting. A shorter chapter, here we will devote time to discussing the use of forecasts in model assessment as well as how to conduct forecast assessment itself. We will conclude this chapter with an introduction to the problem of forecasting with many predictors. For the interested reader, this topic could be further explored once you have completed Prof Benétrix' part of the course on vector-auto-regressions (VARs), as the issues are most interesting in the multivariate context.

Most of your study of econometrics thus far will most likely have centered around linear models. However, when we are studying or trying to model volatility for instance, non-linear models are essential. In fact anything that is not a linear model, is non-linear, so while initially they may seem like special cases, they are actually more general on further reflection. These models are being more pervasively employed by the profession now. While initially they were used in finance, with only the likes of Engle's ARCH paper in 1982 that won him the Nobel prize making any significant noise in macro, the advent of the discovery of the Great Moderation – the decline in aggregate volatility in

in solving non-linear models or the Fortran language for raw power in number crunching areas such as non-linear simulation for estimation and forecasting or OpenMP/MPI/CUDA languages for parallel coding when you have big problems in terms of size and / or speed.

most US time-series data since the early 1980s (see Kim & Nelson (1998), McConnell & Pérez-Quirós (2000) and Blanchard & Simon (2001) for original work and Stock & Watson (2002) for the popularisation of the term they coined the 'great moderation') – has been accompanied by a huge growth in the use of non-linear models in macroeconomics, e.g. the work of Jesús Fernández-Villaverde at University of Pennsylvania, Juan Rubio-Ramírez at Duke University, Martin Uribe at Colombia University, Nicholas Bloom at Stanford University and James Hamilton at UC San Diego to name but a few. Accordingly and given the interest in this area, we will cover the three most popular non-linear models, the ARCH model (and its variants), Markov Switching models and Stochastic Volatility models. These models are the subject of chapter 6.

The typical method in macro is to formulate or cast a model into canonical form (sometimes involving log-linear or nonlinear representations and possibly linearising through Taylor series or log-linear approximations), solve a model through specific techniques (linear: Blanchard & Kahn / Sims' / Klein's / undetermined coefficients approach or nonlinear: iteration (dynamic programming / value-function iterations / policy-function iterations), perturbation or projection (finite elements / orthogonal polynomials) techniques), prepare data (remove the effects of seasonality or equivalently isolate cycles and remove trends for example) including summarising time series behaviour when you can observe all variables (e.g. ARMA / VAR models and summary statistics) and representing it in an amenable form (the state space representation is a popular form) and finally estimate the model and possibly run a number of post-estimation exercises. After a brief discussion of the state-space form at the start of chapter 7, which we will take as given, we will focus on modern estimation and post-estimation techniques. Given a state at time s, say capital, we will want to 'filter' an observable at the same time, say consumption at time s. Once we have estimated our equation(s) through the whole observed time period we may want to 'smooth' backwards the evolution of a variable given the information we have up to and including the final period T. This is useful for instance when estimating through 'filters' and then plotting 'smoothed' estimates through 'smoothers'; filters use information up to but no further than time t while smoothers use information on both sides (up to time T). A study of time-series would not be complete without an exposure to the flip-side of the coin to the time domain, *viz* the frequency domain; note that this is a topic where many results from the first two chapters will be relevant, particularly on complex analysis. We rarely have closed form, analytic solutions in advanced economics. True, there are a number of analytical methods to get around solving complicated multiple integrals too. Traditionally, we used tricks. However, the modern approach makes use of simulation methods hitherto not possible. Recent advances in micro and macro theory, statistics and computers allowed econometricians to develop and apply sophisticated procedures at a relatively low cost to bypass previously insurmountable problems and in some cases even to significantly mitigate issues like the curse of dimensionality. Depending on time, we will conclude this course with a brief introduction to the use of simulation methods for practical purposes including a small bit of theory to provide

justification for what we are doing with our computers. The idea is that when we cannot obtain exact analytical solutions (at least not without great difficulty) or solve complicated multiple integrals, simulation methods provide efficient methods for obtaining very good approximations to the solution we are looking for. Loosely speaking, we can draw thousands of times to simulate random variables from particular distributions and thereby obtain thousands of answers whose distribution approximately converges (in some specific manner) to the distribution of interest; loosely speaking this is a rough application of the analogy principle.

You can find a bibliography at the end of the book in addition to an index of terminology, names, etc. included in the chapters. Any mistakes are mine alone and as this is a first rough draft, please email any comments you may have to mpcurran@tcd.ie.

Chapter 1

Probabilistic Foundations Underlying Econometrics

There are plenty of measure-theoretic definitions to get through in this section that underlie the rest of the course and hence it is important that you understand this material. Let us think first about the real line, $\mathbb{R}$. We know that any interval $[a, b]$ has length $b - a$ and that any two disjoint intervals $[a, b]$ and $[c, d]$ have length $(b - a) + (d - c)$. We will say that $b - a$ is the *measure* of the set $[a, b]$. Measure theory essentially deals with extending these concepts to arbitrary subsets of $\mathbb{R}$. We will look for functions $\mu : F \longrightarrow [0, \infty]$ where $F \subset \mathcal{P}(\mathbb{R})$, the set of all subsets of $\mathbb{R}$ including $\mathbb{R}$ and $\emptyset$. If $C \in F$, then $\mu(C)$ is the *measure* of C. Hopefully μ will be a function satisfying properties such as the ability to measure every possible subset of $\mathbb{R}$, the ability to add lengths of disjoint intervals, to be invariant to translations, rotations and reflections of the set C and to have measure one for the unit interval, i.e. $\mu([0, 1]) = 1$. However, there is no such function that exists on all the sets of $\mathcal{P}(\mathbb{R})$; see for instance the Vitali sets. To solve this problem, we will have to modify the property about being able to measure every possible subset of $\mathbb{R}$ and allow F to be a strict subset of $\mathcal{P}(\mathbb{R})$; we must let it be a σ-algebra, which we need to define – this leads us nicely into first section on probability spaces.

1.1 Probability Spaces

Definition 1.1. The *sample space* is the set of all possible outcomes from an experiment and denoted Ω. Any element of the sample space $\omega \in \Omega$ is called an *elementary event*, while any subset of the sample space $E \subset \Omega$ is called an *event*.

Definition 1.2. For any set S, the *power set of* S, $P(S)$ is the set of all subsets of S including the empty set and S itself.

Remark 1.3. Usually in experiments, we are concerned with computing the probability of occurrence of particular subsets of the sample space. So we would

like to define probabilities for each member of the collection of all subsets of our sample space, i.e. the *power set* of the sample space. However, this is only possible when the sample space is finite or countably infinite. There exist uncountably infinite sample spaces where we can not do this, such as the Cantor set. This motivates the next group of definitions where we need to define particular subcollections of subsets of the sample space where we can compute probabilities. Such subcollections of subsets turn out to have the structure of a *sigma algebra*.

Definition 1.4. A non-empty collection F of subsets of Ω is called a *σ-algebra* when

1. $\Omega \in F$
2. $A \in F \implies A^c \in F$, where A^c is the complement of A.
3. For any countable $\{A_i\}_{i=1}^{\infty}$ such that $A_i \in F$, $\cup_{i=1}^{\infty} A_i \in F$.

Breaking this definition down, the first part says that the $\sigma-$algebra includes the sample space itself; the second part says that the $\sigma-$algebra is closed under complementation; and the third part says that the $\sigma-$algebra is closed under countable unions.

Remark 1.5. It follows from the definition that if $\{A_i\}_{i=1}^{\infty}$ is such that $A_i \in F$, then $\cap_{i=1}^{\infty} A_i \in F$; this is since $\cap_{i=1}^{\infty} A_i \subset \cup_{i=1}^{\infty} A_i \in F$.

Remark 1.6. The powerset of any set is a σ-algebra of that set. While a standard measure for use on such a $\sigma-$algebra is the counting measure (defined shortly), the most important measure, *viz.* the Lebesgue measure (defined shortly) is not defined on the powerset of $\mathbb{R}$.[1] As a rule of thumb, if Ω is at most countably infinite, then the powerset of Ω is a useful $\sigma-$algebra (e.g. used with the counting measure) but if Ω is uncountably infinite, then some other $\sigma-$algebra should be used.

Definition 1.7. For any collection of sets A, let $\sigma(A)$ denote the smallest $\sigma-$algebra that contains all elements in A.

Definition 1.8. A *measurable space* is a pair (Ω, F) of a set Ω and a $\sigma-$algebra F. The subset $E \subset \Omega$ is called a *measurable set* if E belongs to F.

Example 1.9. Define the following experiment relating to the outcome of a soccer match. The sample space Ω is:

$$\Omega = \{win, defeat, tie\}$$

[1] While not part of this course, a simplified explanation – subject to controversy – has to do with the fact that there are sets that cannot have a Lebesgue measure and so cannot be in the $\sigma-$algebra; thus, since the powerset would include this set, the powerset cannot be the $\sigma-$algebra to use with Lebesgue measure.

and each of the three elements is an elementary event. Note that as the sample space is finite, it is possible to think about the $\sigma-$algebra generated by the power set of the sample space. Many different $\sigma-$algebras may be defined from Ω, for instance:

$F_1 = \{\emptyset, \{win,defeat,tie\}\} = \sigma(\Omega)$
$F_2 = \{\emptyset, \{win,defeat,tie\}, \{win\}, \{defeat,tie\}\}$
$F_3 = \{\emptyset, \{win,defeat,tie\}, \{win\}, \{defeat,tie\}, \{defeat\}, \{win,tie\}, \{tie\}, \{win,defeat\}\} = P(\Omega)$

where $F_1 = \sigma(\Omega)$ is the smallest possible $\sigma-$algebra and $F_3 = P(\Omega)$ is the largest possible $\sigma-$algebra. We can form a measurable space by combining the sample space Ω defined above with any of these three examples of $\sigma-$algebras.

Example 1.10. For any $k \in \mathbb{N}$, let $\Omega = \mathbb{R}^k$ and consider the $\sigma-$algebras:

$$F_1 = \{\emptyset, \mathbb{R}^k, \{\mathbf{0}\}, \mathbb{R}^k\{\mathbf{0}\}\}$$
$$F_2 = \sigma(\{\mathbf{y} \in \mathbb{R}^k : \mathbf{y} \leq \mathbf{z}, \mathbf{z} \in \mathbb{R}^k\}\})$$

Here $(\mathbb{R}^k, F_1)$ and $(\mathbb{R}^k, F_2)$ are two measurable spaces. While F_1 is obviously too small for general use in interesting applications, F_2 appears to be more adequate; actually, all subsets of $\mathbb{R}^k$ that are of practical interest are in fact measurable with respect to F_2, which makes F_2 the standard, conventional or implicit $\sigma-$algebra in Euclidean spaces denoted by $\mathcal{B}(\mathbb{R}^k)$.

So far we have focused on sample spaces and sigma-algebras. Now we need to provide a third component to complete basic measure theoretic definitions of probability.

Definition 1.11. Let (Ω, F) be a measurable space. We call $P(\cdot)$ a *probability measure* or *probability* where $P(\cdot)$ is a function mapping sets in the $\sigma-$algebra F to the set of real numbers when it satisfies the following three axioms:

1. $P(E) \geq 0 \; \forall E \in F$
2. $P(\Omega) = 1$
3. If $E_1, E_2, \ldots$ is a countably infinite sequence of pair-wise disjoint sets (i.e. $E_i \cap E_j = \emptyset \forall i \neq j$) with $E_i \in F$ for every $i = 1, 2, \ldots$, then

$$P\left(\cup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$$

Remark 1.12. P is a set-valued function defined on the elements of F.; not all subsets of the sample space are elements of F even though the whole sample space is, which is why it is important that P is defined only on elements of F. Note that the third axiom holds for finite sequences of of pair-wise disjoint sets $E_1, E_2, \ldots, E_n$ since the $\emptyset \in F$ and therefore we can take the 'tail' of the sequence to be empty sets. So we have the following lemma.

Lemma 1.13. *If $E_1, E_2, \ldots, E_n$ is a finite sequence of pair-wise disjoint sets ($E_i \cap E_j = \emptyset \; \forall i \neq j$) with $E_i \in F \; \forall i = 1, 2, \ldots, n$, then*

$$P\left(\cup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i)$$

Proof. Omitted. □

Example 1.14. Let $\Omega = [0, 1]$ and F be the smallest σ-algebra containing all the open sets. This is called the *Borel field*. The measure on $[0, 1]$ is the Lebesgue measure (defined shortly) obtained by setting $P[(a, b)] = b - a$. Note that all ordinary subsets of $[0, 1]$ are Lebesgue measurable (defined shortly).

Example 1.15. For any $A \in F$,

$$(\Omega = A \cup A^c \wedge A \cap A^c = \emptyset) \Longrightarrow P(\Omega) = \underbrace{P(A)}_{\geq 0} + \underbrace{P(A^c)}_{\geq 0} = 1$$

$$\overset{P(A)}{\Longrightarrow} 0 \leq P(A^c \leq 1$$

So probabilities must lie between 0 and 1. It can also be shown that the probability of the set of all rationals on $[0, 1]$ is 0, but that the probability of the set of irrationals on $[0, 1]$ is 1. In the first case, Georg Cantor (1845-1918) showed that while the set of irrationals is uncountable, the set of rationals is countable in 1873. Let $\{r_j\}$ be the set of rationals on $[0, 1]$ and observe that since the length of a point is 0, $P(r_j) = 0$, so

$$\therefore P(\text{set of all } \mathbb{Q}) = P(\cup_{j=1}^{\infty} r_j) = \sum_{j=1}^{\infty} P(r_j) = 0$$

This establishes that the probability of the set of all rationals on $[0, 1]$ is 0. Now let i be an irrational number, so $P(i) = 0$ since the length of a point is zero. The set of all irrationals is the union of i where $i \notin \mathbb{Q}$, which we will call I and we want to check the second equality here:

$$P(I) = P(\cup i) \overset{?}{=} \sum_{i} P(i)$$

Let R be the set of rationals. Note that $[0, 1] = R \cup I$, which is a countable union, I is measurable and $R \cap I = \emptyset$.[2] Since $[0, 1] = \Omega$ has length one:

$$P([0, 1]) = 1 = P(R) + P(I) = 0 + P(I)$$

$$\therefore P(I) = 1$$

[2] I am guilty of a slight amount of hand-waving here in that these results would need to be proved in a longer maths course.

Additional properties can be derived from the axioms and definitions and are given in proposition 1.16.

Proposition 1.16. *1.* $P(A^c) = 1 - P(A)$

2. $P(A) \in [0, 1]$

3. $P(\emptyset) = 0$

4. $A \subset B \implies P(A) \leq P(B)$

5. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Proof. Omitted. □

Definition 1.17. A *measure* μ satisfies the following properties:

1. $A \in \mathcal{A} \implies \mu(A) \geq 0$
2. $\mu(\emptyset) = 0$
3. If $\{A_i\}_{i=1}^{\infty} \in \mathcal{A}$ are disjoint, then $\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$.

The difference between a probability measure and a measure it that for a measure, $P(\Omega) = 1$ is removed while $\mu(\emptyset) = 0$ is added; it can be shown that a probability measure has the property that $P(\emptyset) = 0$. Note that a *measure space* is (Ω, F, μ), where Ω is the sample space, F is a σ-algebra defined on Ω and μ is a measure with respect to the σ-algebra F; notably, a measure*able* is (Ω, F) while the shorter word, measure space is (Ω, F, μ).

Finally, our goal in first part of this section has been to arrive at the definition of a probability space, which is defined as follows.

Definition 1.18. A *probability space* is the triplet (Ω, F, P), where Ω is the sample space, F is the $\sigma-$algebra defined on Ω and P is the probability measure defined on the measurable space (Ω, F).

Example 1.19 (Die roll)**.** An example of a probability space is rolling a die. Since the outcome of this is one element of the set $\{1, 2, 3, 4, 5, 6\}$, $\Omega = \{1, 2, 3, 4, 5, 6\}$. A natural choice for $F = 2^{\Omega} = P(\Omega)$, the powerset of Ω (set of all subsets of Ω); hence, F includes all of the singletons $\{1\}, \{2\}, \ldots, \{6\}$, all of the two element subsets of Ω, etc. and $\emptyset$. A natural choice in the specification of a probability measure here would be to suggest that the probability of any singleton (often called a simple event) is $\frac{1}{6}$. For P to satisfy the additivity property, this pins down the probability of all other events exactly. For example, the probability of $\{1, 3, 6\}$ is $P(\{1, 3, 6\}) = P(\{1\}) + P(\{3\}) + P(\{6\}) = \frac{1}{2}$ since $\{1\}, \{3\}, \{6\} \in F$ are disjoint and their union is $\{1, 3, 6\}$.

Example 1.20 (Coin toss)**.** Another example of a probability space is tossing a coin. Toss a coin:

$$\begin{aligned}
&\Omega = \{H, T\} \\
&F = \{\emptyset, \{H\}, \{T\}, \{HT\}\} \\
&P(\{H, T\}) = 1 \Longrightarrow P(\emptyset) = 0 \\
&P(\{H, T\}) = P(\{H\}) + P(\{T\}) = 1
\end{aligned}$$

Toss two coins:

$$\begin{aligned}
&\Omega = \{HH, HT, TH, TT\} \\
&F = \{\emptyset, HH, \ldots, TT, \{HH, TT\}, \ldots, \{HH, HT, TH\}, \ldots, \Omega\}
\end{aligned}$$

Now for some results.

Definition 1.21. For two events $A, B \in F$, if $P(B) > 0$, the *conditional probability* of the event A given B is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Proposition 1.22. *Let $B \subset \Omega$ be such that $P(B) > 0$. Then the conditional probability $P(\cdot|B)$ is a probability measure.*

Proof. Omitted. □

Example 1.23. Consider sequentially rolling two fair six-sided dice.

1. What is the probability that the sum of the rolls is 8?
2. What is the probability that the sum of the rolls is 8, given that the first die roll is 2?
3. If the first die roll is 1?
4. Given that the sum of the dice rolled is 8, what is the probability that the first die rolled displayed 4?

Denote the outcome of the first die roll by $\omega_1 \in \{1, 2, 3, 4, 5, 6\}$ and the outcome of the second die roll by $\omega_2 \in \{1, 2, 3, 4, 5, 6\}$. The sample space is $\Omega = \{(\omega_1, \omega_2) : \omega_i \in \{1, 2, 3, 4, 5, 6\}, \text{for} i = 1, 2\}$, containing the 36 elementary events $\omega = (\omega_1, \omega_2)$. As both dice are fair, conclude that the elementary events are *equally likely* to occur.

1. Observe that there are five elements in Ω such that $\omega_1 + \omega_2 = 8$ (i.e. (2,6), (3,5), (4,4), (5,3), (6,2)); hence:

$$P(\{\omega_1 + \omega_2 = 8\}) = \frac{5}{36}$$

2.
$$P(\{\omega_1 + \omega_2 = 8\}|\{\omega_1 = 2\}) = \frac{P(\{\omega_1 + \omega_2 = 8\} \cap \{\omega_1 = 2\})}{P(\{\omega_1 = 2\})} = \frac{1}{6}$$

3.
$$P(\omega_1 + \omega_2 = 8|\{\omega_1 = 1\}) = \frac{P(\{\omega_1 + \omega_2 = 8\} \cap \{\omega_1 = 1\})}{P(\{\omega_1 = 1\})} = 0$$

4.
$$P(\{\omega_1 = 4\}|\{\omega_1 + \omega_2 = 8\}) = \frac{P(\{\omega_1 + \omega_2 = 8\} \cap \{\omega_1 = 4\})}{P(\{\omega_1 + \omega_2 = 8\})} = \frac{1}{5}$$

A shorter definition of conditional probability is the following.

Definition 1.24. For $P(B) > 0$, the *conditional probability* of event A given event B is
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

We will use the concept of conditional probability extensively in the topic of identification.

Definition 1.25. For the set A, the collection $B_1, B_2, \ldots, B_n$ of subsets of A is called a *partition of* $A \iff$

1. $B_i \cap B_j = \emptyset \; \forall i \neq j$
2. $\cup_{i=1}^{n} B_i = A$

i.e. the collection is pair-wise disjoint and covers A exactly.

A useful theorem for identification analysis is the Law of Total Probability (LTP).

Theorem 1.26 (Law of Total Probability)**.** *Let $B_1, B_2, \ldots, B_n$ be a partition of the sample space Ω such that $P(B_i) > 0$ for all events in the partition. Then for any event A, we have that*
$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

Proof. Omitted. □

The *law of total probability* (known by other names such as the partition law or law of the extension of conversation) states that for events A and B
$$P(A) = P(A \cap B) + P(A \cap B^c)$$
For discrete random variables
$$P(Y = y) = \sum_{x \in Supp(X)} P(Y = y|X = x)P(X = x)$$

Theorem 1.27 (Bayes' Theorem)**.** *Let the event A be such that $P(A) > 0$ and let $B_1, B_2, \ldots, B_n$ partition Ω such that $P(B_i) > 0 \ \forall i = 1, 2, \ldots, n$. Then we have that*

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

A simplified version of *Bayes Theorem* or *Bayes Rule* states that for any two events A and B where $P(B) > 0$:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

For discrete random variables:

$$P(Y = y|X = x) = \frac{P(X = x|Y = y)P(Y = y)}{\sum_{y \in Supp(Y)} P(X = x|Y = y)P(Y = y)}$$

Proof. By the definition of conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Applying the definition of conditional probability to $P(A \cap B)$ and rearranging, we get Bayes Rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$ □

Now we will move on to the concept of independence.

Definition 1.28. Two events A and B are said to be *independent* $\iff$

$$P(A \cap B) = P(A)P(B)$$

To generalise the definition to any arbitrary finite or infinite collection of events, consider the following definition.

Definition 1.29. The events $E_1, E_2, \ldots, E_n$ are said to be (*collectively*) *independent* $\iff$ for any $2 \leq j \leq n$ and $1 \leq k_1 < \cdots < k_j \leq n$ we have that

$$P\left(\cap_{i=1}^{j} E_{k_i}\right) = \Pi_{i=1}^{j} P(E_{k_i})$$

An infinite collection of events $E_1, E_2, \ldots$ are (*collectively*) *independent* $\iff$ for each finite n, the events $E_1, E_2, \ldots, E_n$ are independent.

Remark 1.30. The definition of independence for arbitrary collections of events involves checking the multiplication condition for all possible (non-singleton) subcollections. For example, with the collection A, B, C, we need to check:

$$\begin{aligned} P(A \cap B) &= P(A)P(B) \\ P(A \cap C) &= P(A)P(C) \\ P(C \cap B) &= P(C)P(B) \\ P(A \cap B \cap C) &= P(A)P(B)P(C) \end{aligned}$$

To show the independence of $E_1, \ldots, E_n$ it is necessary (but no sufficient) to check the condition for all pairs E_i, E_j for $i \neq j$, which leads us to the (weaker) definition of pair-wise independence.

Definition 1.31. Events $E_1, E_2, \ldots, E_n$ are *pair-wise independent* $\iff$ for any pair of different elements in the collection $E_i \neq E_j$:

$$P(E_i \cap E_j) = P(E_i)P(E_j)$$

i.e. any subcollection of two different elements of $\{E_1, E_2, \ldots, E_n\}$ contains independent events.

Remark 1.32. (Collective) independence implies pair-wise independence, but the converse does not necessarily always hold.

1.2 Random Variables

Random variable are functions from the sample space to real numbers, i.e. a random variable assigns a real number to every element of the sample space. But because we want to talk about probabilities in relation to random variables, we will need to be careful about the measurability of the subsets of the sample space on which our random variable takes any particular value, i.e. we will require random variables to be measurable functions.

Definition 1.33. Let (Ω_1, F_1) and (Ω_2, F_2) be measurable spaces. We say that the function $f : \Omega_1 \longrightarrow \Omega_2$ is *measurable* $(F_1/F_2) \iff \forall S \in F_2,\ f^{-1}(S) \in F_1$.

Example 1.34. Any continuous function $f : \mathbb{R}^{k_1} \longrightarrow \mathbb{R}^{k_2}$.

Lemma 1.35. *Let (Ω_1, F_1) be a measurable space, Ω_2 an arbitrary set and A an arbitrary collection of subsets of Ω_2. Then $f : \Omega_1 \longrightarrow \Omega_2$ is measurable $(F_1/\sigma(A)) \iff \forall S \in A,\ f^{-1}(S) \in F_1$*

Proof. Omitted. □

Remark 1.36. When $\Omega_2 = \mathbb{R}^k$, we noticed that $\mathcal{B}(\mathbb{R}^k) = \sigma(\{x : x \leq z, z \in \mathbb{R}^k\})$ is a suitable $\sigma-$algebra for practical purposes. Therefore, for any measurable space (Ω_1, F_1), $f : \Omega_1 \longrightarrow \mathbb{R}^k$ is measurable $\iff$

$$\{\omega | X(\omega) \leq z\} \in F_1 \quad \forall z \in \mathbb{R}^k$$

This is quite a useful check for measurability. The definition of a measurable function and this remark explain the rationale behind the following equivalent definitions of random variable.

Definition 1.37. Let (Ω, F, P) be a probability space. The following are equivalent definitions.

1. The function $X : \Omega \longrightarrow \mathbb{R}^k$ is a *random variable*[3] $\iff$ it is a measurable function $(F/\mathcal{B}(\mathbb{R}^k))$.

2. The function $X : \Omega \longrightarrow \mathbb{R}^k$ is a *random variable* $\iff$ $\forall z \in \mathbb{R}^k$, $\{\omega | X(\omega) \leq z\} \in F$.

So, while a probability measure $P : F \longrightarrow \mathbb{R}$ takes sets and is a function on σ-algebra F, a random variable $X : \Omega \longrightarrow \mathbb{R}$ takes points is is a measure on sample space Ω. The property of measurability is important if we want to be able to integrate, i.e. in probability theory we take its expectation. Given any Borel measurable subset of $\mathbb{R}^k$, Q, we can define the probability that the random variable X belongs to Q as

$$P(X \in Q) = P(\{\omega \in \Omega : X(\omega) \in Q\}) = P(\{\omega \in X^{-1}(Q)\})$$

The set $\{X^{-1}(Q)\}$ is measurable by the definition of a random variable and the choice of Q being Borel-measurable. The following proposition is a very useful result.

Example 1.38. We know that a random variable is a measurable on Ω where measurable means $\{\omega : X(\omega) < z\} \in F \ \forall z \in \mathbb{R}$, i.e. all outcomes $X(\omega) < z$, $\omega \in F$. For example, let X be the number of heads, so $X(HH) = 2$, $X(HT) = 1$, etc.

$$\begin{aligned}
&\{\omega : X(\omega) < z\} = \emptyset \text{ if } z \leq 0 \\
&\{\omega : X(\omega) < z\} = \{TT\} \text{ if } 0 < z \leq 1 \\
&\{\omega : X(\omega) < z\} = \{HT, TH, TT\} \text{ if } 1 < z \leq 2 \\
&\{\omega : X(\omega) < z\} = \Omega \text{ if } z > 2
\end{aligned}$$

Note $F(z) = P(\omega : X(\omega) < z)$ where $P : F \longrightarrow \mathbb{R}$:

$$F(z) = \begin{cases} P(0) = 0 & \text{if } z \leq 0 \\ P(TT) & \text{if } 0 < z \leq 1 \\ P(HT, TH, TT) & \text{if } 1 < z \leq 2 \\ P(\Omega) = 1 & \text{if } z > 2 \end{cases}$$

Proposition 1.39. *Any measurable function of a random variable is also a random variable.*

Proof. Omitted. □

The following example utilises the probability space that we constructed in example **??**.

[3] When $k > 1$, X is a *random vector*. We may assume that X takes also value $\pm\infty$.

Example 1.40 (Example 1.19 continued, part (i))**.** Consider the following lottery. Rolling a six-sided fair die once, if the outcome is an even number then you receive $e\,100$ whereas if the outcome is an odd number then you must pay $e\,100$. How might we represent your net profit as a random variable based on the probability space constructed in example 1.19? Firstly, note that the lottery can be expressed as:

$$X(\omega) = \begin{cases} 100 & \text{if } \omega \in \{2,4,6\} \\ -100 & \text{if } \omega \in \{1,3,5\} \end{cases}$$

A random variable is a real-valued measurable function ($\{\omega : X(\omega) < z\} \in F \forall z \in \mathbb{R}^k$) on Ω.[4] From the definition, it is obvious that $X(\omega)$ is a real-valued function defined on Ω. To check for measurability, recall that we need to ensure $G(t) = \{\omega : X(\omega) < t\} \in F$. When $t \leq -100$, then $G(t) = \emptyset$. When $-100 < t \leq 100$, then $G(t) = \{1,3,5\}$. When $t > 100$, then $G(t) = \Omega = \{1,2,3,4,5,6\}$. Each of these sets is measurable because each set is in F. Therefore, $X(\omega)$ is a random variable.

1.2.1 Distributions

Remark 1.41. It is possible to classify random variables into three categories: *continuous* random variables, *discrete* random variables and *mixed* random variables.[5] Most of what we cover in this course will be expressed in terms of either or both of the first two; however, it is not difficult to extend most of what is said in this course to the case of mixed random variables.

Typically, we refer to the cumulative distribution function (CDF) of a random variable as its 'distribution'. The behaviour of a random variable is completely described by its CDF.

Definition 1.42 (CDF)**.** The *cumulative distribution function* (*CDF*) of a random variable X is given by

$$F_X(z) = P(X \leq z) = P(\{\omega : X_1(\omega) \leq z_1, \ldots, X_k(\omega) \leq z_k\}) \quad \forall z \in \mathbb{R}^k$$

Next consider the properties of the CDF for $k = 1$.

Proposition 1.43. *The function $F_X(z)$ is a CDF* $\iff$

1. *$F_X(z)$ is nondecreasing*
2. *$F_X(z)$ is right-continuous*
3. $\lim_{z \longrightarrow -\infty} F_X(z) = 0,\ \lim_{z \longrightarrow \infty} F_X(z) = 1$

Proof. See Billingsley (2012), Theorem 14.1 [8]. □

[4] Technically, we call this a random vector when $k > 1$

[5] Note that a random variable is *discrete* if it can take on at most countably infinite many values.

So, a distribution function can only have jump discontinuities.

Definition 1.44. The *support of a random variable* X, S_X is defined as

$$S_X = \{z \in \mathbb{R}^k : \forall \epsilon > 0, P(X \in B_\epsilon(z)) > 0\}$$

where $B_\epsilon(z) = \{y \in \mathbb{R}^k : \|y - z\| < \epsilon\}$ is the $k-$dimensional (open) ball of radius ϵ. Intuitively, a point in $\mathbb{R}^k$ belongs to the support of X if with positive probability X falls within any arbitrarily small neighbourhood of it.

Another complete representation of the behaviour of a random variable is given by the probability mass function (PMF) in the discrete case and the probability density function (PDF) in the continuous case. Discrete random variables can take finite or countably infinite values, i.e. $X(\Omega)$ is either finite or countably infinite. We can associate a certain mass of the distribution to each of these value and thereby completely characterise the behaviour of a discrete random variable by a function that indicates which is the probability mass of each value that the random variable can take. This function is the PMF.

Definition 1.45 (PMF)**.** The *probability mass function* (*PMF*) of a discrete random variable X is given by

$$f_X(z) = P(X = z) = P(\{\omega : X(\omega) = z\}) \quad \forall z \in \mathbb{R}^k$$

There is a one-to-one relationship between the CDF and the PMF – from either, you can deduce the other. With discrete random variables, the points in the support of X are those where $f_X(z) > 0$ and they correspond to the discontinuity points of the CDF.

Random variables are customarily denoted by upper case letters while the values they can take are usually written in lower case letters. Let $\{x_i\}_{i=1}^n$ (n may be ∞) be the n possible values that X can take. To go from PMF to CDF:

$$F_X(z) = P(X \leq z) = \sum_{i=1_{x_i \leq z}}^{n} P(X = x_i) = \sum_{i=1_{x_i \leq z}}^{n} f_X(x_i)$$

and to go from CDF to PMF at point z, consider the behaviour of the CDF as approach z from below. If the CDF remains constant at z, then the PMF is zero and if the CDF jumps at z, then the PMF takes the value of the jump. To compute the probability that X is in the set $\{y : y \leq z\}$ for some z, we can use the CDF directly as:

$$P(X \in \{y : y \leq z\}) = P(X \leq z) = F_X(z)$$

To compute the probability that X falls into any arbitrary (measurable) set Q, we can use the PMF:

$$P(X \in Q) = \sum_{i=1_{x_i \in Q}}^{n} P(X = x_i)$$

Example 1.46 (Bernoulli Distribution)**.** This distribution characterises a discrete random variable that can take only two values, zero or one, so its support is $S_X = \{0, 1\}$. Its CDF and PMF are, respectively:

$$F_X(z) = \begin{cases} 0 & \text{if } z \in (-\infty, 0) \\ 1-p & \text{if } z \in [0, 1) \\ 1 & \text{if } z \in [1, \infty) \end{cases}$$

$$f_X(z) = \begin{cases} 0 & \text{if } z \notin \{0, 1\} \\ 1-p & \text{if } z = 0 \\ p & \text{if } z = 1 \end{cases}$$

Example 1.47 (Discrete Uniform Distribution)**.** This distribution characterises a discrete random variable that can take a finite number of values $\{x_i\}_{i=1}^n$ with equal probability. Without loss of generality (WLOG), take the set $\{x_i\}_{i=1}^n$ to be ordered and note that its support is given by $S_X = \{x_i\}_{i=1}^n$. Its CDF and PMF are, respectively:

$$F_X(z) = \begin{cases} 0 & \text{if } z \in (-\infty, x_1) \\ i/n & \text{if } z \in [x_i, x_{i+1}) \text{ for } i = 1, \ldots, n-1 \\ 1 & \text{if } z \in [x_n, \infty) \end{cases}$$

$$f_X(z) = \begin{cases} 0 & \text{if } z \notin \{x_i\}_{i=1}^n \\ 1/n & \text{if } z \in \{x_i\}_{i=1}^n \end{cases}$$

A continuous random variable is a random variable that takes a continuum of or uncountably many values, i.e. if $X(\Omega)$ is uncountably infinite. Unlike the case of discrete random variables, since the probability that a continuous variable is exactly any of the points in the support is zero, none of them have any positive mass. The PDF characterises continuous random variables.

Definition 1.48. The *probability density function* (*PDF*) of a continuous random variable X is given by

$$f_X(z) = \frac{dF_X(z)}{dz} \quad \forall z \in \mathbb{R}^k$$

To go from CDF to PDF, differentiate with respect to every coordinate. To go from PDF to CDF, integrate with respect to every coordinate, where the limit behaviour of the CDF pins down the value of the constant of integration.

Again, like the discrete case, to calculate the probability that X falls in the set $\{y : y \leq z\}$, we can use the CDF directly to compute

$$P(X \in \{y : y \leq z\}) = P(X \leq z) = F_X(z)$$

Similarly to the discrete case, to compute the probability that X falls into any arbitrary set Q (must be Borel measurable), we use the PDF:

$$P(X \in Q) = \int_Q f_X(z) dz$$

Example 1.49 (Continuous Uniform Distribution)**.** This distribution is such that the random variable takes values in an interval $[a, b], a < b$ with equal likelihood (EL). The CDF and PDF are, respectively:

$$F_X(z) = \begin{cases} 0 & \text{if } z \in (-\infty, a) \\ (z-a)/(b-a) & \text{if } z \in [a, b) \\ 1 & \text{if } z \in [b, \infty) \end{cases}$$

$$f_X(z) = \begin{cases} 0 & \text{if } z \notin [a, b] \\ 1/(b-a) & \text{if } z \in [a, b] \end{cases}$$

Example 1.50 (Normal Distribution)**.** Possibly the most important distribution in statistics, its success is mainly due to the Central Limit Theorems (together with Laws of Large Numbers are about the most important theorems of statistics, similar to the importance of the Fundamental Theorem of Algebra for algebra or the Fundamental Theorem of Calculus for mathematical analysis). The support of this continuous distribution is the whole real line, $\mathbb{R}$. The distribution is denoted by $N(\mu, \sigma^2)$, where the two parameters μ (mean) and σ^2 (variance) characterise the distribution. The PDF is given by:

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

The CDF of this distribution has no closed form solution, but we can express it as the integral of the PDF on $(-\infty, z]$ and most statistics books report the relevant values for the standard Normal CDF (i.e. $N(0, 1)$). To deduce the shape of (μ, σ^2) for any value of μ and σ^2 from the quantiles of $N(0, 1)$, we can use the following result.

Lemma 1.51. *For* $\sigma > 0, X \sim N(\mu, \sigma^2) \iff \frac{X-\mu}{\sigma} \sim N(0, 1)$

Proof. Omitted. □

Example 1.52 (Chi-Squared Distribution)**.** The support of this continuous distribution is the non-negative real line, $\mathbb{R}^+$. It is characterised by one parameter, k, the degrees of freedom and denoted by χ^2_k. The chi-squared distribution with k degrees of freedom is simply the sum of the squares of k independent random variables with standard Normal distributions, $N(0, 1)$. Let $X_1, \ldots, X_k$ be independently[6] distributed according to $N(0, 1)$, then $Y = \sum_{i=1}^{k} X_i^2$ is a new random variable with chi-squared distribution having PDF:

$$f_X(z) = \begin{cases} \frac{1}{2}\left(\frac{z}{2}\right)^{\frac{k}{2}-1} \exp\left(\frac{-z}{2}\right) \frac{1}{\Gamma(k/2)} & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases}$$

where $\Gamma(k/2)$ is the Gamma function: $\Gamma(1/2) = \sqrt{\pi}, \Gamma(1) = 1, \Gamma(n) = (n-1)\Gamma(n-1)\ \forall n = m/2, m \in \mathbb{Z}, m > 2$.

[6] Defined shortly.

When $k \geq 2$, we are dealing with *multivariate* random variables. The definitions of CDF and PMF/PDF remain unchanged, but we say *joint* CDF when referring to $F_X(z)\ \forall z \in \mathbb{R}^k$. For now let us restrict attention to bivariate distributions, i.e. where $k = 2$. In this case, the joint CDF may be written as:

$$F_{X,Y}(x, y) = P(X \leq x \wedge Y \leq y)$$

for any $(x, y) \in \mathbb{R}^2$. Everything we will do can be trivially extended to $k \geq 3$.

The *marginal distribution* refers to the distribution of one of the components of a random vector. To obtain the CDF of a marginal distribution, we can take the limit of the joint CDF as the rest of the coordinates go to ∞:

$$F_X(x) = \lim_{y \longrightarrow \infty} P(X \leq x \wedge Y \leq y)$$
$$F_Y(y) = \lim_{x \longrightarrow \infty} P(X \leq x \wedge Y \leq y)$$

for any $(x, y) \in \mathbb{R}^2$. From the marginal CDF, we can obtain the marginal PDF (PMF) as per usual by taking derivatives of the marginal CDF (by deducing it from the jumps of the marginal CDF). We can also deduce the marginal PDF (PMF) from the joint PDF (PMF). For discrete bivariate random variables:

$$f_X(x) = \sum_{y \in S_Y} f_{X,Y}(x, y)$$
$$f_Y(y) = \sum_{x \in S_X} f_{X,Y}(x, y)$$

for any $(x, y) \in \mathbb{R}^2$. For continuous bivariate random variables:

$$f_X(x) = \int f_{X,Y}(x, y)dy$$
$$f_Y(y) = \int f_{X,Y}(x, y)dx$$

We can go from the marginal PDF (PMF) to the marginal CDF as per usual by integration (summation).

Remark 1.53. While we can derive the marginal functions from the joint functions, the reverse is not true. The information in the marginal distributions pertain only to the random behaviour of each component in isolation whereas the joint distribution contains all the relevant information about the joint random behaviour of the random vector. So, in general, the totality of the marginal distributions will not contain all the necessary information to derive the joint distribution and we can see this by looking at an example where the same marginals are derived from different joint distributions.

Example 1.54. Let the random variables X and Y be the result of two-coin tosses, where 1 represents heads and 0 represents tails. Assume both coins are

fair so heads and tails are equally likely outcomes. The marginal PMFS are:

$$P(X=x)=\begin{cases}1/2 & \text{if } x\in\{0,1\}\\ 0 & \text{otherwise}\end{cases}$$
$$P(Y=y)=\begin{cases}1/2 & \text{if } y\in\{0,1\}\\ 0 & \text{otherwise}\end{cases}$$

It can be shown that these marginals can be the result of the following two joint PMFs:

1. X and Y represent independent draws of two fair coins so that the joint PMF is:
$$P(X=x,Y=y)=\begin{cases}1/4 & \text{if } (x,y)\in\{0,1\}^2\\ 0 & \text{otherwise}\end{cases}$$
2. $X=Y$ which is the result of only one draw of a fair coin so that the joint PDF is:
$$P(X=x,Y=y)=\begin{cases}1/2 & \text{if } (x,y)\in\{(0,0),(1,1)\}\\ 0 & \text{otherwise}\end{cases}$$

Therefore, complete knowledge of marginals is obviously not enough to infer the joint distribution.

Example 1.55 (Bivariate Normal Distribution)**.** Let (X_1,X_2) be bivariate normally distributed with mean (μ_1,μ_2) and variance-covariance matrix given by Σ:

$$\Sigma=\begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{pmatrix}$$

Denote this by $(X_1,X_2)\sim BVN(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\sigma_{12})$. By definition, the joint PDF of (X_1,X_2) is given by

$$f_{X_1,X_2}(x_1,x_2)=\frac{\exp\left[\frac{-1}{2\left(1-\left(\frac{\sigma_{12}}{\sigma_1\sigma_2}\right)^2\right)}\left(\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2-\frac{2\sigma_{12}}{\sigma_1\sigma_2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right)\right]}{2\pi\sqrt{\left(1-\left(\frac{\sigma_{12}}{\sigma_1\sigma_2}\right)^2\right)\sigma_1^2\sigma_2^2}}$$

It can be shown that the marginal PDFs satisfy $X_1\sim N(\mu_1,\sigma_1^2)$ and $X_2\sim N(\mu_2,\sigma_2^2)$. Note that a complete knowledge of marginal PDFs does not give any information about the parameter σ_{12}. Therefore, the marginal distributions do not provide complete information about the joint distribution.

By *conditional distribution*, we mean the distribution of a random variable Y given that we already know another random variable, say X, takes a certain value x. Like with conditional probability, the conditional distributions are

only defined when $f_X(x) > 0$. Denote the conditional CDF by $F_{Y|X=x}(y)$ and the conditional PDF (PMF) by $f_{Y|X=x}(y)$. Conditional distributions can be uniquely derived from joint distributions. The conditional PDF of $Y|X$ (or PMF) for any x such that $f_X(x) > 0$ can be derived as follows:

$$f_{Y|X=x}(y) = \frac{f_{Y,X}(y,x)}{f_X(x)}$$

For discrete case, note the interpretation: $f_{Y|X=x}(y) = P(Y = y|X = x)$; the interpretation is the same for the continuous case. The conditional CDF for any x such that $f_X(x) > 0$ can be derived as follows:

$$F_{Y|X=x}(y) = \int_{-\infty}^{y} f_{Y|X=x}(s)ds = \frac{\int_{-\infty}^{y} f_{Y,X}(s,x)ds}{f_X(x)}$$

As we showed the conditional probability is a probability measure, it can be shown that the conditional distribution satisfies all the properties of a distribution as long as $f_X(x) > 0$: $F_{Y|X=x}(y)$ is non-decreasing, right continuous and satisfies the limit properties $\lim_{y\longrightarrow-\infty} F_{Y|X=x}(y) = 0$ and $\lim_{y\longrightarrow\infty} F_{Y|X=x}(y) = 1$. As before with marginal distributions, while conditional distributions can be derived from joint distributions, the converse is not true.

Example 1.56. Consider the following pair of degenerate conditional distributions:

$$f_{Y|X=x}(y) = \begin{cases} 1 & \text{if } y = x \\ 0 & \text{otherwise} \end{cases}$$

$$f_{X|Y=y}(x) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{otherwise} \end{cases}$$

which is the result of the experiment where $X = Y$. Simply by observing the conditional distributions we cannot deduce the joint distribution. As an example, this model may result from either of the following two cases: (i) $X = Y = 0$ where all random variables are degenerate or (ii) $X = Y \sim N(0, 1)$ where there is randomness in the joint distribution but not in the conditional distribution. This example highlights the fact that we can deduce the relationship from conditional distributions that links the components of a random vector but that we are missing how each component behaves separately.

In summary, we cannot go from marginal to joint or from conditional to joint distribution, but we can go from joint to marginal and joint to conditional distribution. Finally, we are able to combine the information contained by the conditional and marginal distributions together to infer the joint distribution. The definition of conditional PDF shows this immediately, if $f_X(x) > 0$:

$$f_{Y|X=x}(y) = \frac{f_{Y,X}(x,y)}{f_X(x)} \Longrightarrow f_{Y,X}(x,y) = f_{Y|X=x}(y)f_X(x)$$

In fact we can extend this to those values of x such that $f_X(x) = 0$. Whenever $f_{Y|X=x}$ is not defined, $f_X(x) = 0$; hence, the definition of the marginal distribution $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy = 0$ together with the fact that $f_{X,Y}(x, y) \geq 0$ implies that $f_{X,Y}(x, y) = 0$.

1.2.2 Moments

While all the relevant information regarding the distribution of a random variable is contained in its CDF/PDF/PMF, for some families of distributions, a set of *moments* of a random variable (quantities we derive from its distribution that contain concise information usually with practical interpretation) may suffice for a complete characterisation, e.g. the Normal distribution (mean and variance); in general, moments provide insufficient information to completely characterise the distribution of a random variable.

Definition 1.57. The *expectation* or *expected value* of a random variable, denoted $E(X)$ is the first moment of the distribution and is a measure of central tendency. For discrete random variables X with $S_X = \{x_1, x_2, \ldots, x_n\}$ and PMF $f_X(\cdot)$, the expectation of X is a weighted average of the possible values of X where the weights are provided by the PMF:

$$E(X) = \sum_{i=1}^{n} x_i f_X(x_i)$$

For continuous random variables X with PDF $f_X(\cdot)$, with an analogous interpretation the expected value is defined by:

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)dx$$

We can extend the definition of expectation of X to random vectors where $E(X)$ corresponds to the vector of expectations of each component.

Definition 1.58. The *conditional mean* of a random variable Y conditional on a realisation of the random variable X is defined as:

$$\begin{aligned} E(y|x) &= \sum_{y} yP(y|X = x) \quad y \text{ discrete} \\ &= \int_{y} ydF(y|x) \quad y \text{ continuous} \end{aligned}$$

Definition 1.59. The *median* is defined as:

$$Med(y|x) = \inf\{t : P(y \leq t|x) \geq \frac{1}{2}\}$$

The textbook definition usually replaces inf with min and $\geq \frac{1}{2}$ with $= \frac{1}{2}$. The textbook definition fails in certain cases, for example if y does not have a

smooth, continuous density function. When y is discrete, there is no solution, e.g. the Bernoulli variable that takes value 0 with probability .2 and 1 with probability .8. With the above definition, there is always a unique solution, irrespective of whether the density is discrete or continuous. So, letting $\mathcal{L}$ denote the loss function, $E(\mathcal{L}(\cdot)|x)$ always has a solution when $L(u) = |u|$. When we are dealing with symmetric distributions, the question of using means versus medians does not matter, but it may matter a lot when we are dealing with asymmetric distributions.

Building on the definition of the median, we can generalise to the α quantile:

$$Q_\alpha(y|x) = \inf\{t : P(y \leq t|x) \geq \alpha\}$$

Remark 1.60. While $E(X)$ may exist and be finite, unlike medians $E(X)$ may exist and be infinite or it may not even exist.

Let

$$X^+ = \begin{cases} X & \text{if } X \geq 0 \\ 0 & \text{if } X < 0 \end{cases}$$

$$X^- = \begin{cases} -X & \text{if } X < 0 \\ 0 & \text{if } X \geq 0 \end{cases}$$

The expectation $E(X)$ *exists* if at least one of $E(X^+)$ or $E(X^-)$ is finite, in which case we define[7]

$$E(X) = E(X^+) - E(X^-)$$

Example 1.61. The Cauchy random variable, which is continuously distributed with PDF

$$f_X(x) = \frac{1}{\pi(1+x^2)}$$

for $x \in \mathbb{R}$, has no expectation.

If both $E(X^+) < \infty$ and $E(X^-) < \infty$ or equivalently $E(\|X\|) < \infty$, then $\exists E(X)$ and is *finite*:

$$E(X) = E(X^+) - E(X^-)$$

Example 1.62. Consider X with $S_X = \{2^i : i \in \mathbb{N}\}$ where $P(X = 2^i) = 1/2^i$. This is a probability distribution since

$$\sum_{i=1}^{\infty} P(X = 2^i) = \sum_{i=1}^{\infty} \left(\frac{1}{2}\right)^i = 1$$

and $\exists E(X)$ since $E(X^-) = 0$. However, the expectation is infinite:

$$E(X) = \sum_{i=1}^{\infty} 2^i P(X = 2^i) = \sum_{i=1}^{\infty} 2^i \left(\frac{1}{2}\right)^i = \infty$$

[7] $E(X)$ may be ∞ or $-\infty$. For the rest of my notes, when I write $E(X)$ I implicitly assume that $\exists E(X)$ (exists) but that $E(X)$ is possibly infinite.

Remark 1.63. Note that $E(X)$ may lie outside S_X. For example, if X represents the result of a fair coin toss and $X = 1$ stands for heads while $X = 0$ stands for tails, $S_X = \{0, 1\}$ and $E(X) = 0.5$.

More realistically, we may have a function of the mean or the median. We have seen above that measurable functions of random variables are random variables. Essentially, we can derive the distribution of a function of a random variable X form the CDF $F_X(\cdot)$. Furthermore, we can think about the expectation of a function of a random variable.

Definition 1.64. For any random variable $X : \Omega \longrightarrow \mathbb{R}^k$ and for any measurable function $g : \mathbb{R}^k \longrightarrow \mathbb{R}^m$:

$$E(g(X)) = \begin{cases} \sum_{i=1}^{n} g(x_i) f_X(x_i) & \text{if } X \text{ discrete }, S_X = \{x_1, x_2, \ldots, x_n\} \text{ and PMF } f_X(\cdot) \\ \int_{-\infty}^{\infty} g(x) f_X(x) dx & \text{if } X \text{ is continuous with PDF } f_X(\cdot) \end{cases}$$

Remark 1.65. What about the relationship between $E(g(X))$ and $g(E(X))$? Firstly, let us look at the case of linearity of g. Let $X : \Omega \longrightarrow \mathbb{R}^k$ be a random variable and let $g : \mathbb{R}^k \longrightarrow \mathbb{R}^m$ be a linear function, i.e. $g_j(x) = \sum_{i=1}^{k} a_{ji} x_i + b_j$ for $j = 1, \ldots, m$. Then

$$E(g_j(X)) = E\left(\sum_{i=1}^{k} a_{ji} X_i + b_j\right) = \sum_{i=1}^{k} a_{ji} E(X_i) + b_j = g_j(E(X))$$

While $E(g(y)|x) = g(E(y|x))$ for g linear, note that we also have results for g concave and for g convex seen in Jensen's Inequality. When a function is both convex and concave, it is linear, so we have equality as above.

Lemma 1.66. *Jensen's Inequality Let $X : \Omega \longrightarrow \mathbb{R}$ and $g : \mathbb{R} \longrightarrow \mathbb{R}$. Then*

1. *If g is concave, $E(g(X)) \leq g(E(X))$.*
2. *If g is convex, $E(g(X)) \geq g(E(X))$.*

A nice, more general result applies to functions of the median. Note that when f is a monotone function of y, $Med(f(y)|x) = f(Med(y|x))$.

Proposition 1.67. *Let $X \sim N(\mu, \sigma^2)$. Then $E(X) = \mu$.*

Proof. Omitted. □

Definition 1.68. If $Y|X$ is a continuously distributed random variable with conditional PDF $F_{Y|X=x}(y)$, then

$$E(Y|X = x) = \int_{-\infty}^{\infty} y f_{Y|X=x}(y) dy$$

while if $Y|X$ is a discrete random variable with conditional PMF $f_{Y|X=x}(y)$, then

$$E(Y|X = x) = \sum_{y \in \Omega_Y} y f_{Y|X=x}(y)$$

Remark 1.69. We can either view the quantity $E(Y|X = x)$ as a deterministic function of the constant x or view the quantity $E(Y|X)$ as a measurable function of the random variable X, which implies that $E(Y|X)$ is a random variable. The next law characterises the expected value of the second interpretation, $E(Y|X)$ as a random variable. The next few definitions and results are very useful tools in the analysis of identification.

Theorem 1.70 (Law of Iterated Expectations (LIE))**.** *Given the existence of expectations, $E(E(Y|X)) = E(Y)$. More concretely, the law of iterated expectations can be stated as*

$$\begin{aligned} E(Y) &= E_X(E(Y|X)) \\ &= \sum_{x \in Supp(X)} E(Y|X = x)P(X = x) \quad \textit{discrete case} \\ &= \int_{x \in Supp(X)} E(Y|X = x)f(x)dx \quad \textit{continuous case} \end{aligned}$$

Proof. For the continuous case, using $f_{X,Y}(x, y) = f_{Y|X=x}(y)f_X(x)$ and $\int_{-\infty}^{\infty} f_{X,Y}(x, y)dx = f_Y(y)$:

$$\begin{aligned} E(E(Y|X)) &= \int_{-\infty}^{\infty} E(Y|X = x)f_X(x)dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{Y|X=x}(y)dy f_X(x)dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{X,Y}(x, y)dydx \\ &= \int_{-\infty}^{\infty} y f_Y(y)dy \\ &= E(Y) \end{aligned}$$

□

As a corollary of LIE, $E(E(h(Y, X)|X)) = E(h(Y, X))$, so we could allow the expression to depend directly on X.

Also note that the *decomposition of variance* is given by:

$$V(Y) = E_X[V(Y|X)] + Var_X[E(Y|X)]$$

Next we will look at the variance, the second centered moment, which is a measure of dispersion of the random variable about its mean.

Definition 1.71. When X is univariate, we can define its *variance* by

$$V(X) = E[(X - E(X))^2]$$

So, if the distribution of X is very concentrated (dispersed) about its mean, $E(X)$, then we would expect low (high) values of $(X - E(X))^2$ and thus, $V(X)$

would be small (big). Note that $V(X) > 0$. By expanding the squares, we get an alternative representation of the variance:

$$V(X) = E(X^2) - E(X)^2$$

We can show that $\exists V(X)$ and $V(X) < \infty$ if $E(|X|^2) < \infty$. But note that unlike expectations, the variance operator is not linear:

Proposition 1.72. *For constants a, b:*

$$V(aX + b) = a^2 V(X)$$

Proof. Using the linearity of the expectation operator:

$$\begin{aligned} V(aX + b) &= E[(aX + b - E(aX + b))^2] \\ &= E[(aX + b - (aE(X) + b))^2] \\ &= a^2 E[(X - E(X))^2] \\ &= a^2 V(X) \end{aligned}$$

□

Definition 1.73. The square root of the variance is the *standard deviation*: $sd(X) = \sqrt{V(X)}$. Note that because the square root of a non-negative number is a strictly increasing function, the standard deviation is also a measure of dispersion of a random number around its mean.

Remark 1.74. The standard deviation is defined in the same units of the expectation, i.e. if we scale the random variable by a certain factor, e.g. for some $a > 0$, $X' = aX$, then the expectation is scaled by the same factor, $E(X') = aE(X)$ by linearity and the standard deviation is scaled by the same factor, $sd(X') = a[sd(X)]$. However, the variance is scaled by the square of the factor: $V(X') = a^2 V(X)$.

For random vectors, the definition of the variance can be extended as follows. Let $X : \Omega \longrightarrow \mathbb{R}^k$ be a column vector. Then the variance of X is:

$$V(X) = E[(X - E(X))(X - E(X))']$$

So, the variance of a vector is a matrix. The standard deviation of a random vector will be the 'square root' of $V(X)$, i.e. it will be a matrix $sd(X)$ such that $sd(X)^2 = V(X)$.

Proposition 1.75. *Let $X \sim N(\mu, \sigma^2)$. Then $V(X) = \sigma^2$.*

Proof. Omitted. □

We can use the definition of the expected value to define higher centered and uncentered moments of a univariate random variable.

Definition 1.76. The k^{th} *uncentered moment* of a random variable is $E(X^k)$. The k^{th} *centered moment* of a random variable is $E[(X - E(X))^k]$.

So, the first moment is the mean and the second centered moment is the variance. If we think of moments as integrals, then we can demonstrate that the first k moments exist and are finite if the k^{th} absolute moment is finite, i.e. $E(|X|^k) < \infty$. This follows from two results in Billingsley (1995) [8]:

- $\exists E(Y) < \infty$ (i.e. YU is integrable with respect to the measure induced by its PDF) $\iff E(|Y|) < \infty$ (Billingsley, 1995: 273) [8].
- If $h \leq k$, then $E(|Y|^k) < \infty \implies E(|Y|^h) < \infty$, which follows from observing that $h \leq k \implies |x|^h \leq 1 + |x|^k$ (Billingsley, 1995: 274) [8].

1.3 Radon-Nikodym Theorem

The next few definitions build up towards the Radon-Nikodym theorem.[8] This theorem allows us to avoid treating separately a mixture of continuous and discrete distributions, e.g. truncated distributions. Most common probability distributions are actually examples of Radon-Nikodym derivatives with respect to a *reference* measure. For discrete random variables (probability functions), the reference measure is the counting measure while for continuous random variables (probability densities) the reference measure is the Lebesgue measure.[9]

Definition 1.77. Let X be a discrete random variable whose support is $\mathbb{N}$. The *counting measure* on $\mathbb{N}$ is

$$m(A) = \sum_{k=0}^{\infty} 1_A(k)$$

where $A \subset \mathbb{N}$ and

$$1_A(k) = \begin{cases} 1 & \text{if } k \in A \\ 0 & \text{if } k \notin A \end{cases}$$

Note that if A is finite then $m(A)$ is the number of elements of A and if A is infinite then $m(A)$ is infinity, i.e. the counting measure 'counts'. The most famous measure – at least for assigning measure to subsets of n-dimensional

[8] Johann Radon was from Austria and lived from 1887 to 1956. Otton Nikodym was from Poland and lived from 1889 to 1974.

[9] Further examples of discrete distributions include binomial and Poisson distributions. Further examples of continuous distributions include the student t and exponential distributions. Mixtures could be for instance a 50-50 mixture of binomial and student t for example. A more realistic example would be a truncated variable. For instance, let us look at the Tobit model. Let Y be expenditure on some durable good, X be income, Z be the sum of all other expenditures and Y_0 be the price of the cheapest durable good that is available. Assume that the relationship between $Y*$ and X is linear, i.e. $Y* = \beta_0 + \beta_1 X + \epsilon$ where $Y*$ is the solution of the associated maximisation problem and ϵ captures the effects of all unobservable variables. We observe $Y = Y*$ if $Y* \geq Y_0$ and $Y = Y_0$ otherwise. When ϵ is Normally distributed, Y is a mixture of a discrete and a continuous random variable.

Euclidean space – is the Lebesgue measure.[10] A thorough development of the concept and properties of the Lebesgue measure is beyond the scope of this course.

Definition 1.78. Let $E \subset \mathbb{R}$. The *Lebesgue outer measure* $\lambda*(E)$ is defined by

$$\lambda*(E) = \inf\{\sum_{k=1}^{\infty} l(I_k) : I_k \text{ is a sequence of open intervals with } E \subset \cup_{k=1}^{\infty} I_k\}$$

Definition 1.79. The *Lebesgue measure* of E is its outer measure $\lambda(E) = \lambda*(E)$ if

$$\lambda*(A) = \lambda*(A \cap E) + \lambda*(A \cap E^c) \quad \forall A \subset \mathbb{R}$$

Example 1.80. Any closed interval $[a, b]$ of real numbers is Lebesgue measurable with Lebesgue measure $b - a$; similarly, for (a, b).

Example 1.81. Any Cartesian product of intervals $[a, b]$ and $[c, d]$ is Lebesgue measurable with Lebesgue measure $(b - a)(d - c)$, the area of the rectangle.

Example 1.82 (A set that is not Lebesgue measurable)**.** Here is an example of a set that it is on the real line but is not the unit interval and is not Lebesgue measurable. Let ζ be an irrational number, which could be the number zero and let R be a sequence $\{r_n\}$ of all rational numbers. Let $E_\zeta = \{\zeta + r; r \in R\}$. Then we have that $E_{\zeta_1} \cap E_{\zeta_2} = \emptyset$ if $E_{\zeta_1} \neq E_{\zeta_2}$. From each distinct set E_ζ, take one number η with the property that $0 \leq \eta \leq \frac{1}{2}$. The set of all such numbers η is not Lebesgue measurable. To see the proof of this result, consult Friedman (1982) [32].

Let μ and ν be measurable on a $\sigma-$algebra $\mathcal{A}$. We now consider a dominance condition.

Definition 1.83. ν is *absolutely continuous* with respect to μ, denoted $\nu << \mu$ if

$$\mu(A) = 0 \Longrightarrow \nu(A) = 0$$

Example 1.84. The probability measure ν is *absolutely continuous* with respect to the counting measure μ ($\mu(A) = |A|$) since the only set that has μ measure zero is the empty set, which has ν measure zero also; hence, $\nu << \mu$.

[10] Another popular alternative and the one Lebesgue wanted to improve upon is the Riemann integral, named after Georg Friedrich Bernhard Riemann (1826-1866) that essentially breaks up a function into steps (along the x-axis). The outer limit tends to the inner limit and vice-versa where both exist. Henri Lebesgue (1875-1941) approached this slightly differently in his 'Theory of measure and integration', which he developed in 1901-1902, essentially moving the reference point to the y-axis. Lebesgue focused on the measure of sets and summing over the value of the function times the measure of the set on which the function has that value. The Lebesgue integral is the limit as the widths of strips (on the y-axis) shrink. It turns out that the Riemann integral does not exist for the Dirichlet function (0 if irrational, 1 if rational), while the Lebesgue integral is one. The Lebesgue measure is σ-finite.

Definition 1.85. Let (X, Σ, μ) be a measure space. If $\mu(X) < \infty$, then we say that the measure μ is *finite*, i.e. all measurable sets have finite measure.

We can weaken this to the following.

Definition 1.86. The measure μ is *σ-finite* if there exists a countable sequence of measurable sets $A_1, A_2, \ldots$ such that $\Omega = \cup_j A_j \wedge \mu(A_j) < \infty \; \forall j$.

Theorem 1.87 (Radon-Nikodym)**.** *Let μ and ν be σ-finite measures on a measurable space (X, Ω) such that $\nu << \mu$. Then there exists a nonnegative function f such that*

$$\nu(A) = \int_A f(x) d\mu(x) \quad \forall A \in \mathcal{A}$$

Furthermore, for two such functions F and G, $\mu(\{x \in X : f(x) \neq g(x)\}) = 0$, i.e. f is unique up to at most a set of $\mu-$measure zero. f is called a Radon Nikodym (RN) density and is often denoted $f = \frac{d\nu}{d\mu}$.

The RN theorem solves the problem of $P(A|B)$ when $P(B) = 0$, saying that it exists under certain conditions, but it does not tell us how to find it. Note that the opposite direction is true: integration under a function yields a measure. If ν, μ, λ are σ-finite, $\nu << \mu$ and $\mu << \lambda$, then $\nu << \lambda \wedge \frac{d\nu}{d\lambda} = \frac{d\nu}{d\mu}\frac{d\mu}{d\lambda}$. To find the Radon-Nikodym derivative of a discrete probability space (one where Ω is at most countably infinite):

1. define the sample space ($\mathbb{N}$), the sigma algebra ($\mathcal{P}(\mathbb{N})$) and two measures ($\nu = \mathcal{P}, \mu = c$) where c is the counting measure and $\mathcal{P}$ is the probability measure;
2. verify conditions (ν, μ are σ-finite and $\nu << \mu$);
3. note that RN implies that there exists a unique f almost everywhere (see definition 2.29) but you need to guess f (note that $\int_{\mathbb{N}} f dc = \sum_{i=0}^{\infty} f(i)$ and show that $\int_A f dc = \mathcal{P}(A)$).

Example 1.88 (Example 1.19 continued, part (ii))**.** The outcome of rolling a die is certainly discrete (random variable takes on at most countably infinite many values). Suppose that a random variable takes only natural number values, so we could think of the sample space being $\mathbb{N}$ and let the σ-algebra be $2^{\mathbb{N}}$. Exercise: verify this is a σ-algebra, i.e. check the conditions from the definition of a σ-algebra in definition 1.4 are satisfied for $\Omega = \mathbb{N}$ and $F = 2^{\mathbb{N}}$. Let ν be a probability measure on $(\mathbb{N}, 2^{\mathbb{N}})$; ν is not a random variable as it is not a function defined on the sample space ($\mathbb{N}$), but it essentially does the same job as our intuition regarding random variables by taking the formal random variable $X(\omega) = \omega$. We want to show (WTS) that ν has a RN derivative with respect to the counting measure μ. On $(\mathbb{N}, 2^{\mathbb{N}})$ $\mu(A) = |A|$, which is not a probability measure since $\mu(\mathbb{N}) = \infty$, so it is simply a measure. We know from the RN theorem that if (Ω, F) is such that F is a σ-algebra of Ω and μ and

ν are σ-finite measures with $\nu << \mu$, then there is a nonnegative measurable function f such that $\nu(A) = \int_A f d\mu$ for each $A \in F$. Furthermore, f is unique up to at most a set of μ-measure zero. To apply the RN theorem, we need to carefully verify the conditions in our application. We claimed that $2^{\mathbb{N}}$ is a σ-algebra of $\mathbb{N}$. To show that μ and ν are σ-finite measures, we need to show that there are a countably infinite or finite sequence of measurable sets $\{A_i\}_{i=1}^{\infty} \in F$ such that $\cup_{i=1}^{\infty} A_i = \Omega$ and $\mu(A_i) < \infty \; \forall i$. Since the measure ν is a probability measure, ν is σ-finite; this can be verified by taking $A_1 = \Omega$. The counting measure μ on $(\mathbb{N}, 2^{\mathbb{N}})$ is also σ-finite, which can be verified by taking $A_i = \{1, 2, \ldots, i\}$, so, $\mu(A_i) = i < \infty$ and $\cup_{i=1}^{\infty} A_i = \mathbb{N}$. We have already demonstrated $\nu << \mu$ earlier, i.e. $\nu(A) = 0 \Longrightarrow \nu(A) = 0$. To repeat, the only set that has μ-measure zero is the empty set, which must have ν-measure zero also; hence, $\nu << \mu$. So, the RN theorem guarantees the existence of a RN derivative.

Example 1.89 (Example 1.19 continued, part (iii))**.** While the RN theorem guarantees the existence of a RN derivative, it does not explicitly tell us what is is. Let's take a guess that it is the ordinary discrete probability function, $f(n) = \nu(\{n\})$. With this definition of f and that of the counting measure μ, if we let $A \in 2^{\mathbb{N}}$, we get that

$$\int_A f d\mu = \sum_{i \in A} f(i) = \sum_{i \in A} \nu(\{i\})$$

which is the same definition of $\nu(A)$ using the countable additivity property of measures since $A = \cup_{i \in A}\{i\}$, so $\nu(A) = \nu(\cup_{i \in A}\{i\}) = \sum_{i \in A} \nu(\{i\})$. So far we have verified that f has the property that for any $F \in 2^{\mathbb{N}}$, $\nu(A) = \int_A f d\mu$. The uniqueness part of the RN theorem shows that f is the only function with this property up to perhaps a set of μ-measure zero. However, the only set of μ-measure zero is the empty set itself, so f is completely unique. Therefore, we have shown that the RN derivative theorem holds in the case of a probability measure on $(\mathbb{N}, 2^{\mathbb{N}})$ with respect to the counting measure and that the RN derivative in this case is the familiar discrete probability function.

