Econometrics II

Taught by Joel Horowitz
Northwestern University, winter 2016

Ludvig Sinander
Northwestern University

This version: 12 December 2018

These notes are based on an econometrics course for first-year PhD students taught by Joel Horowitz at Northwestern in winter 2016. The topics are limit theory and the asymptotic properties of extremum estimators.

I thank Joel for teaching a great class and for agreeing to let me share these notes, and Ahnaf Al Rafi, Bence Bardóczy, Ricardo Dahis and Joe Long for reporting errors.

Contents

1 Outline

Econometrics is about inferring functional relations among variables from data. Schematically, we observe a set of realisations of random vector (x, y), and wish to learn the function $f(\cdot, 0)$ that satisfies $y = f(x, u)$ for some unknown random vector u.[1] One intuitive way of thinking about this problem is to divide it into two parts: learning the parametric form ('shape') of $f(\cdot, 0)$, and learning the values of its parameters. (This intuition underlies parametric methods of estimation. Many nonparametric methods do not divide things up in this way.)

We'll need to sharpen up the question a bit in order to answer it. In general, we cannot learn $f(\cdot, 0)$. As we know from Manski's course, the most that we can hope to learn is the joint distribution $\mathbf{P}(x, y)$ of (x, y). Often, we are interested in some feature of the joint distribution, such as $\mathbf{P}(y|x)$, $\mathbf{E}(y|x)$ or some quantile of y conditional on x. These objects can be thought of as features of the function f when convenient.

A natural approach to estimating $\mathbf{P}(x, y)$ or $\mathbf{P}(y|x)$ when the support is finite is to use the empirical distribution. By a law of large numbers, this gives a pointwise consistent estimate.[2] When the support is uncountable, we could of course discretise the outcome space and apply the same reasoning. One drawback to this approach is that we'll often need a fine grid to provide a good approximation, in which case we'll need an astronomical dataset in order to have more than zero or one observation per cell. Another drawback is that we lose the tractability of analysis. (Discrete maths can be ugly.)

Another issue is that on any finite dataset, there is an infinite number of lines that you can fit through the data. In order for a fitted line to approximate the true relationship more and more closely as the sample size increases, we will therefore require some assumptions on the distribution of (y, x). For concreteness, consider the conditional mean function $g(x) \coloneqq \mathbf{E}(y|x)$. In this case, we can estimate g using nonparametric regression provided that g is continuous. We avoid discretisation by taking local averages, using a bandwidth that shrinks as the sample size increases. Continuity guarantees that whatever weighted local average you take (e.g. which kernel you use in kernel regression), the fitted line will get close to g as the sample size gets large. Continuity is often a pretty weak assumption in economics.[3]

[1] It is wlog to say that we want to learn $f(\cdot, 0)$. If we want to learn $f(\cdot, \xi)$ for $\xi \neq 0$, just reparameterise as $u' \coloneqq u - \xi$.

[2] In fact, this estimator is uniformly strongly consistent by the Glivenko–Cantelli theorem (Billingsley, 1995, p. 269).

[3] In other fields, discontinuity has to be allowed for explicitly. Joel gave the example of

A fundamental problem with nonparametric estimation techniques is the curse of dimensionality. Roughly speaking, this is that the sample size required to get a given level of estimator precision is exponentially increasing in the dimension of x. It can be proved that without stronger assumptions, the curse of dimensionality is unavoidable. Very loosely, the idea is that you're asking a finite dataset to tell you about an infinite-dimensional object.

One way of strengthening the assumptions to avoid the curse of dimensionality is to assume that g belongs to a finite-dimensional family of functions. Schematically, we assume that $g(x) = G(x, \theta)$ where G is a known function and θ is a finite-dimensional, unknown constant, i.e. a parameter. (In this parametric case, it is often natural to do inference directly on θ rather than trying to learn g directly.) It turns out (unsurprisingly) that parametrisation defeats the curse of dimensionality. The price we pay for this victory is the need to specify the function G. If we misspecify G, the math won't break, but the interpretation of results may be way off.

The obvious next conern is the 'accuracy' of our estimate of θ (or g). (If we didn't care about accuracy, there would be no reason to use the data!) Since an estimator of θ is a function of the (random) data, an estimator is a random variable. To characterise accuracy, we have to study this random variable. The problem is that the distribution of an estimator depends on the unknown distribution of (x, y). (If we knew the population distribution, we would once again have no use for a dataset.) Except under very stringent conditions, we cannot consistently estimate (never mind infer with certainty) the distribution of an estimator.

The way we get around this is by using approximations to the distribution of an estimator. If we are to trust these approximations, they must become increasingly good as the data gets increasingly good, in some sense to be made precise. The leading example will be asymptotic approximations, which are approximations that (usually) become increasingly good as the sample size grows. There are many kinds of asymptotic approximation, but the general idea is easily illustrated using the simplest central limit theorem. Suppose that our estimator $\widehat{\theta}$ is an average of n iid random variables (many estimators are), where n is the sample size. Then the central limit theorem says that $n^{1/2}\widehat{\theta}$ converges in distribution to $\mathcal{N}(\theta, \sigma^2)$, a two-dimensional family of distributions!

Our focus will be on asymptotic theory for parametric estimators. Besides this restriction, our treatment will be general, though common special cases will be mentioned along the way. We'll first cover basic (measure-theoretic)

image denoising.

probability theory, then limit theorems. Once the machinery is in place, we will develop the asymptotic theory of extremum estimators.

The course starts off with background probability theory, emphasising concepts required for asymptotic theory. We then state and prove several laws of large numbers and central limit theorems. With the technical machinery in place, we establish the consistency and asymptotic normality of general extremum estimators. We apply these results to maximum likelihood and generalised-method-of-moments estimators, and cover efficiency and specification tests while we're at it. Finally, we study hypothesis testing in the setting of extremum estimation.

The main references for the course are Amemiya (1985) and Newey and McFadden (1994). These texts (unlike most others in econometrics) raise and address all the important technical issues. Joel will also sometimes to refer Rao (1973) and Serfling (1980), two statistics texts that are worth studying for anyone interested in research in econometric theory. Several other texts that will be mentioned along the way, e.g. White (2001).

2 Probability theory

Official reading: Rao (1973, ch. 2).

This section covers basic (measure-theoretic) probability theory. There is a very large number of good texts on measure and probability theory; Joel mentioned Kolmogorov and Fomin (1975) in particular.

2.1 Measurable spaces

Let Ω be an arbitrary set. In the context of probability theory, we call Ω the sample space, and interpret it as the set of all possible states of nature, or outcomes of an experiment.

Definition 1. A collection $\mathcal{A}$ of subsets of Ω is called a σ-algebra (of subsets of Ω) iff

(1) $\Omega \in \mathcal{A}$.

(2) If $A \in \mathcal{A}$ then $A^{\mathrm{C}} \in \mathcal{A}$.[4]

(3) If $A_j \in \mathcal{A}$ for each $j \in \mathbf{N}$, then $\bigcup_{j\in\mathbf{N}} A_j \in \mathcal{A}$.[5]

Some additional properties of σ-algebras are easily derived. For example, if $A_j \in \mathcal{A}$ for each $j \in \mathbf{N}$, then

$$\bigcap_{j\in\mathbf{N}} A_j = \left(\bigcup_{j\in\mathbf{N}} A_j^{\mathrm{C}} \right)^{\mathrm{C}} \in \mathcal{A}$$

by properties (2) and (3). Another one is $\varnothing \in \mathcal{A}$, which follows from (1) and (2).

Definition 2. Let Ω be an arbitrary set, and let $\mathcal{A}$ be a σ-algebra of subsets of Ω. We call $(\Omega, \mathcal{A})$ a measurable space. The elements of $\mathcal{A}$ are called the measurable subsets of Ω; subsets of Ω that are not in $\mathcal{A}$ are called non-measurable.

The idea is that when we start assigning measure to subsets of Ω, we will only assign measure to the subsets of Ω that lie in $\mathcal{A}$; this is why we call these subsets measurable. It might seem like we could make our lives easier by choosing $\mathcal{A} = 2^{\Omega}$, the set of all subsets of Ω. We do not do this

[4] $A^{\mathrm{C}} \coloneqq \Omega \backslash A$ denotes the complement of A in Ω.

[5] $\mathbf{N} = \{1, 2, \ldots\}$ denotes the natural numbers.

because for general uncountable Ω, it leads to paradoxes. This will be less of a problem than it may first appear to be because the subsets of Ω missing from the σ-algebras we will be working with are very strange sets. But can still give rise to difficulties: measurability problems arise fairly frequently in econometric and economic theory.

In the context of probability theory, we sometimes call the measurable sets 'events', and interpret them as 'something that happens'. To illustrate, suppose we draw one coloured ball from an urn, formalised by the measurable space $(\{\text{red}, \text{blue}, \text{green}\}, 2^{\{\text{red},\text{blue},\text{green}\}})$. In ordinary language, one 'event' I might describe is 'I pick a red or a blue ball'. In the formalism, this corresponds to the event (measurable set) $\{\text{red}, \text{blue}\}$.

We might wonder how to choose our σ-algebra. An important criterion is that our σ-algebra contain enough subsets of Ω to allow us to study convergence (of measures, of measurable functions and of integrals). Convergence is a topological notion, so let's equip Ω with a topology. In order to obtain convergence results, we will need the σ-algebra to contain enough topologically interesting subsets of Ω; at the very least, it should contain all of the open subsets of Ω. It turns out that this minimal requirement is enough for most purposes, leading to the following definition.

Definition 3. Let $(\Omega, \mathcal{T})$ be a topological space. The Borel σ-algebra of subsets of Ω (relative to topology $\mathcal{T}$) is the smallest σ-algebra of subsets of Ω that contains $\mathcal{T}$ (all of the open sets). When Ω and its topology are clear from the context, we'll write $\mathcal{B}$ for the Borel σ-algebra.

More broadly, we will sometimes be interested in a certain collection $\mathcal{X}$ of subsets of Ω. In order to say measure-theoretic things about them, we need them to be measurable! To this end, we write $\sigma(\mathcal{X})$ for the smallest σ-algebra of subsets of Ω that contains $\mathcal{X}$; $\sigma(\mathcal{X})$ is called the σ-algebra generated by $\mathcal{X}$. Note that for a topological space $(\Omega, \mathcal{T})$, the σ-algebra $\sigma(\mathcal{T})$ generated by the open sets is exactly the Borel σ-algebra.

2.2 Measures

Let $\overline{\mathbf{R}} = \mathbf{R} \cup \{-\infty, \infty\}$ be the extended real line.

Definition 4. Given a measurable space $(\Omega, \mathcal{A})$, a function $\mu : \mathcal{A} \to \overline{\mathbf{R}}$ is a measure iff

(1) $\mu(A) \geq 0$ for each $A \in \mathcal{A}$.

(2) If $A_j \in \mathcal{A}$ for each $j \in \mathbf{N}$ are disjoint, then

$$\mu\left(\bigcup_{j\in\mathbf{N}} A_j\right) = \sum_{j\in\mathbf{N}} \mu(A_j).$$

(This property is called countable additivity.)

If in addition $\mu(\Omega) = 1$, then μ is a probability measure. We often use $\mathbf{P}$ to denote a probability measure.

Definition 5. Let $(\Omega, \mathcal{A})$ be a measurable space, and let μ be a measure on $(\Omega, \mathcal{A})$. We call $(\Omega, \mathcal{A}, \mu)$ a measure space. If μ is a probability measure, we call it a probability (measure) space.

Example 1 (Lebesgue measure)**.** Let $\Omega = [0, 1]$, let $\mathcal{B}$ be the Borel σ-algebra, and let $\mathcal{I}$ be the intervals of $[0, 1]$. Let $\lambda : \mathcal{I} \to \overline{\mathbf{R}}$ be the probability measure on $(\mathbf{R}, \mathcal{I})$ with the property that the value of an interval is its length, e.g. $\lambda((a, b]) = b - a$. It turns out that there's a unique extension of λ to the rest of $\mathcal{B}$ that respects the measure axioms.[6,7] This measure is called Lebesgue measure on $([0, 1], \mathcal{B})$. We can similarly define Lebesgue measure on other subsets of $\mathbf{R}^n$ for $n \in \mathbf{N}$.

One property of this measure is that any countable set has Lebesgue measure zero. To show this, take a collection of distinct points $x_j \in [0, 1]$ for each $j \in \mathbf{N}$. Since $\lambda(\{x_j\}) = 0$ for each $j \in \mathbf{N}$ (each is an interval of length zero), countable additivity yields

$$\lambda\left(\bigcup_{j\in\mathbf{N}} \{x_j\}\right) = \sum_{j\in\mathbf{N}} \lambda(\{x_j\}) = \sum_{j\in\mathbf{N}} 0 = 0.$$

It follows, for example, that $\lambda(\mathbf{Q} \cap [0, 1]) = 0$.[8]

Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, and consider some property P that does or does not hold at each $\omega \in \Omega$. We say that property P holds μ-almost everywhere (μ-a.e.) on Ω iff it holds everywhere on Ω except possibly on a set of μ-measure zero.[9] For example, a standard real analysis result states

[6] This follows from an extension theorem. Extension theorems provide conditions for the existence, and sometimes uniqueness, of an extension of a measure from a small set (e.g. the intervals) to a larger set (e.g. the Borel σ-algebra).

[7] Actually, there's a unique extension to $([0, 1], \mathcal{L})$, where $\mathcal{L}$ is a larger σ-algebra called the Lebesgue-measurable sets.

[8] $\mathbf{Q}$ denotes the rational numbers.

[9] That is, there is some $A \in \mathcal{A}$ such that P holds at every $\omega \in A$ and $\mu(A^c) = 0$.

that any monotone function $f : \mathbf{R} \to \mathbf{R}$ is continuous λ-a.e., where λ is Lebesgue measure on $(\mathbf{R}, \mathcal{B})$. When μ is a probability measure and P holds μ-a.e., we usually say that P holds μ-almost surely (μ-a.s.) or that P holds with probability 1. More generally, 'μ-almost' is used flexibly as an adverb, e.g. 'μ-almost all' or 'μ-almost every'.

2.3 Measurable functions

For a function $f : F \to G$ and a subset $B \subseteq G$, write

$$f^{-1}(B) \coloneqq \{x \in F : f(x) \in B\}.$$

(This is standard notation, but worth writing down just in case.)

Definition 6. Let $(F, \mathcal{F})$ and $(G, \mathcal{G})$ be measurable spaces. Then $f : F \to G$ is $\mathcal{F}/\mathcal{G}$-measurable iff for any $B \in \mathcal{G}$, $f^{-1}(B) \in \mathcal{F}$. When one or both σ-algebras are clear from the context, we sometimes shorten this to '$\mathcal{F}$-measurable' or simply 'measurable'.[10]

In words, measurable sets of values are generated by measurable sets of arguments. This is very similiar to the definition of continuity from topology, where a function is continuous iff open sets of values are generated by open sets of arguments. Notice that the measurability of a function depends only on the measurable spaces; it has nothing to do with measures defined on those spaces.[11]

Measurability is needed because the whole point of measure theory is to assign measure to things. If a function maps from one measurable space to another, we'd like to be able to say that the measure of a measurable set of values in G of the function is equal to the measure of the set of arguments in F that generate those values. But if our function is not measurable, then there will be sets of values in G that are counted as measurable according to $(G, \mathcal{G})$, but which are generated by a set of arguments in F that is not measurable according to $(F, \mathcal{F})$.

A special case of interest is where $(F, \mathcal{F}) = (\Omega, \mathcal{A})$ is an arbitrary measurable space and $(G, \mathcal{G}) = (\mathbf{R}^k, \mathcal{B})$ where $\mathcal{B}$ is the Borel σ-algebra on $\mathbf{R}^k$. It

[10] For the special case of functions $f : \mathbf{R}^n \to \mathbf{R}$ with the Borel σ-algebras $\mathcal{B}^n$ and $\mathcal{B}$, $\mathcal{B}^n/\mathcal{B}$-measurability is sometimes called Borel-measurability.

[11] Some confusing (in my view) terminology was used at this point in the lecture. Consider two measurable spaces $(F, \mathcal{F})$ and $(G, \mathcal{G})$, a measure μ on $(F, \mathcal{F})$, and a function $f : F \to G$. Above, I defined the property of $\mathcal{F}/\mathcal{G}$-measurability of f. Joel called this same property μ-measurability of f. But as I pointed out, the measure μ has nothing to do with it!

turns out that in this case, a function $f : \Omega \to \mathbf{R}^k$ is measurable iff

$$\{\omega \in \Omega : f(\omega) \leq z\} \in \mathcal{A} \quad \text{for each } z \in \mathbf{R}^k.$$

We can now define random elements, which are principal characters in the sequel.

Definition 7. Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space, and let $(S, \mathcal{S})$ be a measurable space. A random element of $(S, \mathcal{S})$ defined on $(\Omega, \mathcal{A}, \mathbf{P})$ is an $\mathcal{A}/\mathcal{S}$-measurable function $X : \Omega \to S$.

The set S in which a random element takes values can be entirely arbitrary; it need not be a topological or metric space, for example. But often, S will be a metric space. In this case, we will sometimes abuse terminology by saying 'random element of (S, ρ) (defined on $(\Omega, \mathcal{A}, \mathbf{P})$)', on the understanding that S is equipped with a σ-algebra, usually the Borel σ-algebra generated by the topology induced by the metric ρ.

Definition 8. A random variable is a random element of $(\mathbf{R}, \mathcal{B})$. A random n-vector is a random element of $(\mathbf{R}^n, \mathcal{B})$. A random $n \times m$ matrix is a random element of $(\mathbf{R}^{n \times m}, \mathcal{B})$.

For a random element $X : \Omega \to S$ and a measurable subset B of S, $X^{-1}(B)$ is the set of states of the world $\omega \in \Omega$ at which $X(\omega)$ lies in B. We know that $X^{-1}(B) \in \mathcal{A}$ since X is a measurable function. But $\mathcal{A}$ may contain lots of other events that do not correspond to $X^{-1}(B)$ for some $B \in \mathcal{S}$. These other events are not interesting for the study of X, so we sometimes wish to use smaller the σ-algebra that contains all the sets of interest for X but no others. In our previous jargon, what we want is the σ-algebra generated by the interesting sets, viz. $\sigma\big(\{X^{-1}(B)\}_{B \in \mathcal{S}}\big)$. The name of this object is often shortened to 'the σ-algebra generated by X', or $\sigma(X)$.

Definition 9. Let X be a random element of $(S, \mathcal{S})$ defined on $(\Omega, \mathcal{A}, \mathbf{P})$. The law (or distribution) of X is the function $\mathcal{L}_X : \mathcal{S} \to \mathbf{R}$ given by $\mathcal{L}_X(B) \coloneqq \mathbf{P}(X \in B)$ for each $B \in \mathcal{S}$.[12]

If we are interested only in the behaviour of the random element X, then all of the information we need is contained in its law $\mathcal{L}_X$. It does not matter what probability space it is defined on! In fact, $(S, \mathcal{S}, \mathcal{L}_X)$ is itself a probability space, and the random element Y defined by $Y(s) \coloneqq s$ for each $s \in S$ on this probability space has law $\mathcal{L}_Y = \mathcal{L}_X$.

When X is a random vector, there's is an alternative (more tractable) object that fully describes the behaviour of X: the CDF.

[12] $\mathbf{P}(X \in B)$ is shorthand for $\mathbf{P}(\{\omega \in \Omega : X(\omega) \in B\})$.

Definition 10. Let X be a random vector. The cumulative distribution function (CDF) of X is $F_X : \mathbf{R}^n \to [0,1]$ defined by

$$F_X(x_1, \ldots, x_n) \coloneqq \mathcal{L}_X((-\infty, x_1] \times \cdots \times (-\infty, x_n])$$

for each $(x_1, \ldots, x_n) \in \mathbf{R}^n$.

It is intuitive (but not quite obvious, I think) that F_X fully characterises the law of a random vector. Precisely stated: for random vectors X and Y, $\mathcal{L}_X = \mathcal{L}_Y$ (setwise) iff $F_X = F_Y$ (pointwise). See Rosenthal (2006, Proposition 6.0.2) for a (very easy) proof.

Some properties of CDFs are that they are right-continuous, nondecreasing, and satisfy $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$. (In fact, any function $F : \mathbf{R}^n \to [0,1]$ with these four properties is the CDF of some random vector on some probability space.)

We can also define random elements $X : \Omega \to \overline{\mathbf{R}}^n$ that are like random vectors but can take infinite values. If $\mathcal{L}_X(\mathbf{R}^n) < 1$, then the distribution $\mathcal{L}_X$ of X is said to be defective. Conversely, if $\mathcal{L}_X(\mathbf{R}^n) = 1$ then the distribution is called proper. Though we rarely *want* to work with random vectors that take infinite values with positive probability, we sometimes obtain a random vector with a defective distribution as the limit of a sequence of random vectors with proper distributions.

2.4 Independence

Probability theory is basically measure theory plus independence. This will become increasingly clear: all of the interesting theorems that we will state specifically for probability measures (rather than general measures) assume (some weakened form of) independence.

Definition 11. For a probability space $(\Omega, \mathcal{A}, \mathbf{P})$, events $A, B \in \mathcal{A}$ are independent iff $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$.

Sometimes, we wish to impose independence for two whole classes of events. Call $\mathcal{F}$ a sub-σ-algebra of the σ-algebra $\mathcal{A}$ iff it is a σ-algebra of subsets of Ω and $\mathcal{F} \subseteq \mathcal{A}$.

Definition 12. For a probability space $(\Omega, \mathcal{A}, \mathbf{P})$, sub-$\sigma$-algebras $\mathcal{F}$ and $\mathcal{G}$ of $\mathcal{A}$ are independent iff $\mathbf{P}(F \cap G) = \mathbf{P}(F)\mathbf{P}(G)$ for every $F \in \mathcal{F}$ and $G \in \mathcal{G}$.

Fix a probability space $(\Omega, \mathcal{A}, \mathbf{P})$ and a measurable space $(S, \mathcal{S})$. Two random elements X and Y of $(S, \mathcal{S})$ are called independent iff their generated σ-algebras $\sigma(X)$ and $\sigma(Y)$ are independent. This just means that

$$\mathbf{P}(X \in B_X, Y \in B_Y) = \mathbf{P}(X \in B_X)\mathbf{P}(Y \in B_Y)$$

for any measurable subsets B_X and B_Y of S. (So for random vectors, B_X and B_Y are any Borel sets.)

2.5 The Lebesgue integral

A problem with the Riemann integral is that many interesting functions are not Riemann-integrable. This defect is addressed by the Lebesgue integral. The standard Lebesgue integral of a measurable function $f : \Omega \to \mathbf{R}$ is written $\int_\Omega f \mathrm{d}\lambda$, where λ is Lebesgue measure on $\mathbf{R}$. To integrate over a measurable subset $S \subseteq \Omega$, define

$$\int_S f \mathrm{d}\lambda \coloneqq \int_\Omega f \mathbf{1}_S \mathrm{d}\lambda$$

where $\mathbf{1}_S$ is the indicator function for S.[13]

A function must be measurable to be Lebesgue-integrable, but it must also satisfy a boundedness condition to avoid the undefined expression $\infty - \infty$. A necessary and sufficient condition for a measurable function f to be integrable is that $\int_\Omega |f| \mathrm{d}\lambda < \infty$. For the case $\Omega = \mathbf{R}$, any Riemann-integrable function is also Lebesgue-integrable, and in such cases the two integrals coincide. But many functions are Lebesgue- but not Riemann-integrable.[14,15]

It's important that the Lebesgue integral allows the domain of the integrand f to be an arbitrary measure space $(\Omega, \mathcal{A})$ rather than (say) $\mathbf{R}^n$. This provides a powerful generalisation of Riemann integration; for example, we can integrate over functional spaces.

In constructing the ordinary Lebesgue integral, it quickly becomes apparent that we can replace the Lebesgue measure λ with whatever measure μ we like, leading to the generalised Lebesgue integral $\int_S f \mathrm{d}\mu$. Of course, the conditions under which f is integrable depend on what measure we're integrating with respect to. If $\int_S f \mathrm{d}\mu$ exists, we say that f is μ-integrable.

Now consider the case in which μ is a probability measure on a measurable space $(\Omega, \mathcal{A})$, so that the measurable function f is a random variable. Let's use the more familiar notation of $\mathbf{P}$ for the measure and X for the random

[13] I.e. $\mathbf{1}_S(\omega) = 1$ iff $\omega \in S$, 0 otherwise.

[14] For example, consider the function $f : \mathbf{R} \to \mathbf{R}$ such that $f(x) = \mathbf{1}(x \notin \mathbf{Q})$. This function is not Riemann-integrable, but it is Lebesgue-integrable, with $\int_{\mathbf{R}} f \mathrm{d}\lambda = 1$ as we would hope. More generally, there are certain kinds of discontinuity that Lebesgue integration can handle but Riemann integration cannot.

[15] In terms of how they are constructed, the difference between the two integrals is that while the Riemann integral considers the limit of a sequence of approximations to the 'area under f' constructed by discretising the x axis, the Lebesgue integral considers approximations constructed by discretising the y axis.

variable. In this setting, the Lebesgue integral $\int_\Omega X \mathrm{d}\mathbf{P}$ is also called the expected value of X (or the expected value of the distribution $\mathcal{L}_X$). Since not all measurable functions are $\mathbf{P}$-integrable, there are random variables whose expectation is undefined. (One example is a Cauchy-distributed random variable.)

There are various kinds of notation for the expected value, including:

$$\mathbf{E}(X) = \int_\Omega X \mathrm{d}\mathbf{P} = \int_\Omega X(\omega) \mathrm{d}\mathbf{P}(\omega) = \int_\Omega X(\omega) \mathbf{P}(\mathrm{d}\omega).$$

Sometimes, we wish to work with $\mathcal{L}_X$ rather than X and $\mathbf{P}$; in these cases we sometimes write $\mathbf{E}(\mathcal{L}_X)$ (confusingly!). It is simple to show (e.g. Rosenthal (2006, Theorem 6.1.1)) that the expected value can be written as an integral with respect to the law of X,[16] so that

$$\mathbf{E}(X) = \mathbf{E}(\mathcal{L}_X) = \int_{\mathbf{R}} x \mathrm{d}\mathcal{L}_X(x) = \int_{\mathbf{R}} x \mathcal{L}_X(\mathrm{d}x).$$

Finally, we can rewrite the integral in terms of the CDF F_X rather than the law $\mathcal{L}_X$. This is unsurprising in view of the fact that CDFs coincide iff the laws do. The integral w.r.t. a CDF is called a Stieltjes integral, and can be written in various ways:

$$\mathbf{E}(X) = \int_{\mathbf{R}} x \mathrm{d}F_X(x) = \int_{\mathbf{R}} x F_X(\mathrm{d}x).$$

The Stieltjes integral coincides with the Lebesgue integral, but is defined differently (in terms of the CDF).

Let X be a random variable; then X^n for $n \in \mathbf{N}$ is also a random variable (easy proof). $\mathbf{E}\left(X^n\right)$ is called the nth moment of X, and $\mathbf{E}\left((X - \mathbf{E}(X))^n\right)$ is called the nth central moment. (Of course, a given moment of X need not exist or be finite.) The second central moment is called the variance of X, denoted $\mathrm{Var}(X)$.

While we're at it, here are two related concepts. Let X and Y be random variables. Their covariance is

$$\mathrm{Cov}(X, Y) \coloneqq \mathbf{E}\left([X - \mathbf{E}(X)][Y - \mathbf{E}(Y)]\right),$$

and their correlation is

$$\mathrm{Corr}(X, Y) \coloneqq \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}.$$

There are a lot of algebraic shortcuts involving variances, covariances and correlations that are hopefully familiar from undergrad. I won't list them here, but I will make use of them!

[16] Recall that $(\mathbf{R}, \mathcal{B}, \mathcal{L}_X)$ is a probability space.

2.6 The Radon–Nikodým theorem

The Radon–Nikodým theorem gives conditions under which a measure can be represented as a Lebesgue integral w.r.t. another measure. In particular, given measures μ and ν on a measurable space $(\Omega, \mathcal{A})$, we're interested in representations of the form

$$\nu(A) = \int_A f \mathrm{d}\mu \quad \text{for each } A \in \mathcal{A} \tag{1}$$

for some nonnegative, μ-integrable (hence $\mathcal{A}/\mathcal{B}$-measurable) function $f : \Omega \to \mathbf{R}$.

The following concept will turn out to be the key.

Definition 13. Let μ and ν be measures defined on a measurable space $(\Omega, \mathcal{A})$. We say that μ dominates ν, written $\mu \gg \nu$, iff for any $A \in \mathcal{A}$, $\mu(A) = 0$ implies $\nu(A) = 0$. We also sometimes say that ν is absolutely continuous w.r.t. μ.[17]

Suppose that the representation is possible: there is a nonnegative f such that (1) holds. Then if $\mu(A) = 0$ for some $A \in \mathcal{A}$, it follows that

$$\nu(A) = \int_A f(\omega)\mu(\mathrm{d}\omega) = \int_\Omega f(\omega)\mathbf{1}_A(\omega)\mu(\mathrm{d}\omega) = 0$$

since $\mathbf{1}_A(\omega) = 0$ for all $\omega \in \Omega$ outside a set of μ-measure zero (viz. the set A). So we've learned that if (1) holds, then $\mu \gg \nu$. The Radon–Nikodým theorem is the (far less obvious) converse to this result: if $\mu \gg \nu$, then there exists a nonnegative f such that (1) holds. Actually, there's a caveat: in order for the converse to hold, both measures must be σ-finite.

In words, a measure defined on $(\Omega, \mathcal{A})$ is σ-finite iff there is a countable, measurable cover of Ω such that every piece of the cover is assigned finite measure. It should be obvious that every probability measure is σ-finite. For reference, here's a schematic definition:

Definition 14. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space. The measure μ is σ-finite iff there is a countable collection $\{A_j\}_{j \in \mathbf{N}}$ of subsets of Ω such that

[17] There's some unhelpful confusion of terminology here. Most authors (e.g. Kolmogorov and Fomin (1975) and Billingsley (1995)) use 'absolutely continuous' synonymously with 'dominates', as in my definition. But Rosenthal (2006, p. 143) defines 'ν absolutely continuous w.r.t. μ' to mean that there exists a nonnegative f such that representation (1) holds. By the Radon–Nikodým theorem, the two turn out to be equivalent for σ-finite measures, but they are still distinct properties that should have distinct names.

(1) $A_j \in \mathcal{A}$ for each $j \in \mathbf{N}$

(2) $\bigcup_{j \in \mathbf{N}} A_j = \Omega$

(3) $\mu(A_j) < \infty$ for each $j \in \mathbf{N}$.

With definitions in place, we're ready to state the theorem.

Theorem 1 (Radon–Nikodým)**.** Let μ and ν be σ-finite measures on some measurable space $(\Omega, \mathcal{A})$, and suppose that $\mu \gg \nu$. Then there is a nonnegative, μ-integrable function f such that $\nu(A) = \int_A f \mathrm{d}\mu$ for each $A \in \mathcal{A}$.

f in the theorem is usually called the density of ν with respect to μ. (The densities familiar from undergrad are densities w.r.t. Lebesgue measure λ.) It is also called a Radon–Nikodým derivative, denoted $\mathrm{d}\nu/\mathrm{d}\mu$. This name is obviously motivated by the fact that

$$\nu(A) = \int_A \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \mathrm{d}\mu \quad \forall A \in \mathcal{A},$$

an expression analogous to the fundamental theorem of calculus for ordinary derivatives and the Riemann integral.[18]

Another property of representation (1) is that the density f is unique up to sets of measure zero: if f and g are both densities of ν w.r.t. μ then $\mu(f \neq g) = 0$.[19] This is a corollary to a basic property of the Lebesgue integral; a proof can be found in Billingsley (1995, Theorem 16.10).

Example 2 (density w.r.t. counting measure)**.** As already mentioned, the probability density functions familiar from undergrad are Radon–Nikodým derivatives w.r.t. Lebesgue measure. In this example, we'll see that probability *mass* functions (for discrete random variables) are in fact Radon–Nikodým derivatives w.r.t. a measure called counting measure.

Let $(\mathbf{N}, \mathcal{A})$ be a measurable space,[20] and let $\mathbf{P}$ be a probability measure on it. Let $c : \mathcal{A} \to \overline{\mathbf{R}}$ be defined by $c(A) \coloneqq |A|$ for $A \in \mathcal{A}$.[21] c is obviously a σ-finite measure on $(\mathbf{N}, \mathcal{A})$; it is called counting measure. Perhaps unsurprisingly,

[18] There are other properties which Radon–Nikodým derivatives share with ordinary derviatives. One of these is the chain rule: for σ-finite measures $\nu \ll \mu \ll \tau$ on some measurable space, we have $\frac{\mathrm{d}\nu}{\mathrm{d}\tau} = \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \frac{\mathrm{d}\mu}{\mathrm{d}\tau}$.

[19] That is, $\mu(\{\omega \in \Omega : f(\omega) \neq g(\omega)\}) = 0$.

[20] Recall that $\mathbf{N} = \{1, 2, \ldots\}$.

[21] $|A|$ denotes the number of elements in the set A. For A infinite, $|A| = \infty$; moreover $|\varnothing| = 0$.

integration w.r.t. counting measure is equivalent to summation: formally, for any c-integrable (and measurable) function $f : \mathbf{N} \to \mathbf{R}$,

$$\int_A f \mathrm{d}c = \sum_{n \in A} f(n) \quad \text{for every } A \in \mathcal{A}.$$

(If you know how the Lebesgue integral is defined, then this should be trivial.)

Now let's apply the Radon–Nikodým theorem to the σ-finite measures $\mathbf{P}$ and c. The only set to which c assigns measure zero is $\varnothing$, and $\mathbf{P}(\varnothing) = 0$; hence $\mathbf{P} \ll c$. So by the Radon–Nikodým theorem, there is nonnegative and c-integrable function $f : \mathbf{N} \to \mathbf{R}$ such that

$$\mathbf{P}(A) = \int_A f \mathrm{d}c = \sum_{n \in A} f(n) \quad \text{for every } A \in \mathcal{A}.$$

Moreover, f is unique: we already knew it to be unique up to sets of c-measure zero, but the only set of c-measure zero is $\varnothing$.

Let's suppose our σ-algebra is rich enough that it contains all the singletons: $\{n\} \in \mathcal{A}$ for each n. (Using such a rich σ-algebra would cause problems on an uncountable sample space, but it's fine here since $\mathbf{N}$ is countable.) Then for each $n \in \mathbf{N}$ we have $\mathbf{P}(\{n\}) = f(n)$, showing that f is the probability mass function.[22]

The Radon–Nikodým theorem turns out to have various uses. It is used to establish the existence of useful constructs such as conditional probabilities and conditional expectations (see below). In estimation, if the data are distributed according to ν_θ for some parameter θ, then a clever choice of μ can give us a convenient likelihood function $\mathcal{L}(\theta) \coloneqq \mathrm{d}\nu_\theta / \mathrm{d}\mu$.

2.7 Conditional probability

This section may be hard to follow. I recommend Billingsley (1995, sec. 33).

Consider two events A and G of a probability space $(\Omega, \mathcal{A}, \mathbf{P})$. Imagine that we learn that event G obtains, and would like to revise the probability of event A in light of this information. (This scenario is obviously ubiquitous in economics and econometrics.) How do we formalise this?

When $\mathbf{P}(G) > 0$, the answer is intuitive and easy: we say that the conditional probability is $\mathbf{P}(A \cap G)/\mathbf{P}(G)$. (Draw a Venn diagram.) The problem is that we (very) often wish to condition on a probability-zero event,

[22] The Radon–Nikodým theorem doesn't tell us what the function f is, only that it exists. Characterising f will generally require an argument specific to the measure space at hand.

i.e. $\mathbf{P}(G) = 0$. (For example, the realisation of a normally distributed signal.) The ratio formula does not apply in this case, so a subtler construction is called for.

To build some intuition, consider a finite probability space $\left(\Omega, 2^{\Omega}, \mathbf{P}\right)$ in which $\mathbf{P}(\{\omega\}) > 0 \ \forall \omega \in \Omega$. Let $\mathbf{Q}(A|G) \coloneqq \mathbf{P}(A \cap G)/\mathbf{P}(G)$ be the 'ordinary' probability of A conditional on G. Notice that $\mathbf{Q}(A|\{\omega\}) = \mathbf{1}(\omega \in A)$. It follows that for events A and G,

$$\begin{aligned}\mathbf{P}(A \cap G) = \sum_{\omega \in A \cap G} \mathbf{P}(\{\omega\}) &= \sum_{\omega \in G} \mathbf{1}(\omega \in A)\mathbf{P}(\{\omega\}) \\ &= \sum_{\omega \in G} \mathbf{Q}(A|\{\omega\})\mathbf{P}(\{\omega\}) = \int_G \mathbf{Q}(A|\{\omega\})\mathbf{P}(\mathrm{d}\omega).\end{aligned}$$

(The Lebesgue integral is used because unlike the sum, it remains defined when we move to uncountable probability spaces.) We'd like our definition of conditional probability to respect $\mathbf{P}(A \cap G) = \int_G \mathbf{Q}(A|\{\omega\})\mathbf{P}(\mathrm{d}\omega)$. Since this property does not involve division by $\mathbf{P}(G)$, we can require it to hold even when $\mathbf{P}(G) > 0$.

We require an additional bit of machinery. Recall that $\mathcal{A}$ is a σ-algebra of subsets of Ω, we call $\mathcal{G}$ a sub-σ-algebra of $\mathcal{A}$ iff $\mathcal{G}$ is a σ-algebra and $\mathcal{G} \subseteq \mathcal{A}$. Since $\mathcal{G}$ contains only a subset of the events in $\mathcal{A}$, it provides a 'coarser' description of the state of the world $\omega \in \Omega$.

Here's an analogy: I am trying to convey to you the colour of the sky. The sky can be any colour ω in $\Omega = [0, 1]$, where 0 is 'totally blue' and 1 is 'totally red' (say). I have at my disposal a small vocabulary of English phrases that I can use to communicate with, consisting of (any combination of) 'blue' $(= [0, \frac{1}{3}])$, 'purple' $(= (\frac{1}{3}, \frac{2}{3}])$ and 'red' $(= (\frac{2}{3}, 1])$. Formally, my language is the σ-algebra generated by the events 'blue', 'purple' and 'red':

$$\mathcal{A} = \sigma\left(\left\{[0, \tfrac{1}{3}], (\tfrac{1}{3}, \tfrac{2}{3}], (\tfrac{2}{3}, 1]\right\}\right) = \left\{\varnothing, [0, \tfrac{1}{3}], (\tfrac{1}{3}, \tfrac{2}{3}], (\tfrac{2}{3}, 1], [0, \tfrac{2}{3}], (\tfrac{1}{3}, 1], [0, 1]\right\}.$$

Now suppose that my English deteriorates: I forget the words 'blue' and 'purple', and am left only with the coarser term 'blurple' $(= [0, \frac{2}{3}])$. My newly worsened language is

$$\mathcal{G} = \sigma\left(\left\{[0, \tfrac{2}{3}], (\tfrac{2}{3}, 1]\right\}\right) = \left\{\varnothing, [0, \tfrac{2}{3}], (\tfrac{2}{3}, 1], [0, 1]\right\}.$$

Clearly $\mathcal{G}$ is a sub-σ-algebra of $\mathcal{A}$. The example should illustrate the sense in which $\mathcal{G}$ is a coarser language than $\mathcal{A}$.[23]

[23] Aside: a sequence of increasing σ-algebras (each one a sub-σ-algebra of the next one) is called a filtration. A filtration provides formal way to talk about possible histories. They are therefore important for the study of stochastic processes.

Back to conditional probability. In general, we are interested in the probability of A conditional on several different events G. For example, suppose that we want to condition on some random variable X being realised in a Borel set B, i.e. the event $G_B = X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$; usually we'd like to be able to condition on any Borel event of this sort, not just a particular one. The rigorous construction of conditional probabilities requires us to specify in advance what collection of events we want to be able to condition on. It is perhaps not surprising that we require this collection of conditioning events to be a σ-algebra. But we also need it to not lead to measurability problems, and that requires the conditioning σ-algebra to be a sub-σ-algebra $\mathcal{G}$ of $\mathcal{A}$.[24]

To make the sub-σ-algebra $\mathcal{G}$ explicit, we can write the probability of A conditional on $G \in \mathcal{G}$ as $\mathbf{Q}_{\mathcal{G}}(A|G)$, which (for fixed A) is a mapping $\mathcal{G} \to \mathbf{R}$. It turns out, however, to be technically more convenient to define the conditional probability as a $\mathcal{G}$-measurable mapping $\Omega \to \mathbf{R}$, denoted $\mathbf{P}(A|\mathcal{G})(\omega)$. Since it's $\mathcal{G}$-measurable, $\mathbf{P}(A|\mathcal{G})(\omega) = \mathbf{P}(A|\mathcal{G})(\omega')$ whenever ω and ω' lie in all the same sets $G \in \mathcal{G}$; in this sense, the conditional probability can only vary between states of the world that are distinguishable using $\mathcal{G}$.

Let's try to clarify this using the example of conditioning on a random variable. We formalise 'conditioning on a random variable' as conditioning on its generated σ-algebra $\sigma(X)$. Since $\mathbf{P}(A|\sigma(X))$ is $\sigma(X)$-measurable, $\mathbf{P}(A|\sigma(X))(\omega) = \mathbf{P}(A|\sigma(X))(\omega')$ whenever $X(\omega) = X(\omega')$. Conversely, if $X(\omega) \neq X(\omega')$ then in general $\mathbf{P}(A|\sigma(X))(\omega) \neq \mathbf{P}(A|\sigma(X))(\omega')$. Informally, although $\mathbf{P}(A|\sigma(X))(\cdot)$ is really a function of ω, $\sigma(X)$-measurability means (precisely) that it behaves as if it were a function of (the realised value of) X. So this construction allows us to condition on events like $X = x$.[25] But importantly, this conditional probability does *not* give us statements about the probability of A conditional on (say) $X \leq x$. To get conditional probabilities like that, we need a coarser σ-algebra since we're conditioning on 'larger' events.

All this chatting has established that we want our conditional probability $\mathbf{P}(A|\mathcal{G})$ to be a $\mathcal{G}$-measurable mapping $\Omega \to \mathbf{R}$ that satisfies $\mathbf{P}(A \cap G) = \int_G \mathbf{P}(A|\mathcal{G})\mathrm{d}\mathbf{P}$ for every $G \in \mathcal{G}$. The last condition obviously requires that $\mathbf{P}(A|\mathcal{G})$ be $\mathbf{P}$-integrable. (And implies that $\mathbf{P}(A|\mathcal{G}) \in [0,1]$ $\mathbf{P}$-a.s.) Let's put it all together!

[24] If it isn't clear why measurability problems would arise without this requirement, think about it until it's clear!

[25] To construct this probability explicitly, pick some $\omega_x \in \{\omega \in \Omega : X(\omega) = x\}$ for each $x \in \mathbf{R}$ (any will do), and define the 'intuitive conditional probability' $\mathbf{Q}_{\sigma(X)}(A|X = x) \coloneqq \mathbf{P}(A|\sigma(X))(\omega_x)$.

Definition 15. Consider a probability space $(\Omega, \mathcal{A}, \mathbf{P})$ and a sub-σ-algebra $\mathcal{G}$ of $\mathcal{A}$. A random variable $\mathbf{P}(A|\mathcal{G}) : \Omega \to \mathbf{R}$ is a conditional probability of A on $\mathcal{G}$ iff it is (1) $\mathcal{G}$-measurable, (2) $\mathbf{P}$-integrable, and (3) satisfies

$$\mathbf{P}(A \cap G) = \int_G \mathbf{P}(A|\mathcal{G}) \mathrm{d}\mathbf{P} \quad \forall G \in \mathcal{G}.$$

Proposition 1. Consider a probability space $(\Omega, \mathcal{A}, \mathbf{P})$, a sub-$\sigma$-algebra $\mathcal{G}$ of $\mathcal{A}$, and an event $A \in \mathcal{A}$. Then there exists a conditional probability $\mathbf{P}(A|\mathcal{G})$ of A on $\mathcal{G}$.

Proof. $\mu_A(G) \coloneqq \mathbf{P}(A \cap G) \leq \mathbf{P}(G)$ for every $G \in \mathcal{G}$, so $\mu_A \ll \mathbf{P}$. Since both are σ-finite measures, the Radon–Nikodým theorem implies that there exists a (nonnegative,) $\mathcal{G}$-measurable and $\mathbf{P}$-integrable function $f_A : \Omega \to \mathbf{R}$ such that

$$\mathbf{P}(A \cap G) = \mu_A(G) = \int_G f_A \mathrm{d}\mathbf{P} \quad \forall G \in \mathcal{G}. \tag{2}$$

So f_A is a conditional probability of A on $\mathcal{G}$. ■

The use of the Radon–Nikodým theorem highlights that conditional probabilities are generally not unique, though they are unique up to sets of measure zero. This is not ideal, but we cannot fix in it full generality. Instead, we must pick the 'right' conditional probability in each individual application.[26]

So far, we have defined a conditional probability for a single, fixed event $A \in \mathcal{A}$. What we'd really like is a conditional probability *measure* $\mu(\cdot|\mathcal{G})$ that gives the conditional probability of every event in $\mathcal{A}$. We'll obviously construct the random function $\mu(\cdot|\mathcal{G}) : \mathcal{A} \to \mathbf{R}$ according to $\mu(A|\mathcal{G}) \coloneqq \mathbf{P}(A|\mathcal{G})$ for each $A \in \mathcal{A}$,[27] for some collection $\{\mathbf{P}(A|\mathcal{G})\}_{A \in \mathcal{A}}$ of conditional probabilities. Unsurprisingly, $\mu(\cdot|\mathcal{G})(\omega)$ will not be a measure for all $\omega \in \Omega$ unless the collection $\{\mathbf{P}(A|\mathcal{G})\}_{A \in \mathcal{A}}$ is 'consistent' in some way. Happily, it turns out that it is always possible to choose $\{\mathbf{P}(A|\mathcal{G})\}_{A \in \mathcal{A}}$ 'consistently' in this manner.

