Risk, Uncertainty, and Value

Lars Peter Hansen
University of Chicago and NBER

Thomas J. Sargent
New York University and Hoover Institution

December 24, 2022

Contents

List of Figures

Chapter 1

Stochastic Processes and Laws of Large Numbers

1.1 Introduction

A probabilistic form of invariance gives rise to a Law of Large Numbers. The invariance notion is a stochastic counterpart to a steady state of a dynamic economic model. The Law of Large Numbers conditions on a set of special events called *invariant events* that we can interpret as indexing alternative possible statistical models. These ideas allow us to characterize what can be learned from time series evidence and what must originate elsewhere.

1.2 Stochastic Processes

A sequence of random vectors is called a stochastic process. we are interested in time series so we index the sequence 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. The following definition makes reference to Borel sets. Borel sets include open sets, closed sets, finite intersections, and countable unions of such sets.

Definition 1.2.1. *X is an n-dimensional random vector if $X : \Omega \rightarrow \mathbb{R}^n$ has the property that for any Borel set $\mathfrak{b}$ in $\mathbb{R}^n$ $\{X \in \mathfrak{b}\}$ is in $\mathfrak{F}$.*

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

This formal structure facilitates using mathematical analysis to formulate problems in probability theory. A random vector induces a probability distribution over the collection of Borel sets in which the probability assigned to set $\mathfrak{b}$ is given by

$$Pr\{X \in \mathfrak{b}\}$$

By changing the set $\mathfrak{b}$, we trace out a probability distribution implied by the random vector X that is called the *induced distribution.* An induced distribution is what typically interests an applied worker. In practice, an induced distribution is just specified directly without constructing the foundations under study here. However, proceeding at a deeper level as we have by defining a random vector to be a function that satisfies particular measurable properties and imposing the probability measure Pr over the domain of that function has mathematical payoffs that we will exploit in various ways, among them being in construction of stochastic processes.

Definition 1.2.2. *An n-dimensional stochastic process is an infinite sequence of n-dimensional random vectors $\{X_t : t = 0, 1, ...\}$.*

The measure Pr assigns probabilities to a rich and interesting collection of events. For example, consider a stacked random vector

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

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

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

Since the choice of ℓ is arbitrary, Pr implies a distribution over a sequence of random vectors $\{X_t(\omega) : t = 0, 1, ...\}$: given a probability distribution, we can construct a probability space and a random vector that induces this distribution. This insight extends to the study of stochastic processes. Thus, the following way to construct a probability space is particularly enlightening.

Construction 1.2.3. *Let Ω be a collection of infinite sequences in $\mathbb{R}^n$ with an element $\omega \in \Omega$ being a sequence of vectors $\omega = (\mathbf{r}_0, \mathbf{r}_1, ...)$, where $\mathbf{r}_t \in \mathbb{R}^n$. To construct $\mathfrak{F}$, proceed as follows. 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*

$$\Lambda = \{\omega = (\mathbf{r}_0, \mathbf{r}_1, ...) : \mathbf{r}_j \in \mathfrak{b}_j, j = 0, 1, .., \ell\}. \tag{1.1}$$

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.

Next we construct a measure that assigns probabilities to events in $\mathfrak{F}$. For each integer $\ell \geq 0$, let Pr_ℓ assign probabilities to the Borel sets of $\mathbb{R}^{n(\ell+1)}$. A Borel set in $\mathbb{R}^{n(\ell+1)}$ can also be viewed as a Borel set in $\mathbb{R}^{n(\ell+2)}$ with $\mathbf{r}_{n(\ell+1)}$ left unrestricted. Specifically, let $\mathfrak{b}_\ell$ be a Borel set in $\mathbb{R}^{n(\ell+1)}$. Then

$$\mathfrak{b}_\ell^{\ell+1} = \{(\mathbf{r}_0, \mathbf{r}_1, ..., \mathbf{r}_\ell, \mathbf{r}_{\ell+1}) : (\mathbf{r}_0, \mathbf{r}_1, ..., \mathbf{r}_\ell) \in \mathfrak{b}_\ell\}.$$

For the probability measures $\{Pr_\ell : \ell = 0, 1, ...\}$ to be consistent, we require that the probability assigned by $Pr_{\ell+1}$ satisfy

$$Pr_\ell(\mathfrak{b}_\ell) = Pr_{\ell+1}\left(\mathfrak{b}_\ell^{\ell+1}\right)$$

for any $\ell \geq 0$ and any Borel set $\mathfrak{b}_\ell$ in $\mathbb{R}^{n(\ell+1)}$. If consistency in this sense prevails, we can extend this construction to form a probability Pr on the space $(\Omega, \mathfrak{F})$ that is consistent with the probability assigned by Pr_ℓ for all nonnegative integers ℓ.[1]

Finally, we construct the stochastic process $\{X_t : t = 0, 1, ...\}$ by letting

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

for $t = 0, 1, 2, ...$ A convenient feature of this construction is that Pr_ℓ is the probability induced by the random vector $[{X_0}', {X_1}', ..., {X_\ell}']'$.

We refer to this construction as ***canonical****. While this is only one among other possible constructions of probability spaces, it illustrates the flexibility in building sequences of random vectors that induce alternative probabilities of interest.*

[1] This essentially follows from the Kolomorov Extension Theorem or from Theorem 2.26 of Breiman (1968).

The remainder of this chapter is devoted to studying Laws of Large Numbers. What is perhaps the most familiar Law of Large Numbers presumes that the stochastic process $\{X_t : t = 0, 1, ...\}$ is independent and identically distributed (iid). Then

$$\frac{1}{N}\sum_{t=1}^{N}\varphi(X_t) \to E\varphi(X_0)$$

for any (Borel measurable) function φ for which the expectation is well defined. Convergence holds in several senses that we state later. Notice that as we vary the function φ we can infer the (induced) probability distribution for X_0. In this sense, the outcome of the Law of Large Numbers under an iid sequence determines what we will call a *statistical model.*

For our purposes, an iid version of the Law of Large Numbers is too restrictive. First, we are interested in economic dynamics in which model outcomes are temporally dependent. Second, we want to put ourselves in the situation of a statistician who does not know *a priori* what the underlying data generating process is and therefore entertains multiple models. We will present a Law of Large Numbers that covers both settings.

1.3 Constructing a Stochastic Process

We now generalize the canonical construction 1.2.3 of a stochastic process in a way that facilitates stating the Law of Large Numbers that interests us.

We use two objects.[2] The first is a (measurable) transformation $\mathbb{S} : \Omega \to \Omega$ that describes the evolution of a sample point ω. See figure 1.1. Transformation $\mathbb{S}$ 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 $\mathfrak{F}$, as depicted in figure 1.2. The second object is an n-dimensional vector $X(\omega)$ that describes how observations depend on sample point ω. We construct a stochastic process $\{X_t : t = 0, 1, ...\}$ via the formula:

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

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

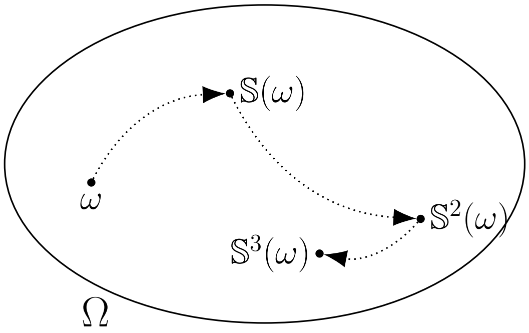


Figure 1.1: The evolution of a sample point ω induced by successive applications of the transformation $\mathbb{S}$. The oval shaped region is the collection Ω of all sample points.

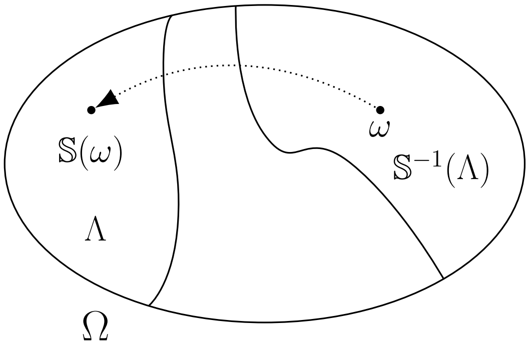


Figure 1.2: An inverse image $\mathbb{S}^{-1}(\Lambda)$ of an event Λ is itself an event; $\omega \in \mathbb{S}^{-1}(\Lambda)$ implies that $\mathbb{S}(\omega) \in \Lambda$.

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*: ω_{t+j} for all $j \geqslant 1$ 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.

This way of constructing a stochastic process might seem restrictive; but actually, it is more general than the canonical construction presented above.

Example 1.3.1. *Consider again our canonical construction 1.2.3. Recall that the set of sample points Ω is the collection of infinite sequences of elements of $\mathbf{r}_t \in \mathbb{R}^n$ so that $\omega = (\mathbf{r}_0, \mathbf{r}_1, ...)$. For this example, $\mathbb{S}(\omega) = (\mathbf{r}_1, \mathbf{r}_2, ...)$. This choice of $\mathbb{S}$ is called the* shift *transformation. Notice that the time t iterate is*

$$\mathbb{S}^t(\omega) = (\mathbf{r}_t, \mathbf{r}_{t+1}, ...)$$

Let the measurement function be: $X(\omega) = \mathbf{r}_0$ so that

$$X_t(\omega) = X\left[\mathbb{S}^t(\omega)\right] = \mathbf{r}_t$$

as posited in construction 1.2.3.

1.4 Stationary Stochastic Processes

We start with a probabilistic notion of invariance. We call a stochastic process *stationary* if any finite integer ℓ, the joint probability distribution induced by the composite random vector $[X_t{}', X_{t+1}{}', ..., X_{t+\ell}{}']'$ is the same for all $t \geq 0$.[3] This notion of stationarity can be thought of as a stochastic version of a steady state of a dynamical system.

We now use the objects $(\mathbb{S}, X)$ to build a stationary stochastic process by restricting construction 1.2.3. Consider the set $\{\omega \in \Omega : X(\omega) \in \mathfrak{b}\} \doteq \Lambda$ and its successors

$$\begin{aligned}\{\omega \in \Omega : X_1(\omega) \in \mathfrak{b}\} &= \{\omega \in \Omega : X\left[\mathbb{S}(\omega)\right] \in \mathfrak{b}\} = \mathbb{S}^{-1}(\Lambda)\\ \{\omega \in \Omega : X_t(\omega) \in \mathfrak{b}\} &= \{\omega \in \Omega : X\left[\mathbb{S}^t(\omega)\right] \in \mathfrak{b}\} = \mathbb{S}^{-t}(\Lambda).\end{aligned}$$

Evidently, if $Pr(\Lambda) = Pr[\mathbb{S}^{-1}(\Lambda)]$ for all $\Lambda \in \mathfrak{F}$, then the probability distribution induced by X_t equals the probability distribution of X for all t. This fact motivates the following definition and proposition.

[3] 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.

Definition 1.4.1. *The pair* $(\mathbb{S}, Pr)$ *is said to be* ***measure-preserving*** *if*

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

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

Proposition 1.4.2. *When* $(\mathbb{S}, Pr)$ *is measure-preserving, probability distributions induced by the random vectors* X_t *are identical for all* $t \geq 0$.

The measure preserving property restricts the probability measure Pr for a given transformation $\mathbb{S}$. Some probability measures Pr used in conjunction with $\mathbb{S}$ will be measure preserving and others not, a fact that will play an important role at several places below.

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

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

We can apply Proposition 1.4.2 to $X^{[\ell]}$ to conclude that the joint distribution function of $(X_t, X_{t+1}, ..., X_{t+\ell})$ is independent of t for $t = 0, 1, \ldots$. That this property holds for any choice of ℓ implying that the stochastic process $\{X_t : t = 1, 2, ...\}$ is stationary. Moreover, $f\left(X^\ell\right)$ where f is a Borel measurable function from $\mathbb{R}^{n(\ell+1)}$ into $\mathbb{R}$ is also a valid measurement function. Such f's include indicator functions of interesting events defined in terms of X^ℓ.

For a given $\mathbb{S}$, we now present examples that illustrate how to construct a probability measure Pr that makes $\mathbb{S}$ measure preserving and thereby brings stationarity. In example 1.4.3, only one Pr makes $\mathbb{S}$ measure preserving, while in example 1.4.4 there are many.

Example 1.4.3. *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 1.4.4. *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* $Pr(\{\omega_1\}) \geqslant 0$ *and* $Pr(\{\omega_2\}) = 1 - Pr(\{\omega_1\})$.

The next example illustrates how to represent an i.i.d. sequence of zeros and ones in terms of an Ω, Pr and an $\mathbb{S}$.

Example 1.4.5. *Suppose that* $\Omega = [0,1)$ *and that* Pr *is the uniform measure on* $[0,1)$. *Let*

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

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

Calculate $Pr\{X_1 = 1 | X_0 = 1\} = Pr\{X_1 = 1 | X_0 = 0\} = Pr\{X_1 = 1\} = 1/2$ *and* $Pr\{X_1 = 0 | X_0 = 1\} = Pr\{X_1 = 0 | X_0 = 0\} = Pr\{X_1 = 0\} = 1/2$. *So* X_1 *is statistically independent of* X_0. *By extending these calculations, it can be verified that* $\{X_t : t = 0, 1, ...\}$ *is a sequence of independent random variables.*[4] *We can alter* Pr *to obtain other stationary distributions. For instance, suppose that* $Pr\{\frac{1}{3}\} = Pr\{\frac{2}{3}\} = .5$. *Then the process* $\{X_t : t = 0, 1, ...\}$ *alternates in a deterministic fashion between zero and one. This provides a version of Example 1.4.3 in which* $\omega_1 = \frac{1}{3}$ *and* $\omega_2 = \frac{2}{3}$.

1.5 Invariant Events and Conditional Expectations

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

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

Figure 1.3 illustrates two invariant events in 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.

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

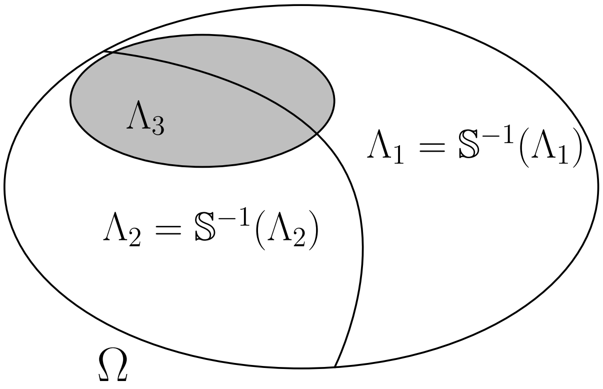

Λ_3
$\Lambda_1 = \mathbb{S}^{-1}(\Lambda_1)$
$\Lambda_2 = \mathbb{S}^{-1}(\Lambda_2)$
Ω

Figure 1.3: Two invariant events Λ_1 and Λ_2 and an event Λ_3 that is not invariant.

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". We begin with a situation in which a conditional expectation is a discrete random vector as occurs when invariant events are unions of sets Λ_j belonging to a countable partition of Ω (together 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

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

Thus, the conditional expectation $E(X|\mathfrak{I})$ is constant for $\omega \in \Lambda_j$ but varies across Λ_j's. Figure 1.4 illustrates this characterization for a finite partition.

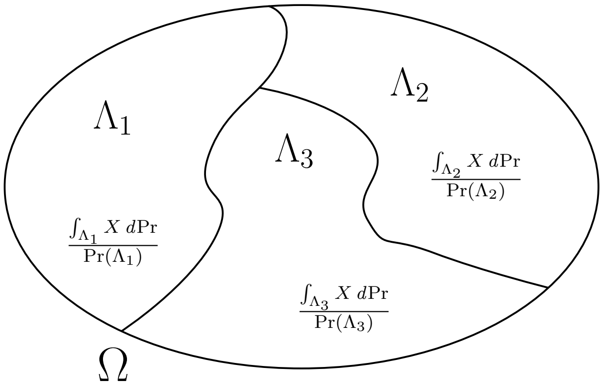

Λ_1
Λ_2
Λ_3
$\frac{\int_{\Lambda_1} X\, d\mathrm{Pr}}{\mathrm{Pr}(\Lambda_1)}$
$\frac{\int_{\Lambda_2} X\, d\mathrm{Pr}}{\mathrm{Pr}(\Lambda_2)}$
$\frac{\int_{\Lambda_3} X\, d\mathrm{Pr}}{\mathrm{Pr}(\Lambda_3)}$
Ω

Figure 1.4: A conditional expectation $E(X|\mathfrak{I})$ is constant for $\omega \in \Lambda_j = \mathbb{S}^{-1}(\Lambda_j)$

Least Squares

Now let X be a random vector with finite second moments $EXX' = \int X(\omega)X(\omega)'dPr(\omega)$. When a random vector X has finite second moments, a conditional expectation is a least squares projection. Let Z be an n-dimensional measurement function that is time-invariant and so satisfies

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

Let $\mathcal{Z}$ denote the collection of all such time-invariant random vectors. 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.[5] Consider the least squares problem

$$\min_{Z \in \mathcal{Z}} E\big[|X - Z|^2\big]. \tag{1.2}$$

Denote the minimizer in problem 1.2 by $\tilde{X} = E(X|\mathfrak{I})$. Necessary conditions for the least squares minimizer $\widetilde{X} \in \mathcal{Z}$ imply that

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

for Z in $\mathcal{Z}$ so that each entry of the vector $X - \widetilde{X}$ of regression errors is orthogonal to every vector Z in $\mathcal{Z}$.

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

A measure-theoretic approach constructs a conditional expectation by extending the orthogonality property of least squares. Provided that $E|X| < \infty$, $E(X|\mathfrak{I})(\omega)$ 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.

1.6 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. To do this, we use a notion of probabilistic invariance that is expressed in terms of the measure-preserving restriction and that implies a Law of Large Numbers applicable to stochastic processes. The following theorem asserts two senses in which averages of intertemporally dependent processes converge to mathematical expectations conditioned on invariant events.

Theorem 1.6.1. *(Birkhoff) Suppose that $\mathbb{S}$ is measure preserving relative to the probability space $(\Omega, \mathfrak{F}, Pr)$.*[6]

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 $E|X(\omega)|^2 < \infty$,

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

[6] See Breiman (1968) chapter 6 for extended discussions and proofs.

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

We have ample flexibility to specify a measurement function $\varphi\left(X^{\ell}\right)$, where φ is a Borel measurable function from $\mathbb{R}^{n(\ell+1)}$ into $\mathbb{R}$. In particular, an indicator functions for event $\Lambda = \{X^{\ell} \in \mathfrak{b}\}$ can be used as a measurement function where:

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

The Law of Large Numbers applies to limits of

$$\frac{1}{N}\sum_{t=1}^{N} \varphi\left[X_t^{\ell}\right]$$

for alternative φ's, so choosing φ's to be indicator functions shows how the Law of Large Numbers uncovers event probabilities of interest.

Definition 1.6.2. *A transformation $\mathbb{S}$ that is measure-preserving relative to Pr is said to be* ***ergodic*** *under 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 1.6.3. *Suppose that the measure preserving transformation $\mathbb{S}$ is ergodic under measure Pr. Then $E(X|\mathfrak{I}) = E(X)$.*

Theorem 1.6.1 describes conditions for convergence in the general case that $\mathbb{S}$ is measure preserving under Pr but in which $\mathbb{S}$ is not necessarily ergodic under Pr. Proposition 1.6.3 describes a situation 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 can call a *standard* Law of Large Numbers. When $\mathbb{S}$ is not ergodic under Pr, limit points of time series averages equal expectations conditioned on invariant events.

The following examples remind us how ergodicity restricts $\mathbb{S}$ and Pr.

Example 1.6.4. *Consider example 1.4.3 again. Ω contains two points and $\mathbb{S}$ maps ω_1 into ω_2 and ω_2 into ω_1: $\mathbb{S}(\omega_1) = \omega_2$ and $\mathbb{S}(\omega_2) = \omega_1$. 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 1.6.5. *Return to example 1.4.4. Ω contains two points, $\Omega = \{\omega_1, \omega_2\}$ and that $\mathbb{S}(\omega_1) = \omega_1$ and $\mathbb{S}(\omega_2) = \omega_2$. $X_t(\omega) = X(\omega)$ so that the sequence is time invariant and equal to its time-series average. A 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.*

1.7 Limiting Empirical Measures

Given a triple $(\Omega, \mathfrak{F}, Pr)$ and a measure-preserving transformation $\mathbb{S}$, we can use Theorem 1.6.1 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. Given an event Λ in $\mathfrak{F}$ and for almost all $\omega \in \Lambda_j$, define the limiting empirical measure Qr_j as

$$Qr_j(\Lambda)(\omega) = \lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} \mathbf{1}_\Lambda \left[\mathbb{S}^t(\omega)\right] = \frac{Pr(\Lambda \cap \Lambda_j)}{Pr(\Lambda_j)}. \tag{1.3}$$

Thus, when $\omega \in \Lambda_j$, $Qr_j(\Lambda)$ is the fraction of time $\mathbb{S}^t(\omega) \in \Lambda$ in very long samples. If we hold Λ_j fixed and let 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 $\{Qr_j\}_{j=1}^{\infty}$. These comprise the set of all measures that can be recovered by applying the Law of Large Numbers. If nature draws an $\omega \in \Lambda_j$, then measure Qr_j describes outcomes.

So far, we started with a probability measure Pr and then constructed the set of possible limiting empirical measures Qr_j's. We now reverse the direction of the logic by starting with probability measures Qr_j and then finding measures Pr that are consistent with them. We do this because Qr_j's are the only measures that long time series can disclose through the Law of Large Numbers: each Qr_j defined by (1.3) uses the Law of Large Numbers to assign probabilities to events $\Lambda \in \mathfrak{F}$. However, because

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

are conditional probabilities, such Qr_j's are silent about the probabilities $Pr(\Lambda_j)$ of the underlying invariant events Λ_j. There are multiple ways to assign probabilities Pr that imply identical probabilities conditioned on invariant events.

Because Qr_j is all that can ever be learned by "letting the data speak", we regard each probability measure Qr_j as a statistical model.[7]

Definition 1.7.1. *A* statistical model *is a probability measure that a Law of Large Numbers can disclose.*

Probability measure Qr_j describes a statistical model associated with invariant set Λ_j.

Remark 1.7.2. *For each j, $\mathbb{S}$ is measure-preserving and ergodic on $(\Omega, \mathfrak{F}, Qr_j)$. The second equality of definition* (1.3) *assures ergodicity by assigning probability one to the event Λ_j.*

Relation (1.3) implies that probability Pr connects to probabilities Qr_j by

$$Pr(\Lambda) = \sum_j Qr_j(\Lambda) Pr(\Lambda_j). \tag{1.4}$$

[7] Marschak (1953), Hurwicz (1962), Lucas (1976), and Sargent (1981) distinguished 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 and change others. A structural model accepts experiments that *alter* statistical models.

While decomposition (1.4) follows from definitions of the elementary objects that comprise a stochastic process and is "just mathematics", it is interesting because it tells 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, to respect evidence from long time series we must hold the Qr_j's fixed, but we can freely assign probabilities Pr to invariant events Λ_j. In this way, we can create a family of probability measures for which $\mathbb{S}$ is measure preserving.

1.8 Ergodic Decomposition

Up to now, we have represented invariant events with a countable partition. Dynkin (1978) deduced a more general version of decomposition (1.4) without assuming a countable partition. Thus, start with a pair $(\Omega, \mathfrak{F})$. Also, 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) constructed limiting empirical measures Qr on $\mathcal{Q}$ and justified the following representation of the set $\mathcal{P}$ of probability measures Pr.

Proposition 1.8.1. *For each probability measure $\widetilde{Pr}$ in $\mathcal{P}$ there is a unique probability measure π over $\mathcal{Q}$ such that*

$$\widetilde{Pr}(\Lambda) = \int_{\mathcal{Q}} Qr(\Lambda)\pi(dQr) \tag{1.5}$$

for all $\Lambda \in \mathfrak{F}$.[8]

[8] Krylov and Bogolioubov (1937) provide 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.

Proposition 1.8.1 generalizes representation (1.4). It asserts a sense in which the set $\mathcal{P}$ of probabilities for which $\mathbb{S}$ is measure-preserving is convex. Extremal points of this set are in the smaller set $\mathcal{Q}$ of probability measures for which the transformation $\mathbb{S}$ is ergodic. Representation (1.5) 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.

To add another perspective, a collection of invariant events $\mathfrak{I}$ is associated with a transformation $\mathbb{S}$. 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 indicator functions of events of interest as well as other random vectors. Alternative probability measures Pr assign different probabilities to the invariant events.

1.9 Risk and uncertainty

An applied researcher typically does not know 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 (1.5) describes uncertainty about statistical models with a probability distribution π over the set of statistical models $\mathcal{Q}$.

For a Bayesian, π is a subjective prior probability distribution that pins down a convex combination of "statistical models."[9] A Bayesian expresses trust in that convex combination of statistical models used to construct a complete probability measure over outcomes and uses it to compute expected utility.[10] A Bayesian decision theory axiomatized by Savage makes no distinction between how decision makers respond to the probabilities described by the component statistical models and the π probabilities that he uses to mix them. All that matters to a Bayesian decision maker is the

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

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

complete probability distribution over outcomes, not how it is attained as a π-mixture of component statistical models.

Some decision and control theorists challenge the complete confidence in a single prior probability assumed in a Bayesian approach.[11] They want to distinguish 'ambiguity', meaning not being able confidently to assign π, from 'risk', meaning prospective outcomes with probabilities reliably described by a statistical model. They imagine decision makers who want to evaluate decisions under alternative π's.[12] We explore these ideas in later chapters.

An important implication of the Law of Large Numbers is that for a given initial π, using Bayes' rule to update the π probabilities as data arrive will eventually concentrate posterior probability on the statistical model that generates the data. Even when a decision maker entertains a family of π's, the updated probabilities conditioned on the data may still concentrate on the statistical model that generates the data.

1.10 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 assumed 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 chapter 2, we want more. To prepare the way for that chapter, in this section we shall also assume that $\mathbb{S}$ is one-to-one and has the property that for any event $\Lambda \in \mathfrak{F}$,

$$\mathbb{S}(\Lambda) = \{\omega \in \Omega : \mathbb{S}^{-1}(\omega) \in \Lambda\} \in \mathfrak{F}. \tag{1.6}$$

Because

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

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

[12] This gives one way to formalize ideas of Knight (1921), who sought to distinguish risk from broader notions of uncertainty.

is well defined for negative values of t, restrictions 1.6 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

$$\mathfrak{A}_t = \left\{ \Lambda_t \in \mathfrak{F} : \Lambda_t = \{\omega \in \Omega : \mathbb{S}^t(\omega) \in \Lambda\} \text{ for some } \Lambda \in \mathfrak{F} \right\}. \tag{1.7}$$

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. Invariant events in $\mathfrak{I}$ are contained in $\mathfrak{A}_t$ for all t.

We construct the following moving-average representation of a scalar process $\{X_t\}$ in terms of an infinite history of shocks.