Recall that for conditional probability

$$P(A|B) = \frac{P(A, B)}{P(B)}$$

What if $P(B) = 0$?[11] We can solve this through the use of the RN theorem as follows. Let $G, A \in F$ with probability space (Ω, F, P) and define the measure $\nu(G) = P(A \cap G)$ such that $\nu << P$. RN implies that

$$P(A \cap G) = \int_G f(A|\zeta) dP(\zeta)$$

[11] As for how $P(B) = 0$ can occur, if we consider the thought experiment of throwing a piece of chalk in a class room and measuring the distance from the blackboard to the chalk, this distance that occurs will not happen twice.

for some f. This makes sense when you think intuitively think about it. Call $f(A|\zeta)$ the probability of A conditional on $\zeta \in G$. So, RN solves the problem of $P(A|B)$ when $P(B) = 0$ saying that it exists under certain circumstances, but RN does not tell us how to find it.

Chapter 2

Asymptotic Theory

2.1 Stochastic Relationships

We will say more now about the relationships between outcomes of random variables. While we have already defined the concept of independence for events, we can also define independence for random variables. When X and Y are independent, usually denoted $X \perp\!\!\!\perp Y$, then having information on any one of them will not provide any relevant information regarding the outcome of the other.

Definition 2.1. When X and Y are random variables, we say they are *independent* $\iff \forall(x, y)$

$$P(X \leq x \wedge Y \leq y) = P(X \leq x)P(Y \leq y) \tag{2.1}$$

or equivalently in terms of the joint CDF and marginal CDFs

$$F_{X,Y}(x, y) = F_X(x)F_Y(y)$$

Remark 2.2. To extend this definition to cases of independence of more than two random variables, note that the components of the vector of random variables (not a random vector) $X = (X_1, X_2, \ldots, X_n)$ are independent $\iff \forall x = (x_1, x_2, \ldots, x_n)$ we have that

$$F_X(x) = \Pi_{i=1}^{N} F_{X_i}(x_i)$$

In addition, we may express the definition of independence in terms of PDFs or PMFs instead of CDFs; we can go from the CDF definition to the PDF one by taking derivatives and vice-versa by taking integrals. In fact, if $X \perp\!\!\!\perp Y$, then for any measurable set B_X and B_Y in $\mathcal{B}(\mathbb{R})$

$$P(X \in B_X \wedge Y \in B_Y) = P(X \in B_X)P(Y \in B_Y)$$

In order to see this, express the condition defining independence in terms of PDFS, i.e. $f_{X,@Y}(x, y) = f_X(x)f_Y(y)$ and integrate both sides of $B_X \times B_Y$

$$\int_{B_x \times B_y} f_{X,Y}(x, y)dxdy = \int_{B_x \times B_y} f_X(x)f_Y(y)dxdy = \int_{B_x} f_X(x)dx \int_{B_y} f_Y(y)dy$$

Furthermore, from the definition of independence with PDFs (or PMFs), we can deduce that if $f_X(x) > 0$ and $X \perp\!\!\!\perp Y$, then

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(y)} = f_Y(y)$$

which (intuitively) means that if $X \perp\!\!\!\perp Y$, then knowing the value of X does not provide any extra information about Y. Thus, the conditional distribution of Y given X coincides exactly with the marginal distribution of Y.

From the concept of independence, we can define the important notion of a random sample also known as an independent identically distributed (iid) sample. This definition applies directly to random vectors as well, i.e. when each X_i maps in to $\mathbb{R}^k$.

Definition 2.3. We call the sample resulting from realisations of the random variables $X_1, X_2, \ldots, X_n$ a *random* or *iid sample* if the following properties hold:

1. $X_1, X_2, \ldots, X_n$ are independent;
2. each component X_i is distributed according to the same (identical) distribution.

The *size* of the random sample is $n \in \mathbb{N}$.

The concepts of covariance and correlation are related to the degree of linear relationship between two univariate random variables.

Definition 2.4. Whenever expectations exist, the *covariance* between two univariate random variables X and Y is given by

$$cov(X,Y) = E[(X - E(X))(Y - E(Y))]$$

Proposition 2.5.

$$\begin{aligned} cov(X,Y) &= E(XY) - E(X)E(Y) \\ V(X) &= cov(X,X) \\ cov(X,Y) &= cov(Y,X) \end{aligned}$$

Covariance is related to the coefficient of best linear relationship between X and Y, where best is in the sense of minimising mean squared error of prediction: the sign of the covariance indicates the sign of the slope of this best linear relationship. When the best linear relationship is positively (negatively) sloped, $(X - E(X))$ and $(Y - E(Y))$ are expected to move in the same (opposite) direction, so $cov(X,Y)$ is positive (negative). While covariance is a good indicator of the direction of the relationship between X and Y, it is a poor indicator of the strength of the relationship since it is not invariant to changes in the scale of measurement; this arises because the covariance is a bilinear operator.

Proposition 2.6. *The covariance operator is bilinear, i.e. for constants* $\{a_{x_1}, a_{x_2}, \ldots, a_{x_k}, b_x\}$ *and* $\{a_{y_1}, a_{y_2}, \ldots, a_{y_m}, b_y\}$ *we have that*

$$cov\left(\sum_{i=1}^{k} a_{x_i} X_i + b_x, \sum_{j=1}^{m} a_{y_j} Y_j + b_y\right) = \sum_{i=1}^{k}\sum_{j=1}^{m} a_{x_i} a_{y_j} cov(X_i, Y_j)$$

In the special case where $k = m = 1$*, this implies that*

$$cov(a_x X + b + x, a_y Y + b_y) = a_x b_y cov(X, Y)$$

Proof. Omitted. □

So, if we change the scale of any two random variables X and Y, e.g. if we compute the covariance of $\tilde{X} = a_x X$ and $\tilde{Y} = a_y Y$ for positive constants a_x, a_y, then

$$cov(\tilde{X}, \tilde{Y}) = cov(a_x X, a_y Y) = a_x a_y cov(X, Y)$$

so we will be changing the value of the covariance. This excludes the absolute value of the covariance from being a suitable measure of the strength of linear relationships.

Proposition 2.7.

$$X \perp\!\!\!\perp Y \implies cov(X, Y) = 0$$

Proof. Omitted. □

Example 2.8 (Memorize this point!)**.** In general, the reverse is not true, i.e. $cov(X, Y) = 0$ does not imply that $X \perp\!\!\!\perp Y$. Let the random variable α take values 0, $\pi/2$ and π with equal probability. Then $\zeta = \sin(\alpha)$ and $\eta = \cos(\alpha)$ are random variables and $E(\zeta) = 1/3$, $E(\eta) = 0$ and $E(\zeta\eta) = -$, so ζ and η are uncorrelated. However,

$$P(\zeta = 1 \wedge \eta = 1) = 0 \neq \frac{1}{9} = P(\zeta = 1)P(\eta = 1)$$

Definition 2.9. The *correlation coefficient* between two random variables X and Y, denoted $\rho(X, Y)$ is given by

$$\rho(X, Y) = \frac{cov(X, Y)}{\sqrt{V(X)}\sqrt{V(Y)}} = \frac{cov(X, Y)}{sd(X)sd(Y)}$$

provided $sd(X) > 0$, $sd(Y) > 0$ and $\exists cov(X, Y)$.

Like the covariance, the correlation coefficient is symmetric and measures the degree of linear relationship between two random variables. When it is defined, the sign of the correlation coefficient coincides with the sign of the covariance so it captures the direction of the linear relationship between two random variables. If $X \perp\!\!\!\perp Y$, then $\rho(X, Y) = 0$, but the reverse is not true in general. However, unlike the covariance, the correlation coefficient has properties that commend it as a suitable measure of the strength of linear relationships. Before considering these properties, we must first define the Cauchy-Schwarz Inequality.

Proposition 2.10 (Cauchy-Scharwz Inequality)**.** *Let X and Y be two random variables such that $\exists E(XY)$, though $E(XY)$ may be infinite. Then we have that*[1]

$$[E(XY)]^2 \leq E(X^2)E(Y^2)$$

Proof. Omitted. □

Corollary 2.11. *Let X and Y be two random variables. Then $|\rho(X,Y)| \in [0,1]$, i.e. $\rho(X,Y) \in [-1,1]$.*

Proof. From Cauchy-Schwarz Inequality:

$$\begin{aligned} 0 \leq |E(XY)| &\leq \sqrt{E(X^2)E(Y^2)} \\ \implies 0 \leq |\rho(X,Y)| &= \frac{|E(XY)|}{\sqrt{E(X^2)E(Y^2)}} \leq 1 \end{aligned}$$

□

This demonstrates the advantage of the correlation coefficient over the covariance, i.e. $\rho(X,Y) \in [-1,1]$. Furthermore, the correlation coefficient is robust to changes in scale and location of any of the two random variables.

Proposition 2.12. *Let X and Y be two random variables. For positive constants, a_x, a_y and arbitrary constants, b_x, b_y, define $\tilde{X} = a_x X + b_x$ and $\tilde{Y} = a_y Y + b_y$. Then we have that*

$$\rho(\tilde{X}, \tilde{Y}) = \rho(X,Y)$$

Proof. Omitted. □

Remark 2.13. The absolute value of the correlation coefficient is a measure of the degree of a linear relationship. When $|\rho(X,Y)| \approx 0$, the degree of the linear relationship between two random variables X and Y is low; when $|\rho| \approx 1$, the degree of linear relationship is high; when $\rho \approx 1$, there is a positive, direct linear relationship; when $\rho \approx -1$, there is a negative, opposite linear relationship; when $|\rho| = 1$, there is a 'perfect' linear relationship.

Proposition 2.14. *Let X be a random variable and let a, b be any constants such that $a \neq 0$. Then for $Y = aX + b$*

$$\rho(X,Y) = \begin{cases} 1 & \text{if } a > 0 \\ -1 & \text{if } a < 0 \end{cases}$$

Proof. Omitted. □

[1] More generally for those interested, the Cauchy-Schwarz Inequality says that $< x, y > \leq < x, x >^{\frac{1}{2}} < y, y >^{\frac{1}{2}}$ for all vectors x and y of an inner product space where $< \cdot, \cdot >$ is the inner product; defining inner product spaces formally is outside the scope of this course.

2.2 Limit Results

We will first introduce some notation for stochastic order dominance, i.e. for the rate of convergence of approximation error.

Definition 2.15. Let $\{x_n\}$ and $\{a_n\}$ be two sequences of constants, i.e. they are not random, i.e. they are nonstochastic. $\{x_n\}$ is a sequence of real numbers and $\{a_n\}$ is a sequence of positive numbers.

$$x_n = O(a_n) \text{ as } n \longrightarrow \infty$$

means that there is a constant $M \in \mathbb{R}, M < \infty : \left|\frac{x_n}{a_n}\right| \leq M \; \forall$ sufficiently large n; sometimes we replace the conditions 'sufficiently large n by $\forall n \in \mathbb{N}$. So, the sequence $\frac{x_n}{a_n}$ is bounded. We say that 'x_n converges no slower than a_n' or 'x_n is at most order a_n.'

Definition 2.16. Let $\{x_n\}$ and $\{a_n\}$ be two sequences of constants, i.e. they are not random, i.e. they are nonstochastic. $\{x_n\}$ is a sequence of real numbers and $\{a_n\}$ is a sequence of positive numbers.

$$x_n = o(a_n) \Longrightarrow \left|\frac{x_n}{a_n}\right| \longrightarrow 0 \text{ as } n \longrightarrow \infty$$

i.e. the sequence $\left|\frac{x_n}{a_n}\right| \longrightarrow 0$ in the standard Euclidean metric space $(\mathbb{R}, d)$ as $n \longrightarrow \infty$. We say that 'x_n converges faster than a_n' or 'x_n is of smaller order than a_n.'

Example 2.17. For the sequence

$$a_n = \frac{3}{n} + \frac{25}{n^2}$$

The following hold:

1. $a_n = O(\frac{1}{n})$
2. $a_n = o(\frac{1}{\sqrt{n}})$
3. $a_n \neq o(\frac{1}{n})$

For 1, we want to show (WTS):

$$\frac{\frac{3}{n} + \frac{25}{n^2}}{\frac{1}{n}}$$

is bounded. Observe that this reduces to showing that $3 + \frac{25}{n}$ is bounded. In fact this is less than 30 for all n. So, we have shown that the desired quantity is bounded; therefore, $a_n = O(\frac{1}{n})$.

For 2, WTS:

$$\frac{\frac{3}{n}+\frac{25}{n^2}}{\frac{1}{\sqrt{n}}} \longrightarrow 0 \text{ as } n \longrightarrow \infty$$

Equivalently, WTS:

$$\frac{3}{\sqrt{n}}+\frac{25}{n^{\frac{3}{2}}}$$

and this goes to zero as $n \longrightarrow \infty$; hence, we get the result.

For 3, WTS:

$$\frac{\frac{3}{n}+\frac{25}{n^2}}{\frac{1}{n}}$$

does not converge to zero as $n \longrightarrow \infty$. Note that this reduces to showing this for $3+\frac{25}{n}$, which of course converges to 3 as $n \longrightarrow \infty$; hence, we have the result.

Definition 2.18. Let $\{X_n\}$ be a sequence of random (stochastic) variables on a probability space (Ω, F, P) and let $\{a_n\}$ be a sequence of positive real numbers.

$$X_n = o_p(a_n) \Longrightarrow \left|\frac{X_n}{a_n}\right| \xrightarrow{p} 0 \text{ as } n \longrightarrow \infty$$

We say that 'X_n converges faster than a_n in probability' or 'X_n is of order smaller than a_n in probability.'

Definition 2.19. Let $\{X_n\}$ be a sequence of random (stochastic) variables on a probability space (Ω, F, P) and let $\{a_n\}$ be a sequence of positive real numbers. We write $X_n = O_p(a_n)$ to mean that $\forall \epsilon > 0, \exists M_\epsilon < \infty \wedge \exists N_\epsilon < \infty$ such that

$$P\left(\left|\frac{X_n}{a_n}\right| \geq M_\epsilon\right) = P(|X_n| \geq M_{epsilon} a_n) \leq \epsilon \quad \forall n \geq N_\epsilon$$

This condition is equivalent to

$$P\left(\left|\frac{X_n}{a_n}\right| \leq M_\epsilon\right) \geq 1-\epsilon \quad \forall n \geq N_\epsilon$$

We say that 'X_n converges no slower than a_n in probability' or '$\left|\frac{X_n}{a_n}\right|$ is bounded in probability.'

Remark 2.20. Note that for each of these concepts, a_n is the 'rate'. Furthermore, we could allow for a_n to be a random variable. We can also allow for a_n being possibly negative by saying $X_n = o_p(a_n)$ if there is a sequence Y_n such that $X_n = Y_n a_n$ and $Y_n \xrightarrow{p} 0$; similarly, we could say that $X_n = O_p(a_n)$ if there is a sequence Y_n such that $X_n = Y_n a_n$ and $Y_n = O_p(1)$.

Example 2.21. Let $\{Z_n\}$ be an iid sequence of random variables from distribution F. Then $\forall M > 0, n \geq 1$,

$$\begin{aligned} P\left(\left|\frac{Z_n}{1}\right| \geq M\right) &= 1 - P(|Z_n| < M) \\ &= 1 - P(-M < Z_n < M) \\ &= 1 - [F(M) - F(-M)] \end{aligned}$$

In order to make $P(|Z_n| \geq M)$ small, say less than some $\epsilon > 0$, all we need to do is to pick a sufficiently large M. It is possible to do this because as $M \longrightarrow \infty$, $F(-M) \longrightarrow 0$ and $F(M) \longrightarrow 1$. Therefore, $Z_n = O_p(1)$.

There are many useful results on relationships between the symbols for order notation. However, this would bring us to far off course. Instead, we will focus on limit results concerning sequences of random variables; Amemiya (1985) is a good reference [2].

Definition 2.22. We say that the sequence of random variables $\{X_n\}_{n=1}^{\infty}$ *converges in probability* to X (possibly a random variable) if $\forall \epsilon > 0$

$$\lim_{n \longrightarrow \infty} P(|X_n - X| < \epsilon) = 1$$

and we express this as $X_n \xrightarrow{p} X$ or equivalently $\text{plim}_{n \longrightarrow \infty} X_n = X$.

Intuitively, the sequence of random variables $\{X_n\}_{n=1}^{\infty}$ converges in probability to another random variable X if the realisations of their difference $X_n - X$ become arbitrarily concentrated around zero as $n \longrightarrow \infty$. Loosely speaking, for any arbitrarily small ball with center zero, the realisation of $X_n - X$ will occur inside the ball with probability that converges towards one. Graphically, the distribution of this difference tends to become more and more concentrated at zero.

Definition 2.23. Let the sequence of random variables $\{X_n\}_{n=1}^{\infty}$ have corresponding CDFs $\{F_n\}_{n=1}^{\infty}$ and consider the random variable X with corresponding CDF F. We say that the sequence $\{X_n\}_{n=1}^{\infty}$ *converges in distribution* to $X \iff$

$$\lim_{n \longrightarrow \infty} F_n(z) = F(z)$$

for all continuity points z of $F(\cdot)$. We express this convergence in distribution as $X_n \xrightarrow{d} X$.

Remark 2.24. While $X_n \xrightarrow{d} X$, this does not require that $X_n - X$ is close to zero in any (stochastic) sense. It is their CDFs that become increasingly similar. Note further that convergence in probability is stronger than convergence in distribution. Intuitively, $X_n \xrightarrow{p} X$ means that the realisations of X_n and X become increasingly close in the stochastic sense, so their CDFs should also become increasingly close. However, while the distributions are getting increasingly closer, this does not necessarily imply that the actual realisations are close in any sense. We see this in the lemma 2.25.

Lemma 2.25.

$$X_n \xrightarrow{p} X \Longrightarrow X_n \xrightarrow{d} X$$

Proof. Omitted. □

Lemma 2.26. *$X_n \xrightarrow{d} X$ does not necessarily imply that $X_n \xrightarrow{p} X$.*

Proof. See the following counter-example.[2] □

Example 2.27. Let $\{X_n\}_{n=1}^{\infty}$ be independent draws from $N(0,1)$ and let $X \sim N(0,1)$ also; hence, $X_n \xrightarrow{d} X$. However, since a linear combination of Normally distributed random variables is also Normally distributed and draws are independent, $X_n - X \sim N(0,2)$. So, taking any arbitrary $\epsilon > 0$:

$$\begin{aligned} P(|Z_n - Z| < \epsilon) &= 1 - P(|Z_n - Z| \geq \epsilon) \\ &= 1 - P(Z_n - Z \geq \epsilon \wedge Z_n - Z \leq \epsilon) \\ &= 1 - [P(Z_n - Z \leq \epsilon) + P(Z_n - Z \leq -\epsilon)] \\ &= 1 - [1 - P(Z_n - Z < \epsilon) + P(Z_n - Z \leq -\epsilon)] \\ &= P(Z_n - Z < \epsilon) - P(Z_n - Z \leq -\epsilon) \\ &= \Phi(\epsilon) - \Phi(-\epsilon) \end{aligned}$$

which is a positive constant. For example with $\epsilon = 0.1$, $P(|Z_n - Z| < \epsilon) = 0.0797 < 1$. As we have shown this for an arbitrary choice of ϵ, it holds for all ϵ.

However, the converse also holds for the special case when the limiting random variable is degenerate at a certain value (i.e. if $X = c$ with probability one). The intuition behind this example is that if F_n is converging to F but F is degenerate at c, then the realisations of X_n are becoming increasingly concentrated at c; hence, $X_n \xrightarrow{p} c$.

Lemma 2.28. *If $X_n \xrightarrow{d} X$ and X is degenerate at c, i.e.*

$$F_X(z) = \begin{cases} 0 & \text{if } z < c \\ 1 & \text{if } z \geq c \end{cases}$$

then $X_n \xrightarrow{p} X$.

Proof. Omitted. □

[2] Some people disagree with the terminology 'counter-example'. If something is true (in the sense that X does not imply Y), then we can show an example of this, rather than use the prefix 'counter'. For purposes of this course in aiding your experience of the idea of proof by contradiction or proof by 'counter-example', I will use the terminology 'counter-example' as I find it tends to be initially more clarifying when studying methods of proof, though I accept the argument against the use of this term.

Definition 2.29. We say that the sequence of random variables $\{X_n\}_{n=1}^{\infty}$ *converges with probability one* (*almost surely, almost everywhere*) to the random variable X if

$$P(\{\omega : \lim_{n\longrightarrow\infty} X_n(\omega) \neq X(\omega)\}) = 0$$

We write this is $X_n \xrightarrow{a.s.} X$.

Definition 2.30. We say that the sequence of random variables $\{X_n\}_{n=1}^{\infty}$ *converges in mean square* to the random variable X if

$$\lim_{n\longrightarrow\infty} E[(X_n - X)^2] = 0$$

We write this as $X_n \xrightarrow{2} X$.

Definition 2.31. Let $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ where X_i is a random variable whose mean is μ_i. We call S_n the *sample mean.*

Note that

$$E(S_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mu_i \equiv \bar{\mu_n}$$

Informally, the Weak Law of Large Numbers (WLLN) states that given certain conditions, the sample mean converges in probability to the population mean as the sample size goes to infinity. The next few definitions and results are very useful tools in analysis of identification.

Proposition 2.32 (Markov's Inequality)**.** *Let X be a random variable. Then, for any $\alpha > 0$ and $k \in \mathbb{N}$*

$$P(|X| \geq \alpha) \leq \frac{1}{\alpha^k} E(|X|^k)$$

Proof.

$$\begin{aligned} E(|X|^k) &= \int_{-\infty}^{\infty} z^k f_{|X|}(z)dz \\ &\geq \int_{\alpha}^{\infty} z^k f_{|X|}(z)dz \\ &\geq \alpha^k \int_{-\infty}^{\infty} f_{|X|}(z)dz \\ &= \alpha^k P(|X| \geq \alpha) \end{aligned}$$

□

A simple version of the *Markov Inequality* is the following: if g is a continuous and nonnegative function and $d > 0$ is a constant, then

$$P(g(x) \geq d) \leq \frac{E[g(x)]}{d}$$

Proof. Observer that $d1[g(X) \geq d] \leq g(X)$. Taking expectations and rearranging yields the Markov Inequality since the expectation of an indicator function is a probability, i.e.

$$E[1[g(X) \geq d]] = P(g(X) \geq d) \qquad \square$$

Sometimes Markov Inequality refers to the case where $g(X) = |X|$.

Corollary 2.33 (Chebyshev's Inequality)**.** *Let X be a random variable. Then, for any $\alpha > 0$*

$$P(|X - E(X)| \geq \alpha) \leq \frac{1}{\alpha^2} V(X)$$

So, when $g(X) = |X - \mu|^2$ where $\mu = E(X)$, we get that:

$$\begin{aligned} P(|X - \mu|^2 \geq \epsilon^2) \leq \epsilon^2) &\leq \frac{E(|X - \mu|^2)}{\epsilon^2} \\ \iff P(|X - \mu| \geq \epsilon) &\leq \frac{Var(X)}{\epsilon^2} \end{aligned}$$

which is *Chebyshev's Inequality.*

Chebyshev's Inequality can thus be seen to be a special case of Markov's Inequality where $g(x) = |X - \mu|^2$ and $d = \epsilon^2$. The *one-sided Chebyshev Inequality* is:

$$P(X - \mu > k\sigma) \leq \frac{1}{1 + k^2}$$

Other (useful) versions include:

$$\begin{aligned} P(|X - \mu| \geq d\sigma) &\leq \frac{1}{d^2} \\ P(|X - c| \geq d) &\leq \frac{E[(X - c)^2]}{d^2} \end{aligned}$$

where $\sigma^2 = Var(X), d > 0$ and $c \in \mathbb{R}$.

Proposition 2.34 (WLLN)**.** *Let $\{X_i\}_{i=1}^n$ be a sequence of uncorrelated random variables, each having finite mean μ_i and variance σ_i^2. Assume further that $\sigma_i^2 \leq M < \infty \; \forall i$. Then $S_n \xrightarrow{p} \bar{\mu_n}$.*

Proof. From Chebyshev's Inequality for any $\epsilon > 0$:

$$P(|S_n - \bar{\mu_n}| \geq \epsilon) \leq \frac{1}{\epsilon^2} V(S_n)$$

Using the bilinearity of covariances

$$\begin{aligned} V(S_n) &= cov(S_n, S_n) \\ &= cov\left(\frac{1}{n}\sum_{i=1}^{n} X_i, \frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} cov(X_i, X_i) + \frac{1}{n^2}\sum_{i\neq j} cov(X_i, X_j) \\ &=\leq \frac{M}{n} \\ \therefore P(|S_n - \bar{\mu_n}| < \epsilon) &\geq 1 - \frac{M}{n\epsilon^2} \\ \implies \lim_{n\longrightarrow\infty} P(|S_n - \mu| < \epsilon) &= 1 \end{aligned}$$

□

Remark 2.35. Note that we only require the random variables to be uncorrelated, not necessarily iid. There are many versions of the LLN (even the WLLN) under various sets of assumptions – this is important to know.

Proposition 2.36 (Strong Law of Large Numbers)**.** *Let $\{X_i\}_{i=1}^n$ be a sequence of iid random variables with finite mean μ. Then $S_n \xrightarrow{a.s.} \mu$.*

Proof. Omitted. □

Let us compare the WLLN with the Strong Law of Large Numbers (SLLN). WLLN (SLLN) assumes that the sequence of random variables is uncorrelated (iid) and that $\mu_i < \infty \wedge \sigma_i^2 < \infty \; \forall i$ (says nothing about the variance, just $\mu < \infty$), so SLLN has more strict assumptions. WLLN (SLLN) predicts that $S_n \xrightarrow{p} \bar{\mu_n}$, i.e. '$p$' and '$E(S_n) < \infty$' ($S_n \xrightarrow{a.s.} \mu$, i.e. 'a.s.' and $\mu_i = \mu \; \forall i < \infty$), so SLLN is a stronger result. As mentioned in remark 2.35, there are lots of versions of the LLN under different assumptions (relevant for specific cases say when we can assume independence or dependence, etc.) and time spent exploring these can be very beneficial for those of you interested in asymptotic theory directly or indirectly; the same holds true for the Central Limit Theorems we will talk about soon. First, we will describe the continuous mapping theorem (CMT), which is a very useful result to show consistency of estimators.

Theorem 2.37 (CMT)**.** *Let X_n be a k-dimensional random vector with the property that $X_n \xrightarrow{p} c$ where c is a constant and let $f : \mathbb{R}^k \longrightarrow \mathbb{R}^m$ be continuous at c. Then $f(X_n) \xrightarrow{p} f(c)$.*

Proof. Omitted. □

Corollary 2.38. *Let $X_n \xrightarrow{p} x$ and $Y_n \xrightarrow{p} y$. Then*

- $dim(X_n) = dim(Y_n) \implies X_n + Y_n \xrightarrow{p} x + y \wedge X_n - Y_n \xrightarrow{p} x - y$
- $dim(Y_n) = 1 \implies X_n Y_n \xrightarrow{p} xy$

- $(dim(Y_n) = 1 \wedge y \neq 0) \implies \frac{X_n}{Y_n} \xrightarrow{p} \frac{x}{y}$

While WLLN shows that under certain assumptions the sample means converges in probability to the (common) expectation μ, it turns out that the rate at which this occurs is $\sqrt{n}$, i.e. $\sqrt{n}(S_n - \mu)$ converges to a non-degenerate distribution, which can be characterised by the Central Limit Theorem (CLT).

Theorem 2.39 (Lindberg-Levy CLT)**.** *Let $\{X_i\}_{i=1}^n$ be iid with mean μ and variance $\sigma^2 < \infty$. Then*

$$\sqrt{n}\left[\frac{S_n - \mu}{\sigma}\right] \xrightarrow{d} N(0,1)$$

Proof. Omitted. □

CLTs are generally more difficult to show than LLNs. Intuitively, this is because we have to show that a whole distribution converges. Slutsky's theorem is very useful to show the limiting distribution of estimators.

Theorem 2.40 (Slutsky's Theorem)**.** *Let X_n and Y_n be two sequences of univariate random variables. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, where c is a constant, then*

$$X_n + Y_n \xrightarrow{d} X + c$$
$$X_n Y_n \xrightarrow{d} Xc$$

Proof. Omitted. □

2.2.1 Delta Method

We will finish this section by looking at the delta method. A full treatment of the delta method, like the LLN and CLT above would require a lengthier course and one that is concerned more with the theoretical side of econometrics. For now, we will present what I hope is an adequate introduction to the delta method in order to provide a basic understanding of how it works. For those interested in seeing the univariate, multivariate and general multivariate method with examples, please consult Amemiya (1985) [2]. Let us start by supposing that

$$\sqrt{n}(X_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

where X_n is a scalar. We would like to know what $\sqrt{n}(g(X_n) - g(\mu))$ converges to in distribution. Let us assume that g is at least once continuously differentiable, where $g'(\mu) \neq 0$. Now take a first order Taylor series approximation of g around μ:

$$g(X_n) = g(\mu) + g'(\bar{X_n})(X_n - \mu)$$

where $\mu \leq \bar{X_n} \leq X_n$. This is called the *Lagrange form* of the remainder in the Taylor series approximation. We can write:

$$\begin{aligned}\sqrt{n}(g(X_n) - g(\mu)) &= \sqrt{n}(g(\mu) + g'(\bar{X_n})(X_n - \mu) - g(\mu)) \\ &= \sqrt{n}(g'(\bar{X_n})(X_n - \mu))\end{aligned}$$

Observe

$$\sqrt{n}(X_n - \mu) \xrightarrow{d} N(0, \sigma^2) \Longrightarrow X_n \xrightarrow{p} \mu$$

This is can be seen by letting $\mu = 0$ without loss of generality. Then because $Z_n = \sqrt{n}X_n$ is asymptotically Normal, it is bounded in probability for sufficiently large n.

$$\therefore \forall \epsilon > 0, \exists M_\epsilon < \infty : P(|X_n| > \frac{M_\epsilon}{\sqrt{n}}) < \epsilon$$

Pick an arbitrary δ. Then we have that $P(|X_n| > \delta) = P(|Z_n| > \sqrt{n}\delta)$. Choose n sufficiently large so that $\sqrt{n}\delta > M_\epsilon$. Then we have that $P(|X_n| > \delta) < \epsilon$. As our choice of ϵ was arbitrary, we have that $P(|X_n| > \delta) \longrightarrow 0$, i.e. boundedness in probability of Z_n implies that $X_n = \frac{Z_n}{\sqrt{n}}$ converges in probability to zero; hence, $X_n \overset{p}{\mu}$. Furthermore, because $\bar{X_n}$ lies between μ and X_n, we have that $\bar{X_n} \overset{p}{\mu}$. Thus, sine we assumed that g' is continuous, $g'(\bar{X_n}) \xrightarrow{p} g'(\mu)$. It follows by Slutsky's theorem that

$$\sqrt{n}(g'(\bar{X_n})(X_n - \mu)) = g'(\bar{X_n})\sqrt{n}(X_n - \mu) \xrightarrow{d} g'(\mu)N(0, \sigma^2) = N(0, g'(\mu)^2\sigma^2)$$

$$\therefore \sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} N(0, g'(\mu)^2\sigma^2)$$

For this to be a non-degenerate distribution, $g'(\mu) \neq 0$.

2.3 Properties of Estimators

Definition 2.41. An *estimator* is simply a function of the sample $(X_1, \ldots, X_n)$.

The aim of an estimator is usually to approximate some underlying parameter that characterises a distribution of interest. One example is the population mean $E(X)$ for which we might choose to use the sample average S_n as an estimator. As it is a function of random variables, an estimator is a random variable.

Definition 2.42. An *estimate* is the particular value that an estimator takes for a particular realisation of the sample $(x_1, \ldots, x_n)$.

An estimate is a number rather than a random variable. For the same example of the mean, the value $s_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ is an estimate for the population mean. We are usually interested in constructing estimators to satisfy certain desirable properties; for instance we are interested in having resulting estimates that are 'centered' at the truth and are as 'precise' as possible.

Definition 2.43. Let $\hat{\theta}$ be an estimate of the parameter θ. We define the *bias of* $\hat{\theta}$ by

$$Bias_\theta(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Definition 2.44. An *unbiased* estimator $\hat{\theta}$ of the parameter θ is such that $Bias_\theta(\hat{\theta}) = 0$.

Remark 2.45. Since unbiasedness does not refer to any sample size, i.e. it should hold for any sample size, we call it a *small sample property*.

Example 2.46. Let the sample $X_1, \ldots, X_n$ be such that $E(X_i) = \mu \ \forall i = 1, \ldots, n$. It turns out that the sample mean S_n is an unbiased estimator of the population mean:

$$E(S_n) = E\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu$$

while the sample variance $\frac{1}{n}\sum_{i=1}^{n}(X_i - S_n)^2$ is biased.

Definition 2.47. We say that the estimator $\hat{\theta}$ is (*weakly*) *consistent* for θ when $\hat{\theta} \xrightarrow{p} \theta$.

Remark 2.48. Since consistency refers to the behaviour of the estimator as the sample tends towards infinity, we call it a *large sample property*.

Example 2.49. Let the sample $X_1, \ldots, X_n$ be such that $E(X_i) = \mu \ \forall i = 1, \ldots, n$. The sample mean S_n is (weakly) consistent for the population mean by the WLLN, i.e. $S_n \xrightarrow{p} \mu$.

Definition 2.50 (MSE)**.** The *mean square error* (MSE) is a natural measure of the distance between an estimator and the parameter and is defined by

$$MSE_\theta(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$$

or equivalently

$$MSE_\theta(\hat{\theta}) = \left[Bias_\theta(\hat{\theta})\right]^2 + V(\hat{\theta})$$

MSE motivates the following definition of relative efficiency between two estimators, which is a finite sample property.

Definition 2.51 (Relative Efficiency)**.** Let $\hat{\theta_1}, \hat{\theta_2}$ denote two estimators of a parameter $\theta \in \Theta$, the parameter space. We say that $\hat{\theta_1}$ is *more efficient relative to* $\hat{\theta_2}$ if it has lower or equal MSE for every value of θ in the parameter space:

$$MSE_\theta(\hat{\theta_1}) \leq MSE_\theta(\hat{\theta_2}) \ \forall \theta \in \Theta \quad \text{smallest MSE}$$

$$\left[Bias_\theta(\hat{\theta_1})\right]^2 + V(\hat{\theta_1}) \leq \left[Bias_\theta(\hat{\theta_2})\right]^2 + V(\hat{\theta_2}) \ \forall \theta \in \Theta \quad \ldots equivalently$$

At a first glance, relative efficiency may seem like a natural concept with which to select best estimators. However, its dependence on the value of the underlying parameter renders the task of finding an estimator that is relatively more efficient than any other estimator infeasible.

Example 2.52. Consider the task of estimating the parameter $\theta = E(\theta)$ and let the estimator be $\hat{\theta} = 5$. A constant function of the sample, it may seem a strange estimator. Observe $MSE_\theta(\hat{\theta}) = (5-\theta)^2$. So, when $\theta = 5$ the constant estimator $\hat{\theta} = 5$ is the best estimator along the lines of the MSE criterion.

Example 2.52 serves to illustrate the point that it is impossible to minimise the MSE over all $\theta \in \Theta$. Let us restrict the class of estimators we consider. In small samples, we restrict attention to the class of unbiased estimators and this leads to the following definition.

Definition 2.53 (Efficiency)**.** We say that an estimator $\hat{\theta_n}$ of θ is *efficient* if it has the smallest variance among all unbiased estimators of θ:

1. $E(\hat{\theta_n} - \theta) = 0 \; \forall \theta$ (unbiased)
2. $V(\hat{\theta_n}) \leq V(\hat{\theta_n}') \; \forall \hat{\theta_n}' \; s.t. \; E(\hat{\theta_n}' - \theta) = 0 \; \forall \theta$ (minimum variance)

Mostly, we do not have exact finite sample results in econometrics, so we focus on large sample results instead.

Definition 2.54 (Asymptotic Efficiency)**.** An estimator $\hat{\theta_n}$ is the *best asymptotically Normal* (BAN) or *asymptotically efficient* estimator $\iff$

1. $\sqrt{n}(\hat{\theta_n} - \theta) \xrightarrow{d} N(0, \hat{\sigma}^2) \; \forall \theta \in \Theta$ (analog to unbiasedness)
2. $\hat{\sigma}^2 \leq \tilde{\sigma}^2 \; \forall \tilde{\theta_n} \; s.t. \; \sqrt{n}(\tilde{\theta_n} - \theta) \longrightarrow N(0, \tilde{\sigma}^2) \; \forall \theta \in \Theta$ (analog to minimum variance)

To know when an estimator is efficient or asymptotically efficient, or whether we are close to efficiency or asymptotic efficiency, we can use the Cramer-Rao (CR) lower bound. This bound imposes a lower bound on the variance for any unbiased estimator under certain regularity conditions. See your notes from the first term on the Cramer-Rao lower bound. Remember that these first two chapters will be useful reference material for the course, which starts in chapter 3.

Chapter 3

Identification

3.1 Problem of Identification

This chapter is based on Manski (2007)[59]. To motivate this topic, let us consider a number of examples.

Example 3.1 (Death Penalty)**.** [1] Looking at a classic two-player game between criminals and society where r_1 denotes how much crime a criminal chooses to commit and r_2 is how touch society can be in terms of sanctions (i.e. toughness against crimes). For each covariate x_j, let $R_{j1}(\cdot)$ be the reaction of criminals to society and $R_{j2}(\cdot)$ be the reaction of society to criminals. So $r_{j1} = R_{j1}(r_{j2})$ and $r_{j2} = R_{j2}(r_{j1})$ are the reactions of criminals and society respectively, which are functions of each other's choices. These will be the equilibrium conditions in a two player game where our objective is $P[R_1(\cdot), R_2(\cdot)|X]$ and our data is $P(r_1, r_2|x)$. Questions include what the crime rate would be if the sanction was r_2 and how tough would society need to be as a function of the crime rate. What is of key interest is the deterrent effect of penalties on the crime rate. This is a very hard *identification* problem since we can only observe the crime rate under a given sanction policy but not under different sanction policies. Then at Chicago University but now at the University at Buffalo, Isaace Ehrlich wrote a highly controversial 1975 AER paper of a 2 person game using a linear model with linear homogeneous reaction functions ($R(r) = \beta r + \epsilon$). The crime of interest was murder and the sanction of interest was execution. So, the focus was on the deterrency of the death penalty and Ehrlich estimated β, finding that one execution deters eight murders. Had Ehrlich provided econometric evidence that the death penalty was a good idea? The National Research Council (NRC) investigated this study and after extensive work reached the conclusion that the identification problem was so severe that it could not be believed. The NRC was recently asked the same question on whether empirical studies have provided scientifically valid evidence to determine if murder rates

[1] Isaac Erhlich's 1975 AER paper 'The Deterrent Effect of Capital Punishment: A Question of Life or Death'.

are affected by the death penalty. They reached the same conclusion.[2] Clearly, identification has huge importance in the realm of policy.

Example 3.2. Let T denote the set of treatments in policy analysis or comparative statics. We will associate the term ambiguity with Knightian uncertainty, i.e. where we can not even quantify the distribution of outcomes. Let $t \in T$ denote a particular treatment and assume that treatments are mutually exclusive and exhaustive; for example with cancer, assume that there are only two treatments that are surgery and chemotherapy and so the treatment set is either, both or neither; hence, the set of treatments consists of four mutually exclusive elements that are exhaustive. First note that we should distinguish $P(y(t)|x)$ and $P(y|x, z = t)$: the first is the hypothetical probability if all people receive treatment t, which is unobserved (e.g. demand function) whereas the second is the realised treatments (observed, e.g. quantity demanded). The essence of the simultaneity problem is that we do not observe demand functions, only the equilibrium price, so we are using a counterfactual (defined shortly) and with heterogeneity, other markets have different market equilibria. So, if we extrapolate (defined shortly), we get counterfactual situations. For example, with macro policy with respect to government and central bank stimuli after the financial crisis, some say this was necessary to prevent a greater depression. However, we cannot refute someone saying that these stimuli did not matter. To ease subsequent conversations about this, I will refer to the Krugman view that the stimulus was not big enough versus (in America) the Republican attitude that the stimulus does not matter. Equivalently, we cannot refute the hypothesis that response $y(t)$ is an increasing function of t, i.e. that increasing the stimulus would have helped. For example, with the fiscal multiplier, say the impact of government spending on macro policy, postulating a high multiplier would imply a more liberal attitude and a small one would reflect a more conservative outlook. These are all examples of non-refutable hypothesis (defined shortly). From data alone, it is impossible to work out the best treatment. If $Tr(A) > Tr(B)$ where $Tr(A)$ is the outcome under treatment A, typically an assumption has been made somewhere. Some papers are clear and some other papers are less clear about what assumptions are made, however. The study of identification will allow you to understand and work through all of these issues and more. With that in mind, let us start our formal study of identification.

Consider the model

$$\begin{aligned} y &= x'\beta + \epsilon \\ E(\epsilon|x) &= 0 \end{aligned}$$

where x is $k \times 1$, alternatively defined by

$$E[y|x] = x'\beta$$

[2] A world leader in the field of identification, Northwestern's Charles Manski was a member of the committee working on the 1978 report then at Carnegie-Mellon and on the 2012 report.

Suppose we have data $(y_i, x_i)_{i=1}^N$ where N is extremely large and an independently and identically distributed (iid) sample. Here with identification, we always work under the assumption that you can observe the population (i.e. $N = \infty$) so $E(y|x) = x'\beta$ is 'known'. Our question relates to the identification of β in the model above with given data; what can we learn about β?

$$\left\{\begin{matrix}\text{Model}\\ + \\ \text{Data}\end{matrix}\right\} \overset{\text{Identification Analysis}}{\Longrightarrow} \text{Information about } \beta$$

Definition 3.3. A parameter $b \in \mathbf{R}^k$ is *identified relative to* β if $P_X\{x : x'b \neq x'\beta\} > 0$.

Definition 3.4. In the model above, β is *point identified* if $\forall b \neq \beta$, b is identified relative to β.

Let us look for a sufficient condition for point identification of β.

Lemma 3.5. *If the matrix Exx' has rank k, then β is point identified.*

Proof. Let $b \neq \beta$.

$$\begin{aligned} E\{(x'(b-\beta))^2\} &= (b-\beta)'Exx'(b-\beta) \geq 0 \\ &= 0 \text{ only if } b = \beta \\ \implies x'(b-\beta) &\neq 0 \text{ on a set of positive measure} \\ \implies x'b &\neq x'\beta \text{ on a set of positive measure} \end{aligned}$$

□

Remark 3.6. A point identified model does not imply that the objective function to estimate β has a unique minimum. However, if the objective function has a unique minimum, then the model is point identified.

Alternative definitions of a parameter being identified include the following.

Definition 3.7. The parameter vector $\boldsymbol{\theta}_0$ is *identified* if for any other parameter vector $\boldsymbol{\theta} \in \Theta$, the set

$$\{y|f(y|\boldsymbol{\theta}) \neq f(y|\boldsymbol{\theta}_0)\}$$

has positive probability.

Definition 3.8. Let Θ be a parameter space and consider a family of parametric probability measures $\{P(\cdot;\theta) : \theta \in \Theta\}$ that are absolutely continuous w.r.t. a σ-finite measure ν (such as Lebesgue or counting). The parameter θ is *identified* if for $\theta_1 \neq \theta_2$, we have that $P(\cdot;\theta_1) \neq P(\cdot;\theta_2)$.

Suppose that we have a set of parameters Θ and a family of probability distributions $P(x;\theta)$. A necessary condition for parameter $\theta^* \in \Theta$ to be identified is that, for any $\theta \in \Theta$, $\theta \neq \theta^*$, $P(\cdot;\theta^*) \neq P(\cdot;\theta)$. Sufficient if $P(x;\theta)$ absolutely continuous w.r.t. σ-finite measure ν.

Example 3.9. Suppose that we observe (x, y) where

$$y = \begin{cases} 0 & \alpha x + \epsilon < \gamma \\ 1 & \alpha x + \epsilon \geq \gamma \end{cases}$$

where $x \perp\!\!\!\perp \epsilon$, $\epsilon \sim N(\mu, \sigma^2)$ and x has some known distribution, which doesn't depend on any of $\alpha, \gamma, \mu, \sigma^2$.

(a) We claim first that $\theta = (\alpha, \gamma, \mu, \sigma^2)$ is not identified:

Proof. We have to check whether, for any θ, there exists $\theta' \neq \theta$ s.t. $\forall x$, $P(y = 0|x; \theta) = P(y = 0|x; \theta')$.

Note that we can restrict ourselves only to conditional distributions instead of the multivariate ones, since we know the distribution of x.

Then $P(y = 0|x; \theta) = P(\alpha x + \epsilon < \gamma|x) = P(\frac{\epsilon - \mu}{\sigma} < \frac{\gamma - \alpha x - \mu}{\sigma}|x) = \Phi(\frac{\gamma - \alpha x - \mu}{\sigma})$ where Φ is the cumulative normal distribution function.

To see that θ is not identified, take $\theta' = (k\alpha, k\gamma, k\mu, k^2\gamma^2)$, for some $k > 0$. Then we have $P(y = 0|x; \theta) = P(y = 0|x; \theta')$. □

(b) Thus, first normalize $\sigma^2 = 1$. Then the parameters (α, γ, μ) are still not identified (although α alone is identified):

Proof. Take $\theta' = (\alpha, \gamma + A, \mu + A)$ and we see that $\Phi(\gamma + A - \alpha x - \mu - A) = \Phi(\gamma - \alpha x - \mu)$. □

(c) If we normalise again $\gamma = 0$, then (α, μ) are finally identified:

Proof. Suppose that $\forall x, \Phi(-(\alpha x + \mu)) = \Phi(-(\alpha' x + \mu'))$. Since Φ is 1-1, it follows that, $\forall x, \alpha x + \mu = \alpha' x + \mu' \implies (\alpha - \alpha')x = \mu' - \mu$. This implies that $\alpha = \alpha'$ and $\mu' = \mu$. □

Identification example: your own movements and your movements in a mirror – which drives which or do both move due to an external stimulus? This reflection problem (Manski, 1993) [58] arises if you try to interpret the common observation that individuals belonging to the same group tend to behave similarly. Three hypotheses have been proposed to explain this phenomenon:

1. *endogenous effects*: 'propensity of an individual to behave in some way varies with the prevalence of the behaviour in the group'.
2. *contextual effects*: 'propensity of an individual to behave in some way varies with the distribution of background characteristics in the group'.
3. *correlated effects*: 'individuals in the same group tend to behave similarly because they face similar environments or have similar individual characteristics'.

Why would we care about what generates observed patterns of group behaviour? One reason is that different processes have different ramifications for *public policy*. Data alone cannot reveal which hypothesis might be correct, so to draw conclusions we need to combine empirical evidence (data) with assumptions. This is an identification problem.

Definition 3.10. *The Law of Diminishing Credibility*: the credibility of inference decreases with the strength of the assumptions made. (Manski, 2007: 3) [59]

Manski (2007) [59] distinguishes identification and statistical inference as follows:

> 'Studies of identification seek to characterize the conclusions that could be drawn if one could use the sampling process to obtain an unlimited number of observations. Studies of statistical inference seek to characterize the generally weaker conclusions that can be drawn from a finite number of observations.' (Manski, 2007: 3) [59]

Logically, identification precedes inference much like the study of probability precedes that of statistics. Koopmans (1949: 132) [55] introduced the term 'identification' into econometric literature as follows.

> 'In our discussion we have used the phrase "a parameter that can be determined from a sufficient number of observations." We shall now define this concept more sharply, and give it the name *identifiability* of a parameter. Instead of reasoning, as before, from "a sufficiently large number of observations" we shall base our discussion on a hypothetical knowledge of the probability distribution of the observations, as defined more fully below. It is clear that exact knowledge of this probability distribution cannot be derived from any finite number of observations. Such knowledge is the limit approachable but not attainable by extended observation. By hypothesizing nevertheless the full availability of such knowledge, we obtain a clear separation between problems of statistical inference arising from the variability of finite samples, and problems of identification in which we explore the limits to which inference even from an infinite number of observations is suspect.'

Data and assumptions lead to conclusions. To paraphrase Manski, we can only overcome identification problems by making stronger assumptions or by initiating new sampling processes that yield different kinds of data rather than gathering more of the same kind of data.

Definition 3.11. *t is in the support of P* if

$$P[t - \delta \leq y \leq t + \delta] > 0 \; \forall \delta > 0$$

An equivalent definition is the following.

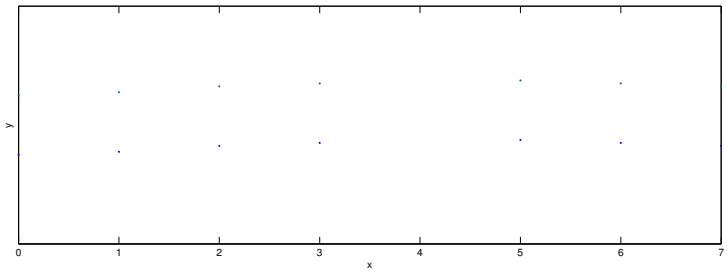

Figure 3.1: Confidence interval for $E(y|x)$ when $n = 100$.

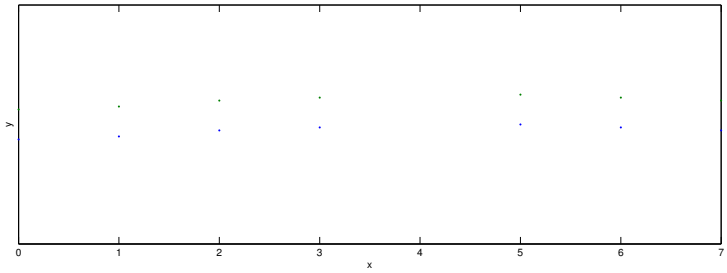

Figure 3.2: Confidence interval for $E(y|x)$ when $n = 1000$.

Definition 3.12. A covariate value x_0 is *on the support of the distribution* $P(x)$ if there is positive probability of observing x arbitrarily close to x_0. A covariate value x_0 is *off the support* of $P(x)$ if there is zero probability of observing x within some neighbourhood of x_0.

Let us now consider the problem of extrapolation – i.e. prediction off the support, in particular the problem of predicting a random variable y conditional on x where x only takes values at $\{0, 1, 2, 3, 5, 6, 7\}$. We can compute a tighter confidence interval for the mean of $y|x$ with 1000 observations of (y, x) than we can with 100 observations. Figures 3.1 and 3.2 represent each case. The width of the confidence intervals relates to a statistical problem, since we can estimate $E(y|x)$ more precisely with more data. However, at $x = 4$, the confidence interval is infinite irrespective of sample size, which means we are dealing with an identification problem.

Extrapolation requires an assumption that restricts $E(y|x)$ globally. Invariance assumptions (e.g. $P(y|x = x_0) = P(y|x = x_1)$ i.e. assume that y behaves in the same way at x_0 as it does at some specified x_1 on the support of $P(x)$ like in RAND study – see example 3.19). Invariance assumptions are special cases of general idea of *shape restrictions*, e.g. linearity, monotonicity. A related concept to extrapolation is generalisability or external validity.

Definition 3.13. An experiment is said to have *external validity* if the distribution of outcomes realized by a treatment group is the same as the distribution of outcomes that would be realized in an actual program. (Manski, 2007:27) [59]

The first function of theory is to allow extrapolation, while the second function is to improve the sampling precision for estimation on the support. As you will be aware from your study of causality in the first term, sometimes you may recognize that theory purports to explain why observed outcomes occur. However, when making predictions, we are less interested in why things happen and more interested in will associations hold. Unfortunately, theory tends to be testable when least necessary, i.e. we can learn $P(y|X)$ on the support of $P(x)$ and theory tends to be least testable when most needed, i.e. for learning $P(y|x)$ off the support of $P(x)$. Essentially, failures off support are inherently not detectable.

With regard to our study of identification in this chapter, we will concentrate on problems arising in prediction and decision. With prediction, our goal is to learn the probability distribution of an outcome y conditional on a covariate x. With decision problems, we focus on cases where the relative merits of alternative actions depend on the outcome distribution $P(y|x)$ and ask how a decision maker might choose an action when available data and credible assumptions only partially identify this distribution. This chapter may challenge your ideas about partial identification:

> 'For most of the twentieth century, econometricians and statisticians commonly thought of identification as a binary event – a parameter is either identified or not.' (Manski, 2007: 11) [59]

Researchers need to be more comfortable with expressing uncertainty and acknowledging ambiguity. Often we only can make partial conclusions: decision makers usually only have part of the info they need to choose unambiguously best actions. It is better to report ranges than point estimates and avoid maintaining only one hypothesis rather than offering predictions under the range of plausible hypotheses that are consistent with the available evidence. We should try to discourage the Johnson syndrome: 'Ranges are for cattle. Give me a number.'

3.2 Conditional Prediction

The joint probability (frequency) distribution of $(y, x \in Y \times X)$ across population is $P(y, x)$. A person is drawn at random from the subpopulation of people with a specified value of x. The problem is to predict his value of y. $P(y|x)$ can be interpreted as:

1. the distribution of y conditional on x, viewed as a function of x;
2. the distribution of y conditional on x, evaluated at a specified value of x;

3. the probability that y takes a given value conditional on x viewed as a function of x;

4. the probability that y takes a given value conditional on x evaluated at a specified value of x.

The particular interpretation will be clear from the context and sometimes I may write $P(y)$.

Whenever we can observe (y, x) at random from the population of interest, we might ask how we can learn about the conditional distribution $P(y|x)$ or at least the value of a best predictor of y given x. Even if we assume nothing about the form of the distribution $P(y, x)$, random sampling will reveal $P(y, x)$. For studying conditional prediction, we will now look at empirical distributions and illustrate use of some of the above concepts, in addition to the analogy principle, which loosely implies using sample statistics for population counterparts and then calling on results from asymptotic theory to justify these sample statistics. The empirical distribution $P_N(y, x)$ is the sample analog and natural estimate of $P(y, x)$. It is a multinomial distribution placing equal mass $\frac{1}{N}$ on each of N observations $[(y_i, x_i) : i = 1, \ldots, N]$; if a particular value of (y, x) recurs in the data, it receives multiple $\frac{1}{N}$ weights. It is natural to use $P_N(y, x)$ to estimate $P(y, x)$ since the empirical distribution estimates the probability $P[(y, x) \in A]$ that $E(y, x)$ falls in some set A by estimating the fraction of observations of (y, x) that fall in the set A:

$$P_N[(y, x) \in A] = \frac{1}{N} \sum_{i=1}^{N} 1[(y_i, x_i) \in A] \xrightarrow{\text{as}} P[(y, x) \in A]$$

where the convergence follows by the SLLN; more specifically, it can be shown that (ICBST) the empirical distribution for this probability converges almost surely to $E[1[(y, x) \in A]]$, which turns out to be $P[(y, x) \in A]$; the proof of the second part of this statement is as follows:

Proof.

$$\begin{aligned} E[1[(y, x) \in A]] &= 1.P[(y, x) \in A] + 0.P[(y, x) \notin A] \\ &= P[(y, x) \in A] \end{aligned}$$ □

This essentially means that we can interpret probability as the expectation of an indicator function. Since with random sampling, we can learn $P[(y, x) \in A]$ even if we knew nothing before about it's value and this holds for every set A, we can learn the distribution $P(y, x)$. Let us now look at the three cases. The lesson from this section will be that we can only do non-parametric estimation on the support.

1. x_0 is on the support of $P(x)$ and $P(x = x_0) > 0$.

2. $P(x = x_0) = 0$ but x_0 is on the support.

3. x_0 is off the support of $P(x)$.

In the first case where covariates have positive probability, i.e. when $P(x = x_0) > 0$, we have that the conditional empirical probability is given by:

$$\begin{aligned} P_N(y \in B | x = x_0) &= \frac{\sum_{i=1}^N 1[y_i \in B, x_i = x_0]}{\sum_{i=1}^N 1[x_i = x_0]} \\ &= \frac{\frac{1}{N}\sum_{i=1}^N 1[y_i \in B, x_i = x_0]}{\frac{1}{N}\sum_{i=1}^N 1[x_i = x_0]} \end{aligned} \tag{3.1}$$

Observe that the numerator converges almost surely to $P(y \in B, x = x_0)$ by SLLN and the denominator converges almost surely to $P(x = x_0)$, which is positive in this case. So, by the Contraction Mapping Theorem and Bayes Theorem, the RHS of (3.1) converges almost surely to $P(y \in B | x = x_0)$:

$$\frac{P(y \in B, x = x_0)}{P(x = x_0)} \overset{\text{Bayes}}{=} P(y \in B | x = x_0)$$

which holds for every set B so we can learn about the conditional distribution $P(y|x = x_0)$. So, (sample) empirical quantiles (e.g. mean and median) converge to population quantiles. Also remember that for the SLLN and CMT, we need functions to be continuous.

Example 3.14.

$$E_N(y|x = x_0) = \frac{\sum_{i=1}^N y_i \cdot 1[x_i = x_0]}{\sum_{i=1}^N 1[x_i = x_0]} = \frac{\frac{1}{N}\sum_{i=1}^N y_i \cdot 1[x_i = x_0]}{\frac{1}{N}\sum_{i=1}^N 1[x_i = x_0]} \tag{3.2}$$

The numerator of the RHS in (3.2) converges almost surely to $E(y \cdot 1[x = x_0])$ as N increases by SLLN, which equals $E(y|x = x_0)P(x = x_0)$ and the denominator converges to $P(x = x_0)$. Given that $P(x = x_0) > 0$, by the CMT:

$$E_N(y|x = x_0) \xrightarrow{\text{a.s.}} E(y|x = x_0)$$

Similarly, $M_N \xrightarrow{\text{a.s.}} M$.

In the second case where covariates have zero probability, i.e. when $P(x = x_0) = 0$ but x_0 on the support of $P(x)$ (e.g. continuous distributions). Let $\rho(x_i, x_0)$ measure distance between covariate of interest x_0 and an observed value x_i. When x is scalar, $\rho(x_i, x_0) = |x_i - x_0|$. When x vector, it could be any reasonable measure of distance between x_i and x_0, e.g. Euclidean distance between these vectors. Let d_N denote *bandwidth*. The subscript N on d_N indicates that bandwidth is a function of sample size. Estimate $E(y|x = x_0)$ by the sample mean of y among the observations for which $\rho(x_i, x_0) < d_N$.

Definition 3.15. The *local average* or *uniform kernel estimate* is:

$$\theta_N(x_0, d_N) \equiv E_N(y|x = x_0) = \frac{\sum_{i=1}^N y_i \cdot 1[\rho(x_i, x_0) < d_N]}{\sum_{i=1}^N 1[\rho(x_i, x_0) < d_N]}$$

where d_N is a sample-size dependent *bandwidth* selected by the researcher conveying the idea of restricting attention to observations where x_i is near x_0; sometimes d_N is written as $d_N(x_0)$ and called the *local bandwidth selection* to emphasise the case where we do not use the same bandwidth everywhere.

A basic finding of modern nonparametric regression analysis is that the uniform kernel estimate $E_N(y|x = x_0) \xrightarrow{\text{a.s.}} E[y|x = x_0]$ provided the following four conditions hold:

1. $E(y|x)$ varies continuously with x near x_0.
2. $V(y|x)$ bounded for x near x_0.
3. Tighten bandwidth d_N as sample size N increases.
4. Do not tighten bandwidth d_N too rapidly as sample size N increases.

Conditions (i) and (ii) are minimum regularity conditions; we can always choose bandwidths so (iii) and (iv) hold.

Remark 3.16 (Curse of Dimensionality)**.** We should choose the bandwidth d_N so MSE tends to zero, i.e. variance and bias tend to zero. Large d_N is good for variance in that we get a big sample and so standard deviation reduces at rate $\sqrt{n}$, but then the bias could be large, i.e. $E(y|\rho(x, x_0) < d_N) - E_N(y|x = x_0)$ could be large when d_N is large. When d_N is small, however, variance increases because there are fewer observations in each cell. The *curse of dimensionality* rears its head in that a larger dimension for x does not affect bias but it does affect variance because there tends to be fewer observations lying inside bandwidths of radius d_N, so variance is still high. With non-parametric estimation, the price to pay is generally that the rate of convergence will be slower than $\sqrt{n}$. Stone (1981) showed that the best rate of convergence you can achieve from non-parametric estimation gets tougher as the dimension of x increases. So the curse of dimensionality takes the following form: the best achievable rate of convergence diminishes as the dimension of covariates increases. Of course, one solution would be to look at semi-parametric models such as linear index models, e.g. the regression of $E(y|x) = g(x)$ where we only know that g is continuous; say x is in a three dimension space, then $g(x) = g(x_1, x_2, x_3) = g'(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)$ for a linear index model where $g \neq g'$, i.e. linear index models wipe out the curse of dimensionality; note that semi-parametric models make assumptions about the regression functions: g is parametric while g' is non-parametric.

In the third case, x_0 is off the support of $P(x)$, i.e.

$$\exists d_0 > 0 : P[\rho(x_i, x_0) < d_0] = 0$$

e.g. when $x_0 = 5$ in figure 3.1. One can only do non-parametric estimation on the support. Now, data alone reveal nothing about $P(y|x = x_0)$. So, we are in the case of *extrapolation*: predicting y when x_0 is off the support, i.e. making

predictions away from data. We need *global* assumptions for identifying power in the case of extrapolation, i.e. we need to restrict $E(\cdot|\cdot)$ globally. We can invoke *invariance* assumptions, such that y behaves the same at x_0 as at x_1 on the support of $P(x)$, i.e.

$$P(y|x = x_0) = P(y|x = x_1)$$

Definition 3.17. Using a *counterfactual* entails expressing what has not happened but what might or would happen if circumstances, i.e. data were different.

Example 3.18. What would the consequences be for the US had the Paulson plan not been enacted? What would the consequences have been for Europe had we not decided to bail out the banks? Surely we would be better off? What about if Senator John McCain had been president instead of Barack Obama? These are hypothetical situations. We can analyse these circumstances, e.g. in DSGE models, but do not have data on them. This has much to do with the inherent problem in economics that experiments are not as readily available as in the natural sciences.

Example 3.19 (Predicting Criminality)**.** Selective incapacitation implies that sentencing of convicts should be linked with predictions regarding their future criminality. The RAND study by Greenwood & Abrahamse (1982) found that using a sample of 2200 prison and jail inmates in 1978 across California, Michigan and Texas, those with backgrounds such as previous convictions, drug use and unemployment predict high rates of future offenses and part of their research team then suggested that those with such backgrounds should receive longer prison terms. This was very controversial, especially when this prediction approach became part of a legislative proposal for selective incapacitation. The part of the controversy that relates to econometrics concerns the external validity from the RAND results to other groups, places and sentencing policies. The findings hold for this particular cohort of prisoners in these three states for the given sentencing policies, but that does not imply that they would still apply to other cohorts of prisoners in other states or to criminals who would be sentenced under alternative policies such as selective incapacitation.

Remark 3.20. Sometimes, hopefully rarely, researchers misinterpret correlation with causation. However, if treatments are randomly assigned, then causation becomes more justifiable. For instance, if there were no selection issues and people were truly randomly assigned to different treatments, then if data showed that outcomes under one treatment were 'better' than outcomes under another treatment, we could conclude that there is a causal link: the first treatment improves the outcome relative to the second.

Failures off the support are inherently not detectable. Theory becomes important. The first function of theory is to allow extrapolation and the second function function of theory is to improve the sampling precision for estimation

on the support. Causal interpretation can be 'dodgy', while prediction is usually better. Unfortunately, cases where theory is most testable are generally least needed – learning $P(y|x)$ on the support of $P(x)$, while cases where theory is least testable are generally most needed – learning $P(y|x)$ off the support of $P(x)$.

Returning to the local average estimate, a more general kernel estimator is the following.

Definition 3.21. The *local weighted average* or *kernel estimate* is

$$E_N(y|x=x_0) = \frac{\frac{1}{N}\sum_{i=1}^{N} y_i K\left[\frac{\rho(x_i,x_0)}{d_N}\right]}{\frac{1}{N}\sum_{i=1}^{N} K\left[\frac{\rho(x_i,x_0)}{d_N}\right]}$$

where $0 \leq K(\cdot)$ is inversely related to $\rho(x_i, x_0)/d_N$. The uniform kernel can be seen to be a special case where

$$1\left[\frac{\rho(x_i,x_0)}{d_N} < 1\right]$$

and here all observations get used and are given the same weight, hence the name 'uniform' kernel.

Remark 3.22. Choosing bandwidth d_N can be extremely subjective. Best practice is to report multiple estimates or use data dependent automated rules to choose the bandwidth, e.g. cross-validation. Cross-validation involves fixing the bandwidth, estimating the regression on each of the N possible subsamples (each of size $N-1$ and then in each case we use the estimate to predict $y|x$ for the observation that was left out. The resulting bandwidth, the cross-validated bandwidth yields the best predictions of the left-out values of y. Increasing the bandwidth typically reduces variance but increases bias, which is not good for the MSE – this is another manifestation of the curse of dimensionality as N increases.

A (more) thorough approach to prediction often stresses the specification of a loss function $L(\cdot)$ that we want to minimise. Letting p be predictor of the random variable Y and X be other random variables, we usually want to minimise the expected loss conditional on the random variates X:

$$\min E[\mathcal{L}(y-p)|x]$$

The *best predictor* solves this minimisation problem, so choosing a best predictor is a decision problem whose solution depends on the objective, i.e. the best predictor is determined by $L(.)$ and $P(y|x)$. It can be shown that when the loss function is a square loss function, the best predictor is the mean. In this case, the solution may not exist, however. It can also be shown that with absolute loss functions, the best predictor is the median. In what follows, let $u \equiv y - p$.

Lemma 3.23. *Under square loss, the best predictor is the mean.*

Proof. Let $L(\cdot)$ be the square loss function, i.e. $L(u) = u^2$, $\mu = E(y)$ and $\mu \neq \theta \in \mathbb{R}$. Then

$$\begin{aligned} E(y-\theta)^2 &= E[(y-\mu)+(\mu-\theta)]^2 \\ &= E(y-\mu)^2 + (\mu-\theta)^2 + 2(\mu-\theta)E(y-\mu) \\ &= E(y-\mu)^2 + (\mu-\theta)^2 > E(y-\mu)^2 \end{aligned}$$

Therefore, μ uniquely minimises the expected loss. □

Lemma 3.24. *Under absolute loss, the best predictor is the median.*

Proof. Let $L(\cdot)$ be the absolute loss function, i.e. $L(u) = |u|$ and $m \equiv \min\{\theta : P(y \leq \theta) \geq \frac{1}{2}\}$ be the median of y where $m \in \mathbb{R}$. Let us first compare the expected loss at m with that at any $\theta < m$:

$$\begin{aligned} E[|y-\theta|] - E[|y-m|] &= E[|y-\theta| - |y-m|] \\ &= (\theta-m)P(y \leq \theta) \\ &\quad + E[2y-(\theta+m)|\theta < y < m]P(\theta < y < m) \\ &\quad + (m-\theta)P(y \geq m) \\ &\geq (\theta-m)P(y \leq \theta) + (\theta-m)P(\theta < y < m) \\ &\quad + (m-\theta)P(y \geq m) \\ &= -(m-\theta)P(y < m) + (m-\theta)P(y \geq m) \\ &= (m-\theta)[P(y \geq m) - P(y < m)] \end{aligned}$$

Since by definition $P(y < m) \leq \frac{1}{2}$, the final expression is nonnegative. Finally, let us compare the expected loss at m with that at any $\theta > m$:

$$\begin{aligned} E[|y-\theta|] - E[|y-m|] &= E[|y-\theta| - |y-m|] \\ &= (\theta-m)P(y \leq m) \\ &\quad + E[(\theta+m)-2y|m < y < \theta]P(m < y < \theta) \\ &\quad + (m-\theta)P(y \geq \theta) \\ &\geq (\theta-m)P(y \leq m) + (m-\theta)P(m < y < \theta) \\ &\quad + (m-\theta)P(y \geq \theta) \\ &= (\theta-m)P(y \leq m) - (\theta-m)P(m < y) \\ &= (\theta-m)[P(y \leq m) - P(m < y)] \end{aligned}$$

Since by definition $P(y \leq m) \geq \frac{1}{2}$, the final expression is nonnegative. □

One advantage of absolute loss functions is that the solution always exists; least absolute deviations (LAD) estimation is necessary for absolute loss functions and its asymptotic theory is complicated due to non-differentiability.