[26] Aside: in game theory, off-equilibrium beliefs (beliefs following events that have probability zero on the equilibrium path) are important for constructing equilibria in dynamic games, since 'fearful' off-equilibrium beliefs can 'deter' players from deviating from equilibrium play. Most equilibrium concepts require that players update their beliefs in accordance with conditional probability. But since conditional probabilities are indeterminate, lots of belief-updating protocols are allowed, leading to equilibrium multiplicity ('fearful' equilibria supported by 'fearful' off-equilibrium beliefs). Much of the refinements literature is concerned with imposing additional, intuitive restrictions on how beliefs are revised in order to kill off implausible equilibria of this sort.

[27] That is, $\mu(\cdot|\mathcal{G})(\cdot) : \mathcal{A} \times \Omega \to \mathbf{R}$ is a function defined by $\mu(A|\mathcal{G})(\omega) \coloneqq \mathbf{P}(A|\mathcal{G})(\omega)$ for each $A \in \mathcal{A}$ and $\omega \in \Omega$.

That is, there exists a random function $\mu(\cdot|\mathcal{G})$ s.t. every $\mu(\cdot|\mathcal{G})(\omega)$ is a probability measure and every $\mu(A|\mathcal{G})(\cdot)$ is a conditional probability of A on $\mathcal{G}$ (Billingsley, 1995, Theorem 33.3). Such a μ is called a regular conditional probability.

Conditional expectation is defined in a similar way.

Definition 16. Consider a probability space $(\Omega, \mathcal{A}, \mathbf{P})$, a sub-$\sigma$-algebra $\mathcal{G}$ of $\mathcal{A}$, and a $\mathbf{P}$-integrable random variable Y.[28] A random variable $\mathbf{E}(Y|\mathcal{G}) : \Omega \to \mathbf{R}$ is a conditional expectation of Y on $\mathcal{G}$ iff it is (1) $\mathcal{G}$-measurable, (2) $\mathbf{P}$-integrable, and (3) satisfies

$$\int_G Y \mathrm{d}\mathbf{P} = \int_G \mathbf{E}(Y|\mathcal{G}) \mathrm{d}\mathbf{P} \quad \forall G \in \mathcal{G}.$$

The proof of existence is similar to the one for conditional probability; see Rosenthal (2006, Proposition 13.1.7). The interpretational subtleties outlined above also apply to conditional expectation. Also, fun fact: for $G = \Omega$ we get

$$\int_\Omega Y \mathrm{d}\mathbf{P} = \int_\Omega \mathbf{E}(Y|\mathcal{G}) \mathrm{d}\mathbf{P},$$

meaning that the 'law of iterated expectation' is actually part of the definition of conditional expectation.

2.8 Inequalities

Probability theory is full of inequalities. The main ones are concentration inequalities, which bound the probability that a random variable deviates away from some value (usually zero or its mean). These are often useful for establishing the convergence of sequences of random variables, the topic of the next section. The two most basic concentration inequalities are Markov's and Chebychev's; both apply to random variables, but extend easily to random vectors.

Proposition 2 (Markov's inequality)**.** Let X be a nonnegative random variable on $(\Omega, \mathcal{A}, \mathbf{P})$. Then $\mathbf{P}(X \geq \varepsilon) \leq \mathbf{E}(X)/\varepsilon \ \forall \varepsilon > 0$.

Proof. Fix $\varepsilon > 0$.

$$\begin{aligned} \mathbf{E}(X) = \int_\Omega X \mathrm{d}\mathbf{P} &= \int_{\{X \geq \varepsilon\}} X \mathrm{d}\mathbf{P} + \int_{\{X < \varepsilon\}} X \mathrm{d}\mathbf{P} \\ &\geq \int_{\{X \geq \varepsilon\}} X \mathrm{d}\mathbf{P} \geq \varepsilon \int_{\{X \geq \varepsilon\}} \mathrm{d}\mathbf{P} = \varepsilon \mathbf{P}(X \geq \varepsilon). \end{aligned}$$ ■

[28] Recall that Y is $\mathbf{P}$-integrable iff $\mathbf{E}(Y)$ exists.

Corollary 1 (Chebychev's inequality)**.** Let X be a random variable on $(\Omega, \mathcal{A}, \mathbf{P})$ such that $\mathbf{E}(X)$ exists and is finite. Then $\mathbf{P}(|X - \mathbf{E}(X)| \geq \varepsilon) \leq \mathrm{Var}(X)/\varepsilon^2$ for every $\varepsilon > 0$.

Proof. Fix $\varepsilon > 0$ and define $Y \coloneqq (X - \mathbf{E}(X))^2$. Y is nonnegative, so by Markov's inequality

$$\mathbf{P}(|X - \mathbf{E}(X)| \geq \varepsilon) = \mathbf{P}\left(Y \geq \varepsilon^2\right) \leq \mathbf{E}(Y)/\varepsilon^2 = \mathrm{Var}(X)/\varepsilon^2. \qquad \blacksquare$$

As an illustration, consider a standard-normal-distributed random variable X. Chebychev's inequality gives us $\mathbf{P}(|X| \geq 1.96) \leq 1/1.96^2 \simeq 0.26$. But we know that $\mathbf{P}(|X| \geq 1.96) \simeq 0.05$, so the Chebychev bound is not very tight in this case. This is typical, and perhaps not very surprising since the bound applies to all probability distributions, even really badly-behaved ones.

Another useful corollary to Markov's inequality is the following.

Proposition 3 (Chernoff bounds)**.** Let X be a random variable on $(\Omega, \mathcal{A}, \mathbf{P})$. Then for every $\varepsilon > 0$,

$$\mathbf{P}(X \geq \varepsilon) \leq \frac{\mathbf{E}\left(\exp(tX)\right)}{\exp(t\varepsilon)} \quad \text{for every } t > 0, \text{ and}$$
$$\mathbf{P}(X \leq \varepsilon) \leq \frac{\mathbf{E}\left(\exp(tX)\right)}{\exp(t\varepsilon)} \quad \text{for every } t < 0.$$

You need to do a bit of work to obtain a useful Chernoff bound. The quantity $f_X(t) \coloneqq \mathbf{E}(\exp(tX))$ is called the moment-generating function (MGF) of X (it is a cousin of the characteristic function introduced in section 3.7). The MGFs of all commonly-used distributions can be looked up on Wikipedia, so when the distribution is known we can compute a Chernoff bound for various values of t. We can also make use of generic properties of MGFs: for example, we can use the fact that

$$f_{\sum_{i=1}^n X_i}(t) = \prod_{i=1}^n f_{X_i}$$

for $\{X_i\}$ independent to derive a Chernoff bound for sums of independent random variables. The fact that the generic bound holds for any t is helpful, since it allows us to pick a t for which the bound is tractable (and tight, hopefully).

There are many refinements of Markov's and Chebychev's inequalities that give tighter bounds than this under additional assumptions. The next two are easy concentration inequalities similar to Markov's above.

Proposition 4 (generalised Markov inequality)**.** Let X be a nonnegative random variable on $(\Omega, \mathcal{A}, \mathbf{P})$, and let $g : \mathbf{R}_+ \to \mathbf{R}_+$ be strictly increasing. Then $\mathbf{P}(X \geq \varepsilon) \leq \mathbf{E}(g(X))/g(\varepsilon)\ \forall \varepsilon > 0$.

Proposition 5 (Cantelli's inequality)**.** Let X be a random variable on $(\Omega, \mathcal{A}, \mathbf{P})$ such that $\mathbf{E}(X)$ exists and is finite. Then

$$\mathbf{P}(X - \mathbf{E}(X) \geq \lambda) \begin{cases} \leq \frac{\mathrm{Var}(X)}{\mathrm{Var}(X)+\lambda^2} & \text{for } \lambda > 0 \\ \geq \frac{\lambda^2}{\mathrm{Var}(X)+\lambda^2} & \text{for } \lambda < 0. \end{cases}$$

The next two concentration inequalities are terribly ugly, but very useful. The former (Kolmogorov's) is a special case of the latter (Hájek–Rényi). We will use the Hájek–Rényi inequality to prove Kolmogorov's first SLLN in section 4.1 (p. 53).

Theorem 2 (Kolmogorov's inequality)**.** Let $\{X_n\}$ be a sequence of mean-zero independent random variables on $(\Omega, \mathcal{A}, \mathbf{P})$. Then for any $m < n$ and $\varepsilon > 0$,

$$\mathbf{P}\left(\max_{k\in[m,n]} \left|\sum_{i=1}^{k} X_i\right| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2} \sum_{i=1}^{n} \mathrm{Var}(X_i).$$

Theorem 3 (Hájek–Rényi inequality)**.** Let $\{X_n\}$ be a sequence of mean-zero independent random variables on $(\Omega, \mathcal{A}, \mathbf{P})$, and let $\{c_n\}$ be a weakly decreasing sequence in $\mathbf{R}_+$. Then for any $m < n$ in $\mathbf{N}$ and any $\varepsilon > 0$,

$$\mathbf{P}\left(\max_{k\in[m,n]} c_k \left|\sum_{i=1}^{k} X_i\right| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2} \left(c_m^2 \sum_{i=1}^{m} \mathrm{Var}(X_i) + \sum_{i=m+1}^{n} c_i^2 \mathrm{Var}(X_i)\right).$$

Next up are two inequalities from the theory of $\mathcal{L}^p$ spaces, a branch of functional analysis (i.e. these are not concentration inequalities).[29] The first is the triangle inequality for $\mathcal{L}^p$ spaces; the second is a generalisation of the Cauchy–Schwarz inequality.

Theorem 4 (Minkowski's inequality)**.** Let X and Y be random variables on $(\Omega, \mathcal{A}, \mathbf{P})$. Then for any $p \geq 1$,

$$\mathbf{E}(|X + Y|^p)^{1/p} \leq \mathbf{E}(|X|^p)^{1/p} + \mathbf{E}(|Y|^p)^{1/p}$$

provided the moments exist.

[29] (It's not important, but) an $\mathcal{L}^p$ space is a set of measurable functions on some measure space $(\Omega, \mathcal{A})$ whose pth power is integrable w.r.t. some measure μ on $(\Omega, \mathcal{A})$. $\mathcal{L}^p$ spaces turn out to be normed vector spaces (you can easily verify this using Minkowski's inequality).

Theorem 5 (Hölder's inequality)**.** Let X and Y be random variables on $(\Omega, \mathcal{A}, \mathbf{P})$. Then for any $p, q \geq 1$ with $p^{-1} + q^{-1} \leq 1$,

$$\mathbf{E}(|XY|) \leq \mathbf{E}\left(|X|^p\right)^{1/p} \mathbf{E}\left(|Y|^q\right)^{1/q}$$

provided the moments exist.

Corollary 2 (Cauchy–Schwarz inequality)**.** Let X and Y be random variables on $(\Omega, \mathcal{A}, \mathbf{P})$. Then

$$\mathbf{E}(|XY|) \leq \mathbf{E}\big(|X|^2\big)^{1/2} \mathbf{E}\big(|Y|^2\big)^{1/2}$$

provided the moments exist.

We'll finish with Jensen's inequality, which is probably familiar.

Theorem 6 (Jensen's inequality)**.** Let X be a random variable on $(\Omega, \mathcal{A}, \mathbf{P})$, and let $g : \mathbf{R} \to \mathbf{R}$ be convex. Then $g\left(\mathbf{E}(X)\right) \leq \mathbf{E}\left(g(X)\right)$ provided the moments exist, with equality iff either g is linear or $\mathcal{L}_X$ is a point mass.

3 Modes of convergence

Official reading: Amemiya (1985, ch. 3), Rao (1973, ch. 2) and Serfling (1980, ch. 1).

Since this section concerns convergence, it will be important that all topologically interesting sets are measurable. So unless otherwise specified, every set will be equipped with (a superset of) its Borel σ-algebra.

3.1 Convergence of random sequences

In econometric theory, we use an estimator $\widehat{\theta}_n$ to estimate some unknown parameter θ. $\widehat{\theta}_n$ is a function of the data (which has sample size n), and so is random. As the sample size grows large, we would like the sequence of estimators $\{\widehat{\theta}_n\}$ to get close to θ in some well-defined sense. This is one of the most basic adequacy criteria for estimators, called consistency. To study consistency, we have to define what it means for a sequence of random variables to converge to a point, or more generally to a random variable.

As a starting point, let's review what 'convergence' means in analysis. We'll restrict attention to convergence in metric spaces (rather than general topological spaces) to make our lives easier.

Definition 17. Let $\{x_n\}$ be a sequence in a metric space (S, ρ). $\{x_n\}$ converges to $x_0 \in S$ iff for any $\varepsilon > 0$, there exists $N_\varepsilon \in \mathbf{N}$ such that $\rho(x_n, x_0) < \varepsilon$ whenever $n \geq N_\varepsilon$. Convergence is typeset as $x_n \longrightarrow x_0$ or $\lim_{n\to\infty} x_n = x_0$.

This notion of convergence can easily be applied to random elements. Let S be the set of all random elements of some measurable space $(T, \mathcal{T})$ defined on $(\Omega, \mathcal{A}, \mathbf{P})$, and let ρ be some metric on S. Let $\{X_n\}$ and X be elements of S. Then the ordinary notion of convergence is perfectly intelligible. For example, if the outcome space T is equipped with a metric d, we can use the sup metric

$$\rho(X, Y) \coloneqq \sup_{\omega\in\Omega} d\left(X(\omega), Y(\omega)\right)$$

on S, in which case $X_n \longrightarrow X$ means uniform convergence of the functions $\{X_n\}$ to the function X.

This concept is so strong, however, that many useful convergence results (to be proved later) are unavailable if we use it. Weaker concepts are therefore called for. One approach would be to try to find a cleverer choice of metric ρ on S, but we will avoid this route because it is less intuitive and because some of our convergence concepts are not metrisable in this manner anyway.

We will instead proceed in an ad-hoc way, dreaming up new convergence concepts with intuition as our guide.

The most obvious way of weakening the ordinary convergence is almost sure convergence. Ordinary convergence in the sup metric required uniform convergence of the measurable functions $\{X_n\}$. We could weaken this to requiring only pointwise convergence of $\{X_n\}$: $X_n(\omega) \longrightarrow X(\omega)$ for all $\omega \in \Omega$ (we might call this 'sure convergence').[30] Almost sure convergence weakens this one step further by requiring that $X_n(\omega) \longrightarrow X(\omega)$ for *almost* all $\omega \in \Omega$, i.e. for all ω outside a set of measure zero. Formally:

Definition 18. Let $\{X_n\}$ and X be random elements of a metric space (S, ρ) defined on $(\Omega, \mathcal{A}, \mathbf{P})$. $\{X_n\}$ converges almost surely (a.s.) to X iff

$$\lim_{n\to\infty} X_n = X \quad \mathbf{P}\text{-a.s.}^{31}$$

Almost sure convergence is typeset $X_n \xrightarrow{\text{a.s.}} X$ or $X_n \longrightarrow X$ a.s., and is also known as convergence with probability 1 (w.p. 1) or (in general measure theory) convergence almost everywhere (a.e.).

This definition of a.s. convergence is intuitive, but it is often difficult to work with. The following lemma gives a more tractable condition. I'll omit the proof becaue it requires machinery I haven't introduced (tail events, continuity of measures).

Lemma 1. Let $\{X_n\}$ and X be random elements of a metric space (S, ρ) defined on $(\Omega, \mathcal{A}, \mathbf{P})$. Then $X_n \xrightarrow{\text{a.s.}} X$ iff

$$\lim_{N\to\infty} \mathbf{P}\left(\sup_{n\geq N} \rho(X_n, X) > \varepsilon\right) = 0.$$

Remark 1. Often, we're interested in the notion that a sequence of random elements converges to a point, not to a random element. To fit this idea into the definition above, simply take the limiting random element X to be a constant function. We'll use the shorthand $X_n \xrightarrow{\text{a.s.}} \alpha$ for a point $\alpha \in S$ to denote a.s. convergence of $\{X_n\}$ to a constant function everywhere equal to α. (Similarly, we will write $X_n \xrightarrow{\text{p}} \alpha$ and $X_n \xrightarrow{\text{m.s.}} \alpha$ for the other two convergence concepts in this section.)

[30] This corresponds to convergence in the product topology on S (a.k.a. the topology of pointwise convergence). This topology is not metrisable!

[31] More explicitly, $\mathbf{P}\left(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}\right) = 1$. The lim operator here is the ordinary one from analysis, since $\{X_n(\omega)\}$ is a sequence in a metric space.

A.s. convergence turns out to be stronger than necessary for most of econometric theory. To weaken it, look at Lemma 1: if we drop the sup operator, we obviously get a less stringent condition. This weaker condition is called convergence in probability.

Definition 19. Let $\{X_n\}$ and X be random elements of a metric space (S, ρ) defined on $(\Omega, \mathcal{A}, \mathbf{P})$. $\{X_n\}$ converges in probability to X iff for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbf{P}(\rho(X_n, X) > \varepsilon) = 0.$$

Convergence in probability is typeset $X_n \xrightarrow{\text{p}} X$ or $\operatorname{plim}_{n\to\infty} X_n = X$, and is also known as convergence with probability approaching 1.

Finally, we introduce a third convergence concept. Like a.s. convergence, it is stronger than convergence in probability, though it neither implies nor is implied by a.s. convergence. Let $\|\cdot\|_2$ denote the Euclidean norm.

Definition 20. Let $\{X_n\}$ and X be random vectors defined on $(\Omega, \mathcal{A}, \mathbf{P})$. $\{X_n\}$ converges in mean square (m.s.) to X iff $\mathbf{E}((\|X_n - X\|_2)^2)$ exists for each $n \in \mathbf{N}$ and

$$\lim_{n\to\infty} \mathbf{E}\left((\|X_n - X\|_2)^2\right) = 0.$$

Convergence in mean square is typeset $X_n \xrightarrow{\text{m.s.}} X$.

We will not use this concept very much because we don't want to assume that the second moment of an estimator exists. But convergence in mean square turns out to be useful for studying the convergence of random functions.

It's clear that convergence in mean square can at most be extended to random elements of normed vector spaces (such as $\mathbf{R}^n$). It cannot be defined for random elements of general metric spaces.

Aside on metrisability. Let's return to the question of whether we can cleverly choose a metric on the set of random elements such that our new convergence concepts are equivalent to convergence in the ordinary sense. Fix a probability space $(\Omega, \mathcal{A}, \mathbf{P})$ and a metric space (T, d), let S be the set of random elements of (T, d) defined on this space, and let $\{X_n\}$ and X lie in S. Endow S with a metric ρ, and let $\to_\rho$ denote ordinary convergence in the metric space (S, ρ).

It should be obvious that $X_n \xrightarrow{\text{m.s.}} X$ is equivalent to $X_n \to_\rho X$ in the metric

$$\rho(X, Y) \coloneqq \mathbf{E}((\|X - Y\|_2)^2).$$

It turns out that there is also a metric ρ such that $X_n \xrightarrow{\text{p}} X$ iff $X_n \to_\rho X$.[32] On the other hand, unless $(\Omega, \mathcal{A})$ is trivial, there is no metric ρ on S such that $X_n \xrightarrow{\text{a.s.}} X$ iff $X_n \to_\rho X$ (see Dudley (2004, p. 289)).

3.2 Convergence of random functions

In analysis, we make a distinction between pointwise and uniform convergence of functions. As in the previous section, we'll make our lives easier by restricting attention to metrisable topologies.

Definition 21. Let (S, ρ) and (S', ρ') be metric spaces, and let $\{f_n\}$ and f be functions $S \to S'$.

(1) $f_n \longrightarrow f$ pointwise iff $\rho'(f_n(x), f(x)) \longrightarrow 0$ for every $x \in S$.

(2) $f_n \longrightarrow f$ uniformly iff $\sup_{x \in S} \rho'(f_n(x), f(x)) \longrightarrow 0$.

I've stated both in terms of convergence (in $\mathbf{R}$) of the distance ρ' because uniform convergence is most naturally expressed in that way, but clearly $f_n \longrightarrow f$ pointwise is equivalent to $f_n(x) \longrightarrow f(x)$ for every $x \in S$.

Uniform convergence obviously implies pointwise convergence. What uniform convergence adds to pointwise convergence is that there must be some rate of convergence that applies to every point. In particular, pointwise convergence says that for each $\varepsilon > 0$ and each $x \in S$, there is $N_{\varepsilon,x}$ such that $\rho'(f_n(x), f(x)) < \varepsilon$ for all $n \geq N_{\varepsilon,x}$; uniform convergence adds the requirement that the set $\{N_{\varepsilon,x}\}_{x \in S}$ has a finite upper bound $\overline{N}_\varepsilon$.

This should make it sound like uniform convergence is quite a lot stronger than pointwise convergence. Indeed, many pointwise convergent sequences of well-behaved functions fail to converge uniformly, as the following example shows.

Example 3. Let $S = [0, 2]$ and let $f_n : S \to \mathbf{R}$ be given by

$$f_n(x) = \begin{cases} nx & \text{for } x \in [0, 1/n) \\ 2 - nx & \text{for } x \in [1/n, 2/n) \\ 0 & \text{for } x \in [2/n, 2]. \end{cases}$$

A few functions in the sequence are drawn in Figure 1. Define $f : S \to \mathbf{R}$ by $f(x) \coloneqq \mathbf{1}(x = 0)$. Obviously $f_n \longrightarrow f$ pointwise. But $\{f_n\}$ does not converge to f uniformly: $\sup_{x \in [0,2]} |f_n(x) - f(x)| = 1$ no matter how large n gets.

[32] For example, $\rho(X, Y) \coloneqq \mathbf{E}\left(|X - Y| / [1 + |X - Y|]\right)$ does the trick. We were asked to prove this in question 3 on problem set 2; a proof can also be found in Dudley (2004, Theorem 9.2.2).

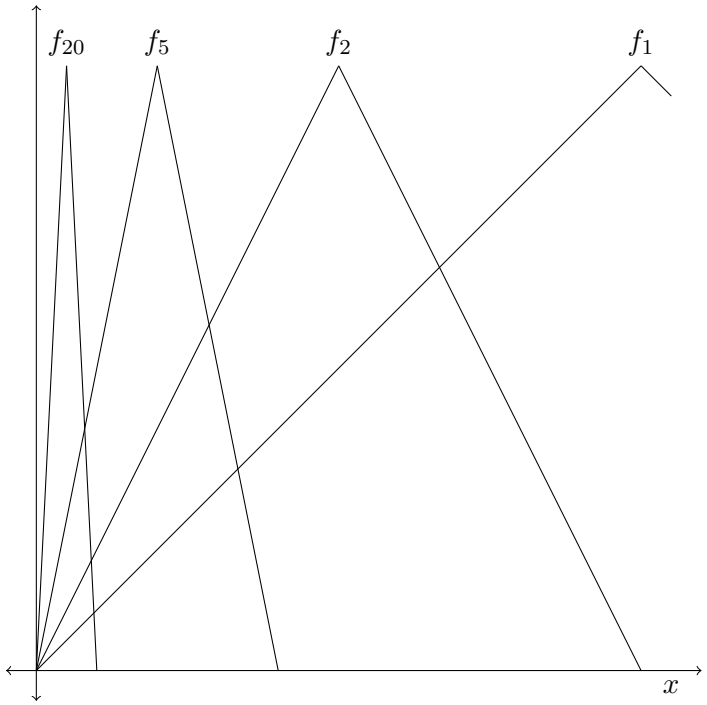

Figure 1 – Some elements of the sequence $\{f_n\}$ from Example 3.

We're defining these convergence concepts using the metric on the image space S'. It is perhaps more natural to define convergence of functions by endowing the functional space F with a topology such that the appropriate convergence concept coincides with convergence in that topology. Convergence in the product topology (a.k.a. the topology of pointwise convergence) on F is equivalent to pointwise convergence as defined above. Convergence in the topology on F induced by the sup metric

$$d(f,g) \coloneqq \sup_{x \in S'} \rho'(f(x), g(x))$$

is equivalent to uniform convergence.

The task in this section is to extend uniform and pointwise convergence to random functions. A random function is simply a random element that takes values in a functional space. As usual, we are using the Borel σ-algebra corresponding to whatever topology we've endowed this functional space with. (There are many topologies we might want to endow a functional space with. We've already seen two, the product topology and the topology induced by the sup metric.)

Here's the (entirely straightforward) extension. We won't bother with convergence in m.s.

Definition 22. Let (S, ρ) and (S', ρ') be metric spaces, and let $\{f_n\}$ and f be random functions $S \to S'$.

(1) $f_n \xrightarrow{\text{a.s.}} f$ pointwise iff $\rho'(f_n(x), f(x)) \xrightarrow{\text{a.s.}} 0$ for every $x \in S$.

(2) $f_n \xrightarrow{\text{a.s.}} f$ uniformly iff $\sup_{x \in S} \rho'(f_n(x), f(x)) \xrightarrow{\text{a.s.}} 0$.

(3) $f_n \xrightarrow{\text{p}} f$ pointwise iff $\rho'(f_n(x), f(x)) \xrightarrow{\text{p}} 0$ for every $x \in S$.

(4) $f_n \xrightarrow{\text{p}} f$ uniformly iff $\sup_{x \in S} \rho'(f_n(x), f(x)) \xrightarrow{\text{p}} 0$.

Remark 2. Obvious equivalences: $f_n \xrightarrow{\text{a.s.}} f$ pointwise iff $f_n(x) \xrightarrow{\text{a.s.}} f(x)$ for every $x \in \mathbf{R}$, and $f_n \xrightarrow{\text{p}} f$ pointwise iff $f_n(x) \xrightarrow{\text{p}} f(x)$ for every $x \in \mathbf{R}$.

Some final (dull) remarks on terminology. Sometimes we're interested in convergence a.s./in probability pointwise/uniformly on some subset $T \subseteq S$; in this case, simply replace 'for every $x \in S$ with 'for every $x \in T$' and '$\sup_{x \in S}$' with '$\sup_{x \in T}$' in the definitions above. When we're holding an argument fixed, as in '$f_n(\cdot, y) \xrightarrow{\text{a.s.}} f(\cdot, y)$ uniformly on $T \subseteq S$', we sometimes say '$f_n(x, y) \xrightarrow{\text{a.s.}} f(x, y)$ uniformly in $x \in T$' instead; similarly for uniform convergence in probability.

3.3 Convergence of measures

All three of the convergence concepts we've given have a similar flavour: they require the random elements X_n to get close to X as n increases. But we might also care about the *distributions* of $\{X_n\}$ getting close to the distribution of X. For example, suppose $X_n \sim \mathcal{N}(0,1)$ and $X \sim \mathcal{N}(0,1)$, all independent.[33] No matter how large n gets, $\mathbf{P}(X_n \neq X) = 1$. Nevertheless, it seems that this is a (trivial) case in which the distributions of $\{X_n\}$ converge to the distribution of X.

For a probability space $(\Omega, \mathcal{A}, \mathbf{P})$ with Ω is endowed with a topology, call $A \in \mathcal{A}$ a $\mathbf{P}$-continuity set iff $\mathbf{P}(\partial A) = 0$.[34]

Definition 23. Let $\{X_n\}$ and X be random elements of a metric space (S, ρ) defined on $(\Omega, \mathcal{A}, \mathbf{P})$. $\{X_n\}$ converges in distribution to X iff $\mathcal{L}_{X_n}(A) \longrightarrow \mathcal{L}_X(A)$ for every $\mathcal{L}_X$-continuity set A. Convergence in distribution is typeset $X_n \xrightarrow{\mathrm{d}} X$ or $X_n \rightsquigarrow X$.

In the case of random vectors, this obviously reduces to the (perhaps more familiar) definition that $F_{X_n} \longrightarrow F_X$ (pointwise) at every continuity point of F_X. The following example illustrates why we do not require convergence at discontinuity points of F_X.

Example 4. Let $\{X_n\}$ be a sequence of logistically distributed random variables. In particular, let them have CDFs

$$F_{X_n}(x) = \left(1 + \exp\left(-x/\theta_n\right)\right)^{-1} \quad \forall x \in \mathbf{R}$$

where the sequence $\{\theta_n\}$ satsifies $\theta_n \longrightarrow 0$. The sequence of functions $\{F_{X_n}\}$ converges pointwise to

$$G(x) \coloneqq \begin{cases} 0 & x < 0 \\ 0.5 & x = 0 \\ 1 & x > 0. \end{cases}$$

G is not a CDF since it isn't right-continuous. But the function F given by $F(x) \coloneqq G(x)$ at $x \neq 0$ and $F(0) \coloneqq 1$ is a CDF, corresponding to a point mass at 0. This is intuitively what the sequence $\{X_n\}$ should converge

[33] $\mathcal{N}\left(\mu, \sigma^2\right)$ denotes the normal distribution; in particular I sometimes use it to mean a normally distributed random variable, sometimes a normal law on $(\mathbf{R}, \mathcal{B})$. '$\sim$' reads 'is distributed as'.

[34] ∂A denotes the boundary of A. It is measurable since it's closed and we're using (a superset of) the Borel σ-algebra.

in distribution to. And since convergence in distribution does not require convergence of the CDFs at discontinuity points, we have $X_n \xrightarrow{\mathrm{d}} X$ where X is any random variable with this CDF.

We say that a sequence of random variables converges in distribution, but it is really more natural to think of the sequence of laws $\{\mathcal{L}_{X_n}\}$ as converging. In this language, convergence in distribution is called weak convergence.

Definition 24. Let $\{\mu_n\}$ and μ be measures defined on a measurable space $(\Omega, \mathcal{A})$, and equip Ω with a topology. $\{\mu_n\}$ converges weakly to μ iff $\mu_n(A) \longrightarrow \mu(A)$ for every μ-continuity set A. Weak convergence is typeset $\mu_n \Rightarrow \mu$. (Occasionally, I may sloppily say that the random variables $\{X_n\}$ converge weakly to X.)

This quite intuitive notion of convergence turns out to be equivalent to several other tractable (but less intuitive) properties. The equivalence is given by the Portmanteau lemma, a small part of which is stated below. As it happens, property (2) below is conventionally taken as the definition of weak convergence.

Lemma 2 (partial Portmanteau lemma)**.** Let $\{\mu_n\}$ and μ be measures on $(\Omega, \mathcal{A})$. The following are equivalent.

(1) $\mu_n(A) \longrightarrow \mu(A)$ for every μ-continuity set A.

(2) $\int_\Omega f \mathrm{d}\mu_n \longrightarrow \int_\Omega f \mathrm{d}\mu$ for every continuous and bounded $f : \Omega \to \mathbf{R}$.

Weak convergence is equivalent to convergence in the weak* topology on the set of probability measures on $(\Omega, \mathcal{A})$. (This is immediate from the definition of the weak* topology!) Moreover, this topology is metrisable (by the Prohorov metric; see Billingsley (1999, pp. 72–3)), so weak convergence corresponds to ordinary convergence in a certain metric on the space of probability measures.

Our interest in weak convergence is motivated by central limit theorems, which concern weak convergence of the laws of normalised sums of random vectors to a normal law. It turns out that for the special case of random vectors, the theory of weak convergence can be studied using the characteristic transform introduced in section 3.7 below. We therefore won't delve any deeper into the theory of weak convergence in general metric spaces here. (For the curious, Billingsley (1999) is a standard book.)

3.4 Relationships between modes of convergence

In this section, we will establish the implication relationships between the modes of convergence we're considering. In particular, we will show that

$$\begin{pmatrix} X_n \xrightarrow{\text{a.s.}} X \\ X_n \xrightarrow{\text{m.s.}} X \end{pmatrix} \Rightarrow \left(X_n \xrightarrow{\text{p}} X \right) \Rightarrow \left(X_n \xrightarrow{\text{d}} X \right).$$

We'll also show that for a constant α, $(X_n \xrightarrow{\text{d}} \alpha) \Rightarrow (X_n \xrightarrow{\text{p}} \alpha)$.

We begin with the first two implications: that a.s. convergence and convergence in m.s. imply convergence in probability. Both have nice, short proofs.

Proposition 6. If $X_n \xrightarrow{\text{a.s.}} X$, then $X_n \xrightarrow{\text{p}} X$.

Proof. Let $X_n \xrightarrow{\text{a.s.}} X$. Fix an $\varepsilon > 0$. Obviously $\rho(X_N, X) > \varepsilon$ implies that $\sup_{n \geq N} \rho(X_n, X) > \varepsilon$. Together with nonnegativity, this yields

$$0 \leq \mathbf{P}\left(\rho(X_N, X) > \varepsilon\right) \leq \mathbf{P}\left(\sup_{n \geq N} \rho(X_n, X) > \varepsilon \right).$$

Since the RHS converges to 0 as $N \to \infty$ by $X_n \xrightarrow{\text{a.s.}} X$, it follows that $\lim_{N \to \infty} \mathbf{P}\left(\rho(X_N, X) > \varepsilon\right) = 0$. Since $\varepsilon > 0$ was arbitrary, $X_n \xrightarrow{\text{p}} X$. ■

Proposition 7. If $X_n \xrightarrow{\text{m.s.}} X$, then $X_n \xrightarrow{\text{p}} X$.

Proof. Let $X_n \xrightarrow{\text{m.s.}} X$. $(\|X_n - X\|_2)^2$ is a nonnegative random variable, so Markov's inequality (p. 22) applies. Together with the fact that probabilities are nonnegative, we have for any $\varepsilon > 0$ that

$$\begin{aligned} 0 \leq \mathbf{P}\left(\|X_n - X\|_2 > \varepsilon\right) &= \mathbf{P}\left((\|X_n - X\|_2)^2 > \varepsilon^2 \right) \\ &\leq \varepsilon^{-2} \mathbf{E}\left((\|X_n - X\|_2)^2 \right). \end{aligned}$$

The RHS converges to 0 since $X_n \xrightarrow{\text{m.s.}} X$. Hence $\mathbf{P}(\|X_n - X\|_2 > \varepsilon) \to 0$ for every $\varepsilon > 0$, i.e. $X_n \xrightarrow{\text{p}} X$. ■

A natural question you might now ask is: convergence in probability plus what property is equivalent to convergence in mean square? The answer is a boundedness property called uniform integrability; see e.g. Williams (1991, sec. 13.7).

To show that a.s. convergence and convergence in m.s. do not imply each other, we give counterexamples.

Example 5 ($\xrightarrow{\text{m.s.}}$ without $\xrightarrow{\text{a.s.}}$)**.** Let $\{X_n\}$ be independent with

$$\mathbf{P}(X_n = 0) = 1 - \tfrac{1}{n} \quad \text{and} \quad \mathbf{P}(X_n = 1) = \tfrac{1}{n} \quad \text{for each } n \in \mathbf{N}.$$

Then $\mathbf{E}\left(X_n^2\right) = 1/n \longrightarrow 0$ as $n \longrightarrow \infty$, so $X_n \xrightarrow{\text{m.s.}} 0$.

It's obvious (but we didn't prove) that if $\{X_n\}$ is a.s.-convergent then the limit must be 0. A.s. convergence to 0 would require that

$$\lim_{N\to\infty} \mathbf{P}\left(\sup_{n\geq N} |X_n| < \varepsilon\right) = 1 \quad \text{for every } \varepsilon > 0.$$

So choose $\varepsilon \in (0,1)$. Then $|X_n| < \varepsilon$ iff $X_n = 0$, so for any $N \in \mathbf{N}$ we have

$$\mathbf{P}\left(\sup_{n\geq N} |X_n| < \varepsilon\right) = \mathbf{P}(X_n = 0 \ \ \forall n \geq N) = \prod_{n=N}^{\infty} \left(1 - \tfrac{1}{n}\right).$$

where the final equality used independence. Taking logs and using the inequality $\ln\left(1 - \frac{1}{n}\right) \leq -\frac{1}{n}$,[35]

$$\ln\left(\mathbf{P}\left(\sup_{n\geq N} |X_n| < \varepsilon\right)\right) = \sum_{n=N}^{\infty} \ln\left(1 - \tfrac{1}{n}\right) \leq -\sum_{n=N}^{\infty} n^{-1} = -\infty$$

since the harmonic series $\sum_{n=N}^{\infty} n^{-1}$ diverges for every N. So by continuity of $\ln(\cdot)$, $\mathbf{P}\left(\sup_{n\geq N} |X_n| < \varepsilon\right) = 0$ for every N, hence

$$\lim_{N\to\infty} \mathbf{P}\left(\sup_{n\geq N} |X_n| < \varepsilon\right) = 0 \neq 1.$$

Example 6 ($\xrightarrow{\text{a.s.}}$ without $\xrightarrow{\text{m.s.}}$)**.** Let $\{X_n\}$ be independent with

$$\mathbf{P}(X_n = 0) = 1 - \tfrac{1}{n^2} \quad \text{and} \quad \mathbf{P}(X_n = n) = \tfrac{1}{n^2} \quad \text{for each } n \in \mathbf{N}.$$

$\mathbf{E}\left(X_n^2\right) = n^2/n^2 = 1$ for every $n \in \mathbf{N}$, so we don't have convergence to 0 in mean square. It's fairly obvious (but we didn't prove) that we cannot have convergence in m.s. to anything other than 0.

Again, a.s. convergence to 0 requires that

$$\lim_{N\to\infty} \mathbf{P}\left(\sup_{n\geq N} |X_n| < \varepsilon\right) = 1 \quad \text{for every } \varepsilon > 0.$$

[35] Since ln is concave, it must lie below all its tangents, i.e. for any $x, x' > 0$, $\ln(x) \leq \ln(x') + (x')^{-1}(x - x')$. Setting $x = 1 - \frac{1}{n}$ and $x' = 1$ yields $\ln\left(1 - \frac{1}{n}\right) \leq -\frac{1}{n}$.

Following the steps in the previous example, for small $\varepsilon > 0$ and any $N \in \mathbf{N}$ we have

$$\ln\left(\mathbf{P}\left(\sup_{n\geq N}|X_n| < \varepsilon\right)\right) = \sum_{n=N}^{\infty} \ln\left(1 - \tfrac{1}{n^2}\right).$$

Using the inequality $\ln\left(1 - \frac{1}{n^2}\right) \geq -\frac{1}{n^2} - \frac{1}{2n^4}$,[36] we obtain

$$\ln\left(\mathbf{P}\left(\sup_{n\geq N}|X_n| < \varepsilon\right)\right) \geq -\sum_{n=N}^{\infty}\left(\frac{1}{n^2} + \frac{1}{2n^4}\right).$$

The p-series $\sum_{n=N}^{\infty} n^{-p}$ is convergent iff $p > 1$ (regardless of N), so the RHS is finite for each N and converges to 0 as $N \to \infty$. So by continuity of $\ln(\cdot)$ we obtain

$$\lim_{N\to\infty} \mathbf{P}\left(\sup_{n\geq N}|X_n| < \varepsilon\right) = 1.$$

Now for another proposition: convergence in probability implies convergence in distribution.

Proposition 8. If $X_n \xrightarrow{\text{p}} X$, then $X_n \xrightarrow{\text{d}} X$.

The statement is true for general random elements, but our proof restricts attention to random vectors in order to make use of the simpler CDF-based definition of convergence in distribution.

Proof for random vectors. Let $X_n \xrightarrow{\text{p}} X$. Define $Z_n \coloneqq X - X_n$, so that $Z_n \xrightarrow{\text{p}} 0$. Fix some $\varepsilon > 0$ and a continuity point $t \in \mathbf{R}$ of F_X. We must show that $F_{X_n}(t) \longrightarrow F_X(t)$.

Using the fact that $A \subseteq B$ implies $\mathbf{P}(A) \leq \mathbf{P}(B)$ and a few other basic

[36] This inequality follows from the fact that in the Taylor series

$$\ln(1+x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n = x - \frac{x^2}{2} + \sum_{n=3}^{\infty} \frac{(-1)^{n+1}}{n} x^n,$$

the remainder $\sum_{n=3}^{\infty} \frac{(-1)^{n+1}}{n} x^n$ can be shown to be nonnegative. Now set $x = -1/n^2$.

facts about probabilities,

$$\begin{aligned}
F_{X_n}(t) &= \mathbf{P}\left(X - (X - X_n) \leq t\right) \\
&= \mathbf{P}\left(X \leq t + Z_n\right) \\
&= \mathbf{P}\left(X \leq t + Z_n, Z_n < \varepsilon\right) + \mathbf{P}\left(X \leq t + Z_n, Z_n \geq \varepsilon\right) \\
&\leq \mathbf{P}\left(X \leq t + \varepsilon, Z_n < \varepsilon\right) + \mathbf{P}\left(X \leq t + Z_n, Z_n \geq \varepsilon\right) \\
&\leq \mathbf{P}\left(X \leq t + \varepsilon\right) + \mathbf{P}\left(X \leq t + Z_n, Z_n \geq \varepsilon\right) \\
&\leq \mathbf{P}\left(X \leq t + \varepsilon\right) + \mathbf{P}\left(X \leq \infty, Z_n \geq \varepsilon\right) \\
&\leq \mathbf{P}\left(X \leq t + \varepsilon\right) + \mathbf{P}\left(Z_n \geq \varepsilon\right) \\
&= F_X\left(t + \varepsilon\right) + \mathbf{P}\left(Z_n \geq \varepsilon\right).
\end{aligned}$$

Since $Z_n \xrightarrow{\mathrm{p}} 0$, $\mathbf{P}\left(Z_n \geq \varepsilon\right) \longrightarrow 0$ as $n \longrightarrow \infty$. It follows that

$$\limsup_{n\to\infty} F_{X_n}(t) \leq F_X(t + \varepsilon).$$

Taking $\varepsilon \longrightarrow 0$ and using the fact that t is a continuity point of F_X,

$$\limsup_{n\to\infty} F_{X_n}(t) \leq F_X(t).$$

Now go through the exactly same steps, replacing F_{X_n} with $1 - F_{X_n}$ and ε with $-\varepsilon$:

$$\begin{aligned}
1 - F_{X_n}(t) &= \mathbf{P}\left(X - (X - X_n) > t\right) \\
&= \mathbf{P}\left(X > t + Z_n\right) \\
&= \mathbf{P}\left(X > t + Z_n, Z_n \leq -\varepsilon\right) + \mathbf{P}\left(X > t + Z_n, Z_n > -\varepsilon\right) \\
&\leq \mathbf{P}\left(X > t + Z_n, Z_n \leq -\varepsilon\right) + \mathbf{P}\left(X > t - \varepsilon, Z_n > -\varepsilon\right) \\
&\leq \mathbf{P}\left(X > t + Z_n, Z_n \leq -\varepsilon\right) + \mathbf{P}\left(X > t - \varepsilon\right) \\
&\leq \mathbf{P}\left(X > -\infty, Z_n \leq -\varepsilon\right) + \mathbf{P}\left(X > t - \varepsilon\right) \\
&\leq \mathbf{P}\left(Z_n \leq -\varepsilon\right) + \mathbf{P}\left(X > t - \varepsilon\right) \\
&= \mathbf{P}\left(Z_n \leq -\varepsilon\right) + \left[1 - F_X\left(t - \varepsilon\right)\right].
\end{aligned}$$

Rearranging, $F_{X_n}(t) \geq F_X(t - \varepsilon) - \mathbf{P}\left(Z_n \leq -\varepsilon\right)$, which yields

$$\liminf_{n\to\infty} F_{X_n}(t) \geq F_X(t - \varepsilon)$$

since $Z_n \xrightarrow{\mathrm{p}} 0$. Taking $\varepsilon \longrightarrow 0$ and using the fact that t is a continuity point of F_X then gives us

$$\liminf_{n\to\infty} F_{X_n}(t) \geq F_X(t).$$

Putting together the pieces,

$$\limsup_{n\to\infty} F_{X_n}(t) \le F_X(t) \le \liminf_{n\to\infty} F_{X_n}(t).$$

Hence $\{F_{X_n}(t)\}$ is convergent and has limit $F_X(t)$. ■

There is a special case in which the converse is true:

Proposition 9. If $X_n \xrightarrow{\text{d}} X$ for X constant, then $X_n \xrightarrow{\text{p}} X$.

The result is pretty obvious, but the proof I've seen use parts of the Portmanteau lemma that I haven't stated, so I won't bother.

3.5 The Borel–Cantelli lemmata

The concentration inequalities in section 2.8 (p. 22) can be used to prove that a sequence of random elements converges in probability. For example, if $\{X_n\}$ are random variables with means μ and variances $n^{-\alpha}\sigma^2$ for some $\alpha > 0$, then Chebychev's inequality (p. 23) yields

$$\mathbf{P}(|X_n - \mu| > \varepsilon) \le \frac{\sigma^2}{n^\alpha \varepsilon^2} \to 0 \quad \text{for any } \varepsilon > 0,$$

so $X_n \xrightarrow{\text{p}} \mu$. (This is how we will prove Chebychev's WLLN in section 4.1 (p. 53).)

It would be nice to have a similarly tractable sufficient condition for almost sure convergence. That is exactly what the first Borel–Cantelli lemma gives us. And there's more: the second Borel–Cantelli lemma says that our sufficient condition is also necessary when the sequence is independent.

Theorem 7 (Borel–Cantelli lemmata)**.** Let $\{X_n\}$ and X be random elements of a metric space (S, ρ) defined on $(\Omega, \mathcal{A}, \mathbf{P})$.

(1) If $\sum_{n=1}^{\infty} \mathbf{P}(\rho(X_n, X) > \varepsilon) < \infty$ for all $\varepsilon > 0$, then $X_n \xrightarrow{\text{a.s.}} X$.

(2) If $\sum_{n=1}^{\infty} \mathbf{P}(\rho(X_n, X) > \varepsilon) = \infty$ for some $\varepsilon > 0$ and $\{X_n\}$ are independent, then X_n does not converge a.s. to X.

The Borel–Cantelli lemmata are actually much more general than what we stated here. If you care, see e.g. Rosenthal (2006, Theorem 3.4.2).