Example 1.10.1. *(Moving average) Suppose that $\{W_t : -\infty < t < \infty\}$ is a vector stationary process for which*[13]

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

and that $E\left(W_t {W_t}' | \mathfrak{I}\right) = I$ for all $-\infty < t < +\infty$. Use a sequence of vectors $\{\alpha_j\}_{j=0}^{\infty}$ to construct

$$X_t = \sum_{j=0}^{\infty} \alpha_j \cdot W_{t-j} \tag{1.8}$$

where

$$\sum_{j=0}^{\infty} |\alpha_j|^2 < \infty. \tag{1.9}$$

Restriction (1.9) *implies that X_t is well defined as a mean square limit. X_t is constructed from the infinite past $\{W_{t-j} : 0 \leqslant j < \infty\}$. 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.*

Remark 1.10.2. *Almost a century ago, both Slutsky (1927) and Yule (1927) used probability models to analyze economic time series. Their models implied moving-average representations like the one in Example 1.10.1. Their idea was to see economic time series as responding linearly to current and*

[13] An i.i.d. sequence is just one example of such a $\{W_t : -\infty < t < \infty\}$ process.

past independent and identically distributed impulses or shocks. In distinct contributions, they showed how such models generate recurrent but aperiodic fluctuations that resemble business cycles and longer-term cycles as well. Yule and Slutsky came from different backgrounds and brought different perspectives. Yule was an eminent statistician who, among other important contributions, managed "effectively to invent modern time series analysis" in the words of Stigler (1986). Yule constructed and estimated what we would now call a second-order autoregression and applied it to study sunspots. Yule's estimates implied α_j coefficients showed damped oscillations at the same periodicity as sunspots. In Russia in the 1920s, Slutsky wrote a seminal paper in Russian motivated by his interest in business cycles. Only later was an English version of his paper published Econometrica. *Even before that, it was already on the radar screen of economists including Ragnar Frisch. Indeed, Frisch was keenly aware of both Slutsky (1927) and Yule (1927) and generously acknowledged both of them in his seminal paper Frisch (1933) on the impulse and propagation problem. Building on insights of Slutsky and Yule, Frisch pioneered impulse response functions. His ambition was to provide explicit economic interpretations for how shocks alter economic time series both now and later.*[14]

1.11 Summary

For a fixed $\mathbb{S}$ there are often many possible probabilities Pr that are measure preserving. A subset of these are ergodic. These ergodic probabilities can serve as building blocks for the other measure preserving probabilities. Thus, each measure preserving Pr can be expressed as a weighted average of the ergodic probabilities. We call the ergodic probabilities statistical models. The Law of Large of Numbers applies to each of the ergodic building blocks with limit points that are unconditional expectations. As embodied in (1.4) and its generalization (1.5), this decomposition interests both frequentist and Bayesian statisticians.

[14] Sims (1980) and others advanced this idea by developing tractable multivariate time series methods and striving to isolate interpretable shocks in multivariate settings.

Chapter 2

Stationary Increments

Logarithms of many economic time series that appear to display stochastic growth can be modeled as having stationary increments. Multivariate versions of these models possess stochastic process versions of balanced growth paths. Applied econometricians seek permanent shocks that contribute to such growth. Furthermore, we shall see that it is convenient to pose central limit theory in terms of processes with stationary increments. The mathematical formulation in this chapter opens door to studying these topics.

2.1 Basic setup

We adopt assumptions from section 1.10 that allow an infinite past and again let $\mathfrak{A}$ be a subsigma algebra of $\mathfrak{F}$ and

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

Let X be a scalar measurement function. Assume that Y_0 is $\mathfrak{A}_0$ measurable and consider a scalar process $\{Y_t : t = 0, 1, ...\}$ with stationary increments $\{X_t\}$:

$$Y_{t+1} - Y_t = X_{t+1} \tag{2.1}$$

for $t = 0, 1, \ldots$. Let

$$\begin{aligned} U_{t+1} &= X_{t+1} - E\left(X_{t+1} | \mathfrak{A}_t\right) \\ \nu &= E\left(X_{t+1} | \mathfrak{I}\right) \\ V_t &= E\left(X_{t+1} | \mathfrak{A}_t\right) - \nu. \end{aligned}$$

Evidently

$$X_{t+1} = U_{t+1} + V_t + \nu.$$

We can interpret the above equation as providing an interesting decomposition of the $\{Y_t : t \geq 0\}$ process. Thus, component U_{t+1} is unpredictable and represents new information about Y_{t+1} that arrives at date $t+1$. Component V_t is the date $t+1$ contribution to Y_{t+1} that can be predicted from time t information net of trend growth. Component ν is the trend rate of growth or decay in $\{Y_t : t \geq 0\}$ conditioned on the invariant events. In the following sections, we present an alternative decomposition that will be useful both in connecting to sources of permanent versus transitory shocks and to central limit theorems.

2.2 A martingale decomposition

A special class of stationary increment processes called additive martingales interests us.

Definition 2.2.1. *The process* $\{Y_t^m : t = 0, 1, ...\}$ *is said to be an* **additive martingale** *relative to* $\{\mathfrak{A}_t : t = 1, 2, ...\}$ *if for* $t = 0, 1, ...$

- Y_t^m *is* $\mathfrak{A}_t$ *measurable, and*
- $E\left(Y_{t+1}^m | \mathfrak{A}_t\right) = Y_t^m$.

Notice that by the Law of Iterated Expectations, for a martingale $\{Y_t^m : t \geq 0\}$, best forecasts satisfy:

$$E\left(Y_{t+j}^m \mid \mathfrak{A}_t\right) = Y_t^m$$

for $j \geq 1$. Under suitable additional restrictions on the increment process $\{X_t : t \geq 0\}$, we can deploy a construction of Gordin (1969) to show that the $\{V_t\}$ process contributes a martingale component to the $\{Y_t^m : t = 0, 1, ...\}$ process.[1] Let $\mathcal{H}$ denote the set of all scalar random variables X such that $E(X^2) < \infty$ and such that[2]

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

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

[2] The random variable H_t somewhat resembles an "undiscounted" version of the resolvent operator that plays an important role in the analysis of Markov processes in chapter 3.

is well defined as a mean-square convergent series. Convergence of the infinite sum on the right side limits temporal dependence of the process $\{X_t\}$. For example, it can exclude so-called long memory processes.[3]

Construct the one-period ahead forecast of H_{t+1}:

$$H_t^+ = E\left(H_{t+1} \mid \mathfrak{A}_t\right)$$

Notice that

$$X_t - \nu = H_t - H_t^+ = G_t + \left(H_{t-1}^+ - H_t^+\right)$$

where

$$G_t = H_t - H_{t-1}^+ = H_t - E\left(H_t \mid \mathfrak{A}_{t-1}\right). \tag{2.2}$$

Since G_t is a forecast error,

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

Assembling these parts, we have

$$Y_{t+1} - Y_t = X_{t+1} = \nu + G_{t+1} + H_t^+ - H_{t+1}^+. \tag{2.3}$$

Let

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

Since Y_t^m is $\mathfrak{A}_t$ measurable, the equality

$$E\left(\sum_{j=1}^{t+1} G_j \mid \mathfrak{A}_t\right) = \sum_{j=1}^{t} G_j$$

implies that the process $\{Y_t^m : t \geq 0\}$ is an *additive martingale.*

For a given stationary increment process, $\{Y_t : t \geq 0\}$, express the martingale increment as

$$\begin{aligned} G_t &= \sum_{j=0}^{\infty} \left[E\left(X_{t+j} \mid \mathfrak{A}_t\right) - E\left(X_{t+j} \mid \mathfrak{A}_{t-1}\right)\right] \\ &= \lim_{j \to \infty} \left[E\left(Y_{t+j} | \mathfrak{A}_t\right) - E\left(Y_{t+j} | \mathfrak{A}_{t-1}\right)\right]. \end{aligned} \tag{2.4}$$

So the increment to the martingale component of $\{Y_t : t \geq 0\}$ is new information about the limiting optimal forecast of Y_{t+j} as $j \to +\infty$.

By accumulating equation (2.3) forward, we arrive at:

[3] See, for instance, Granger and Joyeux (1980), Geweke and Porter-Hudak (1983) and Robinson (1994).

Proposition 2.2.2. *If X is in $\mathcal{H}$, the stationary increments process $\{Y_t : t = 0, 1, ...\}$ satisfies the additive decomposition*

$$Y_t \; = \; \underbrace{t\nu}_{\textbf{\textit{trend}}} \; + \; \underbrace{Y_t^m}_{\textbf{\textit{martingale}}} \; - \; \underbrace{H_t^+}_{\textbf{\textit{stationary}}} \; + \; \underbrace{Y_0 + H_0^+}_{\textbf{\textit{invariant}}}.$$

The martingale component $\{Y_t^m : t \geq 0\}$, $Y_0^m = 0$, and the component $\{H_t^+\}$ is stationary.

We can use the Proposition 2.2.2 decomposition to determine a time trend, a “permanent shock”, and a transitory component of a stationary-increments process like (2.1). The permanent shock is the increment to the martingale. There are multiple ways to construct a transitory component, some of which yield transitory shocks that are correlated with permanent shocks.

Example 2.2.3. *(Moving-average increment process) Consider again the Example 1.10.1 moving-average process:*

$$X_t = \sum_{j=0}^{\infty} \alpha_j \cdot W_{t-j}. \tag{2.5}$$

Use this $\{X_t\}$ process as the increment for $\{Y_t : t \geq 0\}$ in formula (2.1). New information about the unpredictable component of X_{t+j} for $j \geq 0$ that arrives at date t is

$$E\left(X_{t+j} \mid \mathfrak{A}_t\right) - E\left(X_{t+j} \mid \mathfrak{A}_{t-1}\right) = \alpha_j \cdot W_t$$

Summing these terms over j gives

$$G_t = \alpha(1) \cdot W_t$$

where

$$\alpha(1) = \sum_{j=0}^{\infty} \alpha_j$$

provided that the coefficient sequence $\{\alpha_j : j \geq 0\}$ is summable, a condition that restricts temporal dependence of the increment process $\{X_t\}$. Indeed it is possible for $\alpha(1) = \infty$ or for it not to be well defined while

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

ensuring that X_t is well defined. This possibility opened the door to the literature on long-memory processes that allow for $\alpha(1)$ to be infinite as discussed in Granger and Joyeux (1980) and elsewhere.

In what follows, we presume that $\alpha(1)$ is finite. This sum of the coefficients $\{\alpha_j : j \geq 0\}$ in moving-average representation (2.5) *for the* first difference $Y_{t+1} - Y_t = X_{t+1}$ *of* $\{Y_t : t = 0, 1, ...\}$ *tells the* permanent *effect of W_{t+1} on current and future values of the* level *of Y, i.e., the effect on* $\lim_{j \to +\infty} Y_{t+j}$. *Models of Blanchard and Quah (1989) and Shapiro and Watson (1988) build on this property. The variance of the random variable $\alpha(1) \cdot W_{t+1}$ conditioned on the invariant events in $\mathfrak{I}$ is $|\alpha(1)|^2$. The overall variance of X_t is given by*

$$\sum_{j=0}^{\infty} |\alpha_j|^2 \neq |\alpha(1)|^2.$$

2.3 Central limit approximation

Example 2.2.3 starts from a moving average of martingale differences that is used as an increment $\{X_t\}$ to a $\{Y_t : t \geq 0\}$ process, after which it constructs a process of innovations to the martingale component of the $\{Y_t : t \geq 0\}$ process. That analysis illustrates the workings of an operator $\mathbb{D}$ that maps an admissible increment process in $\mathcal{H}$ into the innovation in a martingale component. To construct $\mathbb{D}$, let $\mathcal{G}$ be the set of all random variables G with finite second moments that satisfy the conditions that G is $\mathfrak{A}$ measurable and that $E(G_1|\mathfrak{A}) = 0$ where $G_1 = G \circ \mathbb{S}$. Define $\mathbb{D} : \mathcal{H} \to \mathcal{G}$ via

$$\mathbb{D}(X) = G.$$

Both $\mathcal{G}$ and $\mathcal{H}$ are linear spaces of random variables and $\mathbb{D}$ is a linear transformation. The operator $\mathbb{D}$ plays a prominent role in a central limit approximation.

To form a central limit approximation, construct the following scaled partial sum that nets out trend growth

$$\frac{1}{\sqrt{t}}(Y_t - \nu t) = \frac{1}{\sqrt{t}} Y_t^m - \frac{1}{\sqrt{t}} H_t^+ + \frac{1}{\sqrt{t}}(H_0^+ + Y_0)$$

where

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

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

$$\frac{1}{\sqrt{t}} Y_t^m \Rightarrow \mathcal{N}\left(0, E\left[\mathbb{D}(X)^2 | \mathfrak{I}\right]\right)$$

where $\Rightarrow$ denotes weak convergence, meaning convergence in distribution. Clearly, $\{(1/\sqrt{t})H_t^+\}$ and $\{(1/\sqrt{t})(H_0^+ + Y_0)\}$ both converge in mean square to zero. Thus,

Proposition 2.3.1. *For all stationary increment processes* $\{Y_t : t = 0, 1, 2, ...\}$ *represented by* X *in* $\mathcal{H}$

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

Furthermore,

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

2.4 Cointegration

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

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

The Proposition 2.2.2 martingale component of $\{Y_t : t \geq 0\}$ 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 2.2.2 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 $\{Y_t^2 : t = 0, 1, \ldots\}$.

Proposition 2.2.2 sheds light on the cointegration concept of Engle and Granger (1987) that is associated with 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.[4] That situation prevails when there exist real valued scalars $\mathbf{r}_1$ and $\mathbf{r}_2$ such that

$$\begin{aligned} \mathbf{r}_1 \nu_1 + \mathbf{r}_2 \nu_2 &= 0 \\ \mathbf{r}_1 \mathbb{D}(X^1) + \mathbf{r}_2 \mathbb{D}(X^2) &= 0, \end{aligned}$$

where the ν's correspond to the trend components in Proposition 2.2.2. These two zero restrictions imply that the time trend and the martingale component for the linear combination Y_t are both zero.[5] 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.

This notion of cointegration provides one way to formalize balanced growth paths in stochastic environments through determining linear combination of growing times series for which stochastic growth is absent.

[4] The Box and Tiao (1977) "canonical correlation" approach to linear time series analysis anticipated, at least partially, the co-integration restrictions of time series econometricians and macroeconomists.

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

Chapter 3

Markov Processes

We call a random vector X_t the *state* because it completely describes the position of a dynamic system at time t from the perspective of a model builder or an econometrician. We construct a consistent sequence of probability distributions Pr_ℓ for a sequence of random vectors

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

for all nonnegative integers ℓ by specifying the following two elementary components of a *Markov process*: (i) a probability distribution for X_0, and (ii) a time-invariant distribution for X_{t+1} conditional on X_t for $t \geqslant 0$. All other probabilities are functions of these two distributions. By creatively defining the state vector X_t, a Markov specification includes many models used in applied research.

3.1 Constituents

Assume a state space $\mathcal{X}$ and a transition distribution $P(dx^*|x)$. For example, $\mathcal{X}$ could be $\mathbb{R}^n$ or a subset of $\mathbb{R}^n$. The transition distribution P is a conditional probability measure for each $X_t = x$ in the state space, so it satisfies $\int_{\{x^* \in \mathcal{X}\}} P(dx^*|x) = 1$ for every x in the state space. If in addition we specify a marginal distribution Q_0 for the initial state x_0 over $\mathcal{X}$, then we have completely specified all joint distributions for the stochastic process $\{X_t, t = 0, 1, \ldots\}$.

The notation $P(dx^*|x)$ denotes a conditional probability measure; integration is over x^* and conditioning is captured by x. Thus, x^* is a possible realization of next period's state and x is a realization of this period's state. The conditional probability measure $P(dx^*|x)$ assigns conditional probabilities to next period's state given that this period's state is x. Often, but not always, the conditional distributions have densities against a common distribution $\lambda(dx^*)$ to be used to integrate over states. That lets us use a *transition density* to represent the conditional probability measure.

Example 3.1.1. *A first-order vector autoregression is a Markov process. Here $Q_0(x)$ is a normal distribution with mean μ_0 and covariance matrix Σ_0 and $P(dx^*|x)$ is a normal distribution with mean Ax and covariance matrix BB' for a square matrix A and a matrix B with full column rank.*[1] *These assumptions imply the vector autoregressive (VAR) representation*

$$X_{t+1} = AX_t + BW_{t+1},$$

for $t \geqslant 0$, where W_{t+1} is a multivariate standard normally distributed random vector that is independent of X_t.

Example 3.1.2. *A discrete-state Markov chain consists of a Q_0 represented as a row vector and a transition probability $P(dx^*|x)$ represented as a matrix with one row and one column for each possible value of the state x. Rows contain vectors of probabilities of next period's state conditioned on a realized value of this period's state.*

It is useful to construct an operator by applying a one-step conditional expectation operator to functions of a Markov state. Let $f : \mathcal{X} \to \mathbb{R}$. For bounded f, define:

$$\mathbb{T}f(x) = E\left[f(X_{t+1})|X_t = x\right] = \int_{\{x^* \in \mathcal{X}\}} f(x^*)P(dx^*|x). \tag{3.1}$$

The Law of Iterated Expectations justifies iterating on $\mathbb{T}$ to form conditional expectations of the function f of the Markov state over longer horizons:

$$\mathbb{T}^j f(x) = E\left[f(X_{t+j})|X_t = x\right].$$

[1] When BB' is singular, a density may not exist with respect to Lebesgue measure. The covariance matrix BB' is typically singular for a first-order vector autoregression constructed by rewriting a higher-order vector autoregression.

We can use the operator $\mathbb{T}$ to characterize a Markov process. Indeed, by applying $\mathbb{T}$ to a suitable range of test functions f, we can construct a conditional probability measure.

Fact 3.1.3. *Start with a conditional expectation operator $\mathbb{T}$ that maps a space of bounded functions into itself. We can use $\mathbb{T}$ to construct a conditional probability measure $P(dx^*|x)$ provided that $\mathbb{T}$ is (a) well defined on the space of bounded functions, (b) preserves the bound, (c) maps nonnegative functions into nonnegative functions, and d) maps the unit function into the unit function.*

3.2 Stationarity

We can construct a stationary Markov process by carefully choosing the distribution of the initial state X_0.

Definition 3.2.1. *A probability measure Q over a state space $\mathcal{X}$ for a Markov process with transition probability P is a **stationary distribution** if it satisfies*

$$\int_{\{x \in \mathcal{X}\}} P(dx^*|x)Q(dx) = Q(dx^*).$$

We will sometimes refer to a stationary density q. A density is always relative to a measure. With this in mind, let λ be a measure used to integrate over possible Markov states on the state space $\mathcal{X}$. Then a density q is a nonnegative (Borel measurable) function of the state for which $\int q(x)\lambda(dx) = 1$.

Definition 3.2.2. *A **stationary density** over a state space $\mathcal{X}$ for a Markov process with transition probability P is a probability density q with respect to a measure λ over the state space $\mathcal{X}$ that satisfies*

$$\int P(dx^*|x)q(x)\lambda(dx) = q(x^*)\lambda(dx^*).$$

Various sufficient conditions imply the existence of a stationary distribution. Given a transition distribution P, one such condition that is widely used to justify some calculations from numerical simulations is that the Markov process be *time reversible*, which means that

$$P(dx^*|x)Q(dx) = P(dx|x^*)Q(dx^*) \tag{3.2}$$

for some probability distribution Q on $\mathcal{X}$. Because a transition distribution satisfies $\int_{\{x \in \mathcal{X}\}} P(dx|x^*) = 1$,

$$\int_{\{x \in \mathcal{X}\}} P(dx^*|x)Q(dx) = \int_{\{x \in \mathcal{X}\}} P(dx|x^*)Q(dx^*) = Q(dx^*),$$

so Q is a stationary distribution by Definition 3.2.1. Restriction (3.2) implies that the process is time reversible in the sense that forward and backward transition distributions coincide. Time reversibility is special, so later we will explore other sufficient conditions for the existence of stationary distributions.[2]

Remark 3.2.3. *When a Markov process starts at a stationary distribution, we can construct the process $\{X_t : t = 1, 2, ...\}$ with a measure-preserving transformation $\mathbb{S}$ of the type featured in chapter 1, section 1.3.*

3.3 $\mathcal{L}^2$ and Eigenfunctions

We connected ergodicity to a statistical notion of invariance in chapter 1. The word invariance brings to mind a generalization of eigenvectors called eigenfunctions. Eigenfunctions of a linear mapping characterize an invariant subspace of functions such that the application of a linear mapping to any element of that space remains in the same subspace. Eigenfunctions associated with a unit eigenvalue are themselves invariant under the mapping. So perhaps it is not surprising that such eigenfunctions of $\mathbb{T}$ come in handy for studying ergodicity of Markov processes.

Given a stationary distribution Q, form the space of functions

$$\mathcal{L}^2 = \{f : \mathcal{X} \to \mathbb{R} : \int f(x)^2 Q(dx) < \infty\}.$$

It can be verified that $\mathbb{T} : \mathcal{L}^2 \to \mathcal{L}^2$ and that

$$\|f\| = \left[\int f(x)^2 Q(dx)\right]^{1/2}$$

is a well defined norm on $\mathcal{L}^2$.

We now study eigenfunctions of the conditional expectation operator $\mathbb{T}$.

[2] Numerical Bayesian statistical analysis often computes a posterior probability distribution by iterating to convergence a reversible Markov process whose stationary distribution is that posterior distribution.

Definition 3.3.1. *A function $\tilde{f} \in \mathcal{L}^2$ that solves $\mathbb{T}f = f$ is an eigenfunction of $\mathbb{T}$ associated with a unit eigenvalue.*

The following proposition asserts that an eigenfunction $\tilde{f}(X_t)$ associated with a unit eigenvalue is constant as X_t moves through time.

Proposition 3.3.2. *Suppose that $\tilde{f}$ is an eigenfunction of $\mathbb{T}$ associated with a unit eigenvalue. Then $\{\tilde{f}(X_t) : t = 0, 1, ...\}$ is constant over time with probability one.*

Proof.

$$E\left[\tilde{f}(X_{t+1})\tilde{f}(X_t)\right] = \int (\mathbb{T}\tilde{f})(x)\tilde{f}(x)Q(dx) = \int \tilde{f}(x)^2 Q(dx) = E\left[\tilde{f}(X_t)^2\right]$$

where the first equality follows from the Law of Iterated Expectations. Then because Q is a stationary distribution,

$$\begin{aligned} E\left([\tilde{f}(X_{t+1}) - \tilde{f}(X_t)]^2\right) &= E\left[\tilde{f}(X_{t+1})^2\right] + E\left[\tilde{f}(X_t)^2\right] \\ &\quad -2E\left[\tilde{f}(X_{t+1})\tilde{f}(X_t)\right] \\ &= 0. \end{aligned}$$

□

3.4 Ergodic Markov Processes

Chapter 1 studied special statistical models that, because they are ergodic, are affiliated with a Law of Large Numbers in which limit points are constant across sample points $\omega \in \Omega$. Section 1.8 described other statistical models that are not ergodic and that are components of more general probability specifications that we used to express the idea that a statistical model is unknown.[3] We now explore ergodicity in the context of Markov processes.

From Proposition 3.3.2 we know that time-series averages of an eigenfunction $\mathbb{T}\tilde{f} = \tilde{f}$ are invariant over time, so

$$\frac{1}{N}\sum_{t=1}^{N} \tilde{f}(X_t) = \tilde{f}(X).$$

[3] Unknown parameters manifest themselves as unknown statistical models.

However, when $\tilde{f}(x)$ varies across sets of states x that occur with positive probability under Q, a time series average $\frac{1}{N}\sum_{t=1}^{N}\tilde{f}(X_t)$ can differ from $\int \tilde{f}(x)Q(dx)$. This happens when observations of $\tilde{f}(X_t)$ along a sample path for $\{X_t\}$ convey an inaccurate impression of how $f(X)$ varies across the stationary distribution $Q(dx)$. See Example 3.6.4 below. We can exclude the possibility of such inaccurate impressions by imposing a restriction on the eigenfunction equation $\mathbb{T}f = f$.

Proposition 3.4.1. *When a unique solution to the equation*

$$\mathbb{T}f = f$$

is a constant function (with Q measure one), then it is possible to construct $\{X_t : t = 0, 1, 2, ...\}$ as a stationary and ergodic Markov process with $\mathbb{T}$ as the one-period conditional expectation operator and Q as the initial distribution for X_0.[4]

Evidently, ergodicity is a property that obtains relative to a stationary distribution Q of the Markov process. If there are multiple stationary distributions, it is possible that there is a unique constant function f that solves $\mathbb{T}f = f$ problem for one stationary distribution and that non-constant solutions exist for other stationary distributions.

Invariant events for a Markov process

Consider an eigenfunction $\tilde{f}$ of $\mathbb{T}$ associated with a unit eigenvalue. Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a bounded Borel measurable function. Since $\{\tilde{f}(X_t) : t = 0, 1, 2, ...\}$ is invariant over time, so is $\left\{\varphi\left[\tilde{f}(X_t)\right] : t = 0, 1, 2, \ldots\right\}$ and it is necessarily true that

$$\mathbb{T}(\varphi \circ \tilde{f}) = \varphi \circ \tilde{f}.$$

[4] In particular, the process can be represented using a probability measure Pr defined over events in $\mathfrak{F}$, a transformation $\mathbb{S}$ for which$(\mathbb{S}, Pr)$ is measure preserving, and ergodic and a measurement function $\widetilde{X}$ such that $\left\{\widetilde{X} \circ \mathbb{S}^t : t = 0, 1, \ldots\right\}$ has the same induced distribution as the process $\{X_t : t = 0, 1, 2, \ldots\}$.

Therefore, from an eigenfunction $\tilde{f}$ associated with a unit eigenvalue, we can construct other eigenfunctions,[5] for example

$$\varphi[\tilde{f}(x)] = \begin{cases} 1 & \text{if } \tilde{f}(x) \in \tilde{\mathfrak{b}} \\ 0 & \text{if } \tilde{f}(x) \notin \tilde{\mathfrak{b}} \end{cases} \tag{3.3}$$

for some Borel set $\tilde{\mathfrak{b}}$ in $\mathbb{R}$. It follows that

$$\Lambda = \{\omega \in \Omega : \tilde{f}[X_0(\omega)] \in \tilde{\mathfrak{b}}\}$$

is an invariant event in Ω. Note that by constructing the Borel set, $\mathfrak{b}$ in $\mathcal{X}$

$$\mathfrak{b} = \left\{ x : \tilde{f}(x) \in \tilde{\mathfrak{b}} \right\}$$

we can represent Λ as

$$\Lambda = \{\omega \in \Omega : X_0(\omega) \in \mathfrak{b}\}. \tag{3.4}$$

Thus we have shown how to construct many non-degenerate eigenfunctions, starting from an initial such function.

For Markov processes, all invariant events can be represented as in (3.4), which is expressed in terms of the initial state X_0. See Doob (1953, p. 460, Theorem 1.1). Thus, associated with an invariant event is a Borel set in $\mathcal{X}$. Let $\mathfrak{J}$ denote the collection of Borel subsets of $\mathcal{X}$ for which Λ constructed as in (3.4) is an invariant event. From these invariant events, we can also construct many non-degenerate eigenfunctions as indicator functions of sets in $\mathfrak{J}$. Formally, if $\tilde{\mathfrak{b}} \in \mathfrak{J}$, then the indicator function

$$f(x) = \begin{cases} 1 & \text{if } x \in \mathfrak{b} \\ 0 & \text{if } x \notin \mathfrak{b} \end{cases} \tag{3.5}$$

satisfies

$$\mathbb{T}f = f$$

with Q probability one. Provided that the probability of Λ is neither zero nor one, then we have constructed a nonnegative function f that is strictly positive on a set of positive Q measure and zero on a set with strictly positive Q measure.

[5] This construction also works for unbounded functions φ provided that $\varphi \circ \tilde{f}$ is square integrable under the Q measure.

More generally, when a Markov process $\{X_t : t \geq 0\}$ is not ergodic, there exist bounded eigenfunctions with unit eigenvalues that are not constant with Q measure one. For a non-degenerate eigenfunction $\tilde{f}$ with unit eigenvalue to be constant with Q measure one, it shouldn't be possible for the Markov process permanently to get stuck in a subset of the state space which has probability different from one or zero. Suppose now we consider any Borel set $\mathfrak{b}$ of $\mathcal{X}$ that has Q measure that is neither zero nor one. Let f be constructed as in (3.5) without restricting $\mathfrak{b}$ to be in $\mathfrak{J}$. Then $\mathbb{T}^j$ applied to f is the conditional probability of $\{X_j \in \mathfrak{b}\}$ as of date zero. If we want time series averages to converge to unconditional expectations, we must require that the set $\mathfrak{b}$ be visited eventually with positive probability. To account properly for all possible future dates we use a mathematically convenient *resolvent operator* defined by

$$\mathbb{M}f(x) = (1 - \lambda) \sum_{j=0}^{\infty} \lambda^j \mathbb{T}^j f.$$

for some constant discount factor $0 < \lambda < 1$. Notice that If $\tilde{f}$ is an eigenfunction of $\mathbb{T}$ associated with a unit eigenvalue, then the same is true for $\mathbb{T}^j$ and hence for $\mathbb{M}$. We translate the requirement that X_j be eventually visited to a restriction that applying $\mathbb{M}$ the indicator function f yields a strictly positive function. The following statement extends this restriction to all nonnegative functions that are distinct from zero.

