Chapter 2

Stochastic Processes

2.1 Introduction

A sequence of random vectors is called a stochastic process. Because we are interested in time series, we index sequences by time.

We start with a probability space, namely, a triple $(\Omega, \mathfrak{F}, Pr)$, where $\mathfrak{F}$ is a collection of events (a sigma algebra), and Pr assigns probabilities to events.

Definition 2.1.1. *X is an n-dimensional random vector if $X : \Omega \to \mathbb{R}^n$ such that $\{X \in \mathfrak{b}\}$ is in $\mathfrak{F}$ for any Borel set $\mathfrak{b}$ in $\mathbb{R}^n$.*

A result from measure theory states that for an n-dimensional random vector X to exist, it suffices that $\{X_t \in \mathfrak{o}\}$ is an event in $\mathfrak{F}$ whenever $\mathfrak{o}$ is an open set in $\mathbb{R}^n$.

Definition 2.1.2. *An n-dimensional stochastic process is an infinite sequence of n-dimensional random vectors $\{X_t : t = 0, 1, ...\}$.*

This chapter studies two ways to construct and represent a stochastic process. One way is to let Ω be a collection of infinite sequences of elements of $\mathbb{R}^n$ so that an element $\omega \in \Omega$ is a sequence of vectors $\omega = (\mathbf{r}_0, \mathbf{r}_1, ...)$, where $\mathbf{r}_t \in \mathbb{R}^n$. To construct $\mathfrak{F}$, it helps to observe the following technicalities. Borel sets include open sets, closed sets, finite intersections, and countable unions of such sets. Let $\mathfrak{B}$ be the collection of Borel sets of $\mathbb{R}^n$. Let $\widetilde{\mathfrak{F}}$ denote the collection of all subsets Λ of Ω that can be represented in the following way. For a nonnegative integer ℓ and Borel sets $\mathfrak{b}_0, \mathfrak{b}_1, ..., \mathfrak{b}_\ell$, let

$$\Lambda = \{\omega = (\mathbf{r}_0, \mathbf{r}_1, ...) : \mathbf{r}_j \in \mathfrak{b}_j, j = 0, 1, .., \ell\}. \tag{2.1}$$

Then $\mathfrak{F}$ is the smallest sigma-algebra that contains $\widetilde{\mathfrak{F}}$. By assigning probabilities to events in $\mathfrak{F}$ with Pr, we construct a probability distribution over sequences of vectors.

An alternative approach to constructing a probability space instead starts with an abstract probability space Ω of sample points $\omega \in \Omega$, a collection $\mathfrak{F}$ of events, a probability measure Pr that maps events into numbers between zero and one, and a sequence of vector functions $X_t(\omega), t = 0, 1, 2, \ldots$. We let the time t observation be

$$\mathbf{r}_t = X_t(\omega)$$

for $t = \{0, 1, 2, ...\}$, where $\{\omega \in \Omega : X_t(\omega) \in \mathfrak{b}\}$ is an event in $\mathfrak{F}$. The vector function $X_t(\omega)$ is said to be a random vector. The probability measure Pr implies an *induced probability distribution* for the time t observation X_t over Borel sets $\mathfrak{b} \in \mathfrak{B}$. The induced probability of Borel set $\mathfrak{b}$ is:

$$Pr\{X_t(\omega) \in \mathfrak{b}\}.$$

The measure Pr also assigns probabilities to bigger and more interesting collections of events. For example, consider a composite random vector

$$X^{[\ell]}(\omega) \doteq \begin{bmatrix} X_0(\omega) \\ X_1(\omega) \\ \vdots \\ X_\ell(\omega) \end{bmatrix}$$

and Borel sets $\mathfrak{a}$ in $\mathbb{R}^{n(\ell+1)}$. The joint distribution of $X^{[\ell]}$ induced by Pr over such Borel sets is

$$Pr\{X^{[\ell]} \in \mathfrak{a}\}.$$

Since the choice of ℓ is arbitrary, Pr induces a distribution over a sequence of random vectors $\{X_t(\omega) : t = 0, 1, ...\}$.

The preceding construction illustrates just one way to specify a probability space $(\Omega, \mathfrak{F}, Pr)$ that generates a given induced distribution. There is an equivalence class of the objects $(\Omega, \mathfrak{F}, Pr)$ that implies the same induced distribution.

Two Approaches

The remainder of this chapter describes two ways to construct a stochastic process.

- A first approach specifies a probability distribution over sample points $\omega \in \Omega$, a nonstochastic equation describing the evolution of sample points, and a time invariant measurement function that determines outcomes at a given date as a function of the sample point at that date. Although sample points evolve deterministically, a finite set of observations on the stochastic process fails to reveal a sample point. This approach is convenient for constructing a theory of ergodic stochastic processes and laws of large numbers.

- A second approach starts by specifying induced distributions directly. Here we directly specify a coherent set of joint probability distributions for a list of random vector X_t for all dates t in a finite set of calendar dates T. In this way, we would specify induced distributions for X^ℓ for all nonnegative values of ℓ.

2.2 Constructing Stochastic Processes: I

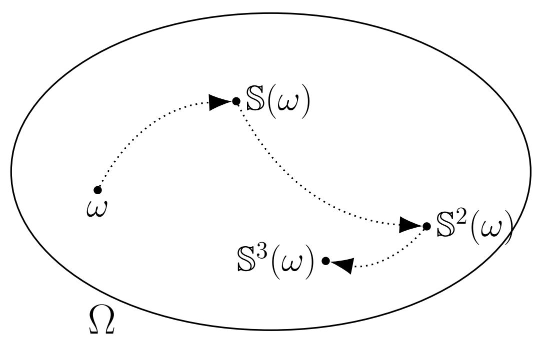

Figure 2.1: The evolution of a sample point ω induced by successive applications of the transformation $\mathbb{S}$. The oval shaped region is the collection Ω of all sample points.

We use the following objects and concepts.[1]

[1]Breiman (1968) is a good reference for these.

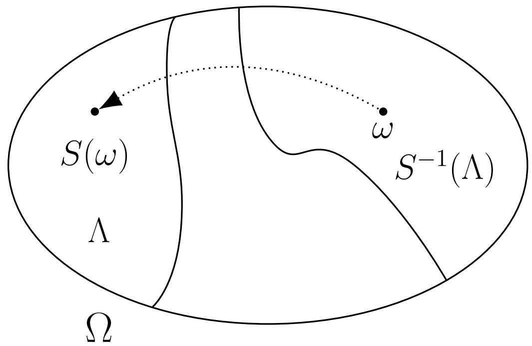

Figure 2.2: An inverse image $\mathbb{S}^{-1}(\Lambda)$ of an event Λ is itself an event; $\omega \in \mathbb{S}^{-1}(\Lambda)$ implies that $\mathbb{S}(\omega) \in \Lambda$.

- A (measurable) transformation $\mathbb{S} : \Omega \to \Omega$ that describes the evolution of a sample point ω.
  - $\{\mathbb{S}^t(\omega) : t = 0, 1, ...\}$ is a deterministic sequence of sample points in Ω.
  - $\mathbb{S}$ has the property that for any event $\Lambda \in \mathfrak{F}$,

$$\mathbb{S}^{-1}(\Lambda) = \{\omega \in \Omega : \mathbb{S}(\omega) \in \Lambda\}$$

    is an event in $\mathcal{F}$, as depicted in figure 2.2.

- An n-dimensional random vector $X(\omega)$ that describes how observations depend on sample point ω.

- A stochastic process $\{X_t : t = 0, 1, ...\}$ constructed via the formula:

$$X_t(\omega) = X[\mathbb{S}^t(\omega)]$$

  or

$$X_t = X \circ \mathbb{S}^t,$$

  where we interpret $\mathbb{S}^0$ as the identity mapping asserting that $\omega_0 = \omega$.

Because a known function $\mathbb{S}$ maps a sample point $\omega \in \Omega$ today into a sample point $\mathbb{S}(\omega) \in \Omega$ tomorrow, the evolution of sample points is *deterministic*: for all $j \geq 1$, ω_{t+j} can be predicted perfectly if we know $\mathbb{S}$ and ω_t.

But we do not observe ω_t at any t. Instead, we observe an $(n \times 1)$ vector $X(\omega)$ that contains incomplete information about ω. We assign probabilities Pr to collections of sample points ω called events, then use the functions $\mathbb{S}$ and X to induce a joint probability distribution over sequences of X's. The resulting stochastic process $\{X_t : 0 = 1, 2, ...\}$ is a sequence of n-dimensional random vectors.

While this way of constructing a stochastic process may seem very special, the following example and section 2.9 show that it is not.

Example 2.2.1. *Let Ω be a collection of infinite sequences of elements of $\mathbf{r}_t \in \mathbb{R}^n$ so that $\omega = (\mathbf{r}_0, \mathbf{r}_1, ...)$, $\mathbb{S}(\omega) = (\mathbf{r}_1, \mathbf{r}_2, ...)$, and $X(\omega) = \mathbf{r}_0$. Then $X_t(\omega) = \mathbf{r}_t$.*

The $\mathbb{S}$ in example 2.2.1 is called the *shift* transformation. The triple $(\Omega, \mathbb{S}, X)$ in example 2.2.1 generates an arbitrary sequence of $(n \times 1)$ vectors. We defer describing constructions of the events $\mathfrak{F}$ and the probability measure Pr until Section 2.9.