It turns out that best predictors associated with convex, symmetric loss functions coincide; the mean is the median when $P(y|x)$ is symmetric; furthermore, $P(y|x) = P(y|h(x))$ when h is injective. $L(u) = u^2$ and $L(u) = |u|$ treat under- and overpredictions of loss symmetrically. However, we don't simply have to look at symmetric loss functions such as $L(u) = |u|$.

Definition 3.25. The *asymmetric α-absolute loss function* is defined as

$$L(u) = \begin{cases} (1-\alpha)|u| & \text{if } u \leq 0 \\ \alpha|u| & \text{if } u \geq 0 \end{cases}$$

An equivalent way of writing this loss function is

$$L(u) = \alpha|y-p|1[y-p>0] + (1-\alpha)|y-p|1[y-p<0]$$

Example 3.26 (GRE scores)**.** Let y denote GRE scores and x be some covariate. Say $\alpha = 0.96$, i.e. the 96^{th} percentile, so $P(\text{GRE score} \leq t|x) = .96$; for example, $t = 790$. Best predictor under this loss function is $Q_\alpha(y|x) = \inf\{t : P(y \leq t|x) \geq \alpha\}$.

The asymmetric α-absolute loss function leads to quantile regression (Koencker & Basset, 1979):

$$Q_\alpha(y|x) = \inf\{t : P(y \leq t|x) \geq \alpha\}$$

Definition 3.27. The *asymmetric α-square loss function* is defined as

$$L(u) = \alpha(y-p)^2 1[y-p>0] + (1-\alpha)(y-p)^2 1[y-p<0]$$

and leads to expectiles.

Example 3.28. See Manski (2007) [59] section 1.5 on predicting high school graduation.

3.3 Incomplete Data

Let (y, z, x) be such that y is an outcome to be predicted, x are covariates and define

$$z = \begin{cases} 1 & \text{if } y \text{ is observed} \\ 0 & \text{else} \end{cases}$$

Draw N people at random from population. For each $i = 1, \ldots, N$, the outcome y_i is observable if $z_i = 1$ and missing if $z_i = 0$. The objective is to use available data to learn about $P(y|x)$ at a specified value of x on $Supp(P(x))$. We can use the LTP to express the missing data problem more clearly:

$$P(y|x) = P(y|x, z=1)P(z=1|x) + \underbrace{P(y|x, z=0)}_{\text{missing}} P(z=0|x)$$

The missing $P(y|x, z = 0)$, which is unknown implies that we are dealing with an identification problem. Denote $P(y|x, z = 0) = \gamma \in \Gamma_Y$, the identification region is:

$$H[P(y|x)] = [P(y|x, z = 1)P(z = 1|x) + \gamma P(z = 0|x), \gamma \in \Gamma_Y]$$

and note that

$$P(z = 0|x) < 1 \Longrightarrow H[\cdot] \subsetneq \Gamma_Y$$

where Γ_Y denotes the set of all probability distributions on the set Y. H is a proper subset of Γ_Y, i.e. $H \subsetneq \Gamma_Y$ when $P(z = 0|x) < 1$ and is the single distribution $P(y|x, z = 1)$ when $P(z = 0|x) = 0$. Therefore, $P(y|x)$ is *partially identified* when $0 < P(z = 0|x) < 1$ and is *point identified* when $P(z = 0|x) = 0$.

Empirical research often has an objective of inferring a parameter of the outcome distribution, e.g. $E(y|x)$. Let $\theta(\cdot)$ be a function mapping probability distributions on Y into $\mathbb{R}$ and consider the parameter $\theta[P(y|x)]$. The identification region for this parameter is the set of all values it may take when $P(y|x)$ varies over all of its feasible values, so $H\{\theta[P(y|x)]\} = \{\theta(\eta), \eta \in H[P(y|x)]\}$.

Now looking at the identification of event probabilities, we can once again use the LTP to express the missing data problem more clearly:[3]

$$P(y \in B|x) \overset{\text{LTP}}{=} P(y \in B|x, z = 1)P(z = 1|x) + \underbrace{P(y \in B|x, z = 0)}_{\in[0,1]} P(z = 0|x)$$

The worst case bound on $P(y \in B|x)$:

$$\begin{aligned} P(y \in B|x, z = 1)P(z = 1|x) &\leq P(y \in B|x) \\ &\leq P(y \in B|x, z = 1)P(z = 1|x) + P(z = 0|x) \end{aligned} \tag{3.3}$$

Equivalently, letting L_B and U_B denote lower and upper bounds, respectively, we have that

$$\begin{aligned} L_B: &\quad P(y \in B|x, z = 1)P(z = 1|x) \\ U_B: &\quad P(y \in B|x, z = 1)P(z = 1|x) + P(z = 0|x) \end{aligned}$$

U_B and L_B are the largest and smallest feasible values of $P(y \in B|x)$ and hence are called *sharp* bounds:

$$H[P(y \in B|x)] = [P(y \in B|x, z = 1)P(z = 1|x), P(y \in B|x, z = 1)P(z = 1|x)+P(z = 0|x)]$$

Observe that the width of the identification interval is given by $U_B - L_B = P(z = 0|x)$; hence, data is informative unless y is always missing. Note that the width of interval may vary with x but it does not vary with the set B.

[3] We will not have time to identify quantiles during this course. For those interested, see Manski (2007:39-40)[59].

Because we are looking at event probabilities, it does not matter whether y has a bounded or an unbounded support: $B = [-\infty, t]$, $P(y \in B) = P(y \leq t|x)$.

$$\begin{aligned} P(y \leq t|x, z = 1)P(z = 1|x) &\leq P(y \leq t|x) \\ &\leq P(y \leq t|x, z = 1)P(z = 1|x) + P(z = 0|x) \end{aligned}$$

Example 3.29 (Bounding Probability of Exiting Homelessness)**.** Piliavin & Sosin (1988) wanted to know the probability of homelessness six months after already being homeless. Let $y = 1$ if the individual has a home six months later and $y = 0$ denote the opposite; x are background attributes. The goal is learn $P(y = 1|x)$ and the missing data problem arises due to not being able to locate part of the original sample six months later. Let $x = sex$. 106 men were sampled originally and 64 of these men were found six months later, 21 of which were no longer homeless. Therefore, the empirical probability estimate of $P(y = 1|male, z = 1) = \frac{21}{64}$ and that of $P(z = 1|male) = \frac{64}{106}$. So, the estimate of the bound on $P(y = 1|male)$ is $[\frac{21}{106}, \frac{63}{106}] \approx [0.20, 0.59]$. 31 women were sampled originally and 14 of these women were found six months later, 3 of which were no longer homeless. Therefore, the empirical probability estimate of $P(y = 1|female, z = 1) = \frac{3}{14}$ and that of $P(z = 1|female) = \frac{14}{31}$. So, the estimate of the bound on $P(y = 1|female)$ is $[\frac{3}{31}, \frac{20}{31}] \approx [0.10, 0.65]$.

With small sample sizes, interpretation of these estimates should be cautious. In fact, we may not actually have tighter bounds on men because we have a smaller sample of women; this is an inferential problem. The attrition rates (equivalently the bounds) for men and women are 0.39 and 0.55, respectively. Importantly, the bounds are informative. Even though no assumptions were placed on the attrition process, we can place meaningful bounds on attrition rates. We will ignore sampling variability while studying identification but return to it for later chapters.

There has been a history of researchers focusing on point identification rather than reporting bounds. The practice of reporting simple bounds did not catch on. Even when they are wide, bounds are useful for two reasons: (i) they establish a domain of consensus and (ii) they highlight the role of credible assumptions leading to tighter findings.

More generally, letting g be a function, LIE yields

$$E[g(y)|x] \overset{\text{LIE}}{=} E[g(y)|x, z = 1]P[z = 1|x] + E[g(y)|x, z = 0]P(z = 0|x)$$

Assume g is bounded and let

$$\begin{aligned} g_0 &= \inf_{y \in Supp(Y)} g(y) \\ g_1 &= \sup_{y \in Supp(Y)} g(y) \end{aligned}$$

$g(y)$ will be bounded as long as $P(z = 0|x) > 0$. If $g_1 = \infty$ or $g_0 = -\infty$, i.e. the case of unboundedness, then we need more assumptions for inference.

The width of the interval will be $(g_1 - g_0)P(z = 0|x)$ and

$$H[E[g(y)|x]] = [E(g(y)|x, z = 1)P(z = 1|x) + g_0 P(z = 0|Zx), \\ E(g(y)|x, z = 1)P(z = 1|x) + g_1 P(z = 0|x)] \tag{3.4}$$

which is a proper subset of $[g_0, g_1]$ when $P(z = 0|x) < 1$, so it is informative. The severity of missing data is directly proportional to $P(z = 0|x)$.

Example 3.30. As an application of (3.4), let $B \subset Y$ and $g(y) = 1[y \in B]$. Then $g_0 = 0, g_1 = 1, E[g(y)|x] = P(y \in B|x)$ and $E[g(y)|x, z = 1] = P(y \in B|x, z = 1)$. So, (3.4) is an alternative version of the bound given by (3.3) on$P(y \in B|x)$.

Remark 3.31. When $g(\cdot)$ is unbounded from above or below, (3.4) still holds but has different implications when $P(z = 0|x) > 0$. The lower bound on $E[g(y)|x]$ is $-\infty$ if $g_0 = -\infty$ and ∞ if $g_1 = \infty$. The identification region has infinite width but is still informative if $g(\cdot)$ is bounded from at least one side. As Manski put it, 'the presence of missing data makes credible assumptions a prerequisite for inference on the mean of an unbounded random variable.' (2007: 44) [59]

An important concept you may come across in microeconomics and econometrics is that of *stochastic dominance*. You may have learned about first order and second order stochastic dominance.

Definition 3.32. Distribution Q *stochastically dominates* distribution Q' if $Q(y \leq t) \leq Q'(y \leq t)\ \forall t$

So one distribution tends to yield larger outcomes than the other. Roughly speaking, if $P(y|x = x_0)$ stochastically dominates $P(y|x = x_1)$, then

$$P(y \leq t|x = x_0) \leq P(y \leq t|x = x_1)\ \forall t \\ < P(y \leq t|x = x_1) \text{ some } t$$

When Γ_R is the space of all distributions on the real line, a real valued parameter $D(\cdot) : \Gamma_R \longrightarrow R$ is said to *respect stochastic dominance* if $D(Q) \geq D(Q')$ whenever Q stochastically dominates Q'. Equivalently, parameters respecting stochastic dominance, say $D(y|x)$ where $D(y|x = x_0)$ stochastically dominates $D(y|x = x_1)$ have the property that for all monotone increasing functions f:

$$E[f(y)|x = x_0] \geq E[f(y)|x = x_1]$$

In fact, every quantile will be 'higher'. For example:

$$M(y|x = x_0) > M(y|x = x_1)$$

So, when Q stochastically dominates Q', we have that $D(Q) \geq D(Q')$ where D is a quantile such as the mean of an increasing function of y. Examples of parameters that respect stochastic dominance are quantiles and means of

increasing functions of y; spread parameters e.g. variance / interquartile range do not respect stochastic dominance. To get identification, we need to look at what happens between the highest and lowest values. Let us investigate how to estimate the bound on $D(P(y))$ via stochastic dominance. Let y_0 and y_1 be the smallest and largest logically possible values for y. The minimum (maximum) value of $D(\cdot)$ that respects stochastic dominance is obtained by presuming that any missing value of y is equal to y_0 (y_1).

Let us now look at some distributional assumptions. We want to learn the identifying power of all distributional assumptions so we can characterise the entire spectrum of inferential possibilities. The first and most common though generally implausible assumption is the *missing at random* or *conditional statistical independence* assumption.

Definition 3.33. The *missing at random* (MAR) or *conditional statistical independence* assumption is:

$$\begin{aligned} &P(y|x, z = 1) = P(y|x, z = 0) = P(y|x) \\ \implies &E[y|x, z = 1] = E[y|x, z = 0] = E[y|x] \end{aligned}$$

The missing at random assumption is an example of a *non refutable* assumption – any assumption about $P(y|x, z = 0)$ is not refutable.

Definition 3.34. 'Any assumption that directly restricts the distribution $P(y|x, z = 0)$ of missing data is *nonrefutable*.' (Manski, 2007: 46; my emphasis) [59]

To see how refutable assumptions may be tested statistically, consider the assumption $E[g(y)|x] \in R_1 \subset \mathbb{R}$. We can reject the this hypothesis if $H_N\{E[g(y)|x)]\}$ is sufficiently far from R_1.

MAR implies that the following identification region contains one element, $P(y|x, z = 1)$:

$$H_0[P(y|x)] \equiv [P(y|x, z = 1)P(z = 1|x) + \gamma P(z = 0|x), \gamma \in \Gamma_{0Y}]$$

Assumptions placed on $P(y|x, z = 0) \in \Gamma_{0Y}$ are not empirically testable (they are nonrefutable though you may be able to argue why it is not plausible for $P(y|x, z = 0)$ to lie in Γ_Y) versus $P(y|x) \in \Gamma_Y$ is better since it can be empirically tested. Interesting assumptions include those positing that $P(y|x)$ lies in a specific set of distributions, Γ_{1Y}. Data alone imply that $P(y|x) \subset H[P(y|x)]$. Combining the data with the assumption $P(y|x) \subset \Gamma_{1Y}$:

$$H_1[P(y|x)] \equiv H[P(y|x)] \cap \Gamma_{1Y}$$

If $\cap\Gamma_{1Y}$ is zero, then $P(y|x)$ cannot lie in Γ_{1Y} and so we have made the wrong assumption, but if $\cap\Gamma_{1Y}$ is nonzero, this does not imply that we accept but rather that we do not reject. Data may be refuted or may 'not be refuted'. When we use the term 'nonrefutable', we ask whether there could be an ex-ante probability that the intersection is the null set – if yes, then our assumption is refutable, else it is nonrefutable. Refutability concerns logic and is a property of

assumptions and data whereas and credibility is a property of assumptions and the researcher, so there is an element of subjectivity involved with credibility.

With no assumption on $P(y|x)$, $P(y|x) \in H[P(y|x)]$. Now assume that $P(y|x) \in \Gamma_Y$. After making this assumption, $P(y|x) \in H[P(y|x)] \cap \Gamma_Y = H_1[P(y|x)]$. If $H_1[P(y|x)]$ is a strict subset of $H[P(y|x)]$, then the assumption is said to have *identifying power*. If $H_1[P(y|x)] = 0$, then the assumption is refutable. If $H_1[P(y|x)] \neq 0$, then the assumption is non refutable; however, this does not mean that the assumption is true!

The sample analogue of ID regions are denoted H_N, where the subscript N emphasises the dependence on the sample size N. As for confidence sets, let C is correspondence taking ψ to $\mathbb{R}$. It is important to understand the correct way to interpret confidence sets. An α-confidence set does not mean that $\theta \in C(\psi)$ for a given set $C(\psi)$ with $\alpha\%$ confidence, but that $\alpha\%$ of such sets will contain the true θ. Generally, to construct a confidence set, we start with a consistent estimate of θ, e.g. $\theta_N(\psi)$ and then construct an interval:

$$[\theta_N(\psi) - \delta_{0N}(\psi), \theta_N(\psi) + \delta_{1N}(\psi)]$$

where $\delta_{0N}(\psi) > 0$ and $\delta_{1N}(\psi) > 0$ are such that

$$P\{\psi : \theta \in [\theta_N(\psi) - \delta_{0N}(\psi), \theta_N(\psi) + \delta_{1N}(\psi)]\} \overset{N \longrightarrow \infty}{\longrightarrow} \alpha$$

What about confidence sets for identification regions? Denote the identification region for θ by $H(\theta)$. Then $C(\cdot)$ gives an $\alpha-$confidence set for $H(\theta)$ provided that $P[\psi : H(\theta) \subset C(\psi)] = \alpha$. Note that an α-confidence for $H(\theta)$ contains θ with a probability at least as large as α, i.e. $P[\psi : \theta \in C(\psi)] \geq P[\psi : H(\theta) \subset C(\psi)]$.[4]

Example 3.35. A confidence interval for the identification region for $E[g(y)|x]$ where $g(\cdot)$ is a bounded function can be given by

$$\begin{aligned} C(\psi) = \big[&E_N[g(y)|x, z = 1]P_N(z = 1|x) \\ &+ g_0 P_N(z = 0|x) - \delta_{0N}(\psi), \\ &E_N[g(y)|x, z = 1]P_N(z = 1|x) \\ &+ g_1 P_N(z = 0|x) + \delta_{1N}(\psi)\big] \end{aligned}$$

Horowitz and Manski (2000) [48] demonstrate how to choose $\delta_{0N}(\psi)$ and $\delta_{1N}(\psi)$ so that the coverage probability of $H\{E[g(y)|x]\}$ converges to α as N increases, while Imbens and Manski (2004) [49] show how to choose $\delta_{0N}(\psi)$ and $\delta_{1N}(\psi)$ so that the coverage probability of $E[g(y)|x]$ converges to α as N increases. Note that in both cases, $\delta_{0N}(\psi)$ and $\delta_{1N}(\psi)$ both converge to zero as N gets larger. So, the confidence set shrinks towards the identification region for $E[g(y)|x]$ asymptotically.

[4] Beyond the scope of the course, Manski (2007:60-1) [59] has a nice section on convergence of sets to sets using the Hausdorff distance for those interested.

Typically, each realization of y is observed either completely or not at all. However, y may be observed to lie in proper but nonunitary subsets of the outcome space Y. Sampling with missing outcomes is a special case of interval measurement. A common source of interval data is measurement devices with bounded ranges of sensitivity.

Up to now, we know the mixture of the distributions, i.e. $P(z = 1|x)$ and $P(z = 0|x)$ are known. What about when we have an unknown mixture of a known $P(y = 1|x, z = 1)$ and an unknown $P(y = 1|x, z = 0)$? The joint missingness of (y, x) exacerbates the identification problem produced by missingness of y alone.

Example 3.36 (Bounding the probability of employment and unemployment rate)**.** See Manski (2007:58-60) [59].

Let us now examine instrumental variables (IV) and describe different distributional assumptions. Expand the distribution of interest (y, z, x) to (y, z, x, v) where $v \in V$ is an observable covariate that may be totally different to (x, z), that may partially overlap with (x, z) or that may be identical to (x, z). The term 'instrumental variable' due to Reiersol (1945) who used it to help identify linear simultaneous equation systems can be defined as follows.[5]

Definition 3.37. v is an *instrumental variable* 'if one poses an assumption that somehow connects the conditional distributions $P(y|v)$ across different values of v.' (Manski, 2007: 63) [59]

v is only useful when we combine observations of v with an assumption that has identifying power. Manski (2007) [59] argues that the discussion of whether a covariate is a 'valid instrument' is imprecise since it neglects to discuss the assumption that accompanies v. He argues that a more precise question would involve querying whether an assumption using an instrumental variable is credible or not.

Let $P(y, z, \alpha, w)$ be such that y denotes outcomes, z denotes whether y_i is observed and α and w are covariates.

The *missing at random* assumption described above

$$P(y|x, z = 0) = P(y|x, z = 1)$$

is an example of a type one assumption and uses IV $w = z$. Sometimes, researchers are uncomfortable about assuming MAR, which uses an instrumental variable ($v = z$), so they sometimes make an alternative assumption, *viz.*:

$$P(y|x, w, z = 0) = P(y|x, w, z = 1) \tag{3.5}$$

Lemma 3.38. *Assumption (3.5) is nonrefutable and point identifies $P(y|x)$.*

[5] Goldberger (1972) claims that this use of instrumental variables dates back at least as far as Wright (1928).

Proof. Assumption (3.5) is nonrefutable precisely because it restricts the distribution of missing data $P(y|x, w, z = 0)$. To see that it is point-identifying, observe that:

$$P(y|x) \overset{\text{LTP}}{=} \sum_{k \in W} P(y|x, w = k)P(w = k|x)$$

further observe that assumption (3.5) implies $P(y|x, w = k) = P(y|x, w = k, z = 1)\ \forall k \in W$.

$$\therefore P(y|x) = \sum_{k \in W} P(y|x, w = k, z = 1)P(w = k|x)$$

The RHS of this equation is revealed (asymptotically) by the sampling process. Therefore, $P(y|x)$ is point-identified. □

Remark 3.39. Note that assumption (3.5) is often written alternatively as

$$P(z = 1|x, w, y) = P(z = 1|x, w) \tag{3.6}$$

Bayes Theorem shows that this is equivalent to assumption (3.5). While proved analogously for the case where y is continuous with the density for y in place of the probability masses, we can see for y discrete by applying Bayes Theorem to the LHS of (3.6) that

$$\frac{P(y = j|x, w, z = 1)P(z = 1|x, w)}{P(y = j|x, w)} = P(z = 1|x, w)$$

for each $j \in Y$ and solving this equation produces $P(y = j|x, w, z = 1) = P(y = j|x, w)$, which is the same as (3.5).

There is no justification for why some researchers believe that MAR conditional on (x, w) is more credible than MAR conditional solely on x – 'controls for' is too vague and lacks theoretical justification in formal probability theory. Sometimes it may be true that outcomes may be MAR conditional on x but not on (x, w) or vice-versa.

Definition 3.40. An example of a type two assumption is *statistical independence*:

$$P(y|u, v) = P(y|u)$$

What is the benefit of using this assumption? Look at identification regions. The identification region for $P(y|u, v = k)$ using the data alone is given by

$$\begin{aligned} H[P(y|u, v = k)] = \big\{ & P(y|u, v = k, z = 1)P(z = 1|u, v = k) \\ & + \gamma_k \cdot P(z = 0|u, v = k), \gamma_k \in \Gamma_Y \big\} \end{aligned}$$

Since the statistical independence assumption states that $P(y|u) = P(y|u, v = k)\ \forall k \in V$, $P(y|u)$ must lie within each identification region $H[P(y|u, v = k)], k \in V$. Furthermore, any particular distribution that lies within all of

these k-specific regions will be a feasible value for $P(y|u)$ and hence with the assumption the identification region for $P(y|u)$ is the intersection, *viz.*:

$$H_1[P(y|u)] = \cap_{k \in V} \big\{ P(y|u, v = k, z = 1)P(z = 1|u, v = k) \\ + \gamma_k \cdot P(z = 0|u, v = k), \gamma_k \in \Gamma_Y \big\}$$

This assumption that $y \perp\!\!\!\perp v|u$ may be refutable since the intersection may be empty. Note that in the binary outcomes case, it turns out that[6]

$$H_1[P(y = 1|u)] \tag{3.7}$$
$$= \big[\max_{k \in V} P(y = 1|u, v = k, z = 1)P(z = 1|u, v = k), \tag{3.8}$$
$$\min_{k \in V} P(y = 1|u, v = k, z = 1)P(z = 1|u, v = k) \tag{3.9}$$
$$+ P(z = 0|u, v = k) \big] \tag{3.10}$$

For binary y, without the assumption:

$$H[P(y = 1|u, v = k)] = [P(y = 1|u, v = k, z = 1)P(z = 1|u, v = k), \\ P(y = 1|u, v = k, z = 1)P(z = 1|u, v = k) + P(z = 0|u, v = k)]$$

With the statistical independence assumption $P(y = 1|u) = P(y = 1|u, v = k)\ \forall k$:

$$H_1[P(y = 1|x)] = \Big[\max_{k \in W} P(y = 1|x, w = k, z = 1)P(z = 1|x, w = k), \\ \min_{k \in W} P(y = 1|x, w = k, z = 1)P(z = 1|x, w = k) + P(z = 0|x, w = k) \Big]$$

Parametric assumptions are weaker than distributional assumptions, so they may be more credible.

Let us now look at different assumptions on means.

Definition 3.41. The assumption of *means missing at random* (MMAR) is:

$$E[g(y)|x, w, z = 0] = E[g(y)|x, w, z = 1] = E[g(y)|x, w]$$

Lemma 3.42. *The means missing at random assumption is non refutable and results in point identification.*

Proof.

$$E[g(y)|x] \overset{\text{LIE}}{=} \sum_{k \in W} E[g(y)|x, w = k]P(w = k|x) \\ \overset{\text{MMAR}}{=} \sum_{k \in W} E[g(y)|x, w = k, z = 1]P(w = k|x) \qquad \square$$

[6] See Manski (2007: 67) [59] for a proof.

Definition 3.43. *Mean independence* (a version of the statistical independence assumption) may be a refutable assumption:

$$E[g(y)|x, w = k] = E[g(y)|x]$$

Without the assumption, the identification region is

$$H[E(g(y)|x, w = k)] = [L(k), U(k)]$$

where $L(k) \equiv LB(w = k)$ and $U(k) \equiv UB(w = k)$ are lower and upper bounds, respectively, given by

$$L(k) = E[g(y)|x, w = k, z = 1]P(z = 1|x, w = k) + g_0 P(z = 0|x, w = k)$$
$$U(k) = E[g(y)|x, w = k, z = 1]P(z = 1|x, w = k) + g_1 P(z = 0|x, w = k)$$

With the assumption of mean independence:

$$H_1[E(g(y)|x)] = [\max_{k \in W} L(k), \min_{k \in W} U(k)]$$

This is a generalisation of (3.10); set $g(y) = 1[y = 1]$ to get (3.10). Here, $E(g(y)|x)$ is not point identified and there is no assumption directly on the distribution of the unobserved variables. The assumption may be refutable if the intersection is empty:

$$H_1[E(g(y)|x)] = \cap_{k \in W} H[E(g(y)|x, w = k)]$$

Another assumption – weaker again – is means missing monotonically.

Definition 3.44. The assumption of *means missing monotonically* (MMM) is:

$$E[g(y)|x, w, z = 1] \geq E[g(y)|x, w, z = 0] \tag{3.11}$$

We need a context to interpret this, e.g. the mean market wage of those that work is no less than that of those who don't work. It is too weak an assumption to point identify the conditional mean $E[g(y)|x]$, but it does have identifying power. To see this, first note that without this assumption we get that

$$E[g(y)|x] \overset{\text{LIE}}{=} \sum_{k \in W} [E[g(y)|x, w = k, z = 1]P(w = k, z = 1|x)]$$
$$+ \sum_{k \in W} \left[\underbrace{E[g(y)|x, w = k, z = 0]}_{\text{unknown}} P(z = 0, w = k|x) \right]$$

Note that LTP implies that $P(w = k|x) = P(w = k, z = 1|x) + P(w = k, z = 0|x)$. With the assumption of means missing monotonically (**??**), we have that:

$$E(g(y)|x) \leq \sum_{k \in W} E[g(y)|x, w = k, z = 1]P(w = k|x)$$

$$\therefore H_1[E[g(y)|x]] = \big[E[g(y)|x, z=1]P(z=1|x) + g_0 P(z=0|x) \sum_{k\in W} E[g(y)|x, w=k, z=1]P(w=k|x)\big]$$

This is a right-truncated subset of the region obtained using the data alone, i.e. the smallest feasible value of $E[g(y)|x]$ is the same as that when using the data alone and the largest is the value that $E[g(y)|x]$ would take under the assumption of means missing at random.

Finally, letting the set V be ordered, we have the weaker assumption of *monotone regressions* or *mean monotonicity*.

Definition 3.45. The assumption of *monotone regressions* or *mean monotonicity* is:

$$E[g(y)|u, v=k] \geq E[g(y)|u, v=k']$$

for all $(k, k') \in V \times V$ such that $k \geq k'$.

The identification region for $E[g(y)|u]$ under the monotone regression assumption follows from a very complicated derivation by Manski & Pepper (2000) [61]:

$$H_1\{E[g(y)|u]\} = \Bigg[\sum_{k\in V} P(v=k)\{\max_{k'\leq k} E[g(y)z + g_0(1-z)|u, v=k']\}, \sum_{k\in V} P(v=k)\{\min_{k'\geq k} E[g(y)z + g_1(1-z)|u, v=k']\}\Bigg]$$

which is a subset of the region obtained using the data alone and a superset of the one obtained under mean independence. This makes sense since the assumption of monotone regressions is a weaker assumption than that of mean independence.

Definition 3.46. *Imputations* assign some logically possible value (e.g. y^*) to each element of the sample that has a missing realization of y.

For instance, we can estimate $E[g(y)]$ using the sample average:

$$\theta_N = \frac{1}{N}\sum_{i=1}^{N} g(y_i)z_i + g(y_i^*)(1-z_i)$$

So, θ_N uses the actual value of y when it is available and the imputation when it is unavailable. By the SLLN, θ_N converges to

$$\theta \equiv E[g(y)|z=1]P(z=1) + E[g(y^*)|z=0]P(z=0)$$

as N increases. Sometimes, the imputed value is y_i^* drawn from the distribution $P(y|z=1)$, so $\theta = E[g(y)]$ if outcomes are MAR; alternatively, y_i^* is sometimes drawn from $P(y|v=v_i, z=1)$ where it is assumed that outcomes are MAR conditional on an IV v. Note that $\theta \in H[E[g(y)]]$ independently of the imputation method used by the nature of imputations, i.e. they are logically possible values of the missing data. $\theta = E[g(y)]$ only if $E[g(y^*)|z=0] = E[g(y)|z=0]$.

3.3.1 Decomposition of mixtures

The *decomposition of mixtures* or *mixing* problem is as follows:

$$A = BC + D(1 - C)$$

We know all A, B, C, D lie in the unit interval and A and C are known. We must determine feasible values of (B, D). Similarly, we want to learn about $P(y|x, w)$ when we only know $P(y|x)$ and $P(w|x)$ from the data. Using LTP, for each covariate value $\zeta \in X$:

$$\underbrace{P(y|x=\zeta)}_{\text{mixture}} = \sum_{\omega \in W} \underbrace{P(y|x=\zeta, w=\omega)}_{\text{components}} \underbrace{P(w=\omega|x=\zeta)}_{\text{mixing probabilities}} \tag{3.12}$$

where $w = \omega$ refers to the mixing covariate. We can restrict $[P(y|x = \zeta, w = \omega), \omega \in W]$ to vectors of distributions that solve the above equation when we know $P(y|x)$ and $P(w|x)$. With no assumptions the identification region is:

$$\begin{aligned} H\{[P(y|x=\zeta, w=\omega), \omega \in W]\} \\ = \Big[& \gamma_\omega \in \Gamma_Y, \omega \in W : \\ & P(y|x=\zeta) = \sum_{\omega \in W} \gamma_\omega P(w=\omega|x=\zeta) \Big] \end{aligned}$$

which is always nonempty. In more familiar notation, the mixing problem is:

$$P(y|x) = \sum_k P(y|x, w=k) P(w=k|x)$$

and the identification regions is given by:

$$H[P(y|w=k), k \in W] = [\gamma_k \in \Gamma_Y, k \in W : P(y) = \sum_{k \in W} \gamma_k P(\gamma = k)]$$

We will soon study the identifying power of different assumptions that may be combined with the data. The distribution $P(y|x = \zeta)$ is a *mixture* of the distributions $[P(y|x = \zeta, w = \omega), \omega \in W]$, which are called *components of the mixture*. $[P(w = \omega|x = \zeta), \omega \in W]$ are the *mixing probabilities*. w is referred to as the *mixing covariate*. Ecological inference and contaminated sample are two of the main motivations behind the decomposition of mixtures problem. Political scientists and sociologists describe the following problem as that of ecological inference. Let us say that we observe two sampling processes, one reveals $P(y|x)$ and the other reveals $P(w|x)$. Ecological inference concerns the problem of inferring $P(y|x, w)$ given knowledge of $P(y|x)$ and $P(w|x)$.

Example 3.47.

$$y = \begin{cases} 1 & \text{if vote democrat} \\ 0 & \text{if not vote democrat} \end{cases}$$

$$w = \begin{cases} 1 & \text{if white} \\ 0 & \text{if not white} \end{cases}$$

Let $y \in \{y_a, y_b\}$, where we are only interested in y_b. Say instead of observing y_a and y_b, we observe y as

$$y \equiv y_a(1-w) + y_b w$$

where w is an unobserved binary random variable, $w \in \{0,1\}$. y is said to be a contaminated measure of y_b. We want to know $P(y_b)$:

$$P(y_b) = P(y_b|w=0)P(w=0) + P(y_b|w=1)P(w=1)$$

Definition 3.48. When $w \perp\!\!\!\perp y_b$, the term *contaminated sampling* is used to describe this observational problem. Without this assumption, the problem is described as that of *corrupted sampling*.

Remark 3.49. If we knew $P(w|x)$, then the sampling process would reveal $P(y|x, w=1)$, but not $P(y_b|x, w=0)$. So, if we knew whether the observations were missing or not, then we would be able to determine $P(y|x, w=1)$, so corrupted sampling would be inference with missing outcomes and contaminated sampling would be inference under the assumption of outcomes MAR.

Definition 3.50. Let $y = y* + u$ where $y*$ is the unobserved outcome of interest and u is an unobserved random variable. Then the observable y measures the unobservable $y*$ with *errors-in-variables*.

Definition 3.51. The problem of inferring $P(y*)$ given $P(y)$ is the *deconvolution problem*. Usually researchers assume that $u \perp\!\!\!\perp$ and $P(u)$ is centered at zero.

Let us assume that the mixing covariate w is binary, $w \in \{0,1\}$. Suppress the conditioning on x, so $P(y|x)$ and $P(w|x)$ are written as $P(y)$ and $P(w)$; also let $p \equiv P(w=0)$. Using the LTP in (3.12):

$$P(y) = pP(y|w=0) + (1-p)P(y|w=1) \tag{3.13}$$

We know $P(y), p$ and $1-p$, but we don't know $P(y|w=0)$ and $P(y|w=1)$. The ID region is:

$$H[P(y|w=0), P(y|w=1)] \\ = \{(\gamma_0, \gamma_1) \in \Gamma_Y \times \Gamma_Y : P(y) = p\gamma_0 + (1-p)\gamma_1\}$$

It is sufficient to study identification regions for either of $P(y|w=0)$ or $P(y|w=1)$ because (3.13) implies that specifying one implies a unique value for the other. Therefore, determining either $H[P(y|w=0)]$ or $H[P(y|w=1)]$ determines the joint identification region $H[P(y|w=0), P(y|w=1)]$. WLOG, consider $P(y|w=1)$. Rearranging (3.13) yields

$$P(y|w=1) = [P(y) - pP(y|w=0)]/(1-p)$$

We can get an identification region for $P(y|w=1)$ by allowing $P(y|w=0)$ to range over all elements of Γ_Y. This is $\{[P(y) - p\gamma_0]/(1-p), \gamma_0 \in \Gamma_Y\}$.

Unfortunately, some of the elements of this set may yield probabilities that lie outside $[0, 1]$ and so are not proper probability distributions. Without these elements, we get the identification region:

$$H[P(y|w = 1)] = \Gamma_Y \cap \{[P(y) - p\gamma_0]/(1 - p), \gamma_{\in}\Gamma_Y\}$$

For event probabilities, ICBST[7]

$$H[P(y \in B|w = 1)] = [0, 1] \cap [[P(y \in B) - p]/(1 - p), P(y \in B)/(1 - p)]$$

The lower bound is positive when $p < 1-P(y \in B)$ and so is informative in this case. The upper bound is positive when $p < P(y \in B)$ and so is informative in this case. Both conditions hold when $p < \frac{1}{2}$ and then the width of the interval is $\frac{p}{1-p}$.

With event probabilities,

$$P(y \in B) = P(y \in B|w = 1)\underbrace{P(w = 1)}_{1-p} + P(y \in B|w = 0)\underbrace{P(w = 0)}_{p}$$

$$P(y \in B|w = 1) = \frac{P(y \in B) - P(y \in B|w = 0)p}{1 - p}$$

The lower bound on $P(y \in B|w = 1)$ is:

$$\frac{P(y \in B) - p}{1 - p}$$

The upper bound on $P(y \in B|w = 1)$ is:

$$\frac{P(y \in B)}{1 - p}$$

So

$$H[P(y \in B|w = 1)] = [0, 1] \cap [\frac{P(y \in B) - p}{1 - p}, \frac{P(y \in B)}{1 - p}]$$

$$\begin{cases} \frac{P(y \in B)-p}{1-p} > 0 \Longrightarrow P(y \in B) > p \\ \frac{P(y \in B)}{1-p} < 1 \Longrightarrow P(y \in B) < 1 - p \end{cases}$$

$$\begin{aligned} p < P(y \in B) < 1 - p \\ p < 1 - p \\ 2p < 1 \\ p < \frac{1}{2} \end{aligned}$$

[7] See Manski (2007: 100) [59] or for a more thorough proof, see Horowitz & Manski (1995, corollary 1.2) [47].

3.4 Treatment Response

Analysis of treatment response is an interesting problem of prediction with missing outcomes: predict outcomes that would occur if alternative treatment rules were applied to a population. We can at most observe the realised outcomes, i.e. outcomes experienced under received treatments, but not the counterfactual outcomes, i.e. outcomes that would be experienced under other treatments. For example, if a group are ill and there are two treatments, *viz.* drugs or surgery where the outcome of interest is life span, then we may wish to predict the life spans that might occur should all patients of a certain type be treated by drugs. However, the only available data on realised life spans would involve some patients that were treated by drugs and the rest by surgery. Another example relates to economic policy where workers displaced from a plant closure were either retrained or assisted in job search where the outcome of interest might be income. We may wish to learn the incomes that might occur if all workers with particular backgrounds were retrained and then compare these incomes with those that would occur if the same workers were given assistance in job search instead. However, available data on realised incomes will most likely involve a subgroup of workers that were retrained and another group, each of who were given job assistance.

More formally, let T be the set of all feasible treatments and each member of the study population possess covariates $x_j \in X$ and a *response function* $y_j(\cdot) : T \longrightarrow Y$ mapping mutually exclusive and exhaustive treatments $t \in T$ into outcomes $y_j(t) \in Y$. Therefore, $y_j(t)$ is the outcome person j would experience if s/he were to receive treatment t.[8] The subscript j on $y_j(\cdot)$ allows treatment response to be heterogeneous across members of the population, i.e. they need not respond to treatment in the same way. Also, treatment response is individualistic, i.e. the outcome that person j experiences is independent of treatments other people receive, hence the notation $y_j(\cdot)$. Let $z_j \in T$ be person j's received treatment, so $y_{\equiv} y_j(z_j)$ is the realised outcome, while counterfactual outcomes are denoted by $[y_j(t), t \neq z_j]$. Observation may reveal $P(y, z|x)$ of realised outcomes and treatments for people with covariates x, while the distribution of outcomes that would occur if all people with covariate x received treatment t is denoted by $P[y(t)|x]$; so, to predict outcomes under a policy of treatment t for people with covariates x, we must infer $P[y(t)|x]$.

Definition 3.52. The *selection problem* refers to the problem of identification of outcome $P[y(t)|x]$ given a knowledge of $P(y, z|x)$.

With treatment response, where $z = t$ is treatment t, from data alone, the

[8] $y_j(t)$ is also called a *potential, latent* or *conjectural* outcome.

problem can be written as:

$$\begin{aligned} P[y(t)|x] &\overset{\text{LTP}}{=} P[y(t)|x, z = t]P(z = t|x) \\ &\quad + \underbrace{P[y(t)|x, z \neq t]}_{\text{unobserved}} P[z \neq t|x] \\ &= P(y|x, z = t)P(z = t|x) \\ &\quad + P[y(t)|x, z \neq t]P(z \neq t|x) \end{aligned}$$

where $P[y(t)|x, z \neq t]$ corresponds to missing outcomes and the remaining quantities after the second equality are known. The identification region using empirical evidence alone is given by

$$H[P(y(t)|x)] = [P(y|x, z = t)P(z = t|x) + \gamma P(z \neq t|x); \gamma \in \Gamma_Y]$$

Note that $P[y(t)|x] = P(y|x, z = t)$ if treatment selection is random, but generally differ otherwise. The primer is the distribution of outcomes that would occur if everyone with covariates x received treatment t, while the latter is the distribution of outcomes that occur for people who have covariates x and actually receive treatment t.

To learn about policies mandating different treatments for people with covariates x, we would like to know about $\{P[y(t)|x], t \in T\}$. The identification region using only data is

$$H\{P[y(t)|x], t \in T\} = \times_{t \in T} H\{P[y(t)|x]\}$$

We learn more about $P[y(t)|x]$ but less about $P[y(t')|x], t' \neq t$, the more often that treatment t is selected in the study population. Furthermore, data alone cannot answer the question as to whether outcomes vary with treatment since counterfactuals are unobservable and observation of realised treatments and outcomes is uninformative regarding all the counterfactual outcome distributions $\{P[y(t)|x, z \neq t], t \in T\}$. It may be possible that in fact for each person j, y_j, the person's realised outcome is the same as the potential outcome under any treatment $y_j(t), t \in T$. Therefore, data alone cannot refute the hypothesis that $P[y(t)|x], t \in T$ are all the same, i.e. that hypothesis is nonrefutable.

Definition 3.53. Focus on two treatments t and t'. The *average treatment effect* (ATE) is

$$E[y(t)|x] - E[y(t')|x]$$

Remark 3.54. Note that the hypothesis $ATE = 0$ is non refutable with data alone since counterfactual observations are missing it is possible that $y_j(t) = y_j(t')\ \forall j$. This hypothesis only becomes refutable if we combine the data with sufficiently strong distributional assumptions, e.g. *randomisation of treatment*:

$$P(y(t)|x, z = t) = P(y(t)|x, z \neq t)$$

which gives point identification. The distributional assumption of statistical independence is the selection problem analog of MAR, i.e.

$$P(y(t)|x) = P(y|x, z = t) = P(y(t)|x, z \neq t)$$

is almost only credible in classical randomized experiments. Equivalently

$$z = t \perp\!\!\!\perp y|x$$

When treatment selection is random, we have point identification of $P(y(t)|x)$ as mentioned before. From LIE and the observability of realised outcomes:

$$\begin{aligned} E[y(t)|x] - E[y(t')|x] & \\ &= E(y|x, z = t)P(z = t|x) \\ &\quad + \underbrace{E[y(t)|x, z \neq t]}_{\in[y_0, y_1])} P(z \neq t|x) \\ &\quad - E(y|x, z = t')P(z = t'|x) \\ &\quad - \underbrace{E[y(t')|x, z = t']}_{\in[y_0, y_1]} P(z \neq t'|x) \end{aligned}$$

So the identification region for ATE is

$$\begin{aligned} H\{E[y(t)|x] - E[y(t')|x]\} & \\ &= \big[E(y|x, z = t)P(z = t|x) + y_0 P(z \neq t|x) \\ &\quad - E(y|x, z = t')P(z = t'|x) - y_1 P(z \neq t'|x), \\ &\quad E(y|x, z = t)P(z = t|x) + y_1 P(z \neq t|x) \\ &\quad - E(y|x, z = t')P(z = t'|x) - y_0 P(z \neq t'|x)\big] \end{aligned}$$

which necessarily contains zero and its width is given by

$$(y_1 - y_0)[P(z \neq t|x) + P(z \neq t'|x)] = (y_1 - y_0)[2 - P(z = t|x) - P(z = t'|x)]$$

i.e. the width of the interval depends on the fraction of the population under study that receive treatments t and t'. Since the sum of the fraction is one, the width of the interval is between $(y_1 - y_0)$ and $2(y_1 - y_0)$. When t and t' are the only feasible treatments, the width is $(y_1 - y_0)$. When there is no data, ATE lies in $[y_0 - y_1, y_1 - y_0]$ and the width is $2(y_1 - y_0)$. Therefore, data alone restricts the ATE to half its logically possible range.

Example 3.55. See section 7.2 in Manski (2007).

Incorporating compliance, let z represent received treatment and d represent assigned treatment. Compliance implies $z_j = d_j$, while non compliance implies $z_j \neq d_j$. If there is no crossover allowed:

$$P(y(a)|x) \overset{\text{randomization}}{=} P(y(a)|x, d = a) \overset{\text{full compliance}}{=} P(y|x, d = a)$$

We have point identification. Allowing cross over:

$$\begin{aligned} P(y(b)|x) &\overset{\text{randomization}}{=} P(y(b)|x, d = b) \\ &\overset{\text{LTP}}{=} P(y(b)|x, d = b, z = b)P(z = b|x, d = b) \\ &+ P[y(b)|x, d = b, z \neq b]P(z \neq b|x, d = b) \\ &= P(y|x, d = b, z = b)P(z = b|x, d = b) \\ &+ \underbrace{P(y(b)|x, d = b, z \neq b)}_{\text{unknown}} P(z \neq b|x, d = b) \end{aligned}$$

Finally note that $P(y(\cdot)|x, z) = P(y(\cdot)|x)$ is a stronger assumption than $P(y(t)|x, z) = P(y(t)|x)$.

When treatment response is linear in the treatment with everyone having the same slope parameter, we have a lot of identifying power but not a lot of credibility. This is the case for linear simultaneous equation models. According to the law of decreasing credibility, weaker assumptions are more credible. One such weaker assumption is the restriction that outcomes vary monotonically with the magnitude of the treatment, which we may have reason to believe in particular circumstances. This is an example of a *shape restriction*.

Definition 3.56. The assumption of *monotone treatment response* (MTR) posits that for all persons j and for all treatment pairs (s, t):

$$t \geq s \implies y_j(t) \geq y_j(s) \; \forall j$$

This is a *non refutable* assumption. Note that

$$y_{0j}(t) = \begin{cases} y_j & \text{if } t \geq z_j \\ y_0 & \text{if } t < z_j \end{cases}$$

$$y_{1j}(t) = \begin{cases} y_1 & \text{if } t \geq z_j \\ y_j & \text{if } t < z_j \end{cases}$$

and

$$y_{0j}(t) \leq y_j(t) \leq y_{1j}(t)$$

and bounds are informative. See diagram 3.3. So $P(y_0(t))$ is stochastically dominated by $P(y(t))$, which in turn is stochastically dominated by $P(y_1(t))$. So, for parameters that respect stochastic dominance:

$$D[y_1(t)] \geq D[y(t)] \geq D[y_0(t)]$$

Example 3.57. See bounds on parameters that respect stochastic dominance, section 9.2 in Manski (2007) [59] and bounds on treatment effects, section 9.3 in Manski (2007) [59]. Graphs will be provided during lectures if this topic is covered.

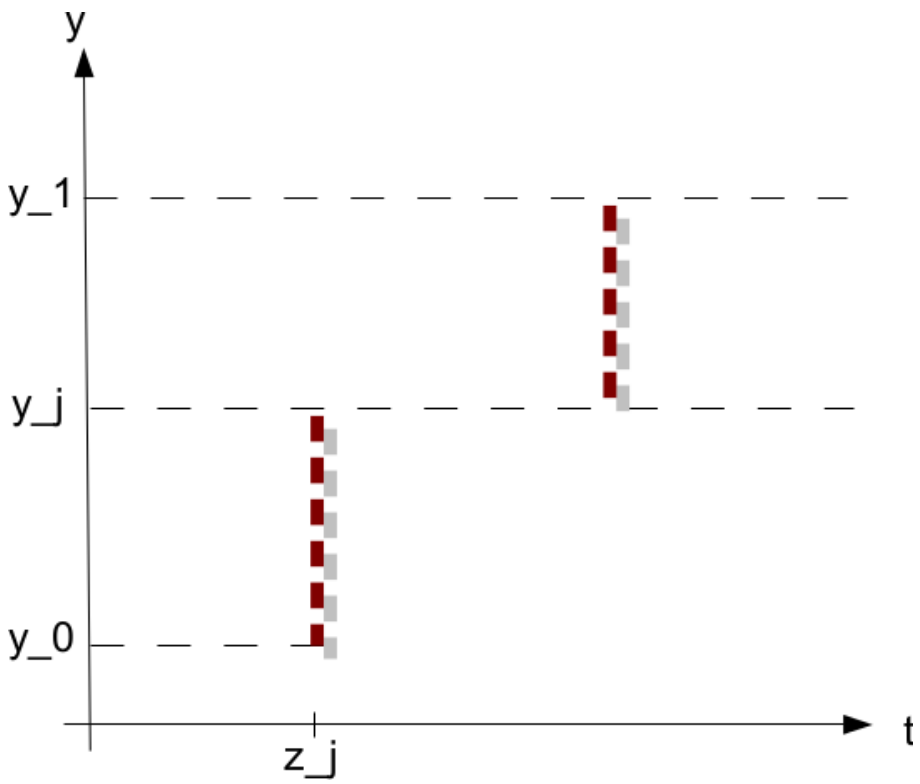

Figure 3.3: Monotone Treatment Response.

MTR allows heterogeneous response functions and is nonrefutable since $\forall j, \exists! y_j(\cdot)$, which is observable, *viz.* $y_j \equiv y_j(z_j)$. So, data alone is consistent with the hypothesis that all response functions are weakly (monotone) increasing. For example, empirical evidence is consistent with the hypothesis that each person's response function is flat, i.e. $\{y_j(t) = y_j, t \in T, \forall j\}$.

Monotonicity is an example of a shape restriction. Other related examples of shape restrictions include semimonotonicity and concave monotonicity.

Example 3.58. Demand analysis is an example of where $y_j(\cdot)$ can be assumed to be monotone. One result from price theory is that market demand is generally a downward-sloping function of price.

Example 3.59. Production analysis contains another such example where we can assume monotonicity. Denote output by $y_j(t)$ and input by t. Then when there is a single input, e.g. labour, $y_j(t)$, the production function is monotone; when there is a vector of inputs, e.g. labour and capital, then treatment response $y_j(t)$ displays semimonotonicity. In both of these cases, production theory generally posits that output weakly increases with the quantity of inputs. Suppose that there are K inputs and let $s \equiv (s_1, s_2, \ldots, s_K)$

and $t \equiv (t_1, t_2, \ldots, t_2)$ be two input vectors. Production theory predicts $y_j(t) \geq y_j(s)$ if $t_k \geq s_k \; \forall k = 1, \ldots, K$, though it does not predict the ordering of $y_j(t)$ and $y_j(s)$ when t and s are unordered, each with some components larger than the other. So, production functions are semimonotone. The assumption of diminishing marginal returns implies that the production function displays concavity in each component holding all other components fixed.

I omit the conditioning on x for simplicity for this section. Let us look at means of increasing functions of outcomes, so allow $f : Y \longrightarrow \mathbb{R}$ be a weakly increasing function and note that $E[f(y(t))]$ respects stochastic dominance. Without any assumptions:

$$\begin{aligned} f(y_0)P(z \neq t) + E[f(y)|z = t]P(z = t) &\leq E[f(y(t))] \\ &\leq f(y_1)P(z \neq t) + E[f(y)|z = t]P(z = t) \\ E[f(y(t))] \overset{\text{LIE}}{=} E[f(y(t))|t = z]P(t = z) &+ \underbrace{E[f(y(t))|t \neq z]}_{\text{unobserved}} P(t \neq z) \end{aligned}$$

We might know (or constrain) $y \in [y_0, y_1]$. With the assumption of MTR

$$\begin{aligned} E[f(y(t))] &= E[f(y(t))|z = t]P(z = t) + E[f(y(t))|z \neq t]P(z \neq t) \\ &= E[f(y(t))|z = t]P(z = t) + E[f(y(t))|z > t]P(z > t) + E[f(y(t))|z < t]P(z < t) \end{aligned}$$

Remember that t in $y(t)$ is the treatment under consideration and $z = t$ is the assigned treatment. Also note that with $E[f(y(t))|z > t]P(z > t)$, these people were assigned a treatment greater than t, so the outcome lies in $[y_0, y]$, while with $E[f(y(t))|z < t]P(z < t)$, these people were assigned a treatment less than t, so the outcome lies in $[y, y_1]$. The lower bound is given by

$$\begin{aligned} &E[f(y)|z = t]P(z = t) + f(y_0)P(z > t) + E[f(y)|z < t]P(z < t) \\ &E[f(y)|z \leq t]P(z \leq t) + f(y_0)P(z > t) \end{aligned}$$

The upper bound is given by

$$\begin{aligned} &E[f(y)|z = t]P(z = t) + E[f(y)|z > t]P(z > t) + f(y_1)P(z < t) \\ &E[f(y)|z \geq t]P(z \geq t) + f(y_1)P(z < t) \end{aligned}$$

Definition 3.60. The *monotone treatment selection* (MTS) assumption states that

$$s' \geq s \Longrightarrow E[y(t)|z = s'] \geq E[y(t)|z = s] \; \forall t \in T$$

So, as illustrated in figure 3.4, MTS means that:

$$E[y(t)|t < z] = \begin{cases} y_0 & \text{lower bound} \\ E[y|t = z] & \text{upper bound} \end{cases}$$

$$E[y(t)|t > z] = \begin{cases} E[y|t = z] & \text{lower bound} \\ y_1 & \text{upper bound} \end{cases}$$

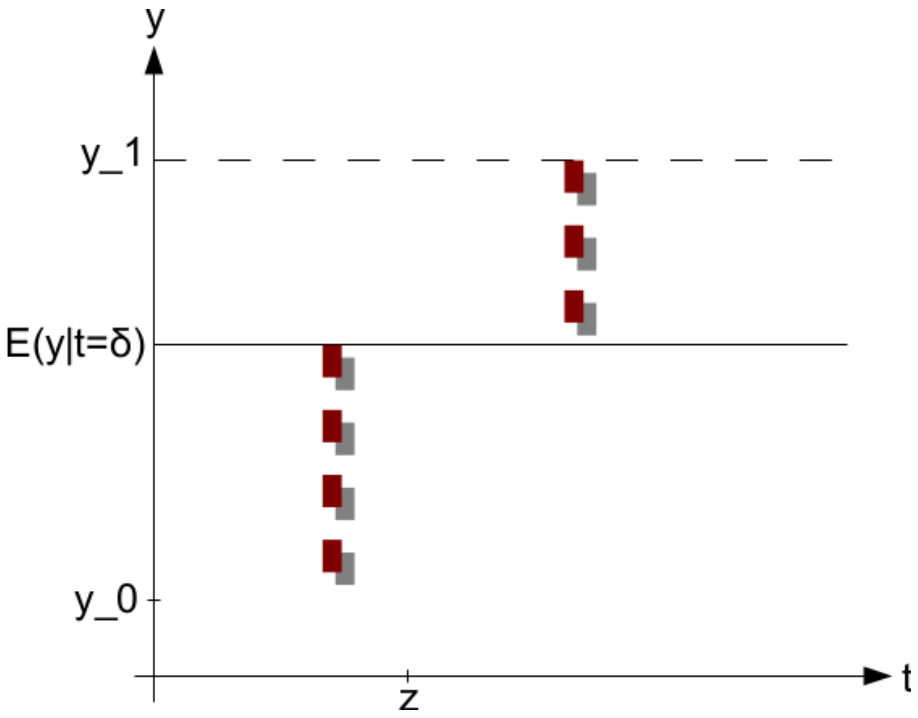

Figure 3.4: Monotone treatment selection.

MTS is a nonrefutable assumption, a version of which was the assumption previously introduced as 'means missing monotonically, e..g when the mean market wage of those that work is at least as large as that of those who don't. To appreciate the subtle distinction between MTR and MTS, consider the statement 'wages increase with schooling.' The MTR interpretation is that each individual's wage is a weakly increasing function of conjectured years of schooling. So, MTR is consistent with the idea that education is a production process. The MTS interpretation is that those who select higher levels of schooling have weakly higher average wage functions than those who select lower levels of schooling. So, MTS is consistent with models that focus on the tendency for people with higher ability to have higher wage functions and higher education. A third interpretation is that observed wages increase with the observed years of schooling, i.e. $E(y|z = t)$ increases with t – a statement regarding the empirical evidence on realised wages, rather than an assumption.

$$E(y(t)) = E[y(t)|z < t]P(z < t)+E[y(t)|z = t]P(z = t)+E[y(t)|z > t]P(z > t)$$

Manski & Pepper (2000) [61] derive bounds on mean outcomes and average treatment effects. Assumption MTS implies a sharp bound on $E[y(t)]$. The

lower bound for $E(y(t))$ is given by

$$y_0 P(t < z) + E[y|t = z]P(t \geq z)$$

The upper bound for $E(y(t))$ is given by

$$E[y|t = z]P(z \leq t) + y_1 P(z > t)$$

Combining MTR and MTS, the sharp bound is then

$$\sum_{s<t} E(y|z = s)P(z = s) + E(y|z = t)P(z \geq t) \leq E[y(t)] \tag{3.14}$$

$$\leq \sum_{s>t} E(y|z = s)P(z = s) + E(y|z = t)P(z \leq t) \tag{3.15}$$

The bound on the ATE with MTR and MTS combined is

$$\begin{aligned} 0 \overset{\text{MTR}}{\leq} & E[y(t)] - E[y(s)] \\ & \leq \sum_{t'>t} E(y|z = t')P(z = t') + E(y|z = t)P(z \leq t) \\ & - \sum_{s'<s} E(y|z = s')P(z = s') - E(y|z = s)P(z \geq s) \end{aligned}$$

where the final inequality follows by subtracting the LHS of (3.14) from the RHS of (3.15).

Combining MTR and MTS increases the identifying power since they are informative even if the outcome space Y is unbounded unlike either assumption alone. Furthermore, the combined assumption is refutable as shown by Manski & Pepper (2000) [61]: $E(y|z = t)$ is a weakly increasing function of t, so if the data show otherwise, then at least one of the two assumptions must be incorrect.

Example 3.61. See section 9.5 in Manski (2007) [59].

3.4.1 Planning Under Ambiguity

Definition 3.62. When treatment varies with observed covariates, we describe this as *screening*, *profiling* or *statistical discrimination*.

Example 3.63. A medical doctor choosing treatment for a population of patients may observe each patient's previous medical history and the results of diagnostic tests. The doctor may make the treatment rule a function of the covariates of the patients. So, acting in the patients' interests, s/he may chose patient health status as the outcome of interest and welfare may be measured as the health status minus the cost of the treatment.

Example 3.64. Reflect on the case of a judge who chooses sentences for a population of convicts. Given the legislative guidelines, the judge may consider observed covariates such as each offender's previous criminal record and demeanor in court when sentencing each convict.

Identification and inference issues prevent complete knowledge of treatment response, so from a planning perspective it is important to know how a planner with partial knowledge of treatment response may reasonably make treatment choices. An actual planner may only have partial knowledge of the distribution of treatment response so s/he may not be able to achieve an optimal policy. This is the problem where the planner faces *ambiguity*, also known as ignorance, uncertainty and Knightian uncertainty.

Definition 3.65. 'A decision maker with a partially known objective function is said to face a problem of choice under ambiguity.' (Manski, 2007: 213) [59]

Let C be the choice set and $f(\cdot) : C \longrightarrow \mathbb{R}$ be the objective function that the decision maker wants to maximise that maps actions into real-valued outcomes. When the choice set and the objective function are known by the decision maker, s/he faces a standard optimisation problem. When only the choice set is known to the decision maker, s/he faces a problem of choice under ambiguity.

Definition 3.66. *Fractional rules* allocate different treatments to observationally identical individuals.[9]

Let $f(\cdot) : C \longrightarrow \mathbb{R}$ and $f(\cdot) \in F$ where F is some set of possible objective functions.

Definition 3.67. An action or decision $d \in C$ is *dominated* if $\exists c \neq d$ feasible action such that

$$g(d) \leq g(c)\ \forall g(\cdot) \in F$$
$$g(d) < g(c) \quad \text{for some } g(\cdot) \in F$$

[9]There is an ethical problem with fractional treatment rules, however since they violate the normative principle of equal treatment of equals. Ex-ante, these rules do not violate this principle since observationally equivalent people have the same chance of receiving a specific treatment. However, ex-post, these rules do violate this principle since observationally equivalent people actually receive different treatments. Ex-ante equal treatment with such rules is found in example such as call for jury service, random drug testing and the Vietnam draft lotteries. See the example given by Manski (2007: 234-5) [59] and also note that once a treatment is known to be significantly effective and another insignificant, medical ethos typically requires cessation of the insignificant treatment. On the other hand, the fractional minimax-regret rule – soon to be defined – is attractive in the sense that it allows society to diversify risk that is privately indivisible; a person receives either treatment a or b, so an individual cannot diversify but a society can diversify by dividing the population and allocating positive fractions of each treatment to parts of the population.

Let $D \subseteq C$ be the set of all undominated actions. Either there is indifference between c and d, i.e.

$$g(c) = g(d) \; \forall g(\cdot) \in F$$

or

$$g(c) > g(d) \;\; \text{some } g(\cdot) \in F$$
$$g(c) < g(d) \;\; \text{some } g(\cdot) \in F$$

A decision maker will never choose a *dominated* action from those in his/her feasible choice set. Let D be the undominated subset of C and $c \in D, d \in D$. Then we either have that $[g(c) = g(d), \; \forall g(\cdot) \in F]$ or $\exists g'(\cdot) \in F \wedge g''(\cdot) \in F : g'(c) > g'(d) \wedge g''(c) < g''(d)$. In the first instance, the decision maker is ambivalent between actions c and d, whereas in the second, c and d are unordered. Note that in an optimisation problem, expanding the set of feasible choices C to $C \cup e$ can't decrease welfare since the decision maker won't choose e if C contains a better action e. However, with planning under ambiguity, expanding the choice set may decrease welfare. To see this, suppose that e neither dominates nor is dominated by any action in the undominated set D. Then the new set of undominated actions will be $D \cup e$ and if the decision maker chooses e, it may be the case that $f(e) < f(c)$. One way of keeping the idea of optimisation, even acknowledging the problem that there is no optimal choice among undominated actions, is to transform the unknown function $f(\cdot)$ into a known function $h(\cdot)$ that may be maximised. Furthermore, as it is difficult to determine the undominated set D, usually the maximisation is over the entire set of feasible actions C. Bayes decision rules (defined shortly) arise from averaging the elements of F and maximising the resulting function. Maximin and minimax-regret criteria (both defined shortly) arise from seeking an action that works uniformly well over all elements of F.

Firstly, let us introduce and examine Bayes decision rules, where the decision maker only knows $f(\cdot) \in F$. Let π be a probability distribution over the elements of $g(\cdot) \in F$ where π expresses the decision maker's personal beliefs about where $f(\cdot)$ lies within F; thus, π is a *subjective* probability distribution or a *prior* distribution. For each feasible action c, let $h(c)$ denote the mean value of $g(c)$ across feasible functions $g(\cdot)$, where the mean is calculated using π:

$$h(c) \equiv \int g(c) d\pi$$

Definition 3.68. A *Bayes decision with respect to* π or *Bayes rule* or *subjective expected utility* (SEU) soles the optimisation problem

$$\max_{c \in C} \int g(c) d\pi$$

where $\int g(c) d\pi$ is the subjective mean of $g(c)$.

Remark 3.69. A basic rationality argument for using a Bayesian criterion is that Bayes decisions are generally undominated when the expectations $\int g(\cdot)d\pi < \infty$. Even if $f(\cdot)$ is not known in practice, we can act as if it is and impute the objective function before choosing an action that is optimal under the imputed function.

Definition 3.70. 'Formally, an *imputation rule* selects some $h(\cdot) \in F$ and chooses an action that maximizes this $h(\cdot)$. Imputation rules are special cases of Bayes rules that place probability on a single element of F.' (Manski, 2007: 216) [59]

Definition 3.71. For every feasible action $c \in C$, let $h(c)$ denote the minimum value of $g(c)$ for all feasible functions $g(\cdot)$, i.e. $h(c) \equiv \min_{g(\cdot)\in F} g(c)$. A *maximin* rule (MM) involves maximising over the worst criterion for all actions and solves the optimisation problem:

$$\max_{c\in C} \min_{g(\cdot)\in F} g(c)$$

where the inner minimisation minimises the value of g while the outer maximisation maximises over the minimum value of g.

Definition 3.72. Denoting the action c and the objective function to be some $g(\cdot)$, the loss in potential welfare from choosing c is called *regret*:

$$\max_{d\in C} g(d) - g(c)$$

So, regret is the difference between outcome under the chosen action and the best possible outcome across all actions.

Denote $h(c) \equiv \max_{g(\cdot)\in F} [\max_{d\in C} g(d) - g(c)]$ to be the maximum regret of action c over all feasible functions $g(\cdot)$.

Definition 3.73. The *minimax-regret* rule (MMR) minimises over the maximum regret and solves the optimisation problem

$$\min_{c\in C} \max_{g(\cdot)\in F} \left[\max_{d\in C} g(d) - g(c)\right]$$

With regard to the MMR, we can calculate it in three steps.

1. Find the regret for any given action. For an action δ

$$\text{regret } = \max(\alpha, \beta) - (\alpha + (\beta - \alpha)\delta)$$

There are two cases:

Case 1: if $\max(\alpha, \beta) = \alpha$:

$$\text{regret } = -\delta(\beta - \alpha) = \delta(\alpha - \beta)$$

Case 2: if $\max \alpha, \beta) = \beta$:

$$\text{regret } = (\beta - \alpha)(1 - \delta)$$

2. Find the max of these regrets:

Case 1: $\delta(\alpha - \beta_L)$
Case 2: $(\beta_U - \alpha)(1 - \delta)$

3.

$$\delta(\alpha - \beta_L) = (\beta_U - \alpha)(1 - \delta)$$
$$\delta_{MMR} = \frac{\beta_U - \alpha}{\beta_U - \beta_L}$$

Example 3.74. MM and MMR are only the same in special cases. For example, let $\max_{d \in C} g(d)$ be constant for all feasible $g(\cdot)$ with value K. Then the maximum regret of action c is given by

$$\max_{g(\cdot) \in F} [\max_{d \in C} g(d) - g(c)] = \max_{g(\cdot) \in F} [K - g(c)] = K - \min_{g(\cdot) \in F} g(c)$$

MMR is unaffected by K that requires choosing an action $c \in C$ to minimise $-\min_{g(\cdot) \in F} g(c)$, which is the same as the MM criterion.