3.6 Convergence of moments

Suppose $X_n \xrightarrow{\text{d}} X$ for random variables $\{X_n\}$ and X. It might seem reasonable to conjecture that $\mathbf{E}(X_n) \to \mathbf{E}(X)$. But upon reflection, it's not a very good conjecture: by the Portmanteau lemma (p. 33), $X_n \xrightarrow{\text{d}} X$ is equivalent to

$$\int_{\mathbf{R}} f \mathrm{d}\mathcal{L}_{X_n} \to \int_{\mathbf{R}} f \mathrm{d}\mathcal{L}_X \quad \text{for every continuous and bounded } f,$$

but we want

$$\int_{\mathbf{R}} I \mathrm{d}\mathcal{L}_{X_n} \to \int_{\mathbf{R}} I \mathrm{d}\mathcal{L}_X$$

where I is the definitely-not-bounded identity function $I(x) \coloneqq x$! So $\{\mathcal{L}_{X_n}\}$ are going to have to be appropriately bounded if the moments are to converge.

I'll give two counterexamples. In the first, no moments exist along the sequence, but the limit distribution has moments. In the (perhaps less trivial) second example, moments exist along the entire sequence, but fail to converge nonetheless.

Example 7. Let $\{X_n\}$ be independent random variables with CDFs

$$F_{X_n}(x) \coloneqq \tfrac{1}{n} C(x) + \left(1 - \tfrac{1}{n}\right) \Phi(x)$$

where Φ is the standard normal CDF and C is the standard Cauchy CDF

$$C(x) \coloneqq \tfrac{1}{2} + \pi^{-1} \arctan(x).$$

It's obvious that $F_{X_n} \to \Phi$ pointwise, so $X_n \xrightarrow{\text{d}} \mathcal{N}(0,1)$. The expectation of the limit is therefore 0. But X_n has no mean for any $n \in \mathbf{N}$ since the Cauchy distribution has no moments. So the sequence $\{\mathbf{E}(X_n)\}$ does not even exist, hence certainly cannot be said to converge to zero.

Example 8. Consider random variables $\{X_n\}$ and X such that

$$\mathbf{P}(X_n = 1) = 1 - \tfrac{1}{n} \quad \text{and} \quad \mathbf{P}(X_n = n) = \tfrac{1}{n}$$

and $X \sim \mathcal{N}(0,1)$, with X independent of $\{X_n\}$. Define $Y_n \coloneqq (X_n X)^2$. Evidently $X_n^2 \xrightarrow{\text{p}} 1$, so by Slutsky's theorem $Y_n = X_n^2 X^2 \xrightarrow{\text{d}} X^2$. But $\mathbf{E}\left(X^2\right) = 1$, whereas (using independence)

$$\mathbf{E}\left(Y_n\right) = \mathbf{E}\left(X_n^2\right) \mathbf{E}\left(X^2\right) = \mathbf{E}\left(X_n^2\right) = \left(1 - \tfrac{1}{n}\right) + \tfrac{1}{n} n^2 \to \infty.$$

But even if $X_n \xrightarrow{\text{d}} X$ is not sufficient for $\mathbf{E}(X_n) \to \mathbf{E}(X)$, surely $X_n \xrightarrow{\text{a.s.}} X$ is sufficient? No again! We can still get the sort of pathological behaviour exhibited by the examples above. To rule this out, we need a boundedness condition on $\{X_n\}$ and X to rule out nonexistence or explosive behaviour.

There are several important theorems giving conditions under which $X_n \xrightarrow{\text{a.s.}} X$ implies $\mathbf{E}(X_n) \to \mathbf{E}(X)$. These include (in order from strongest to weakest assumptions) the monotone convergence theorem, the bounded convergence theorem, the (Lebesgue) dominated convergence theorem and the (Vitali) uniform integrability convergence theorem. The proofs of the last few rely heavily on Fatou's lemma. All of this is covered well by Rosenthal (2006, mainly ch. 9). We'll need the dominated convergence theorem later on, but I won't give a proof.

Theorem 8 (dominated convergence theorem)**.** Let $\{X_n\}$, X and Y be random variables such that $X_n \xrightarrow{\text{a.s.}} X$, $|X_n| \leq Y$ (pointwise) for each $n \in \mathbf{N}$, and $\mathbf{E}(Y)$ exists and is finite. Then $\mathbf{E}(X_n) \to \mathbf{E}(X)$.

In other words, when $X_n \xrightarrow{\text{a.s.}} X$, a sufficient condition for $\int_\Omega X_n \mathrm{d}\mathbf{P} \to \int_\Omega X \mathrm{d}\mathbf{P}$ is that $\{X_n\}$ is dominated by an integrable function (random variable) Y.

3.7 Characteristic functions

When we're working with random vectors, we have access to the following highly convenient object. Let $\mathbf{C}$ denote the complex plane.

Definition 25. Let X be a random n-vector on $(\Omega, \mathcal{A}, \mathbf{P})$. The characteristic function of X is $\phi_X : \mathbf{R}^n \to \mathbf{C}$ given by

$$\phi_X(t) \coloneqq \mathbf{E}\left(\exp\left(it^\top X\right)\right) = \int_{\mathbf{R}^n} \exp\left(it^\top x\right) \mathcal{L}_X(\mathrm{d}x) \quad \text{for each } t \in \mathbf{R}^n.^{37}$$

Above, we took the random variable X as the primitive. Although this is often natural, most of probability theory is concerned with measures, not random variables. It is therefore instructive to study the mapping $\mu \mapsto \phi^\mu$ from probability measures to the characteristic functions of random variables distributed according to those probability measures:

[37] Since $|\exp(ic)| = 1$ for any $c \in \mathbf{R}$, $\exp\left(it^\top X\right)$ is $\mathbf{P}$-integrable for any t. Hence ϕ_X is always well-defined, unlike the otherwise similar moment-generating function.

Definition 26. The characteristic transform is the mapping $\mu \mapsto \phi^\mu$ from probability measures μ on $(\mathbf{R}^n, \mathcal{B})$ to characteristic functions $\phi^\mu : \mathbf{R}^n \to \mathbf{C}$, defined by

$$\phi^\mu(t) \coloneqq \int_{\mathbf{R}^n} \exp\left(it^\top x\right) \mu(\mathrm{d}x) \quad \text{for each } t \in \mathbf{R}^n.^{38}$$

This definition might make you wonder what the space of characteristic functions is. It is by no means the case that every function $\mathbf{R}^n \to \mathbf{C}$ is the characteristic transform of some probability measure on $(\mathbf{R}^n, \mathcal{B})$! It is possible to state 'primitive' necessary and sufficient conditions for a function $\mathbf{R}^n \to \mathbf{C}$ to be a characteristic function. Bochner's theorem (e.g. Rao (1973, p. 141)) gives one set of necessary and sufficient conditions. Tractable sufficient (but not necessary) conditions are given by Pólya's theorem, which I'll only state for the univariate case.

Theorem 9 (Pólya's theorem)**.** Suppose $\varphi : \mathbf{R} \to \mathbf{C}$ is $\mathbf{R}$-valued, even,[39] continuous, convex on $\mathbf{R}_{++}$, and satisfies $\phi(0) = 1$ and $\lim_{t\to\infty} \phi(t) = 0$. Then $\varphi = \phi^\mu$ for some probability measure μ on $(\mathbf{R}, \mathcal{B})$ that is absolutely continuous w.r.t. Lebesgue measure and symmetric about 0.

It turns out that the space of characteristic functions is a dual of the space of probability measures in the following sense. First, the characteristic mapping is a bijection: $\phi^\mu = \phi^\nu$ (pointwise) iff $\mu = \nu$ (setwise). Second, the characteristic transform has a closed-form inverse, and there are many convenient equivalences between properties of probability measures and properties of characteristic functions. Third, the characteristic mapping is continuous in a certain sense.

Let's state two important bits of that formally. Proofs can be found in e.g. Rosenthal (2006, ch. 10).

Theorem 10 (Fourier uniqueness theorem)**.** Let μ and ν be probability measures on $(\mathbf{R}^n, \mathcal{B})$. Then $\phi^\mu = \phi^\nu$ (pointwise) iff $\mu = \nu$ (setwise).

Theorem 11 (Lévy's continuity theorem)**.** Let $\{\mu_n\}$ and μ be probability measures on $(\mathbf{R}^n, \mathcal{B})$. Then $\mu_n \Rightarrow \mu$ iff $\phi^{\mu_n} \longrightarrow \phi^\mu$ pointwise.[40]

[38] The characteristic transform is also sometimes known as (a version of) the Fourier transform. But what exactly is meant by 'Fourier transform' varies hugely between authors and fields, so I won't use this term at all.

[39] A function f is even iff $f(-x) = f(x)$ for every x in its domain.

[40] It's called the continuity theorem because when the space of probability measures is endowed with the topology of weak convergence (the weak* topology) and the space of characteristic functions is endowed with the topology of pointwise convergence (the product topology), the theorem says precisely that the characteristic transform and its inverse are continuous mappings.

These properties mean that any results we prove about characteristic functions, including convergence results, translate directly into results about probability measures (and vice versa). When we face a difficult question about probability measures on $(\mathbf{R}^n, \mathcal{B})$, we will often translate it into a question about characteristic functions, easily find the answer, then translate the answer back into probability-measure space.

The leading example of this strategy is the proof of the Lindeberg–Lévy central limit theorem (p. 63). But we'll also use it to prove part of the continuous mapping theorem on p. 45, and to establish an interesting property of the Cauchy distribution in an example on p. 42.

I mentioned equivalences between properties of measures and of their characteristic transforms. There are many, and they are easy to look up, but here are a few important ones.

Proposition 10. Let X and Y be random variables.

(1) $\phi_X(0) = 1$.

(2) $|\phi_X(t)| = 1$ for every $t \in \mathbf{R}^n$.

(3) $\phi_{aX+b}(t) = \exp(itb)\phi_X(at)$ for any $a \in \mathbf{R}$ and $b, t \in \mathbf{R}$.

(4) $\phi_{X+Y} = \phi_X \phi_Y$ if X and Y are independent. (The converse is not true!)

(5) If $\mathbf{E}\left(X^j\right)$ exists and is finite then $\phi_X^{(j)}(0)$ exists. (There's a partial converse.) Whenever they exist and are finite, $\phi_X^{(j)}(0) = i^j \mathbf{E}\left(X^j\right)$.

(6) ϕ_X is uniformly continuous.

(7) (Riemann–Lebesgue lemma) If $\mathcal{L}_X$ has a density w.r.t. Lebesgue measure, then $\phi_X(t) \longrightarrow 0$ as $|t| \longrightarrow \infty$.

Finally, an illustration.

Example 9 (the Cauchy law is stable)**.** A Cauchy-distributed random variable X is one whose density w.r.t. Lebesgue measure is

$$\frac{\mathrm{d}\mathcal{L}_X}{\mathrm{d}\lambda} = \frac{1}{\sqrt{\pi}} \frac{1}{1+x^2}.$$

A patient reader can verify that the corresponding characteristic function is $\phi_X(t) = \exp(-|t|)$. We know that the Cauchy distribution has no moments, so it shouldn't surprise us that ϕ_X is not differentiable at 0.[41]

[41] Above, we stated the result that when a moment exists and is finite, the corresponding derivative exists. We did not state the partial converse. So this does not constitute a proof that the Cauchy distribution has no moments!

Now consider a sequence $\{X_n\}$ of independent Cauchy-distributed random variables, and write $S_n \coloneqq n^{-1}\sum_{i=1}^{n} X_i$. Then

$$\begin{aligned}
\phi_{S_n}(t) &= \mathbf{E}\left(\exp\left(itn^{-1}\sum_{i=1}^{n} X_i\right)\right) \\
&= \mathbf{E}\left(\prod_{i=1}^{n}\exp\left(itn^{-1}X_i\right)\right) \\
&= \prod_{i=1}^{n}\mathbf{E}\left(\exp\left(itn^{-1}X_i\right)\right) \\
&= \mathbf{E}\left(\exp\left(itn^{-1}X_1\right)\right)^n \\
&= \phi_{X_1}\left(t/n\right)^n \\
&= \exp\left(-|t|/n\right)^n \\
&= \exp\left(-|t|\right) \\
&= \phi_{X_1}\left(t\right).
\end{aligned}$$

So the average of n Cauchy-distributed random variables is itself Cauchy-distributed! n-fold convolution of the Cauchy distribution is itself Cauchy! A distribution with the property that aS_n for some a has the same distribution as X_1 is called a (Lévy or α) stable distribution. Another stable law is the normal distribution (you already knew that—think about it). The theory of stable laws is a very interesting branch of probability theory, I think.[42]

This example also serves as a prelude to our study of laws of large numbers, which give conditions under which S_n converges (a.s. or in probability) to a constant. It should be obvious that S_n converges weakly to a Cauchy law; we don't even need Lévy's continuity theorem to prove this. Convergence to a point fails to happen here because the Cauchy distribution has 'heavy tails', i.e. lots of probability mass in the tails. (The formal definition of 'heavy tail'

[42]Think of a sequence of distributions of S_n as a path in the space of probability distributions. This path is governed by a law of motion. A stable distribution is a steady state of this law of motion: once you're there, you don't leave. Some of these steady states may be attractors in some region: if you start in this region, the sequence converges weakly to the stable law. One theorem in the theory of stable laws is (loosely) that only stable laws can be attractors.

Moreover, there are generalisations of the central limit theorems. CLTs give (large) regions in which the normal law is an attractor; 'generalised central limit theorems' give large regions in which the CLT fails (due to infinite variance), but in which there is another attractor. By the previous result, this attractor must be a stable distribution, but it will not be normal. This material can be found in e.g. Gnedenko and Kolmogorov (1954, ch. 7) and Durrett (2010, sec. 3.7).

is usually that the variance is infinite.) As we will see when we prove LLNs, moment restrictions are required in order to avoid this sort of problem. (At the very least, we'll require the first moment to exist.)

3.8 The continuous mapping theorem

One characterisation of continuity in metric spaces is that a continuous mapping is one that 'preserves convergence': f is continuous at x_0 iff $f(x_n) \longrightarrow f(x_0)$ for any sequence $\{x_n\}$ s.t. $x_n \longrightarrow x_0$. The Mann–Wald continuous mapping theorem (CMT) is the analog for random variables of the 'only if' part of this characterisation: it says that a.s. convergence, convergence in probability and convergence in distribution are all preserved under almost-everywhere continuous transformations.

Theorem 12 (Mann–Wald CMT)**.** Let (S, ρ) and (S', ρ') be metric spaces, let $\{X_n\}$ and X be random elements of (S, ρ), and let $g : S \to S'$ be measurable and continuous $\mathcal{L}_X$-a.e.[43] Then

(1) $X_n \xrightarrow{\text{a.s.}} X$ implies $g(X_n) \xrightarrow{\text{a.s.}} g(X)$.

(2) $X_n \xrightarrow{\text{p}} X$ implies $g(X_n) \xrightarrow{\text{p}} g(X)$.

(3) $X_n \xrightarrow{\text{d}} X$ implies $g(X_n) \xrightarrow{\text{d}} g(X)$.

Remark 3. We did not mention convergence in mean square because it turns out not to be preserved under arbitrary a.e.-continuous mappings! Something much stronger is needed, e.g. g linear.

Proof of (1). We know that there are measurable $\Omega', \Omega'' \subseteq \Omega$ such that $X_n(\omega) \longrightarrow X(\omega)$ for all $\omega \in \Omega'$, g is continuous at all $X(\omega)$ s.t. $\omega \in \Omega''$, and $\mathbf{P}(\Omega') = \mathbf{P}(\Omega'') = 1$. Firstly, $\Omega' \cap \Omega''$ is measurable with $\mathbf{P}(\Omega' \cap \Omega'') = 1$ since

$$\mathbf{P}(\Omega' \cap \Omega'') = 1 - \mathbf{P}\left((\Omega')^{\text{C}} \cup (\Omega'')^{\text{C}}\right) \geq 1 - \mathbf{P}\left((\Omega')^{\text{C}}\right) - \mathbf{P}\left((\Omega'')^{\text{C}}\right) = 1.$$

Secondly, $g(X_n(\omega)) \longrightarrow g(X(\omega))$ at all $\omega \in \Omega' \cap \Omega''$. ■

We won't bother proving (2) in full generality, though it is not hard. Instead, we will content ourselves with the case in which $X_n \xrightarrow{\text{p}} \alpha$ for a constant α. In this case, continuity $\mathcal{L}_X$-a.e. of g reduces to continuity of g at α.

[43] Recall that $\mathcal{L}_X$ is the law of X. So the final requirement is that the underlying probability space $(\Omega, \mathcal{A}, \mathbf{P})$ satisfies $\mathbf{P}(\{\omega \in \Omega : g \text{ continuous at } X(\omega)\}) = 1$.

Proof of (2) for constant X. By continuity of g at α, for each $\varepsilon > 0$, there is a $\delta > 0$ such that $\rho(X_n, \alpha) < \delta$ implies $\rho'(g(X_n), g(\alpha)) < \varepsilon$. So

$$1 \geq \mathbf{P}\left(\rho'(g(X_n), g(\alpha)) < \varepsilon\right) \geq \mathbf{P}(\rho(X_n, \alpha) < \delta).$$

Since $X_n \xrightarrow{\text{p}} \alpha$, the right-hand side converges to 1 regardless of δ. It follows that $\mathbf{P}\left(\rho'(g(X_n), g(\alpha)) < \varepsilon\right) \longrightarrow 1$ for each $\varepsilon > 0$, i.e. $g(X_n) \xrightarrow{\text{p}} g(\alpha)$. ■

For (3), the cleanest general proof that I've seen uses Skorokhod's theorem, then follows the argument for (1). This would take us too far afield, so let's restrict attention to the case in which $\{X_n\}$ and X are random ℓ-vectors and $g : \mathbf{R}^\ell \to \mathbf{R}^m$, so that we can use characteristic functions.

Proof of (3) for $S = \mathbf{R}^\ell$ and $S' = \mathbf{R}^m$. Fix $t \in \mathbf{R}^\ell$; we wish to show that $\phi_{g(X_n)}(t) \longrightarrow \phi_{g(X)}(t)$. We have

$$\begin{aligned}
\phi_{g(X_n)}(t) &= \int_{\mathbf{R}^\ell} \exp\left(it^\top g(y)\right) \mathcal{L}_{X_n}(\mathrm{d}y) \\
&= \int_{\mathbf{R}^\ell} \cos\left(t^\top g(y)\right) \mathcal{L}_{X_n}(\mathrm{d}y) + i \int_{\mathbf{R}^\ell} \sin\left(t^\top g(y)\right) \mathcal{L}_{X_n}(\mathrm{d}y). \\
&\longrightarrow \int_{\mathbf{R}^\ell} \cos\left(t^\top g(y)\right) \mathcal{L}_{X}(\mathrm{d}y) + i \int_{\mathbf{R}^\ell} \sin\left(t^\top g(y)\right) \mathcal{L}_{X}(\mathrm{d}y) \\
&= \int_{\mathbf{R}^\ell} \exp\left(it^\top g(y)\right) \mathcal{L}_{X}(\mathrm{d}y) \\
&= \phi_{g(X)}(t)
\end{aligned}$$

where convergence follows by the Portmanteau lemma (p. 33),[44] since $y \mapsto \cos\left(t^\top g(y)\right)$ and $y \mapsto \sin\left(t^\top g(y)\right)$ are bounded and continuous mappings. Hence $g(X_n) \xrightarrow{\text{d}} g(X)$ by Lévy's continuity theorem (p. 41). ■

The following result is an oft-used corollary to the continuous mapping theorem. It states that the elementary algebraic operations of addition, multiplication and division are preserved under weak convergence. (It's a corollary because these operations are continuous.)

Corollary 3 (Slutsky's theorem)**.** Let $\{X_n\}$ and X be $m \times k$ random matrices, let $\{Y_n\}$ be $k \times k$ random matrices, and let A be a $k \times k$ (constant) matrix. Suppose that $X_n \xrightarrow{\text{d}} X$ and $Y_n \xrightarrow{\text{p}} A$. Then

(1) $X_n + Y_n \xrightarrow{\text{d}} X + A$.

[44] Joel actually appealed to the Helly–Bray theorem, which is a special case of the Portmanteau lemma.

(2) $X_n Y_n \xrightarrow{\text{d}} XA$.

(3) $X_n Y_n^{-1} \xrightarrow{\text{d}} XA^{-1}$ provided A is invertible.[45]

Proof. (X, A) and each (X_n, Y_n) are random elements of the metric space $\mathbf{R}^{m\times k} \times \mathbf{R}^{k\times k}$. $Y_n \xrightarrow{\text{p}} A$ implies $Y_n \xrightarrow{\text{d}} A$ by Proposition 9 (p. 38). The mappings $(x, y) \mapsto x + y$ and $(x, y) \mapsto xy$ are continuous, and $(x, y) \mapsto xy^{-1}$ is continuous whenever y is invertible. The result then follows from part (3) of the continuous mapping theorem. ■

Remark 4. Since the proof of Slutsky's theorem is via the Mann–Wald CMT, the result obviously still holds if we replace $\xrightarrow{\text{d}}$ with $\xrightarrow{\text{p}}$ or $\xrightarrow{\text{a.s.}}$. But be careful here: it's important that Y converges to a constant rather than to a random matrix. When X_n and Y_n both converge to random elements X and Y, it need not be that $X_n + Y_n \xrightarrow{\text{d}} X + Y$, $X_n Y_n \xrightarrow{\text{d}} XY$ or $X_n Y_n^{-1} \xrightarrow{\text{d}} XY^{-1}$. The following example illustrates.

Example 10 (weak convergence of marginals vs. joint)**.** Let $\{X_n\}$, $\{Y_n\}$, X and Y be random variables distributed

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \overset{\text{iid}}{\sim} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) \quad \text{and} \quad \begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \right).$$

The marginal distributions of X_n, Y_n, X and Y are all $\mathcal{N}(0, 1)$. Hence (trivially) we have $X_n \xrightarrow{\text{d}} X$ and $Y_n \xrightarrow{\text{d}} Y$. But

$$X_n + Y_n \sim \mathcal{N}(0, 2(1 + \rho)) \quad \text{and} \quad X + Y \sim \mathcal{N}(0, 2(1 + r)),$$

so it is generally not the case that $X_n + Y_n \xrightarrow{\text{d}} X + Y$! The reason is clear: although the marginal distributions converge weakly, the joint distribution does not, as evidenced by the fact that ρ may differ from r.

In the example, it's clear that if X_n, Y_n are independent ($\rho = 0$) and X, Y are also independent ($r = 0$) then we do in fact have $X_n + Y_n \xrightarrow{\text{d}} X + Y$. This is true in general, since then

$$\phi_{X_n+Y_n}(t) = \phi_{X_n}(t)\phi_{Y_n}(t) \longrightarrow \phi_X(t)\phi_Y(t) = \phi_{X+Y}(t) \quad \text{for arbitrary } t \in \mathbf{R},$$

whence $X_n + Y_n \xrightarrow{\text{d}} X + Y$ follows by Lévy's continuity theorem.

[45] If some $\{Y_n\}$ in (3) are singular, we can replace Y_n^{-1} with a Moore–Penrose pseudo-inverse and still obtain convergence. The Moore–Penrose pseudo-inverse is continuous at invertibility points, so the continuous mapping theorem applies. (But note that unlike the ordinary matrix inverse, the Moore–Penrose pseudo-inverse is not continuous at all points.)

3.9 Stochastic order notation

When we use approximations, we have to control the approximation error. Usually, we want the error to vanish as the sample size grows large. The notation introduced here offers a compact way of keeping track of approximation error. This section will treat sequences in $\mathbf{R}$, on the understanding that the extension to $\mathbf{R}^n$ is straightforward.

Let's start out with order notation from analysis.

Definition 27. Let $\{x_n\}$ and $\{a_n\}$ be sequences in $\mathbf{R}$.

(1) $x_n = \mathrm{O}(a_n)$ iff $\exists M_0 > 0$ s.t. $|X_n/a_n| \leq M_0$ for n sufficiently large.

(2) $x_n = \mathrm{o}(a_n)$ iff $x_n/a_n \longrightarrow 0$.

Intuitively, $x_n = \mathrm{O}(a_n)$ means that $\{x_n\}$ increases no faster than $\{a_n\}$, while $x_n = \mathrm{o}(a_n)$ means that $\{x_n\}$ increases at a slower rate than $\{a_n\}$. Unsurprisingly, these concepts are not well-suited for use with random variables. We therefore use analogous 'in probability' definitions.

Definition 28. Let $\{X_n\}$ be a sequence of random variables and $\{a_n\}$ be a sequence in $\mathbf{R}$.

(1) $X_n = \mathrm{O}_\mathrm{p}(a_n)$ iff $\forall \varepsilon > 0$, $\exists M_\varepsilon > 0$ s.t. $\mathbf{P}\left(|X_n/a_n| \leq M_\varepsilon\right) \geq 1 - \varepsilon$ for n sufficiently large.

(2) $X_n = \mathrm{o}_\mathrm{p}(a_n)$ iff $X_n/a_n \xrightarrow{\mathrm{p}} 0$.

The parallel with O and o is clear. We're weakening them in the 'in probability' way, as opposed to in the 'almost sure' way because the latter would be too strong (but easier, really).

To compare O_p and o_p, use the definition of convergence in probability to see that $X_n = \mathrm{o}_\mathrm{p}(a_n)$ iff $\forall \varepsilon > 0$, $\forall M_0 > 0$, $\mathbf{P}\left(|X_n/a_n| \leq M_0\right) \geq 1 - \varepsilon$ for n sufficiently large. The latter contains 'for all M_0' rather than 'there exists an M_ε'. This should make it clear that $X_n = \mathrm{o}_\mathrm{p}(a_n)$ implies $X_n = \mathrm{O}_\mathrm{p}(a_n)$.

An unfortunate feature of order notation is that it breaks the symmetry of the equality symbol. Concisely put, $x_n = \mathrm{O}(a_n)$ says that x_n is of order a_n; it does not say that the object $\mathrm{O}(a_n)$ is equal to x_n. So $x_n = \mathrm{O}(a_n)$ must be read left-to-right, not right-to-left. This is the convention, and I will be using it. Be forewarned!

Notice that anything that that is bounded (vanishing) is also bounded (vanishing) in probability:

$$\mathrm{O}(a_n) = \mathrm{O}_\mathrm{p}(a_n) \quad \text{and} \quad \mathrm{o}(a_n) = \mathrm{o}_\mathrm{p}(a_n).$$

Of course, the converse is not true, i.e. $\mathrm{O_p}(a_n) = \mathrm{O}(a_n)$ and $\mathrm{o_p}(a_n) = \mathrm{o}(a_n)$ are false in general. (A sequence may be bounded/vanishing in probability without being bounded/vanishing for sure.)

We will do a lot of algebra involving $\mathrm{O_p}$ and $\mathrm{o_p}$ once we start studying estimators, so here's a collection of facts about how $\mathrm{O_p}$ and $\mathrm{o_p}$ can be manipulated. Except for the last one, they are all easily proved from the definitions.

Proposition 11. Some facts about $\mathrm{O_p}$ and $\mathrm{o_p}$:

(1) $\mathrm{o_p}(a_n) = a_n \mathrm{o_p}(1)$ and $\mathrm{O_p}(a_n) = a_n \mathrm{O_p}(1)$.

(2) $\mathrm{o_p}(\mathrm{O_p}(1)) = \mathrm{o_p}(1)$.

(3) $\mathrm{o_p}(1) + \mathrm{O_p}(1) = \mathrm{O_p}(1)$.

(4) $\mathrm{o_p}(1)\mathrm{O_p}(1) = \mathrm{o_p}(1)$.

(5) $(1 + \mathrm{o_p}(1))^{-1} = \mathrm{O_p}(1)$.

(6) If $R(0) = 0$, $R(h) = \mathrm{o}\left(\|h\|^p\right)$ as $h \downarrow 0$, and $a_n = \mathrm{o_p}(1)$, then $R(a_n) = \mathrm{o_p}\left(\|a_n\|^p\right)$.

3.10 The delta method

Suppose you know that a random vector X_n is approximately distributed as $a_n^{-1}W + b$ for large n, where W is a random vector (formally $a_n(X_n - b) \xrightarrow{\mathrm{d}} W$), but that you're actually interested in approximating the distribution of some function $g(X_n)$ of this random vector (e.g. a test statistic). The delta method provides a way of doing this whenever g is smooth near b. Formally, it is based on a Taylor expansion.

Theorem 13 (Taylor's theorem)**.** Let $g : \mathbf{R} \to \mathbf{R}$ be ℓ times differentiable in an open neighbourhood of b.[46] Then

$$g(x) - g(b) = \sum_{j=1}^{\ell} \frac{g^{(j)}(b)}{j!}(x-b)^j + \mathrm{o}\left(|x-b|^\ell\right),$$

where $g^{(j)}$ denotes the jth derivative.

[46] Some authors state Taylor's theorem requiring only differentiability at b, but the proof seems to require differentiability in a neighbourhood.

The theorem extends immediately to any ℓ times differentiable function $g : \mathbf{R}^k \to \mathbf{R}^m$, but the notation becomes ugly fast (tensor products). For $g : \mathbf{R}^k \to \mathbf{R}$, we can go to second order without notational trouble:

$$g(x) - g(b) = \nabla g(b)^\top (x-b) + \tfrac{1}{2}(x-b)^\top \nabla^2 g(b)(x-b) + \mathrm{o}\left(\left(\|x-b\|_2 \right)^2 \right).$$

For $g : \mathbf{R}^n \to \mathbf{R}^m$, only a first-order expansion is easy to write down:

$$g(x) - g(b) = \mathrm{D}g(b)(x-b) + \mathrm{o}\left(\|x-b\|_2 \right).$$

Proposition 12 (delta method)**.** Let $\{X_n\}$ be a sequence of random k-vectors such that $a_n(X_n - b) \xrightarrow{\mathrm{d}} W$ for some constants $\{a_n\}$ and b, and let $g : \mathbf{R}^k \to \mathbf{R}^m$ be differentiable in an open neighbourhood of b, with derivative $\mathrm{D}g(b)$ at b. Then

$$a_n \left(g(X_n) - g(b) \right) \xrightarrow{\mathrm{d}} \mathrm{D}g(b) W.$$

Remark 5. Notice that we did not require $\mathrm{D}g(b)$ to be nonsingular (or even nonzero), nor did we require $\mathrm{D}g$ to be continuous at b. Although W will be normally distributed in the vast majority of applications (by a central limit theorem; see section 5), that is not required, either.

Proof. By Taylor's theorem,

$$g(X_n) - g(b) = \mathrm{D}g(b)(X_n - b) + \mathrm{o_p}\left(\|X_n - b\|_2 \right),$$

so

$$a_n \left(g(X_n) - g(b) \right) = \mathrm{D}g(b) a_n (X_n - b) + \mathrm{o_p}\left(\|a_n \left(X_n - b \right)\|_2 \right).$$

Since $a_n \left(X_n - b \right) \xrightarrow{\mathrm{d}} W$,

$$\mathrm{o_p}\left(\|a_n \left(X_n - b \right)\|_2 \right) = \mathrm{o_p}(\mathrm{O_p}(1)) = \mathrm{o_p}(1),$$

and $\mathrm{D}g(b) a_n (X_n - b) \xrightarrow{\mathrm{d}} \mathrm{D}g(b) W$ by Slutsky's theorem (p. 45). Hence

$$a_n \left(g(X_n) - g(b) \right) = \mathrm{D}g(b) a_n (X_n - b) + \mathrm{o_p}\left(1 \right) \xrightarrow{\mathrm{d}} \mathrm{D}g(b) W. \qquad \blacksquare$$

Remark 6. Suppose instead that we have a sequence $\{b_n\}$ of k-vectors such that $a_n(X_n - b_n) \xrightarrow{\mathrm{d}} W$, and that $b_n \longrightarrow b$. Add the assumption that $\mathrm{D}g$ is continuous at b. Then $\mathrm{D}g(b_n) = \mathrm{D}g(b) + \mathrm{o}(1)$, so the proof above still goes through, giving us

$$a_n (g(X_n) - g(b_n)) \xrightarrow{\mathrm{d}} \mathrm{D}g(b) W.$$

(The same extension is available for the second- and ℓth-order delta methods below.)

Although the first-order delta method above is valid when $\mathrm{D}g(b) = 0$, it isn't very helpful in that case. Unless g is a constant function, $g(X_n)$ is still going to be random, so we'd like our approximating distribution to be nondegenerate. The obvious remedy is to use a second-order Taylor expansion. As noted above, this would require heavy notation for the case $g : \mathbf{R}^k \to \mathbf{R}^m$, so we'll just state it for the case $g : \mathbf{R}^k \to \mathbf{R}$.

Proposition 13 (second-order delta method)**.** Let $\{X_n\}$ be a sequence of random k-vectors such that $a_n(X_n - b) \xrightarrow{\mathrm{d}} W$ for some constants $\{a_n\}$ and b, and let $g : \mathbf{R}^k \to \mathbf{R}$ be twice differentiable in an open neighbourhood of b, with derivatives $\nabla g(b) = 0$ and $\nabla^2 g(b)$ at b. Then

$$a_n^2 \left(g(X_n) - g(b)\right) \xrightarrow{\mathrm{d}} \tfrac{1}{2} W^\top \nabla^2 g(b) W.$$

Proof. By Taylor's theorem and $\nabla g(b) = 0$,

$$g(X_n) - g(b) = \tfrac{1}{2}(X_n - b)^\top \nabla^2 g(b)(X_n - b) + \mathrm{o_p}\left(\left(\|X_n - b\|_2\right)^2\right),$$

so

$$\begin{aligned} &a_n^2 \left(g(X_n) - g(b)\right) \\ &\qquad = \tfrac{1}{2}[a_n(X_n - b)]^\top \nabla^2 g(b)[a_n(X_n - b)] + \mathrm{o_p}\left(\left(\|a_n(X_n - b)\|_2\right)^2\right). \end{aligned}$$

Since $a_n \left(X_n - b\right) \xrightarrow{\mathrm{d}} W$,

$$\mathrm{o_p}\left(\left(\|a_n \left(X_n - b\right)\|_2\right)^2\right) = \mathrm{o_p}\left(\mathrm{O_p}(1)^2\right) = \mathrm{o_p}(\mathrm{O_p}(1)) = \mathrm{o_p}(1),$$

and

$$\tfrac{1}{2}[a_n(X_n - b)]^\top \nabla^2 g(b)[a_n(X_n - b)] \xrightarrow{\mathrm{d}} \tfrac{1}{2} W^\top \nabla^2 g(b) W$$

by Slutsky's theorem (p. 45). Hence

$$\begin{aligned} a_n^2 \left(g(X_n) - g(b)\right) &= \tfrac{1}{2}[a_n(X_n - b)]^\top \nabla^2 g(b)[a_n(X_n - b)] + \mathrm{o_p}\left(1\right) \\ &\xrightarrow{\mathrm{d}} \tfrac{1}{2} W^\top \nabla^2 g(b) W. \end{aligned}$$

■

Remark 7. Combining the first- and second-order delta methods, we get

$$a_n \left(g(X_n) - g(b)\right) \xrightarrow{\mathrm{p}} 0 \quad \text{and} \quad a_n^2 \left(g(X_n) - g(b)\right) \xrightarrow{\mathrm{d}} \tfrac{1}{2} W^\top \nabla^2 g(b) W.$$

(I can write $\xrightarrow{\mathrm{p}}$ rather than $\xrightarrow{\mathrm{d}}$ by Proposition 9 (p. 38).) There is no contradiction between the two: we get different behaviour because we're using different scaling factors ($\{a_n\}$ vs. $\{a_n^2\}$).

Remark 8. Even if $\nabla g(b) \neq 0$, we could use a second-order Taylor expansion to approximate the distribution of $g(X_n)$. But this makes the approximation so complicated that it's rarely worthwhile.

Of course, there's nothing special about the second order: if the first $\ell - 1$ derivatives are zero, we can use the ℓth derivative to approximate the distribution of $g(X_n)$. To duck notational difficulties, I'll only state this for the case $g : \mathbf{R} \to \mathbf{R}$.

Proposition 14 (ℓth-order delta method)**.** Let $\{X_n\}$ be a sequence of random variables such that $a_n(X_n - b) \xrightarrow{\mathrm{d}} W$ for some constants $\{a_n\}$ and b, and let $g : \mathbf{R} \to \mathbf{R}$ be ℓ times differentiable in an open neighbourhood of b, with derivatives $g'(b) = \cdots = g^{(\ell-1)}(b) = 0$ and $g^{(\ell)}(b)$ at b. Then

$$a_n^\ell \left(g(X_n) - g(b)\right) \xrightarrow{\mathrm{d}} \frac{g^{(\ell)}(b)}{\ell!} W^\ell.$$

Proof. By Taylor's theorem,

$$g(X_n) - g(b) = \frac{g^{(\ell)}(b)}{\ell!}(X_n - b)^\ell + \mathrm{o_p}\left(|X_n - b|^\ell\right),$$

so

$$a_n^\ell \left(g(X_n) - g(b)\right) = \frac{g^{(\ell)}(b)}{\ell!}\left[a_n(X_n - b)\right]^\ell + \mathrm{o_p}\left(|a_n(X_n - b)|^\ell\right).$$

Since $a_n \left(X_n - b\right) \xrightarrow{\mathrm{d}} W$,

$$\mathrm{o_p}\left(|a_n \left(X_n - b\right)|^\ell\right) = \mathrm{o_p}\left(\mathrm{O_p}(1)^\ell\right) = \mathrm{o_p}(\mathrm{O_p}(1)) = \mathrm{o_p}(1),$$

and $\left[a_n(X_n - b)\right]^\ell \xrightarrow{\mathrm{d}} W^\ell$ by the continuous mapping theorem (p. 44). So by Slutsky's theorem (p. 45),

$$a_n^\ell \left(g(X_n) - g(b)\right) = \frac{g^{(\ell)}(b)}{\ell!}\left[a_n(X_n - b)\right]^\ell + \mathrm{o_p}(1) \xrightarrow{\mathrm{d}} \frac{g^{(\ell)}(b)}{\ell!} W^\ell. \qquad \blacksquare$$

Before we move on, here's an illustration of how the delta method can be used in econometrics. The example makes use of a law of large numbers and a central limit theorem which will not be covered until sections 4 and 5.

Example 11 ($\exp(\alpha)$ ML estimator)**.** The exponential distribution with parameter $\alpha > 0$ (denoted $\exp(\alpha)$) is any distribution on $(\mathbf{R}, \mathcal{B})$ whose density

w.r.t. Lebesgue measure is $f(x) = \alpha \exp(-\alpha x)$. The mean and variance of this distribution are α^{-1} and α^{-2}.

Suppose we have n iid random variables $\{X_i\}_{i=1}^n$ drawn from the $\exp(\alpha)$ distribution, and wish to estimate α. The obvious analogy estimator, which turns out to also be the maximum likelihood estimator, is

$$\widehat{\alpha}_n \coloneqq \left(n^{-1} \sum_{i=1}^n X_i \right)^{-1}.$$

By Kolmogorov's second SLLN (p. 56),

$$n^{-1} \sum_{i=1}^n X_i \xrightarrow{\text{a.s.}} \mathbf{E}(X_i) = \alpha^{-1},$$

so by the continuous mapping theorem $\widehat{\alpha}_n \xrightarrow{\text{a.s.}} \alpha$, i.e. the estimator is strongly consistent. So $\widehat{\alpha}_n$ will be 'close' to α in a large sample.

But how close? To answer this question, we need to approximate the distribution of $\widehat{\alpha}_n$ in a large sample. The Lindeberg–Lévy CLT (p. 63) gives us

$$n^{-1/2} \sum_{i=1}^n \frac{X_i - \alpha^{-1}}{\sqrt{\alpha^{-2}}} \xrightarrow{\text{d}} \mathcal{N}(0,1),$$

which we can rewrite as

$$n^{1/2} \alpha (\widehat{\alpha}_n^{-1} - \alpha^{-1}) \xrightarrow{\text{d}} \mathcal{N}(0,1).$$

Now use the delta method with $g(x) = 1/x$ (so $g'(x) = -1/x^2$), $a_n = n^{1/2}\alpha$ and $b = \alpha^{-1}$ to obtain

$$n^{1/2} \alpha (\widehat{\alpha}_n - \alpha) \xrightarrow{\text{d}} \left(-1/\alpha^{-2}\right) \mathcal{N}(0,1),$$

or equivalently

$$n^{1/2} (\widehat{\alpha}_n - \alpha) \xrightarrow{\text{d}} \mathcal{N}\left(0, \alpha^2\right).$$

So for n large, the distribution of $\widehat{\alpha}_n$ is well-approximated by $\mathcal{N}\left(\alpha, n^{-1}\alpha^2\right)$.

4 Laws of large numbers

Official reading: Amemiya (1985, ch. 3), Rao (1973, ch. 2) and White (2001, ch. 3).

A law of large numbers (LLN) gives conditions under which the average of n random variables converges as $n \longrightarrow \infty$. They are called weak laws (WLLNs) if convergence is in probability, and strong laws (SLLNs) if convergence is almost sure.[47]

There are a lot of different laws of large numbers. The common theme is that the volatility of the average must be controlled by combining two kinds of restriction. On the one hand, we restrict the individual variances to keep them from getting too large. On the other hand, we restrict the dependence between the random variables, so that one extreme realisation doesn't make further extreme realisations likely. Each LLN imposes some mix of the two, and often we can weaken the one at the expense of strengthening the other.

4.1 Uncorrelated/independent random variables

We begin with an easy-to-prove weak law.

Theorem 14 (Chebychev's WLLN)**.** Let $\{X_n\}$ be a sequence of uncorrelated random variables with

$$\lim_{n\to\infty} n^{-2} \sum_{i=1}^{n} \operatorname{Var}(X_i) = 0.$$

Then $n^{-1} \sum_{i=1}^{n} (X_i - \mathbf{E}(X_i)) \xrightarrow{\text{p}} 0$ as $n \longrightarrow \infty$.

Remark 9. Three separate remarks, really.

(1) Neither $\{n^{-1} \sum_{i=1}^{n} X_i\}$ nor $\{n^{-1} \sum_{i=1}^{n} \mathbf{E}(X_i)\}$ need converge to anything; they could be 'exploding together', for example.

(2) The restriction on the variances implies that each variance is finite, hence that each mean exists and is finite.

(3) The variance condition can be weakened.

Proof. Write

$$S_n \coloneqq n^{-1} \sum_{i=1}^{n} (X_i - \mathbf{E}(X_i));$$

[47] As indicated, we will state our results for random variables. They can of course be applied element-wise to random vectors.

we want to show that $S_n \xrightarrow{\text{p}} 0$.

$$\operatorname{Var}(S_n) = n^{-2}\operatorname{Var}\left(\sum_{i=1}^{n}(X_i - \mathbf{E}(X_i))\right) = n^{-2}\sum_{i=1}^{n}\operatorname{Var}(X_i)$$

by uncorrelatedness. Hence $\operatorname{Var}(S_n) \longrightarrow 0$ by the variance condition. By nonnegativity and Chebychev's inequality, we have for any $\varepsilon > 0$ that

$$0 \leq \mathbf{P}(|S_n| > \varepsilon) \leq \operatorname{Var}(S_n)/\varepsilon^2.$$

Since the RHS converges to 0, it follows that $\mathbf{P}(|S_n| > \varepsilon) \longrightarrow 0$ for every $\varepsilon > 0$, i.e. $S_n \xrightarrow{\text{p}} 0$. ■

Now for an easy strong law. It isn't actually used very often, but it plays an important role in the proof of Kolmogorov's second SLLN.

Theorem 15 (Kolmogorov's first SLLN)**.** Let $\{X_n\}$ be a sequence of independent random variables with

$$\sum_{i=1}^{\infty}\frac{\operatorname{Var}(X_i)}{i^2} < \infty.$$

Then $n^{-1}\sum_{i=1}^{n}(X_i - \mathbf{E}(X_i)) \xrightarrow{\text{a.s.}} 0$ as $n \longrightarrow \infty$.

Remark 10. The Kolmogorov variance condition

$$\sum_{i=1}^{\infty}\operatorname{Var}(X_i)/i^2 < \infty$$

implies the Chebychev variance condition

$$\lim_{n\to\infty} n^{-2}\sum_{i=1}^{n}\operatorname{Var}(X_i) = 0$$

by Kronecker's lemma (below). So the Kolmogorov SLLN strengthens both the variance restriction and the dependence restriction (from uncorrelatedness to independence). Our reward is a stronger result, viz. almost sure convergence.

Our proof will make use of two horrendous inequalities: the Hájek–Rényi inequality (p. 24), and Kronecker's lemma. The latter is

Lemma 3 (Kronecker's lemma)**.** Let $\{x_n\}$ be a sequence in $\mathbf{R}$ such that $\sum_{n=1}^{\infty} x_i$ exists and is finite. Then for any weakly increasing sequence $\{c_n\}$ in $\mathbf{R}_{++}$ such that $c_n \longrightarrow \infty$, $\lim_{n\to\infty} c_n^{-1}\sum_{i=1}^{n} c_i x_i = 0$.