Proposition 3.4.2. *Suppose that for any $f \geq 0$ such that $\int f(x)Q(dx) > 0$, $\mathbb{M}f(x) > 0$ for all $x \in \mathcal{X}$ with Q measure one. Then any solution $\tilde{f}$ to $\mathbb{T}f = f$ is necessarily constant with Q measure one.*

Proof. Consider an eigenfunction $\tilde{f}$ associated with a unit eigenvalue. The function $f = \varphi \circ \tilde{f}$ necessarily satisfies:

$$\mathbb{M}f = f$$

for any φ of the form (3.3). If such an f also satisfies $\int f(x)Q(dx) > 0$, then $f(x) = 1$ with Q probability one. Since this holds for any Borel set $\mathfrak{b}$ in $\mathbb{R}$, $\tilde{f}$ must be constant with Q probability one. □

Proposition 3.4.2 supplies a sufficient condition for ergodicity. A more restrictive sufficient condition is that there exists an integer $m \geqslant 1$ such that

$$\mathbb{T}^m f(x) > 0$$

for any $f \geq 0$ such that $\int f(x)Q(dx) > 0$ on a set with Q measure one.

Remark 3.4.3. *The sufficient conditions imposed in Proposition 3.4.2 imply a property called* irreducibility *relative to the probability measure Q. While this proposition presumes that Q is a stationary distribution,* irreducibility *allows for a more general specification of Q.*

Proposition 3.4.2 provides a way to verify ergodicity. As discussed in Chapter 1, ergodicity is a property of a statistical model. As statisticians or econometricians we often entertain a set of Markov models, each of which is ergodic. For each model we can build a probability Pr using the canonical construction given at the outset of Chapter 1. Convex combinations of these probabilities are measure-preserving but not necessarily ergodic when used in conjunction with the shift transformation $\mathbb{S}$. We can take the ergodic Markov models to be the building blocks for a specification to to be used in a statistical investigation. There can be a finite number of these building blocks or even a continuum of them represented in terms of an unknown parameter vector.

3.5 Periodicity

Next we study a notion of periodicity of a stationary and ergodic Markov process.[6] To define periodicity of a Markov process, for a given positive integer p we construct a new Markov process by sampling an original process every p time periods. This is sometimes called 'skip-sampling' at sampling interval p.[7] With a view toward applying Proposition 3.3.2 to $\mathbb{T}^p$, solve

$$\mathbb{T}^p f = f \tag{3.6}$$

for a function $\tilde{f}$. We know from Proposition 3.3.2 that for an $\tilde{f}$ that solves (3.6), $\{\tilde{f}(X_t) : t = 0, p, 2p, \ldots\}$ is invariant and so is $\{\tilde{f}(X_t) : t = 1, p + 1, 2p + 1, ...\}$. The process $\tilde{f}(X_t)$ is periodic with period p or np for any positive integer n.

[6] Our definition of periodicity is confines to a stationary distribution. Actually, periodicity can be defined more generally. We limit our treatment of periodicity to specifications of transition probabilities for which there exist stationary distributions for convenience here.

[7] See appendix **??** of chapter 1, Hansen and Sargent (1993) and Hansen and Sargent (2013, ch. 14).

Definition 3.5.1. *The* periodicity *of an irreducible Markov process* $\{X_t\}$ *with respect to* $\widetilde{Q}$ *is the smallest positive integer* p *such that there is a solution to equation* (3.6) *that is not constant with* $\widetilde{Q}$ *measure one. When there is no such integer* p*, we say that the process is* aperiodic.

Result 3.5.2. *Consider a counterpart of the resolvent operator* $\mathbb{M}$ *constructed by sampling at interval given by positive integer* p*:*

$$\mathbb{M}_p f(x) = (1-\lambda) \sum_{j=0}^{\infty} \lambda^j \mathbb{T}^{pj} f. \tag{3.7}$$

Provided that $\mathbb{M}_p f(x) > 0$ *with* Q *measure one and all* $p \geq 0$ *for any* $f \geq 0$ *such that* $\int f(x)Q(dx) > 0$*, the Markov process is aperiodic.*

3.6 Finite-State Markov Chains

Suppose that $\mathcal{X}$ consists of n possible states. We can label these states in a variety of ways, but for now we suppose that state x_j is the coordinate vector consisting entirely of zeros except in position j, where there is a one. Let $\mathbb{P}$ be an n by n transition matrix, where entry i, j is the probability of moving from state i to state j in a single period. Thus, the entries of $\mathbb{P}$ are all nonnegative and

$$\mathbb{P}\mathbf{1}_n = \mathbf{1}_n,$$

where $\mathbf{1}_n$ is an n-dimensional vector of ones.

Let $\mathbf{q}$ be an n-dimensional vector of probabilities. Stationarity requires that

$$\mathbf{q}'\mathbb{P} = \mathbf{q}', \tag{3.8}$$

where $\mathbf{q}$ is a row eigenvector (also called a left eigenvector) of $\mathbb{P}$ associated with a unit eigenvalue.

We use a vector $\mathbf{f}$ to represent a function from the state space to the real line. Each coordinate of $\mathbf{f}$ gives the value of the function at the corresponding coordinate vector. Then the conditional expectation operator $\mathbb{T}$ can be represented in terms of the transition matrix $\mathbb{P}$:

$$E\left(\mathbf{f} \cdot X_{t+1} | X_t = x\right) = (\mathbb{T}\mathbf{f}) \cdot x = x'\mathbb{P}\mathbf{f}.$$

Now consider column eigenvectors called “right eigenvectors” of $\mathbb{P}$ that are associated with a unit eigenvalue.

Proposition 3.6.1. *Suppose that the only solutions to*

$$\mathbb{T}\boldsymbol{f} = \boldsymbol{f}$$

are of the form $\boldsymbol{f} \propto \boldsymbol{1}_n$*, where* $\propto$ *means 'proportional to'. Then we can construct a process that is stationary and ergodic by initializing the process with density* $\boldsymbol{q}$ *determined by equation* (3.8)*.*

We can weaken the Proposition 3.6.1 sufficient condition for stationarity and ergodicity to allow nonconstant right eigenvectors. This weakening is of interest when there are multiple stationary distributions.

Proposition 3.6.2. *Assume that there exists a real number* $\boldsymbol{r}$ *such that the right eigenvector* $\boldsymbol{f}$ *and a stationary distribution* $\boldsymbol{q}$ *satisfy*

$$\min_{\mathbf{r}} \sum_{i=1}^{n} (\mathbf{f}_i - \mathbf{r})^2 \mathbf{q}_i = 0.$$

Then the process is stationary and ergodic.

Notice that if $\mathbf{q}_i$ is zero, the contribution of $\mathbf{f}_i$ to the least squares objective can be neglected. This allows for non-constant $\mathbf{f}$'s, albeit in a limited way.

Three examples illustrate ideas in these propositions.

Example 3.6.3. *Recast Example 1.4.3 as a Markov chain with transition matrix* $\mathbb{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$*. This chain has a unique stationary distribution* $q = \begin{bmatrix} .5 & .5 \end{bmatrix}'$ *and the invariant functions are* $\begin{bmatrix} \mathbf{r} & \mathbf{r} \end{bmatrix}'$ *for any scalar* $\mathbf{r}$*. Therefore, the process initiated from the stationary distribution is ergodic. The process is periodic with period two since the matrix* $\mathbb{P}^2$ *is an identity matrix and all two dimensional vectors are eigenvectors associated with a unit eigenvalue.*

Example 3.6.4. *Recast Example 1.4.4 as a Markov chain with transition matrix* $\mathbb{P} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ *. This chain has a continuum of stationary distributions* $\pi \begin{bmatrix} 1 \\ 0 \end{bmatrix} + (1-\pi) \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ *for any* $\pi \in [0,1]$ *and invariant functions* $\begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix}$ *for any scalars* $\mathbf{r}_1, \mathbf{r}_2$*. Therefore, when* $\pi \in (0,1)$ *the process is not ergodic because if* $\mathbf{r}_1 \neq \mathbf{r}_2$ *the resulting invariant function fails to be constant across states* *that have positive probability under the stationary distribution associated with* $\pi \in (0,1)$. *When* $\pi \in (0,1)$, *nature chooses state* $i = 1$ *or* $i = 2$ *with probabilities* $\pi, 1-\pi$, *respectively, at time* 0. *Thereafter, the chain remains stuck in the realized time* 0 *state. Its failure ever to visit the unrealized state prevents the sample average from converging to the population mean of an arbitrary function of the state.*

Example 3.6.5. *A Markov chain with transition matrix* $\mathbb{P} = \begin{bmatrix} .8 & .2 & 0 \\ .1 & .9 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ *has a continuum of stationary distributions* $\pi \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & 0 \end{bmatrix}' + (1-\pi) \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}'$ *for* $\pi \in [0,1]$ *and invariant functions* $\begin{bmatrix} \mathbf{r}_1 & \mathbf{r}_1 & \mathbf{r}_2 \end{bmatrix}'$ *for any scalars* $\mathbf{r}_1, \mathbf{r}_2$. *Under any stationary distribution associated with* $\pi \in (0,1)$, *the chain is not ergodic because some invariant functions are not constant with probability one. But under stationary distributions associated with* $\pi = 1$ *or* $\pi = 0$, *the chain is ergodic.*

3.7 Limited Dependence

Recall the conditional expectations operator $\mathbb{T}$ defined in equation (3.1) for a space $\mathcal{L}^2$ of functions f of a Markov process with transition probability P and stationary distribution Q and for which $f(X_t)$ has a finite second moment under Q:

$$\mathbb{T}f(x) = E\left[f(X_{t+1}) \mid X_t = x\right] = \int_{\{x^* \in \mathcal{X}\}} f(x^*) P(dx^*|x).$$

We suppose that under the stationary distribution Q, the process is ergodic.

Because it is often useful to work with random variables that have been 'centered' by substracting out their means, we define the following subspace of $\mathcal{L}^2$:

$$\mathcal{N} = \left\{ f \in \mathcal{L}^2 : \int f(x) Q(dx) = 0 \right\}. \tag{3.9}$$

We use the same norm $\|f\| = \left[\int f(x)^2 Q(dx)\right]^{1/2}$ on both $\mathcal{L}^2$ and $\mathcal{N}$ too.

Definition 3.7.1. *The conditional expectation operator* $\mathbb{T}$ *is said to be a* strong contraction *on* $\mathcal{N}$ *if there exists* $0 < \rho < 1$ *such that*

$$\|\mathbb{T}f\| \leq \rho \|f\|$$

for all $f \in \mathcal{N}$.

When $\mathbb{T}^m$ is a strong contraction for some positive integer m and some $\rho \in (0,1)$, the Markov process is said to be ρ-mixing conditioned on the invariant events.

Remark 3.7.2. $\mathbb{T}$ *being a strong contraction on* $\mathcal{N}$ *limits intertemporal dependence of the Markov process* $\{X_t\}$.

Let $\mathbb{I}$ be the identity operator. When the conditional expectation operator $\mathbb{T}$ is a strong contraction, the operator $(\mathbb{I}-\mathbb{T})^{-1}$ is well defined, bounded on $\mathcal{N}$, and equal to the geometric sum:[8]

$$(\mathbb{I}-\mathbb{T})^{-1} f(x) = \sum_{j=0}^{\infty} \mathbb{T}^j f(x) = \sum_{j=0}^{\infty} E\left[f(X_{t+j}) | X_t = x\right].$$

Example 3.7.3. *Consider the Markov chain setting of section 3.6 with a transition matrix* $\mathbb{P}$. *A stationary density* $\boldsymbol{q}$ *is a nonnegative vector that satisfies*

$$\boldsymbol{q}'\mathbb{P} = \boldsymbol{q}'$$

and $\boldsymbol{q} \cdot \boldsymbol{1}_n = 1$. *If the only column eigenvector of* $\mathbb{T}$ *associated with a unit eigenvalue is constant over states* i *for which* $\mathbf{q}_i > 0$, *then the process is ergodic. If in addition the only eigenvector of* $\mathbb{P}$ *that is associated with an eigenvalue that has a unit norm (the unit eigenvalue might be complex) is constant over states* i *for which* $\mathbf{q}_i > 0$, *then* $\mathbb{T}^m$ *is a strong contraction for some integer* $m \geqslant 1$.[9] *This implies that the process is ergodic. It also rules out the presence of periodic components that can be forecast perfectly.*

[8] The geometric series after the first equality sign is well defined under the weaker restriction that $\mathbb{T}^m$ is a strong contraction for some integer $m \geqslant 1$.

[9] This follows from Gelfand's Theorem, which asserts the following. Let $\mathcal{N}$ be the $n-1$ dimensional space of vectors that are orthogonal to $\mathbf{q}$. $\mathbb{T}$ maps $\mathcal{N}$ into itself. The spectral radius of $\mathbb{T}$ restricted to $\mathcal{N}$ is the maximum of the absolute values of the eigenvalues. Gelfand's Theorem asserts that the spectral radius governs the behavior as m gets large of the decay factor of the $\mathbb{T}$ transformation applied m times. Provided that the spectral radius is less than one, the strong contraction property prevails for any $\rho < 1$ that is larger than the spectral radius.

3.8 Limits of Multi-Period Forecasts

When a Markov process is aperiodic, there are interesting situations in which

$$\lim_{j\to\infty} \mathbb{T}^j f(x) = \mathbf{r} \tag{3.10}$$

for some $\mathbf{r} \in \mathbb{R}$, where convergence is either pointwise in x or in the $\mathcal{L}^2$ norm. Limit (3.10) asserts that long-run forecasts do not depend on the current Markov state. (Meyn and Tweedie (1993) provide a comprehensive treatment of such convergence.) Let Q be a stationary distribution. Then it is necessarily true that

$$\int \mathbb{T}^j f(x) Q(dx) = \int f(x) Q(dx)$$

for all j. Thus,

$$\mathbf{r} = \int f(x) Q(dx),$$

so that the limiting forecast is necessarily the mathematical expectation of $f(x)$ under a stationary distribution. Here we have assumed that the limit point is a number and not a random variable; we have not assumed that the stationary distribution is unique.

Notice that if (3.10) is satisfied, then any function f that satisfies

$$\mathbb{T} f = f$$

is necessarily constant with probability one. Also, if $\int f(x) Q(dx) = 0$ and convergence is sufficiently fast, then

$$\lim_{N\to\infty} \sum_{j=0}^{N} \mathbb{T}^j f(x) \tag{3.11}$$

is a well-defined function of the Markov state. We shall construct the limit in (3.11) when we extract martingales from additive functionals in chapter 4.

A set of sufficient conditions for the convergence outcome

$$\lim_{j\to\infty} \mathbb{T}^j f(x^*) \to \int f(x) Q(dx) \tag{3.12}$$

for each $x^* \in \mathcal{X}$ and each bounded f is:[10]

Condition 3.8.1. *A Markov process with stationary distribution Q satisfies:*

(i) For any $f \geq 0$ such that $\int f(x)Q(dx) > 0$, $\mathbb{M}_p f(x) > 0$ for all $x \in \mathcal{X}$ with Q measure one and all positive integers $p \geq 0$, where the operator $\mathbb{M}_p$ is defined in (3.7)*.*

(ii) $\mathbb{T}$ maps bounded continuous functions into bounded continuous functions, i.e., the Markov process is said to satisfy the Feller property.

(iii) The support of Q has a nonempty interior in $\mathcal{X}$.

(iv) $\mathbb{T}V(x) - V(x) \leq -1$ outside a compact subset of $\mathcal{X}$ for some nonnegative function V.

We encountered condition (i) in our section 3.4 discussion of Markov processes that are ergodic and aperiodic. Condition *(*iv) is a *drift condition* for stability that requires that we find a function V that satisfies the requisite inequality. Heuristically, the drift condition says that outside a compact subset of the state space, application of the conditional expectation operator pushes the function inward. The choice of -1 as a comparison point is made only for convenience, since we can always multiply the function V by a number greater than one. Thus, -1 could be replaced by any strictly negative number. In section 3.9, we will apply condition 3.8.1 to verify ergodicity of a vector autoregression.

3.9 Vector Autoregressions

A square matrix $\mathbb{A}$ is said to be *stable* when all of its eigenvalues have absolute values that are strictly less than one. For a stable $\mathbb{A}$, suppose that

$$X_{t+1} = \mathbb{A}X_t + \mathbb{B}W_{t+1},$$

where $\{W_{t+1} : t = 1, 2, ...\}$ is an i.i.d. sequence of multivariate normally distributed random vectors with mean vector zero and covariance matrix I

[10] Restriction 3.12 is stronger than ergodicity. It rules out periodic processes, although we know that periodic processes can be ergodic.

and that $X_0 \sim \mathcal{N}(\mu_0, \Sigma_0)$. This specification constitutes a first-order *vector autoregression.*

Let $\mu_t = EX_t$. Notice that

$$\mu_{t+1} = \mathbb{A}\mu_t.$$

The mean μ of a stationary distribution satisfies

$$\mu = \mathbb{A}\mu.$$

Because we have assumed that $\mathbb{A}$ is a stable matrix, $\mu = 0$ is the only solution of $(\mathbb{A} - \mathbb{I})\mu = 0$, so the mean of the stationary distribution is $\mu = 0$.

Let $\Sigma_t = E(X_t - \mu_t)(X_t - \mu_t)'$ be the covariance matrix of X_t. Then

$$\Sigma_{t+1} = \mathbb{A}\Sigma_t\mathbb{A}' + \mathbb{B}\mathbb{B}'.$$

For $\Sigma_t = \Sigma$ to be invariant over time, it must satisfy the discrete Lyapunov equation

$$\Sigma = \mathbb{A}\Sigma\mathbb{A}' + \mathbb{B}\mathbb{B}'. \tag{3.13}$$

When $\mathbb{A}$ is a stable matrix, this equation has a unique solution for a positive semidefinite matrix Σ.

Suppose that $\Sigma_0 = 0$ (a matrix of zeros) and for $t \geqslant 1$ define the matrix

$$\Sigma_t = \sum_{j=0}^{t-1} \mathbb{A}^j\mathbb{B}\mathbb{B}'(\mathbb{A}^j)'.$$

The limit of the sequence $\{\Sigma_t\}_{t=0}^{\infty}$ is

$$\Sigma = \sum_{j=0}^{\infty} \mathbb{A}^j\mathbb{B}\mathbb{B}'(\mathbb{A}^j)',$$

which can be verified to satisfy Lyapunov equation (3.13). Thus, Σ equals the covariance matrix of the stationary distribution.[11] Similarly, for all $\mu_0 = EX_0$

$$\mu_t = \mathbb{A}^t\mu_0,$$

[11] To verify the asserted equality, notice that $\sum_{j=0}^{\infty} \mathbb{A}^j\mathbb{B}\mathbb{B}'\mathbb{A}^{j\prime} = \mathbb{A}(\sum_{j=0}^{\infty} \mathbb{A}^j\mathbb{B}\mathbb{B}'\mathbb{A}^{j\prime})\mathbb{A}' + \mathbb{B}\mathbb{B}'$.

converges to zero, the mean of the stationary distribution.

The linear structure implies that the stationary distribution is Gaussian with mean μ and covariance matrix Σ. To verify ergodicity, we suppose that the covariance matrix Σ of the stationary distribution has full rank and verify conditions 3.8.1. Restriction (*iii*) of Condition 3.8.1 is satisfied. Furthermore, Σ_t has full rank for some t, which guarantees that the process is irreducible and aperiodic so that restriction (*i*) is satisfied. As a candidate for $V(x)$ in condition (*iv*), take $V(x) = |x|^2$. Then

$$\mathbb{T}V(x) = x'\mathbb{A}'\mathbb{A}x + \text{trace}(\mathbb{B}'\mathbb{B})$$

so

$$\mathbb{T}V(x) - V(x) = x'(\mathbb{A}'\mathbb{A} - \mathbb{I})x + \text{trace}(\mathbb{B}'\mathbb{B}).$$

That $\mathbb{A}$ is a stable matrix implies that $\mathbb{A}'\mathbb{A} - \mathbb{I}$ is negative definite, so that drift restriction (*iv*) of Condition 3.8.1 is satisfied for $|x|$ sufficiently large.[12] Thus, having verified conditions 3.8.1, we have verified the ergodicity of the VAR.

We can extend this example to allow the mean of the stationary distribution not to be zero. Partition the Markov state as

$$x = \begin{bmatrix} x_1 \\ x_1 \end{bmatrix}$$

where $x^{[2]}$ is a scalar. Similarly, partition the matrices $\mathbb{A}$ and $\mathbb{B}$ as

$$\mathbb{A} = \begin{bmatrix} [\mathbb{A}_{11} & \mathbb{A}_{12} \\ 0 & 1 \end{bmatrix}$$

$$\mathbb{B} = \begin{bmatrix} \mathbb{B}_1 \\ 0 \end{bmatrix}$$

where A_{11} is a stable matrix. Notice that the dynamics imply

$$X_{t+1}^2 = X_t^2 = \cdots = X_0^2$$

and hence is invariant. Let μ_2 denote the mean of X_t^2 for any t. For a stationary distribution we require that the mean μ_1 of $X_t 1$ satisfy

$$\mu_1 = \mathbb{A}_{11}\mu_1 + \mathbb{A}_{12}\mu_2.$$

[12] The Feller property (*ii*) of Condition 3.8.1 can also be verified.

Hence

$$\mu_1 = (I - \mathbb{A}_{11})^{-1} \mathbb{A}_{12}\mu_2.$$

Imitating our earlier argument, the covariance matrix, Σ_{11} of X_t^1 satisfies

$$\Sigma_{11} = \sum_{j=0}^{\infty} (\mathbb{A}_{11})^j \, \mathbb{B}_1(\mathbb{B}_1)' \, (\mathbb{A}_{11}{}')^j + (\mathbb{I} - \mathbb{A}_{11})^{-1} \mathbb{A}_{12}\Sigma_{22}\mathbb{A}_{12}{}' \, (\mathbb{I} - \mathbb{A}_{11}{}')^{-1}$$

where Σ_{22} is the variance of X_t^2 for all t. Stationarity imposes no restriction on the mean μ_2 and variance Σ_{22}.

Since $\{X_t^2 : t \geq 0\}$ is invariant, the process $\{X_t : t \geq 0\}$ is ergodic only when the variance Σ_{22} is zero. When $\{X_t : t \geq 0\}$ is not ergodic, the limit points in the Law of Large Numbers (Theorem 1.6.1) should be computed by conditioning on X_0^2.

3.10 Inventing a Past Again

In section 1.10, we invented an infinite past for a stochastic process. Here we invent an infinite past for a vector autoregression in a way that is equivalent to drawing an initial condition X_0 at time $t = 0$ from the stationary distribution $\mathcal{N}(0, \Sigma_\infty)$, where Σ_∞ solves the discrete Lyapunov equation (3.13), namely, $\Sigma_\infty = \mathbb{A}\Sigma_\infty\mathbb{A}' + \mathbb{B}\mathbb{B}'$.

Thus, consider the vector autoregression

$$X_{t+1} = \mathbb{A}X_t + \mathbb{B}W_{t+1}$$

where $\mathbb{A}$ is a stable matrix, $\{W_{t+1}\}_{t=-\infty}^{\infty}$ is now a two-sided infinite sequence of i.i.d. $\mathcal{N}(0, I)$ random vectors, and t is an integer. We can solve this difference equation backwards to get the moving average representation

$$X_t = \sum_{j=0}^{\infty} \mathbb{A}^j \mathbb{B} W_{t-j}.$$

Then

$$E\left[X_t \left(X_t\right)'\right] = \sum_{j=0}^{\infty} \mathbb{A}^j \mathbb{B}\mathbb{B}' \left(\mathbb{A}^j\right)' = \Sigma_\infty$$

where Σ_∞ is also the unique positive semidefinite matrix that solves $\Sigma_\infty = \mathbb{A}\Sigma_\infty\mathbb{A}' + \mathbb{B}\mathbb{B}'$.

Chapter 4

Processes with Markovian increments

In this chapter, we use a stationary Markov process to construct a process that displays stochastic arithmetic growth, then show how to extract a linear time trend and a martingale. Eventually, we will explore the implications exponentiating this process to transform an arithmetically growing processes like those described in this chapter to construct a process that displays geometric growth.

4.1 Definition of additive functional

Let $\{W_{t+1} : t \geq 0\}$ be a k-dimensional stochastic process of unanticipated economic shocks. Let $\{X_t : t \geq 0\}$ be a discrete-time stationary Markov process that is generated by initial distribution Q for X_0 and transition equation

$$X_{t+1} = \varphi(X_t, W_{t+1}), \tag{4.1}$$

where φ is a Borel measurable function. Let $\{\mathfrak{A}_t : t = 0, 1, ...\}$ be the filtration generated by histories of W and X; $\mathfrak{A}_t$ serves as the information set (sigma algebra) generated by $X_0, W_1, \ldots, W_t$. We presume that the conditional probability distribution for W_{t+1} conditioned on $\mathfrak{A}_t$ depends only on X_t. To assure that the process $\{W_{t+1} : t \geq 0\}$ represents unanticipated shocks, we restrict it to satisfy

$$E\left(W_{t+1} | X_t\right) = 0.$$

We condition on a statistical model in the sense of section 1.7 and assume that the stationary X_t process is ergodic.[1] The Markov structure of $\{X_t : t \geq 0\}$ makes the distribution of (X_{t+1}, W_{t+1}) conditioned on $\mathfrak{A}_t$ depend only on X_t.[2]

Definition 4.1.1. *A process $\{Y_t\}$ is said to be an* ***additive functional*** *if it can be represented as*

$$Y_{t+1} - Y_t = \kappa(X_t, W_{t+1}) \tag{4.2}$$

for a (Borel measurable) function $\kappa : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}$, or equivalently

$$Y_t = Y_0 + \sum_{j=1}^{t} \kappa(X_{j-1}, W_j),$$

where we initialize Y_0 at some arbitrary (Borel measurable) function of X_0.

When Y_0 is a function of X_0, we can construct Y_t as a function of the underlying Markov process between dates zero and t.

[1] If we wanted to include model uncertainty in the spirit of chapter 1, we could construct a set of statistical models like the one described here, each with its own parameter vector, and then form a weighted average over that set of models.

[2] Like $\{X_t\}$, the pair $\{(X_t, W_t)\}$ is a first-order Markov process restricted so that the joint transition distribution depends only on X_t.

Definition 4.1.2. *An additive functional $\{Y_t : t = 0, 1, ...\}$ is said to be an* ***additive martingale*** *if $E\left[\kappa(X_t, W_{t+1})|X_t\right] = 0$.*

Example 4.1.3. *(Stochastic Volatility) Suppose that*

$$\begin{aligned} Y_{t+1} - Y_t &= \mu(X_t) + \sigma(X_t) W_{t+1} \\ X_{t+1} &= \mathbb{A} X_t + \mathbb{B} W_{t+1} \end{aligned}$$

where $\{W_{t+1} : t \geq 0\}$ is an i.i.d. sequence of standardized multivariate normally distributed random vectors, $\mathbb{A}$ is a stable matrix, and $\mathbb{B}$ has full column rank, and the random vector X_0 is generated by initial distribution Q associated with the stationary distribution for the $\{X_t\}$ process. Here $\mu(X_t)$ is the conditional mean of $Y_{t+1} - Y_t$ and $|\sigma(X_t)|^2$ is its conditional variance. This is called a stochastic volatility model because $|\sigma(X_t)|^2$ is a stochastic process.