With partial compliance, ζ denotes assigned treatment and z denotes received treatment. Say there are two treatments, a status quo treatment, $t = a$ and an innovation, $t = b$. Let the outcome of interest be binary:

$$y(t) = \begin{cases} 1 & \text{success of treatment } t \\ 0 & \text{failure} \end{cases}$$

Assume a randomized experiment is performed on a study population. Individuals assigned to the innovation can choose the status quo treatment instead, as they wish, but those assigned to the status quo treatment cannot cross over to receive the innovation. Data point-identifies $P[y(a) = 1]$ but only partially identifies $P[y(b) = 1]$; see Manski (2007) [59] section 7.4 for the proof of this.

$$\begin{aligned} H\{P[y(b) = 1]\} \\ &= \big[P(y = 1|\zeta = b, z = b)P(z = b|\zeta = b), \\ &\quad P(y = 1|\zeta = b, z = b)P(z = b|\zeta = b) \\ &\quad + P(z \neq b|\zeta = b)\big] \end{aligned}$$

Now imagine the case where a planner has to choose treatments for a new population that are identical to the study distribution in terms of its distribution of treatment response. Noncompliance is no longer an issue and the planner's objective is to maximise the rate of treatment success. One feasible treatment rule assigns the innovation to everyone, another assigns the status quo to everyone and the rest include fractional treatment allocations that randomly assign

the innovation to a specific fraction of the population and the rest are given the status quo treatment. Let $U(\delta, P)$ be the social welfare from assigning a fraction δ of the population to the innovation and $1-\delta$ to the status quo where P is the population distribution of the treatment response. In addition, let $\alpha \equiv P[y(a) = 1]$ and $\beta \equiv P[y(b) = 1]$, so

$$U(\delta, P) = \alpha(1-\delta) + \beta\delta = \alpha + (\beta - \alpha)\delta$$

The planner solves the optimisation rule

$$\max_{\delta \in [0,1]} U(\delta, P)$$

The optimal rule is such that:

$$\delta = \begin{cases} 1 & \beta > \alpha \\ 0 & \beta < \alpha \\ p \in [0,1] & \beta = \alpha \end{cases}$$

The maximum value of social welfare attainable is $\max(\alpha, \beta)$. Consider the case where the planner knows α and only knows that $\beta \in [\beta_L, \beta_U]$, where

$$\begin{aligned} \beta_L &\equiv P(y = 1|\zeta = b, z = b)P(z = b|\zeta = b) \\ \beta_U &\equiv P(y = 1|\zeta = b, z = b)P(z = b|\zeta = b) + P(z \neq b|\zeta = b) \end{aligned}$$

$\delta = 1$ dominates all other treatment allocations if $\alpha \leq \beta_L$ and $\delta = 0$ dominates all other treatment allocations if $\alpha \geq \beta_U$. In the first case, while the planner does not know the exact value of β, s/he knows that $\beta \geq \alpha$; the planner knows that $\alpha \geq \beta$ in the second case. The planner faces the problem of choice under ambiguity if $\beta_L < \alpha < \beta_U$ so s/he cannot order α and β and so all treatment allocations are undominated. The planner can for instance use MM or MMR or Bayes to choose an allocation.

An MM planner acts as if $\beta = \beta_L$ (i.e. the smallest feasible value) and solves the optimisation problem

$$\max_{\delta \in [0,1]} \alpha + (\beta_L - \alpha)\delta$$

With cases where $\beta_L < \alpha < \beta_U$, $\delta_{MM} = 0$, i.e. everyone is assigned the status quo. This is not a good property of the MM rule.

An MMR planner who chooses allocation δ has a regret at β of

$$\begin{aligned} \max(\alpha, \beta) + [\alpha + (\beta - \alpha)\delta] = (\alpha - \beta)\delta \cdot 1[\beta < \alpha] \\ + (\beta - \alpha)(1 - \delta) \cdot 1[\beta > \alpha] \end{aligned}$$

The MMR over all feasible β is given by

$$\max_{\beta \in [\beta_L, \beta_U]} (\alpha - \beta)\delta \cdot 1[\beta < \alpha] + (\beta - \alpha)(1 - \delta) \cdot 1[\beta > \alpha]$$

$$\max[(\alpha - \beta_L)\delta, (\beta_U - \alpha)(1 - \delta)] \ \because \beta_L < \alpha < \beta_U$$

Therefore an MMR rule solves the optimisation problem

$$\min_{\delta \in [0,1]} \max\left[(\alpha - \beta_L)\delta, (\beta_U - \alpha)(1 - \delta)\right]$$

Note that $(\alpha - \beta_L)\delta$ is increasing in δ and $(\beta_U - \alpha)(1 - \delta)$ is decreasing in δ so δ_{MMR} is chosen to equalise these two quantities:

$$\delta_{MMR} = \frac{\beta_U - \alpha}{\beta_U - \beta_L}$$

So, the MMR allocation is fractional, where the fraction assigned to the innovation, δ, is determined by the location of α within $[\beta_L, \beta_U]$. We have that δ_{MMR} decreases linearly from 0 to 1 as α increases from β_L to β_U.

Remark 3.75. The MMR rule is always fractional when in problems that have two undominated treatments. See complement 11A in Manski (2007)[59] for a proof; also for further work on treatment under ambiguity see Manski (2009) [60].

A Bayesian planner places the prior π on $[\beta_L, \beta_U]$ and then computes the subjective mean value of social welfare. Finally, the planner chooses a treatment allocation so as to maximise this subjective mean, i.e. the Bayesian planner solves the optimisation problem

$$\max_{\delta \in [0,1]} \alpha + [E_\pi(\beta) - \alpha]\delta$$

Here $E_\pi(\beta) = \int \beta d\pi$ is the subjective mean of β with respect to the subjective probability distribution π.

$$\delta_{Bayes} = \begin{cases} 1 & E_\pi(\beta) > \alpha \\ 0 & \alpha > E_\pi(\beta) \\ p \in [0,1] & E_\pi(\beta) = \alpha \end{cases}$$

Example 3.76. Let the status quo treatment be the conventional Illinois Unemployment Insurance (UI) and the innovation be UI with a wage subsidy. Let

$$y(t) = \begin{cases} 1 & \text{unemployed person is rehired within 11 weeks} \\ 0 & \text{else} \end{cases}$$

From Dubin & Rivers (1993, table 1):

$$\alpha = 0.35 \quad P(y = 1|\zeta = b, z = b) = 0.38$$
$$P(z = b|\zeta = b) = 0.68$$

Therefore, ICBST $\beta_L = 0.26$ and $\beta_U = 0.58$. Suppose that the objective is to maximise the fraction of unemployed people that are rehired within eleven

weeks. With an MM rule, everyone will be assigned to the conventional UI; there is no room for innovation with MM. With the Bayes rule, everyone is given the UI with the wage subsidy if $E_\pi(\beta) > 0.35$ and everyone is given the UI if $E_\pi(\beta) < 0.35$. Finally, with the MMR rule, 72% of all unemployed people are given the UI with the wage subsidy while 28% are given the UI.

Definition 3.77. *Additive* planning problems refer to a class of planning problems where 'social welfare adds together individual welfare terms across the members of the population.' (Manski, 2007: 222) [59]

Definition 3.78. *Utilitarian* planning is a special case of additive planning 'where the planner's perspective on individual welfare is the same as the perspective that the members of the population hold for themselves.' (Manski, 2007: 222)

Suppose that the planner observes covariates $x_j \in X$ for each member j of the population and for simplicity X has finitely many elements and $P(x = \zeta) > 0 \; \forall \zeta \in X$. Let Δ be the space of functions $\delta(\cdot, \cdot)$ mapping $T \times X$ into the unit interval whose values sum to one across elements of T, i.e. $\sum_{t \in T} \delta(t, \zeta) = 1 \; \forall \zeta \in X$. Elements of Δ are the feasible treatment rules.

Definition 3.79. *Singleton rules* are a special subclass of Δ, which 'assign all persons with the same observed covariates to one treatment.' (Manski, 2007: 222) [59]

Therefore, $\delta(\cdot, \cdot)$ is a singleton rule if the following holds: for each $\zeta \in X, \delta(t, \zeta) = 1$ for some $t \in T$ and $\delta(s, \zeta) = 0 \; \forall s \neq t$. As opposed to singleton rules, nonsingleton fractional rules allocate persons with covariates ζ randomly across multiple treatments where the assignment shares are $[\delta(t, \zeta), t \in T]$. Sometimes planners are only allowed to use a subgroup of covariates to assign treatments; e.g. s/he may not make treatment assignment a function of race or gender. In these cases, define x as the covariates that the planner is allowed to use.

Note that for any feasible treatment rule $\delta(\cdot, \cdot)$, where the first argument is the treatment and the second argument is the covariate, the population mean social welfare realised if the planner chose rule $\delta(\cdot, \cdot)$ has the following additive form:

$$U(\delta, P) \equiv \sum_{\zeta \in X} P(x = \zeta) \sum_{t \in t} \delta(t, \zeta) \cdot E[u(t) | x = \zeta]$$

where $P(x = \zeta)$ is the distribution of treatment responses and $E[u(t)|x = \zeta]$ is the mean welfare realised when the people with covariates ζ receive treatment t. The planner will want to solve the problem

$$\max_{\delta \in \Delta} U(\delta, P)$$

See Manski (2007: 224) [59] for details on the solution.

It is important to understand the correspondence between the study population and the treatment population.

> 'To the degree that treatment response is heterogeneous, a planner must take care when extrapolating research findings from a study population to a treatment population, as optimal treatments in the two may differ. Hence correspondence between the study population and the treatment population assumes considerable importance.' (Manski, 2007: 227) [59]

Definition 3.80. Let Γ be the set of feasible *states of nature*, so $(P_\gamma, \gamma \in \Gamma)$ is the set of values for the distribution of treatment response the planner considers to be feasible. Let $\delta \in \Delta$ and $\delta' \in \Delta$ be two feasible treatment rules. The rule δ *dominates* the rule δ' if

$$U(\delta, P_\gamma) \geq U(\delta', P_\gamma)\ \forall \gamma \in \Gamma$$
$$U(\delta, P_\gamma) > U(\delta', P_\gamma) \text{ for some} \gamma \in \Gamma$$

Remark 3.81. Note that when $U(\delta, P_\gamma) > U(\delta', P_\gamma)$ for some $\gamma \in \Gamma$ and $U(\delta, P_\gamma) < U(\delta', P_\gamma)$ for other $\gamma \in \Gamma$, then the ranking of the two rules is ambiguous.

A Bayesian planner would solve the optimisation problem

$$\max_{\delta \in \Delta} \int U(\delta, P_\gamma) d\pi$$

where π is subjective. The separable-in-covariate structure of $U(\delta, P_\gamma)$ implies that $\forall \zeta, \delta_{\text{SEU}}(\cdot, \cdot)$ solves $\max_{t \in T} \int E_\gamma(u(t)|x = \zeta) d\pi$. A MM planner would solve

$$\max_{\delta \in \Delta} \min_{\gamma \in \Gamma} U(\delta, P_\gamma)$$

A MMR planner would solve

$$\min_{\delta \in \Delta} \max_{\gamma \in \Gamma} U*(P_\gamma) - U(\delta, P_\gamma)$$

where $U*(P_\gamma)$ is the optimal population mean welfare that would be possible if it was known that $P = P_\gamma$:

$$U*(P_\gamma) \equiv \sum_{\zeta \in X} P(x = \zeta) \{\max_{t \in T} E_\gamma[u(t)|x = \zeta]\}$$

where $U*(P_\gamma) - U(\delta, P_\gamma)$ is the regret of choosing the rule δ when the state of nature is γ.

Conditioning on the state of nature $\gamma \in \Gamma$, let $T = \{a, b\}$, $u[y(t), t, \zeta] = y(t)$ and $Y = [0, 1]$. Population welfare in state γ from treatment t is $E_\gamma[y(t)]$. For every person j, outcome $y_j(t)$ is observable $\iff$ $z_j = t$ and combining this with the LIE yields

$$\begin{aligned} E_\gamma[y(a)] &= E(y|z = a)P(z = a) \\ &\quad + E_\gamma[y(a)|z = b]P(z = b) \\ E_\gamma[y(b)] &= E(y|z = b)P(z = b) \\ &\quad + E_\gamma[y(b)|z = a]P(z = a) \end{aligned}$$

Assume that $0 < E(y|z = a) < 1$ and $0 < E(y|z = b) < 1$ and observe that the counterfactual quantities $\{E_\gamma[y(a)|z = b], E_\gamma[y(b)|z = a]\} \in [0, 1]^2$. So

$$\{E_\gamma[y(a)|z = b] = 1, E_\gamma[y(b)|z = 1] = 0\}$$

$$\implies E_\gamma[y(a)] > E_\gamma[y(b)]$$

$$\{E_\gamma[y(a)|z = b] = 0, E_\gamma[y(b)|z = 1] = 1\}$$

$$\implies E_\gamma[y(a)] < E_\gamma[y(b)]$$

Therefore, the ranking of the rules is ambiguous. While confronting the selection problem with data alone is insufficient for a planner to determined an optimal treatment rule, s/he can use a MM, MMR or Bayes approach to choose treatments. See Manski (2007: 230-32) [59] for further details on this and for an example.

While this will not be covered by this course, in section 11.8 of Manski (2007) [59], there is a very interesting discussion of the question of decentralization where individuals make their own choices regarding treatment versus the situation of where planners decide on treatment choice for a population.

3.5 Weak Identification

Recall that $\hat{\delta}_{2SLS} \equiv (S'_{XZ}(S_{XX})^{-1}S_{XZ})^{-1}S'_{XZ}S_{XY} = (\hat{X}'\hat{X})^{-1}\hat{X}'\hat{Y}$ where $\hat{X} = X(X'X)^{-1}X'Z$ and $\hat{\delta}_{2SLS} = \hat{\delta}(\hat{W}) = \hat{\delta}_{GMM,optimal}(\hat{S}^{-1}) \xrightarrow{p} \delta$ under homoscedasticity $E(\epsilon_i^2|x_i) = \sigma^2$. That is, 2SLS is GMM under homogeneity. But there could be a data generating process such that the empirical distribution $F_n(\cdot)$ is a bad approximation to the population distribution $\phi(\cdot)$ or $F_n(\cdot) \overset{n\longrightarrow\infty}{\longrightarrow} \phi(\cdot)$ by CLT but not for a given sample size. The problem is that 2SLS (or $\hat{\delta}(\hat{S}^{-1})$) in general is biased, i.e. $E(\hat{\delta}) \neq \delta$ when expectations are taken with respect to the n-sample distribution of $\hat{\delta}$. Remember $\hat{\delta} \sim F_n(\cdot) \longrightarrow^{n\longrightarrow\infty} \phi(\cdot)$.

Example 3.82. Consider the regression where y, z, β are scalars:

$$y = z\beta + u$$
$$z = x\gamma + v$$

β is identified if $E(zx) \neq 0$, i.e. $\gamma = \frac{Cov(zx)}{V(x)} \neq 0$ so $E(zx) \geq \delta > 0$ is bounded away from zero. Suppose $E(zx) = 0$. Now

$$\frac{1}{\sqrt{n}}\sum_i x_i u_i \sim N_1$$
$$\frac{1}{\sqrt{n}}\sum_i z_i u_i \sim N_2$$

and the difference

$$\hat{\beta} - \beta = \frac{\frac{1}{\sqrt{n}} \sum_i x_i u_i}{\frac{1}{\sqrt{n}} \sum_i z_i x_i} \longrightarrow \frac{N_1}{N_2}$$

which is a ratio of standard Normals, i.e. a Cauchy distribution, and so has no mean. The problem lies in zx, $E(zx) = n^{-1/2}C$ is such that C cannot be estimated.

The point of weak instruments: correlation between x and z is 'small' (not well defined) and this leads to poor approx of CLT. The Anderson-Rubin statistic is another approach that is robust to the problem of weak instruments. Consider

$$\begin{aligned} y &= x_1\gamma + z\beta + \epsilon \quad \epsilon \sim N(0, \sigma^2) \\ z &= x_1\pi_1 + x_2\pi_2 + v \end{aligned} \tag{3.16}$$

where x_1, x_2 are exogenous and z is endogenous; equation (3.16) is in reduced form. Suppose you want to test $\beta = \beta_0$. Then replace z by the reduced form:

$$\begin{aligned} y - z\beta_0 &= x_1\gamma + z(\beta - \beta_0) + \epsilon \\ &= x_1\gamma + (x_1\pi_1 + x_2\pi_2 + v)(\beta - \beta_0) + \epsilon \\ y - z\beta_0 &= x_1\theta_1 + x_2\theta_2 + v^* \end{aligned}$$

where β_0 is known under H_0 and $\theta_1 = \gamma + \pi_1(\beta - \beta_0)$ so importantly $\theta_2 = \pi_2(\beta - \beta_0)$. To test $\beta_0 = \beta$, we just need to test $\theta_2 = 0$ since $\pi_2 = 0$ implies $E(zx) = 0$ so there is no issue with denominator in IV. The Andersen Rubin statistic is given by:

$$\begin{aligned} AR(\beta_0) &= \frac{(SS_0(\beta_0) - SS_1(\beta_0))/K_2}{SS_1(\beta_0)/(N-K)} \sim F(1, N-2) \\ &= \frac{(y - z\beta_0)'M(X)(y - z\beta_0) - (y - z\beta_0)'M(X)(y - z\beta_0)}{(y - z\beta_0)'M(X)(y - z\beta_0)} \end{aligned}$$

where $K_2 = 1$ is the number of restrictions ($\beta = \beta_0$). The confidence interval for β of size α is

$$C_\beta(\alpha) = \{\beta_0 : AR(\beta_0) \leq F_\alpha(1, N-2)\}$$

and this is adaptive to π_2 being 'small'. There is an article by Nelson & Startz (1992, EMA) on why the confidence interval takes the shape that it does and also Stock & Wright (1996, EMA) and Dufour.

Chapter 4

Stationary Time Series

4.1 Introduction to Time Series

There are three forms of data: time series, cross section and panel. As the name indicates, the time series $\{y_t\}_{t=1}^{T}$ is a sequence of T data points observed over time, e.g. quarterly GDP from 1970 to 2000, monthly CPI from 2000 to 2010, annual US private consumption expenditure from 1995 to 2005. Cross sectional data refers to data $\{x\}_{i=1}^{N}$, which varies across units $i = 1, \ldots, N$ such as individuals, firms, industries and countries; for example micro surveys of firm productivity as measured by units produced per worker, average salary differences between players of different teams during a given year. Finally, panel data refers to series $\{z\}_{i=1,t=1}^{N,T}$ where we observe different units or sections i over time t, so intuitively it is a mix of time series and cross section data where the cross sectional data is observed over time. For example, panel data could be annual observations of current account balances across the Eurozone countries between 2000 and 2005, quarterly real GDP growth across OECD countries from 1950 to 2000, etc. In the first case, the members of the Eurozone are the cross section and the time domain is between 2000 and 2005. In the second example, OECD countries form the cross section and the time series they are observed for are the quarters between 1950 and 2000. In this chapter, I focus on time series data, in particular *stationary* time series data.

Definition 4.1. A *time series* typically consists of set of observations on a variable y taken at equally spaced intervals over time. (Harvey, 1993: 1) [46]

Definition 4.2. When several variables are considered together, we have a *multivariate time series*. (Harvey, 1993: 5) [46]

Prof Bénétrix will cover multivariate time series. However, I will restrict attention for this part of the course to univariate (single) time series.

Definition 4.3. A *stochastic process* is a collection of random variables that are ordered in time.

We will consider modeling time series as stochastic processes, i.e. where each observation can be viewed as following a probabilistic process, i.e. they are random variables evolving over time according to a particular law of probability.

Definition 4.4. A *realisation* is a single draw from the process, $\{y_t\}$.

We defined moments (e.g. mean, variance) of a stochastic process with respect to the distribution of the random variables that is the time series $y_1, \ldots, y_T$. For instance, the first moment (mean) of the process at time t is given by:

$$\mu_t = E(y_t) \quad t = 1, \ldots, T \tag{4.1}$$

which is the average value of y_t over all possible realisations. The variance at time t is given by:

$$Var(y_t) = E[(y_t - \mu_t)^2] \quad t = 1, \ldots, T \tag{4.2}$$

and the covariance between y_t and $y_{t-\tau}$ is defined by:

$$Cov(y_t, y_{t-\tau}) = E[(y_t - \mu_t)(y_{t-\tau} - \mu_{t-\tau})] \quad t = \tau + 1, \ldots, T \tag{4.3}$$

With m observations, let $y_t^{(j)}$ denote the j^{th} observation on y_t. So, empirically:

$$\hat{\mu}_t = \frac{1}{m} \sum_{j=1}^{m} y_t^{(j)} \quad t = 1, \ldots, T$$

These 'ensemble' averages are rare since we only observe a single series of observations for most time series. Then, we need to impose restrictions on the *data generating process* (DGP) in order to conduct meaningful inference on (4.1)– (4.3). This leads to our discussion of stationarity.

Definition 4.5. The *autocovariance function* of a stochastic process y_t is given by

$$E[(y_t - \mu)(y_{t-\tau} - \mu)] = \gamma(\tau) \quad \tau = 1, 2, \ldots$$

A more advanced treatment of the autocovariance function is as follows. Given the realisation $\{y_t^{(1)}\}_{t=-\infty}^{\infty}$, let

$$\mathbf{x}_t^{(1)} = \begin{bmatrix} y_t^{(1)} \\ y_{t-1}^{(1)} \\ \vdots \\ y_{t-j}^{(1)} \end{bmatrix}$$

So, each realisation of $\{y_t\}_{t=-\infty}^{\infty}$ generates a particular value of $\mathbf{x}_t$. We want the probability distribution of the vector $\mathbf{x}_t^{(i)}$ across the realisations i, *viz.* the *joint distribution* of $(Y_t, Y_{t-1}, \ldots, Y_{t-j})$.

Definition 4.6. The j^{th} *autocovariance* of Y_t is

$$\begin{aligned}\gamma_{j,t} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(y_t-\mu_)(y_{t-j}-\mu_{t-j}) && (4.4)\\ &\quad \times f_{Y_t,Y_{t-1},\ldots,Y_{t-j}}(y_t,y_{t-1},\ldots,y_{t-j})dy_t dy_{t-1}\cdots dy_{t-j} && (4.5)\\ &= E(Y_t-\mu_t)(Y_{t-j}-\mu_{t-j})\end{aligned}$$

Since the of (4.5) is that of a covariance $Cov(X,Y) = E(X-\mu_X)(Y-\mu_Y)$ – here of Y_t with its lagged self – we use the term 'autocovariance'. Note that $\gamma_{0,t}$, the 0^{th} autocovariance is the variance. Further, the j^{th} autocovariance, $\gamma_{j,t}$ is the $(1, j+1)$ element of the variance-covariance matrix of the vector $\mathbf{x}_t$; hence, autocovariances are referred to as second moments. The j^{th} autocovariance is the probability limit of the ensemble average:

$$\gamma_{j,t} = \operatorname*{plim}_{I\longrightarrow\infty}\frac{1}{I}\sum_{i=1}^{I}[Y_t^{(i)}-\mu_t]\cdot[Y_{t-j}^{(i)}-\mu_{t-j}]$$

Example 4.7. Looking at how to calculate autocovariances, the process in example one (3.1.5) has zero autocovariances except when $j = 0$:

$$\gamma_{j,t} = E(Y_t-\mu)(Y_{t-j}-\mu) = E(\epsilon_t\epsilon_{t-j}) = 0 \quad \text{for } j\neq 0$$

Definition 4.8. A stochastic process is *weakly* or *covariance stationary* if for all t:

$$E(y_t) = \mu \quad \forall t \tag{4.6}$$

$$E[(y_t-\mu)^2] = \sigma_y^2 = \gamma(0) \quad \forall t \tag{4.7}$$

$$E[(y_t-\mu)(y_{t-\tau}-\mu)] = \gamma(\tau) \quad \forall t,\tau \tag{4.8}$$

So, weak stationarity simply requires that neither the mean μ_t nor the autocovariances $\gamma_{j,t}$ depend on time t.

Example 4.9. The process in the first example is covariance stationary:

$$E(Y_t) = \mu$$

$$E(Y_t-\mu)(Y_{t-j}-\mu) = \begin{cases}\sigma^2 & \text{for } j=0\\ 0 & \text{for } j\neq 0\end{cases}$$

However, since its mean βt is a function of time, the process in the second example is not covariance stationary.

Proposition 4.10. *If a process is covariance stationary, then $\gamma_{j,t}$ depends only on j, the length of time separating the observations; $\gamma_{j,t}$ is invariant to t, the observation date. So, $\gamma_j = \gamma_{-j}$.*

Proof. Let Y_t be a covariance stationary process. By definition

$$\gamma_j = E(Y_t - \mu)(Y_{t-j} - \mu) \tag{4.9}$$

Replace t with $t + j$ and observe that since Y_t is covariance stationary, γ_j is the same for all t (does not depend upon t):

$$\gamma_j = E(Y_{t+j}-\mu)(Y_{[t+j]-j}-\mu) = E(Y_{t+j}-\mu)(Y_t-\mu) = E(Y_t-\mu)(Y_{t+j}-\mu) \overset{(4.9)}{=} \gamma_{-j}$$

$$\therefore \gamma_j = \gamma_{-j} \quad \text{for all } j \in \mathbb{Z} \qquad \square$$

Strict stationarity is a stronger condition requiring the joint probability distribution of a set of r observations at $t_1, t_2, \ldots, t_r$ is the same as the joint probability distribution of a shifted set, observations at $t_1 + \tau, t_2 + \tau, \ldots, t_r + \tau$ for any τ. Let us define this more formally.

Definition 4.11. A process Y_t is *strictly stationary* if for any $j_1, j_2, \ldots, j_n$, the joint distribution of $(Y_t, Y_{t+j_1}, Y_{t+j_2}, \ldots, Y_{t+j_n})$ depends only on the intervals separating the dates $(j_1, j_2, \ldots, \ldots, j_n)$ and not on the date itself (t). (Hamilton, 1994:46) [43]

When the first two moments of the distribution exist (e.g. Normal distribution), strict stationarity implies weak stationarity; note that all that we need to parameterise a multivariate Gaussian distribution are the mean and the variance. In fact, we only need to know that the second moment is finite. If the densities we are integrating in the definitions of expectation and j^{th} autocovariance are time invariant, then μ_t and $\gamma_{j,t}$ will not depend upon time. Going in the other direction, covariance stationarity does not imply strict stationarity; and even if the mean and autocovariances do not depend on time, higher moments (e.g. $E(Y_t^3)$) may.

Estimates of (4.6)– (4.8) can be found from a single series of observations via the following formulae:

$$\hat{\mu} = \bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t \tag{4.10}$$

$$\hat{\gamma}(0) = c(0) = \frac{1}{T}\sum_{t=1}^{T} (y_t - \bar{y})^2 \tag{4.11}$$

$$\hat{\gamma}(\tau) = c(\tau) = \frac{1}{T}\sum_{t=\tau+1}^{T} (y_t - \bar{y})(y_{t-\tau} - \bar{y}) \quad \tau = 1, 2, 3, \ldots \tag{4.12}$$

Equations (4.10), (4.11) and (4.12) are the *sample mean*, *sample variance* and *sample autocovariance*, respectively.

An intuitive definition of an ergodic distribution is that such a distribution requires that observations that are 'sufficiently' far apart should have almost no correlation. For such processes, the above statistics yield consistent estimates.

We will focus on models where stationarity implies ergodicity.[1] Ergodicity is related to whether time averages

$$\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t^{(1)} \tag{4.13}$$

converge to an ensemble concept $E(Y_t)$, for stationary processes.

Definition 4.12. A weakly stationary process $\{Y_t\}$ is *ergodic for the mean* if (4.13) converges in probability to $E(Y_t)$ as $T \longrightarrow \infty$, i.e.

$$\operatorname*{plim}_{T\longrightarrow\infty} \bar{y} = \operatorname*{plim}_{T\longrightarrow\infty} \frac{1}{T}\sum_{t=1}^{T} y_t^{(1)} = E(Y_t)$$

What we need for a process to be ergodic for the mean is that γ_j go to zero sufficiently quickly as j becomes large.

Definition 4.13. A weakly stationary process $\{Y_t\}$ is *ergodic for second moments* if

$$\frac{1}{T-j}\sum_{t=j+1}^{T} (Y_t - \mu)(Y_{t-j} - \mu) \xrightarrow{p} \gamma_j \quad \forall j$$

In many cases, stationarity and ergodicity amount to the same requirements.

A trivial example of a covariance stationary stochastic process is white noise:

Definition 4.14. The sequence of random variables ϵ_t is a *white noise* (WN) process if the sequence is uncorrelated, has constant mean and has constant variance.

So a WN process $\{\epsilon_t\}_{t=-\infty}^{\infty}$ is such that its elements have mean zero and variance σ^2, i.e. $E(\epsilon_t) = 0, E(\epsilon_t^2) = \sigma^2$ and the ϵ's are uncorrelated across time, i.e. $E(\epsilon_t\epsilon_\tau) = 0$ for $t \neq \tau$. Occasionally we may want to replace the assumption $E(\epsilon_t\epsilon_\tau) = 0$ for $t \neq \tau$ by a stronger condition that the ϵ's are independent across time, *viz.* $\epsilon_t, \epsilon_\tau$ are independent for $t \neq \tau$. This second 'independence' assumption implies the first 'uncorrelatedness' assumption, but not *vice-versa*.

Definition 4.15. A process $\{\epsilon_t\}$ is called an *independent white noise process* if

$$\begin{aligned} &E(\epsilon_t) = 0 \\ &E(\epsilon_t^2) = \sigma^2 \\ &E(\epsilon_t\epsilon_\tau) = 0 \quad \forall t \neq \tau \\ &\epsilon_t \perp\!\!\!\perp \epsilon_\tau \quad \forall t \neq \tau \end{aligned}$$

[1] Harvey provides an example of a non-ergodic process, *viz.* the cyclical process in equation (6.3.1) in his book.

Definition 4.16. A process $\{Y_t\}$ is *Gaussian* if the joint density

$$f_{Y_t, Y_{t+j_1}, \ldots, Y_{t+j_n}}(y_t, y_{t+j_1}, \ldots, y_{t+j_n})$$

is Gaussian for any $j_1, j_2, \ldots, j_n$. (Hamilton, 1994:46) [43]

Consider the case when we have T observations in a sample from a random variable Y_t, $\{y_t\}_{t=1}^T$.

Definition 4.17. Let $\{\epsilon\}_{t=1}^T$ be a collection of T independently and identically distributed (iid) variables such that

$$\epsilon_t \sim N(0, \sigma^2)$$

We refer to this sample of size T as originating from a *Gaussian white noise process* since the random variable is generated from a Normal (Gaussian) distribution.

Example 4.18. To see an example of a process that is stationary but not ergodic, suppose that the i^{th} realisation $\{y_t^{(i)}\}_{t=-\infty}^{\infty}$ is generated from a $N(0, \lambda^2)$ distribution

$$Y_t^{(i)} = \mu^{(i)} + \epsilon_t \tag{4.14}$$

where $\mu^{(i)}$ denotes the mean of the i^{th} realisation and $\{\epsilon_t\}$ is Gaussian WN independent of $\mu^{(i)}$ with mean 0 and variance σ^2. Observe that

$$\begin{aligned}
\mu_t &= E(\mu^{(i)}) + E(\epsilon_t) = 0 \\
\gamma_{0,t} &= E(\mu^{(i)} + \epsilon_t)^2 = \lambda^2 + \sigma^2 \\
\gamma_{j,t} &= E(\mu^{(i)} + \epsilon_t)(\mu^{(i)} + \epsilon_{t-j}) = \lambda^2 \quad \text{for } j \neq 0
\end{aligned}$$

So, the process of (4.14) is covariance stationary.[2] Also note that the time average

$$\frac{1}{T}\sum_{t=1}^{T} Y_t^{(i)} = \frac{1}{T}\sum_{t=1}^{T} (\mu^{(i)} + \epsilon_t) = \mu^{(i)} + \frac{1}{T}\sum_{t=1}^{T} \epsilon_t$$

converges to $\mu^{(i)}$ instead of zero, the mean of Y_t.

Consider I realisations from the random variable Y_t for y_t, i.e. $\{y_t^i\}_{i=1}^I$. The *unconditional density* $f_{Y_t}(y_t)$ for the Gaussian WN process is given by

$$f_{Y_t}(y_t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-y_t^2}{2\sigma^2}\right)$$

Observe that since WN variables within the sequence are uncorrelated:

$$E(\epsilon_t \epsilon_{t-\tau}) = \begin{cases} \sigma^2 & \tau = 0 \\ 0 & \tau \neq 0 \end{cases}$$

[2] It does not satisfy the sufficient condition $\sum_{j=0}^{\infty} |\gamma_j| < \infty$ for ergodicity of the mean; see Hamilton chapter 7.

Definition 4.19. We call the series Y_t a *martingale* when Y_t follows a martingale process, i.e. if $E(Y_{t+1}|F_t) = Y_t$, where $F_t \subseteq F_{t+1}$ is the time t information set.

Definition 4.20. Let Y_t follows a martingale process $E(Y_{t+1}|F_t) = 0$. $\{Y_t\}$ is called a *martingale difference* (MD) sequence or 'mds'.

Definition 4.21. The *expectation* of the i^{th} observation of a time series is the mean of the probability distribution (if it exists):

$$E(Y_t) = \int_{-\infty}^{\infty} y_t f_{Y_t}(y_t) dy_t$$

and can be expressed as the probability limit of the 'ensemble' average:

$$E(Y_t) = \operatorname*{plim}_{I \longrightarrow \infty} \frac{1}{I} \sum_{i=1}^{I} Y_t^{(i)}$$

Example 4.22. Let

$$Y_t = \mu + \epsilon_t \; \forall t$$

where ϵ_t is Gaussian WN. Then

$$E(Y_t) = \mu + E(\epsilon_t) = \mu$$

Example 4.23. Let Y_t be a time trend plus Gaussian WN:

$$Y_t = \beta t + \epsilon$$

$$\implies E(Y_t) = \beta t$$

Definition 4.24. The *unconditional mean* of Y_t is the expectation $E(Y_t)$ and denoted μ_t, i.e.

$$E(Y_t) = \mu_t$$

Remark 4.25. Note that this definition permits the general case where the mean can be a function of the time t of the observation. The first example has Y_t not to be a function of time while the second example has Y_t to be a function of time.

Definition 4.26. The *variance* of a random variable Y_t is given by

$$\gamma_{0,t} = E(Y_t - \mu_t)^2 = \int_{-\infty}^{\infty} (y_t - \mu_t)^2 f_{Y_t}(y_t) dy_t$$

where $\gamma_{0,t}$ denotes the variance.

Example 4.27 (Example 4.23 continued)**.** For example 4.23, the variance is

$$\gamma_{0,t} = E(Y_t - \beta t)^2 = E(\epsilon_t^2) = \sigma^2$$

The time domain properties of stationary stochastic processes may be summarised by plotting the autocovariance function $\gamma(\tau)$ against τ.

Remark 4.28. It is unnecessary to also plot over negative values of τ. This follows from the observation that $\gamma(\tau) = \gamma(-\tau)$.

Definition 4.29. *Autocorrelations* are a standardisation of the autocovariance function obtained by dividing by the variance:

$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)} \quad \tau = 0, \pm 1, \pm 2, \dots$$

The *autocorrelation function* is simply a plot of autocorrelations (ρ) against non-negative values of τ.

Remark 4.30. Again, observe that we do not need to also plot ρ against negative values of τ since $\gamma(\tau) = \gamma(-\tau) \implies \rho(\tau) = \rho(-\tau)$. Note further that by definition $\rho(0) = 1$.

While the autocovariance and autocorrelation functions provide the same information and are of the identical shape, typically we plot the autocorrelation function because it is dimensionless. As for the case with the theoretical autocovariance, the sample covariance may be standardised similarly.

Definition 4.31. The *sample autocorrelations* are given by:

$$r(\tau) = \frac{c(\tau)}{c(0)} \quad \tau = 1, 2, \dots$$

The *sample autocorrelation function* or *correlogram* is a plot of $r(\tau)$ against the non-negative values of τ.

I will return to discussing autocorrelation functions and partial autocorrelation functions more deeply in section 4.3, where the latter will then be defined.

Definition 4.32. The *lag operator* is defined as:

$$Ly_t = y_{t-1} \tag{4.15}$$

From (4.15), recursion yields

$$L^\tau y_t = y_{t-\tau} \quad \tau = 1, 2, 3, \dots$$

Note that $L^0 y_t = y_t$.

Definition 4.33. The *forward* or *lead operator* is defined as

$$F = L^{-1}$$

Definition 4.34. The *first difference operator* is defined as

$$\Delta = 1 - L$$

Usual algebraic manipulations can be carried out on all these operators. Many stochastic processes can be expressed via infinite lags withing moving averages.

Definition 4.35. An *indeterministic process* or a *linear process* is any model that can be represented as

$$y_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j} \tag{4.16}$$

where $\psi_0, \psi_1, \psi_2, \ldots$ are parameters.

Remark 4.36. Note that the model will only be 'linear' if the ϵ_t's are independent, not just simply uncorrelated.

In order for the process to have finite variance, we need to have the condition

$$\sum_{j=0}^{\infty} \psi_j^2 < \infty$$

or sometimes the stronger condition

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

Theorem 4.37 (Wold decomposition theorem)**.** [3] *Suppose Y_t is generated by a linearly indeterministic covariance stationary process. Then Y_t can be represented as*

$$Y_t = \epsilon_t + c_1\epsilon_{t-1} + c_2\epsilon_{t-2} + \cdots$$

where ϵ is WN with variance σ_ϵ^2, $\sum_{i=1}^{\infty} c_i^2 < \infty$ and $\epsilon = Y_t - Proj(Y_t|lags\ of\ Y_T)$ so that ϵ_t is 'fundamental' because ϵ is given by a linear forecasting rule where we observe the data.

Linear processes are stationary. So as mentioned above, their properties are well summarised by the autocovariance function and these properties may be approximated to any level of accuracy one wishes by a model from the class of autoregressive-moving average (ARMA) processes, which we will define and discuss in section 4.2.3. For now note that an ARMA(p, q) process:

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t + \theta_1\epsilon_{t-1} + \cdots + \theta_q\epsilon_{t-q} \tag{4.17}$$

[3] See for example Brockwell & Davis (1991) [11].

can be written more concisely via *associated polynomials* in the lag operator. Define

$$\phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p \tag{4.18}$$
$$\theta(L) = 1 + \theta_1 L + \cdots + \theta_q L^q \tag{4.19}$$

so (4.17) can be expressed neatly as

$$\phi(L) y_t = \theta(L) \epsilon_t \tag{4.20}$$

Definition 4.38 (Linear Filter)**.** Let $\{c_j\}$ be a sequence of constants and let

$$c(L) = c_{-r} L^{-r} + c_{-r+1} L^{-r+1} + \cdots + c_0 + c_1 L + \cdots + c_s L^s$$

be a polynomial in L. Note that $X_t = c(L) Y_t = \sum_{j=-r}^{s} c_j Y_{t-j}$ is a moving average of Y_t. Sometimes, we can refer to $c(L)$ as a *linear filter* and X is called a *filtered version of* Y.

4.2 AR, MA, ARMA, ADL models

This section explores ARMA models in addition to ADL models. For ARMA models, I am following Harvey (1993) [46] in assuming without loss of generality (WLOG) that the processes have mean zero, i.e. $\mu = 0$.

4.2.1 Autoregressive (AR) models

Definition 4.39. An *autoregressive process of order* p (AR(p)) is defined by

$$y_t = \mu + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t \quad t = 1, \ldots, T \tag{4.21}$$

denoted by $y_t \sim \text{AR}(p)$.

AR processes have been popular since they have a natural interpretation and are easier to estimate than moving average processes (see section 4.2.2) or mixed processes (see section 4.2.3).

4.2.1.1 Conditions for stationarity – AR(1)

Letting $p = 1$, AR(1) is

$$y_t = \phi y_{t-1} + \epsilon_t \quad t = 1, \ldots, T \tag{4.22}$$

Through repeated substitution, it can be shown that:

$$y_t = \sum_{j=0}^{J-1} \phi^j \epsilon_{t-j} + \phi^J y_{t-J} \tag{4.23}$$

The RHS of (4.23) consists of an MA(J-1) of the WN variable and a term depending on the value of y_{t-J}. Treating y_{t-J} as a fixed number and taking expectations of (4.23), we get that

$$E(y_t) = E\left(\sum_{j=0}^{J-1} \phi^j \epsilon_{t-j}\right) + E(\phi^J y_{t-J}) = \phi^J y_{t-J}$$

Remark 4.40. Note that when $|\phi| \geq 1$, $E(y_t)$ depends on starting value y_{t-J} so (4.23) contains a deterministic component, i.e. a knowledge of y_{t-J} allows us to make a non-trivial prediction for future values of the series, irrespective of the horizon. Else when $|\phi| < 1$, as J becomes large, the deterministic component becomes negligible:

$$\lim_{J \longrightarrow \infty} \phi^J y_{t-J} = 0$$

and hence we can regard the process as having started at some point in the distant past. So we can write:

$$y_t = \sum_{j=0}^{\infty} \phi^j \epsilon_{t-j} \quad t = 1, \ldots, T \tag{4.24}$$

Comparing (4.24) with (4.16), we can see that when $|\phi| < 1$, an AR(1) model is indeterministic. This is because:

$$\sum_{j=0}^{\infty} \phi^{2j} = \frac{1}{1-\phi^2}$$

Finally, note that while $E(y_t) = 0$ for all t,

$$\begin{aligned} \gamma(0) = E(y_t^2) = E\left(\sum_{j=0}^{\infty} \phi^j \epsilon_{t-j}\right)^2 &= \sum_{j=0}^{\infty} \phi^{2j} E(\epsilon_{t-j}^2) \\ = \sigma^2 \sum_{j=0}^{\infty} \phi^{2j} &= \frac{\sigma^2}{1-\phi^2} \end{aligned} \tag{4.25}$$

4.2.1.2 Conditions for stationarity – AR(2)

The second order autoregressive process, AR(2) is given by

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t \quad t = 1, \ldots, T \tag{4.26}$$

Similar to the AR(1), we can decompose the AR(2) model into a deterministic part – it now depends on a pair of starting values – and a stochastic part. Again, if the process is stationary, the influence of the starting values (the deterministic part) is negligible once the starting point is in the distant past.

To explore the deterministic part, first suppress the disturbance term ϵ in (4.26) to get the homogeneous difference equation:

$$\bar{y}_t - \phi_1 \bar{y}_{t-1} - \phi_2 \bar{y}_{t-2} = 0 \tag{4.27}$$

where $\bar{y}$ refers to the mean of the process and the solution depends on the roots of its characteristic equation:

$$x^2 - \phi_1 x - \phi_2 = 0 \tag{4.28}$$

It can be shown that the roots m_1, m_2 are given by

$$m_1, m_2 = \frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{2} \tag{4.29}$$

Three cases arise, which depend on the sign of the term under the square root: (i) positive, (ii) zero or (iii) negative. In the first case, when $\phi_1^2 + 4\phi_2 > 0$ the roots are both real and the solution is given by

$$\bar{y}_t = k_1 m_1^J + k_2 m_2^J$$

where k_1 and k_2 are constants and depend on the starting values $\bar{y}_{t-J}$ and $\bar{y}_{t-J+1}$. If $|m_1| < 1 \wedge |m_2| < 1$, $\bar{y}_t$ is close to zero when J is large. In the second case, when $\phi_1^2 + 4\phi_2 = 0$ so the roots are both real and equal, the solution (a different form) implies that the condition necessary for $\bar{y}_t$ to be negligible is that the root – remember they are both equal – is less than one. In the third case, when $\phi_1^2 + 4\phi_2 < 0$ the roots are complex, in particular, they form a pair of complex conjugates, i.e. $m_3 = a + ib$ and $m_4 = a - ib$. The solution is in the same form as the first case except that it can be rewritten:

$$\bar{y}_t = k_3 r^J \cos(\lambda J + k_4)$$

where k_3 and k_4 are constants and depend on the starting values $\bar{y}_{t-J}$ and $\bar{y}_{t-J+1}$, r is the modulus of the roots (i.e. $r = \sqrt{a^2 + b^2}$ when $m = a + ib$) and λ, which is measured in radians is given by:[4]

$$\begin{aligned} \lambda = \tan^{-1}\left[\frac{\text{Im}(m_1)}{\text{Re}(m_1)}\right] &= \tan^{-1}\left[\frac{(-\phi_1^2 - 4\phi_2)^{\frac{1}{2}}}{\phi_1}\right] \\ &= \cos^{-1}\left[\frac{\phi_1}{2\sqrt{-\phi_2}}\right] \end{aligned}$$

$\bar{y}_t$ has a cyclical time path and damped if $|r| < 1$; also, when J is large, $\bar{y}_t$ is negligible.

[4]Note that the modulus of a complex number is the same as the modulus of the complex conjugate of the same complex number.

When the roots of (4.28) are less than one in absolute value, the deterministic component in the AR(2) goes to zero as J tends to infinity and we are left with a linear process (an infinite MA process):

$$y_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j} \tag{4.30}$$

Let us derive the coefficients in this process. First define the following associated lag polynomials:

$$\begin{aligned} \phi(L) &= 1 - \phi_1 L - \phi_2 L^2 \\ \psi(L) &= \psi_0 + \psi_1 L + \psi_2 L^2 + \cdots + \psi_\tau L^\tau + \cdots \end{aligned} \tag{4.31}$$

(4.26) and (4.30) may be expressed as

$$\begin{aligned} y_t &= \phi^{-1}(L)\epsilon_t \\ y_t &= \psi(L)\epsilon_t \end{aligned}$$

From comparing these two equations, we can see that

$$\phi(L)\psi(L) = 1 \tag{4.32}$$

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

$$\begin{aligned} \iff \psi_0 + (\psi_1 - \phi_1\psi_0)L + (\psi_2 - \phi_1\psi_1 - \phi_2\psi_0)L^2 \\ + (\psi_3 - \phi_1\psi_2 - \phi_2\psi_1)L^3 + \cdots = 1 \end{aligned} \tag{4.33}$$

Observe that since lags of y_t do not enter on the right hand side of (4.33), i.e. only $L^0 = 1$ since $L^0 y_t = y_t$, the coefficients of $L, L^2, L^3, \ldots$ on the right hand side of (4.33) are all identically zero and so they will be all identically zero on the left hand side as well. Therefore:

$$\begin{aligned} \psi_0 &= 1 \\ \psi_1 - \phi_1 &= 0 \\ \psi_j - \phi_1\psi_{j-1} - \phi_2\psi_{j-2} &= 0 \quad j \geq 2 \end{aligned} \tag{4.34}$$

The roots of (4.34) are given by (4.29). If they both lie within the unit circle, i.e. are both less than one in absolute value, then $\psi_j \overset{j\longrightarrow\infty}{\longrightarrow} 0$. ICBST the rate of convergence is sufficient for the process to have finite variance and that the condition that the roots of (4.28) have modulus less than one is a sufficient condition for stationarity. ICBST the conditions for stationarity are the following:[5]

$$\begin{aligned} \phi_1 + \phi_2 &< 1 \\ -\phi_1 + \phi_2 &< 1 \\ \phi_2 &> -1 \end{aligned}$$

[5] See Goldberg, 1958: 171-2 [39].

We know that the roots of (4.28) will be complex when the term under the square root is negative – a necessary condition is $\phi_2 < 0$. These conditions can be summarised graphically; see figure 2.2 and the discussion in Harvey (1993: 19-20) [46].

4.2.1.3 Conditions for stationarity – AR(p)

An AR(p) model (4.21) is stationary when the roots of the characteristic equation

$$x^p - \phi_1 x^{p-1} - \cdots - \phi_p = 0 \tag{4.35}$$

are all less than one in absolute value. Alternatively (and equivalently), the associated lag polynomial equation

$$1 - \phi_1 L - \cdots - \phi_p L^p = 0 \tag{4.36}$$

has the stationarity condition that its roots should all be greater than one in absolute value. With (4.35), the stationarity condition requires the roots to lie *within* the unit circle, while the stationarity condition for (4.36) requires the roots to lie *outside* the unit circle since (4.36) is a similar formula except with $\frac{1}{L}$ instead of x and multiplied by L^p.

4.2.1.4 Autocovariance & autocorrelation functions

For $|\phi| < 1$, AR(1) model (4.22) has zero mean and variance given by (4.25). Let us derive the autocovariance at lag τ. We will express y_t as a linear combination of $\epsilon_t, \epsilon_{t-1}, \ldots, \epsilon_{t-\tau+1}$ by setting $J = \tau$ in (4.23) so

$$\gamma(\tau) = E(y_t y_{t-1}) = E\left[\left(\phi^\tau y_{t-\tau} + \sum_{j=0}^{\tau-1} \phi^j \epsilon_{t-j}\right) y_{t-\tau}\right]$$

As $\epsilon_t, \ldots, \epsilon_{t-\tau+1}$ are each uncorrelated with $y_{t-\tau}$, this reduces to

$$\gamma(\tau) = \phi^\tau E(y_{t-\tau}^2) = \phi^\tau \gamma(0) \quad \tau = 1, 2, \ldots \tag{4.37}$$

Since the autocovariances $\gamma(\tau)$ only depend on τ, we can see that the process is stationary. On the other hand, if we started by assuming stationarity, then we could more directly multiply both sides of (4.23) by $y_{t-\tau}$ and take expectations to get

$$E(y_t y_{t-\tau}) = \phi E(y_{t-1} y_{t-\tau}) + E(\epsilon_t y_{t-\tau}) \quad \tau = 0, 1, 2, \ldots \tag{4.38}$$

Note that for a stationary process, $E(y_{t-1} y_{t-\tau}) = E(y_t y_{t-\tau+1}) = \gamma(t-1)$ and $E(\epsilon_t y_{t-\tau}) = 0$ for $\tau > 0$ since ϵ_t is uncorrelated with lagged values of y_t.

$$\therefore \ \gamma(\tau) = \phi\gamma(\tau - 1) \quad \tau = 1, 2, \ldots$$

which is a first-order difference equation, the solution of which is given by (4.37). Similarly, one can derive the variance using the fact that $E(\epsilon_t y_t) = \sigma^2$.

The autocorrelation for an AR(1) processes has the form

$$\rho(\tau) = \phi^{\tau} \quad \tau = 0, 1, 2, \ldots$$

When $\phi > 0$, the ACF will exponentially decline smoothly as in figure 2.3 (a) in Harvey (1993) [46]; the series is 'slowly changing', i.e. differences between successive values are small. When $\phi < 0$, the ACF will exponentially decline but oscillate between negative and positive values with the starting value being negative as in figure 2.3 (b) in Harvey (1993) [46]; an irregular pattern is created because adjacent observations are negatively correlated.

One can derive the variance and autocovariance of the AR(2) process through a generalisation of (4.38) by multiplying (4.26) by $y_{t-\tau}$ and taking expectations:

$$E(y_t y_{t-\tau}) = \phi_1 E(y_{t-1} y_{t-\tau}) + \phi_2 E(y_{t-2} y_{t-\tau}) + E(\epsilon_t y_{t-\tau}) \tag{4.39}$$

$$\text{Note:} \quad E(\epsilon_t y_{t-\tau}) = 0 \quad \ldots \text{when } \tau > 0$$

$$\therefore \ \gamma(\tau) = \phi_1 \gamma(\tau - 1) + \phi_2 \gamma(\tau - 2) \quad \tau = 1, 2, \ldots \tag{4.40}$$

We get the second-order difference equation for the ACF by dividing (4.40) by $\gamma(0)$:

$$\rho(\tau) = \phi_1 \rho(\tau - 1) + \phi_2 \rho(\tau - 2) \quad \tau = 1, 2, \ldots \tag{4.41}$$

Let $\tau = 1$ and observe that since the ACF is symmetric, $\rho(-1) = \rho(1)$ so:

$$\rho(1) = \phi_1 + \phi_2 \rho(1)$$

and the starting values for the ACF are

$$\rho(0) = 1$$

$$\rho(1) = \frac{\phi_1}{1 - \phi_2}$$

Note that since (4.41) has the same form as the homogeneous equation (4.27) its solution will exhibit the same patterns as the deterministic component of the AR(2) process. Specifically, for complex roots, it may display damped cyclical behaviour as in figure 2.4 in Harvey (1993) [46].

Regarding the variance of the AR(2) process, we can obtain this from (4.39) by setting $\tau = 0$. Now the last term is non-zero since

$$\begin{aligned} E(\epsilon_t y_t) &= E[\epsilon_t(\phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t)] \\ &= \phi_1 E(\epsilon_t y_{t-1}) + \phi_2 E(\epsilon_t y_{t-2}) + E(\epsilon_t^2) \\ &= 0 + 0 + \sigma^2 = \sigma^2 \\ \therefore \ \gamma(0) &= \phi_1 \gamma(1) + \phi_2 \gamma(2) + \sigma^2 \\ \iff \gamma(0) &= \frac{\sigma^2}{1 - \rho(1)\phi_1 - \rho(2)\phi_2} \\ \iff \gamma(0) &= \left(\frac{1 - \phi_2}{1 + \phi_2}\right) \frac{\sigma^2}{[(1 - \phi_2)^2 - \phi_1^2]} \end{aligned}$$

where the last line follows by noting that $\rho(1) = \rho_1/(1-\rho_2)$ and $\rho(2) = \phi_1\rho(1) + \phi_2$ and so we have $\gamma(0)$ in terms of ϕ_1 and ϕ_2.

The time domain properties of any AR process may be derived using the same techniques. For instance, multiplying (4.21) by $y_{t-\tau}$, taking expectations and dividing by $\gamma(0)$ yields the p^{th} order difference equation

$$\rho_{tau} = \phi_1\rho(\tau-1) + \cdots + \phi_p\rho(\tau-p) \quad \tau = 1, 2, \ldots \tag{4.42}$$

When $\tau = 0$, it can be shown that the variance is:

$$\gamma(0) = \frac{\sigma^2}{1 - \rho(1)\phi_1 - \cdots - \rho(p)\phi_p}$$

With any stationary AR(p) process, the ACF 'damps down' in that $\rho(\tau) \longrightarrow 0$ as $\tau \longrightarrow \infty$. Its actual behaviour, e.g. regarding cyclical movements, depends upon the roots of the characteristic equation (4.35).

4.2.2 Moving Average (MA) models

Definition 4.41. A *moving average process of order* q (MA(q)) is defined by

$$y_t = \epsilon_t + \theta_1\epsilon_{t-1} + \cdots + \theta_q\epsilon_{t-q} \quad t = 1, \ldots, T \tag{4.43}$$

denoted by $y_t \sim \text{MA}(q)$.

Remark 4.42. Every finite MA process ($q < \infty$) is stationary.

4.2.2.1 Autocovariance and autocorrelation functions

From (4.43)

$$E(y_t) = E(\epsilon_t + \theta_1\epsilon_{t-1} + \cdots + \theta_q\epsilon_{t-q}) = 0 \; \forall t$$

$$\gamma(0) = E(y_t^2) = (1 + \theta_1^2 + \cdots + \theta_q^2)\sigma^2$$

$$\gamma(\tau) = \begin{cases} (\theta_\tau + \theta_1\theta_{\tau+1} + \cdots + \theta_{q-\tau}\theta_q)\sigma^2 & \tau = 1, \ldots, q \\ 0 & \tau > q \end{cases} \tag{4.44}$$

The process can be seen to be stationary since the mean, variance and covariances are independent of t. The fact that autocovariances at lags greater than q are all zero means that it is easy to identify an MA(q) process using the autocovariance function or the autocorrelation function, since they will each have a distinct 'cut-off' at lag length $\tau = q$. This is in sharp contrast to the autocovariance function of an AR process, which slowly decays towards zero.

Example 4.43 (MA(1))**.** The MA(1) process is defined by

$$y_t = \epsilon_t + \theta\epsilon_{t-1} \quad t = 1, \ldots, T$$

where ϵ_t is a sequence of independent random variables from a distribution with zero mean and constant variance and where θ is a parameter. So,

$$\mu = E(\epsilon_t) + \theta E(\epsilon_{t-1}) = 0$$

and we have that the autocovariance function at lag 0 (i.e. variance) is

$$\begin{aligned}\gamma(0) &= E[(\epsilon_t + \theta\epsilon_{t-1})(\epsilon_t + \theta\epsilon_{t-1})] \\ &= E(\epsilon_t^2) + \theta^2 E(\epsilon_{t-1}^2) + 2\theta E(\epsilon_t\epsilon_{t-1}) \\ &= (1+\theta^2)\sigma^2\end{aligned}$$

The autocovariance function at lag 1 is

$$\begin{aligned}\gamma(1) &= E[(\epsilon_t + \theta\epsilon_{t-1})(\epsilon_{t-1} + \theta\epsilon_{t-2})] \\ &= E(\epsilon_t\epsilon_{t-1}) + \theta E(\epsilon_{t-1}^2) + \theta E(\epsilon_t\epsilon_{t-2}) + \theta^2 E(\epsilon_{t-1}\epsilon_{t-2}) \\ &= \theta E(\epsilon_{t-1}^2) \\ &= \theta\sigma^2\end{aligned}$$

The autocovariance function for any lag $\tau \geq 2$ is $\gamma(\tau) = 0$. We see that the process is stationary since the mean, variance and covariance are independent of t. The autocorrelation function is:

$$\rho(1) = \frac{\theta}{1+\theta^2} \tag{4.45}$$

4.2.2.2 Invertibility

Through repeated substitution, the MA(1) process can be expressed as

$$y_t = \theta y_{t-1} - \theta^2 y_{t-2} + \cdots - (-\theta)^J y_{t-J} + \epsilon_t - (-\theta)^{J+1}\epsilon_{t-J-1} \tag{4.46}$$

When $|\theta| < 1$, y_t does not depend on a shock to the system in the distant past. Letting $J \longrightarrow \infty$, the final term in (4.46) disappears and y_t can be expressed as an AR(∞) with declining weights:

$$y_t = -\sum_{j=1}^{\infty}(-\theta)^j y_{t-j} + \epsilon_t$$

Note that an MA(1) model with $|\theta| > 1$ is still stationary, even though it is not invertible; however, its ACF may be reproduced precisely by an invertible process with parameter $\frac{1}{\theta}$, which can be seen from substituting into equation (4.45):

$$\rho(1) = \frac{\frac{1}{\theta}}{1 + \left(\frac{1}{\theta}\right)^2} = \frac{\theta}{1+\theta^2}$$

Relating to identification, other than when $|\theta| = 1$, an ACF will be the same for two processes – one invertible and one non-invertible. We should restrict attention to invertible processes in order to overcome the problem of identifiability. Furthermore, when $|\theta| = 1$, an MA(1) process may not be uniquely identified from the ACF.

Regarding invertibility for higher order MA processes, necessary conditions relate to the MA polynomial equation (4.19), requiring roots of $\theta(L) = 0$ to lie outside the unit circle. Note for MA(1), $1 + \theta L = 0$ implies $L = -\frac{1}{\theta}$ so $|\theta| < 1 \iff |L| > 1$.

4.2.3 Autoregressive Moving Average (ARMA) models

Definition 4.44. An *autoregressive-moving average process of order (p,q)* is defined as

$$y_t = \mu + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$$

denoted by $y_t \sim$ ARMA(p,q).

AR(p) and MA(q) are special cases of ARMA(p,q) where $q = 0$ and $p = 0$, respectively. By a 'mixed process', we mean that $(p, q) >> 0$.

4.2.3.1 Stationarity and invertibility

Stationarity of mixed processes depends entirely on the autoregressive component. Specifically, the roots of $\phi(L)$ must all be greater than one in absolute value. Invertibility of mixed processes is completely determined by the moving average part and the condition is exactly the same as for the MA(q) process, *viz.* the roots of $\theta(L) = 0$ must lie outside the unit circle. Justification for why stationarity for mixed processes is determined solely by the autoregressive component is most apparent when we consider the method of expressing a mixed process as an infinite moving average. The simplest example of a mixed process is the ARMA(1,1):

$$y_t = \phi y_{t-1} + \epsilon_t + \theta \epsilon_{t-1} \quad t = 1, \ldots, T \tag{4.47}$$

and we can substitute repeatedly for y_t as in (4.23) or we can re-express (4.47) as

$$(1 - \phi L) y_t = (1 + \theta L) \epsilon_t$$

$$\iff y_t = \frac{\epsilon_t}{1 - \phi L} + \frac{\theta \epsilon_{t-1}}{1 - \phi L}$$

If $|L| \leq 1$, then when $|\phi| < 1$

$$\begin{aligned} y_t &= \sum_{j=0}^{\infty} (\phi L)^j \epsilon_j + \theta \sum_{j=0}^{\infty} (\phi L)^j \epsilon_{t-1} \\ &= \sum_{j=0}^{\infty} \phi^j \epsilon_{t-j} + \theta \sum_{j=0}^{\infty} \phi^j \epsilon_{t-j-1} \\ &= \epsilon_t + \sum_{j=1}^{\infty} (\theta \phi^{j-1} + \phi^j) \epsilon_{t-j} \end{aligned}$$

Note that when $\theta = 0$, the above is the same as (4.24) and since $|\phi| < 1$, the weights in the above equation decline rapidly enough for the process to have finite variance and for the existence of autocovariances.

We may derive the infinite MA representation of an ARMA process when $p > 1$ similarly by factorising $\phi(L)$ and expanding $\phi^{-1}(L)\theta(L)$ by partial fractions. More conveniently, we may equate the coefficients on the powers of L in

$$\theta(L) = \phi(L)\psi(L) \tag{4.48}$$

NOTE: $\theta(L)$ and $\phi(L)$ are polynomials defined in (4.18)- (4.19). $\psi(L)$ is the infinite MA process defined in (4.31) and (4.48) is a generalisation of (4.32) used to obtain MA coefficients in the AR(2) case. Expanding and rearranging (4.48) and matching coefficients as in the AR(2) case above, it can be shown that

$$\begin{aligned} \psi_0 &= 1 \\ \psi_j &= \theta_j + \sum_{i=1}^{\min j,p} \phi_i \psi_{j-i} \quad j = 1, \ldots, q \\ \psi_j &= \sum_{i=1}^{\min j,p} \phi_i \psi_{j-i} \quad j > q \end{aligned} \tag{4.49}$$

When $j \geq \max(p, q+1)$, difference equation (4.49) determines the ψ_j's and the starting values are given by the previous p values of ψ_j.

Example 4.45. As previously demonstrated, for the AR(2) model the MA coefficients are determined by the difference equation (4.34) for $j \geq 2$ with starting values $\psi_0 = 1$ and $\psi_1 = \phi_1$.

Example 4.46. In the ARMA(1,1) model (4.47) the ψ_j's are defined from the difference equation

$$\psi_j = \phi\psi_{j-1} \quad j \geq 2$$

and the starting value is given by

$$\psi_1 = \theta + \phi\psi_0 = \theta + \phi$$

We may use similar techniques to derive the infinite AR representation of an invertible ARMA process.

4.2.3.2 Autocovariance and autocorrelation functions

Time domain properties of ARMA, or mixed processes are related to some of those belonging to AR and MA processes. Clearly, multiplying an ARMA(1,1), equation (4.47) by $y_{t-\tau}$ and taking expectations yields

$$\gamma(\tau) = \phi\gamma(\tau-1) + E(\epsilon_t y_{t-\tau}) + \theta E(\epsilon_{t-1} y_{t-\tau}) \quad \tau = 0, 1, 2, \ldots$$

It can be shown that

$$E(\epsilon_t y_{t-\tau}) = \begin{cases} \sigma^2 & \tau = 0 \\ 0 & \tau \geq 1 \end{cases}$$

$$E(\epsilon_{t-1} y_{t-\tau}) = \begin{cases} \phi\sigma^2 + \theta\sigma^2 & \tau = 0 \\ \sigma^2 & \tau = 1 \\ 0 & \tau > 1 \end{cases}$$

Thus, the autocovariance function is given by

$$\begin{aligned} \gamma(0) &= \phi\gamma(1) + \sigma^2 + \theta\phi\sigma^2 + \theta^2\sigma^2 \\ \gamma(1) &= \phi\gamma(0) + \theta\sigma^2 \\ \gamma(\tau) &= \phi\gamma(\tau-1) \quad \tau = 2, 3, \ldots \end{aligned}$$

Plugging $\gamma(1)$ from the second equation into the first yields

$$\gamma(0) = \frac{1 + \theta^2 + 2\phi\theta}{1 - \phi^2}\sigma^2 \tag{4.50}$$

$$\therefore\ \gamma(1) = \frac{(1 + \phi\theta)(\phi + \theta)}{1 - \phi^2}\sigma^2 \tag{4.51}$$

ACF can be found by dividing (4.51) and $\gamma(\tau)$ by (4.51):

$$\rho(1) = \frac{(1 + \phi\theta)(\phi + \theta)}{1 + \theta^{+}2\phi\theta} \tag{4.52}$$

$$\rho(\tau) = \phi\rho(\tau-1) \quad \tau = 2, 3, \ldots \tag{4.53}$$

Looking at the ACF, when $\tau > 1$, its behaviour is determined by the first-order difference equation (4.53), so autocorrelations display exponential decay and exhibit exponential decay with oscillatory behaviour when $\phi < 0$. This is exactly like the AR(1) except that instead of a starting value for the difference equation in the AR(1) model of $\rho(0) = 1$, with an ARMA(1,1) model, the starting value is $\rho(1)$. From (4.52), $\rho(1)$ is a function of both ϕ and θ and $\operatorname{sgn}\rho(1)$ depends on $\operatorname{sgn}(\phi + \theta)$.

Example 4.47. Let us illustrate the type of pattern an ACF may display. Let $\phi = 0.3$ and $\theta = 0.9$, so $\rho(1) = 0.649$ and the ACF declines exponentially (see figure 2.5 (a) on page 28 of Harvey). Alternatively, let the MA parameter $\theta = -0.9$, so $\rho(1) = -0.345$ and the ACF declines exponentially but from a negative starting value (see figure 2.5 (b) on page 28 of Harvey). There are six different patterns for the ACF of an ARMA(1,1), which depend on θ and ϕ.

Remark 4.48. For higher order ARMA, properties can be derived similarly and the ACF displays the pattern of having the first q autocorrelations depending on both the AR and the MA parameters. Higher order autocorrelations are given by a p^{th} order difference equation of the form (4.42) with $\rho(q), \rho(q-1), \ldots, \rho(q-p+1)$ as starting values.

4.2.3.3 Autocovariance generating function

Definition 4.49. The *autocovariance generating function* (ACGF) of a stationary process is defined as the polynomial in the lag operator $g(L)$ such that

$$g(L) = \sum_{\tau=-\infty}^{\infty} \gamma(\tau) L^{\tau} \tag{4.54}$$

where the coefficient on L^j corresponds to the autocovariance at lag τ.

Example 4.50. An MA(∞) can be written in terms of a polynomial in the lag operator $\psi(L)$ as

$$g(L) = |\psi(L)|^2 \sigma^2 = \psi(L)\psi(L^{-1})\sigma^2 \tag{4.55}$$

Proof. Consider (4.44) when $q \longrightarrow \infty$. With the obvious change of notation:

$$\gamma(\tau) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+\tau}$$

Plugging this into (4.54) and noting that $\psi_j = 0$ for $j < 0$ yields

$$g(L) = \sigma^2 \sum_{\tau=-\infty}^{\infty} \sum_{j=0}^{\infty} \psi_j \psi_{j+\tau} L^{\tau} = \sigma^2 \sum_{j=0}^{\infty} \sum_{\tau=-j}^{\infty} \psi_j \psi_{j+\tau} L^{\tau}$$

Making the change of variable $j + \tau = h$ (hence $\tau = h - j$):

$$g(L) = \sigma^2 \sum_{j=-\infty}^{\infty} \sum_{h=0}^{\infty} \psi_j \psi_h L^{h-j} = \sigma^2 \sum_{h=0}^{\infty} \psi_h L^h \sum_{j=0}^{\infty} \psi_j L^{-j}$$ □

Following from (4.5), for an ARMA process:

$$g(L) = \frac{|\theta(L)|^2}{|\phi(L)|^2}\sigma^2 = \frac{\theta(L)\theta(L^{-1})}{\phi(L)\phi(L^{-1})}\sigma^2$$

Example 4.51. For an MA(1):

$$\begin{aligned} g(L) &= (1+\theta L)(1+\theta L^{-1})\sigma^2 \\ &= (1+\theta^2)\sigma^2 + \sigma^2\theta L + \sigma^2\theta L^{-1} \\ &= \gamma(0) + \gamma(1)L + \gamma(-1)L^{-1} \end{aligned}$$

The autocovariances are as in example 4.43

4.2.3.4 Common factors

Definition 4.52. If AR and MA polynomials in (4.20) have a root that is the same, then we say that they have a *common factor*.

With a common factor, we are dealing with a model is over-parameterised because we may construct a model with the same properties via reducing p and q both by one. The model is not identifiable.