Proof of Kolmogorov's first SLLN. Write

$$S_n \coloneqq n^{-1} \sum_{i=1}^{n} (X_i - \mathbf{E}(X_i));$$

we want to show that $S_n \xrightarrow{\text{a.s.}} 0$. Fix $\varepsilon > 0$. Using the Hájek–Rényi inequality with $c_i = i^{-1}$,

$$\begin{aligned}\mathbf{P}\left(\max_{k\in[m,n]} |S_k| \geq \varepsilon\right) &= \mathbf{P}\left(\max_{k\in[m,n]} k^{-1} \left|\sum_{i=1}^{k} (X_i - \mathbf{E}(X_i))\right| \geq \varepsilon\right) \\ &\leq \frac{1}{\varepsilon^2}\left(m^{-2}\sum_{i=1}^{m} \operatorname{Var}(X_i) + \sum_{i=m+1}^{n} i^{-2}\operatorname{Var}(X_i)\right).\end{aligned}$$

Taking $n \to \infty$ and using the fact that $\sum_{i=1}^{\infty} \operatorname{Var}(X_i)/i^2$ converges,

$$\mathbf{P}\left(\max_{k\geq m} |S_k| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2}\left(m^{-2}\sum_{i=1}^{m} \operatorname{Var}(X_i) + \sum_{i=m+1}^{\infty} i^{-2}\operatorname{Var}(X_i)\right).$$

Now taking $m \to \infty$,

$$\begin{aligned}&\lim_{m\to\infty} \mathbf{P}\left(\sup_{k\geq m} |S_k| \geq \varepsilon\right) \\ &\qquad \leq \frac{1}{\varepsilon^2}\left(\lim_{m\to\infty} m^{-2}\sum_{i=1}^{m} \operatorname{Var}(X_i) + \lim_{m\to\infty} \sum_{i=m+1}^{\infty} i^{-2}\operatorname{Var}(X_i)\right).\end{aligned}$$

Since $\sum_{i=1}^{\infty} \operatorname{Var}(X_i)/i^2$ exists and is finite, Kronecker's lemma with $c_i = i^2$ yields

$$\lim_{m\to\infty} m^{-2}\sum_{i=1}^{m} \operatorname{Var}(X_i) = \lim_{m\to\infty} m^{-2}\sum_{i=1}^{m} i^2 \frac{\operatorname{Var}(X_i)}{i^2} = 0,$$

i.e. the first term is zero. For the second term,

$$\begin{aligned}\lim_{m\to\infty} \sum_{i=m+1}^{\infty} \frac{\operatorname{Var}(X_i)}{i^2} &= \lim_{m\to\infty} \left(\sum_{i=1}^{\infty} \frac{\operatorname{Var}(X_i)}{i^2} - \sum_{i=1}^{m} \frac{\operatorname{Var}(X_i)}{i^2}\right) \\ &= \sum_{i=1}^{\infty} \frac{\operatorname{Var}(X_i)}{i^2} - \sum_{i=1}^{\infty} \frac{\operatorname{Var}(X_i)}{i^2} = 0.\end{aligned}$$

Hence $\lim_{m\to\infty} \mathbf{P}\left(\sup_{k\geq m}|S_k| \geq \varepsilon\right) \leq 0$. Since probabilities are nonnegative,

$$\lim_{m\to\infty} \mathbf{P}\left(\sup_{k\geq m}|S_k| \geq \varepsilon\right) = 0.$$

Since $\varepsilon > 0$ was arbitrary, we've shown that $S_n \xrightarrow{\text{a.s.}} 0$. ■

There are several refinements of Chebychev's WLLN and Kolmogorov's SLLN. One of these is Markov's SLLN.

4.2 iid random variables

Kolmogorov's second SLLN features a different mix of restrictions on variances and dependence. Relative to Kolmogorov's first SLLN, we drop the variance restriction. But to make up for this, we impose identical distributions. The combination of independence and identical distribution is usually shortened to 'iid'; it is very common in (micro)econometrics.

Theorem 16 (Kolmogorov's second SLLN)**.** Let $\{X_n\}$ be a sequence of iid random variables. Then $n^{-1}\sum_{i=1}^n X_i \xrightarrow{\text{a.s.}} \mu$ if and only if $\mathbf{E}(X_1)$ exists, is finite and equals μ.

Observe that this LLN gives conditions that are *necessary* as well as sufficient for a.s. convergence! We won't provide a proof; you can find one in Rao (1973, pp. 115–6). An obvious corollary is

Corollary 4 (Khinchine's WLLN)**.** Let $\{X_n\}$ be a sequence of iid random variables such that $\mathbf{E}(X_1)$ exist and is finite. Then $n^{-1}\sum_{i=1}^n X_i \xrightarrow{\text{p}} \mathbf{E}(X_1)$.

4.3 Dependent random variables

Finally, we'll present a substantial refinement of Kolmogorov's first SLLN (p. 54) which weakens the variance condition and requires bounded autocorrelation instead of independence. This theorem is useful for time-series econometrics.

Theorem 17 (Serfling (1970) SLLN)**.** Let $\{X_n\}$ be a sequence of random variables with finite variance. Assume that there exist constants $\{\rho_j\}_{j\in\mathbf{N}}$ in $[0,1]$ such that $\sum_{j=1}^\infty \rho_j < \infty$ and $\text{Corr}(X_n, X_m) \leq \rho_{n-m}$ for all $n \geq m$. Further assume that

$$\sum_{i=1}^\infty \left(\frac{\ln(i)}{i}\right)^2 \text{Var}(X_i) < \infty.$$

Then $n^{-1}\sum_{i=1}^n (X_i - \mathbf{E}(X_i)) \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$.

The variance restriction is very similar to (but weaker than) the one in Kolmogorov's first SLLN. The main novelty comes from the fact that we've replaced independence with bounded autocorrelation. Notice that we only need to rule out large and persistent *positive* autocorrelation. Negative autocorrelation is actually helpful: it speeds up 'mixing', leading to faster convergence!

Again we won't give a proof; you can find one in Serfling (1970, Corollary 2.2.1). But we will give an example to verify that there are interesting sequences of random variables that satisfy the hypotheses of the theorem.

Example 12 (AR(1) model)**.** Let $X_0 = 0$ and $X_n = rX_{n-1} + \varepsilon_n$ for $n \in \mathbf{N}$, where $|r| < 1$ and $\{\varepsilon_n\}$ is a white noise process (iid with zero mean and finite variance). The sequence $\{X_n\}$ is called an AR(1) process.

Iterating backward and using $X_0 = 0$, $X_n = \sum_{j=0}^{n-1} r^j \varepsilon_{n-j}$. It follows that $\mathbf{E}(X_n) = 0$. Moreover, writing $\sigma_\varepsilon^2 \coloneqq \mathbf{E}\left(\varepsilon_n^2\right)$, we have

$$\operatorname{Var}(X_n) = \sum_{j=0}^{n-1} r^{2j} \operatorname{Var}\left(\varepsilon_{n-j}\right) = \sigma_\varepsilon^2 \sum_{j=0}^{n-1} r^{2j} = \frac{1 - r^{2n}}{1 - r^2} \sigma_\varepsilon^2$$

where we used $|r^2| < 1$, which follows from $|r| < 1$. Notice that $\operatorname{Var}(X_n) \geq \operatorname{Var}(X_m)$ whenever $n \geq m$, and that $\operatorname{Var}(X_n) < \sigma_\varepsilon^2 / \left(1 - r^2\right)$ for every n.

For $n \geq m$,

$$\begin{aligned}
\operatorname{Cov}(X_n, X_m) &= \operatorname{Cov}\left(\sum_{j=0}^{n-1} r^j \varepsilon_{n-j}, \sum_{j=0}^{m-1} r^j \varepsilon_{m-j}\right) \\
&= \operatorname{Cov}\left(\sum_{j=1}^{n} r^{n-j} \varepsilon_j, \sum_{j=1}^{m} r^{m-j} \varepsilon_j\right) \\
&= \operatorname{Cov}\left(\sum_{j=1}^{m} r^{n-j} \varepsilon_j, \sum_{j=1}^{m} r^{m-j} \varepsilon_j\right) \\
&= r^{n-m} \operatorname{Var}\left(\sum_{j=1}^{m} r^{m-j} \varepsilon_j\right) \\
&= r^{n-m} \operatorname{Var}\left(\sum_{j=0}^{m-1} r^j \varepsilon_{m-j}\right) \\
&= r^{n-m} \operatorname{Var}\left(X_m\right).
\end{aligned}$$

Since $\text{Var}(X_n) \geq \text{Var}(X_m)$, it follows that

$$\begin{aligned}\text{Corr}(X_n, X_m) &= \frac{\text{Cov}(X_n, X_m)}{\sqrt{\text{Var}(X_m)}\sqrt{\text{Var}(X_n)}} \\ &= r^{n-m}\frac{\sqrt{\text{Var}\,(X_m)}}{\sqrt{\text{Var}(X_n)}} \\ &\leq r^{n-m}.\end{aligned}$$

So we have constants $\{\rho_j\} \coloneqq \{r^j\}$ in $[0,1]$ for which $\sum_{j=1}^{\infty}\rho_j < \infty$ and $\text{Corr}(X_n, X_m) \leq \rho_{n-m}$ for all $n \geq m$, as required.

As for the second condition, since $\text{Var}(X_n) < \sigma_\varepsilon^2/\left(1-r^2\right)$, we obtain

$$\sum_{i=1}^{\infty}\left(\frac{\ln(i)}{i}\right)^2\text{Var}(X_i) \leq \frac{\sigma_\varepsilon^2}{1-r^2}\sum_{i=1}^{\infty}\left(\frac{\ln(i)}{i}\right)^2,$$

which can be shown to converge using the integral test. (Or Wolfram Alpha!)

4.4 Uniform laws of large numbers

In this section, we're interested in laws of large numbers for random functions. In particular, consider an iid sequence $\{g_n\}$ of random functions $\Theta \to \mathbf{R}$, and assume $\mathbf{E}(g_1(\theta)) = 0$ for each $\theta \in \Theta$.[48] Kolmogorov's second SLLN tells us that $n^{-1}\sum_{i=1}^{n} g_i \xrightarrow{\text{a.s.}} 0$ pointwise, i.e. for any $\varepsilon > 0$, there is $\{N_{\varepsilon,\theta}\}_{\theta\in\Theta}$ such that for each $\theta \in \Theta$, $\left|n^{-1}\sum_{i=1}^{n} g_i(\theta)\right| < \varepsilon$ whenever $n \geq N_{\varepsilon,\theta}$.

If Θ is finite, the convergence is automatically uniform: for any $\theta \in \Theta$ you like, $\left|n^{-1}\sum_{i=1}^{n} g_i(\theta)\right| < \varepsilon$ whenever $n \geq \max_{\theta\in\Theta} N_{\varepsilon,\theta}$, where the maximum is attained since Θ is finite. But to obtain uniform a.s. convergence of $\{g_n\}$ without assuming that Θ is finite, we need a new theorem. Such theorems are called uniform laws of large numbers. A useful uniform SLLN for the iid case is the following.

Theorem 18 (Jennrich (1969) uniform SLLN)**.** Let $\{g_n\}$ be an iid sequence of random functions $\Theta \to \mathbf{R}$ with $\mathbf{E}(g_1(\theta)) = 0$ for each $\theta \in \Theta$. Assume that $\Theta \subseteq \mathbf{R}^k$ is compact, that g_1 is continuous a.s., and that $\mathbf{E}\left(\sup_{\theta\in\Theta}|g_1(\theta)|\right) < \infty$. Then

$$n^{-1}\sum_{i=1}^{n} g_i \xrightarrow{\text{a.s.}} 0 \quad \text{uniformly over } \Theta.$$

[48] We really just need the mean to exist and be finite. Setting it to zero is a normalisation, for if the mean is μ then we consider $\widetilde{g}_n(\cdot) \coloneqq g_n(\cdot) - \mu(\cdot)$.

More explicitly, the conclusion of the theorem is that

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \left| n^{-1} \sum_{i=1}^{n} g_i(\theta) \right| = 0 \quad \text{a.s.}$$

The method of proof is called a chaining argument, which is used a lot in empirical process theory.[49] The idea is that by compactness, we can cover Θ with a finite number of open balls

$$\{B_\delta(\theta_j)\}_{j=1}^{J(\delta)}$$

of radius $\delta > 0$. For each ball, the centre $\sum_{i=1}^{n} g_i(\theta_j)$ obeys Kolmogorov's second SLLN, and for other points $\theta \in B_\delta(\theta_j)$ it must be that $\sum_{i=1}^{n} g_i(\theta)$ is close to $\sum_{i=1}^{n} g_i(\theta_j)$ by continuity a.s. We then take $\delta \downarrow 0$.

Proof. We want to show that for any $\varepsilon > 0$,

$$\lim_{N\to\infty} \mathbf{P}\left(\sup_{n\geq N} \sup_{\theta\in\Theta} \left| n^{-1} \sum_{i=1}^{n} g_i(\theta) \right| > \varepsilon \right) = 0.$$

Begin by fixing an arbitrary $\delta > 0$. (Further down, we will choose a particular value $\delta(\varepsilon)$ determined by ε). $\{B_\delta(\theta)\}_{\theta\in\Theta}$ is obviously an open cover of Θ, so by compactness it has a finite subcover $\{B_\delta(\theta_1), \ldots, B_\delta(\theta_{J(\delta)})\}$. Then

$$\{\Theta_j^\delta\}_{j=1}^{J(\delta)} \coloneqq \{\operatorname{cl} B_\delta(\theta_j)\}_{j=1}^{J(\delta)}$$

is a finite cover of Θ,[50] and each Θ_j^δ is compact.

[49] Empirical process theory is the asymptotic theory of certain functions of random data, providing (vast) generalisations of classical asymptotic theory. ('Stochastic process' is another name for a random function.) One topic is uniform LLNs: uniform convergence of partial-average functions such as $G_n(\cdot) = n^{-1} \sum_{i=1}^{n} g_i(\cdot)$ to a nonstochastic limit. Another topic is functional CLTs: weak convergence of scaled partial-average processes such as $G_n(\tau) = n^{-1/2} \sum_{i=1}^{\lfloor \tau n \rfloor} X_i$ to Brownian motion. Yet another topic is extensions of the Glivenko–Cantelli theorem: uniform convergence of empirical measures (analogs of empirical CDFs) to the population probability measure.

[50] $\operatorname{cl} A$ denotes the closure of the set A.

Painfully but straightforwardly, compute

$$
\begin{aligned}
&\mathbf{P}\left(\sup_{n\geq N}\sup_{\theta\in\Theta}\left|n^{-1}\sum_{i=1}^{n}g_i(\theta)\right|>\varepsilon\right)\\
&\leq \mathbf{P}\left(\bigcup_{j=1}^{J(\delta)}\left\{\sup_{n\geq N}\sup_{\theta\in\Theta_j^\delta}\left|n^{-1}\sum_{i=1}^{n}g_i(\theta)\right|>\varepsilon\right\}\right)\\
&\leq \sum_{j=1}^{J(\delta)}\mathbf{P}\left(\sup_{n\geq N}\sup_{\theta\in\Theta_j^\delta}\left|n^{-1}\sum_{i=1}^{n}g_i(\theta)\right|>\varepsilon\right)\\
&\leq \sum_{j=1}^{J(\delta)}\mathbf{P}\left(\sup_{n\geq N}\sup_{\theta\in\Theta_j^\delta}\left(\left|n^{-1}\sum_{i=1}^{n}g_i(\theta_j)\right|+\left|n^{-1}\sum_{i=1}^{n}(g_i(\theta)-g_i(\theta_j))\right|\right)>\varepsilon\right)\\
&\leq \sum_{j=1}^{J(\delta)}\mathbf{P}\left(\sup_{n\geq N}\left|n^{-1}\sum_{i=1}^{n}g_i(\theta_j)\right|+\sup_{n\geq N}\sup_{\theta\in\Theta_j^\delta}\left|n^{-1}\sum_{i=1}^{n}(g_i(\theta)-g_i(\theta_j))\right|>\varepsilon\right)\\
&\leq \sum_{j=1}^{J(\delta)}\mathbf{P}\Bigg(\left\{\sup_{n\geq N}\left|n^{-1}\sum_{i=1}^{n}g_i(\theta_j)\right|>\frac{\varepsilon}{3}\right\}\\
&\qquad\qquad\cup\left\{\sup_{n\geq N}\sup_{\theta\in\Theta_j^\delta}\left|n^{-1}\sum_{i=1}^{n}(g_i(\theta)-g_i(\theta_j))\right|>\frac{2\varepsilon}{3}\right\}\Bigg)\\
&\leq \sum_{j=1}^{J(\delta)}\mathbf{P}\left(\sup_{n\geq N}\left|n^{-1}\sum_{i=1}^{n}g_i(\theta_j)\right|>\frac{\varepsilon}{3}\right)\\
&\quad+\sum_{j=1}^{J(\delta)}\mathbf{P}\left(\sup_{n\geq N}\sup_{\theta\in\Theta_j^\delta}\left|n^{-1}\sum_{i=1}^{n}(g_i(\theta)-g_i(\theta_j))\right|>\frac{2\varepsilon}{3}\right)\\
&\leq \sum_{j=1}^{J(\delta)}\mathbf{P}\left(\sup_{n\geq N}\left|n^{-1}\sum_{i=1}^{n}g_i(\theta_j)\right|>\frac{\varepsilon}{3}\right)\\
&\quad+\sum_{j=1}^{J(\delta)}\mathbf{P}\left(\sup_{n\geq N}\left(n^{-1}\sum_{i=1}^{n}\sup_{\theta\in\Theta_j^\delta}|g_i(\theta)-g_i(\theta_j)|\right)>\frac{2\varepsilon}{3}\right).
\end{aligned}
$$

We'll establish separately that both terms on the RHS vanish as $N\longrightarrow\infty$. For the first term, recall that we assumed $\mathbf{E}\left(\sup_{\theta\in\Theta}|g_1(\theta)|\right)<\infty$; hence a fortiori $\mathbf{E}\left(|g_1(\theta_j)|\right)<\infty$. Since $\{g_n(\theta_j)\}$ are iid with mean zero, Kolmogorov's

second SLLN (p. 56) then implies

$$\lim_{N\to\infty} \mathbf{P}\left(\sup_{n\geq N}\left|n^{-1}\sum_{i=1}^{n} g_i(\theta_j)\right| > \frac{\varepsilon}{3}\right) = 0 \quad \text{for each } j \in \{1,\ldots,J(\delta)\},$$

whence it follows that the first term converges to zero:

$$\lim_{N\to\infty} \sum_{j=1}^{J(\delta)} \mathbf{P}\left(\sup_{n\geq N}\left|n^{-1}\sum_{i=1}^{n} g_i(\theta_j)\right| > \frac{\varepsilon}{3}\right) = 0.$$

Notice that this argument works for any fixed $\delta > 0$.

For the second term, we will need to choose a sufficiently small value of δ to ensure convergence. Observe that by decreasing δ, we can shrink Θ_j^δ enough to ensure that any $\theta \in \Theta_j^\delta$ is arbitrarily close to θ_j. Since g_1 is continuous a.s., we can therefore choose $\delta > 0$ small enough to ensure that $\sup_{\theta\in\Theta_j^\delta}|g_1(\theta) - g_1(\theta_j)|$ is arbitrarily small a.s., i.e.

$$\sup_{\theta\in\Theta_j^\delta} |g_1(\theta) - g_1(\theta_j)| \longrightarrow 0 \quad \text{a.s. as } \delta \downarrow 0.$$

Moreover,

$$\sup_{\theta\in\Theta_j^\delta} |g_1(\theta) - g_1(\theta_j)| \leq 2 \sup_{\theta\in\Theta} |g_1(\theta)|,$$

and the right-hand side has finite expectation by assumption. Hence by the dominated convergence theorem (p. 40),

$$\mu_j^\delta \coloneqq \mathbf{E}\left(\sup_{\theta\in\Theta_j^\delta} |g_1(\theta) - g_1(\theta_j)|\right) \longrightarrow 0 \quad \text{as } \delta \downarrow 0.$$

So there exists a $\delta(\varepsilon) > 0$ such that $\mu_j^{\delta(\varepsilon)} < \varepsilon/3$.

Now,

$$\left\{\sup_{\theta\in\Theta_j^{\delta(\varepsilon)}} |g_n(\theta) - g_n(\theta_j)|\right\}$$

is a sequence of iid random variables with finite mean $\mu_j^{\delta(\varepsilon)}$, so by Kolmogorov's second SLLN (p. 56),

$$\lim_{N\to\infty} \mathbf{P}\left(\sup_{n\geq N}\left(n^{-1}\sum_{i=1}^{n} \sup_{\theta\in\Theta_j^{\delta(\varepsilon)}} |g_i(\theta) - g_i(\theta_j)| - \mu_j^{\delta(\varepsilon)}\right) > \frac{\varepsilon}{3}\right) = 0.$$

Since $\mu_j^{\delta(\varepsilon)} < \varepsilon/3$, it follows that

$$\lim_{N\to\infty} \mathbf{P}\left(\sup_{n\geq N}\left(n^{-1}\sum_{i=1}^{n}\sup_{\theta\in\Theta_j^{\delta(\varepsilon)}}|g_i(\theta)-g_i(\theta_j)|\right)>\frac{2\varepsilon}{3}\right)=0,$$

hence

$$\lim_{N\to\infty}\sum_{j=1}^{J(\delta(\varepsilon))}\mathbf{P}\left(\sup_{n\geq N}\left(n^{-1}\sum_{i=1}^{n}\sup_{\theta\in\Theta_j^{\delta(\varepsilon)}}|g_i(\theta)-g_i(\theta_j)|\right)>\frac{2\varepsilon}{3}\right)=0.$$

Putting this all together, we've shown that for any $\varepsilon > 0$, there is a $\delta(\varepsilon) > 0$ such that

$$\begin{aligned}
&\mathbf{P}\left(\sup_{n\geq N}\sup_{\theta\in\Theta}\left|n^{-1}\sum_{i=1}^{n}g_i(\theta)\right|>\varepsilon\right)\\
&\qquad\leq\sum_{j=1}^{J(\delta(\varepsilon))}\mathbf{P}\left(\sup_{n\geq N}\left(n^{-1}\sum_{i=1}^{n}|g_i(\theta_j)|\right)>\frac{\varepsilon}{3}\right)\\
&\qquad\quad+\sum_{j=1}^{J(\delta(\varepsilon))}\mathbf{P}\left(\sup_{n\geq N}\left(n^{-1}\sum_{i=1}^{n}\sup_{\theta\in\Theta_j^{\delta(\varepsilon)}}|g_i(\theta)-g_i(\theta_j)|\right)>\frac{2\varepsilon}{3}\right)\\
&\qquad\to 0\quad\text{as } N\to\infty.
\end{aligned}$$

(The inequality holds for any $\delta > 0$ you like, and the first term on the RHS vanishes as $N \to \infty$ for any $\delta > 0$ you like, but the second term only vanishes when δ is chosen appropriately.) Since probabilities are nonnegative, it follows that

$$\lim_{N\to\infty}\mathbf{P}\left(\sup_{n\geq N}\sup_{\theta\in\Theta}\left|n^{-1}\sum_{i=1}^{n}g_i(\theta)\right|>\varepsilon\right)=0. \qquad \blacksquare$$

Remark 11. The proof extends without much difficulty to the non-iid case. The monstrous inequality holds regardless of how $\{g_n\}$ are distributed. Convergence of the first term requires only that $\{g_n\}$ obeys a SLLN pointwise. Convergence of the second term requires $\mathbf{E}\left(\sup_{\theta\in\Theta}|g_n(\theta)|\right) < \infty$ for *each* n. As far as I can make out, these are the only tweaks that are required to obtain a uniform SLLN for the non-iid case.

As you'd expect, there are many other uniform LLNs. For the iid case, slightly different assumptions can be used to obtain a uniform SLLN, and somewhat weaker assumptions suffice for a uniform WLLN. As I indicated, uniform SLLNs for the non-iid case are also fairly straightforward.

5 Central limit theorems

Official reading: Amemiya (1985, ch. 3), Rao (1973, ch. 2), Billingsley (1995, sec. 27) and White (2001, ch. 5).

In a finite sample, we'd like to have an idea of how far our (consistent) estimator is likely to be from the truth. The exact distribution of an estimator (across repeated samples) will depend on the unknown distribution of the data, and will anyway be extremely complicated. So we'd like an approximation to its distribution that uses only what the econometrician observes, and which is a good approximation in the sense that it becomes arbitrarily accurate as the sample size increases. This may sound like too much to hope for, but it is in fact possible to do precisely this by using the magic of the central limit theorems (CLTs).

A central limit theorem has the following form. Take any sequence $\{X_n\}$ of random variables that satisfy some conditions; then there are sequences of constants $\{a_n\}$ and $\{b_n\}$ such that

$$a_n \sum_{i=1}^{n} (X_i - b_i) \xrightarrow{\text{d}} \mathcal{N}(0,1),$$

a standard-normal-distributed random variable. The conditions on $\{X_n\}$ are analogous to the ones for LLNs: they restrict the variances and the degree of dependence. Generally, the theorem will include a characterisation of a set of sequences $\{a_n\}$ and $\{b_n\}$ for which the result holds; in the simplest CLTs, $b_i = \mathbf{E}(X_i)$ and $a_n = n^{-1/2}$.[51]

5.1 iid random variables

Theorem 19 (Lindeberg–Lévy CLT)**.** Let $\{X_n\}$ be a sequence of iid random variables with $\mathbf{E}(X_1) = 0$ and $\text{Var}(X_1) = 1$. Then $n^{-1/2} \sum_{i=1}^{n} X_i \xrightarrow{\text{d}} \mathcal{N}(0,1)$.

Remark 12. Setting $\mathbf{E}(X_1) = 0$ and $\text{Var}(X_1) = 1$ is wlog, since for $\mathbf{E}(X_n) = \mu$ and $\text{Var}(X_n) = \sigma^2$, we can apply the theorem to $Y_n \coloneqq (X_n - \mu)/\sigma$. The importance of these restrictions is that the mean exists and is finite and that the variance is finite and nonzero.

[51] There are many generalisations of CLTs that don't quite fit this format. A very important example for econometrics is functional central limit theorems, which give conditions under which the random function $S_n(\tau) \coloneqq n^{-1/2} \sum_{i=1}^{\lfloor \tau n \rfloor} X_i$ converges weakly to a Brownian motion. Another example is 'generalised central limit theorems', which give conditions for convergence to a stable law (see footnote 42 on p. 43).

The proof will make use of characteristic functions. In particular, we'll show that the characteristic function of $n^{-1/2}\sum_{i=1}^{n} X_i$ converges pointwise to $\phi_{\mathcal{N}(0,1)}(t) = \exp\left(-\frac{1}{2}t^2\right)$, then appeal to Lévy's continuity theorem.

Proof. Write

$$Z_n \coloneqq n^{-1/2}\sum_{i=1}^{n} X_i.$$

Fix an arbitrary $t \in \mathbf{R}$.

$$\begin{aligned}
\phi_{Z_n}(t) &= \mathbf{E}\left(\exp\left(itn^{-1/2}\sum_{j=1}^{n} X_j\right)\right) \\
&= \mathbf{E}\left(\prod_{j=1}^{n}\exp\left(itn^{-1/2}X_j\right)\right) \\
&= \prod_{j=1}^{n}\mathbf{E}\left(\exp\left(itn^{-1/2}X_j\right)\right) \\
&= \mathbf{E}\left(\exp\left(itn^{-1/2}X_1\right)\right)^n \\
&= \phi_{X_1}\left(t/n^{1/2}\right)^n.
\end{aligned}$$

where we used independence in the third equality and identical distribution in the fourth.

Since $t/n^{1/2} \longrightarrow 0$ for fixed $t \in \mathbf{R}$, only the behaviour of ϕ_X in a shrinking neighbourhood of 0 will matter. Formally, we use Taylor expansion around 0 to approximate $\phi_{X_1}(t/n^{1/2})$ as n grows large:

$$\phi_{X_1}\left(t/n^{1/2}\right) = \phi_{X_1}(0) + \phi'_{X_1}(0)t/n^{1/2} + \tfrac{1}{2}\phi''_{X_1}(0)t^2/n + \mathrm{o}\left(1/n\right),$$

where the derivatives exist since $\mathbf{E}(X_1)$ and $\mathbf{E}\left(X_1^2\right)$ exist and are finite (see part (5) of Proposition 10 (p. 42)). Again by Proposition 10 (p. 42), we have $\phi_{X_1}(0) = 1$, $\phi'_{X_1}(0) = i^{-1}\mathbf{E}(X_1) = 0$ and $\mathbf{E}\left(X_1^2\right) = i^{-2}\mathrm{Var}(X_1) = i^{-2} = -1$, we can write the Taylor expansion as

$$\phi_{X_1}\left(t/n^{1/2}\right) = 1 - \tfrac{1}{2}t^2/n + \mathrm{o}\left(1/n\right).$$

So using the fact that $\lim_{n\to\infty}(1 + x/n)^n = \exp(x)$, we get

$$\begin{aligned}
\phi_{Z_n}(t) = \phi_{X_1}\left(t/n^{1/2}\right)^n = \left(1 - \tfrac{1}{2}t^2/n + \mathrm{o}\left(1/n\right)\right)^n \\
\longrightarrow \exp\left(-\tfrac{1}{2}t^2\right) = \phi_{\mathcal{N}(0,1)}(t).
\end{aligned}$$

Hence $Z_n \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$ by Lévy's continuity theorem (p. 41). ∎

Before moving on, let's give an (unusual) example of how the Lindeberg–Lévy CLT can be used.

Example 13. Let $\{X_n\}$ be independently distributed $X_n \sim \mathcal{N}(0,1)$, and define $S_n \coloneqq \sum_{i=1}^n X_i^2$. In this case, we don't really need to approximate the distribution of S_n since we know that it is $\chi^2(n)$. But the Lindeberg–Lévy CLT still applies, so let's apply it to get an approximation to the distribution of S_n.

Compute $\mathbf{E}\left(Z_n^2\right) = 1$ and $\operatorname{Var}(Z_n^2) = \mathbf{E}(Z_n^4) - \mathbf{E}(Z_n^2)^2 = 3 - 1 = 2$.[52] Hence by the Lindeberg–Lévy CLT,

$$(2n)^{-1/2}\left(S_n - n\right) = n^{-1/2}\sum_{i=1}^{n}\frac{X_i^2 - 1}{\sqrt{2}} \xrightarrow{\text{d}} \mathcal{N}(0,1).$$

The large-n approximation $(2n)^{-1/2}\left(S_n - n\right) \overset{\text{a}}{\sim} \mathcal{N}(0,1)$ lets us approximate the distribution of S_n for n large as $S_n \overset{\text{a}}{\sim} \mathcal{N}\left(n, 2n\right)$.[53]

In most of our applications, we'll actually have a random vector (not variable) whose distribution we'd like to approximate using a multivariate normal distribution. The univariate Lindeberg–Lévy theorem can be applied to any given linear combination of the elements of our random vector, giving weak convergence of a particular marginal distribution, but it isn't obvious that this is sufficient for convergence of the joint distribution. The Cramér–Wold device tells that it *is* in fact sufficient.

Theorem 20 (Cramér–Wold device)**.** Let $\{X_n\}$ and X be random k-vectors. Then $X_n \xrightarrow{\text{d}} X$ iff $\lambda^\top X_n \xrightarrow{\text{d}} \lambda^\top X$ for every $\lambda \in \mathbf{R}^k$.

Proof. Suppose $X_n \xrightarrow{\text{d}} X$ and fix $\lambda \in \mathbf{R}^k$. $x \mapsto \lambda^\top x$ is a continuous mapping, so by the continuous mapping theorem we get $\lambda^\top X_n \xrightarrow{\text{d}} \lambda^\top X$.

Suppose $\lambda^\top X_n \xrightarrow{\text{d}} \lambda^\top X$ for every $\lambda \in \mathbf{R}^k$. By Lévy's continuity theorem, this implies $\phi_{\lambda^\top X_n}(1) \longrightarrow \phi_{\lambda^\top X}(1)$, i.e.

$$\mathbf{E}\left(\exp\left(i\lambda^\top X_n\right)\right) \longrightarrow \mathbf{E}\left(\exp\left(i\lambda^\top X\right)\right) \quad \text{for every } \lambda \in \mathbf{R}^k.$$

But the LHS equals $\phi_{X_n}(\lambda)$, and the RHS equals $\phi_X(\lambda)$! So we've shown that $\phi_{X_n} \longrightarrow \phi_X$ pointwise, which implies $X_n \xrightarrow{\text{d}} X$ by Lévy's continuity theorem (p. 41). ■

Using the Cramér–Wold device, we can easily extend the Lindeberg–Lévy CLT to random vectors.

[52] It is a fact that the fourth moment of the standard normal distribution is 3.

[53] $\overset{\text{a}}{\sim}$ reads 'is approximately distributed as'.

Corollary 5 (vector Lindeberg–Lévy CLT)**.** Let $\{X_n\}$ be a sequence of iid random k-vectors with $\mathbf{E}(X_1) = 0$ and $\mathrm{Var}(X_1) = I$. Then $n^{-1/2} \sum_{i=1}^{n} X_i \xrightarrow{\mathrm{d}} \mathcal{N}(0, I)$.

Proof. Write

$$Z_n \coloneqq n^{-1/2} \sum_{i=1}^{n} X_i,$$

and let Z denote some (any) k-vector distributed $\mathcal{N}(0, I)$. By a property of the normal distribution, $\lambda^\top Z$ is then distributed univariate $\mathcal{N}(0, \lambda^\top \lambda)$ for any $\lambda \in \mathbf{R}^k$.

By the univariate Lindeberg–Lévy CLT,

$$\lambda^\top Z_n \xrightarrow{\mathrm{d}} \mathcal{N}(0, \lambda^\top \lambda) \overset{\mathrm{d}}{=} \lambda^\top Z \quad \text{for any } \lambda \in \mathbf{R}^k.^{54}$$

Hence $Z_n \xrightarrow{\mathrm{d}} Z \overset{\mathrm{d}}{=} \mathcal{N}(0, I)$ by the Cramér–Wold device. ∎

Remark 13. Again, the importance of the restrictions $\mathbf{E}(X_1) = 0$ and $\mathrm{Var}(X_1) = 1$ is that the mean and covariance matrix exist, that $\mathbf{E}(X_1)$ is finite, and that $\mathrm{Var}(X_1)$ is nonsingular and finite. Nonsingularity of $\mathrm{Var}(X_1)$ is the multivariate analog of our previous requirement that $\mathrm{Var}(X_1) > 0$. It fails iff there is a linear combination of X_1 whose variance is zero (i.e. the distribution is degenerate).

To use the theorem for random vectors $\{X_n\}$ with $\mathbf{E}(X_1) = \mu$ and $\mathrm{Var}(X_1) = V$ nonsingular, just apply it to $Y_n \coloneqq V^{-1/2}(X_n - \mu)$, where $V^{-1/2}$ is the unique Choleski factor of V^{-1}.[55]

The rest of the central limit theorems in this section will be stated for random variables only. But all of them can easily be extended to random vectors using Cramér–Wold device in the manner just demonstrated.

5.2 Independent random variables

The CLTs in this section retain the independence assumption of the Lindeberg–Lévy theorem, but drop the requirement of identical distribution in favour

[54] $\overset{\mathrm{d}}{=}$ denotes equality in distribution, i.e. $X \overset{\mathrm{d}}{=} Y$ iff $\mathcal{L}_X = \mathcal{L}_Y$.

[55] A Choleski decomposition of a matrix A is $A = A^{1/2}(A^{1/2})^\top$ where $A^{1/2}$ is lower-triangular with positive diagonal entries. To show that it exists and is unique for V^{-1}, reason as follows. V is real, symmetric and positive semidefinite (p.s.d.) since it's a covariance matrix. A real, symmetric and p.s.d. matrix is nonsingular iff it is positive definite (p.d.), so V is p.d. The inverse of a p.d. matrix is p.d., so V^{-1} is p.d. A matrix has a unique Choleski decomposition iff it is p.d., so V^{-1} has a unique Choleski decomposition.

of restrictions on the variances. (There's a similarity here with Kolmogorov's first SLLN, which also imposes independence and a variance restriction. Continuing the analogy, the Lindeberg–Lévy CLT and Kolmogorov's second SLLN both impose iid.)

The most general theorem along these lines is the following.

Theorem 21 (Lindeberg–Feller CLT)**.** Let $\{X_n\}$ be a sequence of independent random variables with $\mathbf{E}(X_n)$ finite and $0 < \operatorname{Var}(X_n) < \infty$. Write $c_n \coloneqq (\sum_{i=1}^n \operatorname{Var}(X_i))^{1/2}$. Then

$$\lim_{n\to\infty} \max_{i\in[1,n]} \frac{\operatorname{Var}(X_i)}{c_n^2} = 0 \quad \text{and} \quad c_n^{-1} \sum_{i=1}^n (X_i - \mathbf{E}(X_i)) \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$$

hold iff

$$\lim_{n\to\infty} c_n^{-2} \sum_{i=1}^n \mathbf{E}\left[(X_i - \mathbf{E}(X_i))^2 \mathbf{1}\left(|X_i - \mathbf{E}(X_i)| > \varepsilon c_n\right)\right] = 0 \quad \text{for any } \varepsilon > 0.$$

Remark 14. The last condition is called the Lindeberg condition; it restricts the thickness of the tails. The 'only if' part means that the Lindeberg condition is the weakest possible sufficient condition for weak convergence to a normal law when independence and

$$\lim_{n\to\infty} \max_{i\in[1,n]} \operatorname{Var}(X_i) \Big/ c_n^2 = 0$$

hold! But unfortunately, the Lindeberg condition is usually difficult to check.

To deal with the intractability of the Lindeberg condition, we can use the stronger but simpler Liapunov condition.

Theorem 22 (Liapunov CLT)**.** Let $\{X_n\}$ be a sequence of independent random variables with $\mathbf{E}(X_n)$ finite, $0 < \operatorname{Var}(X_n) < \infty$, and

$$\tau_n \coloneqq \mathbf{E}\left(|X_n - \mathbf{E}(X_n)|^3\right) < \infty.$$

Write $c_n \coloneqq (\sum_{i=1}^n \operatorname{Var}(X_i))^{1/2}$ and $b_n \coloneqq (\sum_{i=1}^n \tau_i)^{1/3}$, and assume that $\lim_{n\to\infty} b_n/c_n = 0$. Then

$$c_n^{-1} \sum_{i=1}^n (X_i - \mathbf{E}(X_i)) \xrightarrow{\mathrm{d}} \mathcal{N}(0,1).$$

Proof. Write $Y_n \coloneqq X_n - \mathbf{E}(X_n)$. We wish to show that the Lindeberg condition holds, so fix $\varepsilon > 0$.

$$\begin{aligned} c_n^{-2} \sum_{i=1}^n \mathbf{E}\left[Y_i^2 \mathbf{1}\left(|Y_i| > \varepsilon c_n\right)\right] &= c_n^{-2} \sum_{i=1}^n \int_{\{|y| > \varepsilon c_n\}} y^2 \mathcal{L}_{Y_i}(\mathrm{d}y) \\ &= c_n^{-2} \sum_{i=1}^n \int_{\{|y| > \varepsilon c_n\}} \frac{1}{|y|} |y|^3 \mathcal{L}_{Y_i}(\mathrm{d}y). \end{aligned}$$

Observe that this expression must be nonnegative. To show that it can't be strictly positive, use the nonnegativity of the integrand (together with $c_n > 0$ and $\tau_n < \infty$) to obtain

$$\begin{aligned} c_n^{-2} \sum_{i=1}^n \int_{\{|y| > \varepsilon c_n\}} \frac{1}{|y|} |y|^3 \mathcal{L}_{Y_i}(\mathrm{d}y) &\leq \frac{1}{\varepsilon c_n^3} \sum_{i=1}^n \int_{\{|y| > \varepsilon c_n\}} |y|^3 \mathcal{L}_{Y_i}(\mathrm{d}y) \\ &\leq \frac{1}{\varepsilon c_n^3} \sum_{i=1}^n \tau_i \\ &= \frac{1}{\varepsilon} \left(\frac{b_n}{c_n}\right)^3 \longrightarrow 0 \quad \text{as } n \longrightarrow \infty. \end{aligned}$$

Since $\varepsilon > 0$ was arbitrary, we've shown that the Lindeberg condition holds. Hence the conclusion follows by the Lindeberg–Feller CLT. ∎

Remark 15. It's clear from the proof that we don't really need the third moment to exist; it's enough for the $(2 + \alpha)$th moment to exist for some $\alpha > 0$.

Before moving on, let's give an example of how these theorems are used in econometrics.

Example 14. In the linear model,

$$n^{1/2}(\widehat{\beta} - \beta) = \left(n^{-1} X^\top X\right)^{-1} \left(n^{-1/2} X^\top \varepsilon\right).$$

We assume independent observations, and impose conditions s.t. $n^{-1} X^\top X$ obeys a WLLN, converging in probability to a nonsingular constant matrix A. Then if $n^{-1/2} X^\top \varepsilon$ converges in distribution to $\mathcal{N}(0, \Sigma)$, Slutsky's theorem implies that

$$n^{1/2}(\widehat{\beta} - \beta) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, A^{-1} \Sigma \left(A^{-1}\right)^\top\right),$$

giving us the useful approximation

$$\widehat{\beta} \overset{\mathrm{a}}{\sim} \mathcal{N}\left(\beta, n^{-1} A^{-1} \Sigma \left(A^{-1}\right)^\top\right).$$

Showing that $n^{-1/2}X^\top\varepsilon \xrightarrow{\mathrm{d}} \mathcal{N}(0,\Sigma)$ is easy when $\{X_i\}$ and $\{\varepsilon_i\}$ are iid and independent of each other, for then $\{X_i\varepsilon_i\}$ are iid random vectors and the vector Lindeberg–Lévy CLT can be applied.

But suppose that we want the asymptotic distribution conditional on X, or equivalently that X is nonstochastic ('fixed regressors') with $n^{-1}X^\top X \longrightarrow A$ for some nonsingular A. In this case, more work is required to show that $n^{-1/2}X^\top\varepsilon$ converges in distribution. The observations are still independent since $\{\varepsilon_i\}$ are, but they are no longer identically distributed, since each term in the sum is a different linear combination of the elements of ε. We therefore need to make assumptions sufficient for the Lindeberg condition to be satisfied; loosely, we require that X is sufficiently bounded that no single observation can dominate the variance of the sum. As mentioned above, it is rather hard to check the Lindeberg condition, but in this case the demonstration can be found in Amemiya (1985).

5.3 Dependent random variables

A sequence of random elements is also called a (discrete-time) stochastic process. In this section, we're mainly thinking about time-series applications in which n is a time index, so we'll use the language of stochastic processes. We'll actually consider stochastic processes with no starting date, i.e. sequences $\{X_n\}_{-\infty}^{\infty}$.

'Time-series CLTs' do away with independence, which tends to make things a lot uglier. Many of them also allow for some degree of heterogeneity of distribution. The CLT we will state (which is one of many) replaces replaces identical distribution with strict stationarity and replaces independence with α-mixing.

Definition 29. A stochastic process $\{X_n\}$ is strictly stationary iff for any finite collection of indices $(n_1,\dots,n_T)$, the (joint) distribution of the random vector $(X_{n_1+m},\dots,X_{n_T+m})$ does not depend on $m \in \mathbf{N}$.

Definition 30. A strictly stationary stochastic process $\{X_n\}$ is α-mixing (or strongly mixing) iff there is $\alpha : \mathbf{N} \to \mathbf{R}$ such that $\lim_{k\to\infty}\alpha(k) = 0$ and

$$\sup\Big\{|\mathbf{P}(B\cap C) - \mathbf{P}(B)\mathbf{P}(C)| : B \in \sigma(\dots,X_{n-1},X_n), C \in \sigma(X_{n+k},X_{n+k+1},\dots), n \in (-\infty,\infty)\Big\} \leq \alpha(k) \quad \text{for each } k \in \mathbf{N}.$$

The object that we're taking the supremum of is the degree of 'independence failure'. Since we're taking the supremum (over a very large set), the condition says that the degree of independence failure between 'blocks' k periods apart is uniformly bounded by $\alpha(k)$ for each k. Since $\alpha(k) \longrightarrow 0$, the degree of independence failure must vanish uniformly as blocks are pulled further apart.

The following is one of many 'time-series CLTs'. Joel said that it can be found in White (2001), though I haven't been able to locate it.

Theorem 23 (α-mixing CLT)**.** Let $\{X_n\}$ be a real-valued, strictly stationary process with mean zero. Assume that $\mathbf{E}\left(|X_n|^\gamma\right) < \infty$ for some $\gamma > 2$, and that the process is α-mixing with $\alpha(k) = ak^{-\beta}$ for $a > 0$ and $\beta > \gamma/(\gamma - 2)$. Then

$$\left(\sum_{k=-\infty}^{\infty} \operatorname{Cov}(X_0, X_k)\right)^{-1} \sum_{i=1}^{n} X_i \xrightarrow{\mathrm{d}} \mathcal{N}(0,1).$$

Remark 16. There is an explicit tradeoff here between how heavy tails and how much dependence that can be accommodated: for γ small (heavy tails), β must be large (low dependence).

We (definitely) won't give a proof, but here's a rough indication as to why it's true. We know from the proof of the Lindeberg–Lévy CLT that independence allows us to factor the characteristic function into the product of n characteristic functions. While this no longer holds exactly without independence, α-mixing is sufficient for it to remain a good approximation in the sense that the approximation error vanishes sufficiently fast as $n \to \infty$. This is because for 'blocks' of random variables very far apart, independence 'nearly' holds, and as $n \longrightarrow \infty$ there are many blocks that are very far apart. Formalising this is a delicate business, however!

5.4 The rate of convergence

Central limit theorems tell us that a normalised sum of random variables converges weakly to $\mathcal{N}(0,1)$, but they do no tell us the rate of convergence. If convergence is really slow, CLTs won't provide very good approximations!

Ideally, we'd like a uniform bound on the approximation error from using the standard normal CDF Φ instead of the true (unknown) CDF. The Berry–Esseen theorem does exactly this for the iid case. Its assumptions are exactly those of the Lindeberg–Lévy CLT, except that the existence of the third moment is assumed.

Theorem 24 (Berry–Esseen)**.** Let $\{X_n\}$ be a sequence of iid random variables with $\mathbf{E}(X_1) = 0$ and $\mathrm{Var}(X_1) = 1$ whose third moment $\mathbf{E}(X_1^3)$ exists. Then for every $n \in \mathbf{R}$,

$$\sup_{x \in \mathbf{R}} \left| \mathbf{P}\left(n^{-1/2} \sum_{i=1}^{n} X_i \leq x \right) - \Phi(x) \right| \leq 3n^{-1/2} \mathbf{E}\left(X_1^3\right).$$

There's a cottage industry in probability theory devoted to refining this theorem. On the one hand, it's possible to replace the 3 with some other constant which may be smaller. On the other hand, probabilists have extended the Berry–Esseen theorem to non-iid sequences.

6 Some more limit theory

6.1 Connections between CLTs and LLNs

In this little section, we'll show that whenever a CLT-type property holds for some scaling constants $\{a_n\}$ that don't shrink too fast, a WLLN follows. We'll also show that the converse is false, and that we cannot strengthen the result to obtain a SLLN.

