Econ 2142: Lecture Notes

Ashesh Rambachan *

Contents

*Email: asheshr@g.harvard.edu These notes are based on lectures given by Prof. James H. Stock in Fall 2018 and Fall 2019. Please contact me if you find errors.

1 Second-Order Stationary Stochastic Processes

Let $X_1, \ldots, X_T$ denote a time series observed over $t = 1, \ldots, T$. For now, this is simply a sequence of random variables. As notation, we will sometimes denote this by $\{X_t\}_{t=1}^T$.

Remark 1.1 (Sampling concept)**.** *What does "random sampling" mean for a time series? In cross-sectional settings, we observe the data $X_1, \ldots, X_N$, where each observation is drawn i.i.d. from some infinite super-population. We model the data as a random variables or vectors to capture this random sampling process.*

*In a time series, we imagine that the observed time series $X_1, \ldots, X_T$ are realizations from some infinitely long stochastic process, $\{X_t\}$. The observed data is simply a finite sub-sequence that we observe from this underlying stochastic process. As we will, the assumption that the observed time series is **stationary** or **weakly stationary** will play the role of the usual i.i.d. random sampling assumption in cross-sectional settings.*

Definition 1.1. *Let $\{X_t\}$ for $t = 1, \ldots, T$ be a sequence of random variables.*

- *$\{X_t\}$ is **stationary** if the joint distribution of $(X_{t+1}, \ldots, X_{t+k})$ does not depend on t.*
- *$\{X_t\}$ is **second-order stationary** or **weakly stationary** or **covariance stationary** if*
  1. *$\mathbb{E}[X_t] = \mu \; \forall t$.*
  2. *$Cov(X_t, X_{t-k}) = \gamma_k \; \forall t$.*

  *We call $\gamma_k = Cov(X_t, X_{t-k})$ the **auto-covariance** between X_t, X_{t-k} and*

$$\rho_k = \frac{Cov(X_t, X_{t-k})}{\sqrt{V(X_t)V(X_{t-k})}} \overset{(1)}{=} \frac{\gamma_k}{\gamma_0}$$

  *the **auto-correlation** between X_t, X_{t-k}, where (1) follows under second-order stationarity.*

Definition 1.2. *Let $\mathcal{F}_t$ denote the "filtration" generated by the time series X_t. Heuristically, this is just the time-t "information set" with $\mathcal{F}_t = \{X_t, X_{t-1}, \ldots\}$. We say the time series X_t is a **martingale** if*

$$\mathbb{E}\left[X_t \,|\, \mathcal{F}_{t-1}\right] = X_{t-1}.$$

*We say the time series Z_t is a **martingale difference sequence (mds)** if*

$$\mathbb{E}\left[Z_t \,|\, \mathcal{F}_{t-1}\right] = 0.$$

Notice that if X_t is a martingale, then the series $\Delta X_t \equiv X_t - X_{t-1}$ is a martingale difference sequence.

We next define the lag-operator. We will use this as notation throughout the course.

Definition 1.3. *The **lag operator** L satisfies*

$$LX_t = X_{t-1}.$$

We analogously define the operator L^k as

$$L^k X_t = X_{t-k}, \quad \forall k \geq 0.$$

*The **inverse lag operator** L^{-1} simply shifts the series forward one time period,*

$$L^{-1}X_t = X_{t+1}, \quad where L^{-1}LX_t = X_t.$$

*A **lag polynomial** $a(L)$ is defined as*

$$a(L) = \sum_{j=0}^{\infty} a_j L^j,$$

and so the lag polynomial applied to the time series X_t delivers

$$\begin{aligned} a(L)X_t &= \sum_{j=0}^{\infty} a_j X_{t-j} \\ &= a_0 X_t + a_1 X_{t-1} + a_2 X_{t-2} + \dots \end{aligned}$$

*We also refer to $a(L)X_t$ as a **linear, time invariant filter** of X_t.*

Definition 1.4. *For a lag polynomial $a(L)$, we refer to $a(z)$ as the **z-transform** of the lag polynomial, where*

$$a(z) = \sum_{j=0}^{\infty} a_j z^j.$$

For lag polynomials $a(L), b(L)$, the z-transform satisfies

$$a(z) + b(z) = c(z), \text{ where } c(z) = \sum_{j=0}^{\infty} c_j z^j \text{ and } c_j = a_j + b_j$$

$$a(z)b(z) = c(z), \text{ where } c(z) = \sum_{j=0}^{\infty} c_j z^j, \; c_j = \sum_{i=0}^{\infty} a_i b_{j-i} \text{ and } b_{j-i} = 0 \text{ if } j - i < 0.$$

Remark 1.2 (Inverse of a lag polynomial)**.** *Consider a lag polynomial $a(L)$. Can we define its inverse $a(L)^{-1}$? In other words, suppose that $Y_t = a(L)X_t$. When is it true that $X_t = a(L)^{-1}Y_t$? This will be true whenever the roots of $|a(z)|$ lie outside the unit circle.*

To see this, consider $a(L) = (1 - \alpha L)$ for simplicity. We claim that

$$(1 - \alpha L)^{-1} = \sum_{j=0}^{\infty} \alpha^j L^j.$$

Why? First, consider the case where X_t is a sequence of deterministic scalars. We have that

$$\begin{aligned} \left(\sum_{j=0}^{m} \alpha^j L^j \right) (1 - \alpha L) X_t &= \left(\sum_{j=0}^{m} \alpha^j L^j - \sum_{j=0}^{m} \alpha^{j+1} L^{j+1} \right) X_t \\ &= \left(1 - \alpha^{m+1} L^{m+1} \right) X_t \\ &= X_t - \alpha^{m+1} X_{t-m-1}. \end{aligned}$$

If $|\alpha| < 1$ and $\{X_t\}$ is bounded, then as $m \to \infty$, $\alpha^{m+1} X_{t-m-1} \to 0$. So, we conclude that

$$\left(\sum_{j=0}^{\infty} \alpha^j L^j \right) (1 - \alpha L) X_t = X_t$$

for scalars. Now, let X_t be a mean-zero stochastic process. We have that

$$\begin{aligned} Y_t &= X_t - \alpha X_{t-1} \\ \hookrightarrow X_t &= Y_t + \alpha X_{t-1} \\ \hookrightarrow X_t &= Y_t + \alpha(Y_{t-1} + \alpha X_{t-2}) \\ &\vdots \\ \hookrightarrow X_t &= \sum_{j=0}^{m} \alpha^j Y_{t-j} + \alpha^{m+1} X_{t-m-1}. \end{aligned}$$

With this, we'll show that $X_t - (\sum_j^m \alpha^j L^j)(1 - \alpha L) X_t$ converges in mean-square to zero as $m \to \infty$. We have that

$$\begin{aligned} \mathbb{E}\left[\left(\sum_{j=0}^{m} \alpha^j L^j \right) (1 - \alpha L) X_t - X_t \right]^2 &= \mathbb{E}\left[(-\alpha^{m+1} X_{t-m-1})^2 \right] \\ &= \alpha^{2(m+1)} \mathbb{E}[X_{t-m-1}^2] \\ &= \alpha^{2m+2} \gamma_0. \end{aligned}$$

Then, provided that $|\alpha| < 1$ and $\gamma_0 < \infty$, we have that

$$\mathbb{E}\left[\left(\sum_{j=0}^{m} \alpha^j L^j \right) (1 - \alpha L) X_t - X_t \right]^2 \xrightarrow{m.s.} 0$$

and the claim follows.