Example 4.53. Consider the ARMA(2,1) model

$$y_t = 0.2y_{t-1} + 0.15y_{t-2} + \epsilon_t + 0.3\epsilon_{t-1} \tag{4.56}$$

where the AR polynomial may be factorised as follows:

$$(1 - 0.2L - 0.15L^2) = (1 - 0.5L)(1 + 0.3L)$$

$$\therefore\ y_t = \phi^{-1}(L)\theta(L)\epsilon_t = \frac{(1 + 0.3L)}{(1 - 0.5L)(1 + 0.3L)}\epsilon_t$$

and from this, we can see that (4.54) has the same MA representation as the AR(1) model

$$y_t = 0.5y_{t-1} + \epsilon_t \tag{4.57}$$

So (4.56) and (4.57) will have the same autocovariance functions and therefore (4.56) is overparameterised.

4.2.4 Autoregressive Distributed Lag (ADL) models

A general form for *dynamic regression* is:[6]

$$y_t = \alpha + \sum_{i=0}^{\infty} \beta_i x_{t-i} + \epsilon_t$$

When we believe that the duration of the lagged effects will be very long, we can look at *infinite lag* models allowing these effects to gradually fade over time. However, more typically we have models where changes in x do not have any influence after a certain cut of point usually after only a small number of periods; in this case we are looking at *finite lag* models. As opposed to the classical marginal effect that looks at the response of y to an immediate once off change in x, in a dynamic model, we look at a one-time change in x_t on the equilibrium of y_t. We call β_0 the *impact* or *short-run multiplier* and we call the accumulated effect τ periods later of an impulse at time t the *cumulated effect*, i.e. $\boldsymbol{\beta}_\tau = \sum_{i=0}^{\tau} \beta_t$. Finally, we call $\beta = \sum_{i=0}^{\infty} \beta_t$ the *equilibrium* or *long-run*

[6] When we sum from $i = 0$ to p, we need to choose lag length p. This can be done by the adjusted R^2, Akaike information criterion or the Schwarz information criterion, for example.

multiplier. We usually define the *lag weights* as $\mathbf{w_i} = \frac{\beta_i}{\sum_{j=0}^{\infty} \beta_j}$ so $\sum_{i=0}^{\infty} w_i = 1$ and rewrite the model as:

$$y_t = \alpha + \beta \sum_{i=0}^{\infty} w_i x_{t-i} + \epsilon_t$$

Note that to characterise the period of adjustment to a new equilibrium, we can use the following two statistics: the *median lag*, which is the smallest $q*$ such that $\sum_{i=0}^{q*} w_i \geq 0.5$ and the *mean lag*, which is $\sum_{i=0}^{\infty} i w_i$.[7]

To repeat the definition of a *lag operator*, remember that $Lx_t = x_{t-1}$. Recall that a *polynomial in the lag operator* is:

$$A(L) = 1 + aL + (aL)^2 + (aL)^3 + \cdots = \sum_{i=0}^{\infty} (aL)^i$$

Note that if $|\gamma| < 1$ the *distributed lag* model in the form

$$y_t = \alpha + \beta \sum_{i=0}^{\infty} \gamma^i L^i x_t + \epsilon_t$$

can be written as

$$y_t = \alpha + \beta(1 - \gamma L)^{-1} x_t + \epsilon_t$$

which is called the *moving average* or *distributed lag form*. Multiplying by $(1 - \gamma L)$ and collecting terms yields the *autoregressive form*:

$$y_t = \alpha(1 - \gamma) + \beta x_t + \gamma y_{t-1} + (1 - \gamma L)\epsilon_t$$

Looking at *infinite lag* models, one case is such that we weigh the most recent past greater than the influence of past observations – the latter will fade over time. We would use the *geometric lag model* in this case:

$$\begin{aligned} y_t &= \alpha + \beta \sum_{i=0}^{\infty} (1 - \lambda)\lambda^i x_{t-i} + \epsilon_t \quad 0 < \lambda < 1 \\ &= \alpha + \beta B(L) x_t + \epsilon_t \end{aligned}$$

where

$$B(L) = (1 - \lambda)(1 + \lambda L + \lambda^2 L^2 + \lambda^3 L^3 + \cdots) = \frac{1 - \lambda}{1 - \lambda L}$$

Note that the lag coefficients are $\beta_t = \beta(1 - \lambda)\lambda^i$ and that while the model incorporates *infinite lags*, it only assigns arbitrarily small weights to the distant past, which decline geometrically, i.e. $w_i = (1 - \lambda)\lambda^i$, $0 \leq w_i < 1$. In this case, the *mean lag* is:

$$\bar{w} = \frac{B'(1)}{B(1)} = \frac{\lambda}{1 - \lambda}$$

[7] These results my be meaningless if the lag coefficients don't have the same signs; sometimes this is an indicator of model misspecification.

and the *median lag* is $p*$ such that $\sum_{i=0}^{p*-1} w_i = 0.5$, so

$$p* = \frac{\ln 0.5}{\ln \lambda} - 1$$

The impact multiplier will be $\beta(1-\lambda)$ and the long-run multiplier will be $\beta \sum_{i=0}^{\infty} (1-\lambda)\lambda^i = \beta$.

One issue with finite lag and geometric lag models is that they both impose strong assumptions – and perhaps these assumptions are incorrect too – on the lagged response of the dependent variable with respect to independent variables.

Definition 4.54. An *autoregressive distributed lag* (ADL) model is defined by

$$y_t = \mu + \sum_{i=1}^{P} \gamma_i y_{t-i} + \sum_{j=0}^{r} \beta_j x_{t-j} + \delta w_t + \epsilon_t \tag{4.58}$$

where in the simple case, ϵ_t is serially uncorrelated and homoscedastic (can be relaxed).

The ADL is a general 'compromise' that permits the study of interesting methodological issues and (4.58) may be more neatly written as

$$C(L)y_t = \mu + B(L)x_t + \delta w_t + \epsilon_t$$

where

$$C(L) = 1 - \gamma_1 L - \gamma_2 L^2 - \cdots - \gamma_p L^p$$
$$B(L) = \beta_0 + \beta_1 L + \beta_2 L^2 + \cdots + \beta_r L^r$$

This is an ADL(p,r). The *partial adjustment model* is a special case of the ADL, equivalently ADL(1,0). Other examples are the model of *autocorrelation*, which is an ADL(1,1) with $\beta_1 = -\gamma_1 \beta_0$ and the classical regression model, which is ADL(0,0).

4.2.4.1 Estimation

Apart from the stochastic variables on the right-hand side, ARDL is a linear model with classical disturbances so OLS will be efficient. Conventional tests will be asymptotically valid; hence, for testing linear restrictions we can use Wald statistics though the F statistic is better for finite samples due to its more conservative critical values. However, when $C(1) = 0$, the model is inestimable; in distributed lag term, look at $\frac{\mu}{C(1)}$. Similarly, if $\sum_i \gamma_i = 1$, the stochastic difference equation is unstable and further problems arise; we can test this specification. As a concrete example, consider ARDL(1,0) when $B(L) = 0$:

$$y_t = \mu + \gamma y_{t-1} + \epsilon_t$$

When $\gamma = 1$, we have a *random walk with drift* model:

$$y_t = \mu + y_{t-1} + \epsilon_t$$

Starting time series at time $t = 1$:

$$y_t = t\mu + \sum_s \epsilon_s = t\mu + v_t$$

As the conditional mean increases without limit, the unconditional mean does not exist. The conditional mean of v_t is zero but its conditional variance is $t\sigma^2$ – this illustrates heteroscedasticity. If we consider the OLS estimator of μ, $m = (\mathbf{t}'\mathbf{y})/(\mathbf{t}'\mathbf{t})$ where $\mathbf{t} = [1, 2, \ldots, T]$, then we have that

$$E[m] = \mu + E[(\mathbf{t}'\mathbf{t}^{-1}(\mathbf{t}'\mathbf{v})] = \mu$$

However, we also have that

$$Var[m] = \frac{\sigma^2 \sum_{t=1}^{T} t^3}{\left(\sum_{t=1}^{T} t^2\right)} = \frac{O(T^4)}{[O(T^3)]^2} = O\left(\frac{1}{T^2}\right)$$

Note that the variance is an order of magnitude smaller than usual than usual, so we have a stronger result than that of m being mean square consistent, *viz.* m is mean square *superconsistent*. So, we cannot use the tests whose distributions build on the distribution of $\sqrt{T}(m - \mu)$ since the variance of this normalised statistic will converge to zero. We can still test the hypothesis that $\gamma = 1$ and use the same tests, but now we need to use different critical values.[8]

4.2.4.2 Lag weights

Definition 4.55. The distributed lag form of the ARDL model is called a *rational lag* model:

$$\begin{aligned} y_t &= \frac{\mu}{C(L)} + \frac{B(L)}{C(L)} x_t + \frac{1}{C(L)} \delta w_t + \frac{1}{C(L)} \epsilon_t \\ &= \frac{\mu}{1 - \gamma_1 - \cdots - \gamma_p} + \sum_{j=0}^{\infty} \alpha_j x_{t-j} + \delta \sum_{l=0}^{\infty} \theta_l w_{t-1} + \sum_{l=0}^{\infty} \theta_l \epsilon_{t-l} \end{aligned}$$

This model allows us to approximate very general lag structures, basically producing any shape we desire for the lag distribution with relatively few parameters. The lag coefficients on the x variables are individual terms in the ratio of polynomials appearing in distributed lag form and denoted $\alpha_0, \alpha_1, \alpha_2, \ldots$ as coefficients on $1, L, L^2, \ldots$ in $\frac{B(L)}{C(L)}$. We can write this as $A(L)C(L) = B(L)$

[8] Dickey-Fuller tests are appropriate here. The specific critical values to use will depend on the specifics of the model, e.g. with or without trend, drift and whether the disturbances are white-noise or serially correlated (augmented Dickey-Fuller tests are appropriate for this last consideration if error terms are serially correlated).

and equate coefficients on the powers of L to compute these coefficients, which are the lag weights in the ARDL model. Note that the long-run effect in a rational lag model is given by $\sum_{i=0}^{\infty} \alpha_i$, which can be computed from:[9]

$$\sum_{i=0}^{\infty} \alpha_i = \frac{B(1)}{C(1)}$$

4.2.4.3 Stability

With the geometric lag model, the *stability* condition $|\lambda| < 1$ is necessary for the model to be well behaved. With the AR(1) model, the autocorrelation parameter ρ must be such that $|\rho| < 1$ similarly. The dynamic model (4.58) requires restrictions that are less obvious. To find this restriction, consider the existence of an equilibrium value of y_t. Fix $x_t = \bar{x}$, $w_t = 0$ and $\epsilon = 0$. To find out whether y_t would converge to an equilibrium, consider the following dynamic equation:

$$y_t = \bar{\alpha} + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \cdots + \gamma_p y_{t-p}$$

where we have defined $\bar{\alpha} = \mu + B(1)\bar{x}$. Conditional on y_t actually converging to an equilibrium, the equilibrium would be:

$$\bar{y} = \frac{\mu + B(1)\bar{x}}{C(1)} = \frac{\bar{\alpha}}{C(1)}$$

Whether or not the dynamic equation is stable depends on the *characteristic equation* for the AR part of the model. The roots of this characteristic equation must be greater than one in absolute value to ensure that the model is stable. The characteristic equation is:

$$C(z) = 1 - \gamma_1 z - \gamma_2 z^2 - \cdots - \gamma_p z^p = 0$$

For first-order models, the characteristic equation is:

$$C(z) = 1 - \lambda z = 0$$

and the only root of this equation is $z = \frac{1}{\lambda}$; hence, $|z| > 1 \Longleftarrow |\lambda| < 1$. For more general characteristic equations, the roots are the reciprocals of the characteristic roots of the following matrix:

$$\mathbf{C} = \begin{bmatrix} \gamma_1 & \gamma_2 & \gamma_3 & \cdots & \gamma_{p-1} & \gamma_p \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ & & & \cdots & & \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \tag{4.59}$$

[9] We can use the delta method to compute the standard error for the long-run effect.

Note that the roots may include complex pairs since this is an asymmetric matrix.[10] If one of the roots of $C(z) = 1$, i.e. a unit root, then $\sum_{i=1}^{p} \gamma_i = 1$. As this is an explosive case, it will be a difficult hypothesis to test. In particular, under the null hypothesis of $C(1) = 0$, the F statistic will not have a central F distribution due to the behaviour of the variables in the model.

Example 4.56. See example 20.4 in Greene (2011) [42].

Remark 4.57. Our coverage of time series econometrics does not include vector autoregression (VAR), i.e. we only look at univariate models. Prof Agustín Bénétrix will cover VARs. For now, note that the *univariate autoregression*:

$$y_t = \mu + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \cdots + \gamma_p y_{t-p} + \epsilon_t$$

may be extended with $p - 1$ equations:

$$\begin{aligned} y_{t-1} &= y_{t-1} \\ y_{t-2} &= y_{t-2} \end{aligned}$$

etc. to yield a *vector autoregression* (VAR):

$$\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{C}\mathbf{y}_{t-1} + \boldsymbol{\epsilon}_t$$

where $\mathbf{y}_t$ is $p \times 1$, $\boldsymbol{\epsilon}_t = (\epsilon_t, 0, \ldots,)'$ and $\boldsymbol{\mu} = (\mu, 0, 0, \ldots)'$.

4.2.4.4 Forecasting

Let us now collect the terms in μ, x_t, w_t, etc. into a single term:

$$\mu_t = \mu + \sum_{j=0}^{r} \beta_j x_{t-j} + \delta w_t$$

so the ARDL model reduces to

$$y_t = \mu_t + \gamma_1 y_{t-1} + \cdots + \gamma_p y_{t-p} + \epsilon_t$$

Definition 4.58. Given information up to time T and forecasts of exogenous variable, the *one-period-ahead forecast* of y_t in the ARDL(p, r) model is:

$$\hat{y}_{T+1|T} = \hat{\mu}_{T+1|T} + \gamma_1 y_T + \cdots + \gamma_p y_{T-p+1} + \hat{\epsilon}_{T+1|T}$$

In order to compute the prediction interval, we must first consider the variance of the forecast error:

$$e_{T+1|T} = \hat{y}_{T+1|T} - y_{T+1}$$

[10] For the complex number $a+bi$, its reciprocal is given by $\frac{a}{M} - \left(\frac{b}{M}\right) i$, where $M = a^2 + b^2$ and $i^2 = -1$. So, we need $M < 1$. We will cover some basic complex analysis in the last topic of this chapter, *viz.* the frequency domain approach.

This error comes from three sources. Firstly, since $\mu, \delta, \beta_0, \ldots, \beta_r$ must be estimated, $\hat{\mu}_{T+1|T}$ will differ from μ_{T+1} due to sampling variability in these estimators. Secondly, if x_{T+1} and w_{T+1} have been forecasted, since forecasts are imperfect, there will be another source of error in the forecast. Thirdly, while ϵ_{T+1} is forecast with its expectation of zero, the realisation generally will not be zero and so here we have again a source of error. While in principle, estimating the forecast variance $Var(e_{T+1|T})$ would account for all of these sources of error, in practice it is hard to handle the first two of these errors. So, we focus now on the third source of the errors before looking at the first two sources. Ignoring the variation in $\hat{\mu}_{T+1|T}$, i.e. assuming parameters are known and exogenous variables are perfectly forecasted, the variance of the forecast error is

$$Var(e_{T+1|T}|x_{T+1}, w_{T+1}, \mu, \boldsymbol{\beta}, \delta, y_T, \ldots) = Var(\epsilon_{T+1}) = \sigma^2$$

Now it is easy to form the forecast and compute the forecast variance. Let $\mathbf{z}_{T+1} = [1, x_{T+1}, x_T, \ldots, x_{T-r+1}, w_T, y_T, y_{T-1}, \ldots, y_{T-p+1}]$ and denote the full estimated parameter vector by $\hat{\theta}$. We would then use

$$\text{Estimated } Var(e_{T+1|T}|Z_{T+1}) = s^2 + \mathbf{z}'_{T+1}\{\text{Estimated } AVar(\hat{\theta})\}\mathbf{z}_{T+1}$$

With respect to forecasting further than one period ahead, the two-period ahead forecast is given by:

$$\hat{y}_{T+2|T} = \hat{\mu}_{T+2|T} + \gamma_1 \hat{y}_{T+1|T} + \cdots + \gamma_p y_{T-p+2} + \hat{\epsilon}_{T+2|T}$$

Observe that we use the forecasted y_{T+1} for period $T+1$ and so substituting for $\hat{y}_{T+1}$ we get:

$$\hat{y}_{T+2|T} = \hat{\mu}_{T+2|T} + \gamma_1(\hat{\mu}_{T+1|T} + \gamma_1 y_T + \cdots + \gamma_p y_{T-p+1} + \hat{\epsilon}_{T+1|T}) + \cdots + \gamma_p y_{T-p+2} + \hat{\epsilon}_{T+2|T}$$

We can do similar for subsequent periods. To simplify the method, consider the first forecast period and write the forecast with the previous p lagged values as follows:

$$\begin{bmatrix} \hat{y}_{T+1|T} \\ y_T \\ y_{T-1} \\ \vdots \end{bmatrix} = \begin{bmatrix} \hat{\mu}_{T+1|T} \\ 0 \\ 0 \\ \vdots \end{bmatrix} + \begin{bmatrix} \gamma_1 & \gamma_2 & \cdots & \gamma_p \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} y_T \\ y_{T-1} \\ y_{T-2} \\ \vdots \end{bmatrix} + \begin{bmatrix} \hat{\epsilon}_{T+1|T} \\ 0 \\ 0 \\ \vdots \end{bmatrix}$$

Observe that the coefficient matrix on the right-hand side is simply $\mathbf{C}$ from equation (4.59). We still use the notation $\hat{\mu}_{T+1|T}$ as the forecast for the deterministic part of the model, though for now we know this value along with $\mathbf{C}$. Our forecast is the top element of the forecast vector:

$$\hat{\mathbf{y}}_{T+1|T} = \hat{\boldsymbol{\mu}}_{T+1|T} + \mathbf{C}\mathbf{y}_T + \hat{\boldsymbol{\epsilon}}_{T+1|T}$$

With our assumption that we know each quantity on the right-hand side except for the period $T+1$ disturbance, we can get the covariance matrix for this $p+1$ vector dimensional vector as

$$E[(\hat{\mathbf{y}}_{T+1|T} - \mathbf{y}_{T+1})(\hat{\mathbf{y}}_{T+1|T} - \mathbf{y}_{T+1})'] = \begin{bmatrix} \sigma^2 & 0 & \cdots \\ 0 & 0 & \vdots \\ \vdots & \cdots & \ddots \end{bmatrix}$$

So, the forecast variance of $\hat{y}_{T+1|T}$ is σ^2. We can extend this to forecasting further ahead:

$$\begin{aligned} \mathbf{y}_{T+2|T} &= \hat{\boldsymbol{\mu}}_{T+2|T} = \mathbf{C}\hat{\mathbf{y}}_{T+1|T} + \hat{\boldsymbol{\epsilon}}_{T+2|T} \\ &= \hat{\boldsymbol{\mu}}_{T+2|T} + \mathbf{C}\hat{\boldsymbol{\mu}}_{T+1|T} + \mathbf{C}^2\mathbf{y}_T + \hat{\boldsymbol{\epsilon}}_{T+2|T} + \mathbf{C}\hat{\boldsymbol{\epsilon}}_{T+1|T} \end{aligned}$$

As before, the only unknown quantities are the disturbances; hence, the forecast variance in this case is given by

$$Var(\hat{\boldsymbol{\epsilon}}_{T+2|T} + \mathbf{C}\hat{\boldsymbol{\epsilon}}_{T+1|T}) = \begin{bmatrix} \sigma^2 & 0 & \cdots \\ 0 & 0 & \vdots \\ \vdots & \cdots & \ddots \end{bmatrix} + \mathbf{C} \begin{bmatrix} \sigma^2 & 0 & \cdots \\ 0 & 0 & \vdots \\ \vdots & \cdots & \ddots \end{bmatrix} \mathbf{C}'$$

Let $\boldsymbol{\Psi}(1) = \mathbf{C}\mathbf{j}\mathbf{j}'\mathbf{C}'$, where $\mathbf{j}' = (\sigma, 0, \ldots, 0)$. Then the forecast variance for the two-step-ahead forecast is $\sigma(1 + \boldsymbol{\Psi}(1)_{11})$, where $\boldsymbol{\Psi}(1)_{11}$ is the $(1,1)$ element of $\boldsymbol{\Psi}(1)$. Finally, generalising this device to a forecast F periods beyond the sample:

$$\hat{\mathbf{y}}_{T+F|T} = \sum_{f=1}^{F} \mathbf{C}^{f-1}\hat{\boldsymbol{\mu}}_{T+F-(f-1)|T} + \mathbf{C}^F\mathbf{y}_T + \sum_{f=1}^{F} \mathbf{C}^{f-1}\hat{\boldsymbol{\epsilon}}_{T+F-(f-1)|T} \quad (4.60)$$

which allows straightforward computation of forecasts. The conditional forecast variance can be expressed as:

$$\text{Conditional } Var(\hat{y}_{T+F|T}) = \sigma^2(1 + \boldsymbol{\Psi}(1)_{11} + \boldsymbol{\Psi}(2)_{11} + \cdots + \boldsymbol{\Psi}(F-1)_{11})$$

where $\boldsymbol{\Psi}(i) = \mathbf{C}^i\mathbf{j}\mathbf{j}'\mathbf{C}^{i\prime}$. Looking at the F-period-ahead forecast, i.e. equation (4.60), when the equation is stable (i.e. all roots of matrix $\mathbf{C}$ are less than one in absolute value) $\mathbf{C}^F$ converge to zero and since the forecasted disturbances are zero, the forecast will be dominated by the sum in the first term. If we also suppose that the forecasts of the exogenous variables are simply the period $T+1$ forecasted values and are not revised, then the forecast will converge to

$$\lim_{F \longrightarrow \infty} \hat{\mathbf{y}}_{T+F|T} | \hat{\boldsymbol{\mu}}_{T+1|T} = [\mathbf{I} - \mathbf{C}]^{-1}\hat{\boldsymbol{\mu}}_{T+1|T}$$

So far, we have assumed that parameters are known and exogenous variables are perfectly forecasted. Now let us allow for all sources of variation in forecasts

by letting the forecast variance include variation in the forecasts of exogenous variables and variation in parameter estimates. The first is likely to be intractable, while revision is extremely difficult for the second, especially when we account for $\mathbf{C}$ and $\boldsymbol{\mu}$ to be built from estimated parameters. No longer impossible, it is still an extremely difficult task. We require:

$$\begin{aligned}\text{Estimated Conditional } Var(\hat{y}_{T+F|T}) = \sigma^2[1 + \boldsymbol{\Psi}(1)_{11} + \boldsymbol{\Psi}(2)_{11} + \cdots + \boldsymbol{\Psi}(F-1)_{11}] \\ + \mathbf{g}'\text{Estimated } AVar(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})\mathbf{g}\end{aligned}$$

where we define $\mathbf{g}$ as

$$\mathbf{g} = \frac{\partial \hat{y}_{T+F}}{\partial[\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}]}$$

We can use the bootstrap method for this application, which involves sampling new sets of disturbances from the estimated distribution of ϵ_t and then repeatedly rebuilding the within-sample time series of observations on y_t via

$$\hat{y}_t = \hat{\mu}_t + \gamma_1 y_{t-1} + \cdots + \gamma_p y_{t-p} + e_{bt}(m)$$

where we let $e_{bt}(m)$ be the estimated 'bootstrapped' disturbance in period t during replication m. We repeat the process M times and use new parameter estimates and generate a new forecast for each replication. The estimated forecast variance is given by the variance of these forecasts.

4.3 Autocorrelation and Partial Autocorrelation Functions

4.3.1 Autocorrelation Functions

We can equivalently define the autocovariance function for the process y_t as $\gamma(k) = Cov(y_t, y_{t-k})$ and the autocorrelation function as $\rho(k) = \frac{\gamma(k)}{\gamma(0)}$, where $-1 \leq \rho(k) \leq 1$. Remember that for a stationary process, the ACF is a function of k in addition to the parameters of the process and summarises time domain properties. With a stationary stochastic process, a characteristic of the ACF is that it either suddenly goes to zero at a finite lag or slowly tends to zero.

Example 4.59 (AR(1))**.** The AR(1) process has an ACF of

$$\rho(k) = \phi^k$$

and so this geometric series declines monotonically from $\rho(0) = 1$ when $\phi > 0$ and also declines towards zero with damped oscillations between positive and negative values if $\phi < 0$. For the AR(1), i.e. $y_t = \phi y_{t-1} + \epsilon_t$, recall that $\rho(k) = \phi\rho_{k-1}$ for $k \geq 1$, which interestingly resembles the process itself.

Example 4.60 (AR(2))**.** For AR(2), WLOG $\mu = 0$ as before:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t$$

and if the process is stationary, then since $Var(y_t) = Var(y_{t-s})\ \forall s$, $Var(y_t) = Cov(y_t, y_t)$ and $Cov(\epsilon_t, y_{t-s}) = 0$ if $s > 0$:

$$\gamma(0) = \phi_1\gamma(1) + \phi_2\gamma(2) + \sigma_\epsilon^2$$

ICBST with additional lags:

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

$$\therefore\ \gamma(0) = \sigma_\epsilon^2 \frac{\left[\frac{1-\phi_2}{1+\phi_2}\right]}{(1-\phi_2)^2 - \phi_1^2}$$

And since the variance is constant, dividing by $\gamma(0)$ yields the expression for the autocorrelations:

$$\rho(1) = \phi_1\rho(0) + \phi_2\rho(1)$$

which also uses the starting values $\rho(0) = 1$ and $\rho(1) = \frac{\phi_1}{1-\phi_2}$. ICBST with additional lags:

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

In general for $k \geq 2$:

$$\rho(k) = \phi_1\rho(k-1) + \phi_2\rho(k-2)$$

As with the AR(1) example, the ACF follows the same difference equation as the series itself and the behaviour of the ACF depends on the parameters ϕ_1, ϕ_2 and k. The characteristic equation determines the exact behaviour of the ACF:

$$\rho(k) = \phi_1 \left(\frac{1}{z_1}\right)^k + \phi_2 \left(\frac{1}{z_2}\right)^k$$

and the roots of this equation are given by

$$\frac{1}{z} = \frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{2}$$

As earlier, if we are to have two real roots, then their reciprocals lie within the unit circle (i.e. they will be less than one in absolute value) so $\rho(k)$ will be a sum of two terms decaying to zero; if we are to have two complex roots, then $\rho(k)$ will be a sum of two terms exhibiting damped oscillatory behaviour.

While most applications of AR models focus on the case where $p \leq 2$, higher order AR models can be dealt with similarly.

Example 4.61 (AR(p))**.**

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

the autocovariances are given by the *Yule-Walker equations*

$$\gamma(0) = \phi_1\gamma(1) + \phi_2\gamma(2) + \cdots + \phi_p\gamma(p) + \sigma_\epsilon^2$$
$$\gamma(1) = \phi_1\gamma(0) + \phi_2\gamma(1) + \cdots + \phi_p\gamma(p-1)$$

etc. Once again, the ACF follows the same difference equation as the series itself, *viz.*

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

Example 4.62 (MA(1))**.** For an MA(1) process y_t:

$$y_t = \epsilon_t - \theta\epsilon_{t-1}$$
$$\gamma(0) = (1+\theta^2)\sigma_\epsilon^2$$
$$\gamma(1) = -\theta\sigma_\epsilon^2$$
$$\gamma(k) = 0 \quad k > 1$$

Example 4.63 (MA(2))**.** Multiplying out the terms and taking expectations for an MA(2) process yields

$$\gamma(0) = (1+\theta_1^2+\theta_2^2)\sigma_\epsilon^2$$
$$\gamma(1) = (-\theta_1 + \theta_1\theta_2)\sigma_\epsilon^2$$
$$\gamma(2) = -\theta_2\sigma_\epsilon^2$$
$$\gamma(k) = 0 \quad k > 2$$

For the MA(q) process, the pattern is similar. The key characteristic of an MA process is that the ACF suddenly drops to zero at $q+1$. This will help us distinguish between AR and MA processes.

ARMA(p, q) processes are more complicated because they are mixtures of AR and MA forms.

Example 4.64 (ARMA(1,1))**.** Consider the ARMA(1,1) process, which is defined by

$$y_t = \phi y_{t-1} + \epsilon_t - \theta\epsilon_{t-1}$$

In this case, the Yule-Walker equations are given by

$$\gamma(0) = E[y_t(\phi y_{t-1} + \epsilon_t - \theta\epsilon_{t-1})] = \phi\gamma(1) + \sigma_\epsilon^2 - \sigma_\epsilon^2(\theta\phi - \theta^2)$$
$$\gamma(1) = \phi\gamma(0) - \theta\sigma_\epsilon^2$$
$$\gamma(k) = \phi\gamma(k-1) \quad k > 1$$

Example 4.65 (ARMA(p, q))**.** The characteristic of an ARMA process is that when the MA part is of order q, there will be q terms in the ACF that are complicated functions of both the AR and MA components, but after q periods, the autocorrelation will be given by

$$\rho(k) = \phi_1\rho(k-1) + \phi_2\rho(k-2) + \cdots + \phi_p\rho(k-p) \quad k > q$$

4.3.2 Partial Autocorrelation Functions

While the ACF describes the gross correlation between y_t and y_{t-k}, this can mask a very different underlying relationship. For instance, perhaps we observe a correlation between y_t and y_{t-2} mainly because both variables are correlated with y_{t-1}. Looking at the AR(1) process, $y_t = \phi y_{t-1} + \epsilon_t$, $E(\epsilon_t) = 0$ implies $E(y_t) = E(y_t)/(1-\phi) = 0$. The second gross autocorrelation is $\rho(2) = \phi^2$, but we may be interested in knowing what the correlation between y_t and y_{t-2} actually is *net of the intervening effect of* y_{t-1}. With this AR(1), removing the effect of y_{t-1} from y_t implies that only ϵ_t remains, which is uncorrelated with y_{t-2}. So, the *partial autocorrelation* between y_t and y_{t-2} is zero for the AR(1).

Definition 4.66 (Partial Autocorrelation Coefficient)**.** The *partial correlation* between y_t and y_{t-k} is the simple correlation between y_{t-k} and y_t minus that part explained linearly by the intervening lags, i.e.

$$\rho(k)^* = Corr[y_t - E^*(y_t|y_{t-1}, \ldots, y_{t-k+1}), y_{t-k}]$$

where $E^*(y_t|y_{t-1}, \ldots, y_{t-k+1})$ is the minimum mean-squared error predictor of y_t by $y_{t-1}, \ldots, y_{t-k+1}$. (Greene, 2011:724) [42][11]

The function $E^*(.c)$ may be the linear regression if the conditional mean was linear; but it may not. The optimal *linear* predictor is the linear regression, so we have

$$\rho(k)^* = Corr[y_t - \beta_1 y_{t-1} - \beta_2 y_{t-2} - \cdots - \beta_{k-1} y_{t-k+1}, y_{t-k}]$$

where

$$\begin{aligned}\boldsymbol{\beta} &= [\beta_1, \beta_2, \ldots, \beta_{k-1}] \\ &= \{Var[y_{t-1}, y_{t-2}, \ldots, y_{t-k+1}]\}^{-1} \times Cov[y_t, (y_{t-1}, y_{t-2}, \ldots, y_{t-k+1})]'\end{aligned}$$

So, this equation may be seen to be a vector of regression coefficients and we are computing the correlation between a vector of residuals and y_{t-k}. There are many ways to compute this but let us focus on one equivalent definition for computation.

Definition 4.67 (Partial Autocorrelation Coefficient)**.** The *partial correlation* between y_t and y_{t-k} is the last coefficient in the linear projection of y_t on $[y_{t-1}, y_{t-2}, \ldots, y_{t-k}]$,

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{k-1} \\ \rho(k)^* \end{bmatrix} = \begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(k-2) & \gamma(k-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(k-3) & \gamma(k-2) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma(k-1) & \gamma(k-2) & \cdots & \gamma(1) & \gamma(0) \end{bmatrix}^{-1} \begin{bmatrix} \gamma(1) \\ \gamma(2) \\ \vdots \\ \gamma(k) \end{bmatrix}$$

[11] The minimum mean square error predictor or minimum mean square estimator (MMSE) is the optimal predictor that we will define when we study forecasting in chapter 5; see definition 5.1 together with the proof and the discussion that follows it.

Example 4.68. Consider the AR(p) model

$$y_t = \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \cdots + \gamma_p y_{t-p} + \epsilon_t$$

We want the last coefficient in the projection of y_t on y_{t-1}, then on (y_{t-1}, y_{t-2}), etc. The first will be the simple regression coefficient of y_t on y_{t-1}:

$$\rho_1 * = \frac{Cov(y_t, y_{t-1})}{Var(y_{t-1})} = \frac{\lambda_1}{\lambda_0} = \rho_1$$

Note that the first partial autocorrelation coefficient *for any process* equals the first autocorrelation coefficient. For AR(p), $\rho_1 *$ will be a mixture of all the γ coefficients and when $p = 1$, $\rho_1 * = \rho_1 = \gamma$. For AR(p), the last coefficient in the projection on p lagged values is γ_p and any additional partial autocorrelation must be zero since $k > p \Longrightarrow \rho_k * = Corr(\epsilon_t, y_{t-k}) = 0$. So, for AR(p), ACF ρ_k gradually decays to zero – monotonically if characteristic roots are real or like a sinusoidal pattern if they are complex. We can add to this now that the PACF $\rho_k *$ will be irregular up to lag p and then the PACF will suddenly, permanently drop to zero.

Example 4.69. Recall that for an MA(q) process, the ACF has q irregular spikes and then it drops to zero and remains there. To find the PACF, first write the model as

$$y_t = (1 - \theta_1 L - \theta_2 L^2 - \cdots - \theta_q L^q)\epsilon_t$$

Assuming that the series is invertible:

$$\frac{y_t}{1 - \theta_1 L - \cdots - \theta_q L^q} = \epsilon_t$$

$$\begin{aligned} y_t &= \pi_1 y_{t-1} + \pi_2 y_{t-2} + \cdots + \epsilon_t \\ &= \sum_{i=1}^{\infty} \pi_i y_{t-i} + \epsilon_t \end{aligned}$$

Since the AR form of the MA(q) process has an infinite number of terms, the PACF will not drop to zero as it does for the AR process. Instead, the PACF of an MA process will be similar to the ACF of an AR process. For the MA(1), $y_t = \epsilon_t - \theta \epsilon_{t-1}$, the AR representation is

$$y_t = \theta y_{t-1} + \theta^2 y_{t-2} + \cdots + \epsilon_t$$

This is similar to the AR(∞). So, the PACF of an MA(1) process is identical to the ACF of an AR(∞) process, i.e. $\rho_k * = \theta^k$.

Example 4.70. The ARMA(p,q) will have its ACF and PACF as mixtures of the two forms we have already discussed. This follows since ARMA(p,q) is a mixture of AR and MA processes. Normally, the ACF of an ARMA process will

have a few notable spikes in the early lags, which correspond to the number of MA terms and thereafter they will correspond to the smooth pattern of the AR part of the model. Note that high-order MA process are uncommon and high-order AR processes ($p > 2$) generally are the result of nonstationary processes; the 'workhorses of the applied literature' – for stationary processes – are (2,0) and (1,1) processes. An ARMA(1,1) will have an ACF and PACF that both exhibit a distinctive spike at lag 1 and then decay exponentially after the first lag.

4.4 Identification, Estimation, Testing and Forecasting

In this section, we will present the Box-Jenkins analysis, which is a popular method for univariate time-series modeling and forecasting. We can break the method down into five steps:

1. Assessing if the series are stationary or nonstationary and making suitable data transformations to induce stationarity if the data illustrate otherwise.
2. Identifying an appropriate ARMA model – choosing p and q for ARMA(p, q).
3. Estimating the model parameters.
4. Testing – model diagnostics.
5. Forecasting or repeating steps 2 & 3.

4.4.1 Checking for stationarity

A first pass to check for stationarity is visual inspection of plots. Are there trends? Is heteroscedasticity apparent? These should be fairly straightforward to see in the data but beware that, in general, casual inspection of graphs is no substitute for formal testing.[12] The *correlogram*, which plots sample autocorrelations against lag length is very useful for this purpose. Recall that the autocorrelation function was defined as

$$\rho(y_t, y_{t-j}) = \frac{Cov(y_t, y_{t-j})}{Var(y_t)} \quad j = 1, 2, \ldots$$

[12] Inevitably, when choosing what topics to cover within a relatively short course, trade-offs have to be made and as this course is centered around stationary time series, we will not explore the topic of nonstationarity. Furthermore, most work on nonstationarity and cointegration was carried out during the 1980s and we will be covering even more recent developments in the area of simulation at the very end of these lectures in the final chapter. For now, note the existence of different tests of 'unit roots' (when the AR coefficient is 1) such as Dickey-Fuller and its extensions; also recall the tests from last term for heteroscedasticity. See chapter four in Enders (2008) [27], section 6.1 of Dave & DeJong (2011) [21] and chapter 5 in Harvey (1993) [46]; for the (more) adventurous, take a look at a classic paper on this by Sims & Uhlig (1991) [77] and for the (even more) adventurous, chapter six in Bauwens, Lubrano & Richard (1999) [5] presents a Bayesian perspective on unit root inference.

The sample autocorrelation function is merely the sample counterpart of this by the analogy principle:

$$r_j = \frac{\sum_t (y_t - \bar{y})(y_{t-j} - \bar{y})}{\sum_t (y_t - \bar{y})^2} \quad j = 1, 2, \ldots$$

The correlogram of a stationary series drops off as the number of lags becomes large, but this does not usually happen for a nonstationary series. Usually a correlogram that declines linearly implies that the underlying process is nonstationary; for example, the mean and the variance could have changed. If nonstationarity is found, possible transformations include logarithms to stabilise the variance since outliers get squeezed and first-differences to remove trends. If there are only one or two outliers, we are not doing too badly. Let us look at the latter transformation in example 4.71.

Example 4.71. One example of a nonstationary process is a deterministic trend model:

$$y_t = \alpha + \beta t + u_t$$

This is nonstationary since

$$E(y_t) = \alpha + \beta t$$

This is an easy model to transform into a stationary one. At $t - 1$:

$$y_{t-1} = \alpha + \beta(t - 1) + u_{t-1}$$

$$y_t - y_{t-1} = \Delta y_t = \beta + (u_t - u_{t-1})$$

so the time trend disappears and assuming that $E(u_t) = 0$, we have that $E(\Delta_t) = \beta$. The process is now *difference-stationary*. So, the deterministic trend model can be made stationary by *first-differencing*, which is one example of a transformation that renders certain nonstationary models stationary by removing trends. Higher order trends can be removed by modifying the first-differencing procedure, e.g. by differencing n times.

Other methods to induce stationarity include detrending (in which case we say the resulting series is *trend stationary*) and the use of filters to separate the trend from the cycle such as the Hodrick-Prescott (HP) filter. We will return to the HP filter in the final chapter.[13] If the series displays heteroscedasticity, we can first take logarithms and then difference if there are still trends. We typically plot the correlogram of y_t and successive differences Δy_t, $\Delta^2 y_t$, etc., look at the correlograms at each stage and continue differencing until the correlogram dampens. The very concept of the Box-Jenkins methodology is ARIMA where 'I' stands for 'integrated'. An ARIMA(p, d, q) differenced d times becomes an ARMA(p, q) process, i.e. an ARIMA(p, d, q) is nonstationary, integrated of order d (denoted $I(d)$), while an ARMA(p, q) is stationary, integrated of order zero (denoted $I(0)$).

[13] For those interested in the above mentioned methods, see section 6.1 in Dave & DeJong (2011) [21]. Of course, seasonal peaks in series may indicate nonstationarity as well and other methods are involved to deseasonalise the data; see section 6.2 in Dave & DeJong (2011) [21].

4.4.2 Identification

We identify ARMA models through the use of the correlogram and the partial autocorrelation function by matching the sample correlogram with the theoretical autocorrelation function of a specific AR, MA or AMRA process:

1. If the correlogram cuts off at lag $j = q$, then assume that the series follows an MA(q) process.
2. If the correlogram does not cut off, then assume that the process is AR or ARMA.
3. If the correlogram does not cut off, but the PACF cuts of at lag $j = p$, then assume that the series follows an AR(p) process.
4. If neither the correlogram nor the PACF cuts off, then assume that the series follows an ARMA process. However, it is almost impossible to figure out the values for p and q. Generally, we take a simple version first or a few combinations and assess the model through the diagnostic step and / or the forecasting step of Box-Jenkins.

4.4.3 Estimation

4.4.3.1 MA models

Example 4.72 (MA(1))**.** Let us take the example of the MA(1) model:

$$y_t = \mu + \theta_1 \epsilon_{t-1} \epsilon_t \tag{4.61}$$

It is important to remember that this differs from regression equations because nothing on the right-hand side is known. We neither know μ nor the random shocks. Therefore, we cannot estimate (4.61) by regression. Instead, we estimate MA models by maximum likelihood estimation (MLE). Note that the autocorrelation function we derived earlier for the MA(1) process is

$$\rho_1 = \frac{\theta_1}{1 + \theta_1^2}$$

and consider the sample autocorrelation, where we use $\hat{\theta}_1$ to denote our estimate of θ_1:

$$\begin{aligned} r_1 &= \frac{\hat{\theta}_1}{1 + \hat{\theta}_1^2} \\ \implies r_1 + r_1 \hat{\theta}_1^2 &= \hat{\theta}_1 \\ \iff r_1 \hat{\theta}_1^2 - \hat{\theta}_1 + r_1 &= 0 \end{aligned}$$

There are two solutions for this quadratic equation in $\hat{\theta}_1$:

$$\hat{\theta}_1^2 = \frac{1 \pm \sqrt{1 - 4r_1^2}}{2r_1}$$

We need to choose one of these two solutions. Note that any root that is greater than one puts more weight on the past and we generally rule this out; we also generally rule out any root above 0.5. For MLE, see also appendix 2.1 of Enders (2008) [27]. Note that the formula becomes more complicated for MA(q) processes but computers handle ML with MA(q) with ease.

4.4.3.2 AR models

Example 4.73 (AR(p))**.** Unlike with MA models, the AR(p) is a legitimate regression equation since we know the past values of the y's on the right-hand side

$$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

and therefore we may use OLS, which is asymptotically valid. As long as the disturbances ϵ_t are not autocorrelated, OLS estimates of the ϕ parameters are consistent, even though the regressors are lagged dependent variables. See section 2.7 in Enders (2008) [27].

4.4.3.3 ARMA models

Example 4.74 (ARMA(p, q))**.** Due to the MA component in ARMA models, estimation of say an ARMA(p, q) model

$$y_t = \mu + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$$

is much more complicated and requires ML estimation. Computers do this with ease. See section 2.7 in Enders (2008) [27].

4.4.4 Diagnostics

We might first check how well the estimated models fit the data, e.g. R^2 and the ability of the estimated models to track it, i.e. find turning points. We can also check the randomness of the residuals; if the residuals are not found to be random, then the model is inadequate. For residuals, a first check is to plot and inspect the residuals over time (a time plot). Then we could plot and inspect a correlogram of residuals, which should be one at zero and zero for any lag length after that. With these atheoretic residual checks, we could finally use a test statistic such as the Box-Pierce Q statistic, which tests whether the first M autocorrelations for residuals are significantly different from zero. Under the null hypothesis, autocorrelations for residuals are asymptotically Normal with mean zero and variance $\frac{1}{N}$ where N is the sample size:

$$Q = N \sum_{j=1}^{M} r_j^2 \sim \chi_M^2$$

Thus, if $Q > \chi_M^2$, then we should reject the null hypothesis that the residuals are random and that the model is adequate, at the α-significance level.

4.4.5 Forecasting

Milton Friedman's pragmatic philosophy (1953) [33] asserted that the ultimate test of a model was in its ability to forecast. In this subsection, we will consider forecasting with MA models, AR models and ARMA models.

4.4.5.1 Forecasting with MA models

Example 4.75 (MA(1))**.** A zero mean $\mu = 0$ MA(1) model is written

$$y_t = \epsilon_t + \theta_1 \epsilon_{t-1}$$

We need to generate observations from what is unknown. We can convert MA to AR if we have a stability assumption that $|\theta_1| < 1$:

$$\begin{aligned} \epsilon_N &= y_N - \theta_1 \epsilon_{N-1} \\ &= y_N - \theta_1 (y_{N-1} - \theta_1 \epsilon_{N-2}) \\ &= y_N - \theta_1 y_{N-1} + \theta_1^2 \epsilon_{N-2} \end{aligned}$$

We will follow this technique by considering the one-period-ahead forecast for an MA(1):

$$\hat{y}_{N+1} = \hat{\epsilon}_{N+1} + \hat{\theta}_1 \hat{\epsilon}_N$$

where

$$\begin{aligned} \hat{\epsilon}_N &= y_N - \hat{\theta}_1 y_{N-1} + \hat{\theta}_1^2 y_{N-2} - \cdots + (-\hat{\theta}_1)^k y_{N-k} \qquad (4.62) \\ \hat{\epsilon}_{N+1} &= E(\epsilon_{N+1}) = 0 \\ \therefore \hat{y}_{N+1} &= \hat{\theta}_1 \hat{\epsilon}_N \end{aligned}$$

Similarly, the two-period-ahead forecast will be

$$\begin{aligned} \hat{y}_{N+2} &= \hat{\epsilon}_{N+2} + \hat{\theta}_1 \hat{\epsilon}_{N+1} \\ &= E(\epsilon_{N+2}) + \hat{\theta}_1 E(\epsilon_{N+1}) \\ &= 0 \end{aligned}$$

We could also estimate $\hat{\epsilon}_{N+1}$ by adapting (4.62) and using $\hat{y}_{N+1}$ as the first right-hand side term. So, the two-period-ahead forecast for an MA(1) will be zero, which reflects the short memory of MA processes, which display rapid convergence to the mean (*mean reversion*), i.e. there is rapid reversion of forecasts to the mean value 0 (μ more generally).

Example 4.76 (MA(2) with $\mu = 0$)**.** Using $\hat{\epsilon}_{N+k} = 0,\ k = 1, 2, 3$, the one-, two- and three-period-ahead forecasts are given by:

$$\begin{aligned} \hat{y}_{N+1} &= \hat{\epsilon}_{N+1} + \hat{\theta}_1 \hat{\epsilon}_N + \hat{\theta}_2 \hat{\epsilon}_{N-1} = \hat{\theta}_1 \hat{\epsilon}_N + \hat{\theta}_2 \hat{\epsilon}_{N-1} \\ \hat{y}_{N+2} &= \hat{\epsilon}_{N+2} + \hat{\theta}_1 \hat{\epsilon}_{N+1} + \hat{\theta}_2 \hat{\epsilon}_N = \hat{\theta}_2 \hat{\epsilon}_N \\ \hat{y}_{N+3} &= \hat{\epsilon}_{N+3} + \hat{\theta}_1 \hat{\epsilon}_{N+2} + \hat{\theta}_2 \hat{\epsilon}_{N+1} = 0 \end{aligned}$$

Note that the three-period-ahead forecast says nothing due to short memory – any $q > 2$ periods ahead; the three-period-ahead forecast is essentially the mean 0.

4.4.5.2 Forecasting with AR models

Example 4.77 (AR(1))**.** The one- to j-period ahead forecasts are given by:

$$\begin{aligned}
\hat{y}_{N+1} &= \hat{\mu} + \hat{\phi}_1 y_N + \overbrace{\hat{\epsilon}_{N+1}}^{\text{WN shock}} = \hat{\mu} + \hat{\phi}_1 y_N \\
\hat{y}_{N+2} &= \hat{\mu} + \hat{\phi}_1 \hat{y}_{N+1} + \hat{\epsilon}_{N+2} \\
&= \hat{\mu} + \hat{\phi}_1 \hat{y}_{N+1} = \hat{\mu} + \hat{\phi}_1 \hat{\mu} + \hat{\phi}_1^2 y_N \\
\hat{y}_{N+j} &= \sum_{i=0}^{j-1} \hat{\phi}_1^i \hat{\mu} + \hat{\phi}_1^j y_N
\end{aligned}$$

Note the use of the zero mean for forecasting future disturbances. If the process is stationary, i.e. if $|\phi_1| < 1$, then in the limit as $j \longrightarrow \infty$, we get the mean of the AR(1) process: $\frac{\hat{\mu}}{1-\hat{\phi}}$, i.e. $\frac{\hat{\mu}}{1-\hat{\phi}}$ is the mean, not $\hat{\mu}$. In contrast to the MA model, AR models have something to say about the long-term future because AR processes have long memories. While the correlation reduces, we do not reach zero, which is the residual memory. So, there is a slower tendency for forecasts to settle down to a given value (mean), depending on the size of $\hat{\phi}_1$.

Example 4.78 (AR(2))**.** The one-period-ahead forecast for the AR(2) model is given by

$$\hat{y}_{N+1} = \hat{\mu} + \hat{\phi}_1 y_N + \hat{\phi}_2 y_{N-1} + \hat{\epsilon}_{N+1} = \hat{\mu} + \hat{\phi}_1 y_N + \hat{\phi}_2 y_{N-1}$$

There is more complexity involved in $\hat{y}_{N+1}$ and the general forecast $\hat{y}_{N+j}$, which are left as optional exercises.

4.4.5.3 Forecasting with ARMA models

Example 4.79 (ARMA(1,1))**.** For the ARMA(1,1), the one-period-ahead forecast is

$$\begin{aligned}
\hat{y}_{N+1} &= \overbrace{\hat{\mu} + \hat{\phi}_1 y_N}^{\text{AR component}} + \overbrace{\hat{\epsilon}_{N+1} + \hat{\theta}_1 \hat{\epsilon}_N}^{\text{MA component}} \\
&= \hat{\mu} + \hat{\phi}_1 y_N + \hat{\theta}_1 y \hat{\epsilon}_N
\end{aligned}$$

since we do not know anything about future random shocks. The two-period-ahead forecast is left as an optional exercise. While forecasts further into the future are more complicated, in the distant future ARMA behaves like an AR model, reverting to its mean as the forecast horizons increase (forecasting further and further ahead). This follows because $\hat{\phi}_1^j \longrightarrow 0$ as $j \longrightarrow \infty$ when $|\hat{\phi}_1| < 1$. So, like MA models, forecasts from AR and ARMA models tend towards some given value as the forecast horizon increases. However, they do so more slowly.

Remark 4.80. A caveat is in order here: the above discussion assumes that these models remain valid into the future, but typically these models are only good as short-term forecasting models since parameters change over time. Forecasting can be hazardous. The rule of thumb is that we should weight up more factors than simply what our forecast model 'churns-out'. We will return to the issue of forecasting in more detail in chapter 5.

Chapter 5

Forecasting

5.1 Optimal Prediction

Optimal predictions in ARMA models tend to be conducted recursively. A method for predictions in unobserved components models is discussed in Harvey (1993) chapter 3, which involves the state space form [46]. We can also use the state space form to make exact finite sample predictions about ARMA models given general assumptions on initial conditions. Additionally, though less important for this course, we can use the state space form to obtain the general solution to optimal estimation of unobserved components with finite samples, which is known as *signal extraction* – 'picking out the message from the noise'.

Definition 5.1. The *optimal predictor* l steps ahead, given a set of observations on y_t up to and including y_T, is the expected value of y_{T+l} conditional on the information available at time T:

$$\tilde{y}_{T+l|T} = E(y_{T+l}|Y_T) = E_T(y_{T+l})$$

where Y_T is the information set $\{y_T, Y_{T-1}, \ldots\}$ and E_T denotes expectation conditional on Y_T. The expectation is 'optimal' since it minimises the mean square error.

Proof. WTS: $\tilde{y}_{T+l|T}$ has minimum mean square error (minimum MSE). Observe that for any predictor $\hat{y}_{T+l|T}$ based on information up to and including time T, the estimation error can be split into two parts as follows:

$$y_{T+l} - \hat{y}_{T+l|T} = [y_{T+l} - E(y_{T+l}|Y_T)] + [E(y_{T+l}|Y_T) - \hat{y}_{T+l|T}]$$

The second term on the RHS is fixed at T, so squaring the entire expression and taking expectations (at time T), the cross product term vanishes. So:

$$MSE(\hat{y}_{T+l|T}) = Var(y_{T+l}|Y_T) + [\hat{y}_{T+l|T} - E(y_{T+l}|Y_T)]^2$$

Note that the first term on the RHS, i.e. the conditional variance of y_{T+l} is independent of $\hat{y}_{T+1|T}$. The expression is minimised when $\hat{y}_{T+l|T} = \tilde{y}_{T+l|T}$.

Therefore, the *minimum mean square estimate* (MMSE) – or the *minimum mean squared estimator* (also abbreviated (MMSE)) if we view it as a rule rather than a number – of y_{T+l} is the conditional mean as in the definition of the optimal predictor above and furthermore, it is unique. □

When the optimal predictor is treated as a rule as opposed to a number, it is an estimat*or* of y_{T+l} rather than an estimat*e*. ICBST it is the MMSE since it minimises the MSE when the expectation is taken over all observations in the information set, following from LIE. When observations are normally distributed, the $Var(y_{T+l}|Y_T)$ is independent of the observations and therefore it can directly be interpreted as the MSE of the estimator; however, normality is not necessary for this property of the MSE as we shall see soon. Generally, the MMSE of a random variable is the expectation of the variable conditional on the relevant information set. Conditional expectations do not necessarily have to be linear combinations of the observations. If we restrict to the class of linear estimators, we encounter the familiar best estimator, the MMSLE (L for linear) or BLUE / BLUP (P for predictor) as you may have encountered in introductory econometrics courses.

Let us now concentrate on optimal predictors of the future through MMSEs for ARMA models. Assume the ARMA process is stationary and invertible with known parameters and independent, zero mean, constant variance (σ^2) disturbances. Assume also that for MA and mixed processes, $\epsilon_T, \epsilon_{T-1}, \dots$ are all known, i.e. we know present and all past disturbances, which is equivalent to assuming an infinite realisation of observations going backwards in time. While this is an unrealistic assumption, it can be modified to produce a predictor calculated from a finite sample. At time $T+l$, the ARMA(p,q) is:

$$y_{t+l} = \phi_1 y_{T+l-1} + \cdots + \phi_p y_{T+l-p} + \epsilon_{T+l} + \cdots + \theta_q \epsilon_{T+l-q}$$

The MMSE of a future observation as that above is its expectation conditional on all information up to time T as we described in the definition of the optimal predictor. Taking conditional expectations of the above equation, recognizing that $\epsilon_t = 0$ for all future values since they cannot be predicted as they are independent, we get the following:

$$\tilde{y}_{T+l|T} = \phi_1 \tilde{y}_{T+l-1|T} + \cdots + \phi_p \tilde{y}_{T+l-p|T} + \tilde{\epsilon}_{T+l|T} + \cdots + \theta_q \tilde{\epsilon}_{T+l-q|T},\ l = 1, 2, \dots \tag{5.1}$$

where $\tilde{y}_{T+j|T} = y_{T+j}$ for $j \leq 0$ and

$$\tilde{\epsilon}_{T+j|T} = \begin{cases} 0 & j > 0 \\ \epsilon_{T+j} & j \leq 0 \end{cases}$$

Equation (5.1) provides a recursion for calculating optimal predictions.

Example 5.2. With AR(1), (5.1) yields the following difference equation:

$$\tilde{y}_{T+l|T} = \phi \tilde{y}_{T+l-1|T} \quad l = 1, 2, \dots$$

with starting value $\tilde{y}_{T|T} = y_T$ and so we can solve this difference equation to get

$$\tilde{y}_{T+l|T} = \phi^l y_T \tag{5.2}$$

So, the predicted values decay exponentially towards zero and the forecast function has the same form as the autocovariance function.

Example 5.3. An MA(1) at time $T+1$ can be expressed as

$$y_{T+1} = \epsilon_{T+1} + \theta\epsilon_T$$

The prediction equation makes use of the fact that $\epsilon_{T+1} = 0$ since it is unknown at time T:

$$\tilde{y}_{T+1|T} = \theta\epsilon_T$$

The prediction equation for $T+l$ where $l > 1$ is $\tilde{y}_{T+l|T} = 0$. Therefore, knowledge of the data generating process (DGP) is unhelpful in predicting at horizons greater than one period ahead for the MA(1) model.

Example 5.4. One example of an ARMA(2,2) process is:

$$y_t = 0.6y_{t-1} + 0.2y_{t-2} + \epsilon_t + 0.3\epsilon_{t-1} - 0.4\epsilon_{t-2}$$

Suppose $y_T = 4.0, y_{T-1} = 5.0, \epsilon_T = 1.0$ and $\epsilon_{T-1} = 0.5$. Then

$$\begin{aligned}
\tilde{y}_{T+1|T} &= 0.6y_T + 0.2y_{T-1} + 0.3\epsilon_T - 0.4\epsilon_{T-1} = 3.5 \\
\tilde{y}_{T+2|T} &= 0.6\tilde{y}_{T+1|T} + 0.2y_T - 0.4\epsilon_T = 2.5 \\
\tilde{y}_{T+l|T} &= 0.6\tilde{y}_{T+l-1|T} + 0.2\tilde{y}_{T+l-2|T} \quad l \geq 3
\end{aligned}$$

Splitting the MA(∞) representation of y_t into

$$y_{T+l} = \sum_{j=1}^{l} \psi_{l-j}\epsilon_{T+j} + \sum_{j=0}^{\infty} \psi_{l+j}\epsilon_{T-j} \tag{5.3}$$

Taking conditional expectations yields the MMSE of the forecast y_{T+l}:

$$\tilde{y}_{T+l|T} = \sum_{j=0}^{\infty} \psi_{l+j}\epsilon_{T-j} \tag{5.4}$$

So, the first term on the RHS of (5.3) is the error in predicting l steps ahead and its variance is the prediction MSE:

$$MSE(\tilde{y}_{T+l|T}) = (1 + \psi_1^2 + \cdots + \psi_{l-1}^2)\sigma^2 \tag{5.5}$$

Note that this is independent of the observations so it is the unconditional MSE, i.e. the MSE averaged over all possible realisations of observations rather than the estimator given by (5.4).

Example 5.5. With the AR(1) model, $\psi_{l+j} = \phi^{l+j}$, so:

$$\tilde{y}_{T+l|T} = \sum_{j=0}^{\infty} \phi^{l+j}\epsilon_{T-j} = \phi^l \sum_{j=0}^{\infty} \phi^j \epsilon_{T-j} = \phi^l y_T$$

just as in (5.2).

$$MSE(\tilde{y}_{T+l|T}) = [1 + \phi^2 + \cdots + \phi^{2(l-1)}]\sigma^2 = \frac{1-\phi^{2l}}{1-\phi^2}\sigma^2$$
$$\overset{l\longrightarrow\infty}{\longrightarrow} \frac{\sigma^2}{1-\phi^2} = Var(y_t)$$

Let us now add the assumption that ϵ_t's are normally distributed; thus, the conditional distribution of y_{T+l} is also normal. A 95% prediction interval for y_{T+l} is

$$y_{T+l} = \tilde{y}_{T+l|T} \pm 1.96 \left(1 + \sum_{j=1}^{l-1} \psi_j^2\right)^{\frac{1}{2}} \sigma$$

The correct interpretation of this prediction interval is the following: for a given sample, there is a 95% chance that y_{T+l} will lie within the interval.

As mentioned earlier, the assumption that all present and past disturbances are known for MA and mixed process means that we have a knowledge of ϵ's into the infinite past, which is unrealistic. Alternatively, we could place assumptions on the initial conditions permitting recursive computation of the values for the required disturbances. However, the predictions will still not be optimal if these assumptions are violated; see Harvey 3.3 (1993) [46]. See Harvey 4.4 for a method of computing exact optimal finite sample predictions without making these assumptions [46].

While the assumption that the disturbances are independent is necessary for deriving the MMSE of a future observation in an ARMA model, if the assumption is relaxed from independence to uncorrelatedness, then the result does not necessarily hold since the conditional expectation of future disturbances are not necessarily zero. However, restriction attention to linear predictors, the predictor given by the RHS of (5.4) is still the best in that it minimises the unconditional prediction MSE, i.e. it is the MMSLE or BLUP.

Definition 5.6. A *linear predictor* is a linear function of the observations and so is a linear function of disturbances, past and present and is written as

$$\hat{y}_{T+l|T} = \sum_{j=0}^{\infty} \psi^*_{l+j}\epsilon_{T-j}$$

where ψ^*_{l+j} are pre-specified weights.

The unconditional expectation of the prediction error is given by

$$y_{T+l} - \hat{y}_{T+l|T} = \epsilon_{T+l} + \psi_1\epsilon_{T+l-1} + \cdots + \psi_{l-1}\epsilon_{T+1} + (\psi_l - \psi^*_l)\epsilon_T + (\psi_{l+1} - \psi^*_{l+t})\epsilon_{T-1} + \cdots$$

When the unconditional expectation of the prediction error is zero, the predictor is unbiased.

$$MSE(\hat{y}_{T+l|T}) = \sigma^2(1 + \psi_1^2 + \cdots + \psi_{l-1}^2) + \sigma^2 \sum_{j=0}^{\infty} (\psi_{l+j} - \psi_{l+j}^*)^2$$

which is minimised when $\psi_{l+j}^* = \psi_{l+j}$ and so the MMSLE(y_{T+l}) is given by (5.4) with MSE as in (5.5).

5.2 Forecast Assessment

The outline of this section is as follows. Firstly, I will present a case for using forecast assessment tools before briefly giving a revision of some forecasting basics. Next I will discuss how to estimate parameters for forecasting. Subsequent material focuses on two areas in forecast assessment representing the bulk of this first topic in the chapter, *viz.* (i) comparing forecasts and forecasters to determine which forecast is optimal and (ii) comparing forecasts and models, i.e. using forecast assessment tools to learn about models.

5.2.1 Motivation

Perhaps the first question that may be asked is the following: should we care to study forecast assessment tools, for instance, especially if we are not involved in forecasting? The simple answer is yes, we should care. Not only are forecasting assessment tools appropriate for when we want to know about the future, but they are also very useful as model diagnostics. Recent advances in the research on forecast assessment include methods that focus on both goals, forecasting to know about the future and forecasting as a diagnostic for model evaluation. These two motivations will be important for the remainder of section 5.2.

Forecasts can be helpful when we want to know about or predict the future. Financial markets place great emphasis on the quality of forecasts. The existence of forecasting companies across the world and the various Survey of Professional Forecasters (SPF), e.g. in the US with the oldest quarterly survey of macroeconomic forecasts and in Europe in the ECB testify to the demand for quality forecasts.

In addition to the use of forecasts for trying to predict the future, we may want to evaluate models in terms of the following:

H_0: model is correct

H_A: specific alternative

However, the question remains as to why we should use forecasting methods to evaluate models. The following three points serve to answer this question.

1. Forecasting methods tend to suffer less from problems such as in-sample overfitting and that of data mining. The first point is particularly true

since having a good fit in the sample (in-sample overfitting) tends to hurt you out-of sample. Forecasting being conducted out-of sample means that data mining tends to be penalised.

2. Forecasting assessment methods help to deal with the problem of instability. They can be used to check whether the model stable through time.

3. Sometimes in-sample methods may be very complicated in which case forecast methods can be computationally convenient.

5.2.2 Forecasting 101

5.2.2.1 Terminology

Let Y_{t+h} denote the variable we want to forecast at horizon h. Let X_t be the vector of variables that we will use in making the forecast – this may include such variables as lags of Y_t and possibly other variables too. Let $f_{t+h|t}$ denote the forecast of Y_{t+h} that we make at time t. Allow $e_{t+h} = Y_{t+h} - f_{t+h|t}$ to represent the 'forecast error' and $L(e) = \mathcal{L}(Y - f)$ to be the loss function that is associated with this error. Finally, the 'risk' associated with forecast f is written as $E(\mathcal{L}(e)) = E(\mathcal{L}(Y - f))$.

5.2.2.2 Minimum MSE forecasts

With one forecast, let us assume that the loss function is *quadratic*, i.e.

$$\mathcal{L}(e) = a + be^2$$

This will imply that risk will be the *mean* square error (MSE).[1] So, we need to find the minimum MSE (MMSE) forecast, which will be given by the regression function:

$$f_{t+h|t} = E(Y_{t+h}|X_t)$$

This is an important result and the best function of the data will give the best forecast. The challenge is the figure out what $f_{t+h|t}$ actually is. With a small number of predictor variables (X), this is relatively simple. With a large number of predictor variables (X), this can be complicated and will be the subject of section 5.3.

Now I will present three properties of MMSE forecasts. Firstly, note that strict exogeneity implies week exogeneity:

$$E(e_{t+h}|X_t) = 0 \implies E(e_{t+h}X_t) = 0$$

So, we cannot predict forecast errors using the data we have. Secondly, if current and lagged values of the forecast variable Y_t are used to make forecasts

[1] Exercise: prove that a quadratic loss function implies that associated risk will be the mean square error.

of Y_t, then X_t implicitly incorporates present and past values of e_t; hence, $E(e_{t+h}e_t) = 0$, so

$$e_{t+h} \sim MA(h-1)$$

Thirdly, from the definition of forecast error in the terminology section above, it follows that

$$\begin{aligned} Y_{t+h} &= f_{t+h|t} + e_{t+h} \\ \implies \sigma_Y^2 &= \sigma_f^2 + \sigma_e^2 \quad \because Corr(f_{t+h|t}, e_{t+h}) = 0 \\ \therefore \sigma_Y^2 &\geq \sigma_f^2 \end{aligned} \tag{5.6}$$

Every optimal forecast must satisfy (5.6), else it is not the case that the forecast is optimal. This provides a useful diagnostic check for whether f is an optimal forecast of Y. Note that when σ_f^2 is low and σ_Y^2 is high, it can be difficult to forecast Y. Shortly, we will investigate the case when the loss function is not quadratic. In conclusion, with forecast assessment, we look at the forecast error $Y - f = e$ and see if this behaves as if it arose from an MMSE forecast. Since we know the properties of MMSE forecasts, we can adequately judge if e behaves as such.

What about the case of when we have more than one forecast? Combining forecasts is practically a good idea – remember our twin goals: when the future matters and using forecast assessment for model evaluation. Regarding notation, let f^1 and f^2 denote two forecasts of Y and allow consider a third forecast to be:

$$f^3 = \beta_0 + \beta_1 f^1 + \beta_2 f^2$$

What is critical here is answering the question as to how we should combine the first two forecasts to produce f^3, i.e. what values of β should we choose? We have seen that optimal forecasts correspond to regressions. MMSE forecasts are regressions and combined optimal forecasts will be the best MMSE forecasts ($E(Y|f^1, f^2)$), so β will be the population values from the linear regression of Y_{t+h} on $f^1_{t+h|t}$ and $f^2_{t+h|t}$; see Bates & Granger (1969) [4] and Granger & Ramanathan (1984) [41] for more. Obviously, extending the results to the case where $n \geq 2$ is trivial. For a more advanced treatment, see Timmermann (2006) [79] who adds complications such as correlated series and multiple series.

Before moving onto other loss functions, let us consider one practical issue with combining forecasts, *viz.* using estimates of the β's from sample regressions. To add to the list of the puzzles you will no doubt be studying this year, here we have the 'forecast combining puzzle'. Even with a moderately large number of forecasts to combine, forecasts constructed via estimated β's suffer poor performance and it has been noted in the literature (surveyed by Timmermann) that ad hoc averages such as sample means, medians, 'consensus' forecasts, etc. typically yield superior performance. We will return to this in section 5.3.