2.3 Stationary Stochastic Processes

In a deterministic dynamic system, a *stationary state* or *steady state* remains invariant as time passes. A counterpart of a steady state for a stochastic process is a joint probability distribution that remains invariant over time in the sense that for any finite integer ℓ, the probability distribution of the composite random vector $[X_{t+1}{}', X_{t+2}{}', ..., X_{t+\ell}{}']'$ does not depend on t. We call a process with such a joint probability distribution *stationary*.

Definition 2.3.1. *A transformation $\mathbb{S}$ is **measure-preserving** relative to probability measure Pr if*

$$Pr(\Lambda) = Pr\{\mathbb{S}^{-1}(\Lambda)\}$$

for all $\Lambda \in \mathfrak{F}$.

Notice that the concept of measure preserving is relative to a probability measure Pr. A given transformation $\mathbb{S}$ can be measure preserving relative to one Pr and not relative to others. This feature will play an important role at several points below.

Proposition 2.3.2. *When $\mathbb{S}$ is measure-preserving, probability distribution functions of X_t are identical for all $t \geq 0$.*

Suppose that $\mathbb{S}$ is measure preserving. Given X and an integer $\ell > 1$, form a vector

$$X^{[\ell]}(\omega) \doteq \begin{bmatrix} X_0(\omega) \\ X_1(\omega) \\ \dots \\ X_\ell(\omega) \end{bmatrix}.$$

We can apply Proposition 2.3.2 to $X^{[\ell]}$ to conclude that the joint distribution function of $(X_t, X_{t+1}, ..., X_{t+\ell})$ is independent of t for $t = 0, 1, \dots$. That this property holds for any choice of ℓ is equivalent to the process $\{X_t : t = 1, 2, ...\}$ being stationary.[2] Thus, the restriction that $\mathbb{S}$ is measure preserving relative to Pr implies that the stochastic process $\{X_t : t = 1, 2, ...\}$ is stationary.

For a given $\mathbb{S}$, we now illustrate how to construct a probability measure Pr that makes $\mathbb{S}$ measure preserving and thereby induces stationarity. In example 2.3.3, there is only one Pr that makes $\mathbb{S}$ measure preserving, while in example 2.3.4 there are many.

Example 2.3.3. *Suppose that Ω contains two points, $\Omega = \{\omega_1, \omega_2\}$. Consider a transformation $\mathbb{S}$ that maps ω_1 into ω_2 and ω_2 into ω_1: $\mathbb{S}(\omega_1) = \omega_2$ and $\mathbb{S}(\omega_2) = \omega_1$. Since $\mathbb{S}^{-1}(\{\omega_2\}) = \{\omega_1\}$ and $\mathbb{S}^{-1}(\{\omega_1\}) = \{\omega_2\}$, for $\mathbb{S}$ to be measure preserving, we must have $Pr(\{\omega_1\}) = Pr(\{\omega_2\}) = 1/2$.*

Example 2.3.4. *Suppose that Ω contains two points, $\Omega = \{\omega_1, \omega_2\}$ and that $\mathbb{S}(\omega_1) = \omega_1$ and $\mathbb{S}(\omega_2) = \omega_2$. Since $\mathbb{S}^{-1}(\{\omega_2\}) = \{\omega_2\}$ and $\mathbb{S}^{-1}(\{\omega_1\}) = \{\omega_1\}$, $\mathbb{S}$ is measure preserving for any Pr that satisfies $Pr(\{\omega_1\}) \geq 0$ and $Pr(\{\omega_2\}) = 1 - Pr(\{\omega_1\})$.*

The next example illustrates how to represent an iid sequence of zeros and ones.

[2]Sometimes this property is called 'strict stationarity' to distinguish it from weaker notions that require only that some moments of joint distributions be independent of time. What is variously called wide-sense or second-order or covariance stationarity requires only that first and second moments of joint distributions are independent of calendar time.

Example 2.3.5. *Suppose that $\Omega = [0, 1)$ and that Pr is the uniform measure on this interval. Let*

$$\mathbb{S}(\omega) = \begin{cases} 2\omega & if\ \omega \in [0, 1/2) \\ 2\omega - 1 & if\ \omega \in [1/2, 1), \end{cases}$$
$$X(\omega) = \begin{cases} 1 & if\ \omega \in [0, 1/2) \\ 0 & if\ \omega \in [1/2, 1). \end{cases}$$

By a straightforward calculation, $Pr\{X_1 = 1|X_0 = 1\} = Pr\{X_1 = 1|X_0 = 0\} = Pr\{X_1 = 1\} = 1/2$. An analogous conclusion prevails for the event $\{X_1 = 0\}$, so X_1 is statistically independent of X_0. By extending this argument, it can be shown that $\{X_t : t = 0, 1, ...\}$ is a sequence of independent random variables.[3]

2.4 Invariant Events and Conditional Expectations

In this section, we present a Law of Large Numbers that asserts that when $\mathbb{S}$ is measure-preserving relative to Pr, time series averages converge.

Invariant events

We use the concept of an invariant event to understand how limit points of time series averages relate to a conditional mathematical expectation.

Definition 2.4.1. *An event Λ is **invariant** if $\Lambda = \mathbb{S}^{-1}(\Lambda)$.*

Figure 2.3 illustrates two invariant events on a space Ω. Notice that if Λ is an invariant event and $\omega \in \Lambda$, then $\mathbb{S}^t(\omega) \in \Lambda$ for $t = 0, 1, ..., \infty$.

Let $\mathfrak{I}$ denote the collection of invariant events. The entire space Ω and the null set $\varnothing$ are both invariant events. Like $\mathfrak{F}$, the collection of invariant events $\mathfrak{I}$ is a sigma algebra.

Conditional expectation

We want to construct a random vector $E(X|\mathfrak{I})$ called the "mathematical expectation of X conditional on the collection $\mathfrak{I}$ of invariant events".

[3] This example is from Breiman (1968, p. 108).

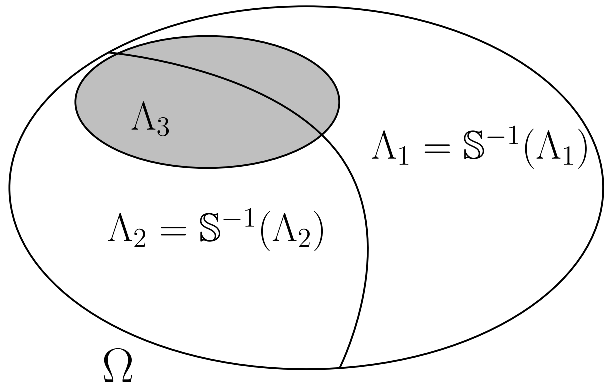

Figure 2.3: Two invariant events Λ_1 and Λ_2 and an event Λ_3 that is not invariant.

We begin with a situation in which a conditional expectation is a discrete random vector, an implication of an assumption that invariant events are unions of sets belonging to a countable partition Λ_j (along with the empty set). Later we'll extend the definition beyond this special setting.

A countable partition consists of a countable collection of nonempty events Λ_j such that $\Lambda_j \cap \Lambda_k = \varnothing$ for $j \neq k$ and such that the union of all Λ_j is Ω. Assume that each set Λ_j in the partition is itself an invariant event. Define the mathematical expectation conditioned on event Λ_j as

$$\frac{\int_{\Lambda_j} X dPr}{Pr(\Lambda_j)}$$

when $\omega \in \Lambda_j$. To extend the definition of conditional expectation to all of $\mathfrak{I}$, take

$$E(X|\mathfrak{I})(\omega) = \frac{\int_{\Lambda_j} X dPr}{Pr(\Lambda_j)} \quad \text{if} \quad \omega \in \Lambda_j.$$

Thus, the conditional expectation $E(X|\mathfrak{I})$ is constant for $\omega \in \Lambda_j$ but varies across Λ_j's. Figure 2.4 illustrates this characterization for a finite partition.

Conditional Expectation as Least Squares

When a random vector X has finite second moments, a conditional expectation is a least squares projection. Let Z be an n-dimensional measurement

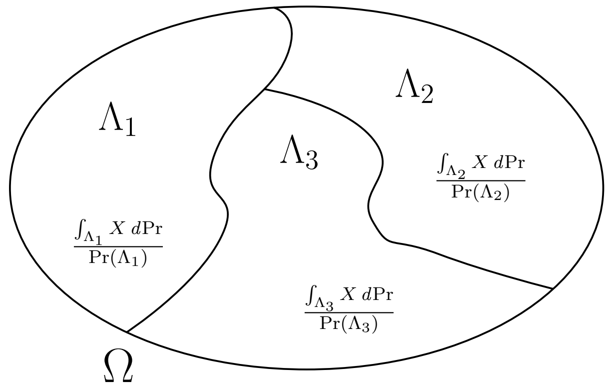

Figure 2.4: A conditional expectation $E(X|\mathfrak{I})$ is constant for $\omega \in \Lambda_j = \mathbb{S}^{-1}(\Lambda_j)$

function that is time-invariant and so satisfies

$$W_t(\omega) = Z[\mathbb{S}^t(\omega)] = Z(\omega).$$

In the special case in which the invariant events can be constructed from a finite partition, Z can vary across sets Λ_j but must remain constant within Λ_j.[4] Let $\mathcal{Z}$ denote the collection of all such time-invariant random vectors, and consider the following least squares problem:

$$\min_{Z \in \mathcal{Z}} E\left[|X - Z|^2\right], \tag{2.2}$$

where we assume that $E|X|^2 < \infty$. Denote the minimizer in the least squares problem 2.2 by $\tilde{X} = E(X|\mathfrak{I})$. An implication of the necessary conditions for the least squares minimizer $\widetilde{X} \in \mathcal{Z}$ is

$$E\left[\left(X - \widetilde{X}\right) Z'\right] = 0$$

for Z in $\mathcal{Z}$, which states that each entry of the vector $X - \widetilde{X}$ of regression errors is orthogonal to any vector Z in $\mathcal{Z}$.