In example (4.1.3), when the conditional mean $\mu(X_t) = 0$, the process $\{Y_t\}$ is a martingale. Note that $E\left[\kappa(X_t, W_{t+1})|X_t\right] = 0$ implies the usual martingale restriction

$$E\left(Y_{t+1}|\mathfrak{A}_t\right) = Y_t, \quad \text{for} \quad t \geq 0.$$

4.2 Extracting Martingales

We can decompose an additive functional into a sum of components, one of which is an additive martingale that encapsulates all long-run stochastic variation as in Proposition 2.2.2. In this section, we show how to extract the martingale component. We adopt a construction like that used to establish Proposition 2.2.2 and proceed in four steps.

i) Construct the trend coefficient is the unconditional expectation:

$$\nu = E\left[\kappa(X_t, W_{t+1})\right].$$

ii) Form the random variable H_t by computing multiperiod forecasts for each horizon and summing these forecasts over all horizons. Start by constructing

$$\overline{\kappa}(x) = E\left[\kappa(X_t, W_{t+1}) - \nu \mid X_t = x\right],$$

Thus

$$E\left[\kappa(X_{t+j-1}, W_{t+j}) - \nu | X_t = x\right] = \mathbb{T}^{j-1}\overline{\kappa}(x).$$

Summing the terms, construct

$$\begin{aligned} H_t &= \sum_{j=0}^{\infty} E\left(\left[\kappa(X_{t-1+j}, W_{t+j} - \nu\right] \mid X_t\right) \\ &= \kappa(X_{t-1}, W_t) - \nu + \sum_{j=0}^{\infty} E\left[\overline{\kappa}(X_{t+j}) \mid X_t\right] \\ &= \kappa_h(X_{t-1}, W_t) \end{aligned}$$

where

$$\begin{aligned} \kappa_h(X_{t-1}, W_t) &= \kappa(X_{t-1}, W_t) - \nu + \sum_{j=0}^{\infty} \mathbb{T}^j \overline{\kappa}(X_t) \\ &= \kappa(X_{t-1}, W_t) - \nu + \left(\mathbb{I} - \mathbb{T}\right)^{-1} \overline{\kappa}(X_t) \end{aligned}$$

where $\mathbb{T}$ is the operator defined in (3.1). The right side becomes a function of only (X_{t-1}, W_t) once we substitute for $\varphi(X_{t-1}, W_t)$ for X_t as implied by (4.1).

This construction requires that the infinite sum

$$\sum_{j=0}^{\infty} \mathbb{T}^j \overline{\kappa}(x) = \left(\mathbb{I} - \mathbb{T}\right)^{-1} \overline{\kappa}(x)$$

converges in mean square relative to the stationary distribution for $\{X_t : t \geq 0\}$. A sufficient condition for this is that $\mathbb{T}^m$ is a strong contraction for some integer $m \geqslant 1$ and $\overline{\kappa} \in \mathcal{N}$ where $\mathcal{N}$ is defined in (3.9).

iii) Compute

$$H_t^+ = E\left(H_{t+1} \mid X_t\right) = \kappa_+(X_t)$$

where[3]

$$\begin{aligned} \kappa_+(x) &\doteq E\left[\kappa(X_t, W_{t+1}) \mid X_t = x\right] - \nu + E\left[\left(\mathbb{I} - \mathbb{T}\right)^{-1} \overline{\kappa}(X_{t+1}) \mid X_t = x\right] \\ &= E\left[\kappa(X_t, W_{t+1}) \mid X_t = x\right] - \nu + \left(\mathbb{I} - \mathbb{T}\right)^{-1} \mathbb{T}\overline{\kappa}(x). \end{aligned}$$

[3] Notice that $\mathbb{T}\left(\mathbb{I} - \mathbb{T}\right)^{-1} \overline{\kappa}(x) = \left(\mathbb{I} - \mathbb{T}\right)^{-1} \mathbb{T}\overline{\kappa}(x)$.

iv) Build the martingale increment:

$$G_t = H_t - H^+_{t-1} = \kappa_m(X_{t-1}, W_t)$$

where

$$\kappa_m(X_{t-1}, W_t) = \kappa_h(X_{t-1}, W_t) - \kappa_+(X_{t-1}).$$

By construction, the expectation of $\kappa_m(X_t, W_{t+1})$ conditioned on X_t is zero.

Armed with these calculations, we now report a Markov counterpart to Proposition 2.2.2.

Proposition 4.2.1. *Suppose that $\{Y_t : t \geq\}$ is an additive functional, that $\mathbb{T}^m$ is a strong contraction on $\mathcal{N}$ for some m, and that $E[\kappa(X_t, W_{t+1})^2] < \infty$. Then*

$$Y_t \quad = \quad \underbrace{t\nu}_{\textbf{\textit{trend}}} \quad + \quad \underbrace{\sum_{j=1}^{t} \kappa_m(X_{j-1}, W_j)}_{\textbf{\textit{martingale}}} \quad - \quad \underbrace{\kappa_+(X_t)}_{\textbf{\textit{stationary}}} \quad + \quad \underbrace{Y_0 + \kappa_+(X_0)}_{\textbf{\textit{invariant}}}.$$

Notice that the martingale component is itself an additive functional. The first is a linear time trend, the second an additive martingale, the third a stationary process with mean zero, and the fourth a time-invariant constant. If we happen to impose the initialization: $Y_0 = -\kappa_+(X_0)$, then the fourth term is zero. We use a Proposition 4.2.1 decomposition as a way to associate a "permanent shock" with an additive functional. The permanent shock is the increment to the martingale.

4.3 Applications

We now compute martingale increments for two models of economic time series.

Application to a VAR

We apply the four-step construction in algorithm 4.2 when the Markov state $\{X_t\}$ follows a first-order VAR

$$X_{t+1} = \mathbb{A}X_t + \mathbb{B}W_{t+1}, \tag{4.3}$$

where $\mathbb{A}$ is a stable matrix and $\{W_{t+1} : t \geq 0\}$ is a sequence of independent and identically normally distributed random variables with mean vector zero and identity covariance matrix. The one-step ahead conditional covariance matrix of the time $t+1$ shocks BW_{t+1} to X_{t+1} equals BB'. Let

$$Y_{t+1} - Y_t = \kappa(X_t, W_{t+1}) = \mathbb{D}X_t + \nu + \mathbb{F}W_{t+1}, \tag{4.4}$$

where D and F are row vectors with the same dimensions as X_t and W_{t+1}, respectively, and the $(\cdot)$ symbol denotes an inner product. For this example, the four steps of algorithm 4.2 become:

(i) The trend growth rate is ν as specified.

(ii)
$$\kappa_h(X_{t-1}, W_t, X_t) = \mathbb{D}X_{t-1} + \mathbb{F}W_t + \mathbb{D}(\mathbb{I} - \mathbb{A})^{-1}X_t$$

(iii)
$$\kappa_+(x) = \mathbb{D}x + \mathbb{D}(\mathbb{I} - \mathbb{A})^{-1}\mathbb{A}x$$

(iv)
$$\begin{aligned} \kappa_m(X_{t-1}, W_t) &= \mathbb{F}W_t + \mathbb{D}(\mathbb{I} - \mathbb{A})^{-1}(X_t - \mathbb{A}X_{t-1}) \\ &= \left[\mathbb{F} + \mathbb{D}(\mathbb{I} - \mathbb{A})^{-1}\mathbb{B}\right] W_t \end{aligned}$$

From Example 1.10.1, we expect the coefficient of martingale increment to be the sum of impulse responses for the increment process $\{\mathbb{D}X_t + \mathbb{F}W_{t+1} : t \geq 0\}$. The impulse response function is the sequence of vectors:

$$\mathbb{F}, \mathbb{D}\mathbb{B}, \mathbb{D}\mathbb{A}\mathbb{B}, \mathbb{D}\mathbb{A}^2\mathbb{B}, \cdots . \tag{4.5}$$

Summing these vectors gives

$$\mathbb{F} + \mathbb{D}\left(\mathbb{I} + \mathbb{A} + \mathbb{A}^2 + \cdots\right)\mathbb{B} = \mathbb{F} + \mathbb{D}(\mathbb{I} - \mathbb{A})^{-1}\mathbb{B}$$

as anticipated.

Growth-Rate Regimes

We construct a Proposition 4.2.1 decomposition for a model with persistent switches in the conditional mean and volatility of the growth rate $Y_{t+1} - Y_t$.

Suppose that $\{X_t : t \geq 0\}$ evolves according to an n-state Markov chain with transition matrix $\mathbb{P}$. Realized values of X_t are coordinate vectors in $\mathbb{R}^n$. Suppose that $\mathbb{P}$ has only one unit eigenvalue. Let $\mathbf{q}$ be the row eigenvector associated with that unit eigenvalue normalized so that $\mathbf{q} \cdot \mathbf{1}_n = 1$ and

$$\mathbf{q}'\mathbb{P} = \mathbf{q}'.$$

Consider an additive functional satisfying

$$Y_{t+1} - Y_t = \mathbb{D}X_t + X_t{}'\mathbb{F}W_{1,t+1},$$

where $\{W_{1,t}\}$ is an i.i.d. sequence of multivariate standard normally distributed random vectors. Evidently, the stationary Markov $\{X_t : t \geq 0\}$ process induces discrete changes in both the conditional mean and the conditional volatility of the growth rate process $\{Y_{t+1} - Y_t\}$.

Observe that $E(X_{t+1}|X_t) = \mathbb{P}X_t$ and let

$$W_{2,t+1} = X_{t+1} - E\left(X_{t+1}|X_t\right). \tag{4.6}$$

Thus we can represent the evolution of the Markov chain as

$$X_{t+1} = \mathbb{P}X_t + W_{2,t+1}$$

$\{W_{2,t+1} : t \geq 0\}$ is an $n \times 1$ discrete-valued vector process that satisfies $E(W_{2,t+1}|X_t) = 0$, which is therefore a martingale increment sequence adapted to $X_t, X_{t-1}, \ldots, X_0$.

We again apply the four-step construction in algorithm 4.2.[4]

i)

$$\nu = \mathbb{D}\mathbf{q}$$

ii)

$$H_t = \mathbb{D}(X_{t-1} - \mathbf{q}) + X_{t-1}{}'\mathbb{F}W_{1,t} + \mathbb{D}\left((\mathbb{I} - \mathbb{P})^{-1} X_t\right.$$

[4] The operator $(\mathbb{I} - \mathbb{P})^{-1}$ applied to zero-means processes is well defined.

iii)

$$H_t^+ = E\left(H_{t+1} \mid X_t\right) = \mathbb{D}\left(X_t - \mathbf{q}\right) + \mathbb{D}\left(\mathbb{I} - \mathbb{P}\right)^{-1} \mathbb{P} X_t$$

which implies that

$$\kappa_+(x) = \mathbb{D}\left(X_t - \mathbf{q}\right) + \mathbb{D}\left(\mathbb{I} - \mathbb{P}\right)^{-1} \mathbb{P} x$$

iv)

$$G_t = H_t - H_{t-1}^+ = X_{t-1}{}' \mathbb{F} W_{1,t} + \mathbb{D}\left(\mathbb{I} - \mathbb{P}\right)^{-1} W_{2,t}$$

where we have substituted from equation (4.6).

The martingale increment has both continuous and discrete components:

$$\kappa_m(X_t, W_{t+1}) \quad = \quad \underbrace{X_t{}' \mathbb{F} W_{1,t+1}}_{\textbf{continuous}} \quad + \quad \underbrace{\mathbb{D}\left(\mathbb{I} - \mathbb{P}\right)^{-1}}_{\textbf{discrete}} W_{2,t+1}.$$

Chapter 5

Hidden Markov Models

5.1 Sufficient Statistics as States

This chapter presents *Hidden Markov Models* that start from a joint probability distribution consisting of a Markov process and a vector of noise-ridden signals about functions of the Markov state. Histories of signals are observed but the Markov state vector is not. *Statistical learning* about the Markov state proceeds by constructing a sequence of probability distributions of the Markov state conditional on histories of signals. Recursive representations of these conditional distributions form auxiliary Markov processes that summarize all information about the hidden state vector contained in histories of signals. A state vector in this auxiliary Markov process is a set of sufficient statistics for the probability distribution of the hidden Markov state conditional on the history of signals. We can construct this auxiliary Markov process of sufficient statistics sequentially.

We present four examples of Hidden Markov Models that are used to learn about

1. A continuously distributed hidden state vector in a linear state-space system

2. A discrete hidden state vector

3. Multiple VAR regimes

4. Unknown parameters cast as hidden invariant states

5.2 Kalman Filter and Smoother

We assume that a Markov state vector X_t and a vector Z_{t+1} of observations are governed by a linear state space system

$$\begin{aligned} X_{t+1} &= \mathbb{A}X_t + \mathbb{B}W_{t+1} \\ Z_{t+1} &= \mathbb{H} + \mathbb{D}X_t + \mathbb{F}W_{t+1}, \end{aligned} \tag{5.1}$$

where the matrix $\mathbb{F}\mathbb{F}'$ is nonsingular, X_t has dimension n, Z_{t+1} has dimension m and is a signal observed at $t+1$, W_{t+1} has dimension k and is a standard normally distributed random vector that is independent of X_t, of $Z^t = [Z_t, \ldots, Z_1]$, and of X_0. The initial state vector $X_0 \sim Q_0$, where Q_0 is a normal distribution with mean $\overline{X}_0$ and covariance matrix Σ_0.[1] To include the ability to represent an unknown fixed parameter as an invariant state associated with a unit eigenvalue in A, we allow A not to be a stable matrix.

Although $\{(X_t, Z_t), t = 0, 1, 2, \ldots\}$ is Markov, $\{Z_t, t = 0, 1, 2, \ldots\}$ is not.[2] We want to construct an affiliated Markov process whose date t state is Q_t, defined to be the probability distribution of the time t Markov state X_t conditional on history $Z^t = Z_t, \ldots, Z_1$ and Q_0. The distribution Q_t summarizes information about X_t that is contained in the history Z^t and Q_0. We sometimes use Q_t to indicate conditioning information that is "random" in the sense that it is constructed from a history of observable random vectors. Because the distribution Q_t is multivariate normal, it suffices to keep track only of the mean vector $\overline{X}_t$ and covariance matrix Σ_t of X_t conditioned on Q_0 and Z^t: $\overline{X}_t$ and Σ_t are sufficient statistics for the probability distribution of X_t conditional on the history Z^t and Q_0. Conditioning on Q_t is equivalent to conditioning on these sufficient statistics.

We can map sufficient statistics $(\overline{X}_{j-1}, \Sigma_{j-1})$ for Q_{j-1} into sufficient statistics $(\overline{X}_j, \Sigma_j)$ for Q_j by applying formulas for means and covariances of a conditional distribution associated with a multivariate normal distribution. This generates a recursion that maps Q_{j-1} and Z_j into Q_j. It enables us to construct $\{Q_t\}$ sequentially. Thus, consider the following three step process.

[1] Many expositions of Kalman filtering assume that $BF' = 0$. We shall study some interesting examples in which $BF' \neq 0$.

[2] The process $\{X_t, t = 0, 1, 2, \ldots\}$ is also Markov.

i) Express the joint distribution of X_{t+1}, Z_{t+1} conditional on X_t as

$$\begin{bmatrix} X_{t+1} \\ Z_{t+1} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ \mathbb{H} \end{bmatrix} + \begin{bmatrix} \mathbb{A} \\ \mathbb{D} \end{bmatrix} X_t, \begin{bmatrix} \mathbb{B} \\ \mathbb{F} \end{bmatrix} \begin{bmatrix} \mathbb{B}' & \mathbb{F}' \end{bmatrix} \right).$$

ii) Suppose that the distribution Q_t of X_t conditioned on Z^t and Q_0 is normal with mean $\overline{X}_t$ and covariance matrix Σ_t. Use the identity $X_t = \overline{X}_t + (X_t - \overline{X}_t)$ to represent $\begin{bmatrix} X_{t+1} \\ Z_{t+1} \end{bmatrix}$ as

$$\begin{bmatrix} X_{t+1} \\ Z_{t+1} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbb{H} \end{bmatrix} + \begin{bmatrix} \mathbb{A} \\ \mathbb{D} \end{bmatrix} \overline{X}_t + \begin{bmatrix} \mathbb{A} \\ \mathbb{D} \end{bmatrix} (X_t - \overline{X}_t) + \begin{bmatrix} \mathbb{B} \\ \mathbb{F} \end{bmatrix} W_{t+1},$$

which is just another way of describing our original state-space system (5.1). It follows that the joint distribution of X_{t+1}, Z_{t+1} conditioned on Z^t and Q_0, or equivalently on $(\overline{X}_t, \Sigma_t)$, is

$$\begin{bmatrix} X_{t+1} \\ Z_{t+1} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ \mathbb{H} \end{bmatrix} + \begin{bmatrix} \mathbb{A} \\ \mathbb{D} \end{bmatrix} \overline{X}_t, \begin{bmatrix} \mathbb{A} \\ \mathbb{D} \end{bmatrix} \Sigma_t \begin{bmatrix} \mathbb{A}' & \mathbb{D}' \end{bmatrix} + \begin{bmatrix} \mathbb{B} \\ \mathbb{F} \end{bmatrix} \begin{bmatrix} \mathbb{B}' & \mathbb{F}' \end{bmatrix} \right).$$

Evidently the marginal distribution of Z_{t+1} conditional on $(\overline{X}_t, \Sigma_t)$ is

$$Z_{t+1} \sim \mathcal{N}(\mathbb{H} + \mathbb{D}\overline{X}_t, \mathbb{D}\Sigma_t\mathbb{D}' + \mathbb{F}\mathbb{F}').$$

This is called the predictive conditional density $\varphi(z^*|Q_t)$, i.e., the distribution of Z_{t+1} conditional on history Z^t and the initial distribution Q_0.

iii) Joint normality implies that the distribution for X_{t+1} conditional on Z_{t+1} and $(\overline{X}_t, \Sigma_t)$ is also normal and fully characterized by a conditional mean vector and a conditional covariance matrix. We can compute the conditional mean by running a population regression of $X_{t+1} - A\overline{X}_t$ on the surprise in Z_{t+1} defined as $Z_{t+1} - \mathbb{H} - \mathbb{D}\overline{X}_t$.[3] Having thus transformed random vectors on both sides of our regression to be independent of past observable information, as ingredients of the

[3] This amounts to dividing the joint distribution for (X_{t+1}, Z_{t+1}) conditioned on Q_t by the marginal density for Z_{t+1} conditional on Q_t.

pertinent population regression, we have to compute the covariance matrices

$$
\begin{aligned}
E\left[\left(Z_{t+1} - \mathbb{H} - \mathbb{D}\overline{X}_t\right)\left(Z_{t+1} - \mathbb{H} - \mathbb{D}\overline{X}_t\right)'\right] &= \mathbb{D}\Sigma_t\mathbb{D}' + \mathbb{F}\mathbb{F}' \equiv \Omega_t \\
E\left[\left(X_{t+1} - \mathbb{A}\overline{X}_t\right)\left(Z_{t+1} - \mathbb{H} - \mathbb{D}\overline{X}_t\right)'\right] &= \mathbb{A}\Sigma_t\mathbb{D}' + \mathbb{B}\mathbb{F}'.
\end{aligned}
$$

These provide what we need to compute the conditional expectation

$$E[(X_{t+1} - \mathbb{A}\overline{X}_t) \mid Z_{t+1} - \mathbb{H} - \mathbb{D}\overline{X}_t, Q_t] = \mathcal{K}(\Sigma_t)(Z_{t+1} - \mathbb{H} - \mathbb{D}\overline{X}_t),$$

where the matrix of regression coefficients $\mathcal{K}(\Sigma_t)$ called the *Kalman gain* is

$$\mathcal{K}(\Sigma_t) = (\mathbb{A}\Sigma_t\mathbb{D}' + \mathbb{B}\mathbb{F}')(\mathbb{D}\Sigma_t\mathbb{D}' + \mathbb{F}\mathbb{F}')^{-1}. \tag{5.2}$$

We recognize formula (5.2) as an application of the population least squares regression formula associated with the multivariate normal distribution.[4] We compute Σ_{t+1} via the recursion

$$
\begin{aligned}
\Sigma_{t+1} =& \mathbb{A}\Sigma_t\mathbb{A}' + \mathbb{B}\mathbb{B}' \\
& - (\mathbb{A}\Sigma_t\mathbb{D}' + \mathbb{B}\mathbb{F}')(\mathbb{D}\Sigma_t\mathbb{D}' + \mathbb{F}\mathbb{F}'^{-1}(\mathbb{D}\Sigma_t\mathbb{A}' + \mathbb{F}\mathbb{B}').
\end{aligned} \tag{5.3}
$$

The right side of recursion (5.3) follows directly from substituting the appropriate formulas into the right side of $\Sigma_{t+1} \equiv E(X_{t+1} - \overline{X}_{t+1})(X_{t+1} - \overline{X}_{t+1})'$ and computing conditional mathematical expectations. The matrix Σ_{t+1} obeys the formula from standard regression theory for the population covariance matrix of the least squares residual $X_{t+1} - \mathbb{A}\overline{X}_t$. The matrix $\mathbb{A}\Sigma_t\mathbb{A}' + \mathbb{B}\mathbb{B}'$ is the covariance matrix of the $X_{t+1} - \mathbb{A}\overline{X}_t$ and the remaining term describes the reduction in covariance associated with conditioning on Z_{t+1}.[5] Thus, the probability distribution Q_{t+1} is

$$X_{t+1} \mid Z_{t+1}, \overline{X}_t, \Sigma_t \sim \mathcal{N}(\overline{X}_{t+1}, \Sigma_{t+1}).$$

[4] Presentations of multivariate regression theory often report the transpose of this matrix. Those presentations pre-multiply coefficients by regressors whereas as Kalman filtering representations post-multiply by regressors.

[5] Let z be an $N \times 1$ random vector with multivariate normal probability density $f(z;\mu,\Sigma) = (2\pi)^{-(\frac{N}{2})}\det(\Sigma)^{-(\frac{1}{2})}\exp\left(-.5(z-\mu)'\Sigma^{-1}(z-\mu)\right)$ where $\mu = Ez \equiv \int z f(z;\mu,\Sigma)\,dz$ is the mean of z and $\Sigma = E(z-\mu)(z-\mu)' \equiv \int (z-\mu)(z-\mu)' f(z;\mu,\Sigma)\,dz$ is the covariance matrix of z. For integer $j \in [2, \ldots, N-1]$, partition z as $z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$, where z_1 is an $(N-j) \times 1$ vector and z_2 is a $j \times 1$ vector. Let $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$

where

$$\overline{X}_{t+1} = \mathbb{A}\overline{X}_t + \mathcal{K}(\Sigma_t)(Z_{t+1} - \mathbb{H} - \mathbb{D}\overline{X}_t) \tag{5.4}$$

Equations (5.2), (5.3), and (5.4) constitute the Kalman filter. They provide a recursion that describes Q_{t+1} as an exact function of Z_{t+1} and Q_t.

Remark 5.2.1. *(Gram-Schmidt) The key idea underying the Kalman filter is recursively to transform the space spanned by a sequence of signals into an a sequence of orthogonal signals. To elaborate, let*

$$U_{t+1} = Z_{t+1} - \mathbb{H} - \mathbb{D}\overline{X}_t.$$

After we condition on $(\overline{X}_0, \Sigma_0)$, $U_t, U_{t-1}, ...U_1$ and $Z_t, Z_{t-1}, ..., Z_1$ generate the same information. The Kalman filter synthesizes U_{t+1} from Z^{t+1} via a Gram-Schmidt process. Conditional on Z^t, $U_{t+1} \sim \mathcal{N}(0, \Omega_t)$, where $\Omega_t = \mathbb{D}\Sigma_t\mathbb{D}' + \mathbb{F}\mathbb{F}'$, so $U^t = U_t, U_{t-1}, ...U_1$ is an orthogonal basis for information contained in Z^t. Step (ii) *computes the innovation U_{t+1} by constructing the predictive density, while step* (iii) *computes the Kalman gain $\mathcal{K}(\Sigma_t)$ by regressing $X_{t+1} - \mathbb{A}\overline{X}_t$ on U_{t+1}.*

Innovations Representation

Taken together, steps (ii) and (iii) present the evolution of $\{Q_{t+1})\}$ as a first-order Markov process. This process is the foundation of an *innovations representation* and its partner the *whitener*. The innovations representation is

$$\begin{aligned} \overline{X}_{t+1} &= \mathbb{A}\overline{X}_t + \mathcal{K}(\Sigma_t)U_{t+1} \\ Z_{t+1} &= \mathbb{H} + \mathbb{D}\overline{X}_t + U_{t+1}. \end{aligned} \tag{5.5}$$

be corresponding partitions of μ and Σ. The marginal densities of the random vectors z_1 and z_2 are $f(z_1; \mu_1, \Sigma_{11})$ and $f(z_2; \mu_2, \Sigma_{22})$, respectively, where $f(z_i; \mu_i, \Sigma_{ii})$ denotes a multivariate normal density with mean vector μ_i and covariance matrix Σ_{ii}. The conditional density of z_1 given z_2, denoted $f(z_1|z_2; \hat{\mu}_1, \hat{\Sigma}_{11})$, is multivariate normal with mean $\hat{\mu}_1 = \mu_1 + \beta(z_2 - \mu_2)$ and covariance matrix $\hat{\Sigma}_{11} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \Sigma_{11} - \beta\Sigma_{22}\beta'$ where $\beta = \Sigma_{12}\Sigma_{22}^{-1}$ is an $(N-j) \times j$ matrix of population regression coefficients of $z_1 - \mu_1$ on $z_2 - \mu_2$. Here $\hat{\mu}_1 = Ez_1|z_2$ and $\hat{\Sigma}_{11} = E[(z_1 - \hat{\mu}_1)(z_1 - \hat{\mu}_1)']|z_2$.

The *whitener* system is

$$\begin{aligned} U_{t+1} &= Z_{t+1} - \mathbb{H} - \mathbb{D}\overline{X}_t \\ \overline{X}_{t+1} &= [\mathbb{A} - \mathbb{D}\mathcal{K}(\Sigma_t)]\,\overline{X}_t + \mathcal{K}(\Sigma_t)(Z_{t+1} - \mathbb{H}) \end{aligned} \tag{5.6}$$

The innovations representation (5.5) and the whitener system (5.6) both take sequences $\{\Sigma_t, \mathcal{K}(\Sigma_t)\}_{t=0}$ as inputs. These can be precomputed from equations (5.2) and (5.3) before observing any Z_{t+1}'s.

Remark 5.2.2. *The covariance matrix Ω_t is presumed to be nonsingular, but it is not necessarily diagonal so that components of the innovation vector U_{t+1} are possibly correlated. We can transform the innovation vector U_{t+1} to produce a new shock process $\overline{W}_{t+1}$ that has the identity as its covariance matrix. To do so construct a matrix Λ_t that satisfies*

$$\Lambda_t = \overline{\mathbb{F}}_t(\overline{\mathbb{F}}_t)'$$

and let

$$\overline{W}_{t+1} = \left(\overline{\mathbb{F}}_t\right)^{-1} U_{t+1}$$

Then

$$\begin{aligned} \overline{X}_{t+1} &= \mathbb{A}\overline{X}_t + \overline{\mathbb{B}}_t\overline{W}_{t+1} \\ Z_{t+1} &= \mathbb{H} + \mathbb{D}\overline{X}_t + \overline{\mathbb{F}}_t\overline{W}_{t+1} \end{aligned} \tag{5.7}$$

where $\overline{\mathbb{B}}_t = \mathcal{K}(\Sigma_t)\overline{\mathbb{F}}_t$ and A Gram-Schmidt process can be used to construct $\overline{W}_{t+1}$.

Please compare the original state space system (5.1) with the innovation representations (5.5) and (5.7). Key differences are

1. In the original system (5.1), the shock vector W_{t+1} can be of much larger dimension than the time $t+1$ observation vector Z_{t+1}, while in the innovation representations (5.5) and (5.7), the dimension of the shock U_{t+1} or $\overline{W}_{t+1}$ equals that of the observation vector.

2. The state vector X_t in the original system (5.1) is not observed while in the innovation representation (5.5) the state vector $\overline{X}_t$ is observed.