5.2.2.3 Other loss functions

Since $Risk(f) = E(\mathcal{L}(Y - f))$ and if $Y_{t+h}|X_t \sim N(\mu_{t+h|t}, \sigma^2_{t+h|t})$, it can be shown that the optimal forecast will be $f_{t+h|t} = \mu_{t+h|t} + \alpha(\sigma^2_{t+h|t})$.[2] Note that with the variance depending on time, the tails of the distribution can move. This result means that the optimal forecast *for any loss function* is given by a regression function and a function of the conditional variance. If the conditional variance is constant, then the best forecast is given by the regression function plus a constant determined by the variance, but this does not matter as much! The implication is that with conditional homoscedastic Gaussian distributions, optimal forecasts are MMSE forecasts, i.e. $\mu_{t+h|t}+$ constant. However, if we have conditional heteroscedasticity, then the constant is time-varying and the result looses its strength. See Granger (1969) [40] for the simple Gaussian case and Christoffersen & Diebold (1997) [15] for the conditionally Gaussian approach.

Delving further into forecast assessment with other loss functions, Elliott, Komunjer & Timmermann (2005) [26] consider a class of loss functions such as those that are non-quadratic, non-symmetric, etc.

$$\mathcal{L}(Y - f) = [\alpha + (1 - 2\alpha) \times \mathbf{1}(Y - f < 0)]|Y - f|^p$$

With the goal of making their tests for forecast efficiency more robust, the authors study properties of optimal linear forecasters $f_{t+h|} = \theta' X_t$. They characterise features features of forecast errors and then see if these features are satisfied in some regressions.

5.2.3 Estimation

This relatively short subsection deals with estimating parameters for use in forecasting models. Two questions may be asked:

1. What estimator should you use? MLE? Others?
2. Should you use real-time data?

We will consider each question in turn, starting with the first. Let us assume an AR(1) model for GDP growth, which has been found for most countries to be approximately true in the data, for instance with the autoregressive parameter of about 0.3:

$$y_t = \phi y_{t-1} + \epsilon_t \tag{5.7}$$

To achieve the goal of forecasting y_{t+2}, we set the optimal forecast $f_{t+2|t} = \beta y_t$ where $\beta = \phi^2$. As for methods to forecast y_{t+2}, one way is is to use the 'iterated' method of estimating ϕ from (5.7) and using

$$\hat{f}^{iterated}_{t+2|t} = \hat{\phi}^2 y_t$$

[2] Exercise: prove this claim.

Lag Length	Horizon			
	3	6	12	24
AR(4)	0.99	0.99	1.00	1.05
AR(12)	1.01	1.01	1.03	1.10
AR(BIC)	0.98	0.97	0.99	1.05
AR(AIC)	1.00	1.01	1.02	1.09

Table 5.1: Relative pseudo-out-of-sample MSE

while the alternative is to employ the 'direct' approach of estimating β from

$$y_t = \beta y_{t-2} + u_t$$

and using

$$\hat{f}^{direct}_{t+2|t} = \hat{\beta} y_t$$

The iterated approach will be preferred if the AR(1) model describes y_t since $\hat{\phi}$ is the MLE and is efficient under AR(1). However, if the model is misspecified, then it may not be the best to predict one period ahead even if we get good predictions for one period ahead. On the other hand the direct approach will be preferred if the model is misspecified since while $\hat{\beta}$ has larger variance than $p\hat{h}i$ under the correct specification, it is robust to misspecification (for the class of forecasts considered). As a rule of thumb, if misspecification is not too bad, the gains from variance reduction will dominate the losses from misspecification, in which case the iterated method is to be preferred.

The literature in this area ranges from Cox (1961) [20] to Schorfheide (2005) [76]. For an empirical comparison, see Marcelino, Stock & Watson (2006) [62] who use 170 monthly US macro series across the time frame from 1959-2002 and construct pseudo-out-of-sample forecasts (POOS). They describe their data with an AR and bivariate VAR for forecast period horizons $h = 3, 6, 12, 24$ months. Table 5.1 reports the relative POOS MSE, which is a ratio of sample MSE:

$$\text{Relative pseudo-out-of-sample MSE} \ = \frac{\sum \left(e^{direct}_{t+h}\right)^2}{\sum \left(e^{iterated}_{t+h}\right)^2}$$

Ratios above one indicate that misspecification is not important so the iterated method is favoured and *vice-versa* for ratios below one.

Note from the table that the Bayesian Information Criterion (BIC) tends to pick a small number of lags, so we may have a case of omitted variable bias in this case implying that misspecification is important. The direct method will be better in this case. However, across lag-length methods, the Akaike Information Criterion (AIC) iterated seems to work the best. So, the iterated approach is better for all other lag length selections than BIC and AIC leads to better forecasts than BIC.

Returning to the second question highlighted at the beginning of this subsection, should we use real time data or historical / revised data? There is no single answer to this question.

Let us consider issues regarding data revisions in y and x from:

$$y_{t+h} = \beta' x_t + u_{t+h}$$

Firstly, regarding y, we can ask the question as to what we actually want to forecast. For example, in 2008, do we want to forecast 2009 values that are announced in 2009 or do we want to forecast 2009 values when we have had time to go back to 2009 to see what really happened? That is, we need to decide whether our goal for forecasting is the first release or final release of variable(s). Secondly, regarding x, since we are conducting real-time forecasting using real-time data, it would appear that $x^{initial}$ should be used in regressions to estimate β. The crucial question is whether the projection of y on x is the same as the projection of y on $x^{initial}$, which is usually the object of interest. Letting

$$x = x^{initial} + x^{revision}$$

we may ask whether revision is 'news' or 'noise' in the dichotomous sense of Mankiw, Runkle & Shapiro (1984) [57]. That is, perhaps $x^{initial}$ and $x^{revision}$ may have different stochastic processes.

There are two approaches to answering this question. The first concerns treating the problem as an 'errors in variables bias'; instead of projecting y on z, we are actually projecting y on z^* where $z^* = z$ measured with error – 'attenuation bias'. These two projections are not the same. We can let $z = z^*$+noise, or *vice-versa*. The second approach is to use 2SLS since projecting y on x is bad even though x may be uncorrelated with the error term. Project x on z to get $\hat{x}$ and then project y on $\hat{x}$ to yield the 2SLS estimate, which will be consistent. However, one con with this approach that must be mentioned is that x has more variability. In conclusion, with errors in $x^{revision}$, if there is noise in the revision, then we should run regressions on $x^{initial}$, whereas if there is news in the revision, then we should run regressions on x^{final}.

5.2.4 Forecast Assessment

In the first subsection, we will look at evaluating forecasts and forecasters, but not models. In the second subsection, we will concentrate on forecast assessment tools in evaluating models using POOS forecasts.

5.2.4.1 Which forecast is optimal?

First we will look at Mincer-Zarnowitz (1969) [66] regressions:

$$Y_{t+h} = \alpha + \beta f_{t+h|t} + \gamma W_t + u_{t+h}$$

If $f_{t+h|t}$ is MMSE forecast, then $\alpha = 0$, $\beta = 1$ and $\gamma = 0$. In order to check of a forecast is optimal, we can test all of these properties or a subset of them. Three issues regarding inference are in order:

1. When the forecast horizon is greater than one period, i.e. $h > 1$, errors u_{t+h} are distributed as MA$(h-1)$, i.e. with serial correlation under the null hypothesis. This warrants the use of heteroscedasticity and autocorrelation consistent standard errors (HAC SEs).

2. If the forecast variable Y_t is persistent, e.g. interest rates, say integrated of order one $I(1)$, then the forecast f will also be persistent (i.e. $I(1)$) and so we are forced to deal with the problems associated with unit root regression inference. We can solve this issue rather easily by taking the difference: $(Y_{t+h} - Y_t) = \alpha + \beta(f_{t+h|t} - Y_t) + \gamma W_t + u_{t+h}$.

3. When we are forecasting far into the future (i.e. when h is large), the error u will be extremely persistent. Unfortunately, HAC SE work poorly here; see Richardson & Stock (1989) [69] for more details.

Moving on from Mincer regressions, with respect to combining or 'encompassing' regressions, let f^1 and f^2 be two forecasts as before. Then the forecast combining regression will be:

$$Y_{t+h} = \beta_0 + \beta_1 f^1_{t+h|t} + \beta_2 f^2_{t+h|t} + u_{t+h} \tag{5.8}$$

If f^1 is the MMSE forecast, then $\beta_0 = \beta_2 = 0$ and $\beta_1 = 1$. If we take f^1 and f^2 as coming from forecasters one and two, respectively, then when the first forecaster produces the MMSE forecast, the forecast from the second forecaster is useless.

Note that these forecast combining regressions will be subject to the same problems of inference as those faced with Mincer regressions. Also note that the above implication from regression (5.8) has three degrees of freedom, i.e. $\beta_0 = \beta_2 = 0$ and $\beta_1 = 1$. We can conduct more powerful tests by reducing the degrees of freedom to two if we impose the more parsimonious restrictions $\beta_0 = 0$ and $\beta_1 + \beta_2 = 1$. This constraint implies

$$Y_{t+h} - f^1_{t+h|t} = \beta_2(f^2_{t+h|t} - f^1_{t+h|t}) + u_{t+h}$$

We can then simply regress the forecast error, which is the left-hand side and do the test.

Next we shall investigate loss-function tests. Let f^1 and f^2 be two forecasts with forecast errors e^1 and e^2, respectively. We would like to compare risk for f^1 and f^2 – which forecast (if any) should be preferred:

$$E(\mathcal{L}(e^1)) \gtreqless E(\mathcal{L}(e^2))$$

Let us consider testing using only quadratic loss functions. Allowing e^1_t and e^2_t to be the realised forecast errors, we want to test whether

$$E[(e^1_t)^2] = E[(e^2_t)^2]$$
$$\text{i.e. } E[(e^1_t)^2 - (e^2_t)^2] = 0$$

To test this, we can use sample moments of differences:

$$\bar{d} = \frac{1}{T}\sum_{t=1}^{T} d_t$$

where $d_t = \{(e_t^1)^2 - (e_t^2)^2\}$ and check whether $\bar{d}$ is statistically significantly different from zero. As before, we encounter similar issues with inference such as when $h > 1$ – implying serial correlation – and again this can be treated via HAC standard errors. For further references, see Diebold & Mariano (1995) [23] and West (2006) [81] for a history.

Consider loss function tests with many competing forecasts. Let f^1 be the benchmark model (e.g. a random walk) and let f^k where $k = 1, \ldots, n$ are n competing models. We want to know whether *any* of the n competing models *dominate* the *benchmark*. Again, we can make use of sample moments of differences:

$$\bar{d}_k = \frac{1}{T}\sum_{t=1}^{T} d_t^k$$

where $d_t^k = \{(e_t^1)^2 - (e_t^k)^2\}$. So $\bar{d}_k$ is the sample MSE improvement from model k over the benchmark prediction. A sensible test statistic is the 'reality check' RC $= \max_k \bar{d}_k$. The 'reality' is that the benchmark model is the right one and we want to 'check' if it is actually the case that this assumption is correct. White (2000) [82] derives the limiting distribution of RC under the null hypothesis that the benchmark model is indeed optimal and in so doing, provides the critical values. Hansen (2005) [44] refines this in that perhaps maximising RC may not be the best way to proceed because if we instead throw out bad models, we will be maximising over a lower dimension, which will change our critical values.

Wrapping up this topic, we will finally consider density forecasts. A density forecast gives us the probability that something will be in a certain range. Sometimes forecasts are not point forecasts. However, we may want to question whether we have the correct density. Diebold, Gunther & Tay (1998) [22] provide an approach to evaluated density forecasts. First note that to generate a Normal number, typically random number generators generate uniform distributions on the $[0, 1)$ interval and then through a Normal distribution cdf, they can plug in the generated uniform numbers as arguments to get values that will be Normally distributed; see diagram in class and note that we can generate any distribution like this by replacing the Normal cdf with the cdf of whatever distribution we are interested in. Their key insight is as follows and essentially goes the other way. Assume Y has cumulative distribution function F, so $U = F(Y) \sim U[0, 1]$ – remember that random number generators often use $Y = F^{-1}(U)$. Therefore, if $F_{t+h|t}$ is in fact the conditional cumulative distribution function of $Y_{t+h|t}$, then $U_{t+h} = F_{t+h|t}(Y_{t+h})$ should be uniformly distributed $U[0, 1]$ and U_{t+h} should be independent of any data from period t and earlier. Interval forecasts give a 95% confidence interval for GDP growth in the final quarter of 2008 – see Christoffersen (1998) [16]. From graphs shown in

class, data obtained should be draws from the predictive distribution function, which should be monotonic. Then observe that the distribution is different date-by-date since plugging in realisations of the data into the density should yield uniform random variables (by an inverse transformation of the predictive cdf yielding a sequence of uniform random variables) that should be independent if $h = 1$ and have some dependence if $h > 1$; see graphs in class. So, these values should be uniform. For one step ahead, we care about the marginals and could implement say a *Kolmogorov-Smirnov* test for uniform distributions; this test requires quite a lot of data points to reject the null hypothesis.

Definition 5.7. Let F_n be the empirical distribution for n iid observations of the random variable X. The *KolmogorovŰSmirnov statistic* for a given cumulative distribution function $F(x)$ is

$$D_n = \sup_x |F_n(x) - F(x)|$$

The *Kolmogorov-Smirnov test* is defined as follows. If F is continuous, then under the null hypothesis that the sample comes from a hypothesised distribution $F(x)$

$$\sqrt{n}D_n \overset{n\longrightarrow\infty}{\longrightarrow} K$$

where K is the Kolmogorov distribution. The goodness-of-fit test or the KolmogorovŰSmirnov test is constructed by using the critical values of the Kolmogorov distribution. We can reject the null hypothesis at level α if

$$\sqrt{n}D_n > K_\alpha$$

where K_α is derived from

$$\Pr(K \leq K_\alpha) = 1 - \alpha$$

The asymptotic power of this test is 1.

5.2.4.2 Using forecast assessment tools to learn about models

Concentrating on forecast assessment in evaluating models using POOS, the most important references for this second subsection are West (1996 [80], 2006 [81]). The main statistical difference between this subsection and the first subsection is that now we are explicitly accounting for sampling variability in estimated model parameters. Continuing with the setup, model 1 has forecasts $f^1(\theta_1)$ and model 2 has forecasts $f^2(\theta_2)$. The forecasts to be evaluated are based on estimated models; hence, we use the hats: $f^1(\hat{\theta}_1)$ and $f^2(\hat{\theta}_2)$. Unlike earlier, now that we are dealing with estimated models, rather than true ones, we encounter extra sampling variability. We must now answer the question: does model 1 *that we do not know* forecast better than model 2 *that we do not know*. The highlighted parts of the previous sentence emphasise the difference from the first subsection.

First, let us consider the POOS forecasting strategy where the sample size $T = R + P$ (final P periods are used for 'prediction' – construction of POOS forecasts):

(i) Estimate θ from observations $1 : R$ to get $\hat{\theta}_R$ – this mimics real time.

(i) Forecast Y_{R+h} using data $1 : R$ and the estimate $\hat{\theta}_R$.

then **Recursive POOS**:

(iii) Estimate θ from observations $1 : R + 1$ to get $\hat{\theta}_{R+1}$.

(iv) Forecast Y_{R+1+h} using data $1 : R + 1$ and the estimate $\hat{\theta}_{R+1}$.

or **Rolling POOS** (convenient / useful in case of instability):

(iii) Estimate θ from observations $2 : R + 1$ to get $\hat{\theta}_{R+1}$.

(iv) Forecast Y_{R+1+h} using data $1 : R + 1$ and the estimate $\hat{\theta}_{R+1}$, which is the same as recursive POOS except that $\hat{\theta}_{R+1}$ will be different.

Conceptually important, here we are interested in:

$$E[\mathcal{L}(Y - f^1(\theta_1))] \gtreqless E[\mathcal{L}(Y - f^2(\theta_2))] \tag{5.9}$$

rather than what we will be interested in shortly:

$$E[\mathcal{L}(Y - f^1(\hat{\theta_1}))] \gtreqless E[\mathcal{L}(Y - f^2(\hat{\theta_2}))] \tag{5.10}$$

Be aware of the existence of cases for which the left-hand side is less than the right-hand side in (5.9) but exceeds the right-hand side (5.10). For more, see exchange rates as random walks for example, Engle & West (2005)[28], Clark & West (2006) [17] and Rossi (2006) [71]. Here we will focus on comparing risk, but note that similarly related issues emerge in combining tests.

One complication to be observed is that it is important to know whether models are nested or non-nested. With nested models, the random walk model 1 is a special case of the AR model 2 where:

$$\begin{aligned} &\text{Model 1: } y_{t+1} = x_t'\beta + \epsilon_{t+1} \\ &\text{Model 2: } y_{t+1} = x_t'\beta + z_t'\gamma + e_{t+1} \end{aligned}$$

Observe that when $\gamma = 0$, model 2 and model 1 are equivalent. With non-nested models, the random walk model 1 is not a special case of the AR model 2 since there is no way of getting model 1 from model 2, where models are defined as:

$$\begin{aligned} &\text{Model 1: } y_{t+1} = x_t'\beta + \epsilon_{t+1} \\ &\text{Model 2: } y_{t+1} = z_t'\gamma + e_{t+1} \end{aligned}$$

Remark: with nested models, the average of models would be nested, while in the case of non-nested models, the average of models would be nested but the other two models would be non-nested.

The following based on West (1996) [80] for non-nested models will not work in nested models. Note that we care about the loss for the true model. Define model 1 as

$$\text{Model 1: } y_{t+1} = x_t\beta + \epsilon$$

where x is a scalar, $h = 1$ step ahead (for clarity). Denote $x_t\beta$ as the true forecast and ϵ_{t+1} as the true forecast error. Also, denote $x_t\hat{\beta}_t$ as the estimated forecast and let the estimated forecast error be given by

$$Y_{t+1} - x_t\hat{\beta}_t = \hat{\epsilon}_{t+1} = \epsilon_{t+1} + x_t(\hat{\beta}_t - \beta)$$

Define model 2 as

$$\text{Model 2: } y_{t+1} = z_t\gamma + e_{t+1}$$

Model 1 and model 2 are non-nested. Denote $z_t\gamma$ as the true forecast and e_{t+1} as the true forecast error. Also, denote $z_t\hat{\gamma}_t$ as the estimated forecast and let the estimated error be given by

$$Y_{t+1} - z_t\hat{\gamma}_t = \hat{e}_{t+1} = e_{t+1} + z_t(\hat{\gamma}_t - \gamma)$$

We would like to know if the risk associated with ϵ is different from the risk associated with e, but we only observe estimates $\hat{\epsilon}$ and $\hat{e}$. It turns out that it is valid to conduct one of the tests using HAC instead of e's and ϵ's.

As for how we conduct loss function tests, let us first consider non-nested models. Before we used averages (over the prediction period) of $d_t = (\epsilon_t^2 - e_t^2)$. So, now that we are using estimates we must use $\hat{d}_t = (\hat{\epsilon}_t^2 - \hat{e}_t^2)$. To see how the sample averages of $\hat{d}_t$ and d_t are related, note that:

$$\begin{aligned}
\frac{1}{\sqrt{P}}\sum_{t=R+1}^{T}(\hat{\epsilon}_t^2 - \hat{e}_t^2) = \frac{1}{\sqrt{P}}\sum_{t=R+1}^{T}(\epsilon_t^2 - e_t^2) \\
&+\frac{1}{\sqrt{P}}\sum_{t=R+1}^{T}(\hat{\beta}_{t-1} - \beta)^2 x_{t-1}^2 + 2\frac{1}{\sqrt{P}}\sum_{t=R+1}^{T}(\hat{\beta}_{t-1} - \beta)x_{t-1}\epsilon_t \\
&+\frac{1}{\sqrt{P}}\sum_{t=R+1}^{T}(\hat{\gamma}_{t-1} - \gamma)^2 z_{t-1}^2 + 2\frac{1}{\sqrt{P}}\sum_{t=R+1}^{T}(\hat{\gamma}_{t-1} - \gamma)z_{t-1}e_t \\
= \frac{1}{\sqrt{P}}\sum_{t=R+1}^{T}(\epsilon_t^2 - e_t^2) + o_p(1)
\end{aligned}$$

Model 1 turns up in the end of the second line and model 2 turns up in the end of the third line. The first term in the second and third terms are very small since we are squaring small numbers and $\hat{\gamma} \sim \gamma$ since we estimate using the true number of observations. West (1996) [80] shows the final equality to

hold when $E(\epsilon_t x_{t-1}) = E(e_t z_{t-1}) = 0$ in addition to extra assumptions. The critical point here is that

$$\frac{1}{\sqrt{P}} \sum_{t=R+1}^{T} (\hat{\epsilon}_t^2 - \hat{e}_t^2) \approx \frac{1}{\sqrt{P}} \sum_{t=R+1}^{T} (\epsilon_t^2 - e_t^2)$$

i.e. there is sampling error if $\hat{\beta}$ and $\hat{\gamma}$ do not matter. Therefore, we can do loss function tests like before, even though now we must estimate parameters.

Still considering conducting loss function tests and moving onto nested models, as before, we have the same identity:

$$\begin{aligned}
\frac{1}{\sqrt{P}} \sum_{t=R+1}^{T} (\hat{\epsilon}_t^2 - \hat{e}_t^2) = \frac{1}{\sqrt{P}} \sum_{t=R+1}^{T} (\epsilon_t^2 - e_t^2) \\
+ \frac{1}{\sqrt{P}} \sum_{t=R+1}^{T} (\hat{\beta}_{t-1} - \beta)^2 x_{t-1}^2 + 2\frac{1}{\sqrt{P}} \sum_{t=R+1}^{T} (\hat{\beta}_{t-1} - \beta) x_{t-1} \epsilon_t \\
+ \frac{1}{\sqrt{P}} \sum_{t=R+1}^{T} (\hat{\gamma}_{t-1} - \gamma)^2 z_{t-1}^2 + 2\frac{1}{\sqrt{P}} \sum_{t=R+1}^{T} (\hat{\gamma}_{t-1} - \gamma) z_{t-1} e_t \\
= \frac{1}{\sqrt{P}} \sum_{t=R+1}^{T} (\epsilon_t^2 - e_t^2) + o_p(1)
\end{aligned}$$

With non-nested models, $\sum_{t=R+1}^{T} (\epsilon_t^2 - e_t^2) \sim O_p(P^{\frac{1}{2}})$ dominates the right-hand side of the above identity. However, with nested models

$$\begin{aligned}
y_{t+1} &= x_t' \beta + \epsilon_{t+1} \\
y_{t+1} &= x_t' \beta + z_t' \gamma + e_{t+1}
\end{aligned}$$

Under equal loss, the models are the same (i.e. model 1 and model 2 perform the same), so $\epsilon_t = e_t$ and so the first term vanishes:

$$\begin{aligned}
\frac{1}{\sqrt{P}} \sum_{t=R+1}^{T} (\hat{\epsilon}_t^2 - \hat{e}_t^2) = 0 \\
+ \frac{1}{\sqrt{P}} \sum_{t=R+1}^{T} (\hat{\beta}_{t-1} - \beta)^2 x_{t-1}^2 + 2\frac{1}{\sqrt{P}} \sum_{t=R+1}^{T} (\hat{\beta}_{t-1} - \beta) x_{t-1} \epsilon_t \\
+ \frac{1}{\sqrt{P}} \sum_{t=R+1}^{T} (\hat{\gamma}_{t-1} - \gamma)^2 z_{t-1}^2 + 2\frac{1}{\sqrt{P}} \sum_{t=R+1}^{T} (\hat{\gamma}_{t-1} - \gamma) z_{t-1} e_t
\end{aligned}$$

This is a more complicated problem, which McCracken (2000) [64] and Clark & McCracken (2001) [19] study by analysing the behaviour of the terms on the right-hand side of the identity. Averaging, the limits tend to be (messy) functions of normals. So, we can utilise parametric bootstrap methods with

Model 1	Model 2
MSPE (under H_0)	
$\frac{1}{P}\sum_{t=R+1}^{P}\epsilon_{t+1}^2$	$\frac{1}{P}\sum_{t=R+1}^{P}\epsilon_{t+1}^2+\frac{1}{P}\sum_{t=r+1}^{P}(x_t'\hat{\beta}_t)^2-2\frac{1}{P}\sum_{t=r+1}^{P}\epsilon_{t+1}x_t'\hat{\beta}_t$
Expectation	
σ_ϵ^2	$\sigma_\epsilon^2+E\left(\frac{1}{P}\sum_{t=r+1}^{P}(x_t'\hat{\beta}_t)^2\right)$

Table 5.2: MSPE under H_0 and expectation

Gaussian errors to approximate the limiting distribution, to compute critical values and so on.

Clark & West (CW) (2006) [18] are able to avoid falling into the problem Clark & McCracken (2001) dealt with. They focus on the situation in which f^1 is a random walk forecast and when f^2 is nested within the random walk. The first difference of a variable of interest can be denoted by y_t, e.g. an exchange rate.

$$\begin{aligned}
&H_0 : y_t \sim \text{ mds}\\
&H_1 : y_t \text{ can be predicted by } x_t\\
&\text{Model 1: } y_{t+1}=\epsilon_{t+1}\\
&\text{Model 2: } y_{t+1}=x_t'\beta+\epsilon_{t+1}
\end{aligned}$$

Let the forecast under model 1 be $\hat{f}_{t+1|t}^1=0$ and that under model two be $\hat{f}_{t+1|t}^2=x_t'\hat{\beta}_t$. Also, allow the errors under H_0 in model 1 to be $\hat{e}_{t+1}^1=\epsilon_{t+1}$ and the errors under H_0 in model 2 to be $\hat{e}_{t+1}^2=\epsilon_{t+1}-x_t'\hat{\beta}_t$. The mean square prediction errors (MSPE) under H_0 for model 1 and model 2 and their expectations are given in table 5.2 So, under the random walk (RW) null hypothesis:

$$E(\text{MSE for RW})=E(\text{MSE for alternative})-E\left(\frac{1}{P}\sum_{t=r+1}^{P}(x_t'\hat{\beta}_t)^2\right)$$

We do not check whether the errors are equal, but rather, we make the subtraction above and ascertain if the expectations are equal. We adjust for sampling error and then we can compare. The random walk forecast does not suffer from issues of overfitting $x_t'\hat{\beta}_t$ and ought to yield a superior forecast than the alternative. The 'overfitting' term can be estimated because model 2 is contaminated by sampling error:

$$E\left(\frac{1}{P}\sum_{t=r+1}^{P}(x_t'\hat{\beta}_t)^2\right)\approx\frac{1}{P}\sum_{t=r+1}^{P}(x_t'\hat{\beta}_t)^2=\frac{1}{P}\sum_{t=r+1}^{P}(f_{t|t-1}^2)^2$$

Then the CW test will be a standardised version of $\hat{\sigma}_1^2-\left(\hat{\sigma}_2^2-\frac{1}{P}\sum_{t=r+1}^{P}(f_{t|t-1}^2)^2\right)$ where $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ denote the POOS MSPE for f^1 (RW) and f^2. We want to

know the distribution of this difference:

$$\sqrt{P}\left(\hat{\sigma}_1^2 - \left(\hat{\sigma}_2^2 - \frac{1}{P}\sum_{t=r+1}^{P}(f_{t|t-1}^2)^2\right)\right) = \frac{1}{\sqrt{P}}\sum_{t-r+1}^{P}\epsilon_t f_{t|t-1}^2$$

Under the null hypothesis $\epsilon_t f_{t|t-1}^2$ is a mds so perhaps we should be looking for a normal limit? No, we must be cautious and we can see why by looking at the following example where $x_t = 1$, $f_{t|t-1}^2 = \hat{\beta}_{t-1} = \frac{1}{t-1}\sum_{i=1}^{t-1} y_i \overset{H_0}{=} \frac{1}{t-1}\sum_{i=1}^{t-1}\epsilon_i$ and $\epsilon_t f_{t|t-1}^2 \overset{H_0}{=} \frac{1}{t-1}\epsilon_t \sum_{i=1}^{t-1}\epsilon_i$. Recall that in the unit root AR model, the numerator of $\hat{\rho} - \rho$ is $\sum_{t=1}^{T}\epsilon_t \sum_{i=1}^{t-1}\epsilon_i$. CW use 'rolling' estimates of β based on R observations to limit this dependence, where R is fixed and not too large. With this in mind, we will now look at an empirical example from CW where monthly changes in US $ exchange rates were forecasted for Canada, Japan, Switzerland and the UK over a POOS period from 1990-2003 with $P = 166$ and $R = 120$; also, $x = (1, 1\text{-month interest differential})$. Table 5.3 taken from Clark & West (2006) [18] gives forecasts of monthly changes in US Dollar exchange rates.[3] Looking at Japan for example, the out-of-sample period is January 1990 to October 2003. The out-of-sample MSE for the random walk is 11.32. The out-of-sample MSE for a model in which they regress the change in exchange rate on a constant term and an interest differential is 11.55. So, the model with the interest rate differential does not forecast as well since $11.32 < 11.55$. However, if the random walk model is true, then the amount of overfitting can be computed as 0.75; so we can compute the sampling error by estimating the β from the null hypothesis that the random walk is true and we subtract to get 0.53, i.e. the model (not as a forecasting tool) would have done better by about 53 basis points.

The first part we covered on forecast assessment related to evaluating forecasts and forecasters and discovering whether

$$E[\mathcal{L}(Y - f^1)] \gtreqless E[\mathcal{L}(Y - f^2)]$$

The second part we covered on forecast assessment related to evaluating models using POOS forecasting and discovering whether

$$E[\mathcal{L}(Y - f^1(\theta_1))] \gtreqless E[\mathcal{L}(Y - f^2(\theta_2))]$$

Finally, a third part on forecast assessment relates to evaluating forecasting models using POOS forecasting and discovering whether

$$E[\mathcal{L}(Y - f^1(\hat{\theta}_1))] \gtreqless E[\mathcal{L}(Y - f^2(\hat{\theta}_2))]$$

Giacomini & White (2006) [37] cover this third topic, though their paper is is more about forecasting parameters rather than models; see table 5.4. They

[3] See the paper for more details, including notation.

(1) Country	(2) Prediction sample	(3) $\hat{\sigma}_1^2$	(4) $\hat{\sigma}_2^2$	(5) adj.	(6) $\hat{\sigma}_2^2$-adj.	(7) MPSE-adjusted $\hat{\sigma}_1^2 - (\hat{\sigma}_2^2 -$ adj.)	(8) MSPE-normal $\hat{\sigma}_1^2 - \hat{\sigma}_2^2$	(9) CCS
Canada	1990 : 1– 2003:10	2.36	2.32	0.09	2.22	0.13 (0.08) 1.78**	0.04 0.54††	3.67
Japan	1990 : 1– 2003:10	11.32	11.55	0.75	10.80	0.53 (0.43) 1.24	-0.23 -0.52	5.23*
Switzerland	1985 : 1– 2003:10	12.27	12.33	0.96	11.37	0.90 (0.48) 1.88**	-0.06 -0.13	2.43
U.K.	1985 : 1– 2003:10	9.73	10.16	0.44	9.72	0.01 (0.33) 0.03	-0.43 -1.27	0.78

Table 5.3: Forecasts of monthly changes in US Dollar exchange rates.

	Data	Forecasting Procedure	Forecast
Forecasting Model 1:	$\{x_i, y_i\}_{i=1}^t$	$\hat{\theta}_1$ etc.	$f_{t+1\|t}^1$
Forecasting Model 2:	$\{x_i, y_i\}_{i=1}^t$	$\hat{\theta}_2$ etc.	$f_{t+1\|t}^2$

Table 5.4: Giacomini & White (2006)

want to discover when using time t information, can they predict which forecast will have a smaller loss, i.e.

$$E[\mathcal{L}(f_{t+1|t}^1) - \mathcal{L}(f_{t+1|t}^2)|g_t] \gtreqless 0$$

Naturally, they convert this into a GMM problem where their moment condition is

$$E[(\mathcal{L}(f_{t+1|t}^1) - \mathcal{L}(f_{t+1|t}^2))q(g_t)] = 0$$

5.3 Forecasting with many predictors

5.3.1 Motivation

We started this course by introducing the problem of identification, which has been a theme across the chapters in terms of working with limited information in macroeconomic time series. However, we have mainly considered models that have relatively few variables, despite the availability of numerous real time economic time series. These could be used for monitoring and forecasting economic phenomena and possibly for estimating single equation models and

multiple equation models. While this seems to violate the principle of parsimony, there are reasons why we would want to use many time series. Let us look at four specific circumstances where extra information may be helpful.

1. With economic monitoring or 'nowcasting', as well as forecasting, we may ask if we could switch from small models with forecasts adjusted using judicial use of extra information to a more scientific oriented system incorporating as much quantitative information as if possible.

2. With IV estimation, more information may lead to stronger instruments.

3. With DSGE estimation, we may achieve better identification with more information.

4. Structural VAR could use more information so that innovations span the space of shocks, e.g. Factor Augmented VAR (FAVAR). We will not cover this as VAR will be part of Prof Agustín Bénétrix's part of the course.

Dynamic factor models (DFMs) that were developed by Geweke (1977) [35] and Sargent & Sims (1977) [73] are very useful for this research. The FED in Chicago uses DFMs for real-time monitoring and forecasting, e.g. Giannone, Reichlin & Small (2008) [38] and Aruoba, Diebold & Scotti (2009) [3]. Other applications include some in the area of SVARs such as Bernanke, Boivin & Eliasz's (2005) [7] FAVAR and some in the area of DSGEs such as Boivin & Giannoni (2006) [9]. Interestingly, the new trend in empirical macro of moving towards using far larger data sets is in line with developments in other sciences, especially experimental sciences such as life sciences and also in observational sciences such as astrophysics.

5.3.2 Dimensionality is not always a curse

Dimensionality need not always be a curse; instead, sometimes it can be a blessing. We will first explore the 'curse' part before discussing the 'blessing' aspect. A VAR with 100 variables and 8 lags has $100^2 \times 8 = 80,000$ coefficients and a further $\frac{100}{2}(1+100) = 5,050$ variance parameters. Exploring the negative consequences for OLS, consider the model:

$$Y_{t+1} = \delta' P_t + \epsilon_{t+1} \quad t = 1, \ldots, T$$

where P contains n orthonormal predictors (stands for principal components) so $\mathbf{P}'\mathbf{P}/T = I_n$. The orthonormality makes calculations easy so we do not need to worry about covariance terms. P_t is strictly exogenous and ϵ_{t+1} is iid $N(0, \sigma_\epsilon^2)$. Let the loss function be quadratic, i.e. $\mathcal{L}(Y_{T+1}, \tilde{Y}_{T+1|t}) = (Y_{T+1} - \tilde{Y}_{T+1|t})^2$ and consider the forecast risk or expected loss of OLS. The frequentist

risk is the expected loss. The forecast risk (how bad on average) is

$$\begin{aligned} E\mathcal{L}(Y_{T+1}, \tilde{Y}_{T+1|t}) &= E(Y_{T+1} - \tilde{Y}_{T+1|t})^2 \\ &= E[(\tilde{\delta} - \delta)P_T + \epsilon_{T+1}]^2 \\ &= E[(\tilde{\delta} - \delta)' P_T P_T' (\tilde{\delta} - \delta)] + \sigma_\epsilon^2 \\ \approx &[(\tilde{\delta} - \delta)'(\tilde{\delta} - \delta)] + \sigma_\epsilon^2 \quad \because P_t \text{ orthonormal} \\ &= R(\tilde{\delta}, \delta) + \sigma_\epsilon^2 \end{aligned}$$

where we have defined

$$R(\tilde{\delta}, \delta) = E[(\tilde{\delta} - \delta)'(\tilde{\delta} - \delta)] = Etr[(\tilde{\delta} - \delta)(\tilde{\delta} - \delta)']$$

which is the frequentist estimation risk that is often called the trace MSE risk since $tr[(\tilde{\delta} - \delta)(\tilde{\delta} - \delta)']$ is the trace MSE loss, i.e. the trace of the MSE matrix of $\tilde{\delta}$. Note that we can affect $R(\tilde{\delta}, \delta)$ since we can change the procedures we use to estimate δ but we cannot do anything about σ_ϵ^2. If we knew δ, we could simply use $\tilde{\delta} = \delta$ so $R(\tilde{\delta}, \delta) = 0$ and $E\mathcal{L}(Y_{T+1}, \tilde{T}_{T+1|t}) = \sigma_\epsilon^2$. Note that if $\tilde{\delta} \xrightarrow{p} \delta$, then $R(\tilde{\delta}, \delta) \longrightarrow 0$, so the forecast risk would also converge to zero and the forecast would be first-order efficient; there would be second order risk due to estimation error. However, if the data set is large, i.e. if n is large, then OLS is not first-order efficient. This is because $\mathbf{P}$ is strictly and ϵ_t is iid $N(0, \sigma_\epsilon^2)$:

$$\tilde{\delta} - \delta \sim N\left(0, \left(\frac{\mathbf{P}'\mathbf{P}}{T}\right)^{-1} \sigma_\epsilon^2\right) = N\left(0, I_n \frac{\sigma_\epsilon^2}{T}\right)$$

$$R(\tilde{\delta}, \delta) = Etr[(\tilde{\delta} - \delta)(\tilde{\delta} - \delta)'] = Etr[I_n \frac{\sigma_\epsilon^2}{T}] = \frac{n}{T}\sigma_\epsilon^2$$

Therefore, the forecast risk of OLS is

$$R(\tilde{\delta}, \delta) + \sigma_\epsilon^2 = (1 + \kappa)\sigma_\epsilon^2$$

where $\kappa = \frac{n}{T}$. But this means that the OLS forecast risk $E\mathcal{L}(Y_{T+1}, \tilde{Y}_{T+1|t}) = (1 + \kappa)\sigma_\epsilon^2 > \sigma_\epsilon^2$. If $\kappa \approx 0$, OLS is almost first-order efficient – this is the case of parsimony (n small relative to T). If $\frac{n}{T}$ is large, then OLS does not achieve first-order forecast efficiency. Furthermore, even when $n \geq 3$, OLS is not admissible under the trace MSE loss. According to Stein (1955), there is a better estimator $\tilde{\tilde{\delta}}$ with frequentist risk $R(\tilde{\tilde{\delta}}, \delta)$ that dominates that of the OLS, i.e. the risk is at least as good as the OLS for some δ and nor worse for all δ, uniformly. James & Stein (1960) developed a shrinkage estimator that dominates the OLS estimator; it does better than the OLS estimator around $\delta = 0$. Other things that do not achieve first order forecast efficiency include the following: (i) discarding all but a few regressors so throwing away information; (ii) allowing only statistically significant regressors; (iii) choosing regressors by information criteria. So, the curse is not really a curse, but rather it was a case of not using the right tools – OLS not being the right tool.

Now let us analyse why using many variables could be a blessing. Improving upon OLS, remember that we could not do anything about σ_ϵ^2 in the forecast risk, which is $R(\tilde{\delta}, \delta) + \sigma_\epsilon^2$, but that we could reduce $R(\tilde{\delta}, \delta)$ since it depends on the estimator and we can choose the estimator. To avoid $R^2 = 1$ forecasting regression in the limit (regression ESS exploding as $T \longrightarrow \infty$), we adopt a local nesting where $\delta_i = \frac{d_i}{\sqrt{T}}$ and let $\{d_i\}$ (unknown scaled coefficient) be distributed according to the empirical distribution G_n; if we observed the true $\{d_i\}$, we would simply construct the empirical CDF of $\{d_i\}$, i.e. G_n, which would be a step function. Restrict attention to estimators that produce the same forecast irrespective of the ordering of the regressors – this is known as *permutation equivariance*. Let us look at the frequentist risk for such estimators, in particular the part we can control:

$$
\begin{aligned}
R(\tilde{\delta}, \delta) &= \sum_{i=1}^{n} E(\tilde{\delta}_i - \delta_i)^2 \quad \text{trace MSE loss} \\
&= \left(\frac{n}{T}\right) \frac{1}{n} \sum_{i=1}^{n} E(\tilde{d}_i - d_i)^2 \quad \because \delta_i = \frac{d_i}{\sqrt{T}} \\
&= \kappa \int E(\tilde{d} - d)^2 dG_n(d) \quad \text{permutation equivalence and CDF } G_n \\
&= \kappa r_{G_n}(\tilde{d}) \quad \text{Bayes risk of estimator } \tilde{d} \text{ wrt } G_n
\end{aligned}
$$

where $\kappa = \frac{n}{T}$. Note that the penultimate equality is the expected loss with respect to the empirical CDF, which is a step function CDF. Note that the Bayes risk $R_{G_n}(\tilde{d})$ is the expected frequentist risk where expectations are taken with respect to a prior distribution. So, we have the result that the frequentist risk for permutation equivariant estimators turns out to be the same as the Bayes risk with respect to the empirical CDF of the d's, i.e. G_n. This is a key result and highlights a strong relation between Bayesian and frequentist inference. It says that if we knew G_n, then we could actually compute the Bayes estimator with respect to G_n, which in turn minimises the Bayes risk over all estimators; however, $R_{G_n}(\tilde{d}) = R(\tilde{d}, d)$, so if we minimise $R_{G_n}(\tilde{d})$, then we are also minimising $R(\tilde{d}, d)$. Therefore, the Bayes estimator that uses the prior G_n is also the optimal frequentist estimator. Obviously, from a subjectivist Bayesian philosophy, one prior can not be better than another, but if we consider using a dogmatic prior for forecasting, e.g. that VAR coefficients are always zero, your opinion would lead to poor forecasts. So, we find a prior that leads to best forecasts and we minimise the trace MSE loss. Note that the *empirical*

Bayes estimator uses the data to choose the prior.[4] Summarising:

$$\begin{aligned}
&\text{Frequentist:} && \min_{\tilde{d}} r_{G_n}(\tilde{d}) = \kappa \int E(\tilde{d}-d)^2 dG_n(d) && \text{CDF of } d_i \\
&\text{Bayes:} && \min_{\tilde{d}} r_{G}(\tilde{d}) = \kappa \int E(\tilde{d}-d)^2 dG(d) && \text{subjective prior} \\
&\text{Empirical Bayes:} && \min_{\tilde{\delta}} r_{\hat{G}}(\tilde{d}) = \kappa \int E(\tilde{d}-d)^2 d\hat{G}(d) && \text{estimated prior}
\end{aligned}$$

Robbins (1964) [70] shows that under certain conditions, the empirical Bayes estimator is asymptotically admissible and asymptotically optimal. Efron & Morris (1973) [25] show that the shrinkage estimator mentioned earlier by James & Stein (1960) [50] is an empirical Bayes estimator. Zhang (2003 [84], 2005 [85]) show minimax properties of empirical Bayes estimators. Note that $\hat{G}$ can be parametric or nonparametric. It turns out that asymptotically, empirical Bayes is minimum risk equivariant (see Edelman (1988) [24] or Knox, Stock & Watson (2001) [54] for a regression context). While these are strong results, they have yet to be proven in time series with predetermined predictors. Even so, they provide a few guidelines:

- Shrinkage, or equivalently Bayes methods can yield good forecasts with many predictors, where good is defined in terms of a frequentist risk perspective.
- Bayes methods that have tuned (estimated) parameters are attractive methods.
- Forecasts with many predictors may actually outperform forecasts built from no or simply a few predictors.
- Choosing regressors by information criteria (AIC, BIC, etc.) is non-optimal; this is a very important point.

Another example of why dimensionality may be a blessing relates to dynamic factor models (DFM). Let X_t contain n variables that are related to some unobservable factors F_t with evolution equations:

$$\begin{aligned}
X_t &= \Lambda F_t + e_t \quad \text{observation equation} \\
F_t &= \Phi(L)F_{t-1} + G\eta_t \quad \text{state/transition equation}
\end{aligned}$$

If observed, the factors could be useful for forecasting, but they are not observed. The early approach to dealing with this problem involved fitting these equations by ML via the Kalman filter. However, a problem with this was that this approach was limited to small sizes n due to the mushrooming of parameters and ML computations in high dimensions. Interestingly, the solution to this problem arose from the suggestion that perhaps using many series could improve the estimates of F_t.

[4] For those interested in books on empirical Bayes work, see Maritz & Lwin (1989) [63], Gelman, Carlin, Stern & Rubin (2003) [34] and Lehmann & Casella (1998, section 4.6) [56].

Example 5.8 (Forni & Reichlin (1998))**.** A simple DFM example involves letting F_t be a scalar so Λ is a vector with elements λ_i. Then

$$X_{it} = \lambda_i f_t + e_{it}$$

where e_{it} is idiosyncratic (uncorrelated across series). So,

$$\frac{1}{n}\sum_{i=1}^{n} X_{it} = \frac{1}{n}\sum_{i=1}^{n}(\lambda_t F_f + e_{it}) = \frac{1}{n}\sum_{i=1}^{n}\lambda_t F_t + \frac{1}{n}\sum_{i=1}^{n} e_{it}$$

If the errors u_{it} have a limited amount of dependence across series, then by LLN, supposing λ's are positive on average:

$$\frac{1}{n}\sum_{i=1}^{n} X_{it} \xrightarrow{p} \bar{\lambda} F_t$$

This is a special case where we are able to recover F_t through the cross-sectional average – an easy nonparametric estimate – as long as n is large, i.e. when there are lots of X's. Thus, we do not require the Kalman filter or state spaces, etc.

All subsequent procedures are based on asymptotic theory for large n assuming that $n \longrightarrow \infty$ typically at a rate relative to T. It is often the case that $\frac{n^2}{T}$ is treated as large in the asymptotics, which makes sense for example in an application where $n = 100$ and $T = 160$. So, by including large n, more sophisticated procedures than example 5.8 are available for allowing consistent estimation of tuning priors (also called prior hyperparameters) in forecasting and for factors within DFMs. This is a very recent area: most of the theory and all of the empirical work has been carried out in the past 15 years or so. For macroeconomic modeling, forecasts with many predictors can be made from a bunch of different models including DFMs and other high-dimensional forecasting methods such as optimal Bayes estimators, hard thresholding, information criteria, false discovery rate (FDR) methods, large VARs, bootstrap aggregation or bagging and Bayesian model averaging. We will not cover these models within this course, but merely mention their existence and the use of large scale models in terms of using many predictors that can be useful within areas such as those using FAVAR (SVARs with factors), using factors as instruments and DSGE estimation.

To conclude, there have been considerable advances towards exploiting large data sets in the recent literature. Having a large number of variables can be a blessing though it seems to question the very principle of parsimony; recall the results as $\frac{n^2}{T} \longrightarrow \infty$. Beyond the scope of this present course, the profession has accumulated a lot of knowledge regarding DFM estimation and has advanced many intriguing applications to forecasting, which have been implemented in real time, e.g. applications to FAVAR, IV estimation and DSGE estimation.

Chapter 6

Nonlinear Volatility Models

In this chapter, we will consider ARCH/GARCH, Markov Switching and Stochastic Volatility models. These are all nonlinear models, which have be used for modeling volatilty. Markov Switching models describe discrete regime change, while the other two are continuous models. Stochastic volatility models allow the variance of a process to be directly determined stochastically rather than being modelled in terms of past observations. Stochastic volatility models and markov switching models each tend to be preferred to ARCH/GARCH models, while the choice of stochastic volatility models over markov switching models is essentially an empirical question; see Fernández-Villaverde & Rubio-Ramírez (2010) [30].

6.1 Modeling volatility

We know that most economic time series neither exhibit constant means nor constant variances. In fact, generally such series display periods of calm followed by periods of turbulence or high volatility. It would therefore appear that homoscedastic or constant variance stochastic variables are less preferable in these contexts to those reflecting heteroscedastic variance. Volatile series may have a constant unconditional variance while for some periods the variance is particularly high. Nonlinear models are necessary when we want to look at volatility; they are useful in other applications too. In this section, we will discuss GARCH, Markov-Switching and Stochastic Volatility models, which are all non-linear models as Harvey (1993) explains in beginning of chapter 8 [46]. Before we begin our study of particular models, let us describe the empirical facts behind many economic time series.

While formal testing is necessary to provide evidence for visual inspection, sometimes visual inspection by itself, though generally perilous, can be sufficient. In fact over-testing can be an issue in certain circumstances. For now, let us consider macroeconomic indicators such as the evolution of real GDP and some of its components such as real consumption, real government expenditure and real investment in the US. Figures 3.1 through 3.6 in Enders

(2008:109-12) [27] indicate that these series are not stationary since the sample mean is not constant, but there is an upward trend over time; there is also a clear presence of heteroscedasticity. Let us present five stylised facts from the data.

1. There is a clear upward trend in most of these series.
2. Shocks to these series are typically persistent.
3. Volatility of these series changes over time.
4. Series appear to meander or exhibit random-walk behaviour, though we need to test if instead there is any mean-reversion for some series say like real effective exchange rates.
5. Some series comove with other series.

To investigate these issues further, we need to formally test for the presence of conditional heteroscedasticity (when the unconditional or long-run variance is constant but there are period where the variance is especially high) or non-stationarity. Obviously, many series clearly display evidence of both, but for some series, this is not so obvious. For the rest of this current section, we will focus on conditional heteroscedasticity. One motivation for studying conditional heteroscedasticity is that asset holders who buy at t and sell at $t+1$ want to predict the rate of return and the variance over the period, so they do not care about the long-run unconditional forecast of variance. So, conditional heteroscedasticity does not imply stationarity to the extent that the unconditional variance could still be constant in the long-run.

6.2 ARCH & GARCH

There are many ways to model changes in variance, the basic set up being to consider the series of interest as a sequence of iid random variables ϵ_t with unit variance multiplied by the standard deviation, a factor σ_t, i.e.

$$y_t = \sigma_t \epsilon_t \quad \epsilon_t \sim iid(0,1)$$

ARCH models the variance in terms of of past observations, while more direct approaches model σ_t as a stochastic process such as an autoregressive process; an example of the later is the SV model we will discuss later. ARCH is attributed to Engle (1982) [**?**] who modeled the variance directly in terms of past observations. The simplest expression of ARCH has

$$\sigma_t^2 = \gamma + \alpha y_{t-1}^2 \quad \gamma > 0, \alpha \geq 0 \tag{6.1}$$

where the constraints ensure the variance is positive. So, the model specifies a predictive distribution for y_t. When ϵ_t is Gaussian

$$y_t = \sigma_t \epsilon_t \quad \epsilon_t \sim NID(0,1) \tag{6.2}$$

and the model is conditionally Gaussian, so $y_t|Y_{t-1} \sim N(0, \sigma_t^2)$. The name ARCH originates from the fact that the model exhibits *autoregressive conditional heteroscedasticity* since the variance has a similar form to the conditional expectation of the mean in a standard AR(1) process. There are a multitude of variations on ARCH; see Bollerslev (2010) [10].

While not independent, the observations form a MD sequence and so the ARCH model has a zero unconditional mean and is serially uncorrelated. The unconditional variance is given by:

$$E_{t-2}E_{t-1}(y_t^2) \overset{\text{LIE}}{=} E_{t-2}[\gamma + \alpha y_{t-1}^2] = \gamma + \gamma\alpha + \alpha^2 y_{t-2}^2$$

Continuing like this until time $t - J$:

$$E_{t-J} \cdots E_{t-1}(y_t^2) = \gamma + \gamma\alpha + \gamma\alpha^2 + \cdots + \gamma\alpha^{J-1} + \alpha^J y_{t-J}^2$$

Provided $\alpha < 1$, we can sum as the infinite geometric progression (letting $J \longrightarrow \infty$) to get

$$Var(y_t) = E(y_t^2) = \gamma/(1 - \alpha) \tag{6.3}$$

So, the ARCH is WN but not strict WN. Furthermore, while conditionally Gaussian, it is not unconditionally Gaussian because if it were, then it would be a linear model. While the unconditional distribution is nonstandard, it is is symmetric because all odd moments are zero and ICBST if $3\alpha^2 < 1$, then the kurtosis is:

$$\frac{3(1 - \alpha^2)}{1 - 3\alpha^2}$$

which is greater than 3 for $\alpha > 0$; thus, the ARCH model produces observations with heavier tails than those coming from a Normal distribution, which is a very appealing property since many financial time series tend to have distributions that are heavy tailed. To see the dynamics of the ARCH model, we can take a look at the ACF of the squared observations:

$$y_t^2 = \sigma_t^2 + (y_t^2 - \sigma_t^2)$$

and use (6.1) and (6.2) to get

$$y_t^2 = \gamma + \alpha y_{t-1}^2 + v_t \tag{6.4}$$

where $v_t = \sigma_t^2(\epsilon_t^2 - 1)$. Note that the disturbance term v_t is a MD since

$$E_{t-1}(v_t) = \sigma_t^2 E_{t-1}(\epsilon_t^2 - 1) = 0$$

and ICBST v_t has constant variance, so it is WN. So, (6.4) shows that the squared observations follow and AR(1) process and therefore the ACF is given by

$$\rho(\tau; y_t^2) = \alpha^\tau \quad \tau = 0, 1, 2, \ldots$$

Regarding prediction, the conditional expectation of any future observation is zero, but the LIE can be used to show that the prediction MSE (here it is just the conditional variance of future observations) is given by

$$MSE(\tilde{y}_{T+l|T}) = \gamma(1 + \alpha + \alpha^2 + \cdots + \alpha^{l-1}) + \alpha^l y_T^2 \tag{6.5}$$

If the series is treated as WN, the prediction would be the unconditional variance (6.3). As $l \longrightarrow \infty$, (6.5) tends to (6.3), but for small horizons – or lead times – it could be quite different.

ARCH has the property that the conditional variance depends only on a single observation. To see why this is unsatisfactory, observe that a high conditional variance at time $t-1$ could generate an observation close to zero so the conditional variance at time t would be rather small. We generally expect variance to change more slowly, which requires spreading the memory of the process over a number of past observations rather than concentrating the memory to the immediately previous period, i.e. we need more lags:

$$\sigma_t^2 = \gamma + \alpha_1 y_{t-1}^2 + \cdots + \alpha_p y_{t-p}^2$$

which is an ARCH(p) and works better when certain restrictions are placed on the coefficients. For example, a linear decline may be represented through the constraint:

$$\alpha_i = \alpha\{(9-i)/36\} \quad i = 1, \ldots, 8$$

which leaves only two free parameters to be estimated. An even better approach introduces lagged values of σ_t^2 into the equation as in the following definition.

Definition 6.1. The *generalised ARCH* (*GARCH*) model is given by

$$\sigma_t^2 = \gamma + \alpha_1 y_{t-1}^2 + \cdots + \alpha_p y_{t-p}^2 + \beta_1 \sigma_{t-1}^2 + \cdots + \beta_q \sigma_{t-q}^2$$

This is also called the GARCH(p, q) model.

The GARCH(p, q) model was introduced by Bollerslev (1986). For the basic GARCH(1,1):

$$\sigma_t^2 = \gamma + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2 \quad \gamma > 0, \alpha, \beta \geq 0, \alpha + \beta < 1$$

Note that all GARCH models are martingale differences and if the sum of the α_i's and β_j's is less than one, then the model has a constant finite variance and so it is white noise. Example 6.2 shows this for GARCH(1,1).

Example 6.2. For the GARCH(1,1):

$$\begin{aligned} E_{t-2}E_{t-1}(y_t^2) &= E_{t-2}[\gamma + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2] \\ &= \gamma + (\alpha + \beta)\sigma_{t-1}^2 \\ &= \gamma + (\alpha + \beta)[\gamma + \alpha y_{t-2}^2 + \beta \sigma_{t-2}^2] \end{aligned}$$

if we repeat this infinitely, then when $\alpha + \beta < 1$, we get that:

$$Var(y_t) = \frac{\gamma}{1 - \alpha - \beta}$$

Next let us look at the ACF of the squared observations of GARCH, which turns out to be similar to that of an ARMA process, though the correspondence is not as direct as with the ARCH case. We can follow (6.4) by writing

$$y_t^2 = \gamma + \sum_{i=1}^{p} \alpha_i y_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 + v_t$$

As before, the error term $v_t = \sigma_t^2(\epsilon_t^2 - 1)$ is WN. By adding and subtracting $\beta_j y_{t-j}^2$ for $j = 1, \ldots, q$, we get that

$$y_t^2 = \gamma + \sum_{i=1}^{p} \alpha_i y_{t-i}^2 + \sum_{j=1}^{q} \beta_j y_{t-j}^2 + \sum_{j=1}^{q} \beta_j (\sigma_{t-j}^2 - y_{t-j}^2) + v_t \qquad (6.6)$$

Rearrange and define $p* = max(p, q)$ to get:

$$y_t^2 = \gamma + \sum_{i=1}^{p*} \phi_i y_{t-i}^2 + v_t + \sum_{j=1}^{q} \theta_j v_{t-j}$$

where we have defined

$$\phi_i = \alpha_i + \beta_i \quad i = 1, \ldots, p* \quad \theta_j = -\beta_j \quad j = 1, \ldots, q$$

So, the ACF of y_t^2 is the same as that of the ARMA(p*,q) process in (6.6). With the GARCH(1,1) model, the ACF is the same of an ARMA(1,1) model:

$$\rho(1) = \frac{(1 + \phi\theta)(\phi + \theta)}{1 + \theta^2 + 2\phi\theta}$$
$$\rho(\tau) = \phi\rho(\tau - 1) \quad \tau = 2, 3, \ldots$$

Therefore, if the sum of α and β is about one, the ACF decays slowly reflecting that the conditional variance changes slowly. This may happen in practice, which motivated the introduction of GARCH and it turns out that GARCH(1,1) with $\alpha + \beta$ close to one often fit the data well.

Drawbacks to GARCH include the following: (i) conditional variance cannot respond asymmetrically to rises and falls in y_t that are sometimes observed in stock returns for instance; (ii) estimated coefficients often violate parameter constraints and these constraints can severely restrict the dynamics of the conditional variance; (iii) it is difficult to asses whether the conditional variance shocks are persistent. Nelson's (1991) [67] *exponential ARCH* (*EGARCH*) overcomes these problems. He assumed that $\log \sigma_t^2$ is a function of past ϵ_t's to keep the conditional variance non-negative:

$$\log \sigma_t^2 = \gamma + \sum_{i=1}^{\infty} \psi_i g(\epsilon_{t-i}) \quad \psi_0 = 1$$

Specifying

$$g(\epsilon_t) = \omega\epsilon_t + \lambda[|\epsilon_t| - E|\epsilon_t|]$$

allows $g(\epsilon_t)$ to be a function of the magnitude and sign of ϵ_{t-1}, which in turn enables σ_t^2 to respond asymmetrically to rises and falls in y_t. Note that when $\epsilon_t > 0$, $g(\epsilon_t)$ is linear in ϵ_t with slope $\omega + \lambda$, while when $\epsilon_t < 0$, $g(\epsilon_t)$ has slope $\omega - \lambda$. EGARCH models are estimated by ML.

We can write a model with a first-order ARCH disturbance as

$$\begin{aligned} y_t &= \mathbf{x}_t'\boldsymbol{\beta} + u_t \quad t = 1, \ldots, T \\ u_t &= \sigma_t \epsilon_t \quad \epsilon_t \sim NID(0,1) \\ \sigma_t^2 &= \gamma + \alpha u_{t-1}^2 \quad \gamma > 0, \alpha \geq 0 \end{aligned} \tag{6.7}$$

As u_t will be WN, the GM theorem implies that OLS of y_t on $\mathbf{x}_t$ will produce the BLUE of $\boldsymbol{\beta}$. However, OLS will be inefficient because disturbances are not independent; they are merely uncorrelated. An efficient estimator will be constructed by ML. For a conditionally Gaussian model, the likelihood function is:

$$\log L(\alpha, \gamma) = -\frac{T}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T}\log(\gamma + \alpha(y_{t-1} - \mathbf{x}_{t-1}'\boldsymbol{\beta})^2) - \frac{1}{2}\sum_{t=1}^{T}\frac{(y_t - \mathbf{x}_t'\boldsymbol{\beta})^2}{\gamma + \alpha(y_{t-1} - \mathbf{x}_{t-1}'\boldsymbol{\beta})^2}$$

ICBST the information matrix is block diagonal with respect to the ARCH parameters γ, α and the regression parameters $\boldsymbol{\beta}$.

Definition 6.3. With the *ARCH-M* model, ARCH effects are present in the mean of the process, so (6.7) will be

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + \delta\sigma_t + u_t$$

ARCH-M models are used when expected return partially depends on risk as is reflected in volatility. We estimate these models by ML. There are many other versions of ARCH/GARCH we will not get into. We will conclude our study of ARCH/GARCH with a discussion of estimation and testing.

We estimate ARCH/GARCH models by maximum likelihood (ML). The joint density of observations is the likelihood:

$$L = \Pi_t p(y_t | Y_{t-1})$$

where we have information at time $t-1$. For a conditionally Gaussian first-order ARCH process, $p(y_t|Y_{t-1})$ is Normally distributed with zero mean and variance given by:

$$\sigma_t^2 = \gamma + \alpha y_{t-1}^2 \quad \gamma > 0, \alpha \geq 0$$

If we further assume that y_0 is arbitrarily fixed equal to zero, then the log-likelihood function is given by:

$$\log L(\alpha, \gamma) = -\frac{T}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T}\log(\gamma + \alpha y_{t-1}^2) - \frac{1}{2}\sum_{t=1}^{T}\frac{y_t^2}{\gamma + \alpha y_{t-1}^2}$$

It turns out that the method of scoring is a feasible procedure for maximising the likelihood function when dealing with ARCH or GARCH models.

Remark 6.4 (Method of Scoring)**.** Methods of optimisation that approximate functions by quadratic forms, hence employing first and second derivatives are called the *Newton-Raphson* method or *Newton's method*. Taylor expanding the criterion function $f(\Psi)$ around the minimum of the parameter vector $\tilde{\Psi}$ yields

$$f(\Psi) = f(\tilde{\Psi}) + (\Psi - \tilde{\Psi})g(\tilde{\Psi}) + \frac{1}{2}(\Psi - \tilde{\Psi})^2 G(\tilde{\Psi}) \tag{6.8}$$

where $g(\Psi)$ is $\partial f(\Psi)/\partial\Psi$ and $G(\Psi)$ is the Hessian $\partial^2 f(\Psi)/\partial\Psi\partial\Psi'$. Differentiating (6.8) with respect to Ψ, we get that

$$g(\Psi) = g(\tilde{\Psi}) + (\Psi - \tilde{\Psi})G(\tilde{\Psi}) \tag{6.9}$$

Since $g(\tilde{\Psi}) = 0$ as $\tilde{\Psi}$ is the minimum of f, we can rearrange (6.9) into:

$$\tilde{\Psi} \approx \Psi - G^{-1}(\tilde{\Psi})g(\Psi) \tag{6.10}$$

which suggests the recursion:

$$\Psi* = \hat{\Psi} - G^{-1}(\tilde{\Psi})g(\hat{\Psi})$$

where our revised estimate is $\Psi*$ and our initial estimate is $\hat{\Psi}$. More generally since we do not know $\tilde{\Psi}$, as the scheme progresses, $\hat{\Psi}$ is the current estimate and is updated to $\Psi*$ via:

$$\Psi* = \hat{\Psi} - G^{-1}(\hat{\Psi})g(\hat{\Psi}) \tag{6.11}$$

See example on pages 128-9 in Harvey (1981) [45]. For the log-likelihood, the Newton-Raphson iteration is:

$$\Psi* = \hat{\Psi} - [D^2 \log L(\hat{\Psi})]^{-1} D \log L(\hat{\Psi})$$

There are many variants of this for ML estimation, exploiting different features of the problem, hence yielding more efficient algorithms; e.g. Gauss-Newton procedure is useful when minimising the sum of squares is equivalent to maximising the likelihood function. Sometimes it can be more efficient to maximise likelihood by looking at the expectation rather than the matrix of second derivatives when maximising likelihood. So, we will get the information matrix if we multiply by minus one and a modified Newton-Raphson iterative procedure will be:

$$\Psi* = \hat{\Psi} + I^{-1}(\hat{\Psi}) D \log L(\hat{\Psi})$$

This is called the *method of scoring* procedure. See Harvey (1981:132) for more details [45].

Let $\rho(\tau; y_t^2)$ denote the ACF of squared observations:

$$\rho(\tau; y_t^2) = \frac{E[(y_t^2 - \sigma_y^2)(y_{t-\tau}^2 - \sigma_y^2)]}{E[(y_t^2 - \sigma_y^2)^2]} \quad \tau 0, 1, 2, \ldots$$

where σ_y^2 is the variance of a zero mean series y_t and let $r(\tau; y_t^2) = \{\rho(\tau)\}^2 \; \forall \tau$ be the sample estimator of $\rho(\tau; y_t^2)$. Then a higher order analog of the portmanteau statistic (test for randomness) is:

$$Q(P) = T(T+2) \sum_{t=1}^{P} (T-\tau)^{-1} \{r(\tau; y_t^2)\}^2 \tag{6.12}$$

which is asymptotically χ_P^2 for Gaussian WN. We can test for ARCH via (6.12).

6.3 Markov-Switching models

Let us now switch over to looking at modeling time series with changes in regime, specifically Hamilton's *Markov Switching* model. Episodes like Mexico's Tequila crisis in 1992 where the Mexican government discouraged the use of dollar-denominated accounts in Mexican banks and Argentina's Corralito crisis in 2001 where the Argentinian government froze bank accounts and transformed dollar-denominated deposits into peso-denominated deposits at an artificial exchange rate lead to dramatic breaks in series. One way of modeling these phenomena would be to introduce a *structural break* in the series for instance:

$$y_t = \mu_1 + D_t(\mu_2 - \mu_1) + \phi y_{t-1} + \epsilon_t$$

where D_t is a dummy variable that is zero until $t = \tau$ and one thereafter. However, this is not very satisfactory since we cannot plausibly forecast from such a model to the extent that a change in regime tends not to be a 'perfectly foreseeable, deterministic event', as Hamilton (1994:677)[43] says [43]. Instead, the regime change is a random variable so for a time series model, we need to model the probability of switching or the transition law from μ_1 to μ_2. We let the unobserved random variable s_t* that we call the *state* or *regime* influence the process, where $s_t* = i$ means the process is in regime i; in the baseline model, $i \in \{1, 2\}$. So we would then have:

$$y_t = \mu_{s_t*} + \phi y_{t-1} + \epsilon_t$$

To model the time series process for s_t*, which is an unbserved variable, remember that it takes only discrete values, e.g. 1 or 2 and so it is different from GARCH and stochastic volatility (defined shortly), which are continuous models. A *Markov chain* is a simple model discrete-valued random variables.

Definition 6.5. Let the random variable s_t take only integer values, say $\{1, 2, \ldots, N\}$. Let

$$P\{s_t = j | s_{t-1} = i, s_{t-2} = k, \ldots\} = P\{s_t = j | s_{t-1} = i\} = p_{ij}$$

This process is called an N-state *Markov chain* with transition probabilities $\{p_{ij}\}_{i,j=1,2,\ldots,N}$ that give the probability of state j in the next period given

state i in the current period, so

$$\sum_{j=1}^{N} p_{ij} = 1$$

Definition 6.6. The *transition matrix* is the $N \times N$ matrix $\mathbf{P}$ of transition probabilities:

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{21} & \cdots & p_{N1} \\ p_{12} & p_{22} & \cdots & p_{N2} \\ \vdots & \vdots & \cdots & \vdots \\ p_{1N} & p_{2N} & \cdots & p_{NN} \end{bmatrix}$$

where the row j column i element of $\mathbf{P}$ is the transition probability p_{ij} – e.g. row 2, column 1 element p_{12} is the probability that state 2 follows state 1.

Markov chains can be conveniently represented by VAR, but for this course, we will neither explore this aspect nor forecasting Markov chains; in addition, we will not go into any depth on reducible Markov chains, ergodic Markov chains, periodic Markov chains and statistical analysis of iid mixture distributions, all important topics by themselves. For those interested, you may wish to consult Hamilton (1994) chapter 22 [43]. The remainder of our discussion of Markov-Switching models in class follows section 22.4 of that book, which you should read.