A measure-theoretic approach constructs a conditional expectation by extending the orthogonality property of least squares. Provided that $E|X| <$

[4]More generally, Z must be measurable with respect to $\mathfrak{I}$.

∞, $E(X|\mathfrak{I})$ is the essentially unique random vector that, for any invariant event Λ, satisfies

$$E\left([X - E(X|\mathfrak{I})]\mathbf{1}_\Lambda\right) = 0,$$

where $\mathbf{1}_\Lambda$ is the indicator function that is equal to one on the set Λ and zero otherwise.

2.5 Law of Large Numbers

An elementary Law of Large Numbers asserts that the limit of an average over time of a sequence of independent and identically distributed random vectors equals the unconditional expectation of the random vector. We want a more general Law of Large Numbers that applies to averages over time of sequences of observations that are intertemporally dependent. The following theorem asserts two senses in which, for measure-preserving $\mathbb{S}, Pr$, averages of intertemporally dependent processes converge to mathematical expectations conditioned on invariant events.

Theorem 2.5.1. *(Birkhoff) Let $(\Omega, \mathfrak{F}, Pr)$ be a probability space. Suppose that $\mathbb{S}$ is measure preserving.*

i) For any X such that $E|X| < \infty$,

$$\frac{1}{N}\sum_{t=1}^{N} X_t(\omega) \to E(X|\mathfrak{I})(\omega)$$

with probability one;

ii) For any $X(\omega)$ such that $E|X(\omega)|^2 < \infty$,

$$E\left[\left|\frac{1}{N}\sum_{t=1}^{N} X_t - E(X|\mathfrak{I})\right|^2\right] \to 0.$$

Part *i*) asserts *almost-sure* convergence and part *ii*) asserts *mean-square* convergence.

Definition 2.5.2. *A measure-preserving transformation $\mathbb{S}$ is said to be **ergodic** under a probability measure Pr if all invariant events have probability zero or one.*

Thus, when a transformation $\mathbb{S}$ is *ergodic* under measure Pr, the invariant events have either the same probability measure as the entire sample space Ω (whose probability measure is one), or the same probability measure as the empty set $\varnothing$ (whose probability measure is zero).

Proposition 2.5.3. *Suppose that* $\mathbb{S}$ *is ergodic under measure* Pr. *Then* $E(X|\mathfrak{I}) = E(X)$.

Theorem 2.5.1 covers convergence in general. Proposition 2.5.3 focuses on the case in which probabilities assigned to invariant events are degenerate in the sense that all invariant events have the same probability as either Ω (probability one) or the null set (probability zero). When $\mathbb{S}$ is *ergodic* under measure Pr, limit points of time series averages equal corresponding unconditional expectations, an outcome we call a *standard* Law of Large Numbers. Without ergodicity, limit points of time series averages equal expectations conditioned on invariant events.

The following examples remind us how ergodicity is a restriction on $\mathbb{S}$ and Pr.

Example 2.5.4. *Consider example 2.3.3 again. Suppose that the measurement vector is*

$$X(\omega) = \begin{cases} 1 & \text{if } \omega = \omega_1 \\ 0 & \text{if } \omega = \omega_2. \end{cases}$$

Then it follows directly from the specification of $\mathbb{S}$ *that*

$$\frac{1}{N}\sum_{t=1}^{N} X_t(\omega) \to \frac{1}{2}$$

for both values of ω. *The limit point is the average across sample points.*

Example 2.5.5. *Returning to example 2.3.4,* $X_t(\omega) = X(\omega)$ *so that the sequence is time invariant and equal to its time-series average. The time-series average of* $X_t(\omega)$ *equals the average across sample points only when* Pr *assigns probability* 1 *to either* ω_1 *or* ω_2.

2.6 Limiting Empirical Measures

Given a triple $(\Omega, \mathfrak{F}, Pr)$ and a measure-preserving transformation $\mathbb{S}$, we can use Theorem 2.5.1 to construct *limiting empirical measures* on $\mathfrak{F}$. To start, we will analyze a setting with a countable partition of Ω consisting of invariant events $\{\Lambda_j : j = 1, 2, ...\}$, each of which has strictly positive probability under Pr. We consider a more general setting later. Define $\mathbf{1}_\Lambda(\omega)$ as the indicator function of the set Λ:

$$\mathbf{1}_\Lambda(\omega) = \begin{cases} 1 & \text{if} \quad \omega \in \Lambda \\ 0 & \text{if} \quad \omega \notin \Lambda. \end{cases}$$

Given an event Λ in $\mathfrak{F}$ and for almost all $\omega \in \Lambda_j$, define the limiting empirical measure Qr_j as

$$Qr_j(\Lambda)(\omega) = \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \mathbf{1}_\Lambda \left[\mathbb{S}^t(\omega)\right] = \frac{Pr(\Lambda \cap \Lambda_j)}{Pr(\Lambda_j)}. \tag{2.3}$$

Thus, when $\omega \in \Lambda_j$, $Qr_j(\Lambda)$ is the fraction of time $\mathbb{S}^t(\omega) \in \Lambda$ in very long samples. Holding Λ_j fixed and letting Λ be an arbitrary event in $\mathfrak{F}$, we can treat Qr_j as a probability measure on $(\Omega, \mathfrak{F})$. By doing this for each $\Lambda_j, j = 1, 2, \ldots$, we can construct a countable set of probability measures $\{Qr_j\}_{j=1}^{\infty}$ that forms the set of all measures that can possibly be approximated by applying the Law of Large Numbers. If nature draws an $\omega \in \Lambda_j$, then measure Qr_j describes outcomes.

Statistical Models

We now reverse the preceding construction that started with a probability measure Pr and then constructed the set of possible limiting empirical measures Qr_j's. Think of starting with probability measures Qr_j and then constructing possible measures Pr that are consistent with them. We do this for the following reason. Each Qr_j defined by (2.3) uses the Law of Large Numbers to assign probabilities to events $\Lambda \in \mathfrak{F}$. Qr_j's are thus the measures that very long time series disclose through the Law of Large Numbers. However, because

$$Qr_j(\Lambda) = Pr(\Lambda \mid \Lambda_j) = \frac{Pr(\Lambda \cap \Lambda_j)}{Pr(\Lambda_j)} \text{ for } j = 1, 2, \ldots, ,$$

such Qr_j's are only conditional probabilities and are silent about the probabilities $Pr(\Lambda_j)$ of the underlying invariant events Λ_j. There are typically multiple choices of probabilities Pr that imply the same probabilities conditioned on invariant events.

Because Qr_j's are all that can ever be learned from "letting the data speak", we regard each probability measure Qr_j as a statistical model.

Definition 2.6.1. *A* statistical model *is a probability measure that a Law of Large Numbers can eventually disclose.*

Therefore, for each invariant set Λ_j, the probability measure Qr_j describes a statistical model.

Remark 2.6.2. *For each j, $\mathbb{S}$ is measure-preserving and ergodic on $(\Omega, \mathfrak{F}, Qr_j)$. In the second equality of definition* (2.3)*, we assured ergodicity by assigning probability one to the event Λ_j.*

Relation (2.3) implies that probability Pr connects to probabilities Qr_j by

$$Pr(\Lambda) = \sum_j Qr_j(\Lambda) Pr(\Lambda_j). \tag{2.4}$$

Following as it does from definitions of the elementary objects comprising a stochastic process, decomposition (2.4) is "just mathematics". But it acquires special interest for us because it shows how to construct alternative probability measures Pr for which $\mathbb{S}$ is measure preserving. Because long data series disclose probabilities conditioned on invariant events to be Qr_j, we must hold the Qr_j's fixed, but we can freely assign probabilities Pr to the underlying invariant events Λ_j and in this way create a family of probability measures for which $\mathbb{S}$ is measure preserving.

Decomposition (2.4) is of interest to both frequentist and Bayesian statisticians.

- Decomposition (2.4) tells frequentist statisticians that when $\omega \in \Lambda_j$ they can learn a particular Qr_j, but not Pr.

- Decomposition (2.4) tells Bayesian statisticians how to represent the probability measure Pr as a mixture of statistical models with weights $Pr(\Lambda_j)$ interpretable as prior probabilities over statistical models.[5]

We extend these ideas in the following section.

[5]The assertions in these two bullet points summarize a conversation between frequentist and Bayesian statisticians imagined by Kreps (1988, Ch. 11).