Likelihood process

Equations (5.2) and (5.3) together with an initial distribution Q_0 for $X_0 \sim \mathcal{N}(\overline{X}_0, \Sigma_0)$ provide components that allow us to construct a recursive representation for a likelihood process for $\{Z_t : t = 1, 2, \ldots\}$. Let $\psi(z^*|\mu, \Sigma)$ denote the density for an m dimensional, normally distributed random vector with mean μ and covariance matrix Λ. With this notation, the density of Z_{t+1} conditional on the on the hidden state X_t is $\psi(z^* \mid \mathbb{H} + \mathbb{D}X_t, \mathbb{B}\mathbb{B}')$, where z^* is an m dimensional vector of real numbers used to represent potential realizations of Z_{t+1}. The distribution of the hidden state X_t conditioned on history Z^{t-1} and ($\overline{X}_0$ and Σ_0) is $Q_t \sim \mathcal{N}(\overline{X}_t, \Sigma_t)$. From these two components, we construct the predictive density $\varphi(z^*|Z^t)$ for Z_{t+1}:

$$\varphi(z^* \mid Z^t, \overline{X}_0, \Sigma_0) = \int \psi(z^* \mid x) Q_t(dx). \tag{5.8}$$

From the Kalman filter, we know that

$$\int \psi(z^* \mid x) Q_t(dx) = \psi(z^* \mid \mathbb{H} + \mathbb{D}\overline{X}_t, \Omega_t)$$

To compute a likelihood process $\{L_t : t = 1, 2, ...\}$, factor the joint density for Z^t into a product of conditional density functions in which a time j density function conditions on past information and the initial $\overline{X}_0, \Omega_0)$. When we evaluate densities at the appropriate random vectors Z_j and the associated histories Z^{j-1} of which $\overline{X}_{j-1}, \Omega_{j-1}$ are functions determined by the Kalman filter, we obtain the likelihood process:[6]

$$L_t = \prod_{j=1}^{t} \psi(Z_j \mid \mathbb{H} + \mathbb{D}\overline{X}_{j-1}, \Omega_{j-1}). \tag{5.9}$$

Via the Kalman filtering formulas for $\{\overline{X}_j, \Omega_j\}_{j=1}^{\infty}$, this construction indicates how the likelihood process depends on the matrices $\mathbb{A}, \mathbb{B}, \mathbb{H}, \mathbb{D}, \mathbb{F}$. Sometimes we regard some entries of these matrices as “free parameters.”

[6] The logarithm of time j component of L_t is evidently

$$\begin{aligned} \log \psi(Z_j \mid H + D\overline{X}_{j-1}) = & -.5m \log(2\pi) - .5 \log\det(\Omega_{j-1}) \\ & - .5(Z_j - H - D\overline{X}_{j-1})'\Omega_{j-1}^{-1}(Z_j - H - D\overline{X}_{j-1}). \end{aligned}$$

Because a likelihood process summarizes information about these parameters, it is the starting point for both frequentist and Bayesian estimation procedures.

1. For fixed values of the parameters that pin down $\mathbb{A}, \mathbb{B}, \mathbb{H}, \mathbb{D}, \mathbb{F}$, $\{L_t\}_{t=1}^{\infty}$ is a stochastic process with some "interesting properties."

2. For a fixed t and a sample of observations Z^t, L_t becomes a *likelihood function* when viewed as a function of the free parameters.

Invariant Kalman gain

If $\overline{\Sigma}$ is a positive definite fixed point of recursion (5.3) and $\Sigma_0 = \overline{\Sigma}$, then $\Sigma_t = \overline{\Sigma}$ for all $t \geq 0$ and

$$\begin{aligned} \mathcal{K}\left(\Sigma_t\right) &= \mathcal{K}\left(\overline{\Sigma}\right) \doteq \overline{\mathcal{K}} \\ \Omega_t &= D\overline{\Sigma}_t D' + FF' \doteq \overline{\Omega} \end{aligned}$$

for all $t \geqslant 1$ simplifies recursive representation (5.9) by making $\mathcal{K}(\Sigma_t)$ and Ω_t both becomes time-invariant. Setting $\Sigma_0 = \overline{\Sigma}$ to the positive semidefinite fixed point of iterations on equation (5.3), sometimes called a matrix Riccati equation, amounts to pretending that at date zero we are conditioning on an infinite history of Z_t's.

Example 5.2.3. *John F. Muth (1960) posed and solved the following inverse optimal prediction problem: for what stochastic process $\{Z_t\}_{t=0}^{\infty}$ is the adaptive expectations scheme of Milton Friedman (1957)*

$$Z_t^* = \lambda Z_t + (1-\lambda) Z_{t-1}^* \quad 0 < \lambda < 1 \tag{5.10}$$

optimal for predicting future Z_{t+k}? And over what horizon k, if any, is Z_t^* a good forecast? Solving difference equation (5.10) *backwards indicates how past data shape Z_t^*:*

$$Z_t^* = \lambda \sum_{j=0}^{\infty} (1-\lambda)^j Z_{t-j}.$$

Although Muth did not use it to solve his problem, we can convey his answers concisely using the Kalman filter. As described above, inventing an infinite *past amounts to initializing the Kalman filter at* $\Sigma_0 = \overline{\Sigma}$. *Set* $\mathbb{A} = \mathbb{D} = 1$, $\mathbb{B} = \begin{bmatrix} \mathbb{B}_1 & 0 \end{bmatrix}$, *and* $\mathbb{F} = \begin{bmatrix} 0 & \mathbb{F}_2 \end{bmatrix}$ *to attain the original state-space system*

$$\begin{aligned} X_{t+1} &= X_t + \mathbb{B}_1 W_{1,t+1} \\ Z_{t+1} &= X_t + \mathbb{F}_2 W_{2,t+1}. \end{aligned}$$

Notice that the best forecast of Z_{t+k} *at time* t *when the state is observed is* X_t *for any* $k \geq 1$. *By the Law of Iterated Expectations, we obtain the mathematical expectation of* Z_{t+k} *conditional on* Z^t *by computing* $\overline{X}_t$. *A time-invariant recursive representation of* $\overline{X}_{t+1}$ *is*

$$\overline{X}_{t+1} = \overline{X}_t + \overline{\mathcal{K}}(Z_{t+1} - \overline{X}_t),$$

where it can be verified that $0 < \overline{\mathcal{K}} < 1$. *Notice that*

$$\overline{X}_{t+1} = \left(1 - \overline{\mathcal{K}}\right) \overline{X}_t + \overline{\mathcal{K}} Z_{t+1} \tag{5.11}$$

Comparing (5.10) *to* (5.11) *shows that "adaptive" expectations become "rational" by setting*

$$\begin{aligned} \overline{X}_t &= Z_t^* \\ \lambda &= \overline{\mathcal{K}}. \end{aligned}$$

Example 5.2.4. *As state variables for the key Bellman equation in his matching model, Jovanovic (1979) deployed sufficient statistics of conditional distribution* Q_t *for a univariate hidden Markov state equal to an unknown constant match quality* θ *drawn from a known initial distribution* $\mathcal{N}\left(\overline{X}_0, \Sigma_0\right)$. *The state-space representation for Jovanovic's model is*

$$\begin{aligned} X_{t+1} &= X_t \\ Z_{t+1} &= X_t + \mathbb{F} W_{t+1} \end{aligned}$$

where $\mathbb{F}$ *and* $X_t = \theta$ *are scalars and* W_{t+1} *is a standardized univariate normal random variable. We fit this model into* (5.1) *by setting* $\mathbb{A} = \mathbb{D} = 1, \mathbb{B} = 0, \mathbb{F} > 0, X_t = \theta$. *Evidently,* $\overline{X}_{t+1} = (1 - \mathcal{K}(\Sigma_t))\overline{X}_t + \mathcal{K}(\Sigma_t) Z_t$ *where* $\Sigma_{t+1} = \frac{\Sigma_t \mathbb{F}^2}{\Sigma_t + \mathbb{F}^2}$ *and* $\mathcal{K}(\Sigma_t) = \frac{\Sigma_t}{\Sigma_t + \mathbb{F}^2}$. *Thus,* $\frac{1}{\Sigma_{t+1}} = \frac{1}{\Sigma_t} + \frac{1}{\mathbb{F}^2} \downarrow 0$ *and* $\mathcal{K}(\Sigma_t) \to 0$. *Thus, partners to an ongoing match who observe* Z^t *eventually learn its true quality* θ. *In Jovanovic's model, especially when* $\mathbb{F}$ *is large, early on in a* *match, Σ_t can be large enough to create a situation in which the "he's just been having a few bad days" excuse prevails to sustain the match in hopes of later learning that it is a good one. Jovanovic put this force to work to help explain why (a) quits and layoffs are negatively correlated with job tenure and (b) wages rise with job tenure.*

Example 5.2.5. *Testing random walk theory of asset prices. We illustrate a classic finding of Working (1934). The price of an asset X_t takes a random walk $X_{t+1} = X_t + BW_{t+1}$, where W_{t+1} is a standardized univariate normal distribution and successive W_{t+j}'s are i.i.d. A researcher wants to test the random walk hypothesis. A data base reports not X_t but a two-period moving average $Z_t = .5(X_t + X_{t-1})$, which evidently implies that $Z_{t+1} = X_t + .5BW_{t+1}$. Here $A = D = 1, F = .5B$. The time-invariant innovations representation for the measured asset price process $\{Z_{t+1} : t = 0, 1, ...\}$ is*

$$\begin{aligned} \overline{X}_{t+1} &= \overline{X}_t + \overline{\mathcal{K}} U_{t+1} \\ Z_{t+1} &= \overline{X}_t + U_{t+1} \end{aligned} \tag{5.12}$$

where $0 < \overline{\mathcal{K}} < 1$. Compute

$$\begin{aligned} Z_{t+1} - Z_t = \overline{X}_t + U_{t+1} - Z_t &= U_{t+1} - U_t + \overline{X}_t - \overline{X}_{t-1} \\ &= U_{t+1} - \left(1 - \overline{\mathcal{K}}\right) U_t. \end{aligned} \tag{5.13}$$

Thus, the first-difference process is temporally dependent so the measured stock price Z_{t+1} does not take a random walk. It is instead a first-order "moving-average process". The time averaging induces serial correlation of a very specific form but alters how an empirical researcher should test the random walk hypothesis about the X_t process. We can deduce a population regression of $Z_{t+1} - Z_t$ on Z^t by using (5.13) to compute U_{t+1}

$$\begin{aligned} U_{t+1} &= \sum_{j=0}^{\infty} \left(1 - \overline{\mathcal{K}}\right)^j (Z_{t+1-j} - Z_{t-j}) \\ &= Z_{t+1} - \overline{\mathcal{K}} \sum_{j=0}^{\infty} \left(1 - \overline{\mathcal{K}}\right)^j Z_{t-j}. \end{aligned}$$

Rearranging terms gives us a so-called autoregressive representation:

$$Z_{t+1} = \overline{\mathcal{K}} \sum_{j=0}^{\infty} \left(1 - \overline{\mathcal{K}}\right)^j Z_{t-j} + U_{t+1},$$

which tells us what coefficients on lagged Z_t's should be if the underlying stock price does indeed follow a random walk. It is straightforward to verify that the regression coefficients on the right side of the above equation sum to one. We also have the following representation for a regression of the first difference $Z_{t+1} - Z_t$ on Z^t

$$Z_{t+1} - Z_t = U_{t+1} + (1 - \overline{\mathcal{K}})[Z_t - \overline{\mathcal{K}} \sum_{j=0}^{\infty} (1 - \overline{\mathcal{K}})^j Z_{t-j-1}].$$

Evidently, measured prices changes $Z_{t+1} - Z_t$ are forecastable from Z^t, which belies the random walk hypothesis for the $\{Z_t\}$ process.

Example 5.2.6. *Skip sampling. What really concerned Working (1934) were the consequences of taking r-period moving averages and then running time series regressions on r-period skip-sampled data. The Kalman filter provides tools for working this out. Let's do it for $r = 2$. The construction works more generally, so we start by iterating once on the original state-space representation* (5.1) *to get:*

$$\begin{aligned} X_{t+2} &= A^2 X_t + B W_{t+2} + A B W_{t+1} \\ Z_{t+2} &= H + D A X_t + F W_{t+2} + D B W_{t+1}, \end{aligned}$$

Consider sampling at even points in time. That is, let $t = 2\tau$ and construct the skip-sampled processes $\{X^s_\tau : \tau = 0, 1, ...\}$ and $\{Z^s_\tau : \tau = 0, 1, ...\}$ where $X^s_\tau = X_{2\tau}$ and $Z^s_\tau = Z_{2\tau}$. Define a new recursive representation:

$$\begin{aligned} X^s_{\tau+1} &= A_s X^s_\tau + B_s W^s_{\tau+1} \\ Z^s_{\tau+1} &= H + D_s X^s_\tau + F_s W^s_{\tau+1} \end{aligned}$$

where

$$W^s_{\tau+1} \doteq \begin{bmatrix} W_{2\tau+2} \\ W_{2\tau+1} \end{bmatrix},$$

$A_s \doteq A^2$, $D_s \doteq DA$ and

$$B_s \doteq \begin{bmatrix} B & AB \end{bmatrix} \quad F_s \doteq \begin{bmatrix} F & DB \end{bmatrix}$$

We can then construct an innovations representation and associated likelihood process for two-period skip-sampled process $\{Z_\tau : \tau = 1, ...\infty\}$. As a special case, we could apply this analysis to study a skip-sampled version of Example 5.2.5 of a process formed as a two-period moving-average of a stock price that, before the moving average, was taking a random walk.

Example 5.2.7. *Two moving-average representations. A first-order moving average process $\{Z_{t+1}\}$ obeys $Z_{t+1} = W_{t+1} - \lambda W_t$, where $\{W_t\}$ is a univariate i.i.d. process of standardized normal random variables and $\lambda > 1$. Use backward recursions on $Z_{t+1} = W_{t+1} - \lambda W_t$ to solve for W_{t+1} as a function of $\{Z_{t+1}\}$ to get*

$$W_{t+1} = \sum_{j=0}^{\infty} \lambda^j Z_{t+1-j}.$$

But λ^j explodes and the sum on the right side is not a (mean-square) convergent series – an indication that the random variable W_{t+1} does not belong to the space spanned by squared summable linear combinations of the history $\{Z_{t+1-j} : j = 0, 1, ...\}$. Although the backward recursion fails to converge, we can write

$$W_t = \frac{1}{\lambda}\left[W_{t+1} + Z_{t+1}\right]$$

and solve forward to indicate how observation of W_t peeks at future Zs.

We construct an alternative moving-average representation using the time invariant Kalman filter. A state-space representation for our first-order moving-average $\{Z_{t+1}\}$ process is

$$X_{t+1} = W_{t+1}$$
$$Z_{t+1} = -\lambda X_t + W_{t+1}.$$

Here $A = 0, B = 1, D = -\lambda, F = 1$. An innovations representation for the $\{Z_{t+1}\}$ process is

$$\overline{X}_t = \overline{\mathcal{K}} U_{t+1}$$
$$Z_{t+1} = -\lambda \overline{X}_t + U_{t+1}.$$

It can be verified that $\overline{\mathcal{K}} = \lambda^{-2}$ so that we have constructed the moving average representation

$$Z_{t+1} = U_{t+1} - \lambda^{-1} U_t.$$

Solve the implied difference equation $U_{t+1} = Z_{t+1} + \lambda^{-1} U_t$ in $\{U_t\}$ backwards to obtain

$$U_{t+1} = \sum_{j=0}^{\infty} \lambda^{-j} Z_{t+1-j},$$

which is well defined as a mean-square limit. This verifies that U_{t+1} can be constructed from $\{Z_{t+1-j}\}_{j=0}^{\infty}$.

We can use the original moving-average to compute second moments $E(Z_{t+1})^2 = (1+\lambda^2), E(Z_{t+1}Z_t) = -\lambda$ and our second one to compute $E(Z_{t+1})^2 = E(U_{t+1})^2(1+\lambda^{-2}), E(Z_{t+1}Z_t) = -E(U_{t+1})^2\lambda^{-1}$. These are consistent because $E(U_{t+1})^2 = \lambda^2$. The steady-state value $\overline{\Sigma} = (1-\lambda^{-2})$. Note that $E(U_{t+1})^2 > E(W_{t+1})^2$.

Kalman smoother

The Kalman filter provides recursive formulas for computing the distribution of a hidden state vector X_t conditional on a signal history $\{Z_\tau\}_{\tau=1}^{t}$ and an initial distribution Q_0 for X_0. This conditional distribution has the form $X_t \sim \mathcal{N}(\overline{X}_t, \Sigma_t)$; the Kalman filtering equations provide recursive formulas for the conditional mean $\overline{X}_t$ and the conditional covariance matrix Σ_t.

Knowing outcomes $\{\overline{X}_\tau, \Sigma_\tau\}_{\tau=1}^{T}$ from the Kalman filter provide the foundation for the *Kalman smoother.* The Kalman smoother uses past, present, and *future* values of Z_τ to learn about *current* values of the state X_τ. The Kalman smoother is a recursive algorithm that computes sufficient statistics for the distribution of X_t conditional on the *entire sample* $\{Z_t\}_{t=1}^{T}$, namely, a mean vector, covariance matrix pair $\widehat{X}_t, \widehat{\Sigma}_t$. The Kalman smoother takes outputs $\{\overline{X}_t, \Sigma_t\}_{t=0}^{T}$ from the Kalman filter as inputs and then works *backwards* on the following steps starting from $t = T$.

- Reversed time regression. Write the joint distribution of (X_t, X_{t+1}, Z_{t+1}) conditioned on $\left(\overline{X}_t, \Sigma_t\right)$ as

$$\begin{bmatrix} X_t \\ X_{t+1} \\ Z_{t+1} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \overline{X}_t \\ \mathbb{A}\overline{X}_t \\ \mathbb{H} + \mathbb{D}\overline{X}_t \end{bmatrix}, \begin{bmatrix} \Sigma_t & \Sigma_t\mathbb{A}' & \Sigma_t\mathbb{D}' \\ \mathbb{A}\Sigma_t & \mathbb{A}\Sigma_t\mathbb{A}' + \mathbb{B}\mathbb{B}' & \mathbb{A}\Sigma_t\mathbb{D}' + \mathbb{B}\mathbb{F}' \\ \mathbb{D}\Sigma_t & \mathbb{D}\Sigma_t\mathbb{A}' + \mathbb{F}\mathbb{B}' & \mathbb{D}\Sigma_t\mathbb{D}' + \mathbb{F}\mathbb{F}' \end{bmatrix} \right)$$

  From this joint distribution, construct the conditional distribution for X_t, given X_{t+1}, Z_{t+1} and $\left(\overline{X}_t, \Sigma_t\right)$. Compute the conditional mean of $X_t - \overline{X}_t$ by using the population least squares formula

$$\widehat{\mathbb{K}}_1\left(X_{t+1} - \mathbb{A}\overline{X}_t\right) + \widehat{\mathbb{K}}_2\left(Z_{t+1} - \mathbb{H} - \mathbb{D}\overline{X}_t\right) \tag{5.14}$$

where the regression coefficient matrix is

$$\begin{bmatrix} \widehat{\mathbb{K}}_1 & \widehat{\mathbb{K}}_2 \end{bmatrix} = \widehat{\mathbb{K}} \doteq \begin{bmatrix} \Sigma_t \mathbb{A}' & \Sigma_t \mathbb{D}' \end{bmatrix} \begin{bmatrix} \mathbb{A}\Sigma_t\mathbb{A}' + \mathbb{B}\mathbb{B}' & \mathbb{A}\Sigma_t\mathbb{D}' + \mathbb{B}\mathbb{F}' \\ \mathbb{D}\Sigma_t\mathbb{A}' + \mathbb{F}\mathbb{B}' & \mathbb{D}\Sigma_t\mathbb{D}' + \mathbb{F}\mathbb{F}' \end{bmatrix}^{-1}$$

and the residual covariance matrix equals

$$\Sigma_t - \begin{bmatrix} \Sigma_t \mathbb{A}' & \Sigma_t \mathbb{D}' \end{bmatrix} \begin{bmatrix} \mathbb{A}\Sigma_t\mathbb{A}' + \mathbb{B}\mathbb{B}' & \mathbb{A}\Sigma_t\mathbb{D}' + \mathbb{B}\mathbb{D}' \\ \mathbb{D}\Sigma_t\mathbb{A}' + \mathbb{F}\mathbb{B}' & \mathbb{D}\Sigma_t\mathbb{F}' + \mathbb{F}\mathbb{F}' \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{A}\Sigma_t \\ \mathbb{D}\Sigma_t \end{bmatrix} \tag{5.15}$$

- Iterated expectations. Notice that the above reverse regression includes $X_{t+1} - \mathbb{A}\overline{X}_t$ among the regressors. Because X_{t+1} is hidden, that is more information than we have. We can condition down to information that we actually have by instead using $\widehat{X}_{t+1} - \mathbb{A}\overline{X}_t$ as the regressor where $\widehat{X}_{t+1}$ is the conditional expectation of X_{t+1} given the full sample of data $\{Z_t\}_{t=1}^T$ and $\widehat{\Sigma}_{t+1}$ is the corresponding conditional covariance matrix. This gives us a backwards recursion for $\widehat{X}_t$:

$$\widehat{X}_t - \overline{X}_t = \widehat{\mathbb{K}}_1 \left(\widehat{X}_{t+1} - \mathbb{A}\overline{X}_t \right) + \widehat{\mathbb{K}}_2 \left(Z_{t+1} - \mathbb{H} - \mathbb{D}\overline{X}_t \right)$$

The law of iterated expectations implies that the regression coefficient matrices $\widehat{\mathbb{K}_1}, \widehat{\mathbb{K}_2}$ equal the ones we have already computed. But since we are using less information, the conditional covariance matrix increases by $\widehat{\mathbb{K}}_1 \widehat{\Sigma}_{t+1} \widehat{\mathbb{K}}_1'$. This implies the backwards recursion:

$$\begin{aligned} \widehat{\Sigma}_t = \Sigma_t - \begin{bmatrix} \Sigma_t \mathbb{A}' & \Sigma_t \mathbb{D}' \end{bmatrix} & \begin{bmatrix} \mathbb{A}\Sigma_t\mathbb{A}' + \mathbb{B}\mathbb{B}' & \mathbb{A}\Sigma_t\mathbb{D}' + \mathbb{B}\mathbb{D}' \\ \mathbb{D}\Sigma_t\mathbb{A}' + \mathbb{F}\mathbb{B}' & \mathbb{D}\Sigma_t\mathbb{D}' + \mathbb{F}\mathbb{F}' \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{A}\Sigma_t \\ \mathbb{D}\Sigma_t \end{bmatrix} \\ + \widehat{\mathbb{K}}_1 \widehat{\Sigma}_{t+1} \widehat{\mathbb{K}}_1' & \end{aligned}$$

- Take $\widehat{\Sigma}_T = \Sigma_T$ and $\widehat{X}_T = \overline{X}_T$ as terminal conditions.

5.3 Recursive Regression

A statistician wants to infer unknown parameters of a linear regression model. By treating regression coefficients as hidden states that are constant over time, we can cast this problem in terms of a hidden Markov model. By assigning a prior probability distribution to statistical models that are indexed by parameter values, the statistician can construct a stationary stochastic process as a mixture of statistical models.[7] From increments to a data history, the statistician learns about parameters sequentially. By assuming that the statistician adopts a conjugate prior á la Luce and Raiffa (1957), we can construct explicit updating formulas.

Consider the first-order vector autoregressive model

$$\begin{aligned} X_{t+1} &= AX_t + BW_{t+1} \\ Z_{t+1} &= H + DX_t + FW_{t+1} \end{aligned} \tag{5.16}$$

where W_{t+1} is an i.i.d. normal random vector with mean vector 0 and covariance matrix I, X_t is an observable state vector, and A, B, D, F, H are matrices containing unknown coefficients. Suppose that Z_{t+1} and W_{t+1} share the same dimensions, that F is nonsingular, and that X_t consists of $Y_t - Y_{t-1} - H$ and a finite number of lags $Y_{t-j} - Y_{t-j-1} - H, j = 1, \ldots, n$.

Conjugate prior updating

By following suggested offered by Zellner (1962), Box and Tiao (1992), Sims and Zha (1999), and especially Zha (1999), we can transform system (5.16) in a way that justifies estimating the unknown coefficients in the matrices A, B, D, F, H by applying least squares equation by equation. Factor the matrix $FF' = J\Delta J'$, where J is lower triangular with ones on the diagonal and Δ is diagonal.[8] Construct

$$J^{-1}(Y_{t+1} - Y_t) = J^{-1}H + J^{-1}DX_t + U_{t+1} \tag{5.17}$$

where

$$U_{t+1} = J^{-1}FW_{t+1}$$

so that $EU_{t+1}U'_{t+1} = \Delta$. The i^{th} entry of U_{t+1} is uncorrelated with, and consequently statistically independent of, the jth components of $Y_{t+1} - Y_t$ for $j = 1, 2, \ldots, i-1$. As a consequence, each equation in system (5.17) can be interpreted as a regression equation in which the left-hand side variable in

[7] This stochastic process is not ergodic, being a mixture of statistical models like those described by Proposition 1.8.1. In the present setting, conditioning on invariant events means knowing parameters, an assumption incompatible with posing a statistical learning problem.

[8] This factorization can be implemented as a Cholesky dcomposition.

equation i is the i^{th} component of $Y_{t+1} - Y_t$. The regressors are a constant, X_t, and the j^{th} components of $Y_{t+1} - Y_t$ for $j = 1, \ldots, i-1$. The ith equation is an unrestricted regression with a disturbance term $U_{t+1,i}$ that is uncorrelated with disturbances $U_{t+1,j}$ to all other equations $j \neq i$.

The system of equations (5.17) is thus recursive. The first equation determines the first entry of $Y_{t+1} - Y_t$, the second equation determines the second entry of $Y_{t+1} - Y_t$ given the first entry, and so forth.

We can construct estimates of the coefficient matrices A, B, D, F, H and the covariance matrix $\Delta = EU_{t+1}U_{t+1}'$ from these regression equations, with the qualification that knowledge of J and Δ determines FF' only up to a factorization of FF' for a nonsingular F. One such factorization is $F = J\Delta^{1/2}$, where a diagonal matrix raised to a one-half power can be built by taking the square root of each diagonal entry. Because matrices F not satisfying this formula also satisfy $FF' = J\Delta J'$, without additional restrictions F is not identified.

Consider, in particular, the ith regression formed in this way and express it as the scalar regression model:

$$Y_{t+1}^{[i]} - Y_t^{[i]} = R_{t+1}^{[i]}{}' \beta^{[i]} + U_{t+1}^{[i]}$$

where $R_{t+1}^{[i]}$ is the appropriate vector of regressors in the ith equation of system (5.17). To simplify notation, we will omit superscripts and understand that we are estimating one equation at a time. The disturbance U_{t+1} is a normally distributed random variable with mean zero and variance σ^2. Furthermore, U_{t+1} is statistically independent of R_{t+1}. Information observed as of date t consists of X_0 and $Y^t = [(Y_t - Y_{t-1})', \ldots, (Y_1 - Y_0)']'$. Suppose that in addition $Y_{t+1} - Y_t$ and R_{t+1} are also observed at date $t+1$ but that β and σ^2 are unknown.

Let the distribution of β conditioned on Y^t, X_0, and σ^2 be normal with mean b_t and precision matrix $\zeta\Lambda_t$ where $\zeta = \frac{1}{\sigma^2}$. Here the precision matrix equals the inverse of a conditional covariance matrix of the unknown parameters. At date $t+1$, information we add $Y_{t+1} - Y_t$ to the conditioning set. So we want the distribution of β conditioned on Y^{t+1}, X_0, and σ^2. It is also normal but now has precision $\zeta\Lambda_{t+1}$, where $\zeta = \frac{1}{\sigma^2}$ and

$$\Lambda_{t+1} = R_{t+1}R_{t+1}{}' + \Lambda_t. \tag{5.18}$$

Recursion (5.18) implies that $\Lambda_{t+1} - \Lambda_t$ is a positive semidefinite matrix, which confirms that additional information improves estimation accuracy.

Evidently from recursion (5.18), Λ_{t+1} cumulates cross-products of the regressors and adds them to an initial Λ_0. The updated conditional mean b_{t+1} for the normal distribution of unknown coefficients can be deduced from Λ_{t+1} via the updating equation:

$$\Lambda_{t+1} b_{t+1} = \left[\Lambda_t b_t + R_{t+1}(Y_{t+1} - Y_t)\right]. \tag{5.19}$$

Solving difference equation (5.19) backwards shows how $\Lambda_{t+1} b_{t+1}$ cumulates cross-products of R_{t+1} and $Y_{t+1} - Y_t$ adds the outcome to an initial condition $\Lambda_0 b_0$.