Suppose that random variables $\{X_n\}$ satisfy a central limit theorem for some constants constants $\{a_n\}$ and $\{b_n\}$:

$$a_n \sum_{i=1}^{n} (X_i - b_i) \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1).$$

(For concreteness, you can think of $a_n = n^{-1/2}$ and $b_n = \mathbf{E}(X_n)$.) Then

$$n^{-1} \sum_{i=1}^{n} (X_i - b_i) = \mathrm{O_p}(1/na_n).$$

If $1/na_n = \mathrm{o_p}(1)$, it follows that $\{X_n\}$ satisfy a weak law of large numbers:

$$n^{-1} \sum_{i=1}^{n} (X_i - b_i) = \mathrm{O_p}(\mathrm{o_p}(1)) = \mathrm{o_p}(1).$$

In particular, this implication holds if $a_n = n^{-1/2}$ as in e.g. the Lindeberg–Lévy CLT. (The argument goes through if $a_n \sum_{i=1}^{n} (X_i - b_i)$ converges weakly to any proper distribution, even if it's nonnormal.)

The converse is false: a CLT result does not follow from a WLLN property. For example, consider $\{X_n\}$ and X with $X = X_n = 0$ a.s. Then $X_n \xrightarrow{\text{a.s.}} X$,

so a fortiori $X_n \xrightarrow{\text{p}} X$, but $a_n \sum_{i=1}^n X_i = 0$ a.s. for *any* sequence of constants $\{a_n\}$. So we cannot obtain a CLT-type result no matter how we choose our scaling constants.

A CLT property does *not* imply a SLLN property, however. I haven't been able to think up a counterexample to illustrate this, sadly. It's not super-intuitive!

6.2 Laws of the iterated logarithm

My understanding of this topic is poor! Proceed with caution.

Laws of the iterated logarithm (LILs) operate 'between' LLNs and CLTs. Consider a sequence of iid random variables with $\mathbf{E}(X_1) = 0$ and $\text{Var}(X_1) = 1$. Kolmogorov's second SLLN and the Lindeberg–Lévy CLT tell us how the partial sums $\{\sum_{i=1}^n X_i\}$ behave when scaled (respectively) by $\text{O}(n)$ and $\text{O}(n^{1/2})$:

$$n^{-1} \sum_{i=1}^n X_i \xrightarrow{\text{a.s.}} 0 \quad \text{and} \quad n^{-1/2} \sum_{i=1}^n X_i \xrightarrow{\text{d}} \mathcal{N}(0,1).^{56}$$

But what happens if we use scaling constants that increase at a rate slower than $\text{O}(n)$ but faster than $\text{O}(n^{1/2})$? In particular, how much slower than $\text{O}(n)$ can we make rate while keeping almost every convergent subsequence of $\{\sum_{i=1}^n X_i\}$ bounded? The Hartman–Wintner LIL says that the answer is $\text{O}([n \ln(\ln(n))]^{1/2})$. (For *any* sequence $\{X_n\}$!)

Theorem 25 (Hartman–Wintner LIL)**.** Let $\{X_n\}$ be a sequence of iid random variables with $\mathbf{E}(X_1) = 0$ and $\text{Var}(X_1) = 1$. Then

$$\limsup_{n\to\infty} [n \ln(\ln(n))]^{-1/2} \sum_{i=1}^n X_i = \sqrt{2} \quad \text{a.s.}$$

Remark 17. The conclusion of the theorem is equivalent to

$$\liminf_{n\to\infty} [n \ln(\ln(n))]^{-1/2} \sum_{i=1}^n X_i = -\sqrt{2} \quad \text{a.s.}$$

To see this, just replace $\{X_n\}$ with $\{-X_n\}$.

[56] If you know Skorokhod's theorem, it will be more helpful to consider that $n^{-1/2} \sum_{i=1}^n X_i \xrightarrow{\text{a.s.}} X$ for some $X \sim \mathcal{N}(0,1)$ (possibly on a different probability space).

Remark 18. A consequence of this LIL is that

$$[n \ln(\ln(n))]^{-1/2} \sum_{i=1}^{n} X_i$$

is bounded a.s. It is therefore obviously bounded in probability:

$$\sum_{i=1}^{n} X_i = \mathrm{O_p}\left([n \ln(\ln(n))]^{-1/2}\right).$$

But observe that our LIL gives us more: it tells us not just the rate of increase, but also the constant of proportionality ($\sqrt{2}$)!

There's a nice equivalent statement in terms of tail events. For events $\{A_n\}$, the event 'A_n occurs infinitely often (i.o.)' is

$$\{A_n \text{ i.o.}\} \coloneqq \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m.$$

An intuitive way of putting this into words is that $\{A_n \text{ i.o.}\}$ obtains iff all but finitely many of the events $\{A_n\}$ occur. (This follows from the deMorgan law; see e.g. Rosenthal (2006, sec. 3.4)).

It turns out (see e.g. Billingsley (1995, pp. 154–6)) that the Hartman–Wintner LIL is equivalent to the following.

Corollary 6. Let $\{X_n\}$ be a sequence of iid random variables with $\mathbf{E}(X_1) = 0$ and $\mathrm{Var}(X_1) = 1$. Then for every $\varepsilon > 0$,

$$\mathbf{P}\left(\sum_{i=1}^{n} X_i \geq (1+\varepsilon)\sqrt{2n \ln(\ln(n))} \quad \text{i.o.}\right) = 0 \quad \text{and}$$
$$\mathbf{P}\left(\sum_{i=1}^{n} X_i \geq (1-\varepsilon)\sqrt{2n \ln(\ln(n))} \quad \text{i.o.}\right) = 1.$$

In words, this reformulated LIL says that $\sum_{i=1}^{n} X_i$ lies in

$$\pm(1-\varepsilon)\sqrt{2n \ln(\ln(n))}$$

infinitely often (with probability 1), and lies outside

$$\pm(1+\varepsilon)\sqrt{2n \ln(\ln(n))}$$

only finitely many times (with probability 1). The LIL therefore bounds the extreme fluctuations of $\{\sum_{i=1}^{n} X_i\}$: fluctuations inside the iterated-log bound

occur infinitely often w.p. 1, and fluctuations big enough to jump outside the iterated-log bound occur at most finitely many times w.p. 1.

It may look like the LIL gives us a usable 100% confidence interval for $\sum_{i=1}^{n} X_i$, but this is not really the case. The LIL says that along $n \in \mathbf{N}$, $\sum_{i=1}^{n} X_i$ lies in

$$\pm(1-\varepsilon)\sqrt{2n\ln(\ln(n))}$$

infinitely many times with probability 1. It doesn't say anything about the probability that $\sum_{i=1}^{n} X_i$ lies in this interval for any *given* (possibly large) n, though!

As with LLNs and CLTs, LILs are also available for independent but not identically distributed random variables, as well as for dependent random variables. Some of these can be found in Serfling (1980, sec. 1.10).

7 Asymptotic properties of extremum estimators

Official reading: Amemiya (1985, sec. 4.1.1–4.1.2) and Newey and McFadden (1994, sec. 2–3).

7.1 Preliminaries

The setting is as follows. There is a $\mathbf{R}^r$-valued stochastic process $\{y_n\}$ defined on a probability space $(\Omega, \mathcal{A}, \mathbf{P})$.[57] We call this stochastic process the data-generating process (DGP), and write μ_0 for its (unknown) law.[58] A dataset of size n is a realisation $\{y_i(\omega)\}_{i=1}^n$ of the first n coordinates of the DGP.

We wish to use a dataset to learn about the law μ_0 of the data-generating process ('the distribution of the data'). In particular, we want to learn about (estimate) a parameter of the law. Formally, a parameter is a mapping $\tau : M \to \Theta$, where M is a set of probability measures to which μ_0 is assumed to belong, and Θ is called the parameter space. Intuitively, τ captures some 'aspect' of the DGP's distribution.[59] Nonparametric econometrics is concerned with the case in which Θ is infinite-dimensional (e.g. a function space). We will focus on parameteric econometrics, meaning that we will be concerned with the finite-dimensional case $\Theta \subseteq \mathbf{R}^k$ for some $k \in \mathbf{N}$.

When studying extremum estimators, we will not say anything about the shape of the map $\tau : M \to \Theta$. Instead, we will study maximisers of a dataset-dependent function of $\theta \in \Theta$. (So θ is properly called a parameter *value*. It is not a parameter.) When we study consistency, we will define a $\theta_0 \in \Theta$ to which the maximisers converge in probability/a.s. as the size of the dataset grows. We leave for the applied researcher the task of establishing that θ_0 as defined below is in fact equal to $\tau(\mu_0)$ for the parameter τ that she wishes to estimate.

An estimator is a mapping from datasets (of arbitrary size) into Θ. As the language suggests, the idea is usually that (for large datasets), the estimator will be close to the true value $\tau(\mu_0)$ of some interesting parameter τ. But don't let the lingo confuse you: an estimator is just a mapping from datasets into Θ, which may or may not be useful for learning about some parameter τ.

[57] Reminder: a (discrete-time) stochastic process is a collection $\{y_n\}_{n \in \mathbf{N}}$ of random variables defined on some (common) probability space.

[58] A stochastic process (taken as a whole) is a random element of a sequence space, so the law of the process is just the law of this random element.

[59] Simple example: if the DGP is $\mathbf{R}$-valued and iid with marginal distribution μ_0^1, then the mean (provided it exists) is a parameter: $\tau(\mu_0) \coloneqq \int_{\mathbf{R}} x \mu_0^1(\mathrm{d}x)$.

An extremum estimator is an estimator constructed by maximising a data-dependent criterion function. Formally, it is a family of mappings $\widetilde{\theta}_n : \mathbf{R}^{n\times r} \to \Theta$, one for each sample size $n \in \mathbf{N}$, such that

$$\widetilde{\theta}_n\left(\{y_i\}_{i=1}^n\right) \in \arg\max_{\theta\in\Theta} \widetilde{Q}_n\left(\{y_i\}_{i=1}^n, \theta\right)$$

for some family $\left\{\widetilde{Q}_n\right\}_{n\in\mathbf{N}}$ of criterion functions $\mathbf{R}^{n\times r} \times \Theta \to \mathbf{R}$. It turns out that the vast majority of estimators in econometrics are extremum estimators.

Example 15 (common extremum estimators)**.** The ordinary least squares, nonlinear least squares, least absolute deviations and maximum likelihood estimators can be written as extremum estimators:

$$\begin{aligned}
\widetilde{\beta}_n^{\mathrm{OLS}}\left(\{y_i, x_i\}_{i=1}^n\right) &\coloneqq \arg\min_{\beta\in\mathbf{R}^k} \sum_{i=1}^n \left(y_i - x_i^\top \beta\right)^2 \\
\widetilde{\beta}_n^{\mathrm{NLS}}\left(\{y_i, x_i\}_{i=1}^n\right) &\coloneqq \arg\min_{\beta\in\mathbf{R}^k} \sum_{i=1}^n \left(y_i - f(x_i, \beta)\right)^2 \\
\widetilde{\beta}_n^{\mathrm{LAD}}\left(\{y_i, x_i\}_{i=1}^n\right) &\coloneqq \arg\min_{\beta\in\mathbf{R}^k} \sum_{i=1}^n |y_i - x_i^\top \beta|. \\
\widetilde{\beta}_n^{\mathrm{ML}}\left(\{y_i, x_i\}_{i=1}^n\right) &\coloneqq \arg\max_{\beta\in\mathbf{R}^k} \mathcal{L}\left(\{y_i\}_{i=1}^n, \beta\right)
\end{aligned}$$

The generalised-method-of-moments estimator is also in the extremum class.

Notice that several of these models are conditional models, i.e. they concern a regression of y_i on x_i (the mean regression in the case of OLS, the median regression for LAD). As the example makes clear, there is nothing special about dependent and independent variables: they're all just part of the dataset.

7.2 Measurability

We've defined the criterion function and extremum estimator as deterministic functions of the data. But for the purposes of asymptotic theory, we'd like to be able to treat them as a random function and a random vector, respectively. (Otherwise concepts like convergence in probability are not defined!) To this end, redefine the criterion function and extremum estimator as mappings directly from the underlying probability space $(\Omega, \mathcal{A}, \mathbf{P})$: for each $n \in \mathbf{N}$ and $\omega \in \Omega$,

$$Q_n(\omega)(\cdot) \coloneqq \widetilde{Q}_n\left(\{y_i(\omega)\}_{i=1}^n, \cdot\right) \quad \text{and} \quad \widehat{\theta}_n(\omega) \in \arg\max_{\theta\in\Theta} Q_n(\omega)(\theta).$$

The necessary and sufficient condition for Q_n to be measurable (hence a random function) is easy: we require $\widetilde{Q}_n$ to be measurable in its first argument w.r.t. your desired σ-algebras on $\mathbf{R}^{n\times r}$ and on the space of functions $\mathbf{R}^{n\times r} \to \Theta$. But the existence of a measurable selection $\widehat{\theta}_n$ from the argmax is not so obvious. The following gives one set of sufficient conditions.

Proposition 15. Suppose that $\widetilde{Q}_n(\cdot, \theta)$ is measurable for each $\theta \in \Theta$, that $\widetilde{Q}_n\left(\{y_i\}_{i=1}^n, \cdot\right)$ is continuous for each $y_n \in \mathbf{R}^{n\times r}$, and that Θ is compact. Then the argmax correspondence $G(\cdot) \coloneqq \arg\max_{\theta\in\Theta} Q_n(\cdot)(\theta)$ admits a measurable selection $\widehat{\theta}_n$.

Remark 19. The main role of continuity and compactness it to ensure that G is nonempty-valued; otherwise G may not admit any selection, measurable or not.

The result is a corollary of the following lemma, which I've adapted from Aliprantis and Border (2006, Theorem 18.19).[60]

Lemma 4 (measurable maximum lemma)**.** Let $(\Omega, \mathcal{A})$ be a measurable space, and let $X \subseteq \mathbf{R}^k$ be compact. Let f be a function $\Omega \times X \to \mathbf{R}$ such that $f(\cdot, x)$ is measurable for each $x \in X$ and $f(\omega, \cdot)$ is continuous for each $\omega \in \Omega$. Define $v : \Omega \to \mathbf{R}$ and $G : \Omega \rightrightarrows X$ by

$$v(\cdot) \coloneqq \max_{x\in X} f(\cdot, x) \quad \text{and} \quad G(\cdot) \coloneqq \arg\max_{x\in X} f(\cdot, x).$$

Then v is $\mathcal{A}/\mathcal{B}_{\mathbf{R}}$-measurable and G admits a $\mathcal{A}/\mathcal{B}_X$-measurable selection, where $\mathcal{B}$ denotes the respective Borel σ-algebras.

Proof. Take any nonempty $X' \subseteq X$. Define $F_{X'}(\cdot) \coloneqq \sup_{x\in X'} f(\cdot, x)$. Then for any $c \in \mathbf{R}$,

$$\begin{aligned}
\{\omega \in \Omega : F_{X'}(\omega) \le c\} &= \left\{\omega \in \Omega : \sup_{x\in X'} f(\omega, x) \le c\right\} \\
&= \{\omega \in \Omega : f(\omega, x) \le c \ \ \forall x \in X'\} \\
&= \bigcap_{x\in X'} \{\omega \in \Omega : f(\omega, x) \le c\} \\
&= \bigcap_{x\in X'\cap \mathbf{Q}^k} \{\omega \in \Omega : f(\omega, x) \le c\} \\
&\in \mathcal{A}
\end{aligned}$$

[60] For now, we only require the second part (the measurability of the argmax). But later on we'll want to use the maximised value of the criterion function as (part of) a test statistic, and a test statistic had better be a random variable!

where the final equality holds since $f(\omega, \cdot)$ is continuous and $\mathbf{Q}^k$ is dense in $\mathbf{R}^k$, and inclusion in $\mathcal{A}$ holds because $f(\cdot, x)$ is measurable and σ-algebras are closed under countable intersection. So $F_{X'}$ is $\mathcal{A}/\mathcal{B}_{\mathbf{R}}$-measurable, no matter what (nonempty) $X' \subseteq X$ you choose. Letting $X' = X$, it follows that $v = F_X$ is $\mathcal{A}/\mathcal{B}_{\mathbf{R}}$-measurable.

Next, we want to show that some selection $g : \Omega \to X$ from G is measurable. Since $X \subseteq \mathbf{R}^k$, this requires precisely that $\{\omega \in \Omega : g(\omega) \leq c\} \in \mathcal{A}$ for every $c \in \mathbf{R}^k$. Write $C \coloneqq \{x \in X : x \leq c\}$. If $C = X$ or $C = \varnothing$ then the result follows trivially, so let c be such that neither C nor $X \backslash C$ is empty. Then

$$\begin{aligned}
\{\omega \in \Omega : g(\omega) \leq c\} &= \left\{\omega \in \Omega : \max_{x \in C} f(\omega, x) \geq \sup_{x \in X \backslash C} f(\omega, x)\right\} \\
&= \left\{\omega \in \Omega : \sup_{x \in C} f(\omega, x) \geq \sup_{x \in X \backslash C} f(\omega, x)\right\} \\
&= \left\{\omega \in \Omega : F_C(\omega) - F_{X \backslash C}(\omega) \geq 0\right\} \\
&\in \mathcal{A}
\end{aligned}$$

where the inclusion follows from the fact that F_C and $F_{X \backslash C}$ are both measurable and that the difference of measurable functions is measurable. ■

From this point on, we will always treat Q_n and $\widehat{\theta}_n$ as random elements, without specifying particular primitive assumptions that guarantee measurability. If you like concreteness, maintain the sufficient conditions given above.[61]

7.3 Consistency

We say that an extremum estimator $\widehat{\theta}_n$ is weakly consistent for $\theta_0 \in \theta$ iff it converges in probability θ_0. We say that it is strongly consistent iff the convergence is almost sure. As mentioned above, this section will give conditions under which extremum estimators are consistent for a θ_0 that we will define from the criterion functions. There is no reason why θ_0 should be equal to the true value of some interesting parameter of the DGP's law μ_0!

Proposition 16 (weak consistency)**.** Assume

[61] Aside: when we move into more complicated econometric problems, measurability sometimes becomes prohibitively difficult to verify. As a result, much of empirical process theory has abandoned ordinary probability measures in favour of outer measure.

(1) $\Theta \subseteq \mathbf{R}^k$ is compact.

(2) Q_n is continuous for each $n \in \mathbf{N}$.

(3) $n^{-1}Q_n \xrightarrow{\text{p}} Q$ uniformly over Θ for some nonstochastic $Q : \Theta \to \mathbf{R}$.

(4) Q has a unique maximum on Θ at θ_0.

Then $\widehat{\theta}_n \xrightarrow{\text{p}} \theta_0$.

Remark 20. Three things.

(1) It is possible to relax compactness by requiring that Q be 'sufficiently concave'; then the argmax eventually lies in some compact set with high probability.

(2) Condition (3) is a high-level assumption. Later on, we will see how a uniform law of large numbers can be used to derive (3) from more primitive conditions on Q_n and the distribution of the data.

(3) Assumption (4) does two things. First, it assumes identification: if there were multiple maxima, we'd have partial identification. Second, it *defines* θ_0.

Proof. To establish convergence in probability, we have to show that $\mathbf{P}(\widehat{\theta}_n \in S) \longrightarrow 1$ as $n \longrightarrow \infty$ for any open neighbourhood S of θ_0. So fix an arbitrary open neighbourhood S in $\mathbf{R}^k$ of θ_0. Since Θ is compact and S is open, $S^{\text{C}} \cap \Theta$ is compact. Since each Q_n is continuous, Q is continuous. Hence $\max_{\theta \in \Theta \cap S^{\text{C}}} Q(\theta)$ exists by the Weierstrass theorem, so we can define

$$\varepsilon \coloneqq Q(\theta_0) - \max_{\theta \in \Theta \cap S^{\text{C}}} Q(\theta). \tag{3}$$

Let A_n be the event

$$A_n \coloneqq \left\{ \sup_{\theta \in \Theta} \left| n^{-1}Q_n(\theta) - Q(\theta) \right| < \varepsilon/2 \right\}.$$

Notice that $\mathbf{P}(A_n) \longrightarrow 1$ as $n \longrightarrow \infty$ by assumption (3).

Since $\widehat{\theta}_n$ and θ_0 lie in Θ, A_n implies

$$Q(\widehat{\theta}_n) > n^{-1}Q_n(\widehat{\theta}_n) - \varepsilon/2 \tag{4}$$

and

$$n^{-1}Q_n(\theta_0) > Q(\theta_0) - \varepsilon/2. \tag{5}$$

Using $Q_n(\widehat{\theta}_n) \geq Q_n(\theta_0)$, (4) implies

$$Q(\widehat{\theta}_n) > n^{-1}Q_n(\theta_0) - \varepsilon/2. \tag{6}$$

Adding (5) and (6) and cancelling $n^{-1}Q_n(\theta_0)$, we see that A_n implies

$$Q(\widehat{\theta}_n) > Q(\theta_0) - \varepsilon = \max_{\theta \in \Theta \cap S^{\mathrm{C}}} Q(\theta)$$

by the definition (3) of ε. It follows that $\widehat{\theta}_n \notin \Theta \cap S^{\mathrm{C}}$, so $\widehat{\theta}_n \in S$.

So A_n implies $\widehat{\theta}_n \in S$, hence $\mathbf{P}(A_n) \leq \mathbf{P}(\widehat{\theta}_n \in S)$. Since $\mathbf{P}(A_n) \longrightarrow 1$, it follows that $\mathbf{P}(\widehat{\theta}_n \in S) \longrightarrow 1$. Since S was an arbitrarily chosen neighbourhood of θ_0, this establishes that $\widehat{\theta}_n \xrightarrow{\mathrm{p}} \theta_0$. ∎

Perhaps unsurprisingly, strengthening assumption (3) to require uniform *almost sure* convergence of $n^{-1}Q_n$ yields strong consistency.

Proposition 17 (strong consistency)**.** Assume

(1) $\Theta \subseteq \mathbf{R}^k$ is compact.

(2) Q_n is continuous for each $n \in \mathbf{N}$.

(3) $n^{-1}Q_n \xrightarrow{\mathrm{a.s.}} Q$ uniformly over Θ for some nonstochastic $Q : \Theta \to \mathbf{R}$.

(4) Q has a unique maximum on Θ at θ_0.

Then $\widehat{\theta}_n \xrightarrow{\mathrm{a.s.}} \theta_0$.

The same method of proof should work, but we will pursue a different argument that was not available for weak consistency. The proof below is very slow because there are subtleties in this argument that were not obvious to me without elaboration.

Proof. Here's the outline of what we're going to do. For fixed $\omega \in \Omega$, $\{\widehat{\theta}_n(\omega)\}$ is a sequence in $\mathbf{R}^k$. From real analysis, we know that if this sequence has convergent subsequences, and if all of these convergent subsequences have the same limit, then the full sequence $\{\widehat{\theta}_n(\omega)\}$ is convergent with the same limit. Since Θ is compact, $\{\widehat{\theta}_n(\omega)\}$ does have at least one convergent subsequence. So we just have to show that for almost all $\omega \in \Omega$, every convergent subsequence has limit θ_0.

Fix an arbitrary $\omega \in \Omega$, and pick an arbitrary convergent subsequence $\{\widehat{\theta}_{n_i^\omega}(\omega)\}$; call its limit θ^ω. Fix $\varepsilon > 0$. By the triangle inequality,

$$\begin{aligned}
&\left|(n_i^\omega)^{-1} Q_{n_i^\omega}\left(\widehat{\theta}_{n_i^\omega}(\omega)\right) - Q(\theta^\omega)\right| \\
&\qquad = \left|(n_i^\omega)^{-1} Q_{n_i^\omega}\left(\widehat{\theta}_{n_i^\omega}(\omega)\right) - Q\left(\widehat{\theta}_{n_i^\omega}(\omega)\right) + Q\left(\widehat{\theta}_{n_i^\omega}(\omega)\right) - Q(\theta^\omega)\right| \\
&\qquad \leq \left|(n_i^\omega)^{-1} Q_{n_i^\omega}\left(\widehat{\theta}_{n_i^\omega}(\omega)\right) - Q\left(\widehat{\theta}_{n_i^\omega}(\omega)\right)\right| + \left|Q\left(\widehat{\theta}_{n_i^\omega}(\omega)\right) - Q(\theta^\omega)\right|
\end{aligned}$$

Since $\widehat{\theta}_{n_i^\omega}(\omega) \longrightarrow \theta^\omega$ and Q is continuous, there exists $N_1 \in \mathbf{N}$ such that

$$\left|Q\left(\widehat{\theta}_{n_i^\omega}(\omega)\right) - Q(\theta^\omega)\right| < \varepsilon/2 \quad \text{for all } i \geq N_1.$$

Since $n^{-1}Q_n \xrightarrow{\text{a.s.}} Q$ uniformly, the subsequence $\{(n_i^\omega)^{-1} Q_{n_i^\omega}\}$ also converges a.s. uniformly to Q. So (regardless of the what the deterministic sequence $\{\widehat{\theta}_{n_i^\omega}(\omega)\}$ happens to look like,) there exists $N_2 \in \mathbf{N}$ such that

$$\left|(n_i^\omega)^{-1} Q_{n_i^\omega}\left(\widehat{\theta}_{n_i^\omega}(\omega)\right) - Q\left(\widehat{\theta}_{n_i^\omega}(\omega)\right)\right| < \varepsilon/2 \quad \text{a.s. for all } i \geq N_2.^{62}$$

Putting this together, there exists N $(= N_1 \vee N_2)$ such that

$$\left|(n_i^\omega)^{-1} Q_{n_i^\omega}\left(\widehat{\theta}_{n_i^\omega}(\omega)\right) - Q\left(\theta^\omega\right)\right| < \varepsilon \quad \text{a.s. for all } i \geq N.$$

Since $\varepsilon > 0$ was arbitrary, we've shown that there is $\Omega' \subseteq \Omega$ such that $\mathbf{P}(\Omega') = 1$ and

$$(n_i^\omega)^{-1} Q_{n_i^\omega}(\omega')\left(\widehat{\theta}_{n_i^\omega}(\omega)\right) \longrightarrow Q\left(\theta^\omega\right) \quad \text{for all } \omega' \in \Omega'.$$

Next, observe that $\widehat{\theta}_n$ maximises Q_n by definition:

$$(n_i^\omega)^{-1} Q_{n_i^\omega}(\omega)\left(\widehat{\theta}_{n_i^\omega}(\omega)\right) \geq (n_i^\omega)^{-1} Q_{n_i^\omega}(\omega)(\theta_0).$$

If ω lies in Ω', the LHS converges to $Q(\theta^\omega)$. For the RHS, the fact that $(n_i^\omega)^{-1} Q_{n_i^\omega} \xrightarrow{\text{a.s.}} Q$ uniformly implies that there is $\Omega'' \subseteq \Omega$ such that $\mathbf{P}(\Omega'') = 1$ and $(n_i^\omega)^{-1} Q_{n_i^\omega}(\omega)(\theta_0) \longrightarrow Q(\theta_0)$ for all $\omega \in \Omega''$. So taking the subsequential limit $i \longrightarrow \infty$ on both sides, we obtain

$$Q(\theta^\omega) \geq Q(\theta_0) \quad \text{for every } \omega \in \Omega' \cap \Omega''.$$

[62] In this expression, $\widehat{\theta}_{n_i^\omega}$ is evaluated at a fixed $\omega \in \Omega$, but the functions $Q_{n_i^\omega}$ are still random; the 'a.s.' is w.r.t. random variation in the functions.

Since θ_0 is the unique maximiser of Q, $Q(\theta^\omega) \geq Q(\theta_0)$ implies $\theta^\omega = \theta_0$; so we have $\theta^\omega = \theta_0$ for every $\omega \in \Omega' \cap \Omega''$. Moreover, $\Omega' \cap \Omega''$ has measure one:

$$\begin{aligned}\mathbf{P}(\Omega' \cap \Omega'') \leq \mathbf{P}(\Omega') = 1 - \mathbf{P}\left((\Omega')^{\mathrm{C}} \cup (\Omega'')^{\mathrm{C}}\right) \\ \geq 1 - \mathbf{P}\left((\Omega')^{\mathrm{C}}\right) - \mathbf{P}\left((\Omega'')^{\mathrm{C}}\right) = 1.\end{aligned}$$

We've now shown that there is a probability-1 event $\Omega' \cap \Omega''$ such that whenever $\omega \in \Omega' \cap \Omega''$, every convergent subsequence of $\{\widehat{\theta}_n(\omega)\}$ converges to θ_0. Moreover, at least one convergent subsequence is guaranteed to exist for any ω by compactness of Θ. It follows that the full sequence $\{\widehat{\theta}_n(\omega)\}$ is convergent with limit θ_0 whenever $\omega \in \Omega' \cap \Omega''$. Since $\mathbf{P}(\Omega' \cap \Omega'') = 1$, this implies that $\widehat{\theta}_n \xrightarrow{\text{a.s.}} \theta_0$ as desired. ■

Example 16 (the incidental parameters problem)**.** Let $\{y_n\}$ be independent random variables with $y_n \sim \mathcal{N}(\mu_n, 1)$. The average log-likelihood at parameter m is

$$\widetilde{Q}_n(\{y_i\}_{i=1}^n, m) \coloneqq -\frac{1}{2}\ln(2\pi) - \frac{1}{2n}\sum_{i=1}^n (y_i - m_i)^2,$$

and

$$\widetilde{Q}(\{y_i\}_{i=1}^n, m) \coloneqq -\frac{1}{2}\ln(2\pi) - \lim_{n\to\infty}\frac{1}{2n}\sum_{i=1}^n (\mu_i - m_i)^2.$$

All the assumptions in our consistency theorems appear to be satisfied. We can take $\mu_n \in M$ for some compact $M \subseteq \mathbf{R}$, so that $\mu \in M^\infty$, which is compact by Tychonoff's theorem. Q_n is measurable and continuous as required. $n^{-1}Q_n$ converges a.s. uniformly to Q by a uniform LLN, though we will not prove this. Finally, Q has a unique maximum at μ, the true means.

But as a matter of fact, the maximum-likelihood estimator is inconsistent in this case. The reason is that the parameters μ live in an infinite-dimensional space, whereas we required Θ to be a subset of the finite-dimensional space $\mathbf{R}^k$. More intuitively, the number of parameters increases as $\mathrm{o}(n)$, whereas we required them to stay fixed at some k. It is perhaps intuitive that we cannot consistently estimate n parameters using n data points!

This is called this incidental parameters problem, and originated with Neyman and Scott (1948). It shows up in many other contexts. An obvious one is the fixed effects model for panel data, where the number of fixed effects is (by construction) equal to the number of cross-sectional observations, so that we cannot estimate them consistently under short-panel asymptotics.

It should be clear that this is a problem of identification: no matter how large your dataset is, you cannot precisely learn the values of the

parameters. We could formalise this in the Manski way by looking at the joint distribution of the data directly. The approach we took above of looking at identification indirectly via what can be consistently estimated from data is the old-fashioned approach to identification.

Example 17 (consistency of MLE for uniform)**.** Let $\{y_i\}_{i=1}^n$ be independent draws from $\mathcal{U}[0, \theta_0]$. The likelihood is

$$\mathcal{L}\left(\theta, \{y_i\}_{i=1}^n\right) = \theta^{-n} \prod_{i=1}^n \mathbf{1}(y_i \in [0, \theta]) = \theta^{-n} \mathbf{1}(y_i \in [0, \theta] \ \ \forall i).$$

The likelihood is discontinuous with a unique maximum at $\max_{i \in \{1,\ldots,n\}} y_i$.

Continuity of the criterion function is violated here, so our consistency theorems do not apply. But as a matter of fact, the maximum likelihood estimator is consistent. The reason is (basically) that despite the discontinuity, the maximum is always attained in this case. It turns out that the MLE is not efficient in this case, however.

Example 18 (consistency of LAD)**.** Suppose $\{y_i\}_{i=1}^n$ are iid. The least absolute deviations (LAD) estimator of the median minimises

$$\widetilde{Q}_n(\{y_i\}_{i=1}^n, \theta) \coloneqq \sum_{i=1}^n |y_i - \theta|.$$

The criterion function is clearly not differentiable everywhere: the derivative fails to exist for θ at which $y_i = \theta$ for some i. When the derivative exists, it is

$$\widetilde{Q}_{n,2}(\{y_i\}_{i=1}^n, \theta) = \sum_{i=1}^n [2 \cdot \mathbf{1}(y_i \leq \theta) - 1].$$

When n is even, there will in general be a continuum of minima, but in fact the measure of minimisers decreases sufficiently quickly with n to ensure than any measurable selection from the argmin is consistent.

Example 19. Let $\{y_i\}_{i=1}^n$ be independent draws from a mixture distribution with density

$$f(y) \coloneqq \lambda \left(2\pi\sigma^2\right)^{-1/2} \exp\left(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right) + (1-\lambda)\left(2\pi\right)^{-1/2} \exp\left(-\frac{1}{2}y^2\right)$$

Mixture distributions arise naturally in models of choice by heterogeneous agents, e.g. in IO (both theoretical and empirical). Let's suppose that we know λ in this case, and wish only to estimate μ and σ^2.

The log-likelihood is

$$\widetilde{Q}_n\left(\{y_i\}_{i=1}^n, \left(\mu, \sigma^2\right)\right) = -\frac{n}{2}\ln(2\pi) + \sum_{i=1}^n \ln\left[\lambda\sigma^{-1}\exp\left(-\frac{1}{2}\left(\frac{y_i-\mu}{\sigma}\right)^2\right) + (1-\lambda)\exp\left(-\frac{1}{2}y_i^2\right)\right]$$

This criterion function is discontinuous: if we set $\mu = y_i$ for some i and decrease σ toward 0, the log-likelihood increases without bound. So in order to satisfy the hypotheses of our consistency theorems, we must rule out $\sigma = 0$. To ensure that the parameter space remains compact, this means we must restrict σ to be bounded away from zero by some $\underline{\sigma} > 0$, perhaps not a very appealing restriction in applications. If we don't do this, consistency fails.

7.4 Asymptotic normality

Suppose $\widehat{\theta}_n$ is consistent for θ_0. We would then like to know how far it is likely to be from θ_0 in our sample. The obvious way to do this is to approximate the distribution of $n^{1/2}(\widehat{\theta}_n - \theta_0)$ with a normal by appeal to a central limit theorem. The asymptotic normality result in this section formalises this idea.

The proposition below is almost exactly Theorem 3.1 in Newey and McFadden (1994); the exception is that condition (3) above is slightly different (in a way that makes the proof easier).

Proposition 18 (asymptotic normality)**.** Assume

(1) $\widehat{\theta}_n \xrightarrow{\text{p}} \theta_0 \in \operatorname{int}\Theta$.

(2) There is a neighbourhood of θ_0 in Θ in which ∇Q_n and $\nabla^2 Q_n$ exist and are continuous for every $n \in \mathbf{N}$.

(3) A defined by

$$A \coloneqq \lim_{n\to\infty} \mathbf{E}\left(n^{-1}\nabla^2 Q_n(\theta_0)\right)$$

is nonsingular, and $n^{-1}\nabla^2 Q_n(\theta_n) \xrightarrow{\text{p}} A$ for any sequence of random vectors $\{\theta_n\}$ such that $\theta_n \xrightarrow{\text{p}} \theta_0$.

(4) $n^{-1/2}\nabla Q_n(\theta_0) \xrightarrow{\text{d}} \mathcal{N}(0, B)$, where

$$B \coloneqq \lim_{n\to\infty} \mathbf{E}\left(n^{-1}[\nabla Q_n(\theta_0)][\nabla Q_n(\theta_0)]^\top\right).$$

Then $n^{1/2}(\widehat{\theta}_n - \theta_0) \xrightarrow{\text{d}} \mathcal{N}\left(0, A^{-1}BA^{-1}\right)$.

Remark 21. θ_0 is defined by assumption (1). We're not taking a stand on why $\widehat{\theta}_n$ is consistent for θ_0, but if we decided to justify it using our weak consistency theorem, then θ_0 would of course be the maximiser of Q.

Partial proof. Assumption (2) says that there is neighbourhood of θ_0 on which each Q_n is twice continuously differentiable. Since θ_0 is interior, it follows that there exists a convex open neighbourhood $H \subseteq \operatorname{int}\Theta$ of θ_0 on which each Q_n is twice continuously differentiable. Let E_n be the event that $\widehat{\theta}_n \in H$. Since $\widehat{\theta}_n \xrightarrow{\text{p}} \theta_0$, $\mathbf{P}(E_n) \longrightarrow 1$.

We will fudge the proof by behaving as though E_n obtains for all n sufficiently large. This does *not* follow from $\mathbf{P}(E_n) \longrightarrow 1$, but proceeding in this manner turns out to be valid nonetheless. For a rigorous proof that avoids this fudge, see Newey and McFadden (1994, 'A complete proof of Theorem 3.1', p. 2152).

When E_n obtains, $\widehat{\theta}_n$ is an interior maximiser on H of the differentiable function Q_n, so must satisfy the first-order condition for a local maximum:

$$\nabla Q_n(\widehat{\theta}_n) = 0.$$

Since Q_n is twice continuously differentiable on H, the mean value theorem lets us replace the left-hand side with an exact Taylor expansion around θ_0:

$$\nabla Q_n(\theta_0) + \nabla^2 Q_n(\widetilde{\theta}_n)(\widehat{\theta}_n - \theta_0) = 0,$$

where the mean value $\widetilde{\theta}_n$ lies between $\widehat{\theta}_n$ and θ_0 (hence in H by convexity). Rearranging,

$$n^{1/2}(\widehat{\theta}_n - \theta_0) = -\left[n^{-1}\nabla^2 Q_n(\widetilde{\theta}_n)\right]^+ \left[n^{-1/2}\nabla Q_n(\theta_0)\right],$$

where $^+$ denotes the Moore–Penrose pseudo-inverse.[63]

Since $\widehat{\theta}_n \xrightarrow{\text{p}} \theta_0$ and $\widetilde{\theta}_n$ lies between $\widehat{\theta}_n$ and θ_0, $\widetilde{\theta}_n \xrightarrow{\text{p}} \theta_0$.[64] Hence by assumption (3),

$$n^{-1}\nabla^2 Q_n(\widetilde{\theta}_n) \xrightarrow{\text{p}} A.$$

[63] Under our assumptions, $n^{-1}\nabla^2 Q_n(\widetilde{\theta}_n)$ may be singular, in which case the ordinary matrix inverse is undefined. That's why we use the Moore–Penrose pseudo-inverse: it is always (uniquely) defined, and coincides with the ordinary inverse for nonsingular matrices. For details, see e.g. Rao (1973, sec. 1.b.5 & 1.c.5).

[64] This statement only makes sense if the mean value $\widetilde{\theta}_n$ is a random vector, i.e. is measurable! It turns out that it is; Newey and McFadden (1994, p. 2141, footnote 25) indicate why, and refer the reader to our friend Jennrich (1969) for a formal proof.

The Moore–Penrose pseudo-inverse operator is continuous at A since A is nonsingular,[65] so

$$\left[n^{-1}\nabla^2 Q_n(\widetilde{\theta}_n)\right]^+ \xrightarrow{\text{p}} A^{-1}$$

by the continuous mapping theorem. Combining this with assumption (4) and Slutsky's theorem, we obtain

$$\begin{aligned} n^{1/2}(\widehat{\theta}_n - \theta_0) &= -\left[n^{-1}\nabla^2 Q_n(\widetilde{\theta}_n)\right]^+ \left[n^{-1/2}\nabla Q_n(\theta_0)\right] \\ &\xrightarrow{\text{d}} -A^{-1}\mathcal{N}(0, B) \overset{\text{d}}{=} \mathcal{N}\left(0, A^{-1}BA^{-1}\right) \end{aligned}$$

where the final equality used the fact that A^{-1} is symmetric by Young's theorem. ■

Remark 22. Two things:

(1) It should be clear from the proof that we don't actually need $\widehat{\theta}_n$ to be a global maximiser; we only need it to satisfy the first-order condition. The result will therefore go through if $\{\widehat{\theta}_n\}$ is a sequence of local minima and/or maxima. This raises two issues, though. First, we'll need to find sufficient conditions for such an object to be measurable. Second, we can no longer appeal to our consistency theorems above to justify assumption (1). But it turns out that there are consistency theorems for local maxima/minima; see e.g. Amemiya (1985, Theorem 4.1.2).

(2) The existence of the second derivative is actually not required for asymptotic normality, though the proof is much harder without this assumption. The kind of proof employed here definitely requires the first derivative, since it is based on the first-order condition.

    The LAD estimator is an example of an asymptotically normal extremum estimator for which even the first derivative fails to exist. In section 7.6 (p. 90), we'll give an idea of how asymptotic normality can be proved without the use of derivatives for certain estimators, including the LAD estimator. The maximum score estimator (Manski,

[65] The ordinary matrix inverse operator is everywhere continuous: for invertible matrices $\{A_n\}$ and A, $A_n \longrightarrow A$ implies $A_n^{-1} \longrightarrow A^{-1}$. (We used this to prove Slutsky's theorem.) But the Moore–Penrose pseudo-inverse is not actually continuous! However, it turns out to be continuous at invertibility points, which is all we need.

1975) is an example in which ∇Q_n does not exist, and the limiting distribution is nonnormal (Kim & Pollard, 1990).[66]

You may wonder what can happen when θ_0 lies on the boundary of the parameter space. The following example shows how this can give rise a nonnormal limiting distribution, even for very simple and otherwise well-behaved estimators.

Example 20 (nonnormality on the boundary)**.** Let $\{y_n\}$ be independent random variables, each with mean $\mu \in \Theta \coloneqq [0, K]$ and variance $\sigma^2 \in (0, \infty)$. Choose K large enough that the sample average is never above K; then the least-squares estimator is

$$\widehat{\mu}_n \coloneqq \operatorname*{arg\,min}_{m \in [0,K]} n^{-1} \sum_{i=1}^{n} (y_i - m)^2 = \max \left\{ n^{-1} \sum_{i=1}^{n} y_i, 0 \right\},$$

what we might call a 'censored average'.

The Lindeberg–Lévy CLT implies that

$$n^{-1/2} \sum_{i=1}^{n} (y_i - \mu) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \sigma^2\right)$$

as usual. If $\mu > 0$ then

$$n^{1/2} (\widehat{\mu}_n - \mu) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \sigma^2\right),$$

but if $\mu = 0$ then the limiting distribution of $n^{1/2} \widehat{\mu}_n$ is $\mathcal{N}\left(0, \sigma^2\right)$ with all negative values censored to zero (the censored normal distribution).

Intuitively, what's going on here is that as n gets large, with high probability the sample average is close to μ by Kolmogorov's first SLLN, so we can restrict attention to the behaviour of $\widehat{\mu}_n$ in an arbitrarily small neighbourhood of μ. When $\mu > 0$, we can choose a neighbourhood that doesn't include zero, and it's as if $\widehat{\mu}_n$ were an ordinary uncensored average. But when $\mu = 0$, every neighbourhood of μ contains negative values at which censoring takes place, so that censoring has a first-order effect on the distribution of $\widehat{\mu}_n$ no matter how large n gets.

This raises another issue. If μ is positive but small, then we'll need a larger n in order for the normal distribution to be a good approximation,

[66] In particular, Q_n is a multidimensional step function, and the limit distribution is the maximum of a multidimensional Gaussian process with quadratic drift that depends on nuisance parameters. Not nice!

i.e. the rate of convergence is slower. This can be formalised by using a Pitman drift, meaning that we let μ drift toward zero as n increases and study how fast the drift can be without breaking asymptotic normality.[67] (See section 9.4 (p. 117) for details on what a Pitman drift is and how it can be used.)

True parameters on the boundary arise frequently in more sophisticated econometric models. One example (currently fashionable) is moment-inequality models, where we're on the boundary whenever an inequality binds in the population.

7.5 Estimating the asymptotic variance

If we are to use our asymptotic normality result from the previous section to approximate the distribution of $\widehat{\theta}_n$, we had better be able to obtain consistent estimates of A and B. This is possible under fairly weak assumptions. In this section, we'll restrict attention to iid data and additively separable criterion functions Q_n.

In particular, assume that the DGP $\{y_i\}$ is iid and that $\widetilde{Q}_n$ can be written

$$\widetilde{Q}_n(\{y_i\}_{i=1}^n, \theta) = \sum_{i=1}^n \widetilde{q}(y_i, \theta)$$

for some function $\widetilde{q}$. We'll want work directly with the random functions $\Theta \to \mathbf{R}$ defined by $q_i(\omega)(\theta) \coloneqq \widetilde{q}(y_i(\omega), \theta)$, so that $Q_n = \sum_{i=1}^n q_i$. Observe that $\{q_i\}$ is an iid sequence of random functions.

Assume that the conditions of the weak consistency and asymptotic normality results (pp. 78 & 84) hold. In the current setting, this implies in particular that each q_i is twice continuously differentiable near θ_0 and that the moments $\mathbf{E}([\nabla q_1(\theta_0)][\nabla q_1(\theta_0)]^\top)$ and $\mathbf{E}\left(\nabla^2 q_1(\theta_0)\right)$ exist. Further suppose that ∇q_1 and $\nabla^2 q_1$ are each bounded by some finite-expectation random variable (so that the dominated convergence theorem applies).