2.7 Ergodic Decomposition

Earlier we started with a triple $(\Omega, \mathfrak{F}, Pr)$ and then studied properties of alternative measure-preserving transformations $\mathbb{S}$. To obtain a generalization of decomposition (2.4), it is fruitful to explore consequences of a different starting point using a mathematical structure of Dynkin (1978). The idea is to begin with a transformation $\mathbb{S}$ and then to study a *set* of probability measures Pr for which $\mathbb{S}$ is both measure-preserving and ergodic.

We start with a pair $(\Omega, \mathfrak{F})$. We assume that there is a metric on Ω and that Ω is separable. We also assume that $\mathfrak{F}$ is the collection of Borel sets (the smallest sigma algebra containing the open sets). Given $(\Omega, \mathfrak{F})$, take a (measurable) transformation $\mathbb{S}$ and consider the set $\mathcal{P}$ of probability measures Pr for which $\mathbb{S}$ is measure-preserving. For some of these probability measures, $\mathbb{S}$ is ergodic, but for others it is not. Let $\mathcal{Q}$ denote the set of probability measures for which $\mathbb{S}$ is ergodic. Under a nondegenerate convex combination of two probability measures in $\mathcal{Q}$, $\mathbb{S}$ is measure-preserving but *not* ergodic. Dynkin (1978) constructs limiting empirical measures Qr on $\mathcal{Q}$ and justifies the following representation of the set $\mathcal{P}$ of probability measures Pr.

Proposition 2.7.1. *For each probability measure $\widetilde{Pr}$ in $\mathcal{P}$ there is a unique probability measure π over $\mathcal{Q}$ such that*

$$\widetilde{Pr}(\Lambda) = \int_{\mathcal{Q}} Qr(\Lambda)\pi(dQr) \tag{2.5}$$

for all $\Lambda \in \mathfrak{F}$.[6]

This proposition generalizes representation (2.4). It asserts a sense in which the set $\mathcal{P}$ of probabilities for which $\mathbb{S}$ is measure-preserving is convex. The extremal points of this set are in the smaller set $\mathcal{Q}$ of probability measures for which the transformation $\mathbb{S}$ is ergodic. Representation (2.5) shows that by forming "mixtures" (i.e., weighted averages or convex combinations) of probability measures under which $\mathbb{S}$ is ergodic, we can represent all probability specifications for which $\mathbb{S}$ is measure-preserving.

[6] Krylov and Bogolioubov (1937) give an early statement of this result. Dynkin (1978) provides a more general formulation that nests this and other closely related results. His analysis includes a formalization of integration over the probability measures in $\mathcal{Q}$. Dynkin (1978) uses the resulting representation to draw connections between collections of invariant events and sets of sufficient statistics.

Practical model building

Like its specialized counterpart (2.4), representation (2.5) offers insights about building and interpreting probability models. As in section 2.6, we can appeal to the Law of Large Numbers to interpret the probabilities in $\mathcal{Q}$ as *statistical models* that are components of a more complete probability specification.[7]

A collection of invariant events $\mathfrak{I}$ is associated with $\mathbb{S}$. Formally, Dynkin (1978) shows that there exists a common conditional expectation operator $\mathbb{J} \equiv E(\cdot|\mathfrak{I})$ that assigns mathematical expectations to bounded measurable functions (mapping Ω into $\mathbb{R}$) conditioned on the set of invariant events $\mathfrak{I}$. The conditional expectation operator $\mathbb{J}$ characterizes limit points of time series averages of $\{\phi(X_t)\}$ for bounded functions ϕ for any $Pr \in \mathcal{P}$. Alternative probability measures Pr assign different probabilities to the invariant events.

Exchangeability

To explore connections between Proposition 2.7.1 and ideas of de Finetti and Hewitt and Savage, we begin with exchangeability property that a joint probability distribution is invariant to rearrangements of the order of observations attained via a permutation of time indexes. de Finetti (1937) and Hewitt and Savage (1955) showed that exchangeability is a natural concept for elementary problems of statistical learning. Proposition 2.7.1 extends exchangeability in ways that are especially useful for time series applications.

Definition 2.7.2. *A permutation of the time index is a one-to-one mapping of the set of nonnegative integers into itself.*

Definition 2.7.3. *A sequence of random vectors* $\{X_t : t = 0, 1, ...\}$ *is said to be exchangeable if joint probability distributions induced by this se-*

[7] Marschak (1953), Hurwicz (1962), Lucas (1976), and Sargent (1981) distinquished between structural econometric models and what we call statistical models. Structural econometric models are designed to forecast outcomes of hypothetical experiments that freeze some components of an economic environment while changing others. This requires more than representation 2.6 of a *given* stochastic process. A structural model invites experiments that *alter* statistical models.

quence equal those for $\{X_{\zeta(0)}, X_{\zeta(1)}, ...\}$ *for any permutation* $\zeta(0), \zeta(1), \ldots$ *that keeps all but a finite number of indices fixed.*[8]

A process is said to be exchangeable if joint distributions of members of the sequence $\{X_t : t = 0, 1, ...\}$ do not depend on their order of appearance.

In section 2.7 we represented a probability measure $\mathcal{P}$ as the (convex) set of all probabilities under which $\mathbb{S}$ is measure-preserving. The set $\mathcal{P}_{ex}$ of all Pr for which $\{X_t : t = 0, 1...\}$ is exchangeable is a (convex) subset of $\mathcal{P}$.[9] More can be said. Let $\mathcal{Q}_{in}$ denote the set of probabilities for which the process $\{X_t : t = 1, 2, ...\}$ is independent and identically distributed. Then $\mathcal{Q}_{in} \subset \mathcal{P}_{ex}$. It happens that $\mathcal{Q} \cap \mathcal{P}_{ex} = \mathcal{Q}_{in}$.

Proposition 2.7.1 represents probabilities in the set $\mathcal{P}$ of probability measures Pr for which $\mathbb{S}$ is measure-preserving in terms of ones in in the restricted set $\mathcal{Q}$ of probability measures under which $\mathbb{S}$ is ergodic. Similarly, a probability in $\mathcal{P}_{ex}$ can be expressed as a weighted average of probabilities in $\mathcal{Q}_{in}$. de Finetti (1937) obtained this result for a setting with binary (zero or one) random variables. Hewitt and Savage (1955) proved the following generalization of de Finetti's result:

Result 2.7.4. *For each* Pr *in* $\mathcal{P}_{ex}$ *there is a unique probability measure* π *over* $\mathcal{Q}_{in}$ *such that*

$$Pr(\Lambda) = \int_{\mathcal{Q}_{in}} Qr(\Lambda)\pi(dQr) \tag{2.6}$$

for all $\Lambda \in \mathfrak{F}$.

Thus, when we restrict the set of probabilities to be exchangeable, extremal points of the set become probabilities under which $\{X_t : t = 1, 2, ...\}$ is independent and identically distributed.

---

[8] An alternative but equivalent characterization of exchangeability due to Ryll-Nardzewski (1957) views exchangeable probabilities over $\{X_t : t = 0, 1, ...\}$ as ones that imply that the transformation $\mathbb{S}$ is measure-preserving under an associated Pr with additional restrictions. Specifically, Ryll-Nardzewski shows that to establish exchangeability, it suffices to verify probabilistic invariance for specifications of ζ's that map the nonnegative integers into the nonnegative integers and that satisfy $\zeta(0) < \zeta(1) < \zeta(2) < \cdots$. To establish stationarity, it suffices to verify this invariance under a more restrictive specification $\zeta(j) = j + \tau$ for $j = 0, 1, 2, ...$ for any $\tau \geq 1$.

[9] Given $(\Omega, \mathfrak{F})$ and the random vector X, the set $\mathcal{P}_{ex}$ could be empty. For instance, provided that X differs across the two values of ω in example 2.3.3, there is no specification of Pr for which $\{X_t : t = 0, 1, ...\}$ is exchangeable. In contrast, all admissible specifications of Pr in example 2.3.4 imply that $\{X_t : t = 0, 1, ...\}$ is exchangeable.

Foundations of statistics

The Law of Large Numbers stated in Theorem 2.5.1 and the Hewitt and Savage Result 2.7.4 provide foundations for both Bayesian and frequentist statistics.[10] Because we are interested in temporally dependent processes, we prefer to rely on foundations cast in a more general setting than allowed by exchangeability. Consequently, we feature stationarity instead of exchangeability, so representation (2.5) in proposition 2.7.1 provides the foundations that interest us.

Risk and uncertainty

We have adopted a frequentist notion of a statistical model based on a law of large numbers. An applied researcher typically does not know the statistical model and so does not know limits of time-series averages.[11] Instead, he or she faces uncertainty about which statistical model generated the data. This situation leads us to specifications of $\mathbb{S}$ that are consistent with a family $\mathcal{P}$ of probability models under which $\mathbb{S}$ is measure preserving and a stochastic process is stationary. Representation (2.5) describes uncertainty about statistical models with a probability distribution π over the set of models $\mathcal{Q}$.

For a Bayesian, π is a subjective prior probability distribution that pins down a convex combination of "statistical models."[12] A Bayesian expresses his trust in that convex combination of statistical models when he constructs a complete probability measure over outcomes and uses it to compute expected utility.[13] A Bayesian decision theory axiomatized by Savage makes no distinction between how decision makers respond to the probabilities embedded in the individual component statistical models and the π probabilities that he uses to mix them. All that matters to a Bayesian decision maker is his complete probability distribution over outcomes, not how it is attained as a π-mixture of component statistical models.