We finally define several useful stochastic processes that will appear frequently.

Definition 1.5. *A stochastic process $\{\epsilon_t\}$ is a **white noise process** if*

1. $\mathbb{E}[\epsilon_t] = 0 \; \forall t$,
2. $V(\epsilon_t) = \sigma^2 < \infty \; \forall t$,
3. $Cov(\epsilon_s, \epsilon_t) = 0 \; \forall s \neq t$.

*We denote a white noise process by $\epsilon_t \sim WN(0, \sigma^2)$. An **order-$p$ autoregressive process**, denoted $AR(p)$, is defined as*

$$X_t = \sum_{j=1}^{p} a_j X_{t-j} + \epsilon_t, \quad \epsilon_t \sim WN(0, \sigma^2).$$

*An **order-q moving average process**, denoted $MA(q)$, is defined as*

$$X_t = \sum_{j=0}^{q} c_j \epsilon_{t-j}, \quad \epsilon_t \sim WN(0, \sigma^2).$$

Example 1.1. *We can easily compute the auto-correlation function of an $MA(q)$ process. For instance, consider an $MA(\infty)$ with*

$$X_t = c(L)\epsilon_t,$$

where

$$\mathbb{E}[\epsilon_t] = 0,$$
$$\mathbb{E}[\epsilon_s \epsilon_t] = \begin{cases} \sigma_\epsilon^2 \text{ if } t = s, \\ 0 \text{ otw.} \end{cases}$$

We have that

$$\begin{aligned} \gamma_k &= \mathbb{E}[X_t X_{t-k}] \\ &= \mathbb{E}\left[\left(\sum_{j=0}^{\infty} c_j \epsilon_{t-j} \right) \left(\sum_{i=0}^{\infty} c_i \epsilon_{t-i-k} \right) \right] \\ &= \sum_{i=0}^{\infty} c_i c_{i+k} \sigma_\epsilon^2 \end{aligned}$$

after some simple algebra.

1.1 The Wold Decomposition

To this point, we introduced the auto-covariance function for a second-order stationary stochastic process, $\{\gamma_k\}$. There are other *equivalent* representations of a second-order stationary process:

- **The moving average representation**: A second-order stationary process X_t can be written as

  $$X_t = c(L)\varepsilon_t,$$

  where ε_t is a serially uncorrelated process, meaning $\mathbb{E}\left[\varepsilon_t \varepsilon_{t-j}\right] = 0$ for all $j \neq 0$.

- **The spectral density representation**: More to come on this later.

In this subsection, we will show how to construct the moving average representation from the auto-covariance function for a second-order stationary process. This is known as the **Wold Decomposition**. The proof of the Wold Decomposition is constructive and we will work through each step.[1]

[1] Chapter 5 of Brockwell and Davis (1991) provides a full statement and careful proof of this result. A rigorous proof of the Wold Decomposition requires an investment in setting up an underlying Hilbert Space upon which the time series X_t lives and in which the projection operator is well-defined. We will take these constructions for granted to focus on the intuitions and insights of this important result.

Theorem 1.1 (Wold Decomposition). *Let $\{X_t\}$ be a mean-zero, second-order stationary process. Define*

$$M_t = span(X_t, X_{t-1}, \ldots).$$

Assume that there is no perfect predictability from the one-step ahead linear forecast, meaning

$$V\left(X_t - \text{Proj}\left\{X_t \mid M_t - 1\right\}\right) > 0.$$

Then,

$$X_t = c(L)\varepsilon_t + v_t,$$

where

1. $\mathbb{E}\left[\varepsilon_t\right] = 0$, $\mathbb{E}\left[\varepsilon_t^2\right] = \sigma_\varepsilon^2 > 0$ *and* $\varepsilon_t \in M_t$.
2. $\mathbb{E}\left[\varepsilon_t \varepsilon_s\right] = 0$ *for all* $s \neq t$.
3. *The coefficients of the lag polynomial $\{c_i\}$ do not depend on t with $c_0 = 1$.*
4. *The coefficients of the lag polynomial are square summable with $\sum_{j=0}^{\infty} c_j^2 < \infty$.*
5. *The series $\{v_t\}$ is deterministic, meaning*

$$v_t \in M_\infty = \cap_{i=0}^{\infty} M_{t-i}.$$

Proof. Define

$$\varepsilon_t = X_t - \text{Proj}\left\{X_t \mid M_{t-1}\right\}.$$

By construction, $\varepsilon_t \perp M_{t-1}$. Denote the residual of the projection of X_t onto its past values as

$$X_t - \text{Proj}\left\{X_t \mid M_{t-1}\right\} = a(L)X_t,$$

where $a_0 = 1$ and a_j is the negative of the coefficient on X_{t-j} in the projection of X_t onto its past values. Since projection coefficients are simply a function of variances and covariances and X_t is second-order stationary, this implies that the projection coefficients and therefore the sequence $\{a_j\}$ are both time-invariant.

With this construction, we now show that each property is satisfied:

1. $\mathbb{E}\left[\varepsilon_t\right] = \mathbb{E}\left[a(L)X_t\right] = 0$ because X_t is mean-zero. Then,

$$\mathbb{E}\left[\varepsilon_t^2\right] = \mathbb{E}\left[\left(a(L)X_t\right)^2\right] = \sigma_\epsilon^2 < \gamma_0 < \infty$$

where the last inequality follows because ϵ_t is defined as a projection residual.

2. Notice that

$$\text{Proj}\left\{\varepsilon_t \mid X_{t-j}\right\} = \frac{\mathbb{E}\left[\varepsilon_t X_{t-j}\right]}{\mathbb{E}\left[X_{t-j}^2\right]} X_{t-j}.$$

We know that $\text{Proj}\left\{\varepsilon_t \mid X_{t-j}\right\} = 0$ because $\varepsilon_t \perp M_{t-1}$ by construction. Therefore, it follows that

$$\mathbb{E}\left[\varepsilon_t X_{t-j}\right] = 0 \quad \forall j.$$

It then follows that

$$\mathbb{E}\left[\varepsilon_t \varepsilon_{t-j}\right] = \mathbb{E}\left[\varepsilon_t a(L) X_{t-j}\right] = 0$$

for all $j > 0$.