$\left\{\nabla^2 q_i(\theta_0)\right\}$ is a sequence of iid $k \times k$ random matrices each with mean $\mathbf{E}\left(\nabla^2 q_1(\theta_0)\right)$, so

$$n^{-1}\nabla^2 Q_n(\theta_0) = n^{-1}\sum_{i=1}^n \nabla^2 q_i(\theta_0) \xrightarrow{\text{a.s.}} \mathbf{E}\left(\nabla^2 q_1(\theta_0)\right)$$

[67] The last point raises a more general issue. There's a fashion in econometrics now for studying estimators whose good properties are uniform in the true parameters. Uniformly good estimators have the nice feature that we don't need to be worried about the true parameter being close to a bad region (e.g. a boundary, a nonidentification region); when it's in a good region, its properties are as good as anywhere else in that region.

by Kolmogorov's second SLLN. Similarly, $\{[\nabla q_i(\theta_0)][\nabla q_i(\theta_0)]^\top\}$ is a sequence of iid $k \times k$ matrices with mean $\mathbf{E}\left([\nabla q_1(\theta_0)][\nabla q_1(\theta_0)]^\top\right)$, so by Kolmogorov's second SLLN we have

$$n^{-1}[\nabla Q_n(\theta_0)][\nabla Q_n(\theta_0)]^\top = n^{-1}\sum_{i=1}^{n}[\nabla q_i(\theta_0)][\nabla q_i(\theta_0)]^\top \xrightarrow{\text{a.s.}} \mathbf{E}\left([\nabla q_1(\theta_0)][\nabla q_1(\theta_0)]^\top\right).$$

Hence by the dominated convergence theorem,

$$A = \lim_{n\to\infty} \mathbf{E}\left(n^{-1}\nabla^2 Q_n(\theta_0)\right) = \mathbf{E}\left(\nabla^2 q_1(\theta_0)\right)$$
$$B = \lim_{n\to\infty} \mathbf{E}\left(n^{-1}[\nabla Q_n(\theta_0)][\nabla Q_n(\theta_0)]^\top\right) = \mathbf{E}\left([\nabla q_1(\theta_0)][\nabla q_1(\theta_0)]^\top\right).$$

Now that we have clean expressions for A and B, we can think about estimating them. Consider the random functions $\Theta \to \mathbf{R}^{r\times r}$

$$\widehat{A}_n \coloneqq n^{-1}\nabla^2 Q_n = n^{-1}\sum_{i=1}^{n}\nabla^2 q_i$$
$$\widehat{B}_n \coloneqq n^{-1}\left[\nabla Q_n\right]\left[\nabla Q_n\right]^\top = n^{-1}\sum_{i=1}^{n}\left[\nabla q_i\right]\left[\nabla q_i\right]^\top.$$

We have just shown that $\widehat{A}_n(\theta_0) \xrightarrow{\text{a.s.}} A$ and $\widehat{B}_n(\theta_0) \xrightarrow{\text{a.s.}} B$. But these are infeasible estimators because they require knowledge of θ_0. The obvious remedy is to plug in our consistent estimator $\widehat{\theta}_n$. By continuous differentiability, $\widehat{A}_n$ and $\widehat{B}_n$ are continuous at θ_0. Hence

$$\widehat{A}_n(\widehat{\theta}_n) \xrightarrow{\text{a.s.}} A \quad \text{and} \quad \widehat{B}_n(\widehat{\theta}_n) \xrightarrow{\text{a.s.}} B$$

by the continuous mapping theorem. It follows by Slutsky's theorem that

$$\widehat{A}_n(\widehat{\theta}_n)^{-1}\widehat{B}_n(\widehat{\theta}_n)\widehat{A}_n(\widehat{\theta}_n)^{-1}$$

is a consistent estimator of the asymptotic variance of our extremum estimator.

As a reward for our toil, we can finally do inference. In particular, what we have learned is the approximate distributional result

$$\widehat{\theta}_n \overset{\text{a}}{\sim} \mathcal{N}\left(\theta_0, n^{-1}\widehat{A}_n(\widehat{\theta}_n)^{-1}\widehat{B}_n(\widehat{\theta}_n)\widehat{A}_n(\widehat{\theta}_n)^{-1}\right).$$

7.6 Asymptotic normality with a nonsmooth objective

This section is basically an aside, so generality and rigour will be sacrificed on the altar of clarity.

Our asymptotic normality result in section 7.4 required the existence and continuity of derivatives of the criterion function in a neighbourhood of θ_0. The second derivative assumption turns out to be unnecessary, but proving this in generality is hard. Getting rid of the first derivative is harder still, but important in certain applications (e.g. auctions). We'll give an example to indicate what can be done.

Suppose that the DGP $\{y_i\}$ is iid and that the parameter θ_0 satisfies the moment condition $\mathbf{E}\left(\widetilde{g}(y_1, \theta_0)\right) = 0$. Estimation based on moment conditions such as these can be done using the generalised method of moments (GMM) covered in section 10, a special case of extremum estimation. We use it here merely as an example. Restrict attention to the univariate case $y_i \in \mathbf{R}$ and $\Theta \subseteq \mathbf{R}$, and write F for the CDF of each y_i.

The population moment can be written $\int_{\mathbf{R}} \widetilde{g}(y, \theta_0) F(\mathrm{d}y) = 0$. Define the empirical distribution function (EDF) by

$$\widehat{F}_n(y) \coloneqq n^{-1} \sum_{i=1}^{n} \mathbf{1}(y \leq y_i).$$

It should be clear that an integral w.r.t. F_n is simply a sample average. We use the method of moments estimator $\widehat{\theta}_n$ which satsifies the analogous sample moment

$$\int_{\mathbf{R}} \widetilde{g}(y, \widehat{\theta}_n) F_n(\mathrm{d}y) = 0.$$

(We simply assume that a solution exists; there are workarounds.)

Now we do some simple algebra:

$$\begin{aligned}
0 &= n^{1/2} \int_{\mathbf{R}} \widetilde{g}(y, \theta_0) F_n(\mathrm{d}y) + n^{1/2} \int_{\mathbf{R}} \widetilde{g}(y, \widehat{\theta}_n) F(\mathrm{d}y) \\
&\qquad - n^{1/2} \int_{\mathbf{R}} \widetilde{g}(y, \theta_0) F_n(\mathrm{d}y) - n^{1/2} \int_{\mathbf{R}} \widetilde{g}(y, \widehat{\theta}_n) F(\mathrm{d}y) \\
&= n^{1/2} \int_{\mathbf{R}} \widetilde{g}(y, \theta_0) F_n(\mathrm{d}y) + n^{1/2} \int_{\mathbf{R}} \widetilde{g}(y, \widehat{\theta}_n) F(\mathrm{d}y) \\
&\qquad + n^{1/2} \int_{\mathbf{R}} \left[\widetilde{g}(y, \widehat{\theta}_n) - \widetilde{g}(y, \theta_0)\right] [F_n - F](\mathrm{d}y) \\
&= \underbrace{n^{1/2} \int_{\mathbf{R}} \widetilde{g}(y, \theta_0) F_n(\mathrm{d}y)}_{\xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \mathbf{E}\left(\widetilde{g}(y, \theta_0)^2\right)\right)} + n^{1/2} \int_{\mathbf{R}} \widetilde{g}(y, \widehat{\theta}_n) F(\mathrm{d}y) + \mathrm{o_p}(1).
\end{aligned}$$

For now, we're merely asserting that the final term is $\mathrm{o_p}(1)$; more on that below. The convergence in distribution of the first term is immediate by the Lindeberg–Lévy CLT.

The usual modus operandi is as follows. Assuming that $\widetilde{g}(y,\cdot)$ is differentiable with derivative $\widetilde{g}_2$, we mean-value expand the remaining term as

$$\begin{aligned} n^{1/2}\int_{\mathbf{R}} \widetilde{g}(y,\widehat{\theta}_n)F(\mathrm{d}y) &= n^{1/2}\int_{\mathbf{R}} \left[\widetilde{g}(y,\theta_0) + \widetilde{g}_2(y,\widetilde{\theta}_n)(\widehat{\theta}_n-\theta_0)\right] F(\mathrm{d}y) \\ &= \left[n^{1/2}(\widehat{\theta}_n-\theta_0)\right]\int_{\mathbf{R}} \widetilde{g}_2(y,\widetilde{\theta}_n)F(\mathrm{d}y) \end{aligned}$$

where the mean value $\widetilde{\theta}_n$ lies between $\widehat{\theta}_n$ and θ_0, and so converges in probability to θ_0. Assuming that the second derivative $\widetilde{g}_2(y,\cdot)$ is continuous at θ_0, it follows by the continuous mapping theorem that

$$\int_{\mathbf{R}} \widetilde{g}_2(y,\widetilde{\theta}_n)F(\mathrm{d}y) \xrightarrow{\mathrm{p}} \int_{\mathbf{R}} \widetilde{g}_2(y,\theta_0)F(\mathrm{d}y) = \mathbf{E}\left(\widetilde{g}_2(y_1,\theta_0)\right),$$

assumed to be nonzero. Finally, rearrange and use Slutsky's theorem to obtain

$$\begin{aligned} n^{1/2}(\widehat{\theta}_n-\theta_0) = & -\left[\int_{\mathbf{R}} \widetilde{g}_2(y,\widetilde{\theta}_n)F(\mathrm{d}y)\right]^{-1}\left[n^{1/2}\int_{\mathbf{R}} \widetilde{g}(y,\theta_0)F_n(\mathrm{d}y)\right] \\ &\xrightarrow{\mathrm{d}} -\mathbf{E}\left(\widetilde{g}_2(y_1,\theta_0)\right)^{-1}\mathcal{N}\left(0,\mathbf{E}\left(\widetilde{g}(y,\theta_0)^2\right)\right) \\ &\overset{\mathrm{d}}{=}\mathcal{N}\left(0,\mathbf{E}\left(\widetilde{g}(y,\theta_0)^2\right)\Big/\,\mathbf{E}\left(\widetilde{g}_2(y_1,\theta_0)\right)^2\right). \end{aligned}$$

This is basically the strategy we followed in proving out asymptotic normality result above, and clearly requires the existence of the first derivative. (We made do without the second derivative by exploiting the moment; for a general exteremum estimator we need the second derivative as well for this strategy to work.)

But when the first derivative does not exist, as in e.g. the LAD case, this strategy is not available to us. Fortunately, the integral is a smoothing operator: the Lebesgue integral of a nonsmooth function may be smooth. (This should be intuitive.) So we can make assumptions sufficient for the integral $\int_{\mathbf{R}} \widetilde{g}(y,\cdot)F(\mathrm{d}y)$ to be differentiable without requiring $\widetilde{g}(y,\cdot)$ to be. Then we can mean-value expand the integral (rather than the integrand)

around θ_0 as

$$\begin{aligned} n^{1/2} &\int_{\mathbf{R}} \widetilde{g}(y, \widehat{\theta}_n) F(\mathrm{d}y) \\ &= n^{1/2} \int_{\mathbf{R}} \widetilde{g}(y, \theta_0) F(\mathrm{d}y) + \left[n^{1/2} (\widehat{\theta}_n - \theta_0) \right] \frac{\partial}{\partial \theta} \int_{\mathbf{R}} \widetilde{g}_2(y, \widetilde{\theta}_n) F(\mathrm{d}y) \\ &= \left[n^{1/2} (\widehat{\theta}_n - \theta_0) \right] \frac{\partial}{\partial \theta} \int_{\mathbf{R}} \widetilde{g}_2(y, \widetilde{\theta}_n) F(\mathrm{d}y). \end{aligned}$$

Assuming that the derivative of the integral is continuous at θ_0, plus some additional boundedness condition to keep the derivative in check, we get

$$\frac{\partial}{\partial \theta} \int_{\mathbf{R}} \widetilde{g}(y, \widetilde{\theta}_n) F(\mathrm{d}y) \xrightarrow{\mathrm{p}} A$$

for some nonstochastic A, assumed nonzero. Now we can rearrange as before to get

$$\begin{aligned} n^{1/2} (\widehat{\theta}_n - \theta_0) = & - \left[\frac{\partial}{\partial \theta} \int_{\mathbf{R}} \widetilde{g}(y, \widetilde{\theta}_n) F(\mathrm{d}y) \right]^{-1} \left[n^{1/2} \int_{\mathbf{R}} \widetilde{g}(y, \theta_0) F_n(\mathrm{d}y) \right] \\ & \xrightarrow{\mathrm{d}} - A^{-1} \mathcal{N} \left(0, \mathbf{E} \left(\widetilde{g}(y, \theta_0)^2 \right) \right) \\ & \stackrel{\mathrm{d}}{=} \mathcal{N} \left(0, \mathbf{E} \left(\widetilde{g}(y, \theta_0)^2 \right) \Big/ A^2 \right). \end{aligned}$$

This approach works for the LAD estimator, for example. There $\widetilde{g}$ is the discontinuous function

$$\widetilde{g}(y, \theta) = 2 \cdot \mathbf{1} (y \leq \theta) - 1,$$

but its integral is

$$\int_{\mathbf{R}} \widetilde{g}(y, \widetilde{\theta}_n) F(\mathrm{d}y) = 2 \int_{-\infty}^{\widetilde{\theta}_n} \mathrm{d}F - 1 = 2F(\widetilde{\theta}_n) - 1,$$

which is differentiable provided the data are continuously distributed. In that case, calling the density f, we have $A = 2f(\theta_0)$. Provided f is strictly positive at θ_0, the LAD estimator has asymptotic distribution

$$\begin{aligned} n^{1/2} (\widehat{\theta}_n - \theta_0) & \xrightarrow{\mathrm{d}} \mathcal{N} \left(0, \mathbf{E} \left([2 \cdot \mathbf{1} (y_1 \leq \theta_0) - 1]^2 \right) \Big/ 4 f(\theta_0)^2 \right) \\ & \stackrel{\mathrm{d}}{=} \mathcal{N} \left(0, 1/4 f(\theta_0)^2 \right). \end{aligned}$$

Another estimator that admits this kind of argument is the maximum rank correlation estimator. (I don't pretend to know what that is.)

Recall that to get to this point, we simply asserted that

$$n^{1/2} \int_{\mathbf{R}} \left[\widetilde{g}(y, \widehat{\theta}_n) - \widetilde{g}(y, \theta_0) \right] [F_n - F](\mathrm{d}y) = \mathrm{o_p}(1).$$

This is not so obvious when $\widetilde{g}$ is nonsmooth, since we'd usually use a mean-value expansion of the integrand to show this. But here empirical process theory comes to the rescue. An empirical process is a random functional, i.e. a random function mapping from a functional space. We can write our troublesome integral as

$$\begin{aligned}
n^{1/2} \int_{\mathbf{R}} & \left[\widetilde{g}(y, \widehat{\theta}_n) - \widetilde{g}(y, \theta_0) \right] [F_n - F](\mathrm{d}y) \\
&= -n^{1/2} \int_{\mathbf{R}} \widetilde{g}(y, \widehat{\theta}_n) F(\mathrm{d}y) - n^{1/2} \int_{\mathbf{R}} \widetilde{g}(y, \theta_0) F_n(\mathrm{d}y) \\
&= -n^{1/2} \int_{\mathbf{R}} \widetilde{g}(y, \theta_0) F(\mathrm{d}y) - n^{1/2} \int_{\mathbf{R}} \widetilde{g}(y, \theta_0) F_n(\mathrm{d}y) + \mathrm{o_p}(1) \\
&= -n^{1/2} \int_{\mathbf{R}} \widetilde{g}(y, \theta_0) F_n(\mathrm{d}y) + \mathrm{o_p}(1) \\
&= -v_n(\widetilde{g}(\cdot, \theta_0)) + \mathrm{o_p}(1)
\end{aligned}$$

where v_n is the empirical process $v_n(f) \coloneqq n^{-1/2} \sum_{i=1}^n f(y_i)$. Provided that $\widetilde{g}$ lives in an appropriate family of functions, theorems in empirical process theory allow us to assert that $v_n(\widetilde{g}(\cdot, \theta_0)) = \mathrm{o_p}(1)$. The requirements involve a concept called stochastic equicontinuity that plays a central role in empirical process theory.

A word of caution is needed here. There are important applications in which the appropriate stochastic equicontinuity condition is violated, in which case the remainder term above does not vanish. In that case we may have a nonnormal asymptotic distribution. An example of this is Manski's maximum score estimator.

8 The (quasi-)maximum-likelihood estimator

Official reading: Amemiya (1985, sec. 4.2.1–4.2.3) and Newey and McFadden (1994, sec. 2.2.1, 2.4, 3.2, 3.3, and 4.2).

8.1 Preliminaries

Recall the estimation setting laid out section 7. The data-generating process $\{y_n\}$ is a $\mathbf{R}^r$-valued stochastic process defined on a probability space $(\Omega, \mathcal{A}, \mathbf{P})$; its law is denoted μ_0. We observe a dataset, meaning a realisation $\{y_i(\omega)\}_{i=1}^n$ of the first n coordinates of the DGP.

We do not know the law $\mu_0 \in M$; a priori we only know that it lies in a set M of probability measures. There is some parameter $\tau : M \to \Theta$ of interest ($\Theta \subseteq \mathbf{R}^k$) whose true value $\tau(\mu_0)$ we wish to learn about using the data.

In our study of extremum estimators, we did not specify a particular parameter τ; instead we looked at estimators that are consistent for some point in $\theta_0 \in \Theta$ which may or may not equal $\tau(\mu_0)$ for some parameter of interest τ. Now that we're looking at a specific class of extremum estimators, we will be able to ensure that $\theta_0 = \tau(\mu_0)$.

So fix a parameter $\tau : M \to \Omega$ of interest. Since we're only interested in estimating $\tau(\mu_0)$, there is no need to distinguish between distinct measures in M that give rise to the same parameter value; so wlog, we will treat τ as a bijection (one-to-one and onto). This means that we can write M as $\{\mu^\theta\}_{\theta\in\Theta}$, a parametric (finite-dimensional) family of distributions. The true value of the parameter is written $\theta_0 \coloneqq \tau(\mu_0)$.[68] To re-iterate, θ_0 is a feature of the unknown population distribution now: it is no longer some hard-to-interpret probability limit of an extremum estimator.

Suppose that there is a measure ν with respect to which every μ^θ possesses a density (Radon–Nikodým derivative). Let μ_n^θ and ν_n^θ denote the marginal distributions that μ^θ and ν induce over the first n coordinates, and write $\boldsymbol{f}_n^\theta \coloneqq \mathrm{d}\mu_n^\theta/\mathrm{d}\nu^n$ for the density governing a sample of size n when the true parameter is θ (the true law is μ^θ).

The likelihood function $\mathcal{L}$ is the random function $\Theta \to \mathbf{R}$ defined by

$$\mathcal{L}_n(\omega)(\theta) \coloneqq \boldsymbol{f}_n^\theta\left(\{y_i(\omega)\}_{i=1}^n\right) \quad \text{for each } \omega \in \Omega \text{ and } \theta \in \Theta.$$

The log-likelihood function is $\ell_n \coloneqq \ln(\mathcal{L}_n)$. (Set $\ell_n(\theta) = -\infty$ whenever

[68] By the way, μ_0 and μ^{θ_0} denote the same thing in my notation. I don't think this should cause any confusion.

$\mathcal{L}_n(\theta) = 0$.) A maximum likelihood estimator is an extremum estimator whose criterion function (Q_n) is ℓ_n.

We will focus on the case of independently and identically distributed data. In this case,

$$\boldsymbol{f}_n^\theta\left(\{y_i\}_{i=1}^n\right) = \prod_{i=1}^n f^\theta\left(y_i\right)$$

where f^θ is the marginal distribution. Define the (log-)likelihood contribution of observation i as

$$\ell_i^1(\omega)(\theta) = \ln\left(f^\theta\left(y_i(\omega)\right)\right) \quad \text{for each } \omega \in \Omega \text{ and } \theta \in \Theta.$$

Then we can write the log-likelihood as $\ell_n = \sum_{i=1}^n \ell_i^1$, a sum of iid random functions to which we can apply Jennrich's uniform SLLN. Pretty much all the results we will derive for the MLE can be extended to the non-iid case, but we won't do it.

Now suppose that our model of μ_0 is misspecified: the $\boldsymbol{f}_n^\theta$ we use to form the likelihood is not actually equal to $\mathrm{d}\mu_n^\theta/\mathrm{d}\nu_n$. The log-likelihood ℓ_n is still well-defined, so we can still obtain an extremum estimator by maximising it. Such an estimator is called a quasi-maximum-likelihood estimator (QMLE). We will see that under appropriate conditions, QMLEs are consistent for θ_0 and asymptotically normal, but that they do not share the efficiency properties of the MLE.

8.2 Consistency for the truth

In this section, we'll establish the consistency of the MLE for $\theta_0 = \tau(\mu_0)$. The (entirely straightforward) argument hinges on the following result.

Proposition 19 (information inequality)**.** Let f and g be densities w.r.t. a measure ν on $(\Omega, \mathcal{A})$. Then $\int_\Omega \ln(g/f) f \mathrm{d}\nu \leq 0$, with equality iff $g = f$ ν-a.e.

Proof. A first-order mean-value theorem expansion of $\ln(g/f)$ around $g/f = 1$ yields

$$\ln(g/f) = (g/f - 1) - \frac{1}{2}\frac{1}{\gamma^2}(g/f - 1)^2,$$

where γ (a function!) lies between g/f and 1.[69] So

$$\begin{aligned}\int_\Omega \ln(g/f) f \mathrm{d}\nu &= \int_\Omega (g/f-1) f \mathrm{d}\nu - \int_\Omega \frac{1}{2\gamma^2}(g/f-1)^2 f \mathrm{d}\nu \\ &= \int_\Omega g \mathrm{d}\nu - \int_\Omega f \mathrm{d}\nu - \int_\Omega \frac{1}{2\gamma^2}\frac{(g-f)^2}{f}\mathrm{d}\nu \\ &= -\int_\Omega \frac{1}{2\gamma^2}\frac{(g-f)^2}{f}\mathrm{d}\nu.\end{aligned}$$

The RHS is evidently nonpositive and equal to zero iff $g = f$ ν-a.e. ■

The maximum likelihood estimator for independent data is an extremum estimator whose criterion function is $\ell_n = \sum_{i=1}^n \ell_i^1$. When the data are iid, $\{\ell_i^1\}$ are iid random functions $\Theta \to \mathbf{R}$. Assume that Θ is compact, that ℓ_1^1 is continuous and that $\mathbf{E}\left(\sup_{\theta\in\Theta}|\ell_1^1(\theta)|\right) < \infty$. Then by Jennrich's uniform SLLN,

$$n^{-1}\ell_n = n^{-1}\sum_{i=1}^n \ell_i^1 \xrightarrow{\text{a.s.}} Q \quad \text{uniformly over } \Theta,$$

where $Q : \Theta \to \mathbf{R}$ is the nonstochastic function

$$\begin{aligned}Q(\theta) &\coloneqq \mathbf{E}\left(\ln\left(f^\theta(y_1)\right)\right) \\ &= \int_{\mathbf{R}^r} \ln\left(f^\theta(y_1)\right) \mu_1^{\theta_0}(\mathrm{d}y_1) \\ &= \int_{\mathbf{R}^r} \ln\left(f^\theta(y_1)\right) f^{\theta_0}(y_1)\, \nu_1(\mathrm{d}y_1).\end{aligned}$$

using the fact that $\mu_1^{\theta_0}$ is the true distribution and that $f^{\theta_0}(y_1)$ is its density w.r.t. ν_1. The information inequality tells us precisely that Q attains a unique maximum at θ_0. Lo and behold, all the assumptions of our strong consistency result (p. 80) are satisfied, so the MLE is strongly consistent for θ_0. And as promised, θ_0 here was defined as $\tau(\mu_0)$, the true value of the parameter τ. 'Consistency for the truth', if you will. To summarise:

Proposition 20 (consistency for the truth)**.** Suppose that Θ is compact, that ℓ_1^1 is continuous and that $\mathbf{E}\left(\sup_{\theta\in\Theta}|\ell_1^1(\theta)|\right) < \infty$. Then the MLE $\widehat{\theta}_n$ is consistent for $\theta_0 = \tau(\mu_0)$.

A similar argument can be applies for the QMLE. But in this case, the information inequality cannot be used to establish that the limit function

[69] We need γ to be $\mathcal{A}$-measurable; otherwise we cannot integrate it. It is in fact measurable, as discussed in footnote 64 (p. 85).

Q has a unique maximum, and certainly not that this maximum is equal to $\tau(\mu_0)$. It is possible, however, to give additional conditions restricting the degree of misspecification in such a way that Q is guaranteed to have a unique maximum at $\tau(\theta_0)$. A QMLE satisfying these extra conditions will also be consistent for the truth.

8.3 Asymptotic normality

Asymptotic normality is even easier than consistency: we just apply our asymptotic normality result for extremum estimators out of the box. Assume that ℓ_1^1 is twice continuously differentiable in a neighbourhood of θ_0. The A and B matrices from the asymptotic normality proposition become (using the fact that the data are iid)

$$\begin{aligned} A &\coloneqq \lim_{n\to\infty} \mathbf{E}\left(n^{-1}\nabla^2\ell_n(\theta_0)\right) \\ &= \lim_{n\to\infty} \mathbf{E}\left(n^{-1}\sum_{i=1}^{n}\nabla^2\ell_i^1(\theta_0)\right) \\ &= \mathbf{E}\left(\nabla^2\ell_1^1(\theta_0)\right) \end{aligned}$$

and

$$\begin{aligned} B &\coloneqq \lim_{n\to\infty} \mathbf{E}\left(n^{-1}\left[\nabla\ell_n(\theta_0)\right]\left[\nabla\ell_n(\theta_0)\right]^\top\right) \\ &= \lim_{n\to\infty} \mathbf{E}\left(\left[n^{-1/2}\sum_{i=1}^{n}\nabla\ell_i^1(\theta_0)\right]\left[n^{-1/2}\sum_{i=1}^{n}\nabla\ell_i^1(\theta_0)\right]^\top\right) \\ &= \lim_{n\to\infty} \mathbf{E}\left(n^{-1}\sum_{i=1}^{n}\left[\nabla\ell_i^1(\theta_0)\right]\left[\nabla\ell_i^1(\theta_0)\right]^\top\right) \\ &= \mathbf{E}\left(\left[\nabla\ell_1^1(\theta_0)\right]\left[\nabla\ell_1^1(\theta_0)\right]^\top\right). \end{aligned}$$

We assume that both of these expectations exist and are finite, and furthermore that A is nonsingular (negative definite will do). The information matrix equality below will tell us that $-A = B$ when the model is correctly specified. $-A$ is called the Hessian form of the information matrix, and B is called the outer product form of the information matrix.

$\{\nabla^2\ell_i^1(\theta_0)\}$ is an iid sequence of random matrices whose mean we assumed exists, so by Khinchine's WLLN (p. 56) we have

$$n^{-1}\nabla^2\ell_n(\theta_0) = n^{-1}\sum_{i=1}^{n}\nabla^2\ell_i^1(\theta_0) \xrightarrow{\text{P}} A.$$

Since $\nabla^2 \ell_n$ is continuous in a neighbourhood of θ_0, any $\{\theta_n\}$ with $\theta_n = \theta_0 + \mathrm{o_p}(1)$ satisfies

$$n^{-1}\nabla^2 \ell_n(\theta_n) = n^{-1}\nabla^2 \ell_n(\theta_0) + \mathrm{o_p}(1) \xrightarrow{\mathrm{p}} A$$

by the continuous mapping theorem.

Assume that $\nabla \ell_1^1(\theta_0)$ is dominated by some random vector with finite expectation. Then the dominated convergence theorem applies, so we can exchange the order of differentiation and integration (see e.g. Rosenthal (2006, sec. 9.2)). Hence

$$\begin{aligned}\mathbf{E}\left(\nabla \ell_1^1(\theta_0)\right) &= \mathbf{E}\left(\frac{\nabla \mathcal{L}_1^1(\theta_0)}{\mathcal{L}_1^1(\theta_0)}\right) = \int_{\mathbf{R}^r} \frac{\frac{\partial}{\partial\theta} f^{\theta_0}(y)}{f^{\theta_0}(y)} f^{\theta_0}(y)\nu_1(\mathrm{d}y) \\ &= \int_{\mathbf{R}^r} \left[\frac{\partial}{\partial\theta} f^{\theta_0}(y)\right] \nu_1(\mathrm{d}y) = \frac{\partial}{\partial\theta}\int_{\mathbf{R}^r} f^{\theta_0}(y)\nu_1(\mathrm{d}y) = \frac{\partial(1)}{\partial\theta} = 0.\end{aligned}$$

In words, the expected score is zero at the truth. So $\{\nabla \ell_i^1(\theta_0)\}$ is an iid sequence of random vectors with mean zero and finite variance B. Hence by the multivariate Lindeberg–Lévy CLT we have

$$n^{-1/2}\nabla \ell_n(\theta_0) = n^{-1/2}\sum_{i=1}^{n} \nabla \ell_i^1(\theta_0) \xrightarrow{\mathrm{d}} \mathcal{N}(0, B).$$

Since $\widehat{\theta}_n \xrightarrow{\mathrm{p}} \theta_0$ by the arguments in the previous section, our asymptotic normality result for extremum estimators (p. 84) applies, giving us

$$n^{1/2}(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, A^{-1}BA^{-1}\right).$$

Actually, it turns out that $-A = B$ when the model is correctly specified, so that the asymptotic variance simplifies to $-A^{-1} = \left[-\nabla^2 \ell_1^1(\theta_0)\right]^{-1}$. This result is called the information matrix equality.

Proposition 21 (information matrix equality)**.** Assume that $\nabla \ell_1^1(\theta_0)$ is dominated by some random vector with finite expectation and that the model is correctly specified. Then

$$-A = -\mathbf{E}\left(\nabla^2 \ell_1^1(\theta_0)\right) = \mathbf{E}\left(\left[\nabla \ell_1^1(\theta_0)\right]\left[\nabla \ell_1^1(\theta_0)\right]^\top\right) = B.$$

Proof. The the proof is similar to the demonstration above that the expected score is zero at the true parameter.

$$
\begin{aligned}
&\mathbf{E}\left(\nabla^2 \ell_1^1(\theta_0)\right) \\
&\quad = \mathbf{E}\left(\frac{\partial}{\partial \theta^\top} \nabla \ell_1^1(\theta_0)\right) \\
&\quad = \mathbf{E}\left(\frac{\partial}{\partial \theta^\top}\left(\frac{\nabla \mathcal{L}_1^1(\theta_0)}{\mathcal{L}_1^1(\theta_0)}\right)\right) \\
&\quad = \int_{\mathbf{R}^r} \frac{\partial}{\partial \theta^\top}\left(\frac{\frac{\partial}{\partial \theta} f^{\theta_0}(y)}{f^{\theta_0}(y)}\right) f^{\theta_0}(y) \mathrm{d}\nu_1(\mathrm{d}y) \\
&\quad = \int_{\mathbf{R}^r} \left(\frac{\frac{\partial^2}{\partial \theta \partial \theta^\top} f^{\theta_0}(y)}{f^{\theta_0}(y)} - \frac{\left[\frac{\partial}{\partial \theta} f^{\theta_0}(y)\right]\left[\frac{\partial}{\partial \theta^\top} f^{\theta_0}(y)\right]}{f^{\theta_0}(y)^2}\right) f^{\theta_0}(y) \mathrm{d}\nu_1(\mathrm{d}y) \\
&\quad = \int_{\mathbf{R}^r} \left(\frac{\partial^2}{\partial \theta \partial \theta^\top} f^{\theta_0}(y)\right) \mathrm{d}\nu_1(\mathrm{d}y) \\
&\qquad\qquad - \int_{\mathbf{R}^r} \left[\frac{\frac{\partial}{\partial \theta} f^{\theta_0}(y)}{f^{\theta_0}(y)}\right]\left[\frac{\frac{\partial}{\partial \theta} f^{\theta_0}(y)}{f^{\theta_0}(y)}\right]^\top f^{\theta_0}(y) \mathrm{d}\nu_1(\mathrm{d}y) \\
&\quad = \int_{\mathbf{R}^r} \left(\frac{\partial^2}{\partial \theta \partial \theta^\top} f^{\theta_0}(y)\right) \mathrm{d}\nu_1(\mathrm{d}y) - \mathbf{E}\left(\left[\nabla \ell_1^1(\theta_0)\right]\left[\nabla \ell_1^1(\theta_0)\right]^\top\right) \\
&\quad = \frac{\partial^2}{\partial \theta \partial \theta^\top} \int_{\mathbf{R}^r} f^{\theta_0}(y) \mathrm{d}\nu_1(\mathrm{d}y) - \mathbf{E}\left(\left[\nabla \ell_1^1(\theta_0)\right]\left[\nabla \ell_1^1(\theta_0)\right]^\top\right) \\
&\quad = \frac{\partial^2 (1)}{\partial \theta \partial \theta^\top} - \mathbf{E}\left(\left[\nabla \ell_1^1(\theta_0)\right]\left[\nabla \ell_1^1(\theta_0)\right]^\top\right) \\
&\quad = -\mathbf{E}\left(\left[\nabla \ell_1^1(\theta_0)\right]\left[\nabla \ell_1^1(\theta_0)\right]^\top\right)
\end{aligned}
$$

where the exchanging of integration and differentiation in the third-last equality is permissible by the dominance condition. ■

When the model is not correctly specified (the QMLE case), we can still obtain asymptotic normality. The only part of the argument that relied on correct specification was our demonstration that the expected score is zero at the truth; this will have to be assumed to ensure asymptotic normality of the QMLE. The information matrix equality does not hold in this case, so we have to stick with the sandwich form $A^{-1}BA^{-1}$ for the asymptotic variance.

8.4 Estimating the asymptotic variance

The matrices A and B are unknown parameters of the DGP. If we're going to use our asymptotic normality result to approximate the distribution of the MLE, we had better be able to estimate them consistently!

Recall from section 7.5 (p. 88) that we already know how to estimate A and B consistently in the general framework of extremum estimation. In particular, define the random functions $\Theta \to \mathbf{R}^{r \times r}$ by

$$\widehat{A}_n \coloneqq n^{-1} \nabla^2 \ell_n = n^{-1} \sum_{i=1}^{n} \nabla^2 \ell_i^1$$
$$\widehat{B}_n \coloneqq n^{-1} \left[\nabla \ell_n\right] \left[\nabla \ell_n\right]^\top = n^{-1} \sum_{i=1}^{n} \left[\nabla \ell_i^1\right] \left[\nabla \ell_i^1\right]^\top;$$

then $\widehat{A}_n(\widehat{\theta}_n)$ and $\widehat{B}_n(\widehat{\theta}_n)$ are strongly consistent for A and B, respectively, under the conditions of the consistency and asymptotic normality results plus a dominance condition on the derivatives to allow the interchange of integration and differentiation.

These were the analogy estimators. Analogy estimation just means using sample averages to estimate population averages (expectations). To motivate an alternative way of estimating A and B, let's restate that in fancier language. Suppose we wish to estimate

$$\mathbf{E}(\psi(y_1, \theta_0)) = \int_{\mathbf{R}^r} \psi(y, \theta_0) \mu_1^{\theta_0}(\mathrm{d}y) = \int_{\mathbf{R}^r} \psi(y, \theta_0) F^{\theta_0}(\mathrm{d}y),$$

where F^θ is the CDF corresponding to the law μ_1^θ and $\psi : \mathbf{R}^r \times \Theta \to \mathbf{R}^\ell$ is some interesting function. (ψ is $\nabla^2 \ell_i^1$ for A and $\left[\nabla \ell_i^1\right] \left[\nabla \ell_i^1\right]^\top$ for B.)

The natural nonparametric estimator of F^{θ_0} is the EDF (empirical distribution function)

$$\widehat{F}_n(y) \coloneqq n^{-1} \sum_{i=1}^{n} \mathbf{1}(y \leq y_i).$$

The EDF is consistent for F^{θ_0} in a strong sense: the Glivenko–Cantelli theorem (e.g. Billingsley (1995, p. 269)) says that $\widehat{F}_n$ converges uniformly a.s. to F^{θ_0}. The estimator is nonparametric in the sense that it does not require us to first estimate θ_0, then plug in our estimate to compute our estimate of F^{θ_0}: we just estimate the function F^{θ_0} directly.

If we knew θ_0, a natural way to estimate $\int_{\mathbf{R}^r} \psi(y, \theta_0) F^{\theta_0}(\mathrm{d}y)$ would be to integrate $\psi(\cdot, \theta_0)$ w.r.t. $\widehat{F}_n$ rather than the unknown F^{θ_0}. We don't know θ_0, but if $\psi(y, \cdot)$ is continuous then we can replace θ_0 with $\widehat{\theta}_n$ without affecting

consistency. The estimator we obtain in this way is (fairly obviously) precisely the analogy estimator:

$$\int_{\mathbf{R}^r} \psi(y, \widehat{\theta}_n) \widehat{F}_n(\mathrm{d}y) = n^{-1} \sum_{i=1}^{n} \psi(y_i, \widehat{\theta}_n).$$

This class of estimators has the virtue that it's easy to prove consistency under weak conditions using a law of large numbers. It is a semiparametric approach: there's a nonparametric step (the EDF $\widehat{F}_n$) and a parametric step (the consistent estimator $\widehat{\theta}_n$).

But in the MLE context, we already have a full parametric model of F^{θ_0}; why not make use of it? In particular, instead of integrating w.r.t. the nonparametric EDF $\widehat{F}_n$, why not integrate w.r.t $F^{\widehat{\theta}_n}$, the plug-in estimate of F^{θ_0} obtained by making use of our model $\theta \mapsto F^\theta$ of the DGP? This suggestion yields a parametric estimator of $\int_{\mathbf{R}^r} \psi(y, \theta_0) F^{\theta_0}(\mathrm{d}y)$:

$$\int_{\mathbf{R}^r} \psi(y, \widehat{\theta}_n) F^{\widehat{\theta}_n}(\mathrm{d}y).$$

Provided $\theta \mapsto F^\theta$ is continuous, this also provides a consistent estimate of F^{θ_0}, and it may be more efficient than the analogy estimator. (But it isn't necessarily more efficient!)

It should be pretty obvious where this is going now. Instead of estimating B by integrating w.r.t. the EDF $\widehat{F}_n$ (the analogy estimator $\widehat{B}_n(\widehat{\theta}_n)$ above), we could integrate w.r.t. $F^{\widehat{\theta}_n}$. To implement this, define the (deterministic) function $\widetilde{B} : \Omega \to \mathbf{R}^{r \times r}$ by

$$\widetilde{B}(\theta) \coloneqq \int_{\mathbf{R}^r} \left[\frac{\partial}{\partial \theta} f^\theta(y)\right] \left[\frac{\partial}{\partial \theta} f^\theta(y)\right]^\top \underbrace{f^\theta(y) \nu_1(\mathrm{d}y)}_{=\mu^\theta(\mathrm{d}y) = F^\theta(\mathrm{d}y)} \quad \text{for each } \theta \in \Theta.$$

Clearly $\widetilde{B}(\theta_0) = B$. Our new estimator of B is the parametric plug-in estimator $\widetilde{B}(\widehat{\theta}_n)$. It is consistent by the continuous mapping theorem provided $\widetilde{B}$ is continuous at θ_0.[70]

How do these three estimators compare? Consider computational issues first. When analytical derivatives of the log-likelihood are available, $\widehat{A}(\widehat{\theta}_n)$ and $\widehat{B}(\widehat{\theta}_n)$ are very fast to compute accurately. By contrast, $\widetilde{B}(\widehat{\theta}_n)$ requires

[70] Defining $\widetilde{A}$ analogously, we have $-\widetilde{A}(\theta) = \widetilde{B}(\theta)$ for *any* $\theta \in \Theta$ (not just θ_0). This holds by the information matrix equality because $\widetilde{A}$ and $\widetilde{B}$ use the same value of the θ in the score/Hessian and the integrating density. Hence there's no point in defining $\widetilde{A}$ formally: just use $-\widetilde{B}$.

numerical integration, which is several orders of magnitude slower than averages (for a given level of accuracy). When analytical derivatives are not available, $\widehat{A}_n(\widehat{\theta}_n)$ suddenly becomes much heavier than $\widehat{B}_n(\widehat{\theta}_n)$ because accurate second derivatives are computationally expensive compared to first derivatives. But $\widetilde{B}(\widehat{\theta}_n)$ is still more expensive than $\widehat{A}_n(\widehat{\theta}_n)$ because accurate numerical integration is a lot slower than numerical differentiation.

Next, how close do these estimators actually tend to be to A and B in a finite sample, assuming that we've computed them (very) accurately? To answer this question, we have to look at Monte Carlo studies. The lighting lit review is that $\widehat{A}_n(\widehat{\theta}_n)$ is worst, $\widehat{B}_n(\widehat{\theta}_n)$ is better, and $\widetilde{B}(\widehat{\theta}_n)$ is best.

However, notice that the latter is only consistent for B (hence for $-A$) under correct specification. In the misspecified (QMLE) case, it will not consistently estimate either $-A$ or B since the integrating density is wrong! By contrast, $\widehat{A}_n(\widehat{\theta}_n)$ and $\widehat{B}_n(\widehat{\theta}_n)$ consistently estimate A and B under incorrect specification because we integrate w.r.t. the EDF, whose consistency does not depend on the correctness or otherwise of the parametric model. It follows that the asymptotic variance estimate

$$\widehat{A}_n(\widehat{\theta}_n)^{-1}\widehat{B}_n(\widehat{\theta}_n)\widehat{A}_n(\widehat{\theta}_n)^{-1}$$

is robust to misspecification of the likelihood, whereas the alternatives

$$-\widehat{A}_n(\widehat{\theta}_n)^{-1}, \quad \widehat{B}_n(\widehat{\theta}_n)^{-1} \quad \text{and} \quad \widetilde{B}(\widehat{\theta}_n)^{-1}$$

are not.

Example 21 (binary choice)**.** Suppose we have data $\{y_i, x_i\}_{i=1}^n$ where y_i takes values in $\{0,1\}$ and $x_i \in \mathbf{R}$ is a covariate. A single-index model specifies $\mathbf{P}(y=1|x,\theta) = \widetilde{p}(\theta^\top x)$ for some known function $\widetilde{p} : \mathbf{R}^k \to [0,1]$ and unknown parameter $\theta \in \Theta \subseteq \mathbf{R}^k$.

Write $p_i(\theta) \coloneqq \widetilde{p}(\theta^\top x_i)$ for short. The ith likelihood contribution (w.r.t. counting measure) is $p_i(\theta)^{y_i}[1-p_i(\theta)]^{1-y_i}$. (Why?) So the log-likelihood is

$$\ell_n(\theta) = \sum_{i=1}^n \left(y_i \ln(p_i(\theta)) + (1-y_i)\ln(1-p_i(\theta))\right).$$

We assume that $\widetilde{p}$, and hence each p_i, are differentiable. Then the score of

the ith log-likelihood contribution is

$$\begin{aligned}\nabla \ell_i^1(\theta) &= y_i \frac{\nabla p_i(\theta)}{p_i(\theta)} - (1-y_i)\frac{\nabla p_i(\theta)}{1-p_i(\theta)} \\ &= \left(\frac{y_i[1-p_i(\theta)] - (1-y_i)p_i(\theta)}{p_i(\theta)[1-p_i(\theta)]}\right)\nabla p_i(\theta) \\ &= \left(\frac{y_i - p_i(\theta)}{p_i(\theta)[1-p_i(\theta)]}\right)\nabla p_i(\theta).\end{aligned}$$

The Hessian is a lot uglier, so let's not even bother.

Now consider a particular single-index model for binary choice, the logit model:

$$\widetilde{p}\left(\theta^\top x\right) = \frac{\exp\left(\theta^\top x\right)}{1+\exp\left(\theta^\top x\right)}.$$

The derivative is

$$\begin{aligned}\nabla p_i(\theta) = \frac{\partial}{\partial\theta}\widetilde{p}\left(\theta^\top x_i\right) &= \frac{\exp\left(\theta^\top x_i\right) x_i \left[1+\exp\left(\theta^\top x_i\right)\right] - \exp\left(\theta^\top x_i\right)^2 x_i}{\left[1+\exp\left(\theta^\top x_i\right)\right]^2} \\ &= \frac{\exp\left(\theta^\top x_i\right)}{1+\exp\left(\theta^\top x_i\right)}\frac{1}{1+\exp\left(\theta^\top x_i\right)}x_i \\ &= p_i(\theta)[1-p_i(\theta)]x_i.\end{aligned}$$

So the score of the ith log-likelihood contribution is

$$\nabla \ell_i^1(\theta) = \left(y_i - p_i(\theta)\right)x_i.$$

Letting $\widehat{\theta}_n$ be the MLE, our analogy estimator of B is therefore

$$\widehat{B}(\widehat{\theta}_n) = n^{-1}\sum_{i=1}^n \left(y_i - p_i(\theta)\right)^2 x_i x_i^\top.$$

This is what e.g. Stata uses by default to estimate the asymptotic variance of the MLE for the logit model. It is numerically different from $\widehat{A}(\widehat{\theta}_n)$, of course. (To give a closed form for the latter we would have to compute some monstrous derivatives, so I'd rather not.)

8.5 The information matrix test

The availability of consistent estimators of A and B raises the possibility of testing whether the model is correctly specified. (These kinds of tests are called specification tests.) Intuitively, $-\widehat{A}_n(\widehat{\theta}_n)$ and $\widehat{B}_n(\widehat{\theta}_n)$ should be

close to each other in a large sample if the model is correctly specified. To formalise what we mean by 'close', we need an asymptotic distribution. It turns out that under the null hypothesis of correct specification, the elements of the $k \times k$ matrix

$$W_n(\widehat{\theta}_n) \coloneqq n^{1/2} \left[\widehat{A}_n(\widehat{\theta}_n) + \widehat{B}_n(\widehat{\theta}_n) \right]$$

are joint normally distributed in the limit, with mean zero and a covariance matrix that can be estimated consistently. $W_n(\widehat{\theta}_n)$ has $k(k+1)/2$ independent elements (since it's symmetric), so we can construct a test statistic as a function of these. Such tests are called information matrix (IM) tests, introduced by White (1982).

There are many ways turning $W_n(\widehat{\theta}_n)$ into a test statistic. A simple one is to take a quadratic form

$$q_n(\widehat{\theta}_n)^\top W_n(\widehat{\theta}_n) q_n(\widehat{\theta}_n)$$

where the vector $q_n(\widehat{\theta}_n)$ is chosen as a function of the estimated covariance. Under the null, this quadratic form converges to a χ^2 distribution with degrees of freedom depending on how many independent elements of $W_n(\widehat{\theta}_n)$ are given positive weight. Many variants have been proposed, some of which are asymptotically equivalent but much easier to compute.

Monte Carlo evidence suggests that the χ^2 approximation to the distribution of IM statistics is very poor. When using the 5% χ^2 critical values, the test often rejects under the null as often as 95% of the time for moderate sample sizes! Fortunately, the bootstrap approximation to the distribution of IM statistics is very accurate, so we can use bootstrap critical values instead.