[10]Kreps (1988, Ch. 11) describes how de Finetti's theorem underlies statistical theories of learning about an unknown parameter.

[11]This statement is half of the point of Kreps (1988, Ch. 11).

[12]This subsection is motivated in part by the intriguing discussions of von Plato (1982) and Cerreia-Vioglio et al. (2013).

[13]Here 'complete' can be taken to be synonymous with 'not conditioning on invariant events'.

Some decision and control theorists challenge the complete confidence in a single prior probability assumed in a Bayesian approach.[14] As an alternative, they imagine decision makers who want to evaluate decisions under alternative π's as a way to distinguish 'uncertainty', meaning ignorance of π, from 'risk', meaning prospective outcomes with probabilities reliably described by a given statistical model. This connects to ideas of Knight (1921) if we regard Knight as using "risk" to refer to ignorance that is completely described by probabilities affiliated with a particular statistical model. A Law of Large Numbers eventually discloses probabilities associated with a particular statistical model. The π probabilities that a Bayesian uses to mix statistical models are not revealed by a Law of Large Numbers. They are 'in the head of the decision maker.' Bayes' rule allows the decision maker to update the π probabilities as data accrue. As data accrue without limit, conditional probabilities over statistical models eventually concentrate on the statistical model that generates the data.

Thus, while a Bayesian statistician knows a prior distribution $\{Pr(\Lambda_j) : j = 1, 2, ...\}$ in representation (2.4), a statistician who doesn't know $\{Pr(\Lambda_j) : j = 1, 2, ...\}$ might want to explore the sensitivity of inferences to alternative prior probability distributions. A robust Bayesian approach studies the consequences of changing the $Pr(\Lambda_j)$'s without altering the baseline statistical models Qr_j. Relatedly, Segal (1990) and Klibanoff et al. (2005) model decision makers who respond differently to ignorance summarized by π than they do to the Knightian risk described by a particular statistical model. Their decision makers refuse to reduce the compound lottery induced by π as is done in decomposition (2.6). Later, we will suggest ways to alter how models are weighted based on decomposition (2.4) along lines sketched in Problem 1.5.1 in Chapter 1. In such cases doubts about the model specification typically will be resolved by the Law of Large Numbers, in effect revealing probabilities conditioned on invariant events. In some the examples that interest us, it will take a long time to reveal these conditional probabilities. We will also consider other formulations that will prevent the Law of Large Numbers from revealing the a unique statistical model governing the data generation. Model ambiguity will not resolved, even asymptotically, in these examples.

---

[14] For example, see Hansen and Sargent (2008).

Relationship to Misspecification Analysis

Sims (1971a,b, 1972, 1974, 1993), White (1994), and Hansen and Sargent (1993) studied limiting behavior of estimators of incorrect probability specifications. To do that, they extended the settings of section 2.7 to capture the idea that the econometrician's ignorance goes beyond 'not knowing a prior'.

To extend our formalism to include a misspecification analysis, let an econometrician's model of a stochastic process for $\{X_t : t = 0, 1, ...\}$ be specified in terms of a sample space and sigma algebra $(\Omega, \mathfrak{F})$, a transformation $\mathbb{S}$, and a set $\mathcal{Q}_m \subset \mathcal{Q}$ of probability measures Pr for which $\mathbb{S}$ is measure-preserving and ergodic. The statistical models considered by the econometrician are members of $\mathcal{Q}_m$. Often these can be represented in terms of alternative possible values of an unknown parameter vector. A Bayesian econometrician constructs a probability measure Pr_m in $\mathcal{P}$ that is a mixture of the probability measures in $\mathcal{Q}_m$.

Unbeknownst to the econometrician, the data are actually generated by a probability model Pr_c for which $\mathbb{S}$ is measure-preserving and ergodic but that is *not* in the econometrician's set of possible models $\mathcal{Q}_m$. The econometrician's specification and the true data generating model share components $(\Omega, \mathfrak{F})$ of the probability space along with the transformation $\mathbb{S}$ and measurement function X. Because the actual data generating mechanism is $Pr_c \in \mathcal{Q}$, it has an associated law of large numbers. But this true statistical model is excluded from the econometrician's set of statistical models $\mathcal{Q}_m$. Papers in the misspecification literature condition on the true statistical model Pr_c and then study the consequences of applying statistical procedures like maximum likelihood to the econometrician's probability model or family of probability models. For example, some of these papers work in a frequentist tradition and show how probability limits of maximum likelihood estimators of parameters for the econometrician's model depend on features of the true model Pr_c missed by the econometrician's probability model.

Papers in this literature investigate consequences of i) assuming an admittedly false too tightly parameterized family of models,[15] and ii) filtering data in ways that eradicate seasonal or low frequency components of time series data to be used to estimate parameters of an economic model that is

[15]See Sims (1971a,b, 1972) and White (1994).

not well designed to confront these components.[16]

2.8 Calibrators and econometricians

Decomposition (2.5) indicates how laws of large numbers condition on invariant events. Unknown parameters that remain constant through time are key invariant events for macroeconomists. Decomposition (2.5) in proposition 2.7.1 sheds light on distinct practices of macroeconomic calibrators and econometricians.

Calibrators

Calibrators appeal to laws of large numbers to justify their ways of comparing theories to observations. Calibrating a statistical model with free parameters $\Theta = [\theta_1, \ldots, \theta_n]$ usually consists of these steps:

- Partition parameters Θ into two sets Θ_1 and Θ_2.

- Import parameters in subset Θ_1 from earlier studies, including ones that may have used *different* statistical models, and treat them as known.[17]

- Given the parameters in Θ_1, compute theoretical values of the Qr_j's implied by an economic model as functions of parameters in subset Θ_2. Here is one place that calibrators apply laws of large numbers.

- Set parameters in subset Θ_2 to make a list of population moments of the pertinent conditional distribution Qr_m equal observed (finite) sample moments.

- Use the Qr_m that emerges from the previous steps to compute population moments of some random variables not used to calibrate the model. (Here is another step that uses a law of large numbers.) Compare these population moments with observed finite sample moments as a basis either for validating the model or for uncovering "puzzles"

[16] See Sims (1974, 1993), Wallis (1974), and Hansen and Sargent (1993).

[17] Kuh and Meyer (1957) and Browning et al. (1999) criticize importing parameter estimates inferred from one econometric specification into another.

(i.e., failures of the statistical model to fit some dimensions of the data.)

A macroeconomic calibrator expresses no uncertainty about parameter values. For him, parameter uncertainty is a distraction that he sets aside in order to answer what he regards as more important questions. He pretends to know parameters (i.e., he conditions on invariant events) and then makes statements about economic outcomes implied by a given statistical model. He uses these outcomes to determine dimensions along which he judges his statistical model to be 'successful' and those along which he judges it to be 'deficient'. In this way he divides outcomes into scientific 'understandings' and 'puzzles'.[18]

Frequentists and Bayesians

While calibrators treat calibrated probabilities Qr_m as known, econometricians confront the fact that finite data don't contain enough information reliably to pin down invariant events. All econometricians seek to quantify uncertainty about parameters and the statistical models associated with them, something that is not part of calibrators' practice. But frequentist and Bayesian econometricians quantify uncertainty in different ways. Unwilling or unable to assign a prior distribution, a frequentist econometrician investigates consequences of conditioning on alternative statistical models in order to make inferences about *sample statistics*. A Bayesian econometrician asserts a more complete probabilistic description of his uncertainty than does a frequentist. He accomplishes this by representing his initial uncertainty over statistical models with a prior probability distribution and then revising those probabilities by conditioning on available data.[19]

Calibrators and Frequentists

Despite differences in their purposes and practices, macroeconomic calibrators and frequentist econometricians both condition on statistical models. As we have mentioned, calibrators appeal to a Law of Large Numbers associated with a particular statistical model to justify how they set parame-

[18] See Mehra and Prescott (1985) and Lucas (2003) for prominent examples.

[19] Kydland and Prescott (1996), Hansen and Heckman (1996), and Sims (1996) discuss calibration.

ters. Frequentists use the Law of Large Numbers to understand the limiting behaviors of sample statistics under alternative models. Thus, after conditioning on invariant events, frequentist econometricians and macroeconomic calibrators proceed to do very different things. A frequentist econometrician conditions on a statistical model, then approximates the sample distributions of particular *statistics* by using both a Law of Large Numbers and a Central Limit Theorem.[20] The frequentist uses those sampling distributions to make probability statements conditioned on alternative models. That is the frequentist econometrician's way of describing his uncertainty about parameters and about the plausibility of alternative statistical models after observing a data sample.