3. We have that

$$\begin{aligned} c(L)\varepsilon_t &= \text{Proj}\left\{X_t \mid \varepsilon_t, \varepsilon_{t-1}, \ldots\right\} \\ &= \frac{\mathbb{E}\left[\varepsilon_t X_t\right]}{\mathbb{E}\left[\varepsilon_t^2\right]}\varepsilon_t + \frac{\mathbb{E}\left[\varepsilon_{t-1} X_t\right]}{\mathbb{E}\left[\varepsilon_{t-1}^2\right]}\varepsilon_{t-1} + \ldots, \end{aligned}$$

where the second equality follows because $\{\varepsilon_t\}$ is serially uncorrelated. Therefore,

$$c_i = \frac{\mathbb{E}\left[\varepsilon_{t-i} X_t\right]}{\mathbb{E}\left[\varepsilon_{t-i}^2\right]} = \frac{\mathbb{E}\left[\varepsilon_{t-i} X_t\right]}{\sigma_\varepsilon^2}.$$

Moreover,

$$\begin{aligned} c_0 &= \frac{\mathbb{E}\left[\varepsilon_t X_t\right]}{\sigma_\varepsilon^2} \\ &= \frac{\mathbb{E}\left[\varepsilon_t(\varepsilon_t + \text{Proj}\left\{X_t \mid M_{t-1}\right\})\right]}{\sigma_\varepsilon^2} \\ &= \sigma_\varepsilon^2 / \sigma_\varepsilon^2 = 1. \end{aligned}$$

4. Next, we have that

$$\begin{aligned} V\left(X_t - c(L)\varepsilon_t\right) &= \gamma_0 - 2\mathbb{E}\left[X_t c(L)\varepsilon_t\right] + \mathbb{E}\left[\left(c(L)\varepsilon_t\right)^2\right] \\ &= \gamma_0 - 2\sum_{j=0}^{\infty} c_j \frac{\mathbb{E}\left[X_t \varepsilon_{t-j}\right]}{\sigma_\varepsilon^2}\sigma_\varepsilon^2 + \sum_{j=0}^{\infty} c_j^2 \sigma_\varepsilon^2 \\ &= \gamma_0 - \sum_{j=0}^{\infty} c_j^2 \sigma_\varepsilon^2 \overset{(1)}{\geq} 0, \end{aligned}$$

where (1) follows from the no perfect predictability assumption. Therefore, it follows that

$$\sum_{j=0}^{\infty} c_j^2 \leq \gamma_0/\sigma_\varepsilon^2 < \infty.$$

5. Finally, we have that $X_t \in M_t$ and $\varepsilon_t \in M_t$. However, notice that

$$\begin{aligned}\mathbb{E}\left[v_t \varepsilon_{t-j}\right] &= \mathbb{E}\left[(X_t - c(L)\varepsilon_t)\varepsilon_{t-j}\right] \\ &= \mathbb{E}\left[X_t \varepsilon_{t-j}\right] - c_j \sigma_\varepsilon^2 = 0\end{aligned}$$

Therefore, $v_t \perp \varepsilon_{t-j}$, and so $v_t \perp span(\varepsilon_t, \varepsilon_{t-1}, \ldots)$. We conclude that $v_t \in \cup_{j=0}^{\infty} M_{t-j}^C$ and an application of DeMorgan's Law implies $v_t \in \cap_{j=0}^{\infty} M_{t-j}$.

□

Heuristically, the Wold Decomposition states that if you are given any second-order stationary process, then we can re-write it as a linear combination of serially uncorrelated innovations. We refer to the series ε_t as the **Wold innovations**.

Remark 1.3 (Wold Decomposition)**.** *We now make a series of remarks about the Wold Decomposition:*

1. *$\{v_t\}$ is a "deterministic" series. What does this mean? A classic example is $v_t = a\cos(bt)$, where a, b are random variables that are independent of the process X_t. Throughout the course, we'll assume that the series we analyze have no deterministic component or, in jargon, that the series are linearly non-deterministic.*

2. *By construction, the Wold Decomposition is unique. We'll return to this important fact when we discuss invertibility.*

3. *The Wold innovations ε_t are serially uncorrelated but they are not independent. This is because the Wold innovations are defined as the residuals from a linear projection.*

4. *The Wold Decomposition shows how to move from the autocovariance function of a second-order stationary stochastic process to the moving average representation. We can also use it to go from the moving average representation to the autocovariance function. Suppose that $X_t = c(L)\varepsilon_t$, where $c(L)$ is known. Then,*

$$\gamma_j = \mathbb{E}\left[X_t X_{t-j}\right] = \sum_{i=j}^{\infty} c_i c_{i-j} \sigma_\varepsilon^2.$$

5. *The Wold Decomposition began its construction by considering the population projection of X_t onto all of its past values and writes X_t as a infinitely long moving average of Wold innovations. What if we truncated this after q lags? That is, define*

$$\begin{aligned}X_t^q &= \text{Proj}\left\{X_t \,|\, \varepsilon_t, \varepsilon_{t-1}, \ldots, \varepsilon_{t-q}\right\} \\ &= c_0\varepsilon_t + c_1\varepsilon_{t-1} + \ldots + c_q\varepsilon_{t-q}.\end{aligned}$$

Then,

$$X_t - X_t^q = \sum_{j=q+1}^{\infty} c_j \varepsilon_{t-j}$$

and

$$V\left(X_t - X_t^q\right) = \sum_{j=q+1}^{\infty} c_j^2 \sigma_\varepsilon^2 \to 0$$

as $q \to \infty$. So, X_t^q converges in mean-square to X_t as q grows large. Therefore, for all $\epsilon > 0$, there exists a $\underline{q}$ such that for all $q > \underline{q}$

$$V\left(X_t - X_t^q\right) < \epsilon.$$

In other words, we can approximate a second-order stationary process arbitrarily well with an $MA(q)$ process.

6. *An analogous result exists for auto-regressive processes. Define*

$$X_t^p = \text{Proj}\left\{X_t \mid X_{t-1}, \ldots, X_{t-p}\right\}.$$

Define

$$X_t^\infty = \text{Proj}\left\{X_t \mid M_{t-1}\right\}.$$

Then, it can be shown that

$$\mathbb{E}\left[\left(X_t^\infty - X_t^p\right)^2\right] \to 0$$

as $p \to \infty$.

1.2 Yule-Walker Equations

The **Yule-Walker Equations** are a useful tool from computing the autocovariances of an auto-regressive moving-average (ARMA) process. We say a time series X_t is an $ARMA(p, q)$ if it can be written as

$$a(L)X_t = b(L)\epsilon_t,$$

where $a(L)$ is a p-th order lag-polynomial and $b(L)$ is a q-th order lag-polynomial.

We will work through the Yule-Walker equations for an $AR(1)$ to illustrate the ideas. Consider

$X_t = \alpha X_{t-1} + \epsilon_t$. Then, the Yule-Walker equations are

$$\begin{aligned}
\mathbb{E}\left[X_t(X_t - \alpha X_{t-1})\right] &= \mathbb{E}\left[X_t \epsilon_t\right] \\
\mathbb{E}\left[X_{t-1}(X_t - \alpha X_{t-1}\right] &= \mathbb{E}\left[X_{t-1} \epsilon_t\right] \\
\mathbb{E}\left[X_{t-2}(X_t - \alpha X_{t-1}\right] &= \mathbb{E}\left[X_{t-2} \epsilon_t\right] \\
&\vdots
\end{aligned}$$

and so on. These are equivalent to

$$\begin{aligned}
\gamma_0 - \alpha\gamma_1 &= \sigma_\epsilon^2 \\
\gamma_1 - \alpha\gamma_0 &= 0 \\
\gamma_2 - \alpha\gamma_1 &= 0.
\end{aligned}$$

Solving these equations, we see that

$$\begin{aligned}
\gamma_0 &= \frac{\sigma_\epsilon^2}{1-\alpha^2} \\
\gamma_1 &= \frac{\alpha\sigma_\epsilon^2}{1-\alpha^2} \\
\gamma_2 &= \frac{\alpha^2\sigma_\epsilon^2}{1-\alpha^2}
\end{aligned}$$

and so on.