So far we pretended that we know σ^2 by conditioning on σ^2, which is equivalent to conditioning on its inverse ζ. Assume now that we don't know σ but instead summarize our uncertainty about it with a date t gamma density for ζ conditioned on Y^t, X_0 so that it is proportional to

$$(\zeta)^{\frac{c_t}{2}} \exp(-d_t \zeta/2),$$

where the density is expressed as a function of ζ, so that $d_t \zeta$ has a chi-square density with $c_t + 1$ degrees of freedom. The implied density for ζ conditioned on time $t+1$ information is also a gamma density with updated parameters:

$$\begin{aligned} c_{t+1} &= c_t + 1 \\ d_{t+1} &= (Y_{t+1} - Y_t)^2 - (b_{t+1})' \Lambda_{t+1} b_{t+1} + (b_t)' \Lambda_t b_t + d_t. \end{aligned}$$

The distribution of β conditioned on Y^{t+1}, X_0, and ζ is normal with mean b_{t+1} and precision matrix $\zeta \Lambda_{t+1}$. The distribution of ζ conditioned on Y^{t+1}, X_0 has a gamma density, so that it is proportional to[9]

$$(\zeta)^{\frac{c_{t+1}}{2}} \exp(-d_{t+1} \zeta/2).$$

Standard least squares regression statistics can be rationalized by positing a prior that is not informative. This is commonly done by using an "improper" priors that does not integrate to unity.[10] Setting $\Lambda_0 = 0$ effectively imposes a uniform but improper prior over β. Although Λ_t's early in

[9] A decision-maker who does not know the underlying parameters in the matrices A, B, D, F, H continues to have a Markov decision problem except that b_t, c_t, d_t must now be included along with the state vector X_t.

[10] Such a procedure can result in estimators that are inadmissible.

the sequence are singular, we can still update $\Lambda_{t+1} b_{t+1}$ via (5.19); b_{t+1} are not be uniquely determined until Λ_{t+1} becomes nonsingular. After enough observations have been accumulated to make Λ_{t+1} become nonsingular, the implied normal distributions for the unknown parameters become proper. When $\Lambda_0 = 0$, the specification of b_0 is inconsequential and b_{t+1} becomes a standard least squares estimator. An "improper gamma" prior over σ that is often associated with an improper normal prior over β sets c_0 to minus two and d_0 to zero. This is accomplished by assuming a uniform prior distribution for the logarithm of the precision ζ or for the logarithm of σ^2. With this combination of priors, d_{t+1} becomes a sum of squared regression residuals.[11]

From the posterior of the coefficients of this transformed system we can compute posteriors of nonlinear functions of those coefficients. We accomplish this by using a random number generator repeatedly to take pseudo random draws from the posterior probability of the coeffcients, forming those nonlinear functions, and then using the resulting histograms of those nonlinear functions to approximate the posterior probability distribution of those nonlinear functions. For example, many applied macroeconomic papers report impulse responses as a way to summarize model features. Impulse responses are nonlinear functions of the $(\mathbb{A}, \mathbb{B})$.

VAR example

In Hansen and Sargent (2021), to identify long-term risk in consumption we imposed cointegration on a VAR. We inferred consequences of this restriction by simulating posterior distributions that measure long-run risk. We turn to that example now.

We adapt the preceding approach along lines suggested by Hansen et al. (2008). We construct a trivariate VAR system in which (1) the logarithm of proprietor's income plus corporate profits, (2) the logarithm of personal dividend income, and (3) the logarithm of consumption have the same trend growth rate and martingale increment. The common martingale increment measures the long-run consumption risk discussed in section 4.4. Figure 5.1 reports log differences in two time series.

[11] Box and Tiao (1992) discuss improper priors that include the specification for the regression model here.

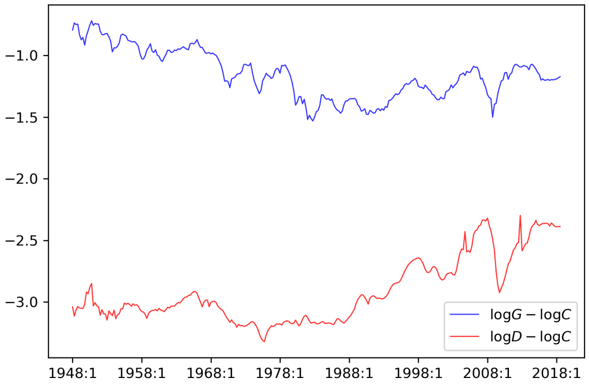


Figure 5.1: Time series for the i) logarithm of proprietor's income plus corporate profits relative to consumption (blue) and ii) the logarithm of personal dividend income relative to consumption (red).

We deployed the following steps.

i) Let

$$Z_{t+1} = \begin{bmatrix} \log C_{t+1} - \log C_t \\ \log G_{t+1} - \log C_{t+1} \\ \log D_{t+1} - \log C_{t+1} \end{bmatrix}$$

where C_t is consumption, G_t is business income, and D_t is personal dividend income. Business income is measured as proprietor's income plus corporate profits per capita. Dividends are personal dividend income per capita. The time series are quarterly data from 1948 Q1 to 2018 Q3.[12] [13]

[12] Our consumption measure is nondurables plus services consumption per capita. The nominal consumption data come from BEA's NIPA Table 1.1.5 and their deflators from BEA's NIPA Table 1.1.4. The business income data with IVA and CCadj are from BEA's NIPA Table 1.12. Personal dividend income data were obtained from from FRED's B703RC1Q027SBEA. Population data comes from FRED's CNP16OV.

[13] By including proprietors' income in addition to corporate profits, we used a broader measure of business income than Hansen et al. (2008) who used only corporate profits. Hansen et al. (2008) did not include personal dividends in their VAR analysis.

ii) Let

$$X_t = \begin{bmatrix} Z_t \\ Z_{t-1} \\ Z_{t-2} \\ Z_{t-3} \\ \log G_{t-4} - \log C_{t-4} \\ \log D_{t-4} - \log C_{t-4} \end{bmatrix}.$$

Express a vector autoregression as

$$\begin{aligned} X_{t+1} &= \mathbb{H} + \mathbb{A}X_t + \mathbb{B}W_{t+1} \\ Z_{t+1} &= \mathbb{D}X_t + \mathbb{F}W_{t+1} \end{aligned}$$

where $\mathbb{A}$ is a stable matrix (i.e., its eigenvalues are all bounded in modulus below unity) and $\mathbb{B}\mathbb{B}'$ is the innovation covariance matrix. Let selector matrix $\mathbb{J}$ verify $Z_{t+1} = \mathbb{J}X_{t+1}$. The implied mean μ of the stationary distribution for X is

$$\mu = (I - \mathbb{A})^{-1}\mathbb{H}.$$

The covariance matrix Σ of the stationary distribution of X solves a discrete Lyapunov equation

$$\Sigma = \mathbb{A}\Sigma\mathbb{A}' + \mathbb{B}\mathbb{B}'.$$

iii) $\log C_t, \log G_t, \log D_t$ are cointegrated. Each of $\log C_t, \log G_t, \log D_t$ is an additive functional in the sense of Chapter 4. Each has an additive decomposition into trend, martingale, and stationary components that can be constructed using a method described in Chapter 4. Trend and martingale components of the three series are identical by construction. The innovation to the martingale process is identified as the only shock having long-term consequences.

The conjugate prior approach described above does not generate a posterior for which either the prior or the implied posteriors for the matrix $\mathbb{A}$ has stable eigenvalues with probability one. We therefore modify that approach to impose that $\mathbb{A}$ is a stable matrix. We do this by rescaling the posterior probability so that it integrates to one over the region of the parameter space for which $\mathbb{A}$ is stable. We in effect condition on $\mathbb{A}$ being stable. This is easy to implement by rejection sampling.[14]

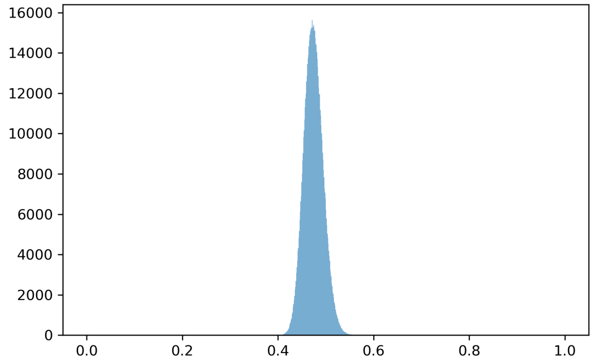
The standard deviation of the martingale increment is a nonlinear function of parameters in $(\mathbb{A}, \mathbb{B})$. We construct a posterior distribution via Monte Carlo simulation. We draw from the posterior of the multivariate regression system and, after conditioning on stability of the $\mathbb{A}$ matrix, compute the nonlinear functions of interest. From the simulation, we construct joint histograms to approximate posterior distributions of functions of interest. [15]

In Figure 5.2, we show posterior histograms for the standard deviations of shocks to short-term consumption growth and of the martingale increment to consumption. The standard deviation of the short-term shock contribution is about one-half that of the standard deviation of the martingale increment. Figure 5.2 tells us that short-term risk can be inferred with much more accuracy than is long-term risk. This evidence says that while there *could* be a long-run risk component to consumption, it is poorly measured. The fat tail in right of the distribution of the long-run standard deviation is induced by Monte Carlo draws for which some eigenvalues of $\mathbb{A}$ have absolute values very close to unity.[16]

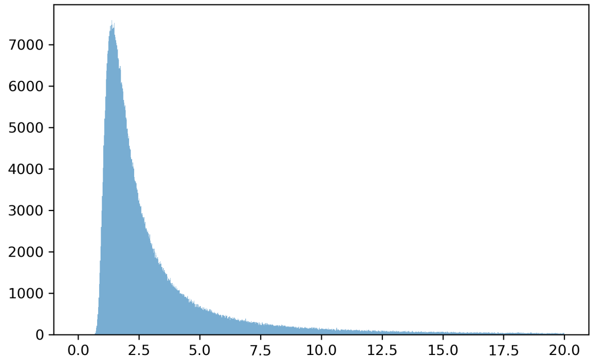
[14] Another approach that we don't use here would be to modify how we construct the likelihood function. Currently the likelihood function conditions on the initial X_0. We could instead impose that X_0 is described by the stationary distribution associated with a stable $\mathbb{A}$ matrix.

[15] We could also have used change in variables formulas to deduce posterior distributions of interest, but that would have involved substantial pencil and paper work and require additional numerical computation.

[16] Bounding absolute values of these eigenvalues to be less than a pre-specified number strictly less than one would thin the right tail. Doing that amounts indirectly to imposing a particular prior on the size of long-run risk.



((a)) Posterior density for conditional standard deviation of consumption growth.



((b)) Posterior distribution for the standard deviation of the martingale increment.

Figure 5.2: Comparing short-run and long-run volatility estimates.

Remark 5.3.1. *Carter and Kohn (1994) proposed an extension of the preceding method that is applicable to situations in which a state vector X_t is hidden. A Carter and Kohn approach would iterate on the following steps:*

- *Conditioned on parameters and a fixed data sample, use inputs into the Kalman smoother to simulate hidden states.*[17]

[17] A Kalman smoother works backward to construct a probability distribution for hidden states X_t for $t = 0, 1, ..., T-1$ conditioned on a complete sample of observations $\{Z_t : t = 1, 2, ...T\}$.

  - *First draw randomly X_T given $\{Z_t : t = 1, 2, ...T\}$ from the solution to the Kalman filtering problem.*
  - *Working backwards, for $t = T-1, T-2, ...1$, draw X_t given X_{t+1} and Z_{t+1} conditioned on $\{Z_\tau : \tau = 1, 2, ...t\}$ using the conditional expectation implied by* (5.14) *and covariance matrix* (5.15)*.*

- *Conditioned on data and hidden states, use the conjugate prior approach described above to simulate unknown parameters.*

Successive iterations on this algorithm form a Markov process with a state vector consisting of the hidden states and the parameters. Under appropriate regularity conditions, the Markov process has a stationary distribution to which the Markov process formed by the preceding iterations converges. That stationary distribution is *the joint posterior distribution of hidden states and parameter values. We are interested in the marginal posterior distributions over parameter values.*[18]

5.4 Mixtures

Suppose now that $\{X_t\}$ evolves as an n-state Markov process with transition probability matrix $\mathbb{P}$. A vector of signals $Y_{t+1} - Y_t$ has density $\psi_i(y^*)$ if state i is realized, meaning that X_t is the i^{th} coordinate vector. We want to compute the probability that X_t is in state i conditional on the signal history. The vector of conditional probabilities equals $Q_t = E\left[X_t | Y^t, Q_0\right]$, where Q_0 is a vector of initial probabilities. We construct $\{Q_t\}_{t=1}^{\infty}$ recursively:

i) Find the joint distribution of $(X_{t+1}, Y_{t+1} - Y_t)$ conditional on X_t. Conditional distributions of $Y_{t+1} - Y_t$ and X_{t+1} are statistically independent by assumption. Write the joint density conditioned on X_t as:

$$\underset{X_{t+1}\ \text{density}}{\underset{\uparrow}{(\mathbb{P}'X_t)}} \quad \times \quad \underset{Y_{t+1} - Y_t\ \text{density}}{\underset{\uparrow}{(X_t)'\text{vec}\,\{\psi_i(y^*)\}}} \tag{5.20}$$

where vec(r_i) is a column vector with r_i in the i^{th} component. We have expressed conditional independence by forming a joint conditional distribution as a product of two conditional densities, one for X_{t+1} and one for $Y_{t+1} - Y_t$.

[18] A Carter and Kohn simulation approach is an example of a Gibbs sampler.

ii) Find the joint distribution of $X_{t+1}, Y_{t+1} - Y_t$ conditioned on Q_t. Since X_t is not observed, we form the appropriate average of (5.20) conditioned on Y^t, Q_0:
$$\mathbb{P}'\text{diag}\{Q_t\} \text{ vec}\left\{\psi_i(y^*)\right\}, \tag{5.21}$$
where $\text{diag}(Q_t)$ is a diagonal matrix with the entries of Q_t on the diagonal. Thus, Q_t encodes all pertinent information about X_t that is contained in the history of signals. Conditional on Q_t, X_{t+1} and $Y_{t+1} - Y_t$ are *not* statistically independent.

iii) Find the distribution of $Y_{t+1} - Y_t$ conditional on Q_t. Summing (5.21) over the hidden states gives
$$(\mathbf{1}_n)'\mathbb{P}' \text{ diag}\{Q_t\}\text{vec}\left\{\psi_i(y^*)\right\} = Q_t \cdot \text{vec}\left\{\psi_i(y^*)\right\}.$$
Thus, Q_t is a vector of weights used to form a mixture distribution. Suppose, for instance, that ψ_i is a normal distribution with mean μ_i and covariance matrix Σ_i. Then the distribution of $Y_{t+1} - Y_t$ conditioned on Q_t is a *mixture of normals* with mixing probabilities given by entries of Q_t.

iv) Obtain Q_{t+1} by dividing the *joint* density of $(Y_{t+1} - Y_t, X_{t+1})$ conditional on Q_t by the *marginal* density for $Y_{t+1} - Y_t$ conditioned on Q_t and then evaluating this ratio at $Y_{t+1} - Y_t$. In this way we construct the density for X_{t+1} conditioned $(Q_t, Y_{t+1} - Y_t)$. It takes the form of a vector Q_{t+1} of conditional probabilities. Thus, we are led to
$$Q_{t+1} = \left(\frac{1}{Q_t \cdot \text{vec}\left\{\psi_i(Y_{t+1} - Y_t)\right\}}\right) \mathbb{P}'\text{diag}(Q_t)\text{vec}\left\{\psi_i(Y_{t+1} - Y_t)\right\} \tag{5.22}$$

Together, steps (iii) and (iv) define a Markov process for Q_{t+1}. As indicated in step (iii), $Y_{t+1} - Y_t$ is drawn from a (history-dependent) mixture of densities ψ_i. As indicated in step (iv), the vector Q_{t+1} equals the exact function of $Y_{t+1} - Y_t$, Q_t described in (5.22).

5.5 VAR Regimes

Following Sclove (1983) and Hamilton (1989), suppose that there are multiple VAR regimes (A_i, B_i, D_i, F_i) for $i = 1, 2, ..., n$, where indices i are governed by a Markov process with transition matrix $\mathbb{P}$. In regime i we have

$$\begin{aligned} X_{t+1} &= A_i X_t + B_i W_{t+1} \\ Y_{t+1} - Y_t &= D_i X_t + F_i W_{t+1}, \end{aligned}$$

where $\{W_{t+1}\}_{t=0}^{\infty}$ is an i.i.d. sequence of $\mathcal{N}(0, I)$ random vectors conditioned on X_0, and F_i is nonsingular.

We can think of X_t and a regime indicator Z_t jointly as forming a Markov process. When regime i is realized, Z_t equals a coordinate vector with one in the i^{th} coordinate and zeros at other coordinates. We study a situation in which regime indicator Z_t is not observed. Let Q_t denote an n-dimensional vector of probabilities over the hidden states Z_t conditioned on Y^t, X_0, and Q_0, where Q_0 is the date zero vector of initial probabilities for Z_0. Equivalently, Q_t is $E(Z_t|Y^t, X_0, Q_0)$.

The vector of conditional probabilities Q_t solves a *filtering problem.* We describe the solution of this problem by representing (X_t, Q_t) as a Markov process via the following four steps.

i) Find the joint distribution for $(Z_{t+1}, Y_{t+1} - Y_t)$ conditioned on (Z_t, X_t). Conditional distributions of Z_{t+1} and $Y_{t+1} - Y_t$ are statistically independent by assumption. Conditioned on Z_t, X_t conveys no information about Z_{t+1} and thus the conditional density of Z_{t+1} is given by entries of $\mathbb{P}'Z_t$. Conditioned on $Z_t = i$, $Y_{t+1} - Y_t$ is normal with mean $D_i X_t$ and covariance matrix $F_i(F_i)'$. Let $\psi_i(y^*, X_t)$ be the normal density function for $Y_{t+1} - Y_t$ conditioned on X_t when Z_t is in regime i. We can write the joint density conditioned on (Z_t, X_t) as:

$$\underbrace{(\mathbb{P}'Z_t)}_{\substack{\uparrow \\ Z_{t+1}\ \text{density}}} \quad \times \quad \underbrace{(Z_t)'\text{vec}\,\{\psi_i(y^*, X_t)\}}_{\substack{\uparrow \\ Y_{t+1} - Y_t\ \text{density}}} \tag{5.23}$$

where vec(r_i) is a column vector with r_i in the i^{th} entry. We have imposed conditional independence by forming a joint conditional distribution as a product of two conditional densities, one for Z_{t+1} and one for $Y_{t+1} - Y_t$.

ii) Find the joint distribution of $Z_{t+1}, Y_{t+1} - Y_t$ conditioned on (X_t, Q_t). Since Z_t is not observed, we form the appropriate average of (5.23) conditioned on the Y^t, X_0, Q_0:

$$\mathbb{P}'\text{diag}\{Q_t\}\ \text{vec}\,\{\psi_i(y^*, X_t)\} \tag{5.24}$$

where diag$\{Q_t\}$ is a diagonal matrix with components of Q_t on the diagonal. Thus, Q_t encodes all pertinent information about the time t regime Z_t that is contained in Y^t, X_0 and Q_0. Notice that conditional on (X_t, Q_t), random vectors $Y_{t+1} - Y_t$ and Z_{t+1} are *not* statistically independent.

iii) Find the distribution of $Y_{t+1} - Y_t$ conditioned on (X_t, Q_t). Summing (5.24) over hidden states gives

$$(\mathbf{1}_n)'\mathbb{P}'\ \text{diag}\{Q_t\}\text{vec}\,\{\psi_i(y^*, X_t)\} = Q_t \cdot \text{vec}\,\{\psi_i(y^*, X_t)\}\,.$$

Thus, the distribution for $Y_{t+1} - Y_t$ conditioned on (X_t, Q_t) is a *mixture of normals* in which, with probability given by the i^{th} entry of Q_t, $Y_{t+1} - Y_t$, is normal with mean $D_i X_t$ and covariance matrix $F_i F_i{}'$. Similarly, the conditional distribution of X_{t+1} is a mixture of normals.

iv) Obtain Q_{t+1} by dividing the *joint* density for $(Y_{t+1} - Y_t, Z_{t+1})$ conditioned on (X_t, Q_t) by the *marginal* density for $Y_{t+1} - Y_t$ conditioned on (X_t, Q_t). Division gives the density for Z_{t+1} conditioned $(Y_{t+1} - Y_t, X_t, Q_t)$, which in this case is just a vector Q_{t+1} of conditional probabilities. Thus, we are led to the recursion

$$Q_{t+1} = \left(\frac{1}{Q_t \cdot \text{vec}\,\{\psi_i(Y_{t+1} - Y_t, X_t)\}}\right) \mathbb{P}'\text{diag}(Q_t)\text{vec}\,\{\psi_i(Y_{t+1} - Y_t, X_t)\}\,. \tag{5.25}$$

Taken together, steps (iii) and (iv) provide the one-step-transition equation for Markov state (X_{t+1}, Q_{t+1}). As indicated in step (iii), $Y_{t+1} - Y_t$ is a mixture of normally distributed random variables. As argued in step (iv) the vector Q_{t+1} is an exact function of $Y_{t+1} - Y_t$, Q_t, and X_t that is given by formula (5.25).

Chapter 6

Likelihoods

This chapter studies likelihood processes and likelihood ratio processes. Derivatives of log-likelihood processes are additive martingales and likelihood ratio processes are multiplicative martingales, assertions that we verify by applying results from chapter 4. We study properties of likelihood ratios as sample size $T \rightarrow +\infty$ and relate them to methods for estimating parameters that pin down a statistical model from within either a discrete set or a manifold of models. These include maximum likelihood, Bayesian, and robust Bayesian methods. A workhorse in this chapter will be the Law of Large Numbers from Chapter 1 that applies in settings in which there are multiple statistical models.

In this chapter, we adopt settings in which state vectors can be inferred perfectly from observations. Chapter 5 studies situations in which some states are hidden and can be inferred only imperfectly.

6.1 Dependent Processes

Suppose that at date $t+1$ we observe a k dimensional random vector Z_{t+1}. We calculate various objects while conditioning on a given probability model. We use some of these calculations to explore alternative models. Each alternative model is presumed to imply a probability measure that is measure-preserving and ergodic. An event collection $\mathfrak{A}_t$ (i.e., a sigma algebra) is generated by the infinite history of Z_t.

We entertain a set of alternative probability models represented with their one-period transition probabilities. Use represent an alternative model as a perturbation of a baseline model. To represent a particular alternative model, we use a nonnegative random variable N_{t+1} to perturb a baseline model's one step transition probabilities. We can characterize an alternative model with a set of implied conditional expectations of all bounded random variable B_{t+1} that are measurable with respect to $\mathfrak{A}_{t+1}$. Such conditional expectations of B_{t+1} under the alternative model can be represented as conditional expectatons of $N_{t+1}B_{t+1}$ under the baseline model:

$$E\left(N_{t+1}B_{t+1} \mid \mathfrak{A}_t\right). \tag{6.1}$$

Thus, mutliplication of B_{t+1} serves in effect to change the baseline probability from the baseline model to the alternative model. To serve this purpose the random variable N_{t+1} must satisfy:

i) $N_{t+1} \geq 0$;

ii) $E\left(N_{t+1} \mid \mathfrak{A}_t\right) = 1$;

iii) N_{t+1} is $\mathfrak{A}_{t+1}$ measurable.

Property i is satisfied because conditional expectations map positive random variables B_{t+1} into positive random variables that are $\mathfrak{A}_t$ measurable. Property ii is satisfied because conditional expectations of random variables B_{t+1} that are $\mathfrak{A}_t$ measurable should equal B_{t+1}. Property iii can be imposed without loss of generality because if it were not satisfied, we could just replace it with $E\left(N_{t+1} \mid \mathfrak{A}_{t+1}\right)$.

This way of representing an alternative probability model is restrictive. Thus, if a nonnegative random variable has conditional expectation zero under the baseline probability, it will also have zero conditional expectation under the alternative probability measure, a version of absolute continuity here applied to transition probabilities. Violating absolute continuity would make possible model decision rules that correctly select models with full confidence from only finite samples.

Example 6.1.1. *Consider a baseline Markov process having transition probability density π_o with respect to a measure λ over the state space $\mathcal{X}$*

$$P_o(dx^+|x)\lambda(dx^+) = \pi_o(x^+ \mid x)\lambda(dx^+)$$

Let π denote some other transition density that we represent as

$$\pi(x^+ \mid x)\lambda(dx^+) = \left[\frac{\pi(x^+ \mid x)}{\pi_o(x^+ \mid x)}\right] \pi_o(x^+ \mid x)\lambda(dx^+)$$

where we assume that $\pi_o(x^+ \mid x) = 0$ implies that $\pi(x^+ \mid x) = 0$ for all x^+ and x in $\mathcal{X}$. Construct the likelihood ratio

$$N_{t+1} = \frac{\pi(X_{t+1} \mid X_t)}{\pi_o(X_{t+1} \mid X_t)}.$$

Example 6.1.2. *Suppose that*

$$\begin{aligned} X_{t+1} &= \mathbb{A}X_t + \mathbb{B}W_{t+1} \\ Z_{t+1} &= \mathbb{D}X_t + \mathbb{F}W_{t+1}, \end{aligned}$$

where $\mathbb{A}$ is a stable matrix, $\{W_{t+1}\}_{t=0}^{\infty}$ is an i.i.d. sequence of $\mathcal{N}(0, I)$ random vectors conditioned on X_0, and $\mathbb{F}$ is a nonsingular square matrix. The conditional distribution of Z_{t+1} is normal with mean $\mathbb{D}X_t$ and nonsingular covariance matrix $\mathbb{F}\mathbb{F}'$. We suppose that $\mathbb{A}$ and $\mathbb{B}$ can be constructed as functions of $\mathbb{D}$ and $\mathbb{F}$.

Since $\mathbb{F}$ is nonsingular, the following recursion connects state and observation vectors:

$$X_{t+1} = \left(\mathbb{A} - \mathbb{B}\mathbb{F}^{-1}\mathbb{D}\right) X_t + \mathbb{B}\mathbb{F}^{-1} Z_{t+1}$$

If $(\mathbb{A} - \mathbb{B}\mathbb{F}^{-1}\mathbb{D})$ is a stable matrix, we can construct X_{t+1} as a linear functon of $Z_{t+1-\tau}$ for $\tau = 0, 1, \ldots$.

Assume a baseline model that has the same functional form with particular settings of the parameters that appear in the matrices $(\mathbb{A}_o, \mathbb{B}_o, \mathbb{D}_o, \mathbb{F}_o)$. Let N_{t+1} be the one-period conditional log-likehood ratio

$$\begin{aligned} \log N_{t+1} = &-\frac{1}{2}(Z_{t+1} - \mathbb{D}X_t)' \left(\mathbb{F}\mathbb{F}'\right)^{-1} (Z_{t+1} - \mathbb{D}X_t) \\ &+ \frac{1}{2}\left(Z_{t+1} - \mathbb{D}_o X_t\right)' \left(\mathbb{F}_o{\mathbb{F}_o}'\right)^{-1} \left(Z_{t+1} - \mathbb{D}_o X_t\right) \\ &- \frac{1}{2}\log\det\left(\mathbb{F}\mathbb{F}'\right) + \frac{1}{2}\log\det\left(\mathbb{F}_o{\mathbb{F}_o}'\right) \end{aligned}$$

Notice how we have subtracted components coming from the baseline model.

6.2 Likelihood Ratio Processes

The random variable N_{t+1} contains the new information in observation Z_{t+1} that is relevant for comparing an alternative statistical model to a baseline model. As data arrive, information accumulates in a way that we describe by compounding the process $\{N_{t+1} : t \geq 0\}$:

$$L_{t+1} = \prod_{j=0}^{t} N_{j+1}$$

so that

$$\log L_{t+1} = \sum_{j=0}^{t} \log N_{j+1}$$

Being functions of a stochastic process of observations $\{Z_{t+1} : t \geq 0\}$, the *likelihood ratio* and *log-likelihood ratios* sequences are both stochastic processes.