2.9 Constructing Stochastic Processes: II

In section 2.2, we studied stochastic processes by using a transformation $\mathbb{S}$ and a measure Pr on a canonical probability space $(\Omega, \mathfrak{F})$. From there we inferred a collection of induced distributions. In practice, it can be more convenient to go the other direction by beginning with a set of induced conditional distributions or joint distributions associated with collections of calendar dates and to restrict the distributions to be temporally consistent. From these, we can work backwards to construct a canonical representation of a stochastic process in terms of a triple $(\Omega, \mathfrak{F}, Pr)$ and the transformation $\mathbb{S}$ that we used to illustrate things in section 2.1. There we let the set of sample points Ω be a collection of infinite sequences of elements of $\mathbb{R}^n$ with typical element $\omega = (\mathbf{r}_0, \mathbf{r}_1, ...)$, $\mathbb{S}(\omega) = (\mathbf{r}_1, \mathbf{r}_2, ...)$, and $X(\omega) = \mathbf{r}_0$. When we take the measurement to be $X_t(\omega) = \mathbf{r}_t$, the transformation $\mathbb{S}$ can appropriately be called a shift transformation. To assign probabilities, fix an integer $\ell > 0$. Let Pr_ℓ assign probabilities to the Borel sets of $\mathbb{R}^{n(\ell+1)}$. Pr_ℓ assigns probabilities to two mathematical objects. The first object is the composite random vector:

$$X^{[\ell]}(\omega) \doteq \begin{bmatrix} X_0(\omega) \\ X_1(\omega) \\ \vdots \\ X_\ell(\omega) \end{bmatrix}.$$

[20] Calibrators rarely use or cite central limit theorems; these contain information about rates at which approximations based on laws of large numbers become reliable.

The second object consists of a set of events in $\mathfrak{F}$ that restrict the first $\ell+1$ entries of $\omega = (\mathbf{r}_0, \mathbf{r}_1, ...)$, namely, events of a form given in (2.1). To construct Pr and a stochastic process $\{X_t : t = 0, 1, ...\}$, we must specify Pr_ℓ for all nonnegative integers ℓ and make sure that $Pr_{\ell+1}$ is consistent with Pr_ℓ in the sense that both assign the same joint distribution to $X^{[\ell]}$. An alternative way to express consistency is to require that both Pr_ℓ and $Pr_{\ell+1}$ assign identical probabilities to events of the form Λ given in (2.1). If consistency in this sense prevails, then the Kolmogorov Extension Theorem guarantees that there exists a probability Pr on the space $(\Omega, \mathfrak{F})$ that is consistent with the probability assignments implied by Pr_ℓ for all nonnegative integers ℓ. In applications, we seek convenient ways to construct probabilities Pr on Ω, while requiring that Pr_ℓ for $\ell = 0, 1, ...$ are consistent so that we can apply mathematical results from sections 2.2-2.7.

When $\mathbb{S}$ is measure-preserving, the probability distribution of the random vector $[X_t{}', X_{t+1}{}', ..., X_{t+\ell}{}']'$ does not depend on t for any nonnegative integer ℓ, so the process is stationary. Under the section 2.2 method of constructing a stochastic process, the result has a converse. Suppose Pr is specified so that $\{X_t : t = 1, 2, ...\}$ is stationary. Then the shift transformation $\mathbb{S}$ in the canonical construction is measure-preserving.

Remark 2.9.1. *We could have used the induced distributions for* $\{X_t : t = 0, 1, \ldots\}$ *to depict a probabilistically equivalent process in terms of a measure-preserving probability on the canonical probability space described at the outset of this chapter. But for central limit approximations and other applications too, it can be more convenient to invent an infinite past.*

2.10 Inventing an infinite past

When Pr is measure preserving and the process $\{X_t : t = 0, 1, ...\}$ is stationary, it can be useful to invent an infinite past. To accomplish this, we reason in terms of the (measurable) transformation $\mathbb{S} : \Omega \to \Omega$ that describes the evolution of a sample point ω. Until now we have presumed that $\mathbb{S}$ has the property that for any event $\Lambda \in \mathfrak{F}$,

$$\mathbb{S}^{-1}(\Lambda) = \{\omega \in \Omega : \mathbb{S}(\omega) \in \Lambda\}$$

is an event in $\mathfrak{F}$. In this section and throughout chapter 5 too we also assume that $\mathbb{S}$ is one-to-one and has the property that for any event $\Lambda \in \mathfrak{F}$,

$$\mathbb{S}(\Lambda) = \{\omega \in \Omega : \mathbb{S}(\omega) \in \Lambda\} \in \mathfrak{F}. \tag{2.7}$$

Because

$$X_t(\omega) = X[\mathbb{S}^t(\omega)] = X_t = X \circ \mathbb{S}^t$$

is well defined for negative values of t, restrictions 2.7 allow us to construct a "two-sided" process that has both an infinite past and an infinite future. Let $\mathfrak{A}$ be a subsigma algebra of $\mathfrak{F}$, and let

$$\mathfrak{A}_t = \left\{\Lambda_t \in \mathfrak{F} : \Lambda_t = \{\omega \in \Omega : \mathbb{S}^t(\omega) \in \Lambda\} \text{ for some } \Lambda \in \mathfrak{F}\right\}.$$

We assume that $\{\mathfrak{A}_t : -\infty < t < +\infty\}$ is a nondecreasing *filtration.* If the original measurement function X is $\mathfrak{A}$-measurable, then X_t is $\mathfrak{A}_t$-measurable. Furthermore, X_{t-j} is in $\mathfrak{A}_t$ for all $j \geq 0$. The set $\mathfrak{A}_t$ depicts information available at date t, including past information. The invariant events in $\mathfrak{I}$ are contained in $\mathfrak{A}_t$ for all t.

We construct the following moving-average process in terms of an infinite history.

Example 2.10.1. *(Moving average) Suppose that* $\{W_t : -\infty < t < \infty\}$ *is a vector stationary process for which*

$$E\left(W_{t+1}|\mathfrak{A}_t\right) = 0$$

and that $E\left(W_t {W_t}'|\mathfrak{I}\right) = I$ *for all* $-\infty < t < +\infty$. *One possibility is that this sequence is i.i.d. Construct*

$$X_t = \sum_{j=0}^{\infty} \alpha_j \cdot W_{t-j} \tag{2.8}$$

where

$$\sum_{j=0}^{\infty} |\alpha_j|^2 < \infty. \tag{2.9}$$

Restriction (2.17) *implies that* X_t *is well defined as a mean square limit.* X_t *is constructed from the infinite past* $\{W_{t-j} : 0 \leq j < \infty\}$. *The process* $\{X_t : -\infty < t < \infty\}$ *is stationary and is often called an infinite-order moving average process. The sequence* $\{\alpha_j : j = 0, 1, ...\}$ *can depend on the invariant events.*

2.11 Processes with stationary increments

Macroeconomic time series such as investment, output, and consumption display geometric growth and are well approximated as processes whose logarithms have stationary increments. For a scalar measurement function X, consider a scalar process $Y = \{Y_t : t = 0, 1, ...\}$ with stationary increments $X = \{X_t : -\infty < t < +\infty\}$:

$$Y_{t+1} - Y_t = X_{t+1} \tag{2.10}$$

for $t = 0, 1, ...,$. Initialize the process at Y_0 where Y_0 is $\mathfrak{A}_0$ measurable. We adopt the assumptions of section 2.10 that rationalize an infinite past, and again let $\mathfrak{A}$ be a subsigma algebra of $\mathfrak{F}$, where

$$\mathfrak{A}_t \doteq \left\{\Lambda_t \in \mathfrak{F} : \Lambda_t = \{\omega \in \Omega : \mathbb{S}^t(\omega) \in \Lambda\} \text{ for some } \Lambda \in \mathfrak{F}\right\}.$$

A Martingale Decomposition

We show how to extract a linear time trend and a martingale from a stationary increment process. The martingale component dominates the long-term variance of the process. As we will see later in example economies, this component has a prominent impact on economic outcomes and valuation. The construction in this section is builds on an insight of Gordin (1969) in his develop of a Central Theorem for stationary processes. [21] The Central Limit Theorem is one of three applications that we will feature.

Let

$$\nu = E\left(X_t \mid \mathfrak{I}\right)$$

and

$$G_t = \sum_{j=0}^{\infty} E\left(X_{t+j} - \nu \mid \mathfrak{A}_t\right) \tag{2.11}$$

where we assume that X has a finite first moment and that the right-hand side of (2.11) converges in mean square. The term ν induces a time trend in the $\{Y_t\}$ process. The random variable G_t captures where you expect the Y process to be in the distant future after subtracting the deterministic time trend and Y_0.

[21] Also see Hall and Heyde (1980).

With the calculation in mind, Let $\mathcal{G}$ denote the set of all scalar random variables X such that $E(X^2) < \infty$ and

$$G_t = \sum_{j=0}^{\infty} E\left(X_{t+j} - \nu | \mathfrak{A}_t\right)$$

is well defined as a mean-square convergent series. The restriction that $X \in \mathcal{G}$ is not merely a technicality as it eliminates some long-memory stationary processes with enduring temporal dependencies. The restriction does allow for the inclusion of Markov specifications used as examples throughout this book.

We will use

Definition 2.11.1. *The process* $\{M_t : t = 0, 1, ...\}$ *is said to be a* **martingale** *relative to* $\{\mathfrak{A}_t : t = 1, 2, ...\}$ *if for* $t = 0, 1, ...$

- M_t *is* $\mathfrak{A}_t$ *measurable;*
- $E\left(M_{t+1} | \mathfrak{A}_t\right) = M_t$.

For $X \in \mathcal{G}$, let

$$F_{t+1} = G_{t+1} - \overline{G}_t$$

where

$$\overline{G}_t = E\left(G_{t+1} \mid \mathfrak{A}_t\right)$$

is the best forecast of G_{t+1} given current information and F_{t+1} is the forecast error or the "'surprise" component of the G_{t+1}.