1.3 Invertibility and Fundamentalness

Time is a strange concept in a stationary or second-order stationary stochastic process. In particular, for these types of time series, the *direction* of time is irrelevant. Consider a second-order stationary process. All of the information about the stochastic process is encoded in its auto-covariance function and by assumption, this function is time invariant. In particular,

$$\gamma_k = Cov(X_t, X_{t-k}) = Cov(X_t, X_{t+k}).$$

So, in this setting, what does "the past" actually mean?

With this strangeness in mind, return to the Wold Decomposition. In the proof of the Wold Decomposition, we immediately defined the Wold innovations to be the residuals from the projection of the time series X_t onto its past values $X_{t-1}, X_{t-2}, \ldots$. But, there was actually no reason for us to do this other than that seems this is a natural construction – we are predicting X_t using its past ("observed") values. In fact, we could have defined the Wold innovations to be the residuals form the projection of X_t onto its *future* values and gone through the exact same proof of the Wold decomposition! So what is going on here?

Consider an example in which X_t is an $MA(1)$ with

$$X_t = \epsilon_t + \theta\epsilon_{t-1} = (1 + \theta L)\epsilon_t.$$

Let us compute the auto-covariances of this process. We have that

$$\begin{aligned}\gamma_0 &= (1+\theta^2)\sigma_\epsilon^2 \\ \gamma_1 &= \theta\sigma_\epsilon^2 \\ \gamma_k &= 0 \quad \forall k \geq 0.\end{aligned}$$

So, we have that

$$\frac{\gamma_1}{\gamma_0} = \frac{\theta}{1+\theta^2} = \frac{\theta^{-1}}{1+\theta^{-2}}.$$

We can construct an *observationally equivalent* series X_t' that has the exact same auto-covariances as X_t. In particular, define

$$X_t' = (1+\theta^{-1}L)\eta_t, \quad \text{where } \eta_t = \epsilon_t/\theta^{-1}.$$

Then, notice that the auto-covariances of X_t' are

$$\begin{aligned}\gamma_0' &= (1+\theta^{-2})(\sigma_\epsilon^2/\theta^{-2}) = (1+\theta^2)\sigma_\epsilon^2 \\ \gamma_1' &= \theta\sigma_\epsilon^2 \\ \gamma_k' &= 0 \quad \forall k \geq 0.\end{aligned}$$

Think of X_t' as the "evil data" of the original time series X_t. From the observed auto-covariances, we cannot tell whether the observed data was generated by X_t or its evil twin X_t'.

Suppose that $|\theta| < 1$. Then, we have that

$$\begin{aligned}X_t = (1+\theta L)\epsilon_t \implies \epsilon_t &= (1+\theta L)^{-1}X_t \\ &= \sum_{j=0}^{\infty}(-\theta)^j X_{t-j} \in M_t.\end{aligned}$$

Since $\epsilon_t \in M_t$, it must be the Wold innovations that we defined in the proof of the Wold decomposition. We call these errors **fundamental**. Now, consider the evil twin X_t'. We have that

$$X_t' = (1+\theta^{-1}L)\eta_t.$$

What happens if we try to invert this lag polynomial, $(1+\theta^{-1}L)$? This operator will not be well-defined because it will not be a convergent series as $|\theta| < 1$ means that $|\theta^{-1}| > 1$. However, we can use the following trick

$$\begin{aligned}\theta L^{-1}X_t' &= \theta L^{-1}(1+\theta^{-1}L)\eta_t \\ &= (\theta L^{-1}+1)\eta_t.\end{aligned}$$

We can invert the lag polynomial $(\theta L^{-1}+1)$, where L^{-1} is the forward-shift operator. Therefore, we

have that

$$\begin{aligned}\eta_t &= (1+\theta L^{-1})^{-1}\theta L^{-1} X_t' \\ &= \theta \sum_{j=0}^{\infty} (-\theta)^j L^{-j} X_{t+1}' \\ &= \theta \sum_{j=0}^{\infty} (-\theta)^j X_{t+j+1}'.\end{aligned}$$

That is, η_t is the residual from the projection of X_t' onto its *future* values! We call this a *non-fundamental error*.

This issue of fundamentalness is important. We will return to it when we discuss structural vector auto-regressions and dynamic causal effects.

1.4 Spectral density and spectral analysis

We typically write a weakly stationary time series in a form like

$$X_t = \mu + \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$$

and then analyze its auto-covariances. This is known as analyzing the time series in the **time domain**. We'll now show that we can also express a weakly stationary time series as

$$X_t = \mu + \int_0^{\pi} \alpha(\omega) cos(\omega t)\, d\omega + \int_0^{\pi} \delta(\omega) sin(\omega t)\, d\omega.$$

We then wish to study the relative importance of cyclical behavior at different frequencies. This is known as analyzing the time series in the **frequency domain** or **spectral analysis**.

Let $\{X_t\}$ be a weakly stationary process with $\mathbb{E}[Y_t] = \mu$, $\gamma_j = Cov(X_t, X_{t-j})$. Additionally assume that the auto-covariances are absolutely summable with

$$\sum_{j=0}^{\infty} |\gamma_j| < \infty.$$

The **auto-covariance generating function** is defined as

$$\gamma(z) = \sum_{j=-\infty}^{\infty} \gamma_j z^j.$$

Example 1.2 ($MA(\infty)$)**.** *The auto-covariance generating function of $X_t \sim MA(\infty)$, where $X_t = c(L)\epsilon_t$ is simply*

$$\gamma(z) = \sigma_\epsilon^2 c(z) c(z^{-1}).$$

If we divide the auto-covariance generating function by 2π and evaluate it at $z = e^{-i\omega}$, then this

is known as the **population spectrum** of X with

$$\begin{aligned} S_X(\omega) &= \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma_j e^{-i\omega j}, \quad i = \sqrt{-1} \\ &= \frac{1}{2\pi} \gamma(e^{-i\omega}). \end{aligned}$$

In other words, the population spectrum $S_X(\omega)$ of X_t is simply the discrete fourier transformation of its auto-covariance generating function. Therefore, from the population spectrum, we can recover the auto-covariances by simply applying the inverse fourier transformation. We have that

$$\begin{aligned} \int_{-\pi}^{\pi} S_X(\omega) e^{i\omega k} d\omega &= \int_{-\pi}^{\pi} \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma_j e^{-i\omega j} e^{i\omega k} d\omega \\ &= \int_{-\pi}^{\pi} \gamma_j \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} e^{-i\omega j} e^{i\omega k} d\omega \\ &= \gamma_k. \end{aligned}$$

Notice that this implies that

$$\gamma_0 = V(X_t) = \int_{-\pi}^{\pi} S_Y(\omega)\, d\omega.$$

This gives rise to the interpretation of the population spectrum as providing a decomposition of the portion of the variance of X_t that is associated with each frequency. We will unpack this idea more in the next section.

Example 1.3 (Spectrum of MA process)**.** *Suppose $X_t = c(L)\epsilon_t$. Recall that $\gamma(z) = c(z)c(z^{-1})\sigma_\epsilon^2$. Then,*

$$\begin{aligned} S_X(\omega) &= \frac{1}{2\pi} c(e^{-i\omega}) c(e^{i\omega}) \sigma_\epsilon^2 \\ &= \|c(e^{i\omega})\|^2 \frac{\sigma_\epsilon^2}{2\pi}, \end{aligned}$$

where $\|\cdot\|$ denotes the complex conjugate.