8.6 Asymptotic efficiency

Consider the class of estimators that are weakly consistent for θ_0 and asymptotically normal. How do we choose between them? The obvious concern is to minimise variance. We don't have formulae for the variance of an estimator in a finite sample, but we do have the asymptotic variance (the variance of the limiting normal distribution).

Since θ_0 could be any point in Θ, we'd like our analysis to be valid no matter what its value happens to be. So we can't treat it as fixed the way we did above; we have to let it vary. To this end, it will be useful to have symbols for the mean and variance of a statistic $\widetilde{T}_n : \mathbf{R}^{n \times r} \to \mathcal{T} \subseteq \mathbf{R}^\ell$ in the

scenario in which some arbitrary $\theta \in \Theta$ is the true value:

$$\mathbf{E}^{\theta}(T_n) \coloneqq \int_{\mathbf{R}^{n\times r}} \widetilde{T}_n\left(\{y_i\}_{i=1}^n\right) \mathrm{d}\mu_n^{\theta}$$

$$\mathrm{Var}^{\theta}(T_n) \coloneqq \int_{\mathbf{R}^{n\times r}} \left(\widetilde{T}_n\left(\{y_i\}_{i=1}^n\right) - \mathbf{E}^{\theta}(T_n)\right)\left(\widetilde{T}_n\left(\{y_i\}_{i=1}^n\right) - \mathbf{E}^{\theta}(T_n)\right)^{\top} \mathrm{d}\mu_n^{\theta}.$$

(Recall that μ_n^{θ} denotes the law of $\{y_i\}_{i=1}^n$ when θ is the true parameter value.) With this notation, we can easily define a function $\mathcal{I} : \Theta \to \mathbf{R}^{k\times k}$ that maps an arbitrary $\theta \in \Theta$ into what the information matrix $-A = B$ would be if θ were the true value:

$$\mathcal{I}(\theta) \coloneqq \mathbf{E}^{\theta}\left(\left[\nabla \ell_1^1(\theta)\right]\left[\nabla \ell_1^1(\theta)\right]^{\top}\right) = -\mathbf{E}^{\theta}\left(\nabla^2 \ell_1^1(\theta)\right).$$

We've seen that when θ is the true value, the asymptotic variance of the MLE is $\mathcal{I}(\theta)^{-1}$. Let $V : \Theta \to \mathbf{R}^{k\times k}$ be the asymptotic variance (at each θ) of some alternative estimator that is also consistent and asymptotically normal. We say that the MLE is asymptotically (weakly) more efficient at $\theta \in \Theta$ than the alternative estimator iff $V(\theta) - \mathcal{I}(\theta)^{-1}$ is positive semidefinite (psd). $V(\theta) - \mathcal{I}(\theta)^{-1}$ being psd says precisely that every linear combination of $\widehat{\theta}_n$ has weakly lower variance than the same linear combination of the alternative estimator. We say that the MLE is asymptotically more efficient (simpliciter) than the alternative estimator iff it is asymptotically more efficient at every $\theta \in \Theta$. Finally, we say that the MLE is asymptotically efficient within some class of estimators iff it is asymptotically more efficient than each other estimator in the class.

A natural conjecture is that the MLE is asymptotically efficient within the class of consistent and asymptotically normal estimators. This conjecture was long believed to be true, but the following (species of) example, due to Hodges, shows that it is not.

Example 22 (super-efficiency)**.** Let $\Theta \subseteq \mathbf{R}$ and let the true value be θ_0. Let $\widehat{\theta}_n$ be the (consistent and asymptotically normal) MLE. Define another estimator $\widetilde{\theta}_n$ by

$$\widetilde{\theta}_n = \begin{cases} 0 & \text{if } |\widehat{\theta}_n| < n^{-1/4} \\ \widehat{\theta}_n & \text{if } |\widehat{\theta}_n| \geq n^{-1/4}. \end{cases}$$

For $\theta_0 \neq 0$, $\widetilde{\theta}_n$ is asymptotically equivalent to the MLE $\widehat{\theta}_n$, so has the same asymptotic variance. But when $\theta_0 = 0$, consistency means that for large n, with high probability, $|\widehat{\theta}_n| < n^{-1/4}$, in which case the asymptotic variance is zero. $\theta_0 = 0$ is called a point of super-efficiency of the estimator $\widetilde{\theta}_n$, and an estimator with super-efficiency points is called a super-efficient estimator.

So the MLE is not efficient: there exists another estimator $\widetilde{\theta}_n$ whose variance is at least as low for every $\theta_0 \in \Theta$ and strictly lower for some $\theta_0 \in \Theta$. In particular, it is always possible to construct a super-efficient estimator that efficiency-dominates the MLE.

But we can salvage a great deal here. Le Cam showed that the set $\Theta_e \subseteq \Theta$ of super-efficiency points of any given super-efficient estimator must have Lebesgue measure zero. This is a formal sense in which we might call super-efficiency a pathology. Moreover, super-efficient estimators turn out to be the only obstacle to calling the MLE asymptotically efficient: Le Cam also showed that the MLE is asymptotically efficient within the class of non-super-efficient, consistent and asymptotically normal estimators. Similarly, if we redefine 'asymptotically more efficient than' to require only that the variance is smaller for a subset of Θ of full Lebesgue measure, then the MLE is asymptotically efficient within the class of all consistent and asymptotically normal estimators.

There's a bunch of other results along the same lines. In efficiency settings, $\mathcal{I}^{-1}$ is called the Cramér–Rao lower bound. A basic result is that no unbiased estimator can achieve asymptotic variance lower than the Cramér–Rao bound at every $\theta \in \Theta$. The MLE is not unbiased in general, but it is asymptotically unbiased under regularity conditions, so this result gives us efficiency of the MLE in the class of consistent, asymptotically normal and asymptotically unbiased estimators. Another result is that the Cramér–Rao bound is a lower bound on the variance of the any consistent and *uniformly* asymptotically normal estimator. The latter means that weak convergence to a normal is uniform in a certain sense.

The most general class of theorems on asymptotic efficiency is exemplified by the Hájek–Le Cam asymptotic minimax theorem (see e.g. Ibragimov and Has'minskii (1981, Theorem 12.1)). Such results give a set of conditions under which an estimator is efficient within some class, and the MLE satisfies these assumptions under regularity conditions.

9 Hypothesis testing

Official reading: Amemiya (1985, sec. 4.5.1) and Newey and McFadden (1994, sec. 9).

9.1 Preliminaries

So far, we've focused on the problem of point estimation: given DGP parameterised by θ_0, we try to find a way of learning the value of θ_0. Now let's turn that on its head: we start with a subset $\Theta_0 \subseteq \Theta$, and want to learn whether $\theta_0 \in \Theta_0$. Intuitively, we should be able to do this using our consistent and asymptotically normal extremum estimator $\widehat{\theta}_n$, for if the hypothesis is true then $\widehat{\theta}_n \in \Theta_0$ with high probability, and if the hypothesis is false then $\widehat{\theta}_n \notin \Theta_0$ with high probability.

The formal setup is as follows. We wish to test the null hypothesis (H_0) that $\theta_0 \in \Theta_0$ against the alternative hypothesis (H_1) that $\theta_0 \notin \Theta_0$. To implement this, we use a test statistic. A real-valued statistic is just a measurable function $T_n : \mathbf{R}^{n \times r} \to \mathbf{R}$ mapping data into the real line; as usual we will work directly with the random variables $T_n(\omega) \coloneqq \widetilde{T}_n\left(\{y_i(\omega)\}_{i=1}^n\right)$.

A rejection region for the statistic T_n is some subset R_n of $\mathbf{R}$. Our testing procedure is as follows:

$$\begin{aligned} &\text{If } T_n \in R_n \text{ then we reject } H_0; \\ &\text{otherwise we fail to reject } H_0. \end{aligned}$$

In many cases, the rejection regions $\{R_n\}$ will take the form

$$R_n = [c_n, \infty) \quad \text{or} \quad R_n = (-\infty, -c_n] \cap [c_n, \infty).$$

The tests discussed below have rejection regions of the former kind; the t test has a rejection region of the latter kind. In either case, we call c_n a critical value for the test.

We'll want a sensible way of choosing the rejection region, or else the test will be useless. There are two problems we have to worry about: rejecting H_0 when it's actually true (type-I error), and failing to reject H_0 when it's false (type-II error). It is much easier to control the probability of type-I error of a test, as we'll see momentarily, so that's what we'll focus on in determining our critical values.

So fix a desired probability α of type-I error; α is called the (desired) size of the test. Then for any fixed $\theta \in \Theta_0$, we can read off the rejection region $R_{T_n}^{\alpha}(\theta)$ from the (approximate) distribution of T_n under $\theta_0 = \theta$. For

the approximate distributions we'll consider, we can summarise the rejection region using a critical value $c^\alpha_{T_n}(\theta_0)$. (E.g. for rejection regions of the form $\left[c^\alpha_{T_n}(\theta_0), \infty\right)$, the critical value is the $(1-\alpha)$th quantile of the (approximate) distribution of T_n.)

Without further restrictions, this test is infeasible because the critical values depend on the unknown value of θ_0. The reason why type-I error is easy to control is that since Θ_0 is generally a fairly small set, it will often be the case that our approximate distribution of T_n under $\theta_0 = \theta \in \Theta_0$ will be the same for each $\theta \in \Theta_0$. That is, we have an approximate distribution that holds under H_0 irrespective of which particular $\theta \in \Theta_0$ is the true value. This gives us critical values c^α_n that can be obtained without knowledge of θ_0.

Useful jargon: a statistic T_n is pivotal under H_0 iff its distribution under H_0 (for finite n) does not depend on θ_0. It is asymptotically pivotal under H_0 iff its asymptotic distribution under H_0 does not depend on θ_0. We are interested in the latter case, since we rarely encounter statistics whose finite-sample distribution can be derived, never mind shown to be independent of θ_0. The previous paragraph says, in short, that our tests will be based on a statistics that are asymptotically pivotal under H_0.

Actually, we will be able to simplify the critical values further. In general, our critical values c^α_n will depend on n since our approximate distribution for T_n might vary with the sample size. But since our approximating distribution will be the asymptotic distribution (obviously independent of n), we can use critical values c^α do not depend on n. However, our asymptotic distribution actually justifies the use of any critical values $c^\alpha + \mathrm{o_p}(1)$ (anything that is asymptotically equivalent to c^α). If we choose to add some n-dependent, asymptotically vanishing term to the critical values, we may obtain a better approximation to the distribution of T_n under H_0 in a finite sample. Many such finite-sample corrections have been proposed for well-known tests, usually justified by Monte Carlo evidence.

So far, we have only dealt with type-I error. But type-II error is also a concern: a test with known (asymptotic) size α but high probability of type-II error will not be able to discriminate between the null and alternative hypotheses. The power $P^\alpha_n(\theta)$ of a test of size α against the (specific) alternative $\theta_0 = \theta \in \Theta^{\mathrm{C}}_0$ is the probability of rejecting H_0 when $\theta_0 = \theta$ (a particular way in which H_1 can be true). (So $P^\alpha_n(\theta)$ is one minus the probability of type-II error.) We say that our test is consistent against the alternative $\theta_0 = \theta$ iff $P^\alpha_n(\theta) \longrightarrow 1$ as $n \longrightarrow \infty$. The test is consistent iff it is consistent against every

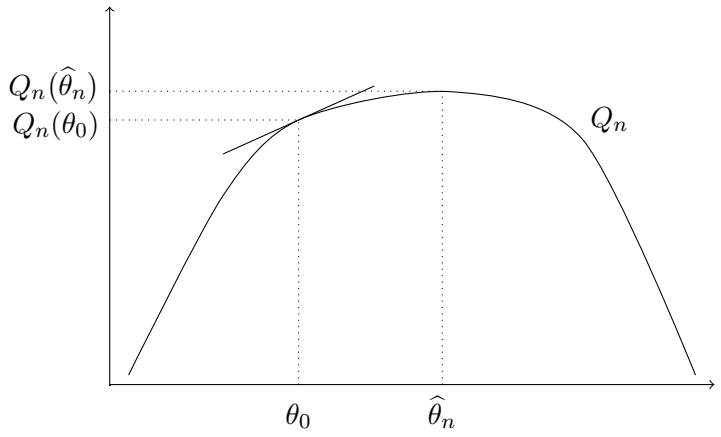

Figure 2 – Graphic illustration of the trinity of test statistics.

alternative $\theta \in \Theta_0^{\mathrm{c}}$.[71]

The power function P_n^α is generally not constant; intuitively, when H_0 is 'nearly' true (θ_0 is close to Θ_0), power will tend to be low. Another thing: you might wonder why we allow type-I errors at all; why not set $\alpha = 0$? The answer is that then power would be zero, making the test useless.

We can classify null hypotheses into two varieties. A simple hypothesis is one for which Θ_0 is a singleton; unsurprisingly, this makes things a lot easier. A composite hypothesis is one for which Θ_0 is not singleton.

9.2 Simple hypotheses

The broad idea behind the three tests in this section is as follows. Suppose you have a consistent and asymptotically normal extremum estimator $\widehat{\theta}_n$ of θ_0, formed by maximising Q_n. Your simple null hypothesis is that the true value is θ_0. (Recall that θ_0 is whatever value uniquely maximised Q, the nonstochastic function satisfying $n^{-1}Q_n \xrightarrow{\text{a.s.}} Q$ uniformly on Θ.) This situation is depicted in Figure 2.

Under the null, $\widehat{\theta}_n - \theta_0$ (the horizontal distance in the figure) had better be small; this forms the basis of the Wald test. Similarly, $Q_n(\widehat{\theta}_n) - Q_n(\theta_0)$ had better be small; this forms the basis of the likelihood ratio (LR) test. Finally, the gradient at $\nabla Q_n(\theta_0)$ at θ_0 had better be close to zero; this is the intuition behind the Lagrange multiplier (LM) test. (The latter is also known as the score test or Rao test.)

[71] Notice that the consistency of a test is a pointwise concept. Uniform consistency is something stronger.

Formally, the three test statistics for a simple null hypothesis with $\Theta_0 = \{\theta_0\}$ are

$$\begin{aligned} LR_n &= 2\left[Q_n(\widehat{\theta}_n) - Q_n(\theta_0)\right] \\ LM_n &= \left[n^{-1/2}\nabla Q_n(\theta_0)\right]^\top \left[\widehat{B}_n(\theta_0)\right]^+ \left[n^{-1/2}\nabla Q_n(\theta_0)\right] \\ W_n &= \left[n^{1/2}(\widehat{\theta}_n - \theta_0)\right]^\top \left[\widehat{A}_n(\widehat{\theta}_n)^+ \widehat{B}_n(\widehat{\theta}_n)\widehat{A}_n(\widehat{\theta}_n)^+\right]^+ \left[n^{1/2}(\widehat{\theta}_n - \theta_0)\right], \end{aligned}$$

where $^+$ denotes the Moore–Penrose pseudo-inverse. Under the hypotheses of our asymptotic normality result for extremum estimators,

$$n^{-1}LR_n \xrightarrow{\text{p}} 0, \quad n^{-1}LM_n \xrightarrow{\text{p}} 0 \quad \text{and} \quad n^{-1}W_n \xrightarrow{\text{p}} 0.$$

In other words, each of the test stats is $\text{o}_\text{p}(n)$. This is the formal sense in which the three test stats must be close to 0 with high probability in a large sample.

To obtain critical values that we can use to perform these tests, we need to derive the asymptotic distributions of the three test statistics. For LM_n and W_n, it should be clear that we just have to apply Slutsky's theorem. (For LR_n, additional structue will be needed.)

Proposition 22. Assume the hypotheses of our asymptotic normality result for extremum estimators (p. 84), and further suppose that $A^{-1}BA^{-1}$ is positive definite.[72] Then $LM_n \xrightarrow{\text{d}} \chi^2(k)$ and $W_n \xrightarrow{\text{d}} \chi^2(k)$.

Proof. $n^{-1/2}\nabla Q_n(\theta_0) \xrightarrow{\text{d}} \mathcal{N}_k(0, B)$ by assumption, and $\widehat{B}_n(\theta_0) \xrightarrow{\text{p}} B$. Hence by Slutsky's theorem,

$$LM_n \xrightarrow{\text{d}} \left[\mathcal{N}_k(0,B)\right]^\top B^{-1} \left[\mathcal{N}_k(0,B)\right] \overset{\text{d}}{=} \left[\mathcal{N}_k(0,I)\right]^\top \left[\mathcal{N}_k(0,I)\right] \overset{\text{d}}{=} \chi^2(k).$$

$n^{1/2}(\widehat{\theta}_n - \theta_0) \xrightarrow{\text{d}} \mathcal{N}_k\left(0, A^{-1}BA^{-1}\right)$ by the asymptotic normality proposition, and $\widehat{A}_n(\widehat{\theta}_n) \xrightarrow{\text{p}} A$ and $\widehat{B}_n(\widehat{\theta}_n) \xrightarrow{\text{p}} B$. $A^{-1}BA^{-1}$ is symmetric and positive definite, hence nonsingular. So by Slutsky's theorem,

$$\begin{aligned} W_n \xrightarrow{\text{d}} \left[\mathcal{N}_k\left(0, A^{-1}BA^{-1}\right)\right]^\top \left[A^{-1}BA^{-1}\right]^{-1} \left[\mathcal{N}_k\left(0, A^{-1}BA^{-1}\right)\right] \\ \overset{\text{d}}{=} \left[\mathcal{N}_k(0,I)\right]^\top \left[\mathcal{N}_k(0,I)\right] \overset{\text{d}}{=} \chi^2(k). \end{aligned}$$

■

In fact, something much stronger is true. Not only are the asymptotic distributions the same; the two statistics are actually numerically close in large samples with high probability.

[72] This just means that the limiting distribution is nondegenerate: no linear combination has variance zero.

Proposition 23. Assume the hypotheses of our asymptotic normality result for extremum estimators (p. 84), and further suppose that $A^{-1}BA^{-1}$ is positive definite. Then $LM_n - W_n = \mathrm{o_p}(1)$.

Partial proof. We will treat the case in which $\widehat{\theta}_n$ lies in the neighbourhood of θ_0 in which the derivatives exist and are continuous. This occurs with probability approaching 1 as $n \to \infty$, but a rigorous proof would proceed more cautiously.

Since $\theta_0 \in \operatorname{int}\Theta$, $\widehat{\theta}_n$ lies in the interior of Θ. Hence it must satisfy the FOC $\nabla Q_n(\widehat{\theta}_n) = 0$. Expanding the derivative around θ_0 using the mean value theorem,

$$0 = \nabla Q_n(\theta_0) + \nabla^2 Q_n(\widetilde{\theta}_n)(\widehat{\theta}_n - \theta_0),$$

where the mean value $\widetilde{\theta}_n$ lies between θ_0 and $\widetilde{\theta}_n$, so that $\widetilde{\theta}_n \xrightarrow{\mathrm{p}} \theta_0$ since $\widehat{\theta}_n$ is consistent. Rearranging and using the definition $\widehat{A}_n = n^{-1}\nabla^2 Q_n$ from section 7.5 (p. 88),

$$-\left[n^{-1/2}\nabla Q_n(\theta_0)\right] = \widehat{A}_n(\widetilde{\theta}_n)\left[n^{1/2}(\widehat{\theta}_n - \theta_0)\right].$$

On the one hand, we can premultiply this by $\widehat{B}_n(\widetilde{\theta}_n)^+$ to get

$$-\widehat{B}_n(\widetilde{\theta}_n)^+\left[n^{-1/2}\nabla Q_n(\theta_0)\right] = \widehat{B}_n(\widetilde{\theta}_n)^+\widehat{A}_n(\widetilde{\theta}_n)\left[n^{1/2}(\widehat{\theta}_n - \theta_0)\right]. \quad (7)$$

On the other hand, we can transpose it to get

$$-\left[n^{-1/2}\nabla Q_n(\theta_0)\right]^\top = \left[n^{1/2}(\widehat{\theta}_n - \theta_0)\right]^\top \widehat{A}_n(\widetilde{\theta}_n) \quad (8)$$

(using the symmetry of the second derivative, which holds by Young's theorem). Premultiplying (7) by (8),

$$\begin{aligned}&\left[n^{-1/2}\nabla Q_n(\theta_0)\right]^\top \widehat{B}_n(\widetilde{\theta}_n)^+\left[n^{-1/2}\nabla Q_n(\theta_0)\right] \\ &\qquad = \left[n^{1/2}(\widehat{\theta}_n - \theta_0)\right]^\top \widehat{A}_n(\widetilde{\theta}_n)\widehat{B}_n(\widetilde{\theta}_n)^+\widehat{A}_n(\widetilde{\theta}_n)\left[n^{1/2}(\widehat{\theta}_n - \theta_0)\right].\end{aligned}$$

By Slutsky's theorem, the LHS is

$$\begin{aligned}&\left[n^{-1/2}\nabla Q_n(\theta_0)\right]^\top \widehat{B}_n(\widetilde{\theta}_n)^+\left[n^{-1/2}\nabla Q_n(\theta_0)\right] \\ &\qquad = \left[n^{-1/2}\nabla Q_n(\theta_0)\right]^\top \widehat{B}_n(\theta_0)^+\left[n^{-1/2}\nabla Q_n(\theta_0)\right] + \mathrm{o_p}(1) \\ &\qquad = LM_n + \mathrm{o_p}(1).\end{aligned}$$

Also by Slutsky's theorem, the RHS is

$$\begin{aligned}
&\left[n^{1/2}(\widehat{\theta}_n - \theta_0)\right]^\top \widehat{A}_n(\widetilde{\theta}_n)\widehat{B}_n(\widetilde{\theta}_n)^+\widehat{A}_n(\widetilde{\theta}_n)\left[n^{1/2}(\widehat{\theta}_n - \theta_0)\right] \\
&= \left[n^{1/2}(\widehat{\theta}_n - \theta_0)\right]^\top \widehat{A}_n(\widehat{\theta}_n)\widehat{B}_n(\widehat{\theta}_n)^+\widehat{A}_n(\widehat{\theta}_n)\left[n^{1/2}(\widehat{\theta}_n - \theta_0)\right] + \mathrm{o_p}(1) \\
&= \left[n^{1/2}(\widehat{\theta}_n - \theta_0)\right]^\top \left[\widehat{A}_n(\widehat{\theta}_n)^+\widehat{B}_n(\widehat{\theta}_n)\widehat{A}_n(\widehat{\theta}_n)^+\right]^+\left[n^{1/2}(\widehat{\theta}_n - \theta_0)\right] + \mathrm{o_p}(1) \\
&= W_n + \mathrm{o_p}(1).
\end{aligned}$$

Together, this says that $LM_n + \mathrm{o_p}(1) = W_n + \mathrm{o_p}(1)$, or equivalently $LM_n - W_n = \mathrm{o_p}(1)$. ■

What about the likelihood ratio stat? Let's just see how far we can get. As in the previous two proofs, let's proceed by simply assuming that $\widehat{\theta}_n$ lies in the neighbourhood of θ_0 in which the derivatives exist and are continuous (which is the case with probability approaching 1 as $n \longrightarrow 1$). A second-order mean value expansion of $Q_n(\theta_0)$ around $\widehat{\theta}_n$ yields

$$Q_n(\theta_0) - Q_n(\widehat{\theta}_n) = \nabla Q_n(\widehat{\theta}_n)(\theta_0 - \widehat{\theta}_n) + \frac{1}{2}(\theta_0 - \widehat{\theta}_n)^\top \nabla^2 Q_n(\widetilde{\theta}_n)(\theta_0 - \widehat{\theta}_n)$$

where the mean value $\widetilde{\theta}_n$ lies between $\widehat{\theta}_n$ and θ_0, hence $\widetilde{\theta}_n \xrightarrow{\mathrm{p}} \theta_0$. By interiority and differentiability, the first-order condition must hold, eliminating the first-order term. Rearranging and using the definition of $\widehat{A}_n$,

$$\begin{aligned}
LR_n = 2\left[Q_n(\widehat{\theta}_n) - Q_n(\theta_0)\right] & \\
&= \left[n^{1/2}(\widehat{\theta}_n - \theta_0)\right]^\top \left[-n^{-1}\nabla^2 Q_n(\widetilde{\theta}_n)\right]\left[n^{1/2}(\widehat{\theta}_n - \theta_0)\right] \\
&= \left[n^{1/2}(\widehat{\theta}_n - \theta_0)\right]^\top \left[-\widehat{A}_n(\widetilde{\theta}_n)\right]\left[n^{1/2}(\widehat{\theta}_n - \theta_0)\right] \\
&= \left[n^{1/2}(\widehat{\theta}_n - \theta_0)\right]^\top \left(-A^{-1}\right)^{-1}\left[n^{1/2}(\widehat{\theta}_n - \theta_0)\right] + \mathrm{o_p}(1).
\end{aligned}$$

But this is the end of the line. The asymptotic variance of $n^{1/2}(\widehat{\theta}_n - \theta_0)$ is $A^{-1}BA^{-1}$, which is not equal to $-A^{-1}$ in general.

But suppose we're in the land of correctly specified MLE. In this case, $A^{-1}BA^{-1} = -A^{-1}$ by the information matrix equality, so we get $LR_n \xrightarrow{\mathrm{d}} \chi^2(k)$. Moreover, we then have

$$\begin{aligned}
&LR_n \\
&= \left[n^{1/2}(\widehat{\theta}_n - \theta_0)\right]^\top \left[\widehat{A}_n(\widehat{\theta}_n)^+\widehat{B}_n(\widehat{\theta}_n)\widehat{A}_n(\widehat{\theta}_n)^+\right]^+\left[n^{1/2}(\widehat{\theta}_n - \theta_0)\right] + \mathrm{o_p}(1) \\
&= W_n + \mathrm{o_p}(1) = LM_n + \mathrm{o_p}(1).
\end{aligned}$$

All of our dreams have come true.

The key to getting the LR stat to be asymptotically equivalent to the Wald and LM stats was the information matrix equality. As we'll see when studying efficient GMM in section 10, generalisations of the information matrix equality are available for certain estimators outside the MLE class. (As in the MLE context, the information matrix equality is tightly linked to efficiency.) We therefore summarise our result in a way that extends beyond the MLE context:

Proposition 24. Assume the hypotheses of our asymptotic normality result for extremum estimators (p. 84), and further suppose that $-A = B$. Then $LR_n = LM_n + \mathrm{o_p}(1) = W_n + \mathrm{o_p}(1)$, and each converges weakly to $\chi^2(k)$.

9.3 Composite hypotheses

Conceptually, not much changes when the null hypothesis is composite. Our null hypothesis is $\theta_0 \in \Theta_0$. (Θ_0 is singleton iff the hypothesis is simple.) To get a handle on this, define the function $h : \Theta \to \mathbf{R}^q$ by $h(\theta) = 0$ iff $\theta \in \Theta_0$. We impose $q \leq k$: you can't have more restrictions than Θ has dimensions. This is wlog since if $q > k$ then either some constraints don't bind at the optimum, or there is no solution. Our null hypothesis can therefore be expressed as $h(\theta_0) = 0$.

We make the crucial (and restrictive) assumption that h is smooth; in particular that it is continuously differentiable in a neighbourhood of θ_0, with $q \times k$ derivative $\mathrm{D}h$. We further assume that $\mathrm{D}h(\theta_0)$ has rank q (full rank). If this were not the case, then there would be redundant constraints encoded in h, so this assumption is wlog.

Define $\widetilde{\theta}_n$ to be a constrained extremum estimator, meaning a solution to

$$\max_{\theta \in \Theta} Q_n(\theta) \quad \text{s.t.} \quad h(\theta) = 0.$$

If the null is true, then with high probability the constraint won't be very binding in a large sample. One implication is that $Q_n(\widehat{\theta}_n) - Q_n(\widetilde{\theta}_n)$ should be small; this is what the LR stat examines. Another implication is that the Lagrange multiplier on the constraint should be small; the LM test checks this. (Hence the name, obviously.) A final implication is that $h(\widehat{\theta}_n)$ should be close to $h(\theta_0) = 0$; this is the basis for the Wald test.

The test statistics are

$$LR_n = 2\left[Q_n(\widehat{\theta}_n) - Q_n(\widetilde{\theta}_n)\right]$$

$$LM_n = \left[n^{-1/2}\nabla Q_n(\widetilde{\theta}_n)\right]^\top \left[\widehat{B}_n(\widetilde{\theta}_n)\right]^+ \left[n^{-1/2}\nabla Q_n(\widetilde{\theta}_n)\right]$$

$$W_n = \left[n^{1/2}h(\widehat{\theta}_n)\right]^\top \left[\mathrm{D}h(\widehat{\theta}_n)\widehat{A}_n(\widehat{\theta}_n)^+\widehat{B}_n(\widehat{\theta}_n)\widehat{A}_n(\widehat{\theta}_n)^+\mathrm{D}h(\widehat{\theta}_n)^\top\right]^+ \left[n^{1/2}h(\widehat{\theta}_n)\right].$$

As far as the asymptotic theory is concerned, the filling in the LM and Wald sandwiches can be any consistent estimators. The advantage of these particular choices is that we can compute LM_n without having to estimate the unrestricted model, and we can compute W_n without having to estimate the restricted model. This flexibility can be a godsend when one $\widehat{\theta}_n$ and $\widetilde{\theta}_n$ is hard to compute. In contrast, we have to estimate both the restricted and unrestricted models to compute LR_n.

Unluckily for us, the Monte Carlo evidence is that the χ^2 approximation is much better for LR_n than it is for either LM_n or W_n. There's some intuition behind this: the later two involve an intermediate step of variance estimation, which introduces an additional source of noise into the test statistics. This noise vanishes asymptotically (of course), but in a finite sample it affects the distribution. Moreover, this noise in the filling of the LM and Wald sandwiches might be correlated with the bread. Again this correlation vanishes asymptotically since the filling converges in probability to a nonstochastic limit, but it may be important in finite samples.

The asymptotic distribution of W_n is basically immediate from the delta method. As in the previous section, consistency of $\widehat{\theta}_n$ and continuity of $\widehat{A}_n$ and $\widehat{B}_n$ yield $\widehat{A}_n(\widehat{\theta}_n) = \widehat{A}_n(\theta_0) + \mathrm{o_p}(1)$ and $\widehat{B}_n(\widehat{\theta}_n) = \widehat{B}_n(\theta_0) + \mathrm{o_p}(1)$ by the continuous mapping theorem, and the right-hand sides converge in probability to A and B (respectively) by Khinchine's WLLN (p. 56). By the delta method (p. 49),

$$n^{1/2}h(\widehat{\theta}_n) = n^{1/2}\left(h(\widehat{\theta}_n) - h(\theta_0)\right) \xrightarrow{\mathrm{d}} \mathcal{N}_q\left(0, \mathrm{D}h(\theta_0)A^{-1}BA^{-1}\mathrm{D}h(\theta_0)^\top\right).$$

Putting this all together using Slutsky's theorem,

$$\begin{aligned}
W_n = & \left[n^{1/2}h(\widehat{\theta}_n)\right]^\top \left[\mathrm{D}h(\theta_0)A^{-1}BA^{-1}\mathrm{D}h(\theta_0)^\top\right]^{-1} \left[n^{1/2}h(\widehat{\theta}_n)\right] + \mathrm{o_p}(1) \\
\xrightarrow{\mathrm{d}} & \mathcal{N}_q\left(0, \mathrm{D}h(\theta_0)A^{-1}BA^{-1}\mathrm{D}h(\theta_0)^\top\right)^\top \left[\mathrm{D}h(\theta_0)A^{-1}BA^{-1}\mathrm{D}h(\theta_0)^\top\right]^{-1} \\
& \times \mathcal{N}_q\left(0, \mathrm{D}h(\theta_0)A^{-1}BA^{-1}\mathrm{D}h(\theta_0)^\top\right) \\
\overset{\mathrm{d}}{=} & \mathcal{N}_q\left(0, I\right)^\top \mathcal{N}_q\left(0, I\right) \\
\overset{\mathrm{d}}{=} & \chi^2(q).
\end{aligned}$$

For the LM stat, first observe that $\widetilde{\theta}_n$ is consistent because we're maintaining the assumptions of our consistency result for extremum estimators (p. 78). So show this carefully, note that we can write

$$\Theta_0 = \Theta \cap \{\theta \in \mathbf{R}^k : h(\theta) = 0\}.$$

Θ is compact and $\{\theta \in \mathbf{R}^k : h(\theta) = 0\}$ is closed, so Θ_0 is compact. Q_n is continuous on Θ, hence continuous on Θ_0. $n^{-1}Q_n \xrightarrow{\mathrm{p}} Q$ uniformly on Θ, hence also on Θ_0. Finally, Q has a unique maximum on Θ at θ_0, hence a fortiori has a unique maximum on Θ_0 at θ_0. So the conditions of the consistency proposition (p. 78) hold on Θ_0, whence it follows that $\widetilde{\theta}_n$ is consistent for θ_0.

By consistency of $\widetilde{\theta}_n$ and continuity of $\widehat{B}_n$, the continuous mapping theorem yields $\widehat{B}_n(\widetilde{\theta}_n) = \widehat{B}_n(\theta_0) + \mathrm{o_p}(1)$. Moreover, $\widehat{B}_n(\theta_0) \xrightarrow{\mathrm{p}} B$ by Khinchine's WLLN (p. 56).

Since we're maintaining the hypotheses of the asymptotic normality result for extremum estimators (p. 84), we have that Q_n is differentiable near θ_0 and that θ_0 is interior to Θ. Since $\widehat{\theta}_n$ is consistent for θ_0, it follows that for large n, with high probability, the FOC holds:

$$\nabla Q_n(\widetilde{\theta}_n) = \mathrm{D}h(\widetilde{\theta}_n)^\top \lambda_n$$

where λ_n is a (random) q-vector of Lagrange multipliers. We will proceed in the same informal manner as in our proof of asymptotic normality (p. 84) by behaving as if the FOC always holds.

Since $\widetilde{\theta}_n$ is consistent for θ_0, and since ∇Q_n and $\mathrm{D}h$ are continuous at θ_0 by assumption, the continuous mapping theorem lets us write

$$\nabla Q_n(\theta_0) = \mathrm{D}h(\theta_0)^\top \lambda_n + \mathrm{o_p}(1),$$

and hence

$$n^{-1/2}\nabla Q_n(\theta_0) = \mathrm{D}h(\theta_0)^\top \left(n^{-1/2}\lambda_n\right) + \mathrm{o_p}\left(n^{-1/2}\right).$$

One of the hypotheses of our asymptotic normality result is that the LHS converges in distribution to $\mathcal{N}_k(0, B)$. Hence the RHS must do the same:

$$\mathrm{D}h(\theta_0)^\top \left(n^{-1/2}\lambda_n\right) \xrightarrow{\mathrm{d}} \mathcal{N}_k(0, B),$$

whence it follows that $n^{-1/2}\lambda_n \xrightarrow{\mathrm{d}} \mathcal{N}_q(0, V)$ for some V that satisfies

$$\mathrm{D}h(\theta_0)^\top V \mathrm{D}h(\theta_0) = B.$$

B is a nondegenerate (i.e. positive definite) variance matrix, so is invertible; therefore

$$[\mathrm{D}h(\theta_0)^\top V \mathrm{D}h(\theta_0)]^{-1} = B^{-1}.$$

It follows that

$$\mathrm{D}h(\theta_0)\left[\mathrm{D}h(\theta_0)^\top V \mathrm{D}h(\theta_0)\right]^{-1}\mathrm{D}h(\theta_0)^\top = \mathrm{D}h(\theta_0)B^{-1}\mathrm{D}h(\theta_0)^\top.$$

Now here's a fun fact that you can (very easily) prove at home: because V is invertible (since it's a nondegerate variance matrix) and $\mathrm{D}h(\theta_0)\mathrm{D}h(\theta_0)^\top$ has full rank q (since $\mathrm{D}h(\theta_0)$ has rank q), we have

$$\mathrm{D}h(\theta_0)\left[\mathrm{D}h(\theta_0)^\top V \mathrm{D}h(\theta_0)\right]^{-1}\mathrm{D}h(\theta_0)^\top = V^{-1}. \tag{9}$$

Hence

$$\mathrm{D}h(\theta_0)B^{-1}\mathrm{D}h(\theta_0)^\top = V^{-1}. \tag{10}$$

Putting together the pieces and using (10),

$$\begin{aligned}
LM_n &= \left[n^{-1/2}\nabla Q_n(\widetilde{\theta}_n)\right]^\top \left[\widehat{B}_n(\widetilde{\theta}_n)\right]^+ \left[n^{-1/2}\nabla Q_n(\widetilde{\theta}_n)\right] \\
&= \left[\mathrm{D}h(\theta_0)^\top \left(n^{-1/2}\lambda_n\right)\right]^\top B^{-1} \left[\mathrm{D}h(\theta_0)^\top \left(n^{-1/2}\lambda_n\right)\right] + \mathrm{o_p}(1) \\
&= \left(n^{-1/2}\lambda_n\right)^\top \mathrm{D}h(\theta_0)B^{-1}\mathrm{D}h(\theta_0)^\top \left(n^{-1/2}\lambda_n\right) + \mathrm{o_p}(1) \\
&= \left(n^{-1/2}\lambda_n\right)^\top V^{-1} \left(n^{-1/2}\lambda_n\right) + \mathrm{o_p}(1).
\end{aligned}$$

Applying Slutsky's theorem then yields

$$LM_n \xrightarrow{\mathrm{d}} \mathcal{N}_q(0, V)^\top V^{-1} \mathcal{N}_q(0, V) \overset{\mathrm{d}}{=} \mathcal{N}_q(0, I)^\top \mathcal{N}_q(0, I) \overset{\mathrm{d}}{=} \chi^2(q).$$

Finally, let's turn to the likelihood ratio statistic. All the steps in our derivation in the previous section (for a simple hypothesis) still go through, giving us

$$LR_n = \left[n^{1/2}(\widehat{\theta}_n - \widetilde{\theta}_n)\right]^\top \left(-A^{-1}\right)^{-1} \left[n^{1/2}(\widehat{\theta}_n - \widetilde{\theta}_n)\right] + \mathrm{o_p}(1).$$

A first-order mean-value expansion of $h(\widehat{\theta}_n)$ around $\widetilde{\theta}_n$ yields

$$h(\widehat{\theta}_n) = h(\widetilde{\theta}_n) + \mathrm{D}h(\bar{\theta}_n)(\widehat{\theta}_n - \widetilde{\theta}_n) = \mathrm{D}h(\bar{\theta}_n)(\widehat{\theta}_n - \widetilde{\theta}_n)$$

where the mean value $\bar{\theta}_n$ lies between $\widehat{\theta}_n$ and $\widetilde{\theta}_n$, so is consistent for θ_0. So

$$\begin{aligned} LR_n &= \left[n^{1/2}h(\widehat{\theta}_n)\right]^\top \left[\mathrm{D}h(\bar{\theta}_n)\left(-A^{-1}\right)\mathrm{D}h(\bar{\theta}_n)^\top\right]^+ \left[n^{1/2}h(\widehat{\theta}_n)\right] + \mathrm{o_p}(1) \\ &= \left[n^{1/2}h(\widehat{\theta}_n)\right]^\top \left[\mathrm{D}h(\widehat{\theta}_n)\left(-A^{-1}\right)\mathrm{D}h(\widehat{\theta}_n)^\top\right]^+ \left[n^{1/2}h(\widehat{\theta}_n)\right] + \mathrm{o_p}(1). \end{aligned}$$

As before, this is not asymptotically χ^2 in general. When $-A = B$ (an information matrix equality holds, e.g. MLE), the asymptotic variance of $\widehat{\theta}_n - \widetilde{\theta}_n$ is $-A^{-1}$, which immediately gives us $LR_n = W_n + \mathrm{o_p}(1)$, hence $LR_n \xrightarrow{\mathrm{d}} \chi^2(q)$.

9.4 Power

We've now figured out the asymptotic distributions of our statistics under the null. But as mentioned above, the tests are not much use if their distributions under the alternative are very similar to their null distributions: then power will be low, so we won't be very likely to reject the null even when it is false.

First, we'd like to show that our tests are consistent (consistent against every alternative $\theta_0 = \theta \in \Theta_0^{\mathrm{c}}$).[73] Since the asymptotic null distribution is χ^2, this will require us to show that our test statistics explode under the null; then in a large sample, with high probability, they will be larger than we would expect a χ^2-distributed random variable to be, leading us to reject the null.

Secondly, we'd like to know *how* powerful they are. We could approach this by studying the rate at which the test stats explode, but that turns out not to be fruitful. Instead, we'll consider the behaviour of the test stat when the null is *nearly* true, using a formal device called a Pitman drift. This will allow us to derive the asymptotic distribution of each test stat explicitly, and then we can read off our tests' asymptotic rejection probabilities from the asymptotic distribution under the alternative. This limiting rejection probability (under local-DGP asymptotics) is called the local power.

The exposition in this section will be a little looser. We'll consider only the case of a simple hypothesis, and we'll limit our derivations to the Wald

[73] This is a nice property to have, and we will have it here. But there are many well-known tests that are not consistent against all alternatives. One of these is the information matrix test, which is inconsistent against DGPs for which the information matrix equality holds despite misspecification. (Which can happen.)

statistic. The null hypothesis is that $\theta_0 = \theta^\star$, but actually the true value θ_0 is $\neq \theta^\star$. The Wald statistic for a simple hypothesis from section 9.2 can be written

$$\begin{aligned}
W_n = & \left[n^{1/2}(\widehat{\theta}_n - \theta_0) + n^{1/2}(\theta_0 - \theta^\star)\right]^\top \left[\widehat{A}_n(\widehat{\theta}_n)^+ \widehat{B}_n(\widehat{\theta}_n) \widehat{A}_n(\widehat{\theta}_n)^+\right]^+ \\
& \qquad \times \left[n^{1/2}(\widehat{\theta}_n - \theta_0) + n^{1/2}(\theta_0 - \theta^\star)\right] \\
= & \left[n^{1/2}(\widehat{\theta}_n - \theta_0) + n^{1/2}(\theta_0 - \theta^\star)\right]^\top \left[A^{-1}BA^{-1}\right]^{-1} \\
& \qquad \times \left[n^{1/2}(\widehat{\theta}_n - \theta_0) + n^{1/2}(\theta_0 - \theta^\star)\right] + \mathrm{o_p}(1) \\
= & \left[\mathcal{N}_k\left(0, A^{-1}BA^{-1}\right) + n^{1/2}(\theta_0 - \theta^\star)\right]^\top \left[A^{-1}BA^{-1}\right]^{-1} \\
& \qquad \times \left[\mathcal{N}_k\left(0, A^{-1}BA^{-1}\right) + n^{1/2}(\theta_0 - \theta^\star)\right] + \mathrm{o_p}(1) \\
= & \chi^2(k) + n(\theta_0 - \theta^\star)^\top \left[A^{-1}BA^{-1}\right]^{-1} (\theta_0 - \theta^\star) \\
& \qquad + n^{1/2}(\theta_0 - \theta^\star)^\top \mathcal{N}_k\left(0, \left[A^{-1}BA^{-1}\right]^{-1}\right) + \mathrm{o_p}(1).
\end{aligned}$$

This is a χ^2 plus something deterministic that explodes plus something stochastic that explodes. So W_n definitely explodes, and hence the Wald test is consistent. The consistency of the LR and LM tests can be demonstrated in pretty much the same way.

To show consistency, we used fixed-DGP asymptotics: we held the truth θ_0 (and the alternative hypothesis $\theta^\star$) fixed and studied the behaviour of W_n as n grew large. Consistency means precisely that for any fixed DGP, the rejection probability converges to unity as $n \longrightarrow \infty$. This is analogous to the consistency of an estimator: it eventually ends up in the right place.

But we'd like an asymptotic approximation that tells us how close it is to the right place. For estimators, we did this by blowing up the estimator by $n^{1/2}$ and showing that this blown-up object converges to a normal rather than a point; we then used this normal to approximate the finite-sample distribution of the estimator. By analogy, we'd like to find a way to mess with our test statistic in such a way that it doesn't explode.

The heuristic reason why W_n explodes under fixed-alternative asymptotics is that as n grows larger, any given false null becomes increasingly easy to reject because of lower sampling variation. So to prevent it from exploding, we need the DGP's law to drift with n in such a way that the (fixed) null $\theta^\star$ becomes increasingly hard to reject as n increases. If we make it harder at just the right rate, we might hope that our test stat will converge in distribution rather than vanishing or exploding.

A DGP law μ^{θ_0} for which the null is hard to reject is precisely one such that θ_0 (the truth) is close to the null hypothesis $\theta^\star$ being tested. So we need to let θ_0 drift toward $\theta^\star$ as n increases. To that end, consider a sequence of DGP laws $\{\mu^{\theta_{0,n}}\}$ such that

$$\theta_{0,n} \coloneqq \theta^\star + n^{-1/2}\mu + \mathrm{o}\left(n^{-1/2}\right)$$

for some fixed $\mu \in \mathbf{R}^k$. (Such a sequence is called a Pitman drift.) The algebra is very easy: rearrange to get $n^{1/2}\left(\theta_{0,n} - \theta^\star\right) = \mu + \mathrm{o}(1)$, then substitute:

$$\begin{aligned}
W_n = &\left[n^{1/2}(\widehat{\theta}_n - \theta_{0,n}) + n^{1/2}\left(\theta_{0,n} - \theta_n^\star\right)\right]^\top \left[\widehat{A}_n(\widehat{\theta}_n)^+ \widehat{B}_n(\widehat{\theta}_n)\widehat{A}_n(\widehat{\theta}_n)^+\right]^+ \\
&\qquad \times \left[n^{1/2}(\widehat{\theta}_n - \theta_{0,n}) + n^{1/2}\left(\theta_{0,n} - \theta_n^\star\right)\right] \\
= &\left[n^{1/2}(\widehat{\theta}_n - \theta_{0,n}) + \mu\right]^\top \left[A^{-1}BA^{-1}\right]^{-1} \left[n^{1/2}(\widehat{\theta}_n - \theta_{0,n}) + \mu\right] + \mathrm{o_p}(1) \\
\xrightarrow{\mathrm{d}} &\left[\mathcal{N}_k\left(0, A^{-1}BA^{-1}\right) + \mu\right]^\top \left[A^{-1}BA^{-1}\right]^{-1} \left[\mathcal{N}_k\left(0, A^{-1}BA^{-1}\right) + \mu\right] \\
\overset{\mathrm{d}}{=} &\chi^2\left(k, \lambda(\mu)\right) \quad \text{where } \lambda(\mu) \coloneqq \mu^\top \left[A^{-1}BA^{-1}\right]^{-1} \mu,
\end{aligned}$$

and $\chi^2(k, \lambda)$ denotes the noncentral χ^2 distribution with k degrees of freedom and noncentrality parameter λ. Note that the convergence in distribution above requires a CLT for triangular arrays.[74] What this tells us that when the truth is close to but not equal to $\theta^\star$, the distribution of W_n is well-approximated by a noncentral χ^2.[75]

The first step toward approximating the power of our test is to figure out the asymptotic rejection probability for a given μ. This is straightforward: for a given size α, obtain the critical value c^α from the $\chi^2(k)$ quantile function (inverse CDF), then plug it into the $\chi^2(k, \lambda(\mu))$ CDF:

$$Q^\alpha(\mu) \coloneqq 1 - F_{\chi^2(k,\lambda(\mu))}\left(F^{-1}_{\chi^2(k)}(1-\alpha)\right).$$

$Q^\alpha(\mu)$ is the limiting rejection probability for size α along the sequence of local DGP laws.