Now for some book-keeping. Notice that by the Law of Large Numbers,

$$\begin{aligned} X_{t+1} - \nu &= G_{t+1} - E\left(G_{t+2} \mid \mathfrak{A}_{t+1}\right) \\ &= G_{t+1} - \overline{G}_{t+1} \\ &= F_{t+1} + \overline{G}_t - \overline{G}_{t+1} \end{aligned}$$

This leads us to write

$$Y_{t+1} - Y_t = \nu + F_{t+1} - \overline{G}_{t+1} + \overline{G}_t$$

Given an initial condition Y_0, solve this difference equation backwards to obtain:

$$Y_t - Y_0 = \nu t + \sum_{j=1}^{t} F_t - \overline{G}_t + \overline{G}_0, \tag{2.12}$$

where we have used the fact that summing a first difference of $\overline{G}_t$ results in the difference over horizon t.

In light of decomposition (2.12), we interpret νt as a time trend component of the stationary increment process Y. Subtracting the trend growth component, leaves us with the martingale component

$$Y_t^m = \sum_{j=1}^{t} F_t.$$

We have established:

Proposition 2.11.2. *If X is in $\mathcal{G}$, the process $\{Y_t : t = 0, 1, ...\}$ admits the additive decomposition*

$$Y_t = \underbrace{t\nu}_{\textbf{\textit{trend}}} + \underbrace{Y_t^m}_{\textbf{\textit{martingale}}} - \underbrace{\overline{G}_t}_{\textbf{\textit{stationary}}} + \underbrace{Y_0 + \overline{G}_0}_{\textbf{\textit{invariant}}}.$$

The martingale component Y^m has stationary increments, $Y_0^m = 0$, and the component $\overline{G}$ is stationary.

Permanent Shock

We can use the decomposition in Proposition 5.1.2 to identify a time trend, a "permanent shock", and a transitory component of a stationary-increments process like (5.1). The permanent shock is the increment to the martingale. This construction is present in univariate analysis of linear time series models by Beveridge and Nelson (1981) and multivariate analyses of Blanchard and Quah (1989) and Shapiro and Watson (1988).

Example 2.11.3. *(Moving-average increment process) Consider again example 2.10.1 of a moving-average process:*

$$X_t = \sum_{j=0}^{\infty} \alpha_j \cdot W_{t-j}. \tag{2.13}$$

Use this process as the increment for $\{Y_t : t = 0, 1, \ldots\}$ in formula (5.1)*. We assume that*

$$\sum_{j=0}^{\infty} j|\alpha_j| < \infty \tag{2.14}$$

which, among other things, implies that

$$\sum_{j=0}^{\infty} \alpha_j$$

is well defined. Restriction (2.14) *limits the temporal dependence of the* X *process.*

For this example

$$\begin{aligned} G_t &= \sum_{j=0}^{\infty} \left(\sum_{\ell=0}^{\infty} \alpha_{j+\ell} \right) \cdot W_{t-j}, \\ F_{t+1} &= \left(\sum_{j=0}^{\infty} \alpha_j \right) \cdot W_{t+1} \\ \overline{G}_t &= \sum_{j=0}^{\infty} \left(\sum_{\ell=0}^{\infty} \alpha_{j+1+\ell} \right) \cdot W_{t-j} \end{aligned}$$

The sum of the coefficients $\{\alpha_j\}_{j=0}^{\infty}$ *in the moving-average representation* (5.5) *tells the* permanent *effect of* W_{t+1} *on current and future values of the* level *of* Y*, i.e., on* $\lim_{j \to +\infty} Y_{t+j}$*. Later we will discuss how to use these calculations to identify permanent shocks to macro time series such as the logarithm of consumption and output. The variance of the random variable* $\left(\sum_{j=0}^{\infty} \alpha_j\right) \cdot W_{t+1}$ *is* $\left(\sum_{j=0}^{\infty} \alpha_j\right) \cdot \left(\sum_{j=0}^{\infty} \alpha_j\right)$ *conditioned on invariant events in* $\mathfrak{I}$*.*

Cointegration

In macroeconomics we often discuss balanced growth paths in models with explicit growth. The theory of cointegration allows us to formalize this using the decomposition that we just developed. The construct of cointegration has a variety of origins was developed in a more systemic way by Engle and Granger (1987).

Linear combinations of stationary increment processes $Y_t^{[1]}$ and $Y_t^{[2]}$ have stationary increments. For real valued scalars $\mathbf{r}_1$ and $\mathbf{r}_2$, form

$$Y_t = \mathbf{r}_1 Y_t^{[1]} + \mathbf{r}_2 Y_t^{[2]}$$

where

$$\begin{aligned} Y_{t+1}^{[1]} - Y_t^{[1]} &= X_{t+1}^{[1]} \\ Y_{t+1}^{[2]} - Y_t^{[2]} &= X_{t+1}^{[2]}. \end{aligned}$$

The increment in $\{Y_t : t = 0, 1, \ldots\}$ is

$$X_{t+1} = \mathbf{r}_1 X_{t+1}^{[1]} + \mathbf{r}_2 X_{t+1}^{[2]}.$$

and

$$Y_0 = \mathbf{r}_1 Y_0^{[1]} + \mathbf{r}_2 Y_0^{[2]}.$$

The Proposition 5.1.2 martingale component of $\{Y_t : t = 0, 1, \ldots\}$ is the corresponding linear combination of the martingale components of $\{Y_t^{[1]} : t = 0, 1, ...\}$ and $\{Y_t^{[2]} : t = 0, 1, ...\}$. The Proposition 5.1.2 trend component of $\{Y_t : t = 0, 1, \ldots\}$ is the corresponding linear combination of the trend components of $\{Y_t^{[1]} : t = 0, 1, \ldots\}$ and $\{Y_t^{[2]} : t = 0, 1, \ldots\}$.

Engle and Granger (1987) focus on linear combinations of stationary increment processes whose trend and martingale components are both zero. Engle and Granger call two processes *cointegrated* if there exists a linear combination of them that is stationary, which is true when there exist real valued scalars $\mathbf{r}_1$ and $\mathbf{r}_2$ such that

$$\begin{aligned} \mathbf{r}_1 \nu_1 + \mathbf{r}_2 \nu_2 &= 0 \\ \mathbf{r}_1 F^{[1]} + \mathbf{r}_2 F^{[2]} &= 0, \end{aligned}$$

where the ν's correspond to the trend components in Proposition 5.1.2. These two zero restrictions imply that the time trend and the martingale component for the linear combination Y_t are both zero.[22] When $\mathbf{r}_1 = 1$ and $\mathbf{r}_2 = -1$, the component stationary increment processes $Y_t^{[1]}$ and $Y_t^{[2]}$ share a common growth component as may occur in macroeconomic models with balanced growth paths.

Central Limit Theory

Using these same tools, we now explore a Central Theorem to complement our previously discussed Law of Large Numbers. Consider a measurement

[22] The cointegration vector $(\mathbf{r}_1, \mathbf{r}_2)$ is determined only up to scale.

function X in $\mathcal{H}$. To understand this result, apply Proposition and write

$$\frac{1}{\sqrt{t}}(Y_t - Y_0 - t\nu) = \frac{1}{\sqrt{t}}\sum_{j=0}^{t} F_t - \frac{1}{\sqrt{t}}\overline{G}_t + \frac{1}{\sqrt{t}}\overline{G}_0$$

From Billingsley (1961)'s central limit theorem for martingales

$$\frac{1}{\sqrt{t}}\sum_{j=0}^{t} F_t \Rightarrow \mathcal{N}\left(0, E\left[(F_t)^2 \,|\mathfrak{I}\right]\right),$$

where $\Rightarrow$ denotes weak convergence or equivalently convergence in distribution. Given the scaling by the horizon t, $\{(1/\sqrt{t})\overline{G}_t : t = 1, 2, ...\}$ and $\{(1/\sqrt{t})\overline{G}_0\}$ both converge in mean square to zero. Thus,

Proposition 2.11.4. *For all stationary increment processes* $\{Y_t : t = 0, 1, 2, ...\}$ *represented by* X *in* $\mathcal{H}$

$$\frac{1}{\sqrt{t}}(Y_t - \nu t) \Rightarrow \mathcal{N}\left(0, E\left[(F_t)^2|\mathfrak{I}\right]\right).$$

Furthermore,

$$E\left[(F_t)^2|\mathfrak{I}\right] = \lim_{t\to\infty} E\left(\left[\frac{1}{\sqrt{t}}(Y_t - Y_0 - t\nu)\right]^2 \Big|\mathfrak{I}\right).$$

Notice that an implication of this proposition is that

$$\sqrt{t}\left(\sum_{j=1}^{t} X_t - \nu\right) \Rightarrow \left(0, E\left[(F_t)^2 \,|\mathfrak{I}\right]\right)$$

This carries over to vector stochastic processes.[23]

[23] Moreover, there is an elegant extension of this establishing weak convergence to a Brownian motion.