Example 1.4 (Spectrum of AR process)**.** *Suppose that $a(L)X_t = \epsilon_t$ and $a(L)^{-1}$ exists. Then, $X_t = a(L)^{-1}\epsilon_t$ and applying the previous result, we see that*

$$S_X(\omega) = \frac{1}{\|a(e^{i\omega})\|^2} \frac{\sigma_\epsilon^2}{2\pi}.$$

Example 1.5 (Spectrum of an $AR(1)$)**.** *Consider $X_t = \alpha X_{t-1} + \epsilon_t$ and define $a(L) = (1 - \alpha L)$. Then,*

applying our previous result, we have that

$$\begin{aligned} S_X(\omega) &= \frac{\sigma_\epsilon^2/2\pi}{\|1-\alpha e^{i\omega}\|^2} \\ &= \frac{\sigma_\epsilon^2/2\pi}{(1-\alpha e^{i\omega})(1-\alpha e^{-i\omega})} \\ &= \frac{\sigma_\epsilon^2/2\pi}{1+\alpha^2-2\alpha\cos(\omega)}, \end{aligned}$$

where the last equality followed by multiplying the denominator out and applying Euler's formula (recall $e^{-i\omega j} = \cos(\omega j) - i\sin(\omega j)$*).*

Proposition 1.1. *Properties of the population spectrum The population spectrum* $S_X(\omega)$ *has the following properties:*

1. $S_X(\omega)$ *is a real-valued function.*
2. $S_X(\omega)$ *is symmetric around* $\omega = 0$ *and is periodic with period equal to* 2π*, meaning*

$$S_X(\omega) = S_X(\omega + 2\pi k)$$

for $k = \pm 1, \pm 2, \ldots$

Proof. We can re-write the population spectrum as

$$\begin{aligned} S_X(\omega) &= \frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\gamma_j e^{-i\omega j} \\ &= \frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\gamma_j\left(\cos(\omega j) - i\sin(\omega j)\right). \end{aligned}$$

Since X_t is second-order stationary, $\gamma_j = \gamma_{-j}$. Recall that $\sin(x) = -\sin(-x)$. Therefore, the population spectrum simplifies to

$$\frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\gamma_j\cos(\omega j).$$

The results then follow from properties of the cosine function. □

Because the population spectrum is symmetric around $\omega = 0$ and is periodic with period 2π, it is sufficient to examine its properties over $\omega \in [0, \pi]$.

1.4.1 The population spectrum as an "asymptotic diagonalization" of the auto-covariance matrix

Define the auto-covariance matrix Γ of X_t as

$$\Gamma_T = \mathbb{E}\left[\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_T \end{pmatrix}\begin{pmatrix} X_1 & X_2 & \dots & X_T \end{pmatrix}\right],$$

whose (i,j)-th entry is $\gamma_{|i-j|}$. We will now show that the spectrum is an "asymptotic diagonalization" of this covariance matrix. That is, we'll show that we can construct a basis of eigenvectors that will be made up of cosine and sine functions

$$e_j'(\omega) = (e^{-i\omega}, \dots, e^{-i\omega T}),$$

whose lengths will equal the spectrum evaluated at ω asymptotically, $\|e_j'(\omega)\|^2 \approx S_X(\omega)$. Moreover, we'll consider a particular orthonormal eigenbasis in which we choose the values of ω to be $\omega_j = 2\pi j/T$.

We begin with a simple result that we will use throughout this subsection.

Lemma 1.1. *Let X_t be a second-order stationary process with absolutely summable auto-covariances. That is,*

$$\sum_{u=0}^{\infty} |\gamma_u| < \infty.$$

Let $W_T(u) \leq 1$ be a sequence of weight functions and assume that

$$W_T(u) \to W(u)$$

pointwise in u. Then,

$$\sum_{u=-(T-1)}^{T-1} W_T(u)\gamma_u \to \sum_{u=-\infty}^{\infty} W(u)\gamma_u.$$

Proof. This result is an application of the **dominated convergence theorem**.[2] We have that

$$W_T(u)\gamma_u \to W(u)\gamma_u$$

pointwise for all u. Moreover,

$$|W_T(u)\gamma_u| \leq |\gamma_u|$$

for all u and T. The result follows by dominated convergence. □

[2]Heuristically, the dominated convergence theorem states: Suppose that $f_n \to f$ pointwise and that f_n is dominated by an integrable function g, $|f_n| \leq g$ for all n. Then, $\int f_n d\mu \to \int f d\mu$, where μ is some measure.

Now, define the function

$$d_X(\omega) = \sqrt{\frac{1}{T}} \sum_{t=1}^{T} X_t e^{-i\omega t}.$$

Remark 1.4. *Where does this come from? Define the series*

$$e(\omega)' = (e^{-i\omega\cdot 1}, e^{-i\omega\cdot 2}, \ldots, e^{-i\omega T}).$$

Then, the coefficient in the projection of X onto $e(\omega)$ is given by

$$\beta = \frac{\langle e(\omega), X\rangle}{\langle e(\omega), e(\omega)\rangle} = \frac{1}{T}\sum_{t=1}^{T} X_t e^{-i\omega t}.$$

Therefore, the function $d_X(\omega)$ is simply $\sqrt{T}\beta$.

With this in hand, we'll consider two calculations. The first calculation computes $\mathbb{E}\left[\frac{1}{2\pi}\|d_X(\omega)\|^2\right]$. We have that

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{2\pi}\|d_X(\omega)\|^2\right] &= \frac{1}{2\pi}\mathbb{E}\left[\left\|\sqrt{\frac{1}{T}}\sum_{t=1}^{T} X_t e^{-i\omega t}\right\|^2\right] \\
&\overset{(1)}{=} \frac{1}{2\pi}\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} X_t X_s e^{-i\omega t} e^{i\omega s}\right] \\
&= \frac{1}{2\pi}\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\gamma_{t-s}e^{-i\omega(t-s)} \\
&= \frac{1}{2\pi}\frac{1}{T}\sum_{u=-(T-1)}^{T-1}(T-|u|)\,\gamma_u e^{-i\omega u}, \quad u = t-s \\
&= \frac{1}{2\pi}\sum_{u=-(T-1)}^{T-1}\underbrace{(1-|u/T|)\,e^{-i\omega u}}_{W_T(u)}\gamma_u \\
&\overset{(2)}{\to} \frac{1}{2\pi}\sum_{u=-\infty}^{\infty} e^{-i\omega u}\gamma_u = S_X(\omega),
\end{aligned}$$

where (1) uses that $\|d_X(\omega)\|^2 = d_X(\omega)'\overline{d_X(\omega)}$ and (2) applies Lemma 1.1, where $W_T(u) = (1-|u/T|)\,e^{-i\omega u} \xrightarrow{T\to\infty} e^{-i\omega u}$ pointwise in u. Now, we can use Remark 1.4 to interpret this calculation. Recall that $d_X(\omega)$ can

be thought of as $\sqrt{T}$ times the coefficient in the projection of X onto the series $e(\omega)$. We have that

$$\begin{aligned}
\|d_X(\omega)\|^2 &= \left\| \sqrt{\frac{1}{T}} \sum_{t=1}^{T} X_t e^{-i\omega t} \right\|^2 \\
&= \left\| \sqrt{\frac{1}{T}} \sum_{t=1}^{T} X_t (\cos(\omega t) - i \sin(\omega t)) \right\|^2 \\
&= \left(\sqrt{\frac{1}{T}} \sum_{t=1}^{T} X_t (\cos(\omega t) - i \sin(\omega t)) \right) \left(\sqrt{\frac{1}{T}} \sum_{t=1}^{T} X_t (\cos(\omega t) + i \sin(\omega t)) \right) \\
&= \left(\sqrt{\frac{1}{T}} \sum_{t=1}^{T} X_t \cos(\omega t) \right)^2 + \left(\sqrt{\frac{1}{T}} \sum_{t=1}^{T} X_t \sin(\omega t) \right)^2
\end{aligned}$$