[74] A triangular array is a double-indexed sequence $\{\{y_{n,N}\}_{n=1}^N\}_{N=1}^\infty$ of random variables. In our case, the triangular array is 'iid within rows', meaning that $\{y_{n,N}\}_{n=1}^N$ is an iid sequence for each $N \in \mathbf{N}$. Most of our central limit theorems apply to triangular arrays; for example, the Lindeberg–Feller CLT extends without change to triangular arrays that are independent within rows.

[75] I've tried to make this clear, but it's worth repeating: the Pitman drift is just a technique that lets us approximate the power of a test. We're not actually interested in something weird like the asymptotic behaviour of a test when the DGP changes with the sample size to make your life harder.

Recall that what we want is to approximate the power function P_n^α for a fixed n. Expressed more longwindedly, for any $\theta_0 \neq \theta^\star$, we want an approximation to the power $P_n^\alpha(\theta_0)$ of our test of size α of the null $\theta_0 = \theta^\star$ when the sample size is n. So for fixed n and θ_0, choose $\mu \in \mathbf{R}^k$ so that the nth DGP law $\mu^{\theta_{0,n}}$ in the Pitman sequence has $\theta_{0,n} = \theta_0$:

$$\theta_0 = \theta^\star + n^{-1/2}\mu,$$

or $\mu = n^{1/2}(\theta_0 - \theta^\star)$. Then our local-DGP asymptotics tell us that when θ_0 is the truth and the sample size is n, the approximate distribution of the Wald stat is $\chi^2(k, \lambda)$, where the noncentrality parameter is

$$\lambda = \left[n^{1/2}(\theta_0 - \theta^\star)\right]^\top \left[A^{-1}BA^{-1}\right]^{-1} \left[n^{1/2}(\theta_0 - \theta^\star)\right].$$

So the approximate distribution of the Wald stat under the alternative θ_0 will be very different from the central χ^2 distribution from which we compute critical values whenever n is large, θ_0 is very different from the null $\theta^\star$, and the asymptotic variance $A^{-1}BA^{-1}$ is small. Intuitive!

The local power against the alternative θ_0 of our test of size α of the null $\theta^\star$ with sample size n is defined

$$\begin{aligned}\widehat{P}_n^\alpha(\theta_0) &\coloneqq Q^\alpha\left(n^{1/2}(\theta_0 - \theta^\star)\right) \\ &= 1 - F_{\chi^2\left(k, \lambda\left(n^{1/2}(\theta_0 - \theta^\star)\right)\right)}\left(F^{-1}_{\chi^2(k)}(1-\alpha)\right) \quad \text{for each } \theta_0 \in \Theta.\end{aligned}$$

By varying θ_0, we trace out a function $\widehat{P}_n^\alpha : \Theta \to [0,1]$ that approximates the power envelope P_n^α. As we would hope, $\widehat{P}_n^\alpha(\theta^\star) = \alpha$, the asymptotic rejection probability when the null is true; formally this is because $\theta_0 = \theta^\star$ corresponds to $\mu = 0$, which is how we did the asymptotics under the null. As we vary n, we trace out a family $\{\widehat{P}_n^\alpha\}_{n \in \mathbf{N}}$ of approximate power curves corresponding to different sample sizes. (We can also vary α and $\theta^\star$ as desired. I did not index P_n^α and $\widehat{P}_n^\alpha$ by $\theta^\star$, but that was only to avoid clutter.)

To get an idea of what we've obtained, a typical family $\{\widehat{P}_n^\alpha\}_{n \in \mathbf{N}}$ of local power envelopes is depicted in Figure 3. The rejection probability at the null is α because we constructed our test to control size asymptotically. The power against alternatives close to the null is low; formally this is because the noncentrality parameter is then small, so our noncentral χ^2 distribution is close to the (central) χ^2 null distribution. Power increases as we move away from the null, and also increases with sample size.[76,77]

[76]The local power envelope of an arbitrary test need not be symmetric about $\theta^\star$, nor

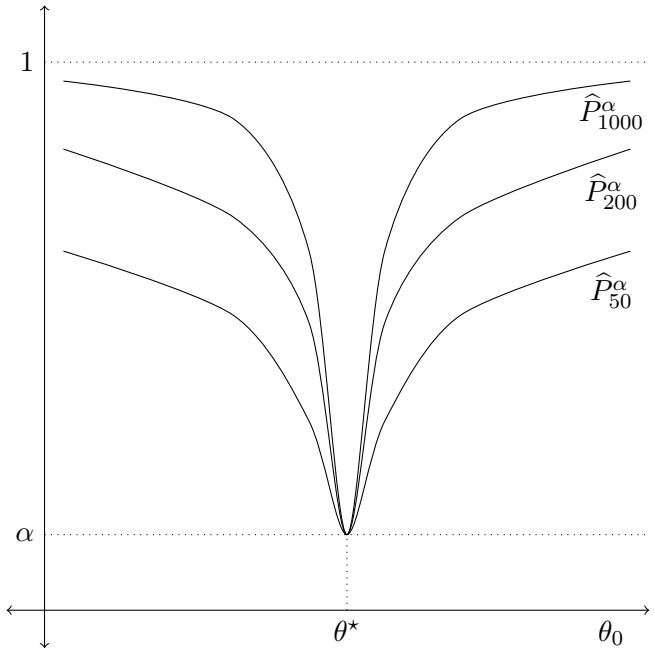

Figure 3 – A family of local power envelopes for $\Theta \subseteq \mathbf{R}$ for a given α and $\theta^\star$. Power is high when θ_0 is far from the null and n is large.

Finally, the derivations in section 9.2 showing that the LR, LM and Wald statistics are asymptotically equivalent (i.e. within $o_p(1)$ of each other) apply with almost no changes to this new asymptotic environment. It follows that the LR and LM statistics are also noncentral-χ^2-distributed, with the same noncentrality parameter, under local-DGP asymptotics. Their local power is therefore the same as that of the Wald test. Moreover, all of the results extend to (smooth) composite hypotheses: the trinity tests are still consistent, and their power can be approximated using a Pitman drift.

---

need it be monotonic on either side of $\theta^\star$. These properties do hold for the Wald test, however, because the noncentrality varies monotonically and symmetrically with θ_0.

[77] It is important that Figure 3 depicts the local power envelopes. The true power envelope family $\{P_n^\alpha\}_{n\in\mathbf{N}}$ may not be as well-behaved, since the DGP could be weird. The true rejection probability under the null will not be α, though it should be 'close' to α when n is large. Moreover, the power curves need not have the nice symmetric and monotonic shape depicted, though again they will 'nearly' have these properties for n large.

10 The generalised method of moments estimator

Official reading: Amemiya (1985, sec. 8.1.1 and 8.2.2) and Newey and McFadden (1994, sec. 2.2.3, 2.5, 3.3 and 4.3).

10.1 Preliminaries

The generalised method of moments (GMM), introduced by Hansen (1982), is a pretty general (haha) technique for estimation and inference that subsumes most parametric methods as special cases. It occupies an important place in the history of econometrics: before GMM, there were many disparate methods such as 2SLS, 3SLS and so on (as you can see by looking at Amemiya (1985), which was written before GMM was introduced). GMM subsumed these classical methods as special cases.

We will treat the case in which the DGP $\{y_i\}$ is iid. We have some (economic) model that gives us the moment conditions $\mathbf{E}(\widetilde{g}(y_1, \theta_0)) = 0$ for some known function $\widetilde{g} : \mathbf{R}^r \times \Theta \to \mathbf{R}^q$.[78] Even if the model imposes additional structure on the data, GMM makes use only of these moment conditions.[79] As usual we define the random functions $g_i : \Theta \to \mathbf{R}^q$ by $g_i(\omega)(\theta) \coloneqq \widetilde{g}(y_i(\omega), \theta)$, allowing us to write the moment conditions as $\mathbf{E}(g_1(\theta_0)) = 0$.

So we have q restrictions on the k-dimensional parameter θ_0. We assume that θ_0 is point-identified, meaning that θ_0 is the unique solution in Θ to $\mathbf{E}(g_1(\theta)) = 0$. A necessary condition for this is (obviously) $q \geq k$. The case $q > k$ is called overidentification; in the lingo, the model gives us overidentifying restrictions on θ_0 in this case.[80] When the moment conditions are misspecified and $q > k$, it should be clear that $\mathbf{E}(g_1(\theta)) = 0$ may not have a solution in Θ. This indicates that we can formulate a specification test for the model by expoiting the overidentifying restrictions; we will work out the details in section 10.5 below.[81]

[78] A rather nice thing is that a lot of economic models give rise to moment conditions; Euler equations are one example. By contrast, it's very rare that a convincing economic model gives rise to a more restrictive statistical model such as a density (as required for maximum likelihood).

[79] This is another advantage of GMM over e.g. ML. Suppose we have a model that gives us a likelihood from which we can derive moment conditions. If the likelihood is misspecified then our estimates will in general be inconsistent. But if the moment conditions hold (a much weaker condition in general), then GMM will give us consistent estimates.

[80] Correspondingly, the case $q = k$ is called 'exact identification'. Then we have enough restrictions to identify θ_0, but no more.

[81] More generally, we could allow $\mathbf{E}(g_1(\theta)) = 0$ to have multiple solutions, in which case we obtain partial identification. Another form of partially-identified GMM comes from replacing the moment equalities with moment *in*equalities. The latter is currently a hot

Every parameteric estimator that we have mentioned so far is a GMM estimator for some choice of $\widetilde{g}$. In particular, any extremum estimator for which the FOC $\nabla Q_n(\widehat{\theta}_n) = 0$ holds is a GMM estimator. Since we're only covering GMM for the iid case, restrict attention to separable criterion functions, so that the FOC can be written

$$\sum_{i=1}^{n} \nabla q_i(\widehat{\theta}_n) = 0.$$

Any such estimator is an exactly-identified ($q = k$) GMM estimator with $g_i = \nabla q_i$.

One example is the MLE, for which $g_i = \ell_i^1$. Another is the nonlinear least squares estimator, for which

$$\widetilde{g}((y, x), \theta) = (y - f(x, \theta))\, x,$$

and the moment condition is derived from the assumption $\mathbf{E}(y|x) = f(x, \theta_0)$. Yet another is the LAD estimator, where

$$\widetilde{g}((y, x), \theta) = 2 \cdot \mathbf{1}\,(y \leq \theta) - 1,$$

and the moment condition is derived from the assumption that the median of the distribution of y conditional on x is $f(x, \theta)$.

When $q = k$ and $\widetilde{g}$ is well-behaved, the sample moment condition

$$n^{-1} \sum_{i=1}^{n} g_i(\theta) = 0$$

will have a unique solution. This value is called the method-of-moments estimator, and the sample moment condition is sometimes called an estimating equation in this context. But what can we do when $q > k$ or $\widetilde{g}$ is ill-behaved? One possibility is to throw away $q - k$ moment conditions and choose $\widetilde{\theta}_n$ to make $n^{-1} \sum_{i=1}^{n} g_i(\theta)$ as close as possible to zero in some metric (exactly equal to zero if $\widetilde{g}$ is well-behaved). More generally, we could set k linear combinations of the sample moments as close as possible to zero (in some metric).

GMM is similar to the latter suggestion. We keep all q moment conditions (instead of combining them into k conditions), and we minimise their distance from zero in some metric. GMM uses a particular metric: a quadratic form

topic in econometric theory.

in the sample moments. Writing $G_n \coloneqq n^{-1/2}\sum_{i=1}^n g_i$, the GMM objective function is

$$Q_n(\theta) \coloneqq \frac{1}{2}G_n(\theta)^\top W_n G_n(\theta) = \frac{1}{2}\left[n^{-1/2}\sum_{i=1}^n g_i(\theta)\right]^\top W_n \left[n^{-1/2}\sum_{i=1}^n g_i(\theta)\right].^{82}$$

The $q\times q$ weight matrix W_n is allowed to be stochastic (a function of the data), but must satisfy $W_n = W + \mathrm{o_p}(1)$ for some symmetric, positive definite, nonstochastic $q \times q$ matrix W. The GMM estimator (for given moment conditions $\mathbf{E}(g_1(\theta_0)) = 0$ and given weight matrix W_n) minimises Q_n. (So the GMM estimator is an extremum estimator.)

10.2 Consistency

Conditions for consistency of GMM can be obtained (essentially) from our consistency results for extremum estimators (pp. 78 and 80). We'll give primitive conditions for strong consistency in the iid case.

Assume that Θ is compact and that each g_i is continuous (so that Q_n is). Further assume that $\mathbf{E}\left(\sup_{\theta\in\Theta}|g_1(\theta)|\right) < \infty$. Then Jennrich's uniform SLLN (p. 58) applies, giving us

$$n^{-1/2}G_n = n^{-1}\sum_{i=1}^n g_i \xrightarrow{\text{a.s.}} g \quad \text{uniformly over } \Theta,$$

where $g : \Theta \to \mathbf{R}^q$ is the nonstochastic function $g(\theta) \coloneqq \mathbf{E}(g_1(\theta))$. Hence

$$n^{-1}Q_n = \frac{1}{2}\left[n^{-1/2}G_n\right]^\top W_n \left[n^{-1/2}G_n\right] \xrightarrow{\text{a.s.}} \frac{1}{2}g^\top W g =: Q \quad \text{uniformly over } \Theta.$$

We assumed point identification, which says precisely that

$$g(\theta) = \mathbf{E}(g_1(\theta)) = 0 \quad \text{iff} \quad \theta = \theta_0.$$

Hence $Q(\theta) = \frac{1}{2}g(\theta)^\top W g(\theta) > 0$ for $\theta \neq \theta_0$ (remember that W is positive definite) and $Q(\theta_0) = 0$, so Q is uniquely minimised at θ_0.

We've now verified all of the conditions of our strong consistency result for extremum estimators (p. 80), so $\widehat{\theta}_n \xrightarrow{\text{a.s.}} \theta_0$.

[82] Most authors, including Joel, scale the objective function differently. I think that my scaling makes by far the most sense. First, having $n^{-1/2}\sum_{i=1}^n g_i(\theta)$ instead of $n^{-1}\sum_{i=1}^n g_i(\theta)$ or $\sum_{i=1}^n g_i(\theta)$ allows us to directly apply our results for extremum estimators. Second, the $1/2$ means that Q_n is similar to the likelihood: the efficient GMM estimator will satisfy a generalised information matrix equality, and the LR stat will be $2\big(Q_n\big(\widetilde{\theta}_n\big) - Q_n\big(\widehat{\theta}_n\big)\big)$.

10.3 Asymptotic normality

Unsurprisingly, we will appeal to our asymptotic normality result for extremum estimators (p. 84). Again we treat the iid case and give primitive conditions.

So maintain the assumptions we imposed to obtain strong consistency in the previous section, and assume that $\theta_0 \in \operatorname{int}\Theta$. Further assume that each g_i is twice continuously differentiable in a neighbourhood of θ_0, so that Q_n is too. Write $\mathrm{D}g_i$ for the $q \times k$ first derivative, and $\mathrm{D}^2 g_i$ for the $q \times k \times k$ second derivative. The latter is a three-dimensional array!

The derivatives of Q_n are

$$\begin{aligned}\nabla Q_n &= \left[\mathrm{D}G_n\right]^\top W_n G_n \\ \nabla^2 Q_n &= \left[\mathrm{D}G_n\right]^\top W_n \left[\mathrm{D}G_n\right] + \left[\mathrm{D}^2 G_n\right]^\top W_n G_n.\end{aligned}$$

Don't forget that $\left[\mathrm{D}^2 G_n\right]^\top W_n G_n$ is the product of a three-dimensional array with a matrix, yielding a matrix. My notation for this is not ideal (e.g. it doesn't tell the reader along what dimensions we're transposing the array), but it won't matter because this term is going to vanish. The first derivative of G_n is of course

$$\mathrm{D}G_n = n^{-1/2} \sum_{i=1}^n \mathrm{D}g_i.$$

We'll want $n^{-1/2}\mathrm{D}G_n$ to converge uniformly to $\mathrm{D}g$, where

$$g(\theta) = \mathbf{E}(g_1(\theta))$$

as in the previous section. We already have that $\{\mathrm{D}g_i\}$ are iid and continuous and that Θ is compact, so we only have to add the assumption that $\mathbf{E}\left(\sup_{\theta\in\Theta}|\mathrm{D}g_1(\theta)|\right) < \infty$. Jennrich's uniform SLLN (p. 58) then tells us that

$$n^{-1/2}\mathrm{D}G_n(\theta) = n^{-1}\sum_{i=1}^n \mathrm{D}g_i(\theta) \xrightarrow{\text{a.s.}} \mathrm{D}g \quad \text{uniformly over } \Theta.$$

The uniform boundedness assumption on g_1 (that we used to derive consistency) is sufficient for the dominated convergence theorem, and hence for the interchanging of integration and differentiation. Therefore

$$\mathrm{D}g(\theta) = \frac{\partial}{\partial\theta^\top}\mathbf{E}(g_1(\theta)) = \mathbf{E}\left(\frac{\mathrm{d}}{\mathrm{d}\theta^\top}g_1(\theta)\right) = \mathbf{E}\left(\mathrm{D}g_1(\theta)\right).$$

Good to know.

It follows by the dominated convergence theorem and independence that

$$\begin{aligned}
B &= \lim_{n\to\infty} \mathbf{E}\left(n^{-1}[\nabla Q_n(\theta_0)][\nabla Q_n(\theta_0)]^\top\right) \\
&= \lim_{n\to\infty} \mathbf{E}\Big(\left[\left[n^{-1/2}\mathrm{D}G_n(\theta_0)\right]^\top W_n G_n(\theta_0)\right] \\
&\qquad\qquad \times \left[\left[n^{-1/2}\mathrm{D}G_n(\theta_0)\right]^\top W_n G_n(\theta_0)\right]^\top \Big) \\
&= \lim_{n\to\infty} \mathbf{E}\left(\left[n^{-1/2}\mathrm{D}G_n(\theta_0)\right]^\top W_n G_n(\theta_0) G_n(\theta_0)^\top W_n^\top \left[n^{-1/2}\mathrm{D}G_n(\theta_0)\right]\right) \\
&= [\mathrm{D}g(\theta_0)]^\top W \left\{\lim_{n\to\infty} \mathbf{E}\left(G_n(\theta_0)G_n(\theta_0)^\top\right)\right\} W\,[\mathrm{D}g(\theta_0)] \\
&= [\mathrm{D}g(\theta_0)]^\top W \left\{\lim_{n\to\infty} \mathbf{E}\left(n^{-1}\sum_{i=1}^n\sum_{j=1}^n g_i(\theta_0)g_j(\theta_0)^\top\right)\right\} W\,[\mathrm{D}g(\theta_0)] \\
&= [\mathrm{D}g(\theta_0)]^\top W \left\{\lim_{n\to\infty} \mathbf{E}\left(n^{-1}\sum_{i=1}^n g_i(\theta_0)g_i(\theta_0)^\top\right)\right\} W\,[\mathrm{D}g(\theta_0)] \\
&= [\mathrm{D}g(\theta_0)]^\top W \mathbf{E}\left(g_1(\theta_0)g_1(\theta_0)^\top\right) W\,[\mathrm{D}g(\theta_0)]\,.
\end{aligned}$$

(The penultimate equality holds by independence.)

Now let's do something similar for $\nabla^2 Q_n$. Add the Jennrich boundedness condition $\mathbf{E}\left(\sup_{\theta\in\Theta}|\mathrm{D}^2 g_1(\theta)|\right) < \infty$ on the second derivative; then $n^{-1/2}\mathrm{D}^2 G_n \xrightarrow{\text{a.s.}} \mathrm{D}^2 g$ uniformly by Jennrich's uniform SLLN. So

$$\begin{aligned}
n^{-1}\nabla^2 Q_n &= \left[n^{-1/2}\mathrm{D}G_n\right]^\top W_n \left[n^{-1/2}\mathrm{D}G_n\right] + \left[n^{-1/2}\mathrm{D}^2 G_n\right]^\top W_n \left[n^{-1/2}G_n\right] \\
&\xrightarrow{\text{a.s.}} [\mathrm{D}g]^\top W\,[\mathrm{D}g] + \left[\mathrm{D}^2 g\right]^\top W g \quad \text{uniformly over } \Theta.
\end{aligned}$$

So using $g(\theta_0) = 0$, we have

$$\begin{aligned}
n^{-1}\nabla^2 Q_n(\theta_0) &\xrightarrow{\text{a.s.}} [\mathrm{D}g(\theta_0)]^\top W\,[\mathrm{D}g(\theta_0)] + \left[\mathrm{D}^2 g(\theta_0)\right]^\top W g(\theta_0) \\
&= [\mathrm{D}g(\theta_0)]^\top W\,[\mathrm{D}g(\theta_0)]\,.
\end{aligned}$$

Hence by the dominated convergence theorem,

$$A = \lim_{n\to\infty} \mathbf{E}\left(n^{-1}\nabla^2 Q_n(\theta_0)\right) = [\mathrm{D}g(\theta_0)]^\top W\,[\mathrm{D}g(\theta_0)]\,.$$

The fact that

$$n^{-1}\nabla^2 Q_n \xrightarrow{\text{a.s.}} [\mathrm{D}g]^\top W\,[\mathrm{D}g] + \left[\mathrm{D}^2 g\right]^\top W g \quad \text{uniformly over } \Theta$$

gives us that for any sequence $\{\theta_n\}$ of random k-vectors with $\theta_n \xrightarrow{\mathrm{p}} \theta_0$, we have

$$n^{-1}\nabla^2 Q_n(\theta_n) \xrightarrow{\mathrm{p}} [\mathrm{D}g(\theta_0)]^\top W [\mathrm{D}g(\theta_0)] = A.$$

You may recall that this is what assumption (3) in the asymptotic normality result for extremum estimators (p. 84) requires.

To verify the fourth and final hypothesis of the asymptotic normality result, first observe that

$$\begin{aligned} n^{-1/2}\nabla Q_n(\theta_0) &= \left[n^{-1/2}\mathrm{D}G_n(\theta_0)\right]^\top W_n [G_n(\theta_0)] \\ &= [\mathrm{D}g(\theta_0)]^\top W \left[n^{-1/2}\sum_{i=1}^n g_i(\theta_0)\right] + \mathrm{o_p}(1). \end{aligned}$$

$\{g_i(\theta_0)\}$ are iid with mean zero and variance $\mathbf{E}\left(g_1(\theta_0)g_1(\theta_0)^\top\right)$, so by the multivariate Lindeberg–Lévy CLT and Slutsky's theorem,

$$\begin{aligned} n^{-1/2}\nabla Q_n(\theta_0) &\xrightarrow{\mathrm{d}} [\mathrm{D}g(\theta_0)]^\top W \mathcal{N}_q\left(0, \mathbf{E}\left(g_1(\theta_0)g_1(\theta_0)^\top\right)\right) \\ &\overset{\mathrm{d}}{=} \mathcal{N}_q\left(0, [\mathrm{D}g(\theta_0)]^\top W \mathbf{E}\left(g_1(\theta_0)g_1(\theta_0)^\top\right) W [\mathrm{D}g(\theta_0)]\right) \\ &\overset{\mathrm{d}}{=} \mathcal{N}_q\left(0, B\right). \end{aligned}$$

So we've satisfied all the conditions of our asymptotic normality result (p. 84). Hence

$$n^{1/2}(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, A^{-1}BA^{-1}\right).$$

The asymptotic variance is the beast

$$\begin{aligned} A^{-1}BA^{-1} = \left([\mathrm{D}g(\theta_0)]^\top W [\mathrm{D}g(\theta_0)]\right)^{-1} [\mathrm{D}g(\theta_0)]^\top W \mathbf{E}\left(g_1(\theta_0)g_1(\theta_0)^\top\right) \\ \times W [\mathrm{D}g(\theta_0)] \left([\mathrm{D}g(\theta_0)]^\top W [\mathrm{D}g(\theta_0)]\right)^{-1}. \end{aligned}$$

10.4 Asymptotic efficiency

In general, there's no reason to think that the GMM estimator is asymptotically efficient within (say) the class of consistent and asymptotically normal estimators. For if the moment conditions are uninformative then there's no way to obtain a low-variance estimator.

Instead, we fix the moment conditions and search for the asymptotically most efficient estimator that uses only this information. Since the only parameter in GMM estimation that can be varied is the weight matrix W_n, the theory of efficient GMM amounts to the theory of how to optimally

choose W_n. Of course only the probability limit W of $\{W_n\}$ matters, so really we're choosing W.

It turns out that an (infeasible) optimal choice of weight matrix is $W = \mathbf{E}\left(g_1(\theta_0)g_1(\theta_0)^\top\right)^{-1}$. (We won't show this, but it's straightforward matrix algebra.) So the asymptotic variance of an efficient GMM estimator is

$$\begin{aligned} A^{-1}BA^{-1} &= \left(\left[\mathrm{D}g(\theta_0)\right]^\top W \left[\mathrm{D}g(\theta_0)\right]\right)^{-1} \left[\mathrm{D}g(\theta_0)\right]^\top WW^{-1} \\ &\qquad \times W\left[\mathrm{D}g(\theta_0)\right]\left(\left[\mathrm{D}g(\theta_0)\right]^\top W \left[\mathrm{D}g(\theta_0)\right]\right)^{-1}. \\ &= \left(\left[\mathrm{D}g(\theta_0)\right]^\top W \left[\mathrm{D}g(\theta_0)\right]\right)^{-1} \\ &= \left(\left[\mathrm{D}g(\theta_0)\right]^\top \mathbf{E}\left(g_1(\theta_0)g_1(\theta_0)^\top\right)^{-1}\left[\mathrm{D}g(\theta_0)\right]\right)^{-1} \\ &= A^{-1} = B^{-1}. \end{aligned}$$

There's a close analogy with the efficiency of (correctly-specified) MLE here. The efficient choice of weight matrix (resp. density) causes a bunch of cancellations in the asymptotic variance, leaving us with A^{-1}. (It's A rather than $-A$ because we're minimising Q_n, so $\nabla^2 Q_n$ is *positive* definite here.) Moreover, we obtain $A = B$, a generalisation of the information matrix equality.[83] Clearly B^{-1} is the lower bound on the variance of any GMM estimator using these moment conditions, analogous to the Cramér–Rao bound.

So far we only have an infeasible procedure, since it requires knowledge of $\mathbf{E}\left(g_1(\theta_0)g_1(\theta_0)^\top\right)^{-1}$. But provided we can consistently estimate the latter, we can obtain a feasible estimator. The standard way of doing this (proposed by Hansen (1982)) is called two-step GMM. First pick some arbitrary weight matrix W_n,[84] and obtain the GMM estimate $\widetilde{\theta}_n$. Define the $\Theta \to \mathbf{R}^{q\times q}$ function

$$\widehat{W}_n \coloneqq \left(n^{-1}\sum_{i=1}^n g_i g_i^\top\right)^+,$$

and estimate $\mathbf{E}\left(g_1(\theta_0)g_1(\theta_0)^\top\right)^{-1}$ by $\widehat{W}_n(\widetilde{\theta}_n)$. This estimator is obviously consistent under the maintained assumptions. Now do GMM again using the estimated optimal weight matrix to obtain the two-step GMM estimate $\widehat{\theta}_n$. Since $\widehat{W}_n$ is asymptotically equivalent to $\mathbf{E}\left(g_1(\theta_0)g_1(\theta_0)^\top\right)^{-1}$, $\widehat{\theta}_n$ is

[83] If you remember how we made use of the information matrix equality to derive the asymptotic distribution of the LR statistic, then you should realise that the LR-type statistic $2\big(Q_n\big(\widetilde{\theta}_n\big) - Q_n\big(\widehat{\theta}_n\big)\big)$ will be asymptotically χ^2 in the efficient GMM setting just as in correctly specified MLE.

[84] There's some evidence on what is and is not a good idea for a first-step weight matrix. The identity is very bad; the 2SLS weight matrix is pretty good. (The 2SLS weight matrix is consistent for B^{-1} in the linear homoskedastic case.)

asymptotically equivalent to the infeasible efficient GMM estimator above. The two-step GMM estimator $\widehat{\theta}_n$ is therefore asymptotically efficient within the class of GMM estimators that use these moment conditions.

There are serious finite-sample problems with the two-step procedure. The intermediate step of estimating $\mathbf{E}\,(g_1(\theta_0)g_1(\theta_0)^\top)^{-1}$ lowers the asymptotic variance, but at the cost of introducing an additional source of noise into the estimator. Worse, the noise in estimating the optimal weight matrix is generally correlated with the noise in the sample moments G_n, which introduces bias into the two-step GMM estimator. In Monte Carlo studies, this bias is quite severe for nonlinear DGPs, even in large samples.

There are two obvious ways of dealing with finite-sample bias. On the one hand, we could just eschew two-step GMM in favour of one-step GMM, which is asymptotically less efficient but allows for more reliable inference. (And may be more efficient in a finite sample!) On the other hand, we could incorporate finite-sample corrections to our estimates. Many have been proposed, and some of them are quite helpful.

Another solution is to use a one-step procedure that nevertheless delivers an estimator that is asymptotically equivalent to the infeasible efficient GMM estimator. The continuously-updated (CUE) GMM estimator maximises

$$G_n(\theta)^\top \widehat{W}_n(\theta) G_n(\theta),$$

obviating the need for a second step. And it turns out (perhaps unsurprisingly) to be asymptotically equivalent to the infeasible efficient GMM estimator that uses weight matrix $\mathbf{E}\,(g_1(\theta_0)g_1(\theta_0)^\top)^{-1}$. The lack of an initial step turns out to make a big difference: the Monte Carlo evidence is that the finite-sample behaviour of the CUE GMM estimator is much better than that of two-step GMM. Unfortunately, obtaining the CUE GMM estimator is in general a pretty hard problem: even when $\widetilde{g}$ is linear, it is a nonlinear optimisation problem.

Yet another one-step method that is asymptotically equivalent to efficient GMM is the empirical likelihood (EL) estimator. The EL estimator is the first part of the argmax in the problem

$$\max_{(\theta,p)\in\Theta\times(0,1)^n} \sum_{i=1}^n \ln(p_i) \quad \text{s.t.} \quad n^{-1}\sum_{i=1}^n g_i(\theta) = 0 \quad \text{and} \quad \sum_{i=1}^n p_i = 1.$$

Again, the asymptotics are the same as for efficient GMM, but the finite-sample properties are much better. And again, this estimator is computationally troublesome, as we're now maximising over $k+n$ variables with an additional constraint. There's a broader class of estimators called generalised

empirical likelihood (GEL) estimators which share these good finite-sample properties. We won't delve into (G)EL estimation here; see Imbens (2002) for a nice intro-level survey.

10.5 The J test

Recall that we defined θ_0 to be the unique solution to the population moment condition $\mathbf{E}(g_1(\theta_0)) = 0$. So we know that the model is misspecified (in the precise sense that the moment conditions are inconsistent with each other) if $\mathbf{E}(g_1(\theta)) = 0$ has no solution in Θ.

When $q = k$ (exact identification), provided g_1 is moderately well-behaved, our GMM procedure will force all q sample moment conditions to hold. Intuitively, whenever a moment condition fails, we will have a degree of freedom (a parameter whose estimate we have not yet determined) that we can play with until the last sample moment is satisfied.

But consider the overidentified case $q > k$. Intuitively, we can only force k of the sample moment conditions to hold, leaving another $q - k$ free to do as they please. If the model is correctly specified (the moment conditions hold in the population) then the additional moments should be close to zero. This provides the basis for a specification test, i.e. a test of the null hypothesis that the model is correctly specified.[85] The test is sometimes called the Hansen J test, or else simply the test of overidentifying restrictions.

To that end, consider the LR-type statistic

$$J_n \coloneqq 2Q_n(\widehat{\theta}_n) = G_n(\widehat{\theta}_n)^\top W_n G_n(\widehat{\theta}_n).$$

Though it has the flavour of an LR statistic, note that we are not testing the null hypothesis that the true parameter satisfies some restriction. Instead, our null is that $Q_n(\theta_0)$ is $\mathrm{o_p}(1)$, meaning that all the moment conditions are satisfied at the truth.

To derive the asymptotic distribution under the null, begin with a mean-value expansion:

$$\begin{aligned} G_n(\widehat{\theta}_n) &= G_n(\theta_0) + \left[n^{-1/2}\mathrm{D}G_n(\widetilde{\theta}_n)\right]\left[n^{1/2}(\widehat{\theta}_n - \theta_0)\right] \\ &= G_n(\theta_0) + \left[\mathrm{D}g(\theta_0)\right]\left[n^{1/2}(\widehat{\theta}_n - \theta_0)\right] + \mathrm{o_p}(1) \end{aligned}$$

[85] The information matrix test is another specification test. The J test is conceptually distinct from the information matrix test, however. Given that we derived a generalised information matrix equality for efficient GMM, we could formulate an information matrix test for GMM if desired.

where $\widetilde{\theta}_n$ is the mean value. Recall from our proof of asymptotic normality for extremum estimators (p. 84) that

$$\begin{aligned} n^{1/2}(\widehat{\theta}_n - \theta_0) = & -\left[n^{-1}\nabla^2 Q_n(\widetilde{\theta}_n)\right]^+ \left[n^{-1/2}\nabla Q_n(\theta_0)\right] \\ = & -\left(\left[n^{-1/2}\mathrm{D}G_n(\widetilde{\theta}_n)\right]^\top W_n \left[n^{-1/2}\mathrm{D}G_n(\widetilde{\theta}_n)\right]\right)^+ \\ & \times \left(\left[n^{-1/2}\mathrm{D}G_n(\theta_0)\right]^\top W_n G_n(\theta_0)\right) \\ = & -\left([\mathrm{D}g(\theta_0)]^\top W [\mathrm{D}g(\theta_0)]\right)^{-1} [\mathrm{D}g(\theta_0)]^\top W G_n(\theta_0) + \mathrm{o_p}(1) \end{aligned}$$

where $\widetilde{\theta}_n$ is also a mean value. So

$$\begin{aligned} G_n(\widehat{\theta}_n) = G_n(\theta_0) - [\mathrm{D}g(\theta_0)] \left([\mathrm{D}g(\theta_0)]^\top W [\mathrm{D}g(\theta_0)]\right)^{-1} \\ \times [\mathrm{D}g(\theta_0)]^\top W G_n(\theta_0) + \mathrm{o_p}(1). \end{aligned}$$

You may be tempted to use the matrix algebra result in equation (9) (p. 116) here to write

$$[\mathrm{D}g(\theta_0)] \left([\mathrm{D}g(\theta_0)]^\top W [\mathrm{D}g(\theta_0)]\right)^{-1} [\mathrm{D}g(\theta_0)]^\top = W^{-1},$$

but this is a mistake! The reason is that $[\mathrm{D}g(\theta_0)]\,[\mathrm{D}g(\theta_0)]^\top$ must have full rank in order for this identity to hold. But since $\mathrm{D}g(\theta_0)$ is $q \times k$ and $q > k$ by assumption, $[\mathrm{D}g(\theta_0)]\,[\mathrm{D}g(\theta_0)]^\top$ can have rank at most k!

Instead, premultiply $G_n(\widehat{\theta}_n)$ by $W^{1/2}$:

$$\begin{aligned} & W^{1/2} G_n(\widehat{\theta}_n) \\ & \quad = W^{1/2} G_n(\theta_0) - W^{1/2} [\mathrm{D}g(\theta_0)] \left([\mathrm{D}g(\theta_0)]^\top W [\mathrm{D}g(\theta_0)]\right)^{-1} \\ & \qquad\qquad \times [\mathrm{D}g(\theta_0)]^\top W G_n(\theta_0) + \mathrm{o_p}(1) \\ & \quad = \{I - W^{1/2} [\mathrm{D}g(\theta_0)] \left([\mathrm{D}g(\theta_0)]^\top W [\mathrm{D}g(\theta_0)]\right)^{-1} [\mathrm{D}g(\theta_0)]^\top W^{1/2}\} \\ & \qquad\qquad \times W^{1/2} G_n(\theta_0) + \mathrm{o_p}(1) \\ & \quad = M W^{1/2} G_n(\theta_0) + \mathrm{o_p}(1), \end{aligned}$$

where

$$M \coloneqq I - W^{1/2} [\mathrm{D}g(\theta_0)] \left([\mathrm{D}g(\theta_0)]^\top W [\mathrm{D}g(\theta_0)]\right)^{-1} [\mathrm{D}g(\theta_0)]^\top W^{1/2}$$

M is symmetric (obvious) and idempotent (trivial to verify, just compute MM and cancel terms to recover M).[86]

[86] A square matrix M is idempotent iff $MM = M$.

So the test statistic is

$$\begin{aligned} J_n &= \left[W_n^{1/2} G_n(\widehat{\theta}_n)\right]^\top \left[W_n^{1/2} G_n(\widehat{\theta}_n)\right] \\ &= \left[W^{1/2} G_n(\widehat{\theta}_n)\right]^\top \left[W^{1/2} G_n(\widehat{\theta}_n)\right] + \mathrm{o_p}(1) \\ &= \left[MW^{1/2} G_n(\theta_0)\right]^\top \left[MW^{1/2} G_n(\theta_0)\right] + \mathrm{o_p}(1) \\ &= \left[W^{1/2} G_n(\theta_0)\right]^\top M^\top M \left[W^{1/2} G_n(\theta_0)\right] + \mathrm{o_p}(1) \\ &= \left[W^{1/2} G_n(\theta_0)\right]^\top M \left[W^{1/2} G_n(\theta_0)\right] + \mathrm{o_p}(1). \end{aligned}$$

It's immediate by the Lindeberg–Lévy CLT that

$$G_n(\theta_0) \xrightarrow{\mathrm{d}} \mathcal{N}_q\left(0, \mathbf{E}\left(g_1(\theta_0) g_1(\theta_0)^\top\right)\right).$$

Now suppose (this is important!) that we choose the weight matrix optimally: $W = \mathbf{E}\left(g_1(\theta_0) g_1(\theta_0)^\top\right)^{-1}$. Then we obtain

$$W^{1/2} G_n(\theta_0) \xrightarrow{\mathrm{d}} \mathcal{N}_q\left(0, W^{1/2} W^{-1} W^{1/2}\right) \stackrel{\mathrm{d}}{=} \mathcal{N}_q\left(0, I\right).$$

It then follows by Slutsky's theorem that

$$\begin{aligned} J_n &= \left[W^{1/2} G_n(\theta_0)\right]^\top M \left[W^{1/2} G_n(\theta_0)\right] + \mathrm{o_p}(1) \\ &\xrightarrow{\mathrm{d}} \left[\mathcal{N}_q\left(0, I\right)\right]^\top M \left[\mathcal{N}_q\left(0, I\right)\right]. \end{aligned}$$

Now here's a fact for you: for $\xi \sim \mathcal{N}_q(0, I)$ and any symmetric and idempotent $q \times q$ matrix M,

$$\xi^\top M \xi \sim \chi^2\left(\operatorname{rank} M\right).$$

To see why, eigen-decompose M as $P^\top \Lambda P$, where Λ is a diagonal matrix with the eigenvalues of M on the diagonal and P contains the eigenvectors. Wlog, arrange the rows so that the zero eigenvalues are last. Then

$$\xi^\top M \xi = (P\xi)^\top \Lambda (P\xi) = \sum_{j=1}^{q} \lambda_j \left((P\xi)_j\right)^2 = \sum_{j=1}^{\operatorname{rank} M} \lambda_j \left((P\xi)_j\right)^2.$$

Since M is idempotent, all its eigenvalues are either zero or unity, so

$$\xi^\top M \xi = \sum_{j=1}^{\operatorname{rank} M} \left((P\xi)_j\right)^2.$$

Since $P\xi$ is a linear combination of normals, it is normally distributed. Since the eigenvectors are orthogonal and the components of ξ are independent, the components of $P\xi$ are independent. Moreover

$$\operatorname{Var}((P\xi)_j) = P_{j,\cdot}\operatorname{Var}(\xi_j)P_{j,\cdot}^\top = P_{j,\cdot}P_{j,\cdot}^\top = 1$$

since P is an orthogonal matrix. So we've shown that $\{(P\xi)_j\}_{j=1}^{\operatorname{rank} M}$ are independent standard-normal-distributed random variables. Therefore

$$\xi^\top M\xi = \sum_{j=1}^{\operatorname{rank} M} \left((P\xi)_j\right)^2 \sim \chi^2\left(\operatorname{rank} M\right).$$

So let's compute the rank of our matrix

$$M = I - W^{1/2}\left[\mathrm{D}g(\theta_0)\right]\left(\left[\mathrm{D}g(\theta_0)\right]^\top W\left[\mathrm{D}g(\theta_0)\right]\right)^{-1}\left[\mathrm{D}g(\theta_0)\right]^\top W^{1/2}.$$

Using the fact that the rank and trace of an idempotent matrix are equal, and writing I_m for an $m \times m$ identity matrix to make the dimensions explicit,

$$\begin{aligned}
&\operatorname{rank} M \\
&\quad = \operatorname{tr} M \\
&\quad = \operatorname{tr} I_q - \operatorname{tr} W^{1/2}\left[\mathrm{D}g(\theta_0)\right]\left(\left[\mathrm{D}g(\theta_0)\right]^\top W\left[\mathrm{D}g(\theta_0)\right]\right)^{-1}\left[\mathrm{D}g(\theta_0)\right]^\top W^{1/2} \\
&\quad = \operatorname{tr} I_q - \operatorname{tr}\left(\left[\mathrm{D}g(\theta_0)\right]^\top W\left[\mathrm{D}g(\theta_0)\right]\right)^{-1}\left[\mathrm{D}g(\theta_0)\right]^\top W^{1/2}W^{1/2}\left[\mathrm{D}g(\theta_0)\right] \\
&\quad = \operatorname{tr} I_q - \operatorname{tr} I_k \\
&\quad = q - k.
\end{aligned}$$

So using our fun fact, we obtain

$$J_n \xrightarrow{\mathrm{d}} \left[\mathcal{N}_q\left(0, I\right)\right]^\top M\left[\mathcal{N}_q\left(0, I\right)\right] \stackrel{\mathrm{d}}{=} \chi^2\left(\operatorname{rank} M\right) \stackrel{\mathrm{d}}{=} \chi^2\left(q - k\right).$$

That was the distribution under the null (correct specification). It's pretty clear that the test is consistent, for the same sort of reason as the trinity tests in section 9 were. We can also construct a Pitman drift such that the asymptotic distribution under local-DGP asymptotics is noncentral χ^2.

References

Aliprantis, C. D. & Border, K. C. (2006). *Infinite dimensional analysis: A hitchhiker's guide* (3rd). Berlin: Springer.

Amemiya, T. (1985). *Advanced econometrics*. Cambridge, MA: Harvard University Press.

Billingsley, P. (1995). *Probability and measure* (3rd). New York, NY: Wiley.

Billingsley, P. (1999). *Convergence of probability measures* (2nd). New York, NY: Wiley.

Dudley, R. M. (2004). *Real analysis and probability*. Cambridge: Cambridge University Press.

Durrett, R. (2010). *Probability: Theory and examples* (4th). Cambridge: Cambridge University Press.

Gnedenko, B. V. & Kolmogorov, A. N. (1954). *Limit distributions for sums of independent random variables*. Cambridge, MA: Addison-Wesley.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, *50*(4), 1029–1054.

Ibragimov, I. A. & Has'minskii, R. (1981). *Statistical estimation: Asymptotic theory*. Berlin: Springer.

Imbens, G. W. (2002). Generalized method of moments and empirical likelihood. *Journal of Business & Economic Statistics*, *20*(4), 493–506.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Annals of Mathematical Statistics*, *40*(2), 633–643.

Kim, J. & Pollard, D. (1990). Cube root asymptotics. *Annals of Statistics*, *18*(1), 191–219.

Kolmogorov, A. N. & Fomin, S. V. (1975). *Introductory real analysis*. New York, NY: Dover.

Manski, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics*, *3*(3), 205–228.

Newey, W. K. & McFadden, D. L. (1994). Large sample estimation and hypothesis testing. In R. F. Engle & D. L. McFadden (Eds.), *Handbook of econometrics* (Chap. 36, Vol. 4, pp. 2111–2245). Amsterdam: North-Holland.

Neyman, J. & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16*(1), 1–32.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd). New York, NY: Wiley.

Rosenthal, J. S. (2006). *A first look at rigorous probability theory* (2nd). Singapore: World Scientific.

Serfling, R. J. (1970). Convergence properties of S_n under moment restrictions. *Annals of Mathematical Statistics*, *41*(4), 1235–1248.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. New York, NY: Wiley.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *1*(50), 1–25.

White, H. (2001). *Asymptotic theory for econometricians* (Revised). San Diego, CA: Academic Press.

Williams, D. (1991). *Probability with martingales*. Cambridge: Cambridge University Press.