Appendix 2.A Frequency Domain Representation

Processes having an infinite past can be analyzed in the frequency domain. Where z is a complex variable, the right side of

$$\alpha(z) = \sum_{j=0}^{\infty} \alpha_j z^j$$

is a power series expansion of the complex-valued function α. We assume that the power series converges for $|z| < 1$, a condition that is satisfied when the function α is analytic on the set $|z| < 1$.

We care about the behavior of $\alpha(z)$ near the boundary $z = 1$. To study this, it is convenient to invent a probability space and use it to construct a process having the same second moment processes as the original process (conditioned on invariant events). To show how to do this, suppose initially that $\{W_t : -\infty < t < +\infty\}$ is scalar. Let $\tilde{\omega} \in \widetilde{\Omega} = [0, 2\pi)$ and let $\widetilde{Pr}$ denote the uniform measure over $\widetilde{\Omega}$. Introduce a scalar complex valued process

$$\widetilde{W}_t(\tilde{\omega}) = \exp(-i\tilde{\omega}t)$$

for all integers t. Where $*$ denotes complex conjugation, notice that

$$\begin{aligned} \widetilde{W}_t \left(\widetilde{W}_s\right)^* &= \exp\left[i\tilde{\omega}(s-t)\right] \\ \left|\widetilde{W}_t\right|^2 &= 1. \end{aligned}$$

Since

$$\frac{1}{2\pi} \int_{\widetilde{\Omega}} \exp\left[i\tilde{\omega}(s-t)\right] = 0$$

for nonzero integer values of $s - t$, the process $\left\{\widetilde{W}_t : -\infty < t < +\infty\right\}$ is serially orthogonal. Moreover, the process has a unit second moment for all integer t. We construct a process

$$\widetilde{X}_t(\tilde{\omega}) = \exp(-\tilde{\omega}t)\alpha[\exp(i\tilde{\omega})]$$

that is well defined as a mean-square limit. Moreover, if we set $z = \rho\exp(i\tilde{\omega})$, then for almost all $\tilde{\omega}$, $\alpha[\exp(i\tilde{\omega})]$ is the limit of $\alpha(z)$ as ρ increases to one. Using the convenient $(\widetilde{\Omega}, \widetilde{Pr})$ probability space, we compute

$$\begin{aligned}\widetilde{E}\left|\widetilde{X}_t\right|^2 &= \frac{1}{2\pi}\int_{\widetilde{\Omega}} |\alpha[\exp(i\tilde{\omega})]|^2\, d\tilde{\omega} \\ &= \sum_{j=0}^{\infty} (\alpha_j)^2\,,\end{aligned}$$

where the second equality is known as Parseval's formula. Moreover,

$$\widetilde{E}\left[\widetilde{X}_t\left(\widetilde{X}_{t-s}\right)^*\right] = \frac{1}{2\pi}\int_{\widetilde{\Omega}} \exp(-i\tilde{\omega}s) \mid \alpha[\exp(i\tilde{\omega})] \mid^2 d\tilde{\omega}$$

for any s. By construction:

$$\begin{aligned}E\left(X_t X_{t-s}|\mathfrak{I}\right) &= \widetilde{E}\left[\widetilde{X}_t\left(\widetilde{X}_{t-s}\right)^*\right] \\ &= \frac{1}{2\pi}\int_{\widetilde{\Omega}} \exp(-i\tilde{\omega}s) \mid \alpha[\exp(i\tilde{\omega})] \mid^2 d\tilde{\omega}\end{aligned}$$

The function $\mid \alpha[\exp(i\tilde{\omega})] \mid^2$ is called the spectral density of the $\{X_t\}$ process.

Up to now we have assumed a scalar shock process. If there are multiple shocks that are uncorrelated, then we can repeat our calculations shock by shock. Now $\alpha(z)$ is vector of analytic functions

$$E\left(X_t X_{t-s}|\mathfrak{I}\right) = \frac{1}{2\pi}\int_{\widetilde{\Omega}} \exp(-i\tilde{\omega}s) \mid \alpha[\exp(i\tilde{\omega})] \mid^2 d\tilde{\omega},$$

and $\mid \alpha[\exp(i\tilde{\omega})] \mid^2 = \mid \alpha[\exp(i\tilde{\omega})]\alpha'[\exp(-i\tilde{\omega})] \mid$ is the spectral density matrix.

Appendix 2.B Hidden Periodicity

Example 2.B.1. *(Hidden periodicity) Suppose again that* $\{W_t : -\infty < t < \infty\}$ *is a vector stationary process for which*[24]

$$E\left(W_{t+1}|\mathfrak{A}_t\right) = 0$$

[24] This example extends Hansen and Sargent (1993) and Hansen and Sargent (2013, ch. 14), where we started with periodic autoregressive representations.

and that $E\left(W_t W_t' | \Im\right) = I$ *for all* $-\infty < t < +\infty$*. Where* $p \geq 2$ *is an integer that we call the periodicity, let there be* p *matrix sequences of moving average coefficients, one for each "season"* $\ell \in \{1, 2, \ldots, p\}$*. Suppose that*

$$X_t = \sum_{j=0}^{\infty} \alpha_j^{\ell(t)} \cdot W_{t-j} \tag{2.15}$$

where $\ell(t+1) = \ell(t) + 1$ *unless* $\ell(t) = p$*, in which case* $\ell(t+1) = 1$*. It follows that*

$$\ell(t+p) = \ell(t) \quad \forall t = \{0, \pm 1, \pm 2, \ldots\}. \tag{2.16}$$

Assume that for each $\ell \in \{1, 2, \ldots, p\}$

$$\sum_{j=0}^{\infty} |\alpha_j^{\ell}|^2 < \infty. \tag{2.17}$$

and calculate z*-transforms*

$$\alpha^{\ell}(z) = \sum_{j=0}^{\infty} \alpha_j^{\ell}(z)^j$$

for $\ell = 1, 2, ..., p$*. When* z*-transform* $\alpha^{\ell}(z)$ *applies at date* t*,* $\alpha^{\ell+1}(z)$ *applies at date* $t+1$*, modulo* p*, by which we mean that when* $\alpha^p(z)$ *is the transform at date* t*,* $\alpha^1(z)$ *is the transform at date* $t+1$*.*

Whether the stochastic process $\{X_t\}$ *in* (2.15) *is stationary depends on how we think of initializing* ℓ *at date* $t = 0$*. The following procedure initializes* ℓ *so that the process* $\{X_t\}$ *is stationary: randomize* ℓ *at date* $t = 0$ *by letting* ℓ *be each possible value* $\{1, \ldots, p\}$ *with probability* $1/p$*. Then if the shock process is an i.i.d. vector sequence of standard normally distributed random vectors, the* $\{X_t\}$ *process in* (2.15) *is stationary and ergodic. This stationary process has spectral density*

$$\frac{1}{p} \sum_{\ell=1}^{p} |\alpha^{\ell}\left[\exp\left(i\tilde{\omega}\right)\right]|^2$$

and associated inversion formula is

$$E\left(X_t X_{t-s}\right) = \frac{1}{2\pi} \int_{\tilde{\Omega}} \exp(-i\tilde{\omega}s) \left(\frac{1}{p} \sum_{\ell=1}^{p} |\ \alpha^{\ell}\left[\exp\left(i\tilde{\omega}\right)\right] |^2 \right) d\tilde{\omega}$$

for integer s.

Take such a stationary $\{X_t\}$ process and express the idea that we observe it only every p periods by constructing a skip-sampled *process*

$$\{X_{p\tau}, \quad \tau = 0, \pm 1, \pm 2 \ldots\}$$

that is also stationary but not ergodic. Use a value taken by the random variable ℓ to define an invariant event. The spectral density of the skip-sampled process conditioned on ℓ connects to the spectral density of the original $\{X_t\}$ process by the following "folding formula".[25]

$$\frac{1}{p}\sum_{k=1}^{p}\left|\alpha^{\ell}\left[\exp\left(i\left[\frac{\tilde{\omega}+2\pi(k-1)}{p}\right]\right)\right]\right|^2. \tag{2.18}$$

[25]To deduce formula (2.18), suppose that $s = pj$ for some integer j. Because $\exp(-i\tilde{\omega}s)$ is a strictly periodic function of $\tilde{\omega}$ with period 2π, notice that for any integer k $\exp(-i\tilde{\omega}s) = \exp\left[-i\left(\tilde{\omega}+\frac{2\pi k}{p}\right)s\right]$. The inversion formula is

$$EX_tX_{t-s} = \frac{1}{2\pi}\int_0^{\frac{2\pi}{p}} \exp(-i\tilde{\omega}s)\sum_{k=1}^{p}\left|\alpha^{\ell}\left(\exp\left[i\left(\tilde{\omega}+\frac{2\pi(k-1)}{p}\right)\right]\right)\right|^2 d\tilde{\omega}$$

After a change of variables $\tilde{\omega}p = \hat{\omega}$, we obtain

$$EX_tX_{t-pj} = \frac{1}{2\pi}\int_0^{2\pi} \exp(-i\hat{\omega}j)\,\frac{1}{p}\sum_{k=1}^{p}\left|\alpha^{\ell}\left[\exp\left(i\left[\frac{\hat{\omega}+2\pi(k-1)}{p}\right]\right)\right]\right|^2 d\hat{\omega}$$

for integer j, which identifies (2.18) as the spectral density of the skip-sampled process for $\ell = 1, 2, \ldots, p$.