6.4 Stochastic Volatility models

Taylor (1986) [78] provides an early example of the SV model as a flexible alternative to ARCH and GARCH models for capturing heteroscedasticity in innovations in time series work. Taylor was interested in incorporating persistent, time-varying volatility in financial returns data as well as accommodating fat-tailed behaviour. The SV model Taylor used was a simple one:

$$\begin{aligned} y_t &= u_t \beta \exp(s_t/2) \\ s_{t+1} &= \phi s_t + v_t \end{aligned}$$

where $u_t \perp\!\!\!\perp v_t$ are both Normal random variables with variances 1 and σ^2, respectively. β here clearly represents the model volatility while ϕ and σ^2 control the degree of persistence and variance of shocks to volatility. Regarding empirical implementation, the nonlinear aspect of the model requires more complicated methods.

Again, when we model volatility, we may consider certain time series that have time-varying variances, where the changes seem to be serially correlated, for instance with groups of more volatile observations occurring on after another. The intuition behind this is that when markets face uncertainty – possibly due to an international crisis, then it takes time for tranquility to reign once again, or for prices to settle down. The basic set up for modeling changes

in variance σ_t^2 is to let the series be subject to iid unit variance shocks that are multiplied by a time varying factor, which is the standard deviation:

$$y_t = \sigma_t \epsilon_t \quad \epsilon_t \sim IID(0,1) \tag{6.13}$$

When σ_t is modeled by a stochastic process, e.g. an autoregression, the model is called a stochastic variance model. This contrasts to the ARCH/GARCH models we studied above where σ_t is a function of past observations. However, they are both similar in that they are both an MD sequence even though they are not independent. This subsection focuses on the case where σ_t in (6.13) follows a stochastic process and is an unobserved variable, i.e. we look at stochastic volatility/variance (SV) models.

SV models appear frequently in finance, especially in generalisations of the Black-Scholes option pricing theory result. It is very hard to write down the exact likelihood function and this was their 'principal disadvantage.' (Harvey, 1993:281) [46] However, with recent advances in non-linear filtering, this problem has been greatly overcome; see for example, the work of Jesus Fernández-Villaverde who makes use of particle filtering to overcome the problem of non-linearities in estimating the likelihood function. Furthermore, they do not suffer from many disadvantages of GARCH models.

Rather than directly set up a stochastic process for σ_t^2, the stochastic process is assumed for $\log(\sigma_t)$ so that σ_t^2 is always positive similar to EGARCH. The general expression for a SV model is as follows:

$$y_t = \sigma_t \epsilon_t \quad \sigma_t^2 = \exp(h_t) \quad t = 1, \ldots, T \tag{6.14}$$

$$h_t = \gamma + \phi h_{t-1} + \eta_t \quad \eta_t \sim NID(0, \sigma_\eta^2) \tag{6.15}$$

where η_t may be a function of t or may not be.

Looking at the properties of SV models, let assume throughout that $\eta_t \perp\!\!\!\perp \epsilon_t$. Note that h_t is strictly stationary if $|\phi| < 1$ in (6.15). Then h_t will have mean $\gamma_h = \gamma/(1-\phi)$ and variance $\sigma_h^2 = \sigma_\eta^2/(1-\phi^2)$. Note that the product of two strictly stationary processes is also strictly stationary, so y_t is also strictly stationary. Therefore, the conditions for stationarity of y_t will be those that guarantee the stationarity of the process generating h_t.

Remark 6.7. Observe that y_t is WN. Given the independence of ϵ_t and η_t, note that the mean is obviously zero and since $E(\epsilon_t \epsilon_{t-\tau}) = 0$,

$$E(y_t y_{t-\tau}) = E(\epsilon_t \epsilon_{t-\tau}) E[\exp(h_t/2) \exp(h_{t-\tau}/2)] = 0$$

As long as ϵ_t is symmetric, all the odd moments of y_t are zero, while if ϵ_t is Normal, then the even moments can be obtained in a formulaic fashion using the following result.

Lemma 6.8. *If $\exp(h_t)$ is log-normal (i.e. h_t is Gaussian), then the j^{th} moment around the origin is:*

$$\exp\{j\gamma_h + \frac{1}{2} j^2 \sigma_h^2\}$$

From this lemma:

$$V(y_t) = E(\epsilon_t^2)E\{\exp(h_t)\} = \exp\{\gamma_h + \frac{1}{2}\sigma_h^2\} \tag{6.16}$$

$$E(y_t^4) = E(\epsilon_t^4)E\{\exp(2h_t)\} = 3\exp\{2\gamma_h + 2\sigma_h^2\} \tag{6.17}$$

So the kurtosis is $3\exp\{\sigma_h^2\}$, which is greater than 3 when $\sigma_h^2 > 0$; hence the model displays excess kurtosis as compared with the normal distribution.

It is easier to investigate the dynamic properties of the model in $\log y_t^2$ rather than y_t^2. Logarithms of squared observations in (6.14) are given by

$$\log y_t^2 = h_t + \log \epsilon_t^2$$

When $\epsilon_t \sim N(0,1)$, $E(\log \epsilon_t^2) = -1.27$ and $V(\log \epsilon_t^2) = 4.93$, so

$$\log y_t^2 = -1.27 + h_t + \epsilon_t^* \tag{6.18}$$

where $\epsilon_t^* = \log \epsilon_t^2 + 1.27$. Then $\log y_t^2$ is the sum of an AR(1) component and WN and its ACF is given by

$$\rho(\tau; \log y_t^2) = \phi^\tau/(1 + 4.93/\sigma_h^2) \quad \tau = 1, 2, \ldots \tag{6.19}$$

As $\log y_t^2$ is equivalent to an ARMA(1,1), its properties are similar to a GARCH(1,1). For instance, if σ_h^2 is small and / or $\phi \sim 1$, then the correlogram of y_t^2 is very close to that of an ARMA(1,1) process.

We may generalise the model so h_t follows a stationary ARMA process. In this case, y_t is stationary with variance and fourth moment given by (6.16) & (6.17), respectively. We can work out the ACF of $\log y_t^2$ from (6.18) along with the dynamic properties of h_t.

Furthermore, we can model ϵ_t as a student t-distribution, as we could for ARCH models. The importance of this can be seen for ARCH in that the kurtosis of many financial time-series is greater than that from using a conditionally heteroscedastic Gaussian process. With the SV model, we can again show that if h_t is stationary, then y_t is WN and from the properties of the t-distribution we get that the formula for $Var(\epsilon_t)$ in (6.16) becomes $\{\nu/(\nu-2)\}\exp(\gamma_h + \frac{1}{2}\sigma_h^2)\}$, where ν is the degrees of freedom; the kurtosis is $3\{(\nu-2)/(\nu-4)\}\exp(\sigma_h^2)$. When $\epsilon_t \sim t$ distribution

$$\epsilon_t = \zeta_t/\kappa_t^{\frac{1}{2}}$$

where $\zeta_t \sim N(0,1)$ and $\zeta_t \perp\!\!\!\perp \nu\kappa_t \sim \chi_\nu^2$.

$$\therefore \log \epsilon_t^2 = \log \zeta_t^2 - \log \kappa_t \tag{6.20}$$

and[1]

$$\begin{aligned} E(\log \kappa_t) &= \psi(\nu/2) - \log(\nu/2) \\ Var(\log \kappa_t) &= \psi'(\nu/2) \end{aligned} \tag{6.21}$$

[1] See Abramowitz and Stegun (1970: 260) [1].

where $\psi(\cdot)$ and $\psi'(\cdot)$ are digamma and trigamma functions, respectively. So (6.18) becomes

$$\log y_t^2 = -1.27 - \psi(\nu/2) + \log(\nu/2) + h_t + \epsilon_t^*$$

where

$$E(\epsilon_t^*) = 0$$
$$Var(\epsilon_t^*) = 4.93 + \psi'(\nu/2)$$

The ACF of $\log y_t^2$ has the same form as before except now $\psi'(\nu/2)$ is added to 4.93 in the expression for $\rho(\tau; \log y_t^2)$ in (6.19).

The state space form is particularly helpful in dealing with SV models. Equations (6.18) & (6.15) make up the measurement and transition equations, respectively. ICBST η_t and the disturbances ϵ_t^* are uncorrelated, even if η_t and ϵ_t are not. The key problem is that ϵ_t^* in (6.18) is non-Gaussian. The Kalman filter is therefore inappropriate since it will only yield MMSLEs of the state and future observations rather than MMSEs.[2] Moreover, because the model is not conditionally Gaussian, we cannot obtain the exact likelihood from the Kalman filter. One approach is to compute the estimates by treating the model as if it was Gaussian and maximising the resulting quasi-likelihood function. Ruiz (1992) [72] shows that there is little gain in efficiency when making the assumption that ϵ_t is Gaussian even when it is true when using this procedure. So it might make more sense to estimate the variance of ϵ_t^* rather than setting it to 4.93. However, this leads to an identification problem in that when the distribution of ϵ_t is not specified, γ_h is not identified because the expected value of $\log \epsilon_t^2$ is unknown. Therefore, the level of volatility is not determined. Under the assumption that ϵ_t follows a t-distribution, the estimated variance of ϵ_t^* implies a value of ν when set to $4.93 + \psi'(\nu/2)$, which yields the expectation of $\log \epsilon_t^2$ from (6.20) & (6.21). As an alternative to quasi-ML, the GMM estimation procedure was employed by Melino & Turnbull (1990) [65]. A final alternative involves a recent approach using the particle filter; there are even more sophisticated methods making use of efficient importance samplers, which are beyond the scope of this course.

We may assume a non-stationary process for the variance such as a random walk:

$$h_t = h_{t-1} + \eta_t \quad \eta \sim NID(0, \sigma_\eta^2) \tag{6.22}$$

Here $\log y_t^2$ will be a random walk plus noise. Harvey shows in section 5.3 of TSM that the optimal predictor is an EWMA of past observations, so there is a parallel with the IGARCH model where the unconditional variance

$$h_t = \gamma + \alpha y_{t-1}^2 + (1-\alpha) h_{t-1}$$

is also an EWMA. The critical difference between the IGARCH and this SV model is that in the IGARCH model, conditional variance is known *exactly*,

[2] Minimum mean square linear estimators (MMSLE) and minimum mean square estimators (MMSE) are defined in chapter 5; see definition 5.1 together with the proof and the discussion that follow it.

whereas now the variance is an unobserved component and a better estimate can be obtained by a smoothing algorithm. Though the model based on (6.22) doesn't lead to an exact form of the likelihood (unlike IGARCH), it does contain one less parameter and can be (relatively) easily estimated by the quasi-ML procedure mentioned above. Similar to IGARCH, it seems to provide a good fit to many data sets and it generalises easily to multivariate series.

Example 6.9. See example 1 on page 284 of Harvey (1993) [46].

Chapter 7

Filtering and Simulation

Most of what we will do in the first two sections pertains to linear filtering theory. First we study the state space form, which is useful in that linear state space models allow us to employ Kalman filtering. After discussing Kalman filters and smoothers, we will look at frequency related filtering, which necessitates a study of the frequency domain approach. In the final section we will discuss simulation methods and we will explore some non-linear filters.

7.1 State space form and Kalman filters & smoothers

We will first look at the state space form (SSF), which is extremely powerful as a tool that can be used for time series. Once we write a model in state space form, we can apply the Kalman filter for prediction and smoothing. With Gaussian models, the likelihood function can be constructed by prediction error decomposition via the Kalman filter. This technique can be applied for instance to exact ML estimation of ARMA and TVP regression models. While this course concerns itself mainly with univariate time series, the SSF can be applied to multivariate time series.[1] Here, I will present the model for multivariate time series and stress univariate time series concerns arise.

Definition 7.1. The $N \times 1$ vector of observed variables at time t, $\mathbf{y}_t$ is related to the $m \times 1$ *state vector* $\boldsymbol{\alpha}_t$ through a *measurement equation*:

$$\mathbf{y}_t = \mathbf{Z}_t \boldsymbol{\alpha}_t + \mathbf{d}_t + \boldsymbol{\epsilon}_t \quad t = 1, \ldots, T$$

where $\mathbf{Z}_t$ is an $N \times m$ matrix, $\mathbf{d}_t$ is an $N \times 1$ vector and $\boldsymbol{\epsilon}_t$ is an $N \times 1$ vector of serially uncorrelated disturbances satisfying

$$E(\boldsymbol{\epsilon}_t) = \mathbf{0}$$
$$Var(\boldsymbol{\epsilon}_t) = \mathbf{H}_t$$

[1] Professor Agustin Bénétrix will discuss multivariate models, in particular vector autoregression (VAR) models for time series.

In general we do not observe the elements of $\boldsymbol{\alpha}_t$, but we know that they are generated by a first-order Markov process also known as the *transition equation*:[2]

$$\boldsymbol{\alpha}_t = \mathbf{T}_t\boldsymbol{\alpha}_{t-1} + \mathbf{c}_t + \mathbf{R}_t\boldsymbol{\eta}_t \quad t = 1, \ldots, T \tag{7.1}$$

where $\mathbf{T}_t$ is an $m \times m$ matrix, $\mathbf{c}_t$ is an $m \times 1$ vector, $\mathbf{R}_t$ is an $m \times g$ matrix and $\boldsymbol{\eta}_t$ is a $g \times 1$ vector of serially uncorrelated disturbances satisfying

$$\begin{aligned} E(\boldsymbol{\eta}_t) &= \mathbf{0} \\ Var(\boldsymbol{\eta}_t) &= \mathbf{Q}_t \end{aligned} \tag{7.2}$$

In order to complete the specification of the state space system, we need two further assumptions:

1. The initial state vector $\boldsymbol{\alpha}_0$ has the properties that:

$$\begin{aligned} E(\boldsymbol{\alpha}_0) &= \mathbf{a}_0 \\ Var(\boldsymbol{\alpha}_0) &= \mathbf{P}_0 \end{aligned}$$

2. The disturbances are mutually uncorrelated across time periods and with the initial state, i.e.

$$E(\boldsymbol{\epsilon}_t\boldsymbol{\eta}_s') = \mathbf{0} \quad \forall s, t = 1, \ldots, T$$

$$E(\boldsymbol{\epsilon}_t\boldsymbol{\alpha}_0') = \mathbf{0} \quad E(\boldsymbol{\eta}_t\boldsymbol{\alpha}_0') = \mathbf{0} \quad \forall t = 1, \ldots, T$$

   Various algorithms will require modifications when the first assumption is relaxed.

Definition 7.2. We refer to matrices $\mathbf{Z}_t, \mathbf{d}_t$ and $\mathbf{H}_t$ in the measurement equations and matrices $\mathbf{T}_t, \mathbf{c}_t, \mathbf{R}_t$ and $\mathbf{Q}_t$ in the transition equation as the *system matrices*.

The system matrices are assumed to be non-stochastic, unless explicitly stated otherwise; hence, while they may change with time, they change deterministically. So, the system is *linear* and for any t, $\mathbf{y}_t$ may be expressed as a linear combination of past and present $\boldsymbol{\epsilon}_t$'s and $\boldsymbol{\eta}_t$'s and $\boldsymbol{\alpha}_0$.

Definition 7.3. The model is *time invariant* or *time homogeneous* if the system matrices $\mathbf{Z}_t, \mathbf{d}_t, \mathbf{H}_t, \mathbf{T}_t, \mathbf{c}_t, \mathbf{R}_t, \mathbf{Q}_t$ do not change over time.

Stationary models form a special case and the transition equation in a time invariant model is a first-order vector autoregressive process (VAR(1)); see Prof Bénétrix's part of the course for more on VAR.

[2]The disturbance term may be redefined to have covariance matrix $\mathbf{R}_t\mathbf{Q}_t\mathbf{R}_t'$, though the representation given of the transition equation is typically more natural when $\boldsymbol{\eta}_t$ is identified with a particular set of disturbances in the model.

Example 7.4. The AR(1) plus noise model can be written as a time invariant state space model where μ_t is the state:

$$\begin{aligned} y_t &= \mu_t + \epsilon_t \quad Var(\epsilon_t) = \sigma_\epsilon^2 \\ \mu_t &= \phi\mu_{t-1} + \eta_t \quad Var(\eta_t) = \sigma_\eta^2 \end{aligned}$$

For any given statistical model, the definition of $\boldsymbol{\alpha}_t$ is determined by construction. The elements of $\boldsymbol{\alpha}$ may or may not be identifiable with components that have a substantive interpretation. The aim of the SSF is to set up $\boldsymbol{\alpha}_t$ such that $\boldsymbol{\alpha}_t$ contains all relevant information on the system for time t and that it achieves this goal by having as small a number of elements as possible. Importantly, the fact that the transition equation is a first-order process is not restrictive because higher order processes may easily be cast in the Markov form.

Example 7.5. For the AR(2), one possible state space representation is as follows:

$$y_t = (1 \;\; 0)\boldsymbol{\alpha}_t$$

$$\boldsymbol{\alpha}_t = \begin{bmatrix} y_t \\ \phi_2 y_{t-1} \end{bmatrix} = \begin{bmatrix} \phi_1 & 1 \\ \phi_2 & 0 \end{bmatrix} \boldsymbol{\alpha}_{t-1} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \epsilon_t$$

Another possible state space representation for the AR(2) model is

$$y_t = (1 \;\; 0)\boldsymbol{\alpha}_t^*$$

$$\boldsymbol{\alpha}_t^* = \begin{bmatrix} y_t \\ y_{t-1} \end{bmatrix} = \begin{bmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{bmatrix} \boldsymbol{\alpha}_{t-1}^* + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \epsilon_t$$

However, we do not need to confine ourselves to AR models in order to explore the SSF. Any ARMA model can be put into SSF; we will see this for ARMA models towards the end of this subsection. For now, let us consider the MA(1) model.

Example 7.6. Let us consider the MA(1) model

$$y_t = \epsilon_t + \theta\epsilon_{t-1} \quad t = 1, \ldots, T$$

To put this model in state space form, define the state vector $\boldsymbol{\alpha}_t = (y_t, \theta\epsilon_t)'$ and write:

$$\begin{aligned} y_t &= (1 \;\; 0)\boldsymbol{\alpha}_t \quad t = 1, \ldots, T \\ \boldsymbol{\alpha}_t &= \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \boldsymbol{\alpha}_{t-1} + \begin{bmatrix} 1 \\ \theta \end{bmatrix} \epsilon_t \end{aligned}$$

If we let $\boldsymbol{\alpha}_t = (\alpha_{1t}, \alpha_{2t})'$, then $\alpha_{2t} = \theta\epsilon_t$ and $\alpha_{1t} = \alpha_{2,t-1} + \epsilon_t = \epsilon_t + \theta\epsilon_{t-1}$. Therefore, the first element in the state is y_t, which is extracted by the measurement equation. With this representation, there is no measurement equation

noise. Also note that there are other state space representations where the dimension of the state vector is only one; however, in these representations, the cost is introducing correlation between the measurement and transition equation disturbances.

Building on our study of the SSF, we will now concentrate on filtering. Once the model is in SSF, we can apply the Kalman filter, a recursive procedure for computing the optimal estimator of the state vector conditional on all currently available information; see Kalman (1960) [52] and Kalman & Bucy (1961) [53]. Once we reach the end of the series, we can compute the optimal predictions of future observations. Similarly, we can make use of a backward recursion called smoothing to calculate optimal estimators of the state vectors at all points in time using the full sample.

Definition 7.7. Let $\mathbf{a}_t$ denote the optimal estimator of the state vector $\boldsymbol{\alpha}_t$ based on all the observations up to and including $\mathbf{y}_t$. Let $\mathbf{P}_t$ be the $m \times m$ covariance matrix of the associated estimation error, also called the *mean square error* (MSE) matrix of $\mathbf{a}_t$:

$$\mathbf{P}_t = E[(\boldsymbol{\alpha}_t - \mathbf{a}_t)(\boldsymbol{\alpha}_t - \mathbf{a}_t)']$$

Remark 7.8. Note that the MSE matrix can't be called the covariance matrix of $\mathbf{a}_t$ since elements of the vector we are interested in estimating, $\boldsymbol{\alpha}_t$ are random variables instead of fixed parameters.

Definition 7.9. At time $t-1$ with $\mathbf{a}_{t-1}$ and $\mathbf{P}_{t-1}$ given, the optimal estimator of $\boldsymbol{\alpha}_t$ is found from the *prediction equations*

$$\mathbf{a}_{t|t-1} = \mathbf{T}_t \mathbf{a}_{t-1} + \mathbf{c}_t \tag{7.3}$$

and

$$\mathbf{P}_{t|t-1} = \mathbf{T}_t \mathbf{P}_{t-1} \mathbf{T}_t' + \mathbf{R}_t \mathbf{Q}_t \mathbf{R}_t' \quad t = 1, \ldots, T \tag{7.4}$$

Note that the corresponding estimator of $\mathbf{y}_t$ is

$$\tilde{\mathbf{y}}_{\mathbf{t}|\mathbf{t-1}} = \mathbf{Z}_t \mathbf{a}_{t|t-1} + \mathbf{d}_t \quad t = 1, \ldots, T$$

Definition 7.10. The prediction error or *innovation* vector is

$$\mathbf{v}_t = \mathbf{y}_t - \tilde{\mathbf{y}}_{t|t-1} = \mathbf{Z}_t(\boldsymbol{\alpha}_t - \mathbf{a}_{t|t-1}) + \boldsymbol{\epsilon}_t \quad t = 1, \ldots, T$$

The MSE of the prediction error is:

$$\mathbf{F}_t = \mathbf{Z}_t \mathbf{P}_{t|t-1} \mathbf{Z}_t' + \mathbf{H}_t$$

We are able to update the estimator of the state as new observations become available. The prediction error vector $\mathbf{v}_t$ plays a crucial role in updating – the further the predictor of observation deviates from its realised value, the bigger the change that will be made to the estimator of the state.

Definition 7.11. The *updating equations* are:

$$\mathbf{a}_t = \mathbf{a}_{t|t-1} + \mathbf{P}_{t|t-1}\mathbf{Z}_t'\mathbf{F}_t^{-1}(\mathbf{y}_t - \mathbf{Z}_t\mathbf{a}_{t|t-1} - \mathbf{d}_t) \tag{7.5}$$

$$\mathbf{P}_t = \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{Z}_t'\mathbf{F}_t^{-1}\mathbf{Z}_t\mathbf{P}_{t|t-1} \quad t = 1, \ldots, T \tag{7.6}$$

Definition 7.12. The *Kalman filter* (KF) is made up of the prediction equations (7.3) & (7.4) and the updating equations (7.5) & (7.6).

Given the initial conditions $\mathbf{a}_0$ and $\mathbf{P}_0$, the KF produces the optimal estimator of the state as each new observation arrives. After T observations have been processed, all the information needed to make predictions of future observations are contained in the estimator $\mathbf{a}_T$.

Now that we have defined the KF, let us look at its use in prediction and smoothing. Starting with prediction, the formula for predicting more than one step ahead makes use of the updating equations. In particular, the optimal estimator of the state vector at $T + l$ based on information at time T is:

$$\mathbf{a}_{T+l|T} = \mathbf{T}_{T+l}\mathbf{a}_{T+l-1} + \mathbf{c}_{T+l} \quad l = 1, 2, \ldots$$

with $\mathbf{a}_{T|T} = \mathbf{a}_T$. The associated MSE matrix is:

$$\mathbf{P}_{T+l|T} = \mathbf{T}_{T+l}\mathbf{P}_{T+l-1|T}\mathbf{T}_{T+l}' + \mathbf{R}_{T+l}\mathbf{Q}_{T+l}\mathbf{R}_{T+l}' \quad l = 1, 2, \ldots \tag{7.7}$$

with $\mathbf{P}_{T|T} = \mathbf{P}_T$. The predictor of $\mathbf{y}_{T+l}$ is given by

$$\tilde{\mathbf{y}}_{T+l|T} = \mathbf{Z}_{T+l}\mathbf{a}_{T+l|T} + \mathbf{d}_{T+l} \quad l = 1, 2, \ldots$$

with prediction MSE given by

$$MSE(\tilde{\mathbf{y}}_{T+l|T}) = \mathbf{Z}_{T+l}\mathbf{P}_{T+l|T}\mathbf{Z}_{T+l}' + \mathbf{H}_{T+l} \tag{7.8}$$

Note that (7.8) can be used to compute prediction intervals when the model is Gaussian.

Example 7.13. For an AR(1) plus noise model of example 7.4:

$$\tilde{y}_{T+l|T} = \phi^l a_T \quad l = 1, 2, \ldots \tag{7.9}$$

So the forecast function dampens exponentially towards zero as in the vanilla AR(1) model, but it happens to start from an estimator of the unobserved component μ_T rather than from the last observation. We get the forecast MSE from solving (7.7) and substituting in (7.8):

$$MSE(\tilde{y}_{T+l|T}) = \phi^{2l}P_T + (1 + \phi^2 + \cdots + \phi^{2(l-1)})\sigma_\eta^2 + \sigma_\epsilon^2 \tag{7.10}$$

which compares to

$$MSE(\tilde{y}_{T+l|T}) = \frac{1 - \phi^{2l}}{1 - \phi^2}\sigma^2$$

Now let us apply the KF for smoothing. Remember that the objective of filtering is to estimate α_t given information available at time t while that of smoothing is to take account of information available after time t. The smoothed estimator is called the *smoother* and is denoted by $\mathbf{a}_{t|T}$. It is based on more information than the filtered estimator, so its MSE matrix $\mathbf{P}_{t|T}$ will be smaller in general than that of the filter estimator $\mathbf{P}_{T+l|T}$. For the linear model, there are three smoothing algorithms. The first, *fixed-point* smoothing, relates to smoothed estimators of the state vector at some fixed point in time; hence, it yields $\mathbf{a}_{\tau|t}$ for particular values of τ at all time periods $t > \tau$. *Fixed-lag* smoothing is concerned with smoothed estimators for a fixed delay, i.e. $\mathbf{a}_{t-j|t}$ for $j = 1, \ldots, M$, where M is some maximum lag. These algorithms are both applicable online, while *fixed-interval* smoothing, which relates to the full set of smoothed estimators for a fixed span of data is an offline technique, producing $\mathbf{a}_{t|T}, t = 1, \ldots, T$; hence fixed-interval smoothing is the most widely used algorithm for social and economic data. We focus on this last algorithm below.

The fixed-interval smoothing algorithm is a backward recursion initiated with $\mathbf{a}_T$ and $\mathbf{P}_T$ from the KF. The equations are the following:

$$\mathbf{a}_{t|T} = \mathbf{a}_t + \mathbf{P}_t * (\mathbf{a}_{t+1|T} - \mathbf{T}_{t+1}\mathbf{a}_t - \mathbf{c}_{t+1})$$
$$\mathbf{P}_{t|T} = \mathbf{P}_t + \mathbf{P}_t * (\mathbf{P}_{t+1|T} - \mathbf{P}_{t+1|t})\mathbf{P}_t{*}'$$

where $\mathbf{P}_t{*} = \mathbf{P}_t\mathbf{T}'_{t+1}\mathbf{P}^{-1}_{t+1|t} \quad t = T-1, \ldots, 1$, $\mathbf{a}_{T|T} = \mathbf{a}_T$ and $\mathbf{P}_{T|T} = \mathbf{P}_T$. The algorithm requires the storage of $\mathbf{a}_t$ and $\mathbf{P}_t$ for all t in order that they can be combined with $\mathbf{a}_{t+1|T}$ and $\mathbf{P}_{t+1|T}$. Note that state space smoothing algorithms are more general than the classical signal extraction problem because they can be used for finite samples and for systems that are not necessarily time invariant.

Initialisation of the KF is critical. When the state follows a stationary process, the initial conditions for the KF are given by its unconditional mean and variance. In example 7.13 with the AR(1) plus noise model, the unconditional mean is zero and the unconditional variance is $P_0 = \sigma^2_\eta/(1-\phi^2)$. Moving to the more general case of a stationary, time invariant transition equation of the form (7.1) & (7.2), the mean and unconditional covariance matrix are given by:

$$\mathbf{a}_0 = (\mathbf{I} - \mathbf{T})^{-1}\mathbf{c} \tag{7.11}$$
$$\text{vec}(\mathbf{P}_0) = [\mathbf{I} - \mathbf{T} \otimes \mathbf{T}]^{-1}\text{vec}(\mathbf{RQR}') \tag{7.12}$$

This remains valid even if the matrices $\mathbf{Z}_t$, $\mathbf{H}_t$ and $\mathbf{d}_t$ are not time invariant. When the state does not follow a stationary process, the initial conditions must be estimated from the observations and there are two approaches for this. One approach is to assume that $\boldsymbol{\alpha}_0$ is fixed, which implies that its distribution is degenerate since $\mathbf{P}_0 = 0$. Elements of $\boldsymbol{\alpha}_0$ must be estimated being treated as though they are unknown model parameters since $\boldsymbol{\alpha}_0$ is unknown. The other approach is to assume $\boldsymbol{\alpha}_0$ is random and has a *diffuse* distribution with

covariance matrix $\mathbf{P}_0 = \kappa\mathbf{I}$ where $\kappa \longrightarrow \infty$. However, this means that we know nothing about the initial state, so starting values are constructed from the initial observations, effectively.

Example 7.14. Say $\phi = 1$ in example (7.13) so μ_t follows a random walk. Then the KF equations produce a_1, the estimator of μ_1, which is given by

$$a_1 = a_0 + \frac{P_0 + \sigma_\eta^2}{P_0 + \sigma_\eta^2 + \sigma_\epsilon^2}(y_1 - a_0)$$

Associated with this is the MSE, given by

$$P_1 = P_0 + \sigma_\eta^2 - \frac{(P_0 + \sigma_\eta^2)^2}{P_0 + \sigma_\eta^2 + \sigma_\epsilon^2}$$

As $P_0 \longrightarrow \infty$, $a_1 = y_1$ and $P_1 = \sigma_\epsilon^2$, independently of the value of a_0. So, a diffuse prior for μ_0 produces the same result as directly using y_1 as an estimator of μ_1 having an associated estimation error associated of

$$E[(y_1 - \mu_1)^2] = E(\epsilon_t^2) = \sigma_\epsilon^2$$

We can initiate the KF with κ being a large, finite number and doing so will produce an approximation to the filter that would be found with diffuse initial conditions. However, this is not very satisfactory, especially since large numbers within the filter tends to lead to numerical instability.[3]

Let us now look at Gaussian models and the likelihood function, which can be constructed from the prediction errors from the KF that is relatively straightforward to derive in Gaussian models. In Gaussian state space models, $\boldsymbol{\epsilon}_t$, $\boldsymbol{\eta}_t$ and $\boldsymbol{\alpha}_0$ are Normally distributed. Now we will derive the KF.

The initial state is Normally distributed and has mean $\mathbf{a}_0$ and covariance matrix $\mathbf{P}_0$. Note that at $t = 1$, the state vector is

$$\boldsymbol{\alpha}_1 = \mathbf{T}_1\boldsymbol{\alpha}_0 + \mathbf{c}_1 + \mathbf{R}_1\boldsymbol{\eta}_1$$

So, $\boldsymbol{\alpha}_1$ is a linear combination of a vector of constants and two vectors of random variables, which both have (multivariate) Normal distributions and so $\boldsymbol{\alpha}_1$ is also Normal with a mean and a covariance matrix of

$$\begin{aligned}\mathbf{a}_{1|0} &= \mathbf{T}_1\mathbf{a}_0 + \mathbf{c}_1 \\ \mathbf{P}_{1|0} &= \mathbf{T}_1\mathbf{P}_0\mathbf{T}_1' + \mathbf{R}_1\mathbf{Q}_1\mathbf{R}_1'\end{aligned}$$

where $\mathbf{a}_{1|0}$ is the mean of the distribution of $\boldsymbol{\alpha}_1$ conditional on the information available at time $t = 0$. To derive the distribution of $\boldsymbol{\alpha}_1$ conditional on $\mathbf{y}_1$, write

$$\boldsymbol{\alpha}_1 = \mathbf{a}_{1|0} + (\boldsymbol{\alpha}_1 - \mathbf{a}_{1|0}) \tag{7.13}$$

$$\mathbf{y}_1 = \mathbf{Z}_1\mathbf{a}_{1|0} + \mathbf{d}_1 + \mathbf{Z}_1(\boldsymbol{\alpha}_1 - \mathbf{a}_{1|0}) + \boldsymbol{\epsilon}_t \tag{7.14}$$

[3] Ansley & Kohn (1985) and DeJong (1991) have created algorithms that overcome these problems.

Equation (7.14) is a re-arrangement of the measurement equation. From (7.13) & (7.14), it can be seen that $(\boldsymbol{\alpha}_1', \mathbf{y}_1')$, which is Normally distributed, has a mean and a covariance matrix that are given by

$$\begin{bmatrix} \mathbf{a}_{1|0} \\ \mathbf{Z}_1\mathbf{a}_{1|0} + \mathbf{d}_1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \mathbf{P}_{1|0} & \mathbf{P}_{1|0}\mathbf{Z}_1' \\ \mathbf{Z}_1\mathbf{P}_{1|0} & \mathbf{Z}_1\mathbf{P}_{1|0}\mathbf{Z}_1' + \mathbf{H}_1 \end{bmatrix}$$

From the properties of the multivariate distribution, the distribution of $\boldsymbol{\alpha}_1$ conditional on a realised value of $\mathbf{y}_1$ is multivariate normal with mean and covariance matrix given by

$$\begin{aligned} \mathbf{a}_1 &= \mathbf{a}_{1|0} + \mathbf{P}_{1|0}\mathbf{Z}_1'\mathbf{F}_1^{-1}(\mathbf{y}_1 - \mathbf{Z}_1\mathbf{a}_{1|0} - \mathbf{d}_1) \\ \mathbf{P}_1 &= \mathbf{P}_{1|0} - \mathbf{P}_{1|0}\mathbf{Z}_1'\mathbf{F}^{-1}\mathbf{Z}_1\mathbf{P}_{1|0} \end{aligned}$$

where

$$\mathbf{F}_1 = \mathbf{Z}_1\mathbf{P}_{1|0}\mathbf{Z}_1' + \mathbf{H}_1$$

The KF is obtained by repeating the above steps for $t = 2, \ldots, T$. Note that we can interpret $\mathbf{a}_t$ and $\mathbf{P}_t$ as the mean and covariance matrix of the conditional distribution of $\boldsymbol{\alpha}_t$. This conditional mean is the MMSE of $\boldsymbol{\alpha}_t$ and if $\mathbf{a}_t$ is regarded as an estimat*or* rather than an estimat*e*, then $\mathbf{a}_t$ minimises the MSE where expectation is taken over all the observations in the information set rather than being conditional on a particular set of values. Therefore, the conditional mean estimator is the MMSE of $\boldsymbol{\alpha}_t$.

As the expectation of the estimation error is zero, the estimator $\mathbf{a}_t$ is unbiased. This property is usually referred to as unconditional unbiasedness since the expectation is taken over all observations in the information set. Note also that $\mathbf{P}_t$ is the unconditional error covariance matrix associated with $\mathbf{a}_t$ since it is independent of the observations, i.e. the expectation

$$\mathbf{P}_t = E[(\boldsymbol{\alpha}_t - \mathbf{a}_t)(\boldsymbol{\alpha}_t - \mathbf{a}_t)']$$

does not need to be conditional on the realised observation up to and including time t.

However, when $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\eta}_t$ are non-Normal, it is no longer the case in general that the KF produces the conditional mean of the state vector; although, if we look at estimators that are linear combinations of observations, then $\mathbf{a}_t$ is MMSLE of $\boldsymbol{\alpha}_t$ based on observations up to and including time t and is unconditionally biased the the unconditional covariance matrix of the estimation error is $\mathbf{P}_t$, given by the KF.

Finally note that the above points apply identically to $\mathbf{a}_{t|t-1}$ and $\mathbf{P}_{t|t-1}$. In addition, the conditional mean of $\mathbf{y}_t$ at time $t-1$, *viz.* $\tilde{\mathbf{y}}_{t|t-1}$ may be interpreted as the MMSE of $\mathbf{y}_t$ in a Gaussian model and otherwise is the MMSLE.

Regarding estimation via maximum likelihood, system matrices in state space models typically depend on unknown parameters, say the $n \times 1$ vector $\boldsymbol{\Psi}$ called *hyperparameters*.

Example 7.15. With the AR(1) plus noise model of example 7.4, the hyperparameters are σ_η^2, ϕ and σ_ϵ^2.

To estimate the hyperparameters – via maximum likelihood – we use the Kalman filter to construct the likelihood function and maximise this using some numerical optimisation procedure. For a multivariate model, the joint density of a set of T observations expressed in terms of conditional distributions is

$$L(\mathbf{y};\mathbf{\Psi}) = \Pi_{t=1}^{T} p(\mathbf{y}_t|Y_{t-1})$$

where we let $p(\mathbf{y}_t|Y_{t-1})$ be the distribution of y_t conditional on the information available at time $t-1$, i.e. $Y_{t-} = \{\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \ldots, \mathbf{y}_1\}$. Recall that conditional on Y_{t-1}, $\boldsymbol{\alpha}_t \sim N(\mathbf{a}_{t|t-1}, \mathbf{P}_{t|t-1})$. The measurement equation is

$$\mathbf{y}_t = \mathbf{Z}_t\mathbf{a}_{t|t-1} + \mathbf{Z}_t(\boldsymbol{\alpha}_t - \mathbf{a}_{t|t-1}) + \mathbf{d}_t + \boldsymbol{\epsilon}_t$$

from which it is clear that the conditional distribution of y_t is Normal, having a mean of $\tilde{y}_{t|t-1} = \mathbf{Z}_t\mathbf{a}_{t|t-1} + \mathbf{d}_t$ and a covariance matrix of $\mathbf{F}_t = \mathbf{Z}_t\mathbf{P}_{t|t-1}\mathbf{Z}_t' + \mathbf{H}_t$. Thus, for a Gaussian model, the log-likelihood function is

$$\log L(\mathbf{\Psi}) = -\frac{NT}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T}\log|\mathbf{F}_t| - \frac{1}{2}\sum_{t=1}^{T}\mathbf{v}_t'\mathbf{F}_t^{-1}\mathbf{v}_t \tag{7.15}$$

where $\mathbf{v}_t$ denotes the vector of prediction errors, i.e.

$$\mathbf{v}_t = \mathbf{y}_t - \tilde{\mathbf{y}}_{t|t-1} = \mathbf{Z}_t(\boldsymbol{\alpha} - \mathbf{a}_{t|t-1}) + \boldsymbol{\epsilon}_t \quad t = 1, \ldots, T$$

Equation (7.15) is called the *prediction error decomposition* form of the likelihood.

We can reparameterise a univariate model so $\mathbf{\Psi} = (\mathbf{\Psi}_*', \sigma_*^2)'$ where we let $\mathbf{\Psi_*}$ be a vector of $n-1$ parameters and a scale factor σ_*^2 (usually the variance of one distribution in the model). The measurement equation is

$$y_t = \mathbf{z}_t'\boldsymbol{\alpha}_t + d_t + \epsilon_t \quad Var(\epsilon_t) = \sigma_*^2 h_t \quad t = 1, \ldots, T \tag{7.16}$$

where h_t is a scalar and $\mathbf{z}_t$ is an $m \times 1$ vector. The only change we make for the transition equation is to redefine the covariance matrix of the disturbance $\boldsymbol{\eta}_t$ to $\sigma_*^2\mathbf{Q}_t$. The Kalman filter runs independently of σ_*^2, if the initial covariance matrix $\mathbf{P}_0$ is also specified up to the factor of proportionality (σ_*^2), i.e. $Var(\boldsymbol{\alpha}_0) = \sigma_*^2\mathbf{P}_0$. We do this since we can concentrate σ_*^2 out of the likelihood function, if it is an unknown parameter. Prediction errors are invariant to the omission of σ_*^2 from the Kalman filter. Their variances are:

$$Var(v_t) = \sigma_*^2 f_t \tag{7.17}$$

Relative to (7.15), $\mathbf{F}_t = \sigma_*^2 f_t$, so:

$$\begin{aligned}\log L(\mathbf{\Psi}_*, \sigma_*^2) = &-\frac{T}{2}\log 2\pi - \frac{T}{2}\log\sigma_*^2 \\ &-\frac{1}{2}\sum_{t=1}^{T}\log f_t - \frac{1}{2\sigma_*^2}\sum_{t=1}^{T}\frac{v_t^2}{f_t}\end{aligned} \tag{7.18}$$

Differentiating (7.18) this with respect to σ_*^2 noting that v_t and f_t are independent of σ_*^2 and putting this equal to zero yields

$$\tilde{\sigma}_*^2(\boldsymbol{\Psi}_*) = \frac{1}{T}\sum_{t=1}^{T}\frac{v_t^2}{f_t} \tag{7.19}$$

where $\tilde{\sigma}_*(\boldsymbol{\Psi}_*)$ is the ML estimator of σ_*^2 given the value $\boldsymbol{\Psi}_*$. Substituting $\tilde{\sigma}_*(\boldsymbol{\Psi}_*)$ into (7.18) yields the concentrated log-likelihood function

$$\log L_c(\boldsymbol{\Psi}_*) = -\frac{T}{2}(\log 2\pi + 1) - \frac{1}{2}\sum_{t=1}^{T}\log f_t - \frac{T}{2}\log\tilde{\sigma}_*^2(\boldsymbol{\Psi}_*)$$

We can either maximise this with respect to the elements of $\boldsymbol{\Psi}_*$ or minimise the sum of squares function

$$S(\boldsymbol{\Psi}_*) = \left(\Pi_{t=1}^{T} f_t\right)\sum_{t=1}^{T}\left(\frac{v_t^2}{f_t}\right)$$

Example 7.16. The AR(1) plus noise model in example 7.4 can be reparameterised by allowing σ_ϵ^2 to be σ_*^2:

$$Var(\epsilon_t) = \sigma_\epsilon^2 \quad Var(\eta_t) = \sigma_\epsilon^2 q$$

Thus, $\boldsymbol{\Psi}_*$ contains only q, which is the signal to noise ratio. Note that the Kalman filter depends only on q with the initial conditions such that instead of $P_0 = \frac{\sigma_\eta^2}{1-\phi^2}$, we have

$$P_0 = \frac{q}{1-\phi^2}$$

So, the Kalman filter can be initialised with the mean and covariance matrix of the unconditional distribution of $\boldsymbol{\alpha}_t$ when $\boldsymbol{\alpha}_t$ is stationary. With non-stationary state vectors, we can form a likelihood from all T prediction errors as in (7.15) only if we have prior information where $\boldsymbol{\alpha}_0$ has a proper distribution with a known mean $\mathbf{a}_0$ and bounded covariance matrix $\mathbf{P}_0$. Such an initial state has a *diffuse prior*. Typically if the state vector has d non-stationary elements, then a proper distribution for the state can be constructed at time $t = d$ using the first d observations. Then summing (7.18) from $t = d+1$ instead of $t = 1$ gives the joint density function of $y_{d+1}, \ldots, y_T$ conditional on $y_1, \ldots, y_d$.

Example 7.17. If the parameter ϕ in example 7.4 is not necessarily less than one in absolute value and if σ_ϵ^2 is zero so there is no measurement error, then we can construct a Gaussian likelihood function conditional on the first observation and the ML estimator is given by regressing y_t on y_{t-2} for $t = 2, \ldots, T$. If $\sigma_\epsilon^2 \neq 0$, then using a diffuse prior for μ_0 implies that μ_1 will have the proper distribution

$$\mu_1 \sim N(y_1, \sigma_\epsilon^2)$$

Then we can construct the likelihood from the prediction error decomposition (7.18) with the summation running from 2 to T instead of from 1 to T.

Instead to handle the initialisation problem if we treat $\boldsymbol{\alpha}_0$ as fixed parameters to be estimated, then we can concentrate these parameters out of the likelihood. However, the properties of these estimators are not as appealing as those from the diffuse prior approach. Moreover, the diffuse prior likelihood is the marginal likelihood for the model holding $\boldsymbol{\alpha}_0$ fixed. Generally, marginal likelihood is recommended with models having nuisance parameters.

Example 7.18. Continuing from example 7.4, assume again that ϕ is not necessarily less than one in absolute value and assume that $\sigma_\epsilon^2 = 0$. Let y_0 be an unknown parameter. The likelihood function will be

$$\log L(\phi, \sigma_\eta^2) = -\frac{T}{2}\log 2\pi - \frac{T}{2}\log \sigma_\eta^2 - \frac{1}{2\sigma_\eta^2}\sum_{t=1}^{T}(y_t - \phi y_{t-1})^2$$

Note that the ML estimator of y_0 is $\frac{y_1}{\phi}$, while the ML estimator of ϕ does not change. All that changes here is that the estimator of σ_η^2 is now divided by T rather than by $T-1$. If we instead assume that $\sigma_\epsilon^2 \neq 0$ and assume $\phi = 1$, the estimate of q displays some troublesome characteristics when μ_0 is treated as an unknown parameter.

Considering residuals and diagnostic checking, we know that in a Gaussian model, the innovations $\mathbf{v}_t \sim NID(\mathbf{0}, \mathbf{F}_t),\ t = 1, \ldots, T$. Therefore, the standardised residuals $\mathbf{F}_t^{-\frac{1}{2}}$ can be useful for diagnostic checking; however, if system matrices contain unknown hyperparameters replaced by estimators, then we do not necessarily have exactly that $\mathbf{v}_t \sim NID(\mathbf{0}, \mathbf{F}_t),\ t = 1, \ldots, T$. With a univariate model formulated as in (7.16) having d non-stationary elements in the state, from (7.17) we get that

$$\tilde{v}_t = \frac{v_t}{\sqrt{f_t}} \sim NID(0, \sigma_*^2) \quad t = d+1, \ldots, T$$

Without the normality assumption, the mean of $\mathbf{v}_t$ is $\mathbf{0}$ and the covariance matrix at time t is $\mathbf{F}_t$; innovations are uncorrelated across time. While other residuals that can be constructed will not enjoy the independence property of the innovations, they can be useful diagnostics. For instance, we can detect outliers and structural breaks by looking at the estimator of the measurement equation noise, $\boldsymbol{\epsilon}_t$ calculated from the smoother as

$$\tilde{\boldsymbol{\epsilon}}_{t|T} = \mathbf{y}_t - \mathbf{Z}_t\mathbf{a}_{t|T} - \mathbf{d}_t$$

and the estimator the transition equation noise, $\boldsymbol{\eta}_t$ calculated from the smoother as

$$\tilde{\boldsymbol{\eta}}_{t|T} = \mathbf{a}_{t|T} - \mathbf{T}_t\mathbf{a}_{t-1|T} - \mathbf{c}_t \quad t = 1, \ldots, T$$

Remark 7.19. The state space framework handles missing observations problems with ease, e.g. if an observation is missing at time $t = \tau$, then

$$\mathbf{a}_t = \mathbf{a}_{t|t-1} \quad \text{and} \quad \mathbf{P}_t = \mathbf{P}_{t|t-1}$$

The prediction and updating equations can be applied where the primer yields a two-step ahead predictor and the latter can be used once $\mathbf{y}_{t+1}$ becomes available. The smoother estimates the state at time τ. So, we can compute the estimator of the missing observation along with its MSE. The likelihood function is computed as standard with innovations from the Kalman filter, except that there will be no innovation associated with the missing observation. There will be no conditional distribution at time τ and the distribution at time $\tau + 1$ will be conditional on the information at time $\tau - 1$.

Example 7.20. With the AR(1) plus noise model, the distribution of $y_{\tau+1}|Y_{\tau-1}$ will be normal with mean $\phi^2 a_{\tau-1}$ and variance $\phi^4 P_{\tau-1} + (1+\phi^2)\sigma_\eta^2 + \sigma_\epsilon^2$. This follows from (7.9) & (7.10).

Remark 7.21 (Remark 7.19 continued)**.** We can generalise the results to any linear Normal state space model for any number of missing observations (and where they are located). So, if m consecutive observations are missing, then the mean and variance of an $m+1$-step ahead predictive distribution will taken into the likelihood. In addition, the Kalman filter can deal with cases where observations are aggregated intertemporally and where we can only observe the aggregate, e.g. if a flow variable (national income say) has some annual observations and some quarterly observations, we can solve this by extending the state vector to include a variable that is able to cumulate the series at points where the series is unobserved; this phenomenon of quarterly observations following annual observations is inherent in many national income and product accounting data sets internationally and is an issue most empiricists must deal with. The state space approach is capable of handling other data irregularities too, such as when there are data revisions by government agencies to time series.

We will conclude our section on state space forms by working through an example – an application to ARMA models. We can construct an exact likelihood function of an ARMA model via the Kalman filter. First for ARMA(p, q), by defining

$$m = \max(p, q + 1)$$

we can write the ARMA(p, q) as

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_m y_{t-m} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_{m-1}\epsilon_{t-m+1} \tag{7.20}$$

where unless $p = q+1$, some coefficients will be zero. We can represent (7.20) in a Markovian way by defining $\boldsymbol{\alpha}_t$ to be an $m \times 1$ vector following the multivariate

AR(1) model:

$$\boldsymbol{\alpha}_t = \begin{bmatrix} \phi_1 & \\ \phi_2 & \\ \vdots & \mathbf{I}_{m-1} \\ & \\ \phi_m & \mathbf{0}' \end{bmatrix} \boldsymbol{\alpha}_{t-1} + \begin{bmatrix} 1 \\ \theta_1 \\ \vdots \\ \vdots \\ \theta_{m-1} \end{bmatrix} \epsilon_t$$

which can be see to be a transition equation (7.1)-(7.2) where $\mathbf{T}_t$ and $\mathbf{R}_t$ are constant and $\mathbf{Q}_t = \sigma^2$. We can easily recover the original ARMA model since the first element of $\boldsymbol{\alpha}_t$ is equal to y_t, which can be seen by repeated substitution; the measurement equation extracts the first element of the state vector; so:

$$y_t = \mathbf{z}_t{}'\boldsymbol{\alpha}_t \quad t = 1, \ldots, T$$

where we define $\mathbf{z}_t' = (1\ \mathbf{0}_{m-1}{}')'$. The disturbance term is zero unless the model includes measurement error.

Example 7.22 (MA(1))**.** When the model is stationary, we get the initial conditions for the state from (7.11) and (7.12) and the likelihood via the Kalman filter in the form (7.18). Looking at an MA(1), the transition equation is

$$\boldsymbol{\alpha}_t = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \boldsymbol{\alpha}_{t-1} + \begin{bmatrix} 1 \\ \theta \end{bmatrix} \epsilon_t$$

and the initial state vector is $\mathbf{a}_0 = \mathbf{a}_{1|0} = \mathbf{0}$. As $\boldsymbol{\alpha}_t = (y_t, \theta\epsilon_t)'$, the initial matrix $\mathbf{P}_0 = \mathbf{P}_{1|0}$ is

$$\mathbf{P}_{1|0} = \mathbf{P}_0 = \frac{1}{\sigma^2} E(\boldsymbol{\alpha}_t \boldsymbol{\alpha}_t{}') = \begin{bmatrix} 1+\theta^2 & \theta \\ \theta & \theta^2 \end{bmatrix}$$

The first prediction error is $v_1 = y_1$ and $f_1 = 1 + \theta^2$. Updating, we get

$$\mathbf{a}_1 = \begin{pmatrix} y_1 \\ \frac{\theta y_1}{1+\theta^2} \end{pmatrix} \quad \text{and} \quad \mathbf{P}_1 = \begin{pmatrix} 0 & 0 \\ 0 & \frac{\theta^4}{1+\theta^2} \end{pmatrix}$$

We get the following prediction equations for α_2:

$$\begin{aligned} \mathbf{a}_{2|1} &= \begin{pmatrix} \frac{y_1\theta}{1+\theta^2} \\ 0 \end{pmatrix} \\ \text{and} \quad \mathbf{P}_{2|1} &= \begin{pmatrix} \frac{\theta^4}{1+\theta^2} & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 1 & \theta \\ \theta & \theta^2 \end{pmatrix} \\ &= \begin{pmatrix} \frac{1+\theta^2+\theta^4}{1+\theta^2} & \theta \\ \theta & \theta^2 \end{pmatrix} \\ \therefore v_2 &= y_2 - \frac{\theta y_1}{1+\theta^2} \\ \text{and} \quad f_2 &= \frac{1+\theta^2+\theta^4}{1+\theta^2} \end{aligned}$$

Further repetition reveals that the Kalman filter essentially computes the prediction errors from the following recursion:

$$v_t = y_t - \frac{\theta v_{t-1}}{f_{t-1}} \quad t = 1, \ldots, T$$

where we initialise $v_0 = 0$ and define

$$f_t = 1 + \frac{\theta^{2t}}{1 + \theta^2 + \cdots + \theta^{2(t-1)}}$$

Finally, note that while there are efficient filtering algorithms for ARMA models, computing the exact likelihood is always more time consuming than computing the conditional sum of squares (CSS). Moreover, given the availability of analytic derivatives for the CSS case, we can carry out numerical optimisation more efficiently. Nonetheless, exact ML enjoys statistical advantages especially when MA parameters lie near or on the boundary of the invertibility region.

7.2 Frequency Domain Approach

In order to study frequency-related filters, we will now shift focus to frequency domain time series. Lots of early research in time series was carried out in the frequency domain, especially since other sciences contributed here. Today, while such courses are often left to more advanced masters/PhD courses since they can form the basis for advanced simulation/filtering time series second year PhD 'field courses' internationally, it seems to be the case that most leading theoretical econometric research in time series is now conducted within the time series domain approach as opposed to the frequency domain approach since most of the issues within the frequency domain approach have been worked out a long time ago! Most of the spectral preliminaries and applications including linear filtering theory is relatively old. We will study the frequency domain approach in order to get a taste of a somewhat complete, chronological, history of thought approach. This section on frequency domain descriptive statistics is layed out as follows: (i) a review of complex analysis; (ii) a review of elementary time-series; (iii) spectral representation theorems of stationary processes; (iv) spectral properties of filters; (v) multivariate spectra; (vi) spectral estimation; and further frequency related filtering.

7.2.1 Complex analysis

Data preparation is a critical part of all empirical work. DeJong & Dave (2011) [21] characterise three steps in this process. Firstly, there must be a clearly established correspondence between what is being modeled and what the data measures. The remaining two steps involve removing trends and isolating cycles. If the model we are looking at (say a business cycle model) focuses on cyclical behaviour, then we must remove the trend from a time series

that involves both trends and cycles. Even after removing the trend, sometimes we must go further to isolate cycles by their recurring frequency. If the model is supposed to describe patterns of medium run, or business cycle fluctuations (i.e. 6-40 quarters), then we are not interested in seasonal fluctuations, so we need to deseasonalise the data. We will return to the discussion of trend removal in the last chapter (filtering), briefly mentioning three approaches to trend removal: detrending, differencing and filtering (e.g. Hodrick-Prescott (HP) filter, band pass filter). Isolating cycles motivates an understanding of the frequency domain, which in turn begets a digression into complex analysis.[4] We will return to discuss filters used to isolate cycles such as the HP filter and the band pass filter in addition to discussing seasonal adjustment in the chapter on filtering.

Definition 7.23. An *imaginary variable* i is such that

$$i^2 = -1$$

Definition 7.24. A *complex* variable $z \in \mathbb{C}$ is one that can be represented as

$$z = x + iy$$

where $x \in \mathbb{R}, y \in \mathbb{R}$ and $x = Re(z), y = Im(z)$, where Re and Im stand for real and imaginary components, respectively. This is the rectangular coordinates representation of z.

Definition 7.25. The *modulus* of a complex number z is the distance of z from the origin:

$$\begin{aligned}\sqrt{x^2 + y^2} &= \sqrt{(x + iy)(x - iy)} \\ &\equiv |z|\end{aligned}$$

If $|z| = 1$, then z can be said to lie on the unit circle.

Definition 7.26. The *complex conjugate* of $z \in \mathbb{C}$, $z = x + iy$ is $x - iy$.

Another representation of z is in terms of polar coordinates. Here ω denotes the radian angle of z, i.e. the distance in radians counterclockwise from the x-axis to z; see diagram. Here z is represented by

$$z = |z|(\cos\omega + i\sin\omega) = |z|e^{i\omega} \tag{7.21}$$

Proof.

$$WTS : \cos\omega + i\sin\omega = e^{i\omega}$$

Taking first order Taylor series expansions around zero:

$$e^{i\omega} \approx e^0 + ie^0\omega = 1 + i\omega$$

[4] Brown & Churchill (2003) [12] and Palka (1991) [68] are decent undergraduate and graduate references on this topic. Regarding extra references on isolating cycles, see also Sargent (1987) [75], Harvey (1993) [46], Hamilton (1994) [43] and Kaiser & Maravall (2001) [51].

$$\cos(\omega) \approx \cos(0) - \omega \sin(0) = 1$$
$$\sin(\omega) \approx \sin(0) + \omega \cos(0) = \omega$$
$$\therefore \cos(\omega) + i \sin(\omega) \approx 1 + i\omega \qquad \square$$

Corollary 7.27.

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

See figure 6.5 in DeJong & Dave (2011) [21].

Theorem 7.28 (DeMoivre's Theorem)**.**

$$z^j = |z|^j e^{i\omega j} = |z|^j (\cos(\omega j) + i \sin(\omega j))$$

Proof. Proof is trivial, from (7.21). $\square$

Definition 7.29. The *square summability condition* requires that for a sequence of complex numbers $\{a_j\}_{j=-\infty}^{\infty}$

$$\sum_{j=-\infty}^{\infty} |a_j|^2 < \infty$$

Definition 7.30. A series $\{\lambda_k\}_{k=0}^{\infty}$ is *absolutely summable* if $\sum_{i=0}^{\infty} |\lambda_k| < \infty$.

Theorem 7.31 (Riesz-Fischer Theorem)**.** *For any sequence of complex numbers $\{a_j\}_{j=-\infty}^{\infty}$ satisfying the square summability condition, there exists a complex function $f(\omega)$:*

$$f(\omega) = \sum_{j=-\infty}^{\infty} a_j e^{-i\omega j} \quad \omega \in [-\pi, \pi] \tag{7.22}$$

Definition 7.32. The *Fourier transform* of $\{a_j\}_{j=-\infty}^{\infty}$ is $f(\omega)$ in (7.22).

Remark 7.33. Given $f(\omega)$, the inverse to the Fourier transform yields $\{a_j\}_{j=-\infty}^{\infty}$:

$$a_j = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\omega) e^{i\omega j} d\omega$$

This is known as the *Fourier inversion theorem.*

Two properties of Fourier transform relate to additivity of elements and multiplication by a scalar. For any two complex functions $f(\omega)$ and $g(\omega)$ such that

$$f(\omega) = \sum_{j=-\infty}^{\infty} a_j e^{-i\omega j}$$
$$g(\omega) = \sum_{j=-\infty}^{\infty} b_j e^{-i\omega j}$$

we have that

$$f(\omega) + g(\omega) = \sum_{j=-\infty}^{\infty} (a_j + b_j)e^{-i\omega j}$$
$$\alpha f(\omega) = \sum_{j=-\infty}^{\infty} \alpha a_j e^{-i\omega j}$$

i.e. the Fourier transform of the sum of sequences is the sum of the Fourier transforms of the individual sequences and the Fourier transform of $\{\alpha a_j\}_{j=-\infty}^{\infty}$ is α times the Fourier transform of $\{a_j\}_{j=-\infty}^{\infty}$.

Now consider

$$y_t^{\omega} = \alpha(\omega)\cos(\omega t) + \beta(\omega)\sin(\omega t) \tag{7.23}$$

where $\alpha(\omega)$ and $\beta(\omega)$ are uncorrelated random variables with zero mean and identical variances. Here ω determines the frequency with which $\cos(t)$ completes a cycle relative to $\cos(t)$ as t changes (0 to 2π, 2π to 4π, etc. – t is fixed so the frequency of $\cos(t)$ would be 1). Now consider a time series y_t constructed from a continuum of y_t^{ω}'s where ω varies over $[0, \pi]$:[5]

$$y_t = \int_0^{\pi} \alpha(\omega)\cos(\omega t)d\omega + \int_0^{\pi} \beta(\omega)\sin(\omega t)d\omega \tag{7.24}$$

Theorem 7.34 (Spectral/Cramér Representation Theorem)**.** *Given the appropriate specifications for $\alpha(\omega)$ and $\beta(\omega)$, uncorrelated, zero-mean random variables with identical variances, any time series y_t may be represented as in (7.24), i.e. y_t is represented as resulting from the influence of a continuum of cyclical components of different frequencies.*

Definition 7.35. The *spectrum* of a time series y_t is a tool measuring the effect of the cyclical components y_t^{ω} over the continuum $[0, \pi]$ to the overall variance of y_t. In particular, the spectrum is a frequency decomposition of the variance of y_t. To see this, consider the autocovariance $\gamma(\tau)$ between y_t and $y_{t+\tau}$ where $E(y_t) = \mu_t$ and $\gamma(0) = Var(y_t)$. Assuming the sequence $\{\gamma(\tau)\}_{\tau=-\infty}^{\infty}$ is square-summable, the Fourier transform exists by the Riesz-Fischer Theorem and is given by

$$f_y(\omega) = \sum_{\tau=-\infty}^{\infty} \gamma(\tau)e^{-i\omega t} \tag{7.25}$$

By the Fourier inversion theorem:

$$\gamma(\tau) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f_y(\omega)e^{i\omega t}d\omega \tag{7.26}$$

The *(power) spectrum* or *spectral density function* is defined by

$$s_y(\omega) = \frac{1}{2\pi}f_y(\omega) \tag{7.27}$$

[5] We can concentrate on $[0, \pi]$ due to the symmetry of the sine and cosine functions between $t \in [0, \pi]$ and $t \in [\pi, 2\pi]$.

Alternatively, the spectrum can be represented by:

$$s_y(\omega) = \left(\frac{1}{2\pi}\right)\left[\gamma(0) + 2\sum_{\tau=1}^{\infty}\gamma(\tau)\cos(\omega\tau)\right] \quad \omega \in [0,\pi] \tag{7.28}$$

Definition 7.36. The Fourier transform with a series of autocovariances (7.25), i.e. where $a_\tau = \gamma(\tau)$ is known as the *autocovariance generating function* for the time series y_t.

Remark 7.37. From (7.26) and (7.27), we can see how the spectrum may be interpreted as a frequency decomposition of $Var(y_t)$. Note:

$$\gamma(\tau) = \int_{-\pi}^{\pi} s_y(\omega)cos(\omega\tau)d\omega$$

So the sequences of autocovariances and the spectrum are two different ways of looking at y_t – time domain in the first case and frequency domain in the second case, which gives rise to the name for this analysis. When $\tau = 0$:

$$\gamma(0) = \int_{-\pi}^{\pi} s_y(\omega)d\omega$$

and the relative importance of fluctuations at different frequencies in influencing variations in y_t can be obtained from comparing the height of $s_y(\omega)$ for different values of ω. So, the total variance can also be interpreted as the area under the spectral density. Again from symmetry of the cosine function and integrating over only some frequencies:

$$\frac{2}{\gamma(0)}\int_0^{\omega_j} s_y(\omega)d\omega = \lambda(\omega_j) \quad 0 < \omega_j \leq \pi, 0 < \lambda(\omega_j) \leq 1$$

$\lambda(\omega_j)$ may be interpreted as the proportion of total variance of y_t due to frequencies no greater than ω_j.

Example 7.38. For the ARMA(p,q):

$$\Phi(L)(y_t - \mu) = \Theta(L)\epsilon_t$$

the autocovariance generating function (ACGF) is given by

$$f_y(\omega) = \frac{\sigma^2\Theta(e^{i\omega})\Theta(e^{-i\omega})}{\Phi(e^{i\omega})\Phi(e^{-i\omega})} = \sigma^2\Pi(e^{i\omega})\Pi(e^{-i\omega})$$

where $\Pi(e^{i\omega})$ represents the sequence of coefficient in the MA(∞) representation of the series, i.e. $\Pi(e^{i\omega}) = \frac{\Theta(e^{i\omega})}{\Phi(e^{i\omega})}$. The spectral density can be derived from this relationship as:

$$s_y(\omega) = \frac{\sigma^2}{2\pi}\Pi(e^{i\omega})\Pi(e^{-i\omega})$$

Example 7.39. For an AR(1) process:

$$y_t = \rho y_{t-1} + \epsilon_t \quad \epsilon_t \sim N(0,1)$$

the lag polynomials are $\Theta(e^{i\omega}) = 1$ and $\Phi(e^{i\omega}) = 1 - \rho e^{i\omega}$. The ACGF is

$$\begin{aligned} f_y(\omega) &= \frac{\sigma^2}{(1-\rho e^{i\omega})(1-\rho e^{-i\omega})} \\ &= \frac{\sigma^2}{1+\rho^2-\rho(e^{i\omega}+e^{-i\omega})} \\ &= \frac{\sigma^2}{1+\rho^2}\sum_{i=0}^{\infty}\left(\frac{\rho}{1+\rho^2}\right)^i\left(\frac{1+e^{2i\omega}}{e^{i\omega}}\right)^i \end{aligned}$$

The spectrum is

$$s_y(\omega) = \frac{\sigma^2}{2\pi[1-\rho e^{-i\omega}][1-\rho e^{i\omega}]} = \frac{\sigma^2}{2\pi[1+\rho^2-2\rho\cos(\omega)]}$$

Definition 7.40. The *period* p of y_t^ω is defined as the number of units of time necessary for y_t^ω in (7.23) to complete a cycle: $p = \frac{2\pi}{\omega}$.

The period helps us to interpret frequency in units of time. Similarly, $\frac{1}{p} = \frac{\omega}{2\pi}$ is the number of cycles completed by y_t^ω per period.

Example 7.41. For a period representing a quarter, a 10 year (40 quarter) cycle has an associated frequency of $\omega = \frac{2\pi}{40} = 0.157$. A six quarter cycle has a frequency of $\omega = \frac{2\pi}{6} = 1.047$ Therefore, business cycle frequencies lie within the interval $[0.157, 1.047]$.

So far, we have concentrated on theoretical results. What about empirical counterparts? Obviously, the lowest frequency we have will be once in the entire sample period we have of y_t from $t = 1, \ldots, T$. Therefore, we can map ω_1 to $\frac{2\pi}{T}$. Likewise, the highest frequency would be $\omega = 2\pi$ and the intermediate values will be $\frac{2\pi j}{T}, j = 2, \ldots, T-1$. The number of periods per cycle is $T/j = 2\pi/\omega_j$, with the lowest frequency ω_1 corresponding to the highest period T dates (e.g. days, months, quarters, years, etc.).

Definition 7.42. Denote the sample autocovariance at lag τ by c_τ:

$$c_\tau = c_{-\tau} = \frac{1}{T}\sum_{t=\tau+1}^{T}(y_t - \bar{y})(y_{t-\tau} - \bar{y})$$

where

$$\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t \quad \tau = 0, 1, \ldots, T-1$$

Then the sample counterpart of the population spectral density function called the *sample periodogram* is

$$\hat{s}_y(\omega) = \frac{1}{2\pi}\left[c_0 + 2\sum_{\tau=1}^{T-1} c_\tau \cos\omega\tau\right]$$

The problem with estimating the sample periodogram is that there are always zero degrees of freedom since we are estimating T parameters (variance and $T-1$ autocovariances) with T observations. Truncation and windowing have been suggested as remedies for this problem. Truncation is based on a subset of the first $L < T$ autocovariances; the choice of L is subjective, though Chatfield (1996) [13] suggests $L \approx 2\sqrt{T}$. The set of weights $\{w_\tau, \tau = 0, \ldots, L\}$ called a *lag window* leads to the revised estimator:

$$\hat{s}_y(\omega) = \frac{1}{2\pi}\left[w_0 c_0 + 2\sum_{\tau=1}^{L} w_\tau c_\tau \cos(\omega\tau)\right]$$

The Bartlett window is one choice of weights:

$$\hat{s}_{y,\text{Bartlett}}(\omega) = \frac{1}{2\pi}\left[c_0 + 2\sum_{\tau=1}^{L} w(\tau, L) c_\tau \cos(\omega\tau)\right] \quad w(\tau, L) = 1 - \frac{\tau}{L+1}$$

The Tukey window has $w_\tau = \frac{1}{2}[1 + \cos(\pi\tau/L)]$. The Parzen window has

$$w_k = \begin{cases} 1 - 6[(\tau/L)^2 - (\tau/L)^3] & \tau \leq L/2 \\ 2(1-\tau/L)^3 & \text{else} \end{cases}$$

With the empirical estimate of the spectrum, the variance decomposition may be approximated via summing the values about the frequencies of interest.