Fact 6.2.1. *Since $E\left(N_{t+1} \mid \mathfrak{A}_t\right) = 1$, a likelihood ratio process satisfies*

$$E\left(L_{t+1} \mid \mathfrak{A}_t\right) = L_t.$$

Therefore, it is a martingale *relative to the information sequence $\{\mathfrak{A}_t : t \geq 0\}$.*

Fact 6.2.2. *A log-likelihood ratio process $\{\log(L_{t+1}) : t = 0, 1, \ldots, t\}$ is a stationary increment process with increment*

$$\log L_{t+1} - \log L_t = \log N_{t+1}.$$

The log likelihood process is additive in how it accumulates stationary increments $\log N_{t+1}$. Consequently, the likelihood ratio process is what we call a multiplicative process.

Our next fact uses Jensen's inequality for the concave function $\log(N)$ illustrated in Figure 6.1.

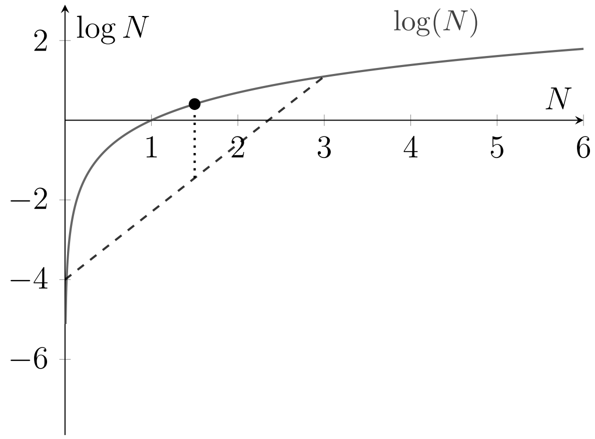


Figure 6.1: Jensen's Inequality. The logarithmic function is the concave function that equals zero when evaluated at unity. An interior average of the endpoints of the straight line lies below the logarithmic function. Jensen's Inequality asserts that the line segment lies below the logarithmic function.

Fact 6.2.3. *By Jensen's inequality,*

$$E\left(\log N_{t+1} \mid \mathfrak{A}_t\right) \le \log E\left(N_{t+1} \mid \mathfrak{A}_t\right) = 0,$$

where the mathematical expectation is again under the baseline model parameterized by θ_o*. Thus*

$$E\left(\log L_{t+1} \mid \mathfrak{A}_t\right) \le \log L_t.$$

This implies that that under the baseline model the log-likelihood ratio process is a super martingale *relative to the information sequence* $\{\mathfrak{A}_t : t \ge 0\}$.

Notice that if N_{t+1} is not identically one, then

$$E\left(\log N_{t+1}\right) < 0.$$

From the Law of Large Numbers, the population mean is well approximated by a sample average from a long time series. That opens the door to discriminating between two models. Under the baseline model, the log likelihood ratio process scaled by the inverse of the sample size $t + 1$ converges to a negative number. After changing roles of the baseline and alternative models, we can do an analogous calculation that entails using $\frac{1}{N_{t+1}}$ instead of N_{t+1} as an increment. Then the scaled-by-$(t+1)^{-1}$ log likelihood ratio would converge to the expectation of $-\log N_{t+1}$ under the alternative model that is now in the denominator of the likelihood ratio. This limit would be positive under the assumption that the alternative model generated the data. These calculations justify selecting between the two models by calculating $\log L_{t+1}$ and checking if it is positive or negative. This procedure amounts to a special case of the method of maximum likelihood.

Remark 6.2.4. *Suppose that data are not generated by the baseline model. Instead, suppose that the statistical model implied by the change of measure* N_{t+1} *governs the stochastic evolution of the observations. Define* ***conditional entropy*** *relative to baseline model* θ_o *as the following conditional expectation:*

$$E\left(N_{t+1}\log N_{t+1} \mid \mathfrak{A}_t\right).$$

Here multiplication of $\log N_{t+1}$ *by* N_{t+1} *changes the conditional probability distribution from the misspecified baseline model to the alternative statistical model that we assume generates the data. The function* $n\log n$ *is convex and equal to zero for* $n = 1$*. Therefore, Jensen's inequality implies that conditional relative entropy is nonnegative and equal to zero when* $N_{t+1} = 1$*. An unconditional counterpart of relative entropy is the Large of Large Numbers limit*

$$\lim_{t\to+\infty}\frac{1}{t+1}\sum_{j=0}^{t}\log N_{t+1} = \lim_{t\to+\infty}\frac{1}{t+1}\sum_{j=0}^{t}E\left(N_{t+1}\log N_{t+1} \mid \mathfrak{A}_t\right) \geq 0$$

under the data generating process. Relative entropy is often used to analyze model misspecifications. It is also a key component for studying the statistical theory of "large deviations" for Markov processes, as we shall discuss later.

Bayes' law and likelihood ratio processes

Suppose now that we attach a prior probability π_o to the baseline model with probability $1-\pi_o$ on the alternative. Then after observing $Z_{j+1} : 0 \leq j \leq t$, conditional probabilities for the baseline and alternative models are

$$\frac{\pi_o}{L_{t+1}(1-\pi_o)+\pi_o} \quad \text{and} \quad \frac{L_{t+1}(1-\pi_o)}{L_{t+1}(1-\pi_o)+\pi_o}.$$

When $\frac{1}{t+1}\log L_{t+1}$ converges to a negative number, the first probability converges to one, and when $\frac{1}{t+1}\log L_{t+1}$ converges to a positive number, it converges to zero.

6.3 Parameterizing Likelihoods

Let Θ be a set of parameter vectors. Each $\theta \in \Theta$ indexes an alternative transition probability as represented by the $N_{t+1}(\theta)$ that belongs in formula (6.1). We presume that a particular θ, denoted θ_o, indexes the transition probability that generates the data and is used as to calculate the conditional expectation in formula (6.1). Accordingly, $N_{t+1}(\theta_o) = 1$. Since $N_{t+1}(\theta)$ is a likelihood increment, a recursion that defines a likelihood ratio process for each θ is

$$L_{t+1}(\theta) = N_{t+1}(\theta)L_t(\theta)$$

Setting $L_0(\theta) = 1$ for each θ completes a parameterized family of likelihood ratios.

But the applied researcher does not know θ_o. For that reason, it is convenient now to use a model with some arbitrary known parameter vector $\tilde{\theta} \in \Theta$ as a baseline model in place of the θ_o model. We can accomplish this by defining an increment process $\widetilde{N}_{t+1}(\theta)$ as

$$\widetilde{N}_{t+1}(\theta) = \frac{N_{t+1}(\theta)}{N_{t+1}(\tilde{\theta})}.$$

Notice that when we use $\widetilde{N}_{t+1}(\theta)$ in formula (6.1), we must also change the transition probability used to take the expectation in (6.1) to be the transition probability implied by $\tilde{\theta}$. This is evident because $\widetilde{N}_{t+1}(\tilde{\theta}) = 1$. Our change of baseline model leads us now to construct likelihood ratios with the recursion:

$$\widetilde{L}_{t+1}(\theta) = \widetilde{N}_{t+1}(\theta)\widetilde{L}_t(\theta),$$

where we set $\widetilde{L}_0(\theta) = 1$. In this way, we construct parameterized likelihoods without knowing the θ_o model that generates the data.

In order to apply the Law of Large Numbers to the logarithm of the likelihood ratio process divided by $t + 1$, namely to

$$\frac{1}{t+1}\log \widetilde{L}_{t+1}(\theta) = \frac{1}{t+1}\sum_{j=0}^{t}\log \widetilde{N}_{j+1}(\theta)$$

for $t \geq 0$, we want to compute expectations under the θ_o model that actually generates the data. Under the θ_o model's expectation operator

$$\tilde{\nu}(\theta) = E\left[\log \widetilde{N}_{t+1}(\theta)\right] = E\left[\log N_{t+1}(\theta)\right] - E\left[\log N_{t+1}(\tilde{\theta})\right] = \nu(\theta) - \nu(\tilde{\theta}).$$

Maximum Likelihood

We want the law of large numbers from Chapter 1 eventually to disclose the parameter vector θ_o. The following argument shows that it will. The law of large numbers leads us to expect that

$$\lim_{t\to+\infty} \frac{1}{1+t} \log \widetilde{L}_{t+1} = \tilde{\nu}(\theta).$$

Since θ_o generates the data, the super martingale property of the log likelihood ratio process implies that

$$\nu(\theta) \leq \nu(\theta_o) = 0.$$

Therefore

$$\tilde{\nu}(\theta) = \nu(\theta) - \nu(\tilde{\theta}) \leq \nu(\theta_o) - \nu(\tilde{\theta}) = \tilde{\nu}(\theta_o).$$

This implies that θ_o is a maximizer of $\tilde{\nu}(\theta)$, and gives a "population counterpart" to maximum likelihood estimation. By a population counterpart we imagine a setting in which via the law of large numbers, sample averages have converged to their population counterparts. Formally, a population counterpart to maximum likelihood estimation solves:

$$\max_{\theta\in\Theta} \tilde{\nu}(\theta).$$

We have shown that the set of θ's that solve $\max_{\theta\in\Theta} \nu(\theta)$ includes θ_o. We say that the model is **identified** if θ_o is the unique maximizer. The maximum likelihood estimator from a finite data sample uses a sample counterpart of the above equation, namely,

$$\text{argmax}_{\theta\in\Theta} \frac{1}{t+1} \log \widetilde{L}_{t+1}(\theta).$$

Remark 6.3.1. *(Reverse relative entropy) Continuing to index conditional distributions by parameter vectors $\theta \in \Theta$, form the ratio*

$$N^o_{t+1}(\theta) = \frac{\widetilde{N}_{t+1}(\theta_o)}{\widetilde{N}_{t+1}(\theta)}$$

Using a version of formula (6.1), we can use this ratio of likelihood increments to represent the transition distribution for statistical model θ_o relative to that for an arbitrary statistical model θ. The ratio of likelihood increments effectively changes the baseline model from $\tilde{\theta}$ to an arbitrary θ. Then the associated unconditional relative entropy defined in remark 6.2.4 becomes

$$D(\theta) = \lim_{t \to +\infty} \frac{1}{t+1} \sum_{j=0}^{t} \log N^o_{t+1}(\theta) \geq 0$$

where the θ_o model generates the data. We can then express the population counterpart to maximum likelihood as the solution to a minimum relative entropy problem. Thus, since the model indexed by parameter vector θ_o generates the data, the population maximum likelihood estimator solves

$$\min_{\theta \in \Theta} D(\theta) = 0.$$

6.4 Score Process

We assume that parameter vector θ_o in the interior of a parameter space Θ. Moreover, for each $\theta \in \Theta$,

$$E\left(N_{t+1}(\theta) \mid \mathfrak{A}_t\right) = 1$$

where $N_{t+1}(\theta_o) = 1$ by construction. Provided that we can differentiate inside the mathematical expectation:[1]

$$E\left(\frac{\partial N_{t+1}}{\partial \theta}\Big|_{\theta=\theta_o} \mid \mathfrak{A}_t\right) = 0.$$

Since expectations are taken under the θ_o probability

$$\frac{\partial N_{t+1}}{\partial \theta}\Big|_{\theta=\theta_o} = \frac{\partial \log N_{t+1}}{\partial \theta}\Big|_{\theta=\theta_o}$$

Since the right side is a logarithmic derivative

$$\frac{\partial \log N_{t+1}}{\partial \theta}\Big|_{\theta=\theta_o} = \frac{\partial \log \widetilde{N}_{t+1}}{\partial \theta}\Big|_{\theta=\theta_o}$$

where we used $\widetilde{N}_{t+1}(\theta)$ to build an operational likelihood process for alternative $\theta \in \Theta$.

[1] Formally, we define the derivative of a family $\{\log N_{t+1}(\theta) : \theta \in \Theta\}$ in terms of mean square limits, and we let $\frac{\partial N_{t+1}}{\partial \theta} = N_{t+1} \frac{\partial N_{t+1}}{\partial \theta}$.

Definition 6.4.1. *The* ***score increment*** *is*

$$S_{t+1} - S_t = \frac{\partial \log N_{t+1}}{\partial \theta}\Big|_{\theta=\theta_o}$$

and the ***score process*** *is*

$$S_{t+1} = \frac{\partial}{\partial \theta} \log L_{t+1}\Big|_{\theta=\theta_o}$$

Fact 6.4.2. *The score process* $\{S_{t+1} : t \geq 0\}$ *is a (multivariate) martingale with stationary increments. Consequently*

$$\frac{1}{\sqrt{t}} S_{t+1} \Rightarrow \mathcal{N}(0, \mathbb{V})$$

where $\mathbb{V} = E\left[(S_{t+1} - S_t)(S_{t+1} - S_t)'\right]$.

Fact 6.4.2 motivates characterizing the large sample behavior of the score process by utilizing the martingale central limit theorem stated in Proposition 2.3.1. In effect, the matrix $\mathbb{V}$ measures curvature of the log-likelihood process in the neighborhood of the "true" parameter value θ_o. The more curvature there is – i.e., the "larger" is the variance matrix $\mathbb{V}$ of the score vector – the more information the data contain about θ. The matrix $\mathbb{V}$ is called the Fisher information matrix in honor of R.A. Fisher.

An associated central limit approximation yields a large sample characterization of the maximum likelihood estimator of θ in a Markov setting. Let θ_t maximize the log-likelihood function $\log L_t(\theta)$. Under some regularity conditions

$$\sqrt{t}(\theta_t - \theta_o) \to \mathcal{N}\left(0, \mathbb{V}^{-1}\right).$$

This limit justifies interpreting the covariance matrix $\mathbb{V}$ of the martingale increment of the score process as quantifying information that data contain about the parameter vector θ_o.

6.5 Nuisance parameters

Consider a situation in which to learn about one parameter, we have to estimate other parameters too. Suppose that θ is a vector and $\mathbb{V}$ is a matrix. We seek a notion of "Fisher information" about a single component of θ that interests us – a single parameter $\bar{\theta}$. A natural guess might be simply to take as our measure the appropriate diagonal entry of $\mathbb{V}$. It turns out that this measure of our uncertainty is misleading because it ignores the fact that in order to estimate the parameter of interest to us we have to "spend" some of the information in the sample to estimate "nuisance parameters" that we also had to estimate in order make inferences about $\bar{\theta}$. It turns out that a better way to summarize our uncertainty about the parameter of interest is to define its "Fisher information" as the reciprocal of an appropriate entry of $\mathbb{V}^{-1}$.

Thus, partition

$$\theta = \begin{bmatrix} \bar{\theta} \\ \tilde{\theta} \end{bmatrix}$$

where $\bar{\theta}$ is the scalar parameter of interest and $\tilde{\theta}$ is an associated unknown nuisance parameter vector. Write the multivariate score process as

$$\left\{ \begin{bmatrix} \overline{S}_{t+1} \\ \widetilde{S}_{t+1} \end{bmatrix} : t = 0, 1, ... \right\}$$

Partition the covariance matrix $\mathbb{V}$ of the score process increment conformably with $(\bar{\theta}', \tilde{\theta}')'$:

$$\begin{bmatrix} E\left(\overline{S}_{t+1} - \overline{S}_t\right)^2 & E\left(\overline{S}_{t+1} - \overline{S}_t\right)\left(\widetilde{S}_{t+1} - \widetilde{S}_t\right)' \\ E\left(\widetilde{S}_{t+1} - \widetilde{S}_t\right)\left(\overline{S}_{t+1} - \overline{S}_t\right) & E\left(\widetilde{S}_{t+1} - \widetilde{S}_t\right)\left(\widetilde{S}_{t+1} - \widetilde{S}_t\right)' \end{bmatrix} \equiv \begin{bmatrix} \mathbb{V}_{11} & \mathbb{V}_{12} \\ \mathbb{V}_{21} & \mathbb{V}_{22} \end{bmatrix}.$$

We claim that taking $E\left(\overline{S}_{t+1} - \overline{S}_t\right)^2$ as a measure of "Fisher information" about $\bar{\theta}$ would overstate our information. Instead, the appropriate Fisher information about $\bar{\theta}$ is the inverse of the $(1,1)$ component of the asymptotic covariance matrix $\mathbb{V}^{-1}$. Applying a partitioned inverse formula for a symmetric matrix to compute that measure of Fisher information yields

$$I_{\bar{\theta}} = \mathbb{V}_{11} - \mathbb{V}_{12}\mathbb{V}_{22}^{-1}\mathbb{V}_{12}'. \tag{6.2}$$

An enlightening interpretation of the $(1,1)$ component $I_{\bar{\theta}}$ of $\mathbb{V}^{-1}$ comes from recognizing that it is the residual variance of a population least squares regression of the score vector increment of $\bar{\theta}$ on the score vector increment for the nuisance parameter vector $\widetilde{\theta}$. Thus, a population least squares regression is

$$\overline{S}_{t+1} - \overline{S}_t = \beta(\widetilde{S}_{t+1} - \widetilde{S}_t) + U_{t+1}, \tag{6.3}$$

where β is a population regression coefficient vector and U_{t+1} is a population regression residual that by construction is orthogonal to the regressor $(\widetilde{S}_{t+1} - \widetilde{S}_t)$. The least squares regression coefficient vector is

$$\beta = \mathbb{V}_{12}\mathbb{V}_{22}^{-1}$$

and the residual variance is

$$EU_{t+1}^2 = \mathbb{V}_{11} - \mathbb{V}_{12}\mathbb{V}_{22}^{-1}\mathbb{V}_{12}'$$

which equals the Fisher information measure $I_{\bar{\theta}}$ defined above. From the orthogonality of least squares residuals to regressors, the variability of the left side variable $\overline{S}_{t+1} - \overline{S}_t$ in the projection equation (6.3) cannot exceed that of the least squares residual so that

$$E(U_{t+1})^2 \leq E\left(\overline{S}_{t+1} - \overline{S}_t\right)^2,$$

an inequality that confirms that information about $\bar{\theta}$ is lost by not knowing the nuisance parameters $\tilde{\theta}$.

Chapter 7

GMM estimation

7.1 Formulation

We study a family of GMM estimators of an unknown parameter vector β constructed from theoretical restrictions on conditional or unconditional moments of functions φ. The functions φ depend on an unknown parameter vector β and on a random vector X_t that is observable to an econometrician and has expectation zero. This property opens the door to the construction of estimating equations to be used in constructing an estimator b_N of β and making inferences.

Data generation

For much of the analysis in this chapter implicitly "conditions on a model" of the data generation. This data generation is presumed to be stationary and ergodic. We do not presume that this model is known to the investigator, which would make the analysis uninteresting. In many settings that interest us, the parameter vector β incompletely characterizes that statistical model. This latter feature is important, as the methods we consider only presume that the economic model is "partially specified." This is meant to apply to situations in which a researcher "wishes to do something without having to do everything." In contrast, likelihood and Bayesian methods require a full specification of the data generating process. Formally, we implement a version of what is known as *semi-parametric* estimation: while β is a finite-dimensional parameter vector that we want to estimate, we acknowledge that, in addition to β, a potentially infinite dimensional nuisance parameter vector might be required to pin down the complete statistical model on which we condition when we apply the law of large numbers and central limit theorems. For the estimation problems that we consider, the nuisance parameter vector needed to complete model specification is left in the background. We will come back to the formal semi-parametric interpretation later in this chapter when we discuss statistical efficiency.

Restrictions on the data generation

As a starting point, we consider a class of restrictions large enough to include examples of both the conditional and the unconditional moment restrictions that interest us. Members of this class take the form

$$E\left[{A_t}'\varphi(X_t, b)\right] = 0 \text{ if and only if } b = \beta \tag{7.1}$$

for all sequences of selection matrices $A \in \mathcal{A}$ where $A = \{A_t : t \geq 1\}$ and where

- the vector of functions φ is r-dimensional.
- the unknown parameter vector β is k-dimensional, as is b.
- $\mathcal{A}$ is a collection of time series of (possibly random) selection matrices characterizing valid moment restrictions.
- A_t denotes a time t $r \times k$ selection matrix for a subset of the valid moment restrictions that is used to construct a particular statistical estimator b of β.
- the mathematical expectation is taken with respect to the statistical model that generates the $X = \{X_t : t \geq 1\}$ process.

Applying a Law of Large Numbers to the population moment condition (7.1) motivates a "generalized method of moments" b_N estimator of the $k \times 1$ vector β that solves the following k equations:

$$\frac{1}{N}\sum_{t=1}^{N} {A_t}'\varphi(X_t, b_N) = 0.$$

Different processes of selection matrices $\{A_t : t \geq 1\}$ and $\{\widetilde{A}_t : t \geq 1\}$ typically give rise to different properties for the estimator $\{b_N\}$, but in some cases they do not. For instance, suppose that

$$\widetilde{A}_t = A_t \mathbb{K}$$

for some $k \times k$ nonsingular matrix $\mathbb{K}$. Although the selection matrices $\widetilde{A}_t$ and A_t could be distinct, the set of moment conditions used to identify and estimate β are effectively the same. Satisfying (7.1) for A is equivalent satisfying (7.1) for $\widetilde{A}$.

Example 7.1.1. ***Unconditional moment restrictions*** *Suppose that*

$$E\left[\varphi(X_t, \beta)\right] = 0$$

where $r \geq k$. Let $\mathcal{A}$ be the set of all constant (time invariant) $r \times k$ matrices $\mathbb{A}$. Rewrite the restrictions as:

$$\mathbb{A}' E\left[\varphi(X_t, \beta)\right] = 0$$

for all $r \times k$ matrices $\mathbb{A}$. Sargan (1958) and Hansen (1982) assumed moment restrictions like these.

Example 7.1.2. ***C****onditional moment restrictions Assume the conditional moment restictions*

$$E\left[\varphi(Y_t, \beta) \mid \mathfrak{J}_{t-\ell}\right] = 0$$

for a particular $\ell \geq 1$ and $Y_t = X_t$. Let $\mathcal{A}_t$ be the set of all $r \times k$ matrices, A_t, of bounded random variables that are $\mathfrak{J}_{t-\ell}$ measurable. Then the preceding conditional moment restrictions are mathematically equivalent to the unconditional moment restrictions

$$E\left[{A_t}' \varphi(Y_t, \beta)\right] = 0$$

for all random matrices $A_t \in \mathcal{A}_t$. This formulation is due to Hansen (1985) and closely related to analysis of Chamberlain (1987).

It is common in practice to use the idea provided in Example 7.1.2 while substantially restricting the set of moment conditions considered for estimation. Specifically, we take a collection of conditional moment restrictions and from them create unconditional moment restrictions like those in Example of 7.1.1. In this way we can reduce the class of GMM estimators under consideration.

Example 7.1.3. *Let $A_t^1, A_t^2, ..., A_t^m$ be m* ad hoc *choices of selection matrices. Form*

$$\varphi^+(X_t, b) = \begin{bmatrix} A_t^{1\prime} \\ A_t^{2\prime} \\ ... \\ A_t^{m\prime} \end{bmatrix} \varphi(X_t, b)$$

where X_t now includes variables used to construct A_t^j and A_t^2. We presume that no linear combination of columns of any A_t^j duplicate any columns of the A_t's. Otherwise, we would omit such columns and adjust φ^+ accordingly. Let $r^+ \geq r$ denote the remaining non-redundant columns.

$$\mathbb{A}' E \left(\varphi^+(X_t, b) \right] = 0$$

and study an associated family of GMM estimators. This strategy reduces the moment conditions from an infinite to a finite dimensional collection as in Example 7.1.1.[1]

Example 7.1.4. *"Moment matching" is another special case of Example 7.1.1. Suppose that*

$$\varphi(X_t, b) = \psi(X_t) - \kappa(b)$$

where

$$E\left[\psi(X_t)\right] = \kappa(\beta).$$

Here $\psi(Y)$ defines moments to be matched and $\kappa(b)$ gives model-predicted moments as functionals of a parameter vector b. The function κ is often computed by simulating the model for alternative values of parameter vector β. See Lee and Ingram (1991) *and* Duffie and Singleton (1993).[2] *In contrast to other applications of GMM estimation, this one presumes that, given b, the model completely determines the simulated data. The method is applied either for reasons computational simplicity or because the research wants to focus on moments believed to be robust to model misspecification.*

Collections $\mathcal{A}$ of selection processes for all of these examples satisfy the following "linearity" restriction.

[1] More generally, construct φ^+ using columns from alternative selection matrices.

[2] Important related approaches use a misspecified maximum likelihood (Smith (1993) and Gourieroux et al. (1993)) or the score increment of of such a likelihood (Gallant and Tauchen (1996)) to summarize empirical evidence and use model simulation to account for the misspecification.

Restriction 7.1.5. *If A^1 and A^2 are both in $\mathcal{A}$ and $\mathbb{J}_1$ and $\mathbb{J}_2$ are $k \times k$ matrices of real numbers, then $A^1\mathbb{J}_1 + A^2\mathbb{J}_2$ is in $\mathcal{A}$.*

7.2 Central limit approximation

The process

$$\left\{ \sum_{t=1}^{N} {A_t}'\varphi(X_t, \beta) : N \geq 1 \right\}.$$

can be verified to have stationary and ergodic increments conditioned on the statistical model. So there exists a Proposition 2.2.2 decomposition of the process. Provided that

$$E\left[{A_t}'\varphi(X_t, \beta)\right] = 0$$

under the statistical model that generates the data, the trend term in the decomposition of Proposition 2.2.2 is zero, implying that the martingale dominates the behavior of sample averages for large N. In particular, Proposition 2.3.1 gives a central limit approximation for

$$\frac{1}{\sqrt{N}} \sum_{t=1}^{N} {A_t}'\varphi(X_t, \beta)$$

provided that we restrict the family of selection matrices.

Restriction 7.2.1. *For any $A \in \mathcal{A}$,*

$$E\left[\sum_{j=0}^{\infty} {A_{t+j}}'\varphi(X_{t+j}, \beta) \mid \mathfrak{J}_t\right]$$

converges in mean square.

Define the one-step-ahead forecast error:

$$G_t(A) = E\left[\sum_{j=0}^{\infty} {A_{t+j}}'\varphi(X_{t+j}, \beta) \mid \mathfrak{J}_t\right] - E\left[\sum_{j=0}^{\infty} {A_{t+j}}'\varphi(X_{t+j}, \beta) \mid \mathfrak{J}_{t-1}\right]$$

Paralleling the construction of the martingale increment in Proposition 2.2.2,

$$\frac{1}{\sqrt{N}} \sum_{t=1}^{N} {A_t}'\varphi(X_t, \beta) \approx \frac{1}{\sqrt{N}} \sum_{t=1}^{N} G_t(A)$$

where by the approximation sign $\approx$ we intend to assert that the difference between the right side and left side converges in mean square to zero as $N \to \infty$. Consequently, the covariance matrix in the central limit approximation is $E\left[G_t(A)G_t(A)'\right]$.

Recall Restriction 7.1.5. For the preceding construction of the martingale increment, it is straightforward to verify that

$$G_t(A^1\mathbb{J}_1 + A^2\mathbb{J}_2) = (\mathbb{J}_1)'G_t(A^1) + (\mathbb{J}_1)'G_t(A^2)$$

follows from the linearity of conditional expectations.

Example 7.2.2. *Consider again example 7.1.1 in which $A_t = \mathbb{A}$ for all $t \geq 0$ and*

$$G_t(A) = \mathbb{A}'F_t$$

where

$$F_t = E\left[\sum_{j=0}^{\infty} \varphi(X_{t+j}, \beta) \mid \mathfrak{J}_t\right] - E\left[\sum_{j=0}^{\infty} \varphi(X_{t+j}, \beta) \mid \mathfrak{J}_{t-1}\right].$$

Define the covariance matrix

$$\mathbb{V} = E\left(F_t {F_t}'\right)$$

and note that

$$E\left[G_t(A)G_t(A)'\right] = \mathbb{A}'\mathbb{V}\mathbb{A}.$$

Example 7.2.3. *In Example 7.1.2*

$$E\left[\varphi(Y_t, \beta) \mid \mathfrak{J}_{t-\ell}\right] = 0$$

and hence

$$E\left[{A_t}'\varphi(Y_t, \beta) \mid \mathfrak{J}_{t-\ell}\right] = 0$$

whenever entries of A_t are restricted to be $\mathfrak{J}_{t-\ell}$ measurable. Consequently

$$E\left[{A_{t+j}}'\varphi(Y_{t+j}), \beta) \mid \mathfrak{J}_t\right] = 0$$

for $j \geq \ell$ so that the infinite sums used to construct $G_t(A)$ simplify to finite sums.