Now, consider the choice $\omega = \frac{2\pi j}{T}$ for some integer j. Consider the regression of X_t onto $\frac{1}{\sqrt{j\pi}} \cos(\omega t)$ and $\frac{1}{\sqrt{j\pi}} \sin(\omega t)$ for this choice of ω. We have that

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{T} \sum_{t=1}^{T} (j\pi)^{-1} \cos^2(\omega t) & \frac{1}{T} \sum_{t=1}^{T} (j\pi)^{-1} \cos(\omega t) \sin(\omega t) \\ \frac{1}{T} \sum_{t=1}^{T} (j\pi)^{-1} \cos(\omega t) \sin(\omega t) & \frac{1}{T} \sum_{t=1}^{T} (j\pi)^{-1} \sin^2(\omega t) \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{T} \sum_{t=1}^{T} \frac{1}{\sqrt{j\pi}} X_t \cos(\omega t) \\ \frac{1}{T} \sum_{t=1}^{T} \frac{1}{\sqrt{j\pi}} X_t \sin(\omega t). \end{pmatrix}$$

Notice that as T grows large, $\frac{1}{T} \sum_{t=1}^{T} (j\pi)^{-1} \cos^2(\omega t) \to (j\pi)^{-1} \int_0^{2\pi j} \cos^2(z) dz = 1$, $\frac{1}{T} \sum_{t=1}^{T} (j\pi)^{-1} \sin^2(\omega t) \to (j\pi)^{-1} \int_0^{2\pi j} \sin^2(z) dz = 1$, $\frac{1}{T} \sum_{t=1}^{T} (j\pi)^{-1} \sin(\omega t) \cos(\omega t) \to (j\pi)^{-1} \int_0^{2\pi j} \sin(z) \cos(z) dz = 0$. So, for large T, we have that

$$\sqrt{j\pi}\hat{\beta} \approx \begin{pmatrix} \frac{1}{T} \sum_{t=1}^{T} X_t \cos(\omega t) \\ \frac{1}{T} \sum_{t=1}^{T} X_t \sin(\omega t). \end{pmatrix}$$

So, with this calculation, we can interpret $\|d_X(\omega)\|^2$ as

$$\|d_X(\omega)\|^2 \approx \pi j T \hat{\beta}_1^2 + \pi j T \hat{\beta}_2^2,$$

In the second calculation, we will consider the "off-diagonal" cross-terms. For this calculation, additionally assume that $\omega_j = \frac{2\pi j}{T}$ and we'll compute $\frac{1}{2\pi} \mathbb{E}\left[d_X(\omega_j) \overline{d_X(\omega_k)} \right]$. We have that

$$\begin{aligned}
\frac{1}{2\pi} \mathbb{E}\left[d_X(\omega_j)' \overline{d_X(\omega_k)} \right] &= \frac{1}{2\pi} \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} \mathbb{E}\left[X_t X_s \right] e^{-i\frac{2\pi j}{T} t} e^{i\frac{2\pi k}{T} s} \\
&= \frac{1}{2\pi} \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} \mathbb{E}\left[X_t X_s \right] e^{-i\frac{2\pi j}{T}(t-s)} e^{i\frac{2\pi(k-j)}{T} s}
\end{aligned}$$

Now, let $u = t - s$. Then, substituting this in, we get that

$$\frac{1}{2\pi} \mathbb{E}\left[d_X(\omega_j) \overline{d_X(\omega_k)} \right] = \frac{1}{2\pi} \sum_{u=-(T-1)}^{T-1} \gamma_u e^{-i\frac{2\pi j}{T} u} \left(\frac{1}{T} \sum_{s=1}^{T} e^{i\frac{2\pi(k-j)}{T} s} \mathbb{1}\left\{ 1 \le s, s+u \le T \right\} \right),$$

where

$$W_T(u) = e^{-i\frac{2\pi j}{T}u}\left(\frac{1}{T}\sum_{s=1}^{T} e^{i\frac{2\pi(k-j)}{T}s}\mathbb{1}\{1 \leq s, s+u \leq T\}\right) \to \int_0^1 e^{-i(2\pi)(j-k)r}\,dr = 0$$

as $T \to \infty$ because $j \neq k$. Therefore, applying Lemma 1.1, we see that the diagonal terms converge to zero in expectation as $T \to \infty$.

Remark 1.5. *Where does this choice of ω_j come from? For $j = 1, \ldots, q$, consider the discrete series*

$$e(j)' = \left(e^{-i\frac{2\pi j}{T}\cdot 1}, e^{-i\frac{2\pi j}{T}\cdot 2}, \ldots, e^{-i\frac{2\pi j}{T}\cdot T}\right).$$

Notice that for $j \neq k$, these series are orthogonal. That is,

$$\begin{aligned}\langle e(j), e(k)\rangle &= \sum_{t=1}^{T} e^{-i\frac{2\pi j}{T}t}e^{i\frac{2\pi k}{T}t} \\ &= \sum_{t=1}^{T} e^{-i\frac{2\pi}{T}(j-k)t} \\ &\overset{(1)}{=} \frac{1 - e^{-i(2\pi)(j-k)}}{1 - e^{-i\frac{2\pi}{T}(j-k)}}.\end{aligned}$$

For $j \neq k$, Euler's formula shows that the numerator equals zero and the denominator is non-zero. Therefore, these are orthogonal if $j \neq k$. Next, notice that for $j = k$, we have that

$$\langle e(j), e(j)\rangle = \sum_{t=1}^{T} e^{-i\frac{2\pi j}{T}t}e^{i\frac{2\pi j}{T}t} = T.$$

Collect these series into a $T \times q$ matrix

$$E = \begin{pmatrix} e(1) & e(2) & \ldots & e(q)\end{pmatrix}$$

By the calculation above, we have shown that

$$E'E = T \cdot I_q.$$

Therefore, the coefficients in the projection of X onto the matrix E are defined as

$$(E'E)^{-1}E'X = \begin{pmatrix} \frac{1}{T}\sum_{t=1}^{T} X_t e^{-i\frac{2\pi}{T}t} \\ \frac{1}{T}\sum_{t=1}^{T} X_t e^{-i\frac{2\pi}{T}\cdot 2\cdot t} \\ \vdots \\ \frac{1}{T}\sum_{t=1}^{T} X_t e^{-i\frac{2\pi}{T}\cdot T\cdot t}\end{pmatrix}$$

By this construction, we have that $\sqrt{T}$ times the coefficient on the series $e(j)$ in the projection of X_t onto E is given by $d_X(\omega_j)$.