Example 7.43. See example 21.3 in Greene (2011) [42]. This is a very instructive example.

High frequency, disaggregated (across time and individuals) data especially in financial econometrics with microlevel data are generally not smooth and tricky to analyse, e.g. stock market data. Here, tools of spectral analysis and the frequency domain have proved useful to analysts.

Remark 7.44. On a computational note, the discrete Fourier transform of the series of autocovariances used in the both definitions of the spectrum, (7.27) & (7.28) involves computations on the order of T^2 sets of computations. This is troublesome for large data sets with series having many thousands of observations (e.g. daily stock returns). In this case, the fast Fourier transform may be advantageous as it reduces the computational level to $O(T\log_2 T)$. MATLAB has FFT (fast Fourier transform) algorithms in versions of `fft`.

Before proceeding further, we will first present a quick review of elementary time-series to establish notation for the remainder of this chapter.

7.2.2 Time-Series Review

The terminology for this section is presented below.

1. $\{Y_t\}$: a sequence of random variables.
2. 'Stochastic Process': the probability law governing $\{Y_t\}$
3. 'Realisation': a single draw from the process, $\{y_t\}$.
4. 'Strict stationarity': the process is strictly stationary if the probability distribution of $(Y_t, Y_{t+1}, \ldots, Y_{t+k})$ is identical to the probability distribution of $(Y_t, Y_{t+1}, \ldots, Y_{t+k})$ for all t, τ and k. Therefore, all joint distributions are invariant to time.
5. 'Autocovariances': $\gamma_{t,k} = cov(Y_t, Y_{t+k})$.
6. 'Autocorrelations': $\rho_{t,k} = corr(Y_t, Y_{t+k})$.
7. 'Covariance Stationarity': the process is covariance stationary if $\mu_t = E(Y_t) = \mu$ and $\gamma_{t,k} = \gamma_k$ for all t and k.
8. 'White noise' (WN): a process is called white noise if it is covariance stationary and $\mu = 0$ and $\gamma_k = 0$ for $k \neq 0$.
9. 'Martingale': Y_t follows a martingale process if $E(Y_{t+1}|F_t) = Y_t$, where $F_t \subseteq F_{t+1}$ is the time t information set.
10. 'Martingale Difference Process': Y_t follows a martingale process if $E(Y_{t+1}|F_t) = 0$. $\{Y_t\}$ is called a martingale difference sequence or 'mds'.
11. 'Lag Operator': L lags the elements of a sequence by one period, e.g. $Ly_t = y_{t-1}$, $L^2 y_t = y_{t-2}$, $\ldots$, $L^k y_t = y_{t-k}$. Note that $bLY_t = L(bY_t) = bY_{t-1}$ if b is a constant.
12. 'Linear filter': letting $\{c_j\}$ be a sequence of constants and
$$c(L) = c_{-r}L^{-r} + c_{-r+1}L^{-r+1} + \cdots + c_0 + c_1 L + \cdots + c_s L^s$$
be a polynomial in L. Note that $X_t = c(L)Y_t = \sum_{j=-r}^{s} c_j Y_{t-j}$ is a moving average of Y_t. Sometimes, we can refer to $c(L)$ as a 'linear filter' (see below) and X is called a *filtered version of* Y.
13. 'AR(p) process': $\phi(L)Y_t = \epsilon_t$ where
$$\phi(L) = (1 - \phi_1 L - \cdots - \phi_p L^p)$$
and ϵ_t is WN.
14. 'MA(q) process': $Y_t = \theta(L)\epsilon_t$ where
$$\theta(L) = (1 - \theta_1 L - \cdots - \theta_q L^q)$$
and ϵ_t is WN.

15. 'ARMA(p,q)': $\phi(L)Y_t = \theta(L)\epsilon_t$.

16. 'Wold decomposition theorem' (e.g. Brockwell & Davis (1991) [11]): suppose Y_t is generated by a linearly indeterministic covariance stationary process. Then Y_t can be represented as

$$Y_t = \epsilon_t + c_1\epsilon_{t-1} + c_2\epsilon_{t-2} + \cdots$$

where ϵ_t is WN with variance σ^2_ϵ, $\sum_{i=1}^{\infty} c_i^2 < \infty$ and $\epsilon_t = Y_t - Proj(Y_t|\text{lags of } Y_T)$ so that ϵ_t is 'fundamental' because ϵ_t is given by a linear forecasting rule where we observe the data.

17. 'Spectral Representation Theorem'/'Cramér Representation Theorem' (a frequency domain decomposition, e.g. Brockwell & Davis (1991)): suppose Y_t is a covariance stationary zero mean process. Then there exists an orthogonal-increment process $Z(\omega)$ such that:

    a) $\text{Var}(Z(\omega)) = F(\omega)$
    b) $X_t = \int_{-\pi}^{\pi} e^{it\omega} dZ(\omega)$

    where F is the 'spectral distribution function of the process'. The 'spectral density' $S(\omega)$ is the density associated with F. This is an extremely useful, important decomposition that merits further discussion. The Wold decomposition theorem started by taking the Y's and breaking them into pieces ϵ's that are uncorrelated with each other and they are homoscedastic – each has variance σ^2_ϵ – and they have time associated with them, so ϵ_{t-2} is the forecast error you made at $t-2$. This theorem is another decomposition theorem where we start by taking X's and breaking them into pieces Z's, where each piece corresponds to different frequency components (high, business cycle, low, etc.) that are uncorrelated with each other and strictly periodic and each have their own different variance – one component corresponding to the business cycle might have a big variance so realisations from that process will have a big business cycle Z and another process will have a big low frequency component so a realisation generated from these will look very trendy.

At this juncture, we will raise some important questions related to the frequency domain literature. We may ask how important are the seasonal or business cycle components in Y_t? Seasonal components tend to spike and fall in regular patters; business cycle components tend to oscillate about a trend. Can we measure the variability at a particular frequency? Frequency zero, i.e. the long run, will be particularly important since this is the essence of heterocedasticity- and autocorrelation- adjusted (HAC) covariance matrices. Can we isolate or eliminate the 'seasonal' component or 'business cycle' component? Frequency domain research can highlight the relative importance of components. Filtering concentrates on what we can do to extract a certain component. What about in real time? Can we estimate the business cycle or

'gap' component in real time? If we can do this, then perhaps we want to know how accurate our estimate is. With low frequencies, we can use HAC, robust SE. Note that in estimating spectra, we estimate the variance component of the low frequency parts, which correspond to the variance of sample averages.

7.2.3 Spectral representations of stationary processes

In this subsection we consider four models for Y_t, *viz.* a deterministic process, a stochastic process, a stochastic process with more components and a stochastic process with even more components. We will soon let Y_t be a covariance stationary stochastic process.

Firstly, let Y_t be a deterministic process such as

(a) $Y_t = \cos(\omega t)$

(b) $Y_t = a \times \cos(\omega t) + b \times \sin(\omega t)$

In the first case, Y_t is strictly periodic with a period of $\frac{2\pi}{\omega}$, an amplitude of 1 and a starting value of $Y_0 = 1$. In the second case, Y_t is strictly periodic with the same period $\frac{2\pi}{\omega}$, an amplitude of $\sqrt{a^2 + b^2}$ and a starting value of $Y_0 = a$. We can change ω, we can shift the origin and we can change the amplitude. If ω is big (high frequency), then the graph repeats often; else, the graph repeats less often.

Now let us consider the stochastic process:

$$Y_t = a \times \cos(\omega t) + b \times \sin(\omega t)$$

where a and b are random variables with zero mean and that are mutually uncorrelated with the same variance ω^2. This is all we need to know for spectral processes. So our (first and second) moments can be given by:

$$E(Y_t) = 0$$
$$Var(Y_t) = \sigma^2 \times \{\cos^2(\omega t) + \sin^2(\omega t)\} = \sigma^2$$
$$Cov(Y_t, Y_{t-k}) = \sigma^2\{\cos(\omega t)\cos(\omega(t-k)) + \sin(\omega t)\sin(\omega(t-k))\} = \sigma^2\cos(\omega k)$$

So far we have been focusing on one component, i.e. one ω, say seasonal. What about low frequencies, high frequencies, business cycle frequencies, seasonal frequencies, etc.? We can write a stochastic process that follows more components as:

$$Y_t = \sum_{j=1}^{n} \{a_j \cos(\omega_j t) + b_j \sin(\omega_j t)\}$$

where $\{a_j, b_j\}$ are zero-mean, uncorrelated random variables having the property that $Var(a_j) = Var(b_j) = \sigma_j^2$, i.e. heteroscedastic variance that is frequency related. We can let seasonals be more important than business cycle frequencies by allowing them to have bigger variances σ_j^2 and so a_s, b_s are then

more important than a_{bc}, b_{bc}. So our (first and second) moments can be given by:

$$E(Y_t) = 0$$

$$Var(Y_t) = \sum_{j=1}^{n} \sigma_j^2 \quad \text{decomposition of variance}$$

$$Cov(Y_t, Y_{t-k}) = \sum_{j=1}^{n} \sigma_j^2 \cos(\omega_j k) \quad \text{decomposition of auto-covariances}$$

We can go even further by examining stochastic processes with even more components. Including all frequencies between zero and π:

$$Y_t = \int_0^{\pi} \cos(\omega t) da(\omega) + \int_0^{\pi} \sin(\omega t) db(\omega)$$

where $da(\omega)$ and $db(\omega)$ are zero mean random variables that are mutually uncorrelated, uncorrelated across frequency and with common variance that is a function of frequency. Remember that time data is still discrete. The variance function (i.e. how variance depends on frequency) is called the *spectrum*:

$$\sigma_j^2 = \sigma^2(\omega) = S(\omega)$$

Let us change notation for convenience:

$$\begin{aligned} Y_t &= a \times \cos(\omega t) + b \times \sin(\omega t) \\ &= \frac{1}{2} e^{i\omega t}(a - ib) + \frac{1}{2} e^{-i\omega t}(a + ib) \\ &= e^{i\omega t} g + e^{-i\omega t} \bar{g} \end{aligned}$$

where $i = \sqrt{-1}$, $e^{i\omega t} = \cos(\omega t) + i \sin(\omega t)$ from Euler, $g = \frac{1}{2}(a - ib)$ and $\bar{g}$ is the complex conjugate of g, i.e. $\bar{g} = \frac{1}{2}(a + ib)$. Likewise

$$\begin{aligned} Y_t &= \int_0^{\pi} \cos(\omega t) da(\omega) + \int_0^{\pi} \sin(\omega t) db(\omega) \\ &= \frac{1}{2} \int_0^{\pi} e^{i\omega t}(da(\omega) - i db(\omega)) + \frac{1}{2} \int_0^{\pi} e^{-i\omega t}(da(\omega) + i db(\omega)) \\ &= \int_{-\pi}^{\pi} e^{i\omega t} dZ(\omega) \end{aligned}$$

where

$$dZ(\omega) = \begin{cases} \frac{1}{2}(da(\omega) - i db(\omega)) & \omega \geq 0 \\ \bar{dZ}(-\omega) & \omega < 0 \end{cases}$$

Note that the mean of dZ is zero since da and db have zero mean. Let $Var(dZ(\omega)) = E(dZ(\omega)\overline{dZ(\omega)}) = S(\omega)$ and observe that $E(dZ(\omega)\overline{dZ(\omega')}) = 0$ for $\omega \neq \omega'$ because we assume da and db are uncorrelated across frequency. So

$E(dZ(\omega)\overline{dZ(\omega')}) = 0$ implies $\omega \neq \omega'$. Now (first and second) moments of Y_t can be given by

$$E(Y_t) = E\left\{\int_{-\pi}^{\pi} e^{i\omega t} dZ(\omega)\right\} = \int_{-\pi}^{\pi} e^{i\omega t} E(dZ(\omega)) = 0$$

$$\begin{aligned}\gamma_k = E(Y_t Y_{t-k}) = E(Y_t \overline{Y_{t-k}}) &= E\left\{\int_{-\pi}^{\pi} e^{i\omega t} dZ(\omega) \int_{-\pi}^{\pi} e^{-i\omega(t-k)} \overline{dZ(\omega)}\right\} \\ &= \int_{-\pi}^{\pi} e^{i\omega t} e^{-i\omega(t-k)} E(dZ(\omega)\overline{dZ(\omega)}) \\ &= \int_{-\pi}^{\pi} e^{i\omega k} S(\omega) d\omega\end{aligned}$$

Observe that when $k = 0$, $\gamma_0 = Var(Y_t) = \int_{-\pi}^{\pi} S(\omega) d\omega$.

To summarise:

1. $S(\omega)d\omega$ may be interpreted as the variance of the cyclical component of Y corresponding to frequency ω. As usual, the period of this component is given by $\frac{2\pi}{\omega}$.
2. $S(\omega) \geq 0$ since $S(\omega)$ is a variance.
3. $S(\omega) = S(-\omega)$. Therefore, plots of the spectrum are typically presented across the range $0 \leq \omega \leq \pi$ due to this symmetry.
4. We may invert $\gamma_k = \int_{-\pi}^{\pi} e^{i\omega k} S(\omega) d\omega$ to yield

$$S(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} e^{-i\omega k} \gamma_k = \frac{1}{2\pi}\left\{\gamma_0 + 2\sum_{k=1}^{\infty} \gamma_k \cos(\omega k)\right\}$$

   So we see that the spectrum is the sum of the autocovariances, which are easy to compute for WN (zero) – a flat spectrum, i.e. equal to zero.

The long-run variance is $S(0)$, which is the variance of the zero-frequency or ∞ period component.

$$S(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} e^{-i\omega k} \gamma_k \Longrightarrow S(0) = \frac{1}{2\pi} \sum_{k=-\infty}^{\omega} \gamma_k$$

Let us examine the important role this plays in statistical inference. Suppose Y_t is a covariance stationary stochastic process with mean μ. Then we have that

$$\begin{aligned}Var(\sqrt{T}(\bar{Y} - \mu)) &= Var\left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \omega_t\right) \\ &= \frac{1}{T}\{T\gamma_0 + (T-1)(\gamma_1 + \gamma_{-1}) + (T-2)(\gamma_2 + \gamma_{-2}) + \cdots 1(\gamma_{T-1} + \gamma_{1-T})\} \\ &= \sum_{j=-T+1}^{T-1} \gamma_j - \frac{1}{T} \sum_{j=1}^{T-1} j(\gamma_j + \gamma_{-j})\end{aligned}$$

The last term on the third line goes to zero as $j \longrightarrow \infty$. If autocovariances are '1-summable' so that $\sum j|\gamma_j| < \infty$, then we have that

$$Var\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\omega_t\right) \longrightarrow \sum_{j=-\infty}^{\infty}\gamma_j = 2\pi S(0)$$

Sometimes by 'long-run variance' we mean $2\pi S(0)$ and sometimes we mean $S(0)$.

7.2.4 Spectral properties of filters

This subsection is broken into two parts: (i) band pass filters and (ii) one-sided filters. To motivate these, consider estimating the output gap in real time (e.g. finding the output gap in the last quarter). To get the output gap, we filter the data to find its trend and then subtract this from the series. Two-sided filters require data from the future and the past to figure out what the value of the trend is today. While this is useful if you are conducting historical analysis, but not if you are engaged in real time analysis. We will look at how we can modify these two-sided filters so we can work with real time data issues where we only have data up to the present and we will examine how to quantify our uncertainty about the current output gap, for example.

Filters are really just moving averages defined by $x_t = c(L)y_t$, where we say that x_t is a filter of y_t and $c(L) = c_{-r}L^{-r} + \cdots + c_sL^s$, where L^{-r} is the forward operator, so x is a moving average of y having the c's for weights.[6] So, the filter is composed of forward and backward operators. They are called filters because they filter out what you don't want, similar to turning on and off or up and down the volume of different frequencies on your car's radio, e.g. adjusting the bass and treble. We will be interested in using $c(L)$ to focus on seasonal frequencies and business cycle frequencies – we can filter out the frequency we are looking for. We want to know how $c(L)$ changes the cyclical properties of y. Suppose that y is strictly periodic:

$$y_t = 2\cos(\omega t) = e^{i\omega t} + e^{-i\omega t}$$

with period $p = \frac{2\pi}{\omega}$. We put y through the filter to get x, which will also be strictly periodic because y is strictly periodic. x will be shifted in time because of the backward and forward operators. Furthermore, due to the presence of c's, which generally will not all be ones, x will be dampened or amplified. We want to find out how much x is shifted in time and how much it is attenuated or amplified. Two examples of filters, i.e. what x is could be the first difference of y or the seasonal difference of y. While y is really composed of a bunch of different ω's, it is intuitive to start by looking at one component at of y at a

[6] Older literature used the notation B for backward and F for forward operators. Now, L^+ represents backward operators and L^- represents forward operators.

time. We can represent x as the following:

$$\begin{aligned} x_t &\overset{MA}{=} \sum_{j=-r}^{s} c_j y_{t-j} \\ &= \sum_{j=-r}^{s} c_j [e^{i\omega(t-j)} + e^{-i\omega(t-j)}] \\ &= e^{i\omega t} \sum_{j=-r}^{s} c_j e^{-i\omega j} + e^{-i\omega t} \sum_{j=-r}^{s} c_j e^{i\omega j} \\ &= e^{i\omega t} c(e^{-i\omega}) + e^{-i\omega t} c(e^{i\omega}) \end{aligned}$$

where the exponentials in the last line represent y while the c's are the MA weights. Note that $c(e^{i\omega}) \in \mathbb{C}$, say $c(e^{i\omega}) = a + ib$ where $a = \mathbb{Re}[c(e^{i\omega})]$ and $b = Im[c(e^{i\omega})]$. We can write this number in polar form as:

$$c(e^{i\omega}) = (a^2 + b^2)^{\frac{1}{2}} [\cos(\theta) + i \sin(\theta)] = g e^{i\omega}$$

where we have defined $g = (a^2 + b^2)^{\frac{1}{2}} = [c(e^{i\omega}) c(e^{-i\omega})]^{\frac{1}{2}}$ and $\theta = \tan^{-1}\left(\frac{b}{a}\right) = \tan^{-1}\left(\frac{Im[c(e^{i\omega})]}{\mathbb{Re}[c(e^{i\omega})]}\right)$; for example, $c(L) = c_1 L + c_2 L^2$ would mean that the sum would be be $c(e^{-i\omega}) = c_1 e^{-i\omega} + c_2 e^{-i\omega 2}$. So g is the distance from the origin to the point (a, b) in the $\mathbb{Re} - Im$ space, where θ makes the counter-clockwise angle from the $\mathbb{Re}$ axis. So

$$\begin{aligned} x_t &= e^{i\omega t} g e^{-i\theta} + e^{-i\omega t} g e^{i\theta} \\ &= g[e^{i\omega[t - \frac{\theta}{\omega}]} + e^{-i\omega[t - \frac{\theta}{\omega}]}] \\ &= 2g \cos\left(\omega\left(t - \frac{\theta}{\omega}\right)\right) \end{aligned}$$

So we can see here that the filter $c(L)$ has two effects: (i) it amplifies y by the factor g and (ii) it shifts y back in time by $\frac{\theta}{\omega}$ units of time. We will write $g(\omega)$ and $\theta(\omega)$ to emphasise their dependence on ω. $g(\omega)$ is the *filter gain* or *amplitude gain* and $\theta(\omega)$ is the *filter phase* or *phase shift*. $g(\omega)^2 = c(e^{i\omega}) c(e^{-i\omega})$ is called the *power transfer function* of the filter. Note that we have done this for one component ω, e.g the business cycle frequency; if we do it for another cyclical component, e.g. the seasonal, we will have different g and θ as they depend on ω, which will be different. We will now look at g and θ for different ω. We will look at different filters.

Example 7.45. Consider

$$c(L) = L^2$$

This filter shifts y back in time by two periods. So

$$c(e^{i\omega}) = e^{2i\omega} = \cos(2\omega) + i \sin(2\omega)$$

so that we have that

$$\theta(\omega) = \tan^{-1}\left[\frac{\sin(2\omega)}{\cos(2\omega)}\right] = 2\omega$$

so we have that the filter shifts y back in time by $\frac{\theta}{\omega} = \frac{2\omega}{\omega} = 2$ time periods. Also note that

$$\begin{aligned} g(\omega) &= |c(e^{i\omega})| \\ &= |\cos(2\omega) + \sin(2\omega)| \\ &= \sqrt{\cos^2(2\omega) + \sin^2(2\omega)} \\ &= 1 \end{aligned}$$

So the gain is one.

Example 7.46 (Kuznets Filter)**.** In Sargent's textbook (1979) [74], he refers to the *Kuznets filter* for annual data, which (i) eliminates trends through centered ten-year differences (see $b(L)$ below) thereby yielding a volatile series and (ii) smooths out the rest (MA). To smooth out the series, define

$$a(L) = \frac{1}{5}(L^{-2} + L^{-1} + L^0 + L^1 + L^2)$$

and to get rid of trends, define

$$b(L) = (L^{-5} - L^5)$$

which is like $y_{t+5} - y_{t-5}$. The *Kuznets filter* is

$$c(L) = b(L)a(L)$$

We can compute the gain:

$$g(\omega) = |c(e^{i\omega})| = |b(e^{i\omega})||a(e^{i\omega})|$$

easily for a grid of values through MATLAB. This filter eliminates low frequencies (gets rid of the trend), attenuates high frequencies and amplifies particular frequencies of about 0.3, which correspond to frequencies of about 20 years. So, the Kuznets filter effectively 'turns up the volume' on 20 year cycles (approximately the period) which are called *Kuznets cycles*. Note that we get twenty year cycles even if they are not in the data since we are only turning up the volume for these frequencies. For instance, we will get these cycles even if we look at WN. One interpretation is that we have done something that was seemingly sensible but arguably ultimately wrong. Let us next look at some 'better' filters.

Example 7.47 (X-11 Seasonal Adjustment)**.** At the US Bureau of Labour Statistics in the early 1960s, the statistician Julius Shiskin developed the first

computer program, the X-1 to automate the process of seasonally adjusting data. Today, X-11 ARIMA is the standard program, with a few improvements in X-12 ARIMA and X-13 ARIMA-SEATS versions. X-12 and X-13 can be downloaded and used for free from http://www.census.gov/.[7] Let us focus on the X-11 since it is simpler to describe and X-12 and X-13 are based on it. The linear operations in X-11 can be described by

$$x_t^{sa} = X11(L)x_t$$

where X11(L) is a two-sided filter that is constructed in multiple steps involving linear filters:

1. Initial estimate of TC: $\hat{TC}_t^1 = A_1(L)x_t$ where $A_t(L)$ is the centered 12-month MA filter $A_t(L) = \sum_{j=-6}^{6} b_j L^j$ with $b_{|6|} = \frac{1}{24}$ and $b_j = \frac{1}{12}$ for $j \in \{-5, \ldots, 5\}$.

2. Initial estimate of $S + I$: $\hat{SI}_t^1 = x_t - \hat{TC}_t^1$.

3. Initial estimate of S_t: $\hat{S}_t^1 = S_1(L^{12})\hat{SI}_t^1$ where $S_1(L^{12}) = \sum_{j=-2}^{2} c_j L^{12j}$ with c_j as weights from a 3×3 MA, i.e. $\frac{1}{9}, \frac{2}{9}, \frac{3}{9}, \frac{2}{9}, \frac{1}{9}$.

4. Adjust estimates of S so they approximately sum to zero over any 12 month period: $\hat{S}_t^2 = S_2(L)\hat{S}_t^1$ where $S_2(L) = 1 - A_1(L)$ with the definition of $A_1(L)$ given in step 1.

5. Second estimate of TC: $\hat{TC}_t^2 = A_2(L)(x_t^1 - \hat{S}_t^2)$ where $A_2(L)$ is a 'Henderson' MA filter. 13-term Hendersen MA filter is $A_2(L) = \sum_{i=-6}^{6} A_{2,i} L^i$ where $A_{2,0} = .2402$, $A_{2,|1|} = .2143$, $A_{2,|2|} = .1474$, $A_{2,|3|} = .0655$, $A_{2,|4|} = 0$, $A_{2,|5|} = -.0279$, $A_{2,|6|} = -.0194$.

6. Third estimate of S: $\hat{S}_t^3 = S_3(L^{12})(x_t - \hat{TC}_t^2)$ where $S_3(L^{12}) = \sum_{j=-3}^{3} d_j L^{12j}$ with D_j as weights from a 3×5 MA, i.e. $\frac{1}{15}, \frac{2}{15}, \frac{3}{15}, \frac{3}{15}, \frac{2}{15}, \frac{1}{15}$.

7. Adjust estimates of S so they approximately sum to zero over any 12 month period: $\hat{S}_t^4 = S_2(L)\hat{S}_t^3$ where $S_2(L)$ has been defined in step 4.

8. Final seasonally adjusted value: $x_t^{sa} = x_t - \hat{S}_t^4$.

Looking at the gain, this filter effectively turns down the volume on seasonals. As we said for the Kuznets filter, one interpretation is that we can get things wrong. So, we would like to be smart about what we are doing and get things right, but how?

Example 7.48 (HP Filter)**.** The HP filter eliminates trends.

[7] Typically one uses the X-12 ARIMA program, which is heavily based on X-11. However, the transition to X-13 ARIMA-SEATS is well underway with most groups inside the US Census Bureau using the X-13 ARIMA-SEATS to produce seasonal adjustments.

Let $x_t = c(L)y_t$ where y has spectrum $S_y(\omega)$. We want to find the spectrum of X. Since the frequency components of x are simply scaled frequency components of y, where the scaling is by the factor $g(\omega)e^{i\theta(\omega)}$, we have the following relation between the spectra of x and y:

$$S_x(\omega) = g(\omega)^2 S_y(\omega) = c(e^{i\omega})c(e^{-i\omega})S_y(\omega)$$

where if $g < 1$, we have attenuated y and if $g > 1$, we have amplified it. Since the spectrum is like the variance, as we are simply scaling the variance of y, we can easily find the spectrum of x. This is an important formula and it is useful because it allows us to compute the spectra of ARMA models, say, with a very simple formula. Furthermore, with heteroscedastic consistent standard errors that are estimates of the spectra, we can use this formula to figure out the spectra.

Example 7.49. The spectrum of the WN process ϵ_t is:

$$S_\epsilon(\omega) = \frac{1}{2\pi}\left\{\gamma_0 + 2\sum_{k=1}^{\infty}\gamma_k \cos(\omega k)\right\} = \frac{\sigma_\epsilon^2}{2\pi}$$

Recall that when $y_t \sim$ ARMA, $\phi(L)y_t = \theta(L)\epsilon_t$ or $y_t = c(L)\epsilon_t$ with $c(L) = \frac{\theta(L)}{\phi(L)}$. Then the spectrum of y is

$$\begin{aligned} S_y(\omega) &= c(e^{i\omega})c(e^{-i\omega})S_\epsilon(\omega) \\ &= \sigma_\epsilon^2 \frac{\theta(e^{i\omega})\theta(e^{-i\omega})}{\phi(e^{i\omega})\phi(e^{-i\omega})}\frac{1}{2\pi} \\ &= \sigma_\epsilon^2 \frac{(1-\theta_1 e^{i\omega} - \cdots - \theta_q e^{iq\omega})(1-\theta_1 e^{-i\omega} - \cdots - \theta_q e^{-iq\omega})}{(1-\phi_1 e^{i\omega} - \cdots - \phi_p e^{ip\omega})(1-\phi_1 e^{-i\omega} - \cdots - \phi_p e^{-ip\omega})}\frac{1}{2\pi} \end{aligned}$$

When $\omega = 0$, i.e. at zero frequency:

$$S_y(0) = \sigma_\epsilon^2 \frac{(1-\theta_1 - \cdots - \theta_q)(1-\theta_1 - \cdots - \theta_q)}{(1-\phi_1 - \cdots - \phi_p)(1-\phi_1 - \cdots - \phi_p)}\frac{1}{2\pi}$$

which is a simple formula for figuring out the long-run variance of this process.

Now let us suppose we want to construct a filter that keeps certain frequencies, say the trend. This is a classical problem in signal processing, i.e. constructing a *band-pass* filter. Let $c(L) = \sum_{j=-\infty}^{\infty} c_j L^j$ and allow the phase of $c(L)$ to be zero so there is no shift in time, i.e. $c(L)$ is symmetric: $c_j = c_{-j}$ (treat future and past the same way) and we want to make

$$gain(c(L)) = |c(e^{i\omega})| \overset{symm}{=} c(e^{i\omega}) = \begin{cases} 1 & -\underline{\omega} \leq \omega \leq \underline{\omega} \\ 0 & \omega \notin [-\underline{\omega}, \underline{\omega}] \end{cases}$$

The gain function will be one from the origin and drop to zero from frequency ω onwards. So, we only keep frequencies between zero and $\underline{\omega}$ and turn down

the volume or eliminate the rest. Thus, we have a band-pass filter: we want to pass a band of certain frequencies. We have constructed a filter $c(L)$ that gives us this. Note that since $c(e^{-i\omega}) = \sum_{j=-\infty}^{\infty} c_j e^{-i\omega j}$, we get the identity $c_j = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\omega j} c(e^{-i\omega}) d\omega$. If we set the gain to one over the desired frequencies, the integration produces the following weights:

$$c_j = \frac{1}{2\pi}\frac{1}{ij} e^{i\omega j}|_{-\underline{\omega}}^{\underline{\omega}} = \begin{cases} \frac{1}{j\pi}\sin(\underline{\omega} j) & j \neq 0 \\ \frac{\underline{\omega}}{\pi} & j = 0 \end{cases}$$

Observe that the values of c_j, which due to symmetry are both the weight you put on the the lag of the series j periods ago and the lead of the series j periods ahead will die out at the rate j^{-1}, so they decline slowly. Furthermore, if you want everything except say the low frequency (if that was what you got in the first place with this filter), note that $1 - c(L)$ passes everything except for $-\underline{\omega} \leq \omega \leq \underline{\omega}$, so $1 - c(L)$ will be a high pass filter and note that difference of low-pass filters can be used to pass any set of frequencies. We take the original series and subtract what we got from our filter. We can get say business cycle frequencies by subtracting low pass filters and high pass filters where we define the filters cleverly. Essentially, we are subtracting rectangles to get whatever rectangles we want. Finally, Baxter and King (1999) [6] show that $c_k(L) = \sum_{j=-k}^{k} c_j L^j$ is an optimal finite order approximation to $c(L)$ in that the gain of $c_k(L)$ is as close ($L^2 - norm$) as possible to the gain of $c(L)$ for a k-order filter; this means that since generally there will be an infinite number of leads and lags, we can approximate the gain function (best in terms of L^2 approximation) where we truncate over a certain number of periods.

What if we want to pass periods less than eight years? We can look at the *gap filter*. Periods greater than eight years may be thought of as trends, while periods less than eight years might be thought of as deviations from trends, say output gaps. This tends not to die out very quickly in that while most weight is put around contemporaneous observations, there is still a significant amount of weight far away and we cannot simply truncate this.

An interesting exercise is to apply band-pass filters to the log of real GDP, say in the US. Take a plot of the series, its low frequency components, i.e. the trend (say periods above 32 quarters), business cycle frequencies (periods between 6 and 32 quarters) and high frequency components (periods below 6 quarters). Each component will look like its frequency: the trend looks like a trend, the business cycle looks like a business cycle and the high frequency component certainly will display high frequency. An incorrect interpretation would be to say that there seems to be repetition every few quarters, say for the business cycle frequencies because we have reached this result by construction. What is more interesting is the clearly observed volatility decline, especially since 1980s.

It is infeasible to use two-sided filters for the present time. For instance, in 2013 if we were asked what would be the output gap for the next 20 years, if we had access to data on GDP (i.e. perfectly foreseeable), then we could plug in the data for GDP and use a two-sided filter. What we could do instead is

to try and use the information we have up to today. The issue therefore with two-sided filters with weights that die out slowly is the essentially 'endpoint' problems. Geweke (1978) [36] showed how to implement a minimum MSE one-sided filter (estimator) for using data that is available in real time. Let

$$x_t = c(L)y_t = \sum_{i=-\infty}^{\infty} c_i y_{t-i}$$

We will use information up to the latest available time period, T. The optimal estimate of x_t given $\{y_j\}_{j=1}^T$ is

$$\begin{aligned} E(x_t|\{y_j\}_{j=1}^T) &= \sum_{i=-\infty}^{\infty} c_i E(y_{t-i}|\{y_j\}_{j=1}^T) \\ &= \sum_{i=-\infty}^{0} c_{t-i}\hat{y}_i + \sum_{i=1}^{T} c_{t-i}y_i + \sum_{i=T+1}^{\infty} c_{t-i}\hat{y}_i \end{aligned}$$

where $\hat{y}_i$ are unknown – we forecast them and backcasts of y_i are also constructed from the data $\{y_j\}_{j=1}^T$. So, we have applied an optimal two-sided filter to series for which we use actual data for the pre-sample period (best backcast is the data itself) and use forecasts for the post-sample period. Geweke talked about this in the context of seasonal adjustment and it is what X-12 ARIMA does, which is simply X-11 ARIMA applied to series with forecasts and backcasts appended with ARIMA models. More generally, we can apply this technique to any models, say AR, VAR, DSGE, etc. We can compute the errors, i.e. how uncertain we are about the output gap today, since we have the forecasts from some model, say AR / ARMA / VAR / DSGE models and so we know the variance and covariance properties. Findley *et al* (1991) describes how this is done in X-12. The variance of the error from using $\{y_j\}_{j=1}^T$ will be

$$var(x_t - E(X_t|\{y_j\}_{j=1}^T)) = var\left(\sum_{i=-\infty}^{\infty} c_i\{E(y_{t-i}|\{y_j\}_{j=1}^T) - y_{t-i}\}\right)$$

While Geweke was concerned with the X11 filter, his result is general and applies to any linear filter, e.g. band-pass, HP, etc. You will find that you get higher standard errors around the backcasts before you have data and forecasts after you have data as you would expect.

Let us now look briefly at regressions using filtered data:

$$y_t = x_t'\beta + u_t \quad E(u_t x_t) = 0$$

which allows our OLS estimates to be consistent. Now let us use some filter (e.g. first difference / HP / seasonal adjustment). Denote filtered data by:

$$\begin{aligned} y_t^{\text{filtered}} &= c(L)y_t \\ x_t^{\text{filtered}} &= c(L)x_t \end{aligned}$$

where we use the same filter on both. Now write

$$y_t^{\text{filtered}} = x_t^{\text{filtered}\prime}\beta + u_t^{\text{filtered}}$$

To use OLS, we need to know if $E(x_t^{\text{filtered}} u_t^{\text{filtered}}) = 0$. This holds when x is strictly exogenous. Note that x^{filtered} and u^{filtered} include lagged, current and future values. For $E(x_t^{\text{filtered}} u_t^{\text{filtered}}) = 0$, x and u need to be uncorrelated at all leads and lags, but this never happens unless say $x = 1$ or something trivial. The reasoning is analogous to the argument against using GLS in time series regression There, we may attempt to correct for serial correlation with an AR(1) error. But that is equivalent to applying a filter to both sides, which in turn changes the orthogonality conditions. So, correcting for AR(1) errors – or other time series errors – may lead to other problems; hence, you should use heteroscedastic consistent standard errors instead. The takeaway from this is that when using filtered data or conducting GLS in time series regressions, you must think rather hard about how and whether to do so.

7.2.5 Multivariate spectra

We will not cover this section in class. This subsection simply extends what we have done so far to vectors. When y_t is a scalar, the spectrum associated with frequency ω of the process y_t is $S(\omega) = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty} \lambda_j e^{-i\omega j}$. This represents the variance of the complex valued Cramér increment $dZ(\omega)$ in $y_t = \int_{-\pi}^{\pi} e^{i\omega t} dZ(\omega)$. We can generalise this.

Generalising, let Y_t be an $n \times 1$ vector with j^{th} autocovariance matrix $\Gamma_j = V(Y_t Y'_{t-j})$ and let $S(\omega) = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty} \Gamma_j e^{-i\omega j}$ so $S(\omega)$ will be an $n \times n$ matrix. The spectral representation is the same as above except that $dZ(\omega)$ is an $n \times 1$ complex-valued random vector with associated spectral density matrix $S(\omega)$. Note that $S(\omega)$ may be interpreted as a covariance matrix for the increments $dZ(\omega)$. The diagonal elements of $S(\omega)$ are the univariate spectra of the series, while the off-diagonal elements are the 'cross-spectra' (covariance between two increments). Cross-spectra are complex valued and such that $S_{ij}(\omega) = \overline{S_{ij}(\omega)}$. $Re(S_{ij}(\omega))$ is called the *co-spectrum* and $Im(S_{ij}(\omega))$ is called the *quadrature* spectrum. Also:

$$\begin{aligned}
Coherence(\omega) &= \frac{S_{ij}(\omega)}{\sqrt{S_{ii}(\omega) S_{jj}(\omega)}} \\
Gain_{ij}(\omega) &= \frac{|S_{ij}(\omega)|}{S_{jj}(\omega)} \\
Phase_{ij}(\omega) &= tan^{-1}\left(\frac{-Im(S_{ij}(\omega))}{Re(S_{ij}(\omega))}\right)
\end{aligned}$$

In order to interpret these definitions, consider scalars Y and X with spectra S_Y and S_X and cross-spectra S_{XY}. Furthermore, consider the regression of Y_t

onto leads and lags of X_t:

$$Y_t = \sum_{j=-\infty}^{\infty} c_j X_{t-j} + u_t = c(L)X_t + u_t$$

Observe that since u and X are uncorrelated at all leads and lags

$$S_Y(\omega) = |c(e^{i\omega})|^2 S_Y(\omega) + S_u(\omega)$$

In addition, we can see

$$E(Y_t X_{t-k}) = \sum_{j=-\infty}^{\infty} c_j E(X_{t-j} X_{t-k}) = \sum_{j=-\infty}^{\infty} c_j E(X_t X_{t-k+j}) = \sum_{j=-\infty}^{\infty} c_j \gamma_{k-j}$$

where γ represents the autocovariance of X.

$$\begin{aligned}
\therefore S_{YX}(\omega) &= \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} e^{-i\omega k} \sum_{j=-\infty}^{\infty} c_j \gamma_{k-j} \\
&= \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} c_j e^{-i\omega j} \sum_{l=-\infty}^{\infty} e^{-i\omega l} \gamma_l \\
&= c(e^{-i\omega}) S_X(\omega)
\end{aligned}$$

Therefore, the gain and phase of the cross-spectrum is the gain and phase of $c(L)$.

7.2.6 Spectral Estimation

In this final, brief subsection, we will say a few words about AR/VAR/ARMA parametric estimation. Let $\Phi(L)Y_t = \Theta(L)\epsilon_t$ where Y can be a vector and ϵ_t is WN with variance-covariance matrix Σ_ϵ. The spectral density matrix of Y is given by:

$$S_Y(\omega) = \Phi(e^{i\omega})^{-1} \Theta(e^{i\omega}) \Sigma_\epsilon \Theta(e^{-i\omega})' \Phi(e^{-i\omega})'^{-1}$$

Parametric estimators use estimates of the AR and MA parameters and Σ_ϵ. For example, consider the VAR(1) model given by $(I - \Phi L)Y_t = \epsilon_t$. The estimate of the spectral density matrix of Y is given by:

$$\hat{S}_Y(\omega) = (I - \hat{\Phi} e^{i\omega})^{-1} \hat{\Sigma}_\epsilon (I - \hat{\Phi} e^{-i\omega})'^{-1}$$

7.2.7 Further frequency related filtering

Let us continue with frequency related filtering. Denoting the growth rate of the time series y_t by x_t, if the frequency of the data on y_t is monthly, then x_t is the monthly growth rate:

$$x_t = 100 \cdot [\log(y_t) - \log(y_{t-1})]$$

So, other than the scale parameter (100), the monthly growth rate x_t is obtained from $\log(y_t)$ by applying the following filter, where L is the lag operator:

$$x_t = (1 - L)\log(y_t) \tag{7.29}$$

Let Y_t be a covariance-stationary time series with absolutely summable autocovariances where $g_Y(z)$ is the ACGF of Y and $s_Y(\omega)$ is the spectral density function of Y and recall that the spectrum can be expressed as

$$s_Y(\omega) = \frac{1}{2\pi} g_Y(e^{-i\omega})$$

Now transform Y by $X_t = h(L)Y_t$ where $h(L) = \sum_{j=-\infty}^{\infty} h_j L^j$ and $\sum_{j=-\infty}^{\infty} |h_j| < \infty$. Recall also that the ACGF of Y can yield the ACGF of X through:

$$g_X(z) = h(z)h(z^{-1})g_Y(z)$$

$$\therefore s_X(\omega) = \frac{1}{2\pi} g_X(e^{-i\omega}) = \frac{1}{2\pi} h(e^{-i\omega})h(e^{i\omega})g_Y(e^{-i\omega})$$

which is the population spectrum of X and substituting $s_Y(\omega)$ into this shows the following relationship between the population spectra of X and Y:

$$s_X(\omega) = h(e^{-i\omega})h(e^{i\omega})s_Y(\omega)$$

The filter $h(L)$, which operates on the series Y_t effectively multiplies the spectrum by the function $h(e^{-i\omega})h(e^{i\omega})$. The difference operator in (7.29) has a filter of $h(L) = 1 - L$, so the function $h(e^{-i\omega})h(e^{i\omega})$ would be given by:

$$\begin{aligned} h(e^{-i\omega})h(e^{i\omega}) &= (1 - e^{-i\omega})(1 - e^{i\omega}) \\ &= 1 - e^{-i\omega} - e^{i\omega} + 1 \\ &= 2 - 2 \cdot \cos(\omega) \end{aligned}$$

where the last line follows from observing that

$$e^{-i\omega} + e^{i\omega} = \cos(\omega) - i\sin(\omega) + \cos(\omega) + i\sin(\omega) = 2\cos(\omega)$$

So, if $X_t = (1 - L)Y_t$, then to find the population spectrum of X at any frequency ω, we find the value of $s_Y(\omega)$ and then multiply it by $2 - 2\cos(\omega)$; e.g. the spectrum at $\omega = 0$ is multiplied by zero, that at $\omega = \pi/2$ is multiplied by 2 and that at $\omega = 4$ is multiplied by 4. In effect, differencing the data removes low frequency components and accentuates high frequency components. If on the other hand Y_t is non stationary, then the differenced data $(1 - L)Y_t$ will not generally have a zero population spectrum at $\omega = 0$.

Looking at the year to year growth rates or the percentage change in y_t between t and its value for the same month of the previous year:

$$w_t = 100[\log(y_t) - \log(y_{t-12})]$$

the seasonal difference filter would be $h(L) = 1 - L^{12}$, so

$$\begin{aligned} h(e^{-i\omega})h(e^{i\omega}) &= (1 - e^{-12i\omega})(1 - e^{12i\omega}) \\ &= 1 - e^{-12i\omega} - e^{12i\omega} + 1 \\ &= 2 - 2\cos(12\omega) \end{aligned}$$

The function is zero when $12\omega = 0, 2\pi, 4\pi, 6\pi, 8\pi, 10\pi, 12\pi$, i.e. it is zero at frequencies $\omega = 0, 2\pi/12, 4\pi/12, 6\pi/12, 8\pi/12, 10\pi/12, \pi$. Therefore, seasonally differencing eliminates the low frequency ($\omega = 0$) components of a stationary process as well as any contribution from cycles that have periods of $12, 6, 4, 3, 2.4$ or 2 months.

A slowly evolving trend can be thought of as a low frequency cycle. A constant trend can be thought of as a cycle with zero frequency. Filters are tools that can eliminate the influence of cyclical variations at different frequencies. Earlier when I mentioned trend removal, I said that there were three basic techniques: detrending, differencing and filtering. To target low frequencies, we can use detrending filters such as the first-difference and HP filter; to target seasonal frequencies, we can use seasonal filters, etc.

Definition 7.50. A *linear filter* applied to y_t that produces y_t^f is given by

$$y_t^f = \sum_{j=-r}^{s} c_j y_{t-j} \equiv C(L)y_t$$

i.e. the filter series y_t^f is a linear combination of the original, unfiltered series y_t. This is the time domain approach. In the frequency domain, we replace L^j in $C(L)$ with with $e^{-i\omega j}$ and get the *frequency response function* $C(e^{-i\omega})$.

Let us derive $s_{y_t^f}$ to see how we isolate cycles through filters. Assume $\{y_t\}$ is a mean-zero process with autocovariance sequence $\{\gamma(\tau)\}_{t=-\infty}^{\infty}$. Write the autocovariance function between y_t^f and $y_{t-\tau}^f$ as

$$\begin{aligned} E(y_t^f y_{t-\tau}^f) &= E\left(\sum_{j=-r}^{s} c_j y_{t-j}\right)\left(\sum_{k=-r}^{s} c_k y_{t-k-\tau}\right) \\ &= E\sum_{j=-r}^{s}\sum_{k=-r}^{s} c_j c_k y_{t-j} y_{t-k-\tau} \\ &= \sum_{j=-r}^{s}\sum_{k=-r}^{s} c_j c_k \gamma(\tau + k - j) \\ &\equiv \gamma_{y^f}(\tau) \end{aligned}$$

Now apply the Fourier transform of $\gamma_{y^f}(\tau)$ to get the spectral density function of y_t^f:

$$\begin{aligned} s_{y^f}(\omega) &= \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty}\gamma_{y^f}(\tau)e^{-i\omega\tau} \\ &= \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty}\sum_{j=-r}^{s}\sum_{k=-r}^{s}c_j c_k\gamma(\tau+k-j)e^{-i\omega\tau} \end{aligned} \tag{7.30}$$

Let $h=\tau+k-j$ and rewrite $e^{-i\omega\tau}$ in (7.30) as

$$\begin{aligned} e^{-i\omega\tau} &= e^{-i\omega(h+j-k)} \\ &= e^{-i\omega h}e^{-i\omega j}e^{i\omega k} \end{aligned} \tag{7.31}$$

Lastly, using (7.31), substitute into (7.30) for $e^{-i\omega\tau}$ to get

$$\begin{aligned} s_{y^f}(\omega) &= \frac{1}{2\pi}\sum_{j=-r}^{s}c_j e^{-i\omega j}\sum_{k=-r}^{s}c_k e^{i\omega k}\sum_{h=-\infty}^{\infty}\gamma(h)e^{-i\omega h} \\ &= \sum_{j=-r}^{s}c_j e^{-i\omega j}\sum_{k=-r}^{s}c_k e^{i\omega k}s_y(\omega) \\ &= C(e^{-i\omega})C(e^{i\omega})s_y(\omega) \end{aligned} \tag{7.32}$$

Definition 7.51. The *gain function* is

$$G(\omega) = |C(e^{-i\omega})|$$

where $|C(e^{-i\omega})|$ is the modulus of $C(e^{-i\omega})$, i.e.

$$|C(e^{-i\omega})| = \sqrt{C(e^{-i\omega})C(e^{i\omega})}$$

Example 7.52. For the first-difference filter, $(1-L)$, the gain function is

$$\begin{aligned} G(\omega) &= \sqrt{(1-e^{-i\omega})(1-e^{i\omega})} \\ &= \sqrt{2}\sqrt{1-\cos(\omega)} \end{aligned} \tag{7.33}$$

where the second equality follows from the fact that $e^{-i\omega}+e^{i\omega}=2\cos(\omega)$.

Having defined the gain function, we are now in a better position to interpret the relationship between $s_{y^f}(\omega)$ and $s_y(\omega)$ given by (7.32):

$$s_{y^f}(\omega) = |C(e^{-i\omega})|^2 s_y(\omega) \equiv G(\omega)^2 s_y(\omega)$$

$G(\omega)^2$ is the squared gain of the filter and this relationship shows how filters isolate cycles, *viz.* they strengthen or weaken the spectrum of the original series on a frequency-by-frequency basis. For example, from (7.33) the first-difference filter, $(1-L)$, essentially shuts down zero frequency cycles.

Having studied the first-difference detrending filter for isolating low frequency cycles, let us look at another detrending frequency for isolating low frequency cycles, *viz.* the HP filter. The HP filter is a special case of the third type of detrending, i.e those that use filters that are designed to separate trend from cycles given the possibility of a slowly evolving rather than a constant trend; this third type of detrending contrasts from the assumption underlying the first two types of trend removal, detrending and differencing, that the data follow approximately constant growth rates. The HP filter is very popular in business cycle applications; the band pass filter is a leading alternative to the HP filter and will be discussed soon.

First decompose $\log y_t$ as

$$\log y_t = g_t + c_t \tag{7.34}$$

where g_t is the growth component and c_t is the cyclical component.

Definition 7.53. The *HP filter* estimates g_t and c_t, taking λ as given in the minimisation:

$$\sum_{t=1}^{T} c_t^2 + \lambda \sum_{t=3}^{T} [(1-L)^2 g_t]^2 \tag{7.35}$$

Trend removal is achieved by using the estimates from the HP filter, $\hat{g}_t$ and $\hat{c}_t$ as follows:

$$\tilde{y}_t = \log y_t - \hat{g}_t = \hat{c}_t$$

λ in (7.35) is the parameter that determines the degree of importance of the smoothness of the evolving growth component: a smoother g_t will have a smaller second difference. When $\lambda = 0$, no weight is placed on smoothness so all variation in $\log y_t$ is due to the trend component, c_t, whereas as $\lambda \longrightarrow \infty$, the trend is as smooth as possible, i.e. linear. As a rule of thumb, $\lambda = 100$ for annual data, $\lambda = 1600$ for quarterly data (business cycle data) and $\lambda = 14400$ for monthly data. To explain these choices, we will explore the frequency domain once more. But as an aside before doing this, we will illustrate some of the trajectories of $\hat{g}_t$ that result from this specification for sample data, including hours. Typically in business cycle applications, all (or most) series are subjected to the HP filter. See figure 6.3 of DeJong & Dave (2011) [21]. Detrended output $\tilde{y}_t$ by the three different trend removal procedures are presented in figure 6.4 of DeJong & Dave (2011) [21]. Of interest is the difference in volatility across the three measures: the linearly detrended series has the highest standard deviation, while the HP filtered series has the smallest standard deviation; that linearly detrended series have higher volatility than differenced series should be obvious from the above discussion of the HP filter. The behaviour in 6.4 provides additional support for the trend break in the series.

Now moving into the frequency domain to further discuss the HP filter, note that the specification of λ determines the influence of ω-specific y_t^{ω} on y_t

between g_t and c_t in (7.34). The gain function of the HP filter is given by

$$G(\omega) = \left[1 + \left(\frac{\sin(\omega/2)}{\sin(\omega_0/2)}\right)^4\right]^{-1}$$

where

$$\omega_0 = 2\text{arcsin}\left(\frac{1}{2\lambda^{\frac{1}{4}}}\right)$$

See Kaiser and Maravall (2001) [51] for more details on the derivation of this, etc. The parameter ω_0 is selected through the exact specification of λ and determines the frequency at which 50% of the filter gain has been completed, i.e. at which $G(\omega) = 0.5$. You can play around with this in a computing language such as MATLAB. The specification $\lambda = 1600$ for quarterly data implies that 50% of the filter gain has been completed at a 40-quarter cycle ($\omega = .157$); similarly, $\lambda = 400$ implies that the 50% completion point is at 20 quarters and $\lambda = 6,400$ implies that the point is at a 56-quarter cycle. However, while first-difference and HP filters can eliminate trends, they are not designed to remove seasonals. With quarterly data, seasonal frequencies are associated with $\frac{1}{4}$ and $\frac{1}{2}$ cycles per quarter and the squared gains associated with each of these filters are positive at these values. Business cycle models are usually not designed to explain seasonal variation, so we need to go further than the HP filter or work with variables that have had seasonal variations eliminated; aggregate variables are often reported in seasonally adjusted (SA) form. One such seasonal adjustment filter is the US Census Bureau's X-12 ARIMA found at http://www.census.gov/srd/www/x12a/; therefore, usually seasonal adjustment is not a big issue in preparing data for empirical analysis, but it is nonetheless instructive to understand the issue to appreciate the importance of the seasonal adjustment step; in addition, the issue will motivate the use of the band pass filter, which is the leading alternative to the HP filter. See figures 6.8 and 6.9 in DeJong & Dave (2011) [21] for an illustration of the importance of seasonal adjustment – non seasonally adjusted (NSA) versus SA form, including HP trends (and HP filtered series in figure 6.9) for both; also see figure 6.10 for spectra and remember that business cycle fluctuations lie roughly between 1/40 and 1/6 cycles per quarter.

The band pass (BP) filter shuts down all fluctuations *outside* of a chosen frequency band. Say we are interested in cycles with periods between p_l and p_u, e.g. 6 and 40 quarters in business cycles applications. Then the ideal BP filter has a squared gain such that

$$G(\omega)^2 = \begin{cases} 1 & \omega \in [2\pi/p_u, 2\pi/p_l] \\ 0 & \text{else} \end{cases}$$

However, implementation of the ideal BP filter is infeasible since this requires an infinite number of observations of unfiltered series as an input. This is clear from (7.36) below. Yet again, we can approximate BP filters using

different techniques proposed in the literature, for instance that by Baxter & King (1999) [6], Woitek (1998) [83] and Christiano & Fitzgerald (2003) [14]. We will concentrate on the Baxter & King (1999) [6] version here.

Definition 7.54. The *ideal symmetric BP filter* for a chosen frequency range is defined by

$$\alpha(L) = \sum_{j=-\infty}^{\infty} \alpha_j L^j \tag{7.36}$$

where $\alpha_{-j} = \alpha_j \ \forall j$ by symmetry.

This symmetry property is an important one for filters since it allows us to avoid the 'phase effect', which DeJong & Dave (2011) [21] characterise as follows:

> 'Under a phase effect, the timing of events between the unfiltered and filtered series, such as the timing of business cycle turning points, will be altered.' (DeJong & Dave, 2011: 132) [21]

A symmetric filter has a very simple form of Fourier transform. Here, it is given by

$$\begin{aligned}
\alpha(e^{-i\omega}) \equiv \alpha(\omega) &= \sum_{j=-\infty}^{\infty} \alpha_j e^{-i\omega j} \\
&= \alpha_0 + \sum_{j=1}^{\infty} \alpha_j (e^{-i\omega j} + e^{i\omega j}) \quad \because \alpha_{-j} = \alpha_j \ \forall j \\
&= \alpha_0 + 2\sum_{j=1}^{\infty} \alpha_j \cos(\omega) \quad \because e^{-i\omega} + e^{i\omega} = 2\cos\omega
\end{aligned}$$

Baxter & King approximate $\alpha(\omega)$ by the symmetric finite-ordered filter[8]

$$A(\omega) = \alpha_0 + 2\sum_{j=1}^{K} a_j \cos(\omega)$$

where at zero frequency

$$A(0) = \sum_{j=-K}^{K} a_j = 0$$

which ensures that $A(\omega)$ can remove a trend from the unfiltered series. $A(\omega)$ solves

$$\min_{a_j} \int_{-\pi}^{\pi} |\alpha(\omega) - A(\omega)|^2 d\omega \quad \text{subject to } A(0) = 0$$

[8] So, ∞ is replaced by K and α is replaced by a.

i.e. $A(\omega)$ minimizes deviations from $\alpha(\omega)$ (measured in terms of squared errors) that are accumulated over different frequencies. The solution is given by

$$a_j = \alpha_j + \theta \quad j = -K, \ldots, K$$
$$\alpha_j = \begin{cases} \frac{\omega_u - \omega_l}{\pi} & j = 0 \\ \frac{\sin(\omega_2 j) - \sin(\omega_1 j)}{\pi j} & j = \pm 1, \ldots, K \end{cases}$$
$$\theta = \frac{-\sum_{j=-K}^{K} \alpha_j}{2K+1}$$

where $\omega_l = 2\pi/p_u$ and $\omega_u = 2\pi/p_l$. Baxter & King suggest $K = 12$ for quarterly data, which results in the loss of 12 filtered observations at the beginning and end of the sample period. See figure 6.11 in DeJong & Dave (2011) that depicts squared gains associated with the ideal (optimal) filter and the Baxter-King approximated BP filter constructed over the 1/40 and 1/6 cycles per quarter (i.e. business cycle) range. See also figure 6.12 on BP filters using $K = 12$ with SA and NSA consumption series; these are smooth series and quite similar with no obvious trends or seasonal variations; spectra are shown in 6.13 and confirm the absence of trend and seasonal variation on the variations in the series. When graphing to show trend removal, graph SA and NSA versions and the estimated spectra.

7.3 Simulation Methods

The final lecture 'simulation methods' is based on chapters 17 and 18 in Greene (2011) [42], which is required reading. Please refer to these chapters in addition to any notes you may take during the lecture.

Bibliography

[1] I. Abramowitz, M. & Stegun. *Handbook of Mathematical Functions*. Dover Publications, New York, 1970.

[2] T. Amemiya. *Advanced econometrics*. Harvard university press, 1985.

[3] Diebold F.X. & Scotti C. Aruoba, S.B. Real-time measurement of business conditions. *Journal of Business & Economic Statistics*, 27(4):417–427, 2009.

[4] C.W.J. Bates, J.M. & Granger. The combination of forecasts. *OR*, pages 451–468, 1969.

[5] Lubrano M. & Richard J.F. Bauwens, L. *Bayesian inference in dynamic econometric models*. Oxford University Press, 2000.

[6] R.G. Baxter, M. & King. Measuring business cycles: Approximate band-pass filters for economic time series. *The Review of Economics and Statistics*, 81(4):575–593, November 1999.

[7] Boivin J. & Eliasz P.S. Bernanke, B. Measuring the effects of monetary policy: A factor-augmented vector autoregressive (favar) approach. *The Quarterly Journal of Economics*, 120(1):387–422, January 2005.

[8] P. Billingsley. *Probability and measure*, volume 939. Wiley, 2012.

[9] M. Boivin, J. & Giannoni. Dsge models in a data-rich environment. NBER Technical Working Papers 0332, National Bureau of Economic Research, Inc, December 2006.

[10] T. Bollerslev. *Glossary to ARCH (GARCH*)*. Oxford University Press, 2010.

[11] R.A. Brockwell, P.J. & Davis. *Time Series: Theory and Methods*. Springer, New York, 1991.

[12] R.V. Brown, J.W. & Churchill. *Complex Variables and Applications*. McGraw-Hill, Boston, 7th edition, 2003.

[13] C. Chatfield. *The analysis of time series – an introduction*. Chapman and Hall, London, 5th edition, 1996.

[14] T.J. Christiano, L.J. & Fitzgerald. The band pass filter. *International Economic Review*, 44(2):435–465, May 2003.

[15] F.X. Christoffersen, P.F. & Diebold. Optimal prediction under asymmetric loss. *Econometric Theory*, 13(6):pp. 808–817, December 1997.

[16] P.F. Christoffersen. Evaluating interval forecasts. *International Economic Review*, 39(4):841–62, November 1998.

[17] K.D. Clark, T.E. & West. Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. *Journal of Econometrics*, 135(1-2):155–186, 2006.

[18] K.D. Clark, T.E. & West. Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. *Journal of Econometrics*, 135(1âĂŞ2):155 – 186, 2006.

[19] M.W. Clark, T.E. & McCracken. Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, 105(1):85–110, November 2001.

[20] D.R. Cox. Tests of separate families of hypotheses. *Proc. Fourth Berkeley Symp.*, 1:105–123, 1961.

[21] C. DeJong, D.N. & Dave. *Structural macroeconometrics*. Princeton University Press, 2011.

[22] Gunther T.A. & Tay A.S. Diebold, F.X. Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39(4):863–83, November 1998.

[23] R.S. Diebold, F.X. & Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–63, July 1995.

[24] D. Edelman. Estimation of the mixing distribution for a normal mean with applications to the compound decision problem. *The Annals of Statistics*, 16(4):1609–1622, 1988.

[25] C. Efron, B. & Morris. Stein's estimation rule and its competitors–an empirical bayes approach. *Journal of the American Statistical Association*, 68(341):117–130, 1973.

[26] Komunjer I. & Timmermann A. Elliott, G. Biases in macroeconomic forecasts: Irrationality or asymmetric loss? *Journal of the European Economic Association*, 6(1):122–157, 2008.

[27] W. Enders. *Applied econometric time series*. John Wiley & Sons, 2008.

[28] K.D. Engel, C. & West. Exchange rates and fundamentals. *Journal of Political Economy*, 113(3):485–517, June 2005.

[29] R.F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, pages 987–1007, 1982.

[30] J. Fernández-Villaverde and J. Rubio-Ramírez. Macroeconomics and volatility: Data, models, and estimation. Technical report, National Bureau of Economic Research, 2010.

[31] L. Forni, M. & Reichlin. Let's get real: A factor analytical approach to disaggregated business cycle dynamics. *Review of Economic Studies*, 65(3):453–73, July 1998.

[32] A. Friedman. *Foundations of modern analysis*. Dover publications, 2010.

[33] M. Frieman. *Essays in positive economics*. University of Chicago Press, 1953.

[34] Carlin J.B. Stern H.S. & Rubin D.B. Gelman, A. *Bayesian Data Analysis, Second Edition (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, 2 edition, 2003.

[35] J. Geweke. *The dynamic factor analysis of economic time series*. North-Holland, Amsterdam, 1977.

[36] J. Geweke. The temporal and sectoral aggregation of seasonally adjusted time series. In *Seasonal Analysis of Economic Time Series*, NBER Chapters, pages 411–432. National Bureau of Economic Research, Inc, July 1978.

[37] H. Giacomini, R. & White. Tests of conditional predictive ability. *Econometrica*, 74(6):1545–1578, November 2006.

[38] L. & Small D. Giannone, D. & Reichlin. Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4):665–676, May 2008.

[39] S. Goldberg. John Wiley, New York, 1958.

[40] C W J Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–38, July 1969.

[41] R. Granger, C.W.J. & Ramanathan. Improved methods of combining forecasts. *Journal of Forecasting*, 3(2):197–204, 1984.

[42] W.H. Greene. *Econometric Analysis (7th Edition)*. Prentice Hall, 7 edition, 2011.

[43] J.D. Hamilton. *Time series analysis*, volume 2. Cambridge Univ Press, 1994.

[44] L.P. Hansen. Model uncertainty and policy evaluation: some theory and empirics - comments. *Proceedings*, 2005.

[45] A.C. Harvey. *The Econometric Analysis of Time Series*. Philip Allan, 1983.

[46] A.C. Harvey. Time series models. 1993.

[47] C.F. Horowitz, J.L. & Manski. Identification and robustness with contaminated and corrupted data. *Econometrica: Journal of the Econometric Society*, pages 281–302, 1995.

[48] C.F. Horowitz, J.L. & Manski. Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association*, 95(449):77–84, 2000.

[49] C.F. Imbens, G.W. & Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.

[50] C. James, W. & Stein. Estimation with quadratic loss. *Proc. Fourth Berkeley Symp.*, 1:361–379, 1960.

[51] A. Kaiser, R. & Maravall. *Measuring Business Cycles in Economic Time Series*. Springer-Verlag, Berlin, 2001.

[52] R.E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[53] R.S. Kalman, R.E. & Bucy. New results in linear filtering and prediction theory. *Journal of Basic Engineering*, 83(1):95–108, March 1961.

[54] Stock J.H. & Watson M.H. Knox, T. Empirical bayes forecasts of one time series using many predictors. NBER Technical Working Papers 0269, March 2001.

[55] T. Koopmans. Identification problems in economic model construction. *Econometrica*, 17:125–44, 1949.

[56] G. Lehmann, E.L. & Casella. *Theory of point estimation*. Springer, 2 edition, 1998.

[57] Runkle D.E. & Shapiro M.D. Mankiw, N.G. Are preliminary announcements of the money stock rational forecasts? *Journal of Monetary Economics*, 14(1):15 – 27, 1984.

[58] C.F. Manski. Identification of endogenous social effects: The reflection problem. *The review of economic studies*, 60(3):531–542, 1993.

[59] C.F. Manski. *Identification for prediction and decision*. Harvard University Press, 2007.

[60] C.F. Manski. The 2009 lawrence r. klein lecture: Diversified treatment under ambiguity*. *International Economic Review*, 50(4):1013–1041, 2009.

[61] J.V. Manski, C.F. & Pepper. Monotone instrumental variables: With an application to the returns to schooling. *Econometrica*, 68(4):pp. 997–1010, 2000.

[62] Stock J.H. & Watson M.W. Marcellino, M. A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series. *Journal of Econometrics*, 135(1-2):499–526, 2006.

[63] T. Maritz, J.S. & Lwin. *Empirical Bayes Methods*. Chapman and Hall, London, 2 edition, 1989.

[64] M.W. Mc Cracken. Robust out-of-sample inference. *Journal of Econometrics*, 99(2):195–223, December 2000.

[65] S.M. Melino, A. & Turnbull. Pricing foreign currency options with stochastic volatility. *Journal of Econometrics*, 45(1-2):239–265, 1990.

[66] V. Mincer, J.A. & Zarnowitz. The evaluation of economic forecasts. In *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, NBER Chapters, pages 1–46. National Bureau of Economic Research, Inc, July 1969.

[67] D.B. Nelson. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society*, pages 347–370, 1991.

[68] B. Palka. *An Introduction to Complex Function Theory*. Springer, Berlin, 1991.

[69] J.H. Richardson, M. & Stock. Drawing inferences from statistics based on multiyear asset returns. *Journal of Financial Economics*, 25(2):323–348, 1989.

[70] H. Robbins. The empirical bayes approach to statistical problems. *Annals of Mathematical Statistics*, 35:1–20, 1964.

[71] B. Rossi. Are exchange rates really random walks? some evidence robust to parameter instability. *Macroeconomic Dynamics*, 10(01):20–38, February 2006.

[72] E. Ruiz. Quasi-maximum likelihood estimation of stochastic variance models. *LSE Econometrics discussion paper*, 1992.

[73] C.A. Sargent, T.J. & Sims. Business cycle modeling without pretending to have too much a priori economic theory. Working Papers 55, Federal Reserve Bank of Minneapolis, 1977.

[74] T.J. Sargent. *Macroeconomic Theory.* Academic Press, New York, 1st edition, 1979.

[75] T.J. Sargent. *Macroeconomic Theory.* Academic Press, London, 2nd edition, 1987.

[76] F. Schorfheide. Learning and monetary policy shifts. *Review of Economic Dynamics*, 8(2):392–419, April 2005.

[77] H. Sims, C.A. & Uhlig. Understanding unit rooters: A helicopter tour. *Econometrica: Journal of the Econometric Society*, pages 1591–1599, 1991.

[78] S. Taylor. *Modelling Financial Time Series.* John Wiley & Sons, New York, 1986.

[79] A. Timmermann. *Forecast Combinations*, volume 1 of *Handbook of Economic Forecasting*, chapter 4, pages 135–196. Elsevier, 2006.

[80] K.D. West. Asymptotic inference about predictive ability. *Econometrica*, 64(5):1067–84, September 1996.

[81] K.D. West. *Forecast Evaluation*, volume 1 of *Handbook of Economic Forecasting*, chapter 3, pages 99–134. Elsevier, 2006.

[82] H. White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, September 2000.

[83] U. Woitek. A note on the baxter-king filter. Working Papers 9813, Business School - Economics, University of Glasgow, June 1998.

[84] C.H. Zhang. Compound decision theory and empirical bayes methods. *The Annals of Statistics*, 31(2):379–390, 2003.

[85] C.H. Zhang. General empirical bayes wavelet methods and exactly adaptive minimax estimation. *The Annals of Statistics*, 33(1):54–100, 2005.

Index