7.3 Mean value approximation

Write

$$\begin{aligned}\frac{1}{\sqrt{N}}\sum_{t=1}^{N} A_t{}'\varphi(X_t, b_N) \approx & \frac{1}{\sqrt{N}}\sum_{t=1}^{N} A_t{}'\varphi(X_t, \beta) \\ & + \frac{1}{N}\sum_{t=1}^{N} A_t{}'\left[\frac{\partial\varphi}{\partial b'}(X_t, \beta)\right]\sqrt{N}(b_N - \beta) \\ \approx & \frac{1}{\sqrt{N}}\sum_{t=1}^{N} A_t{}'\varphi(X_t, \beta) + \nabla(A)'\sqrt{N}(b_N - \beta)\end{aligned}$$

where

$$\nabla(A) \doteq E\left(\left[\frac{\partial\varphi}{\partial b'}(X_t, \beta)\right]' A_t\right)$$

$$\frac{1}{\sqrt{N}}\sum_{t=1}^{N} A_t{}'\varphi(X_t, b_N) \approx 0,$$

$$\nabla(A)'\sqrt{N}(b_N - \beta) \approx -\frac{1}{\sqrt{N}}\sum_{t=1}^{N} A_t{}'\varphi(X_t, \beta)$$

So long as $\nabla(A)$ is nonsingular,

$$\sqrt{N}(b_N - \beta) \approx -\left[\nabla(A)'\right]^{-1}\frac{1}{\sqrt{N}}\sum_{t=1}^{N} A_t{}'\varphi(X_t, \beta).$$

This approximation underlies an "efficiency bound" for GMM estimation. Notice that the covariance matrix in a central limit approximation is:

$$\mathbf{cov}(A) = \left[\nabla(A)'\right]^{-1} E\left[G_t(A)G_t(A)'\right]\left[\nabla(A)\right]^{-1}$$

We want to know how small we can make this matrix by choosing a selection process.

Example 7.3.1. *Consider again Example of 7.1.1. In this case $A_t = \mathbb{A}$ for all $t \geq 0$ and*

$$\nabla(A) = \mathbb{D}'\mathbb{A}$$

where

$$\mathbb{D} \doteq E\left[\frac{\partial \varphi}{\partial b'}(X_t, \beta)\right]$$

and

$$\boldsymbol{cov}(\mathbb{A}) = (\mathbb{A}'\mathbb{D})^{-1}\, \mathbb{A}'\mathbb{V}\mathbb{A}\, (\mathbb{D}'\mathbb{A})^{-1}$$

7.4 GMM Efficiency Bound

Recall

$$\mathbf{c}\mathrm{ov}(A) = [\nabla(A)']^{-1}\, E\left[G_t(A)G_t(A)'\right] [\nabla(A)]^{-1}$$

We seek a greatest lower bound on the covariance matrix on the right.

i) Suppose that $[\nabla(A)']^{-1}$ is nonsingular and impose that

$$[\nabla(A)] = \mathbb{I}$$

If not post multiply A by a nonsingular matrix $\mathbb{K}$. That leaves the GMM estimator unaltered. Thus, we have

$$\mathbf{cov}(A) = E\left[G_t(A)G_t(A)'\right]$$

subject to $[\nabla(A)] = \mathbb{I}$

ii) Find an A^d such that for all $A \in \mathcal{A}$

$$\nabla(A) = E\left[G_t(A^d)G_t(A)'\right].$$

iii) Form

$$A_t^* = A_t^d \left(E\left[G_t(A^d)G_t(A^d)'\right]\right)^{-1}$$

for all $A \in \mathcal{A}$. These form a set of first-order sufficient conditions for our constrained minimization problem. Then

$$G_t(A^*) = \left(E\left[G_t(A^d)G_t(A^d)'\right]\right)^{-1} G_t(A^d)$$

and

$$E\left[G_t(A^*)G_t(A)'\right] = \left(E\left[G_t(A^d)G_t(A^d)'\right]\right)^{-1}$$

provided that $[\nabla(A)] = \mathbb{I}$.

iv) Therefore,

$$\begin{aligned}0 &\leq E\left([G_t(A) - G_t(A^*)]\,[G_t(A) - G_t(A^*)]'\right)\\ &= \mathbf{cov}(A) - \mathbf{cov}(A^*)\\ &= \mathbf{cov}(A) - \left(E\left[G_t(A^d)G_t(A^d)'\right]\right)^{-1}.\end{aligned}$$

Result 7.4.1. *Given a solution to equation* (ii)

$$\inf_{A \in \mathcal{A}} \boldsymbol{cov}(A) = \left(E\left[G_t(A^d)G_t(A^d)'\right]\right)^{-1} \tag{7.2}$$

Remark 7.4.2. *In the result 7.4.1 efficiency bound, we might be tempted to think that $G_t(A^d)$ plays the same role that the "score vector" increment does in maximum likelihood estimation. But because there is potentially a set of infinite dimensional nuisance parameters here, a better analogy is that $G_t(A^d)$ acts much like the residual vector in a regression of the score increments for parameters of interest on score increments of nuisance parameters. By taking conditional or unconditional moment restrictions as the starting point for estimation of parameter vector β, we have purposefully pushed all nuisance parameters into the background.*

Remark 7.4.3. *Consider two GMM estimators, one with a selection process A and the other with A^*. Transform A :*

$$\widetilde{A} = A\mathbb{K}$$

and choose $\mathbb{K}$ so that

$$\nabla\left(\widetilde{A}\right) = \nabla\left(A\right)\mathbb{K} = I. = I$$

Thus $\mathbb{K} = \left[\nabla(A) \doteq E\left(\left[\frac{\partial\varphi}{\partial b'}(X_t, \beta)\right]' A_t\right)\right]^{-1}$. Since the selection processes are asymptotically equivalent, we may use $\widetilde{A}$ to characterize the limiting distribution of the corresponding GMM estimator. Let $\{b_T : T \geq 1\}$ denote the corresponding GMM and let $\{b_T^* : T \geq 1\}$ be the asymptotically efficient GMM estimator. Then

$$\sqrt{N}\left(b_T - b_T^*\right) \approx \frac{1}{\sqrt{T}} \sum_{t=t}^{N} G_t\left(\widetilde{A}\right) - G_t\left(A^*\right)$$

with a limiting covariance matrix

$$E\left([G_t(A) - G_t(A^*)]\,[G_t(A) - G_t(A^*)]'\right) = \boldsymbol{cov}(A) - \boldsymbol{cov}(A^*)$$

Example 7.4.4. *Consider Example 7.1.1 in which we assumed that $A_t = \mathbb{A}$. Then*

$$\mathbb{A}'\mathbb{V}\mathbb{A}^d = \mathbb{A}'\mathbb{D}.$$

Therefore,

$$\mathbb{A}^d = \mathbb{V}^{-1}\mathbb{D}$$

and the GMM efficiency bound is

$$\left(\mathbb{D}'\mathbb{V}^{-1}\mathbb{D}\right)^{-1}.$$

Example 7.4.5. *Consider again Example 7.1.2 in the special case in which $\ell = 1$. Let*

$$E\left[\varphi(X_t,\beta)\varphi(X_t,\beta)' \mid \mathfrak{I}_{t-1}\right] = V_{t-1}$$

wish to solve the following equation for A_t^d

$$E\left(A_t^{d\prime} V_{t-1} A_t\right) = \nabla(A) = E\left(\left[\frac{\partial\varphi}{\partial b'}(X_t,\beta)\right]' A_t\right). \tag{7.3}$$

Given the flexibility in the choice of the random A_t with entries that are $\mathcal{A}_{t-1}$ measurable, this equation is equivalent to

$$V_{t-1}A_t^d = E\left(\left[\frac{\partial\varphi}{\partial b'}(X_t,\beta)\right] \mid \mathfrak{I}_{t-1}\right)$$

where we have taken transposes of the expressions in (7.3). *Thus*

$$A_t^d = \left(V_{t-1}\right)^{-1} E\left(\left[\frac{\partial\varphi}{\partial b'}(X_t,\beta)\right] \mid \mathfrak{I}_{t-1}\right)$$

and the efficiency bound is:

$$\left[E\left(\left[\frac{\partial\varphi}{\partial b'}(X_t,\beta)\right]' \mid \mathfrak{I}_{t-1}\right)\left(V_{t-1}\right)^{-1} E\left(\left[\frac{\partial\varphi}{\partial b'}(X_t,\beta)\right] \mid \mathfrak{I}_{t-1}\right)\right]^{-1}.$$

Example 7.4.6. ***Two-stage least squares***. *Add the following special restrictions to example 7.4.5. Suppose that $r = 1$ and that $V_{t-1} = \mathbf{v} > 0$ where $\mathbf{v}$ is constant. Further suppose that*

$$\varphi(X_t, b) = Y_t^1 - Y_t^2 \cdot b$$

Finally, suppose that

$$E\left(Y_t^2 \mid \mathfrak{J}_{t-1}\right) = \Pi Z_{t-1}$$

where Z_{t-1} has more entries than Y_t^2. Notice that Π can be computed as a least squares regression. Then

$$A_t^d = \left(\frac{1}{\mathbf{v}}\right) Z_{t-1}{}'\Pi'$$

The scaling by $\frac{1}{\mathbf{v}}$ is inconsequential to the construction of a selection process. The matrix of regression coefficients can be replaced by the finite sample least squares regression coefficients without altering the statistical efficiency.

Example 7.4.6 has a special structure that does not prevail in some important applications. For instance, suppose that V_{t-1} depends on conditioning information so that a form conditional heteroskedasticity is present. That dependence shows up in essential ways in how A_t^d should be constructed. Further, suppose that the expectation $E\left(X_t^2 \mid \mathfrak{J}_{t-1}\right)$ potentially depends nonlinearly on Z_{t-1}. In that case, to attain or to approximate the efficiency bound, a least squares regression should account for potential nonlinearity. Finally, suppose that $\ell > 1$. Then even if the covariance structure is homoskedastic and conditional expectations are linear, the two-squares least square approach will no longer be statistically efficient. We again have to deploy an appropriate martingale central limit approximation. In these circumstances, simply by mapping into the framework of Example 7.1.1, we can improve efficiency relative to least squares or two-stage least squares, for instance, by letting

$$\varphi(X_t, b) = Z_{t-\ell}\left[Y_t^1 - \left(Y_t^2\right)' b\right]$$

Hansen and Singleton (1996) construct the efficiency bound in Example 7.1.2 for a linear data generating process.

Remark 7.4.7. *Consider again Example 7.1.1. Instead of "solving a system of equations," form estimators by optimiziation:*

$$\min_{b \in \Pi} \left[\frac{1}{N}\sum_{t=1}^{N} \varphi(X_t, b)\right]' \mathbb{W} \left[\frac{1}{N}\sum_{t=1}^{N} \varphi(X_t, b)\right]'$$

where β is an interior point of Π and $\mathbb{W}$ is a positive definite weighting matrix. The first-order conditions for this minimization problem are:

$$\frac{1}{N}\left[\sum_{t=1}^{N}\frac{\partial\varphi}{\partial b'}(X_t,b_N)\right]'\mathbb{W}\left[\frac{1}{N}\sum_{t=1}^{N}\varphi(X_t,b_N)\right]=0$$

where b_N is the minimizer. The efficiency bound is attained by replacing $\mathbb{W}$ with $\mathbb{V}^{-1}$ or a consistent estimator of this $\mathbb{V}^{-1}$.

7.5 Statistical tests

For purposes of devising a test of the "over-identifying restrictions," let $B = \{B_t : t \geq 0\}$ be an $r \times \tilde{k}$ matrix process constructed to verify

$$E\left[{B_t}'\varphi(X_t,\beta)\right]=0.$$

Suppose that

$$E\left[\sum_{j=0}^{\infty}{B_{t+j}}'\varphi(X_{t+j},\beta)\mid\mathfrak{J}_t\right]$$

converges in mean square so that we can apply a central limit approximation. Construct

$$\widetilde{\nabla}(B)\doteq E\left(\left[\frac{\partial\varphi}{\partial b'}(X_t,\beta)\right]'B_t\right).$$

By imitating the earlier argument

$$\begin{aligned}\frac{1}{\sqrt{N}}\sum_{t=1}^{N}{B_t}'\varphi(X_t,b_N)\approx&\frac{1}{\sqrt{N}}\sum_{t=1}^{N}{B_t}'\varphi(X_t,\beta)+\widetilde{\nabla}(B)'\sqrt{N}(b_N-\beta)\\ \approx&\frac{1}{\sqrt{N}}\sum_{t=1}^{N}{B_t}'\varphi(X_t,\beta)\\ &-\widetilde{\nabla}(B)'\nabla(A)^{-1}\frac{1}{\sqrt{N}}\sum_{t=1}^{N}{A_t}'\varphi(X_t,\beta)\\ \approx&\frac{1}{\sqrt{N}}\sum_{t=1}^{N}\left[{B_t}'-\widetilde{\nabla}(B)'\left[\nabla(A)'\right]^{-1}{A_t}'\right]\varphi(X_t,\beta)\end{aligned}$$

Notice that if $A_t = B_t$, then the right side is zero and the limiting distribution is degenerate. This approximation is used to construct tests that account for having used GMM to estimate a parameter vector β.

Example 7.5.1. *Consider again unconditional moment restrictions specified in Example 7.1.1. Let the selection process for testing be constant over time so that $B_t = \mathbb{B}$. Then*

$$\frac{1}{\sqrt{N}} \sum_{t=1}^{N} {B_t}' \varphi(X_t, b_N) \approx \frac{1}{\sqrt{N}} \sum_{t=1}^{N} \left[\mathbb{B}' - \mathbb{B}'\mathbb{D} \left(\mathbb{A}'\mathbb{D}\right)^{-1} \mathbb{A}'\right] \varphi(X_t, \beta).$$

Testing with a statistically efficient estimator

First suppose that we have statistically efficient selection process. Recall the approximation

$$\frac{1}{\sqrt{N}} \sum_{t=1}^{N} {B_t}' \varphi(X_t, b_N) \approx \frac{1}{\sqrt{N}} \sum_{t=1}^{N} \left[{B_t}' - \widetilde{\nabla}(B)' \left[\nabla(A^d)'\right]^{-1} A_t^{d'}\right] \varphi(X_t, \beta).$$

Let $\widetilde{G}_t(B)$ denote the increment in the martingale approximation for

$$\sum_{t=1}^{N} {B_t}' \varphi(X_t, \beta).$$

From the restrictions that we have imposed on the process B used for constructing tests

$$\widetilde{\nabla}(B) = E\left[G_t(A^d) G_t(B)'\right].$$

Using both of these representations:

$$\frac{1}{\sqrt{N}} \sum_{t=1}^{N} {B_t}' \varphi(X_t, b_N) \approx \frac{1}{\sqrt{N}} \sum_{t=1}^{N} \widehat{G}_t(B) \tag{7.4}$$

where

$$\widehat{G}_t(B) \doteq \widetilde{G}_t(B) - E\left[\widetilde{G}_t(B) G_t(A^d)'\right] \left(E\left[G_t(A^d) G_t(A^d)'\right]\right)^{-1} G_t(A^d)$$

The term, $\widehat{G}_t(B)$, that appears inside the sum on the right side of (7.4) is the population least squares residual from regressing $\widetilde{G}_t(B)$ onto $G_t(A^d)$.

This regression residual can also be interpreted as a martingale increment for a stationary increments process.

Suppose that $\widehat{G}_t(B)$ has a nonsingular covariance matrix. Consider the quadratic form used for building a test:

$$\frac{1}{N}\left[\sum_{t=1}^{N}\varphi(X_t,b_N)'B_t\right]\left(E\left[\widehat{G}_t(B)\widehat{G}_t(B)'\right]\right)^{-1}\left[\sum_{t=1}^{N}B_t{}'\varphi(X_t,b_N)\right]\Rightarrow\chi^2(\tilde{k}).$$

This test can be implemented in practice by replacing $E\left[\widehat{G}_t(B)\widehat{G}_t(B)'\right]$ with a statistically consistent estimator of it. There is an equivalent way to represent this quadratic form:

$$\frac{1}{N}\sum_{t=1}^{N}\varphi(X_t,b_N)'\begin{bmatrix}B_t & A_t^d\end{bmatrix}\left[E\left(\begin{bmatrix}\widetilde{G}_t(B)\\ G_t(A^d)\end{bmatrix}\begin{bmatrix}\widetilde{G}_t(B)' & G_t(A^d)'\end{bmatrix}\right)\right]^{-1}$$
$$\left[\sum_{t=1}^{N}\begin{bmatrix}B_t{}'\\ A_t^{d\prime}\end{bmatrix}\varphi(X_t,b_N)\right]$$

This equivalence follows because the inverse of the covariance matrix for the regression error $\widehat{G}_t(B)$ is the upper diagonal block of the inverse of the covariance matrix:

$$E\left(\begin{bmatrix}\widetilde{G}_t(B)\\ G_t(A^d)\end{bmatrix}\begin{bmatrix}\widetilde{G}_t(B)' & G_t(A^d)'\end{bmatrix}\right)$$

Example 7.5.2. *Consider Example 7.1.1 again. We have already shown that*

$$\mathbb{A}^d=\mathbb{V}^{-1}\mathbb{D}.$$

Suppose that we choose $\mathbb{B}$ *with dimension* $r\times(r-k)$ *so that*

$$\begin{bmatrix}\mathbb{A}^d & \mathbb{B}\end{bmatrix}$$

has full rank. Then

$$\frac{1}{N}\sum_{t=1}^{N}\varphi(X_t,b_N)'\mathbb{V}^{-1}\sum_{t=1}^{N}\varphi(X_t,b_N)'\Rightarrow\chi^2(r-k).$$

If we replace b_N *with* β *on the left side of the above limit we find*

$$\frac{1}{N}\sum_{t=1}^{N}\varphi(X_t,\beta)'\mathbb{V}^{-1}\sum_{t=1}^{N}\varphi(X_t,\beta)'\Rightarrow\chi^2(r)$$

The difference in the resulting χ^2 distribution emerges because estimating k free parameters reduces degrees of freedom by k. It is straightforward to show that

$$\frac{1}{N}\sum_{t=1}^{N}\varphi(X_t,\beta)'\mathbb{V}^{-1}\sum_{t=1}^{N}\varphi(X_t,\beta)' - \frac{1}{N}\sum_{t=1}^{N}\varphi(X_t,b_N)'\mathbb{V}^{-1}\sum_{t=1}^{N}\varphi(X_t,b_N)' \Rightarrow \chi^2(k),$$

an approximation that is useful for constructing confidence sets for GMM estimates of parameter vector β.

Bibliography

Bansal, Ravi and Amir Yaron. 2004. Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles. *Journal of Finance* 59 (4):1481–1509.

Billingsley, P. 1961. The Lindeberg-Levy Theorem for Martingales. *American Mathematical Monthly* 12.

Blackwell, D. and M. A. Girshick. 1954. *Theory of Games and Statistical Decisions*. New York: Wiley Publications in Statistics.

Blanchard, O. J and D. Quah. 1989. The Dynamic Effects of Aggregate Demand and Supply Disturbances. *American Economic Review* 79:655–673.

Box, G. E. P. and G. C. Tiao. 1992. *Bayesian Inference in Statisical Analysis*. New York: John Wiley and Sons, Inc.

Box, George E.P. and George C. Tiao. 1977. A Canonical Analysis of Multiple Time Series. *Biometrika* 64 (2):355–365.

Breiman, Leo. 1968. *Probability Theory*. Reading, Massachusetts: Addison-Wesley Publishing Company.

Carter, C. K. and R. Kohn. 1994. On Gibbs sampling for state space models. *Biometrika* 81 (3):541–553.

Cerreia-Vioglio, Simone, Fabio Maccheroni, Massimo Marinacci, and Luigi Montrucchio. 2013. Ambiguity and Robust Statistics. *Journal of Economic Theory* 148:974–1049.

Chamberlain, Gary. 1987. Asymptotic Efficiency in Estimation with Conditional Moment Restrictions. *Journal of Econometrics* 34 (3):305–334.

Doob, J. L. 1953. *Stochastic Processes.* New York: John Wiley and Sons.

Duffie, Darrell and Kenneth J. Singleton. 1993. Simulated Moments Estimation of Markov Models of Asset Prices. *Econometrica* 61 (4):929–952.

Dynkin, E. B. 1978. Sufficient Statistics and Extreme Points. *Annals of Probability* 6:705–730.

Engle, R. and C. W. J. Granger. 1987. Co-integration and Error Correction: Representation, Estimation and Testing. *Econometrica* 55:251–276.

Fan, K. 1952. Fixed Point and Minimax Theorems in Locally Convex Topological Linear Spaces. *Proceedings of the National Academy of Sciences* 38:121–126.

Ferguson, T. S. 1967. *Mathematical Statistics: A Decision Theoretic Approach.* New York: Academic Press.

Friedman, Milton. 1957. *A Theory of the Consumption Function.* Princeton, New Jersey: Princeton University Press.

Frisch, Ragnar. 1933. Propagation Problems and Impulse Problems in Dynamic Economics. In *Economic Essays in Honour of Gustav Cassel*, 171–205. Allen and Unwin.

Gallant, A. Ronald and George Tauchen. 1996. Which Moments to Match? *Econometric Theory* 12 (4):657–681.

Geweke, John and Susan Porter-Hudak. 1983. The Estimation and Application of Long Memory Time Series Models. *Journal of Time Series Analysis* 4 (4):221–238.

Gilboa, Itzhak and David Schmeidler. 1989. Maxmin Expected Utility with Non-Unique Prior. *Journal of Mathematical Economics* 18:141–153.

Gordin, M. I. 1969. The Central Limit Theorem for Stationary Processes. *Soviet Mathematics Doklady* 10:1174–1176.

Gourieroux, C, A Monfort, and E Renault. 1993. Indirect Inference. *Journal of Applied Econometrics* 8 (1S):85–118.

Granger, C. W. J. and Roselyne Joyeux. 1980. An Introduction To Long-Memory Time Series Models And Fractional Differencing. *Journal of Time Series Analysis* 1 (1):15–29.

Hall, P. and C. C. Heyde. 1980. *Martingale Limit Theory and Its Application*. Boston: Academic Press.

Hamilton, J. D. 1989. A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica* 57 (2):357–384.

Hansen, Lars Peter. 1982. Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica* 50 (4):1029–1054.

———. 1985. A Method for Calculating Bounds on the Asymptotic Covariance Matrices of Generalized Method of Moments Estimators. *Journal of Econometrics* 30 (1):203–238.

Hansen, Lars Peter and Thomas J. Sargent. 1993. Seasonality and approximation errors in rational expectations models. *Journal of Econometrics* 55 (1-2):21–55.

———. 2008. *Robustness*. Princeton, New Jersey: Princeton University Press.

———. 2013. *Recursive Models of Dynamic Linear Economies*. Princeton, New Jersey: Princeton University Press.

———. 2021. Macroeconomic Uncertainty Prices when Beliefs are Tenuous. *Journal of Econometrics* 223 (1):222–250.

Hansen, Lars Peter and Kenneth J. Singleton. 1996. Efficient Estimation of Linear Asset-Pricing Models with Moving Average Errors. *Journal of Business & Economic Statistics* 14 (1):53–68.

Hansen, Lars Peter, John C. Heaton, and Nan Li. 2008. Consumption Strikes Back?: Measuring Long Run Risk. *Journal of Political Economy* .

Hurwicz, Leonid. 1962. On the Structural Form of Interdependent Systems. In *Logic, Methodology and Philosophy of Science*, 232–239. Stanford, CA: Stanford University Press.

Jovanovic, Boyan. 1979. Job Matching and the Theory of Turnover. *Journal of Political Economy* 87 (5):972–990.

Klibanoff, Peter, Massimo Marinacci, and Sujoy Mukerji. 2005. A Smooth Model of Decision Making under Ambiguity. *Econometrica* 73 (6):1849–1892.

Knight, Frank H. 1921. *Risk, Uncertainty, and Profit.* Houghton Mifflin.

Krylov, N. and N. Bogolioubov. 1937. La Theorie Generale de la Mesure dans son Application a letude des systemes de la Mecanique non Lineaires. *Annals of Mathematics* 38:65–113.

Lee, Bong Soo and Beth Fisher Ingram. 1991. Simulation estimation of time-series models. *Journal of Econometrics* 47 (2-3):197–205.

Lucas, Robert E. Jr. 1976. Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy* 1:19 – 46.

Luce, R.D. and H. Raiffa. 1957. *Games and Decisions.* J. Wiley, New York.

Maccheroni, Fabio, Massimo Marinacci, and Aldo Rustichini. 2006. Ambiguity Aversion, Robustness, and the Variational Representation of Preferences. *Econometrica* 74 (6):1447–1498.

Marschak, Jacob. 1953. Economic Measurements for Policy and Prediction. In *Studies in econometric method*, edited by Tjalling Charles Koopmans William C. Hood, chap. 1, 1–26. John Wiley and Sons, Inc.

Meyn, S. and R. Tweedie. 1993. *Markov Chains and Stochastic Stability.* London: Springer-Verlag.

Muth, John F. 1960. Optimal Properties of Exponentially Weighted Forecasts. *Journal of the American Statistical Association* 55:299–306.

von Plato, Jan. 1982. The Significance of the Ergodic Decomposition of Stationary Measures for the Interpretation of Probability. *Synthese* 53:419–432.

Robinson, P. M. 1994. Semiparametric Analysis of Long-Memory Time Series. *The Annals of Statistics* 22 (1):515–539.

Sargan, J. D. 1958. The Estimation of Economic Relationships Using Instrumental Variables. *Econometrica* 26 (3):393–415.

Sargent, Thomas J. 1981. Interpreting Economic Time Series. *Journal of Political Economy* 89 (2):213–48.

Savage, L. J. 1954. *The Foundations of Statistics.* New York: John Wiley and Sons.

Sclove, S. L. 1983. Time-Series Segementation: A Model and a Method. *Information Sciences* 29 (1):7–25.

Shapiro, M. and M. Watson. 1988. Sources of Business Cycle Fluctuations. *NBER Macroeconomics Annual* 3:111–148.

Sims, Christopher A. 1980. Macroeconomics and Reality. *Econometrica* 48 (1):1–48.

Sims, Christopher A. and Tao Zha. 1999. Error Bands for Impulse Responses. *Econometrica* 67 (5):1113–1155.

Slutsky, Eugen. 1927. The Summation of Random Causes as the Source of Cyclic Processes. In *Problems of Economic Conditions*, vol. 3. Moscow: The Conjuncture Institute.

Smith, Anthony A. 1993. Estimating nonlinear time-series models using simulated vector autoregressions. *Journal of Applied Econometrics* 8 (1S):63–84.

Stigler, Stephen M. 1986. *The History of Statistics: The Measurement of Uncertainty Before 1900.* Harvard University Press.

Wald, Abraham. 1939. Contributions to the Theory of Statistical Estimation and Testing Hypotheses. *The Annals of Mathematical Statistics* 10 (4):299–326.

———. 1949. Statistical Decision Functions. *Annals of Mathematical Statistics* 20 (2):165–205.

———. 1950. *Statistical Decision Functions.* New York: John Wiley and Sons.

Working, Holbrook. 1934. A Random Difference Series for Use in the Analysis of Time Series. *Journal of the American Statistical Association* 29:11–24.

Yule, G. Udny. 1927. On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 226:267–298.

Zellner, Arnold. 1962. An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. *Journal of the American Statistical Association* 57.

Zha, Tao. 1999. Block recursion and structural vector autoregressions. *Journal of Econometrics* 90 (2):291–316.