1.4.2 The spectrum and long-run variance

Another the spectrum is an important object is because of its connection to the **long-run variance** of the time series. In particular, consider the variance of the sample mean, $\bar{X} = T^{-1}\sum_{t=1}^{T} X_t$. Assume that X_t is mean-zero for simplicity, and so we have that

$$\begin{aligned} T \cdot V(\bar{X}) &= T\mathbb{E}\left[\left(\frac{1}{T}\sum_{t=1}^{T} X_t\right)^2\right] \\ &= T\left(T^{-2}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}\left[X_t X_s\right]\right) \\ &= \sum_{-(T-1)}^{T-1}(1-|\frac{u}{T}|)\gamma_u \to \sum_{-\infty}^{\infty}\gamma_u = S_X(0). \end{aligned}$$

We refer to the spectrum evaluated at zero as the **long-run variance**. Constructing an estimator of this object will be central to constructing standard -errors in time series settings.

1.5 Linear filtering

A **linear filter** $b(L)$ is a linear operator on the time series X_t. The filtered time series is simply

$$Y_t = b(L)X_t.$$

It is much easier to understand what a filter is doing by analyzing its behavior in the frequency domain.

Lemma 1.2. *Let X_t be a second-order stationary time series and define $Y_t = b(L)X_t$. Then,*

$$S_Y(\omega) = \|b(e^{i\omega}\|^2 S_X(\omega),$$

*where $\|b(e^{i\omega}\|^2$ is called the **gain** of the filter.*

Proof. Apply the Wold Decomposition to write $X_t = c(L)\varepsilon_t$. Then, $Y_t = b(L)c(L)\varepsilon_t$, and so Y_t is a linear filter of the Wold innovations and just a moving average. Therefore, we know that

$$\begin{aligned} S_X(\omega) &= \frac{1}{2\pi}\|c(e^{i\omega}\|^2\sigma_\epsilon^2 \\ S_Y(\omega) &= \frac{1}{2\pi}\|b(e^{i\omega}c(e^{i\omega}\|^2\sigma_\epsilon^2, \end{aligned}$$

and the result follows directly. □

Lemma 1.2 is stated for a second-order stationary time series but can be generalized to cover non-stationary time series processes. We assumed second-order stationarity to simplify the proof (it allowed us to use the Wold decomposition). Why is this an interesting result? We can think of Y_t as heightening or dampening the spectrum of X_t depending on what gain $\|b(e^{i\omega}\|^2$ that is selected. In other words, the filter will screen out variation in the time series X_t at different frequencies according

to what gain function is selected. In macroeconomic applications, we often select a filter that is associated with a gain that will filter out high frequency variation in the time series, which we often think is measurement error and noise. We also may select a filter to select only the variation in the time series that is associated with business cycles.

Example 1.6. *Year-over-year filter Suppose that X_t is a quarterly time series. The year-over-year filter is defined as*

$$Y_t = \frac{X_t + X_{t-1} + X_{t-2} + X_{t-3}}{4} - \frac{X_{t-4} + X_{t-5} + X_{t-6} + X_{t-7}}{4}.$$

It is simply the difference in the average of the time series over the last year relative to one year ago. We can write this more compactly as

$$Y_t = \frac{1}{4}(1 - L^4)(1 + L + L^2 + L^3)X_t.$$

Example 1.7. *Baxter-King Filter Let $b(L) = \sum_{k=-q}^{q} b_k L^k$ be a forward and backward looking lag polynomial. Let $[\omega_0, \omega_1] \subseteq [0, \pi]$ be the range of frequencies of interest. The **Baxter-King filter** chooses the coefficients of $b(L)$ to solve*

$$b^* = \min_b \int_0^\pi \left(\|b(e^{i\omega})\|^2 - \mathbb{1}\{\omega_0 \leq \omega \leq \omega_1\} \right) d\omega.$$

That is, the ideal filter would simply pick out the frequencies between ω_0, ω_1. The Baxter-King filter computes the moving average whose gain provides the best mean-square approximation to the ideal "band-pass" filter. The filtered series is then

$$Y_t = b^*(L)X_t.$$

Example 1.8. *Butterworth Filter Let $a(L) = \sum_{k=0}^{q} a_k L^k$ be a lag polynomial and define $a(L)^{-1}$ to be its associated inverse. Again, define $[\omega_0, \omega_1] \subseteq [0, \pi]$ be the range of frequencies of interest. The **Butterworth filter** solves an analogous problem to the Baxter-King filter*

$$\min_a \int_0^\pi \left(\|a(e^{i\omega})^{-1}\|^2 - \mathbb{1}\{\omega_0 \leq \omega \leq \omega_1\} \right) d\omega.$$

That is, the Butterworth filter computes the auto-regression whose gain provides the best mean-square approximation to the ideal band-pass filter. The filter series satisfies

$$a^*(L)Y_t = X_t.$$

Example 1.9. *The **Hodrick-Prescott filter** solves*

$$\min_{g_t} \sum_{t=1}^{T} (X_t - g_t)^2 + \lambda \sum_{t=1}^{T} (\Delta g_t - \Delta g_{t-1})^2,$$

where $\lambda > 0$ is a tuning parameter that determines its smoothness penalty, which penalizes g_t based on a

discrete version of the second derivative. Hamilton (2018) recently notes several problems with the Hodrick-Prescott filter: (1) it often introduces spurious dynamics and (2) The fit of the resulting filtered values at the end of the sample are quite poor.

Moreover, Hamilton (2018) shows that the HP filter is the optimal solution to a particular signal extraction problem. Suppose that $X_t = g_t + c_t$, where g_t satisfies $\Delta g_t = \Delta g_{t-1} + \nu_t$, $\nu_t \sim (0, \sigma_\nu^2)$ and $c_t \sim (0, \sigma_c^2)$ The HP filter is the optimal filter for extracting the unobserved component g_t in this model, where the penalty parameter λ is set as a function of the variances of the noise σ_ν^2, σ_c^2. While it is nice that the HP filter can be given this interpretation, it is unclear whether this signal extraction problem is particularly relevant to the problem of extracting variation associated business cycle frequencies in macroeconomic data.

1.6 Multivariate extensions

We provide simple extensions of our results for second-order stationary time series to a random vector. A more detailed discussion of these ideas can be found in Chapter 10 of Hamilton (1994) and Chapter 11 of Brockwell and Davis (1991). Let X_t be an $n \times 1$ random vector. We say that X_t is **second-order stationary** or **covariance stationary** or **weakly stationary** if its mean vector is time invariant and its auto-covariance matrices are also time-invariant. That is,

$$\begin{aligned} \mathbb{E}[X_t] &= \mu, \\ Cov(X_t, X_{t-j}) = \mathbb{E}\left[(X_t - \mu)(X_{t-j} - \mu)'\right] &= \Gamma_j \quad \forall t. \end{aligned}$$

Note that for a covariance stationary vector process, we have that

$$\Gamma_j' = \Gamma_{-j}. \tag{1}$$

Why is this the case? We have that

$$Cov(X_{t+j}, X_t) = \mathbb{E}\left[(X_{t+j} - \mu)(X_t - \mu)'\right] = \Gamma_j \implies \Gamma_j' = \mathbb{E}\left[(X_t - \mu)(X_{t+j} - \mu)'\right] = \Gamma_{-j}.$$

Importantly, we can define the **population spectrum** for a multivariate time series as well. If the autocovariances are absolutely summable, then the **autocovariance generating function** is

$$\Gamma(z) = \sum_{k=-\infty}^{\infty} \Gamma_k z^k,$$

where z is a complex scalar. Note that this maps a complex scalar into a $K \times K$ matrix of complex numbers. Then, with the same calculations as the univaraite case, we define the **population spectrum** as

$$\begin{aligned} S_X(\omega) &= \frac{1}{2\pi}\Gamma(e^{-i\omega}) \\ &= \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} \Gamma_k e^{-i\omega k}. \end{aligned}$$

The population spectrum has the same properties as before. For example, we can show that

$$\int_{-\pi}^{\pi} S_X(\omega) e^{i\omega k} \, d\omega = \Gamma_k.$$

For a complex number z, let $\bar{z}$ denote its complex conjugate. Moreover, we have that the multivariate population spectrum satisfies

$$\begin{aligned}
\overline{S_X(\omega)}' &= \frac{1}{2\pi} \overline{\Gamma_X(e^{-i\omega})}' \\
&= \frac{1}{2\pi} \Gamma_X(e^{i\omega})' \\
&\overset{(1)}{=} \frac{1}{2\pi} \Gamma_X(e^{-i\omega}) = S_X(\omega),
\end{aligned}$$

where (1) used the fact that $\Gamma_j = \Gamma'_{-j}$ for a covariance stationary vector process.

Example 1.10. *Let $X_t = c(L)\epsilon_t$ be a vector moving average. Then,*

$$S_X(\omega) = \frac{1}{2\pi} c(e^{i\omega}) \Sigma_\epsilon c(e^{-i\omega})',$$

where $\Sigma_\epsilon = \mathbb{E}[\epsilon_t \epsilon_t']$.

Similarly, the long-run variance of a second-order stationary vector process is

$$\begin{aligned}
\Omega &= \lim_{T \to \infty} 2\pi T \cdot V(\bar{X}) \\
&= \Gamma(1) = 2\pi S_X(0) = \sum_{j=-\infty}^{\infty} \Gamma_j.
\end{aligned}$$

We can also derive a multivariate extension of the Wold Decomposition. For a second-order stationary vector process, this will take the form

$$\underbrace{X_t}_{n \times 1} = \underbrace{c(L)}_{n \times n} \underbrace{\varepsilon_t}_{n \times 1},$$

where ε_t is the vector of **Wold innovations** and defined as $\varepsilon_t = X_t - \text{Proj}\{X_t \mid X_{t-1}, X_{t-2}, \ldots\}$.

Finally, we define a **vector auto-regression**. We say that X_t is a **vector autoregression of order** p, $VAR(p)$ if

$$A(L)X_t = \epsilon_t, \quad A(L) = A_0 - A_1 L - A_2 L^2 - \ldots - A_p L^p.$$

1.7 A central limit theorem for weakly dependent processes

To this point, we focused on population objects and have not considered estimation and inference. We will now develop a central limit theorem for a stationary stochastic time series. This will be one of main tools for constructing tests and confidence intervals in time series. To do so, we will make sufficient assumptions such that the familiar asymptotic results from cross-sectional settings with independent data apply to a second-order stationary time series. Intuitively, these assumptions will place restrictions on the degree of dependence in the time series process. Our discussion in this sec-

tion follows closely the presentation on pp. 402–405 in Hayashi (2000).

Definition 1.6. *Let $\{Z_t\}$ be a time series. We say it is **ergodic** if for any two bounded functions $f : \mathbb{R}^k \to \mathbb{R}$ and $g : \mathbb{R}^l \to \mathbb{R}$*

$$\lim_{n\to\infty} |\mathbb{E}\left[f(Z_t, \ldots, Z_{t+k})g(Z_{t+n}, \ldots, Z_{t+n+l})\right] = |\mathbb{E}\left[f(Z_t, \ldots, Z_{t+k})\right]\mathbb{E}\left[g(Z_{t+n}, \ldots, Z_{t+n+k})\right]$$

In words, ergodicity means that at a long enough horizon, the time series process becomes essentially independent. This places restrictions on the degree of dependence over time and enforces that it dies off at a long enough horizon, enabling us to provide a central limit theorem for averages of a time series as T grows large.

We now state a central theorem for a stationary, ergodic stochastic process.

Theorem 1.2. *Let Z_t be a mean-zero, stationary and ergodic stochastic process. Suppose that Z_t satisfies **Gordin's conditions**:*

1. $\mathbb{E}[Z_t^2] < \infty$,
2. $\mathbb{E}\left[\mathbb{E}(Z_t|Z_{t-j}, Z_{t-j-1}, \ldots)\right] \to 0$ *as* $j \to \infty$,
3. $\sum_{j=0}^{\infty} \mathbb{E}[r_{tj}^2]^{1/2} < \infty$, *where* $r_{tj} = \mathbb{E}[Z_t|Z_{t-j}, Z_{t-j-1}, \ldots] - \mathbb{E}[Z_t|Z_{t-j-1}, Z_{t-j-2}, \ldots]$.

Then,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \xrightarrow{d} N(0, 2\pi S_Z(0))$$

as $T \to \infty$, where

$$2\pi S_Z(0) = \sum_{j=-\infty}^{\infty} \gamma_j$$

*is known as the **long-run variance** of Z_t.*

Remark 1.6. *Gordin's condition limit the degree of serial dependence in the series Z_t. Condition (1) is simply a finite variance condition. Condition (2) states that we allow serial dependence but impose that the dependence dies off as periods become further apart. In other words, data from j periods ago becomes uninformative in forcasting Z_t as j grows large. Condition (3) imposes that moving from time $t-j-1$ to $t-j$ reveals information about time t but again, this dies off as j gets large.*

Example 1.11. *Suppose that the random variables Z_t are i.i.d. across t. Then, Gordin's conditions hold trivially.*

Example 1.12. *Suppose that Z_t is a martingale difference sequence. Then, since $\mathbb{E}\left[Z_t \mid Z_{t-j}, \ldots\right] = 0$, Condition (2) is satisfied in Gordin's conditions. Moreover, Condition (3) is satisfied because $r_{tj} = 0$ for all $j > 0$.*

Example 1.13. *Suppose that Z_t is a moving average of order $q < \infty$, meaning $Z_t = c(L)\epsilon_t$, where ϵ_t is a martingale difference sequence. Then, once $j > q$, the dependence in the series dies off completely and so, Condition (2) and Condition (3) are satisfied.*

Example 1.14. *Suppose that Z_t is an infinite-order moving average, with $Z_t = c(L)\epsilon_t$, where ϵ_t is a martingale difference sequence and $c(L)$ is invertible. First, we will check Condition (2).*

$$\begin{aligned}\mathbb{E}\left[Z_t \mid Z_{t-j}, \ldots\right] &\overset{(1)}{=} \mathbb{E}\left[c(L)\epsilon_t \mid \epsilon_{t-j}, \ldots\right] \\ &= \sum_{l=j}^{\infty} c_l \epsilon_{t-l},\end{aligned}$$

where we were able to changed the conditioning in (1) because $c(L)$ is invertible. Therefore, we have that

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\left[Z_t \mid Z_{t-j}, \ldots\right]^2\right] &= \mathbb{E}\left[\left(\sum_{l=j}^{\infty} c_l \epsilon_{t-l}\right)^2\right] \\ &= \sum_{l=j}^{\infty} c_l^2 \sigma_\epsilon^2.\end{aligned}$$

Therefore, if the coefficients in the lag polynomial $c(L)$ are square summable meaning $\sum_{j=0}^{\infty} c_j^2 < \infty$, then Condition (2) is satisfied. Next, note that $r_{tj} = c_j \epsilon_{t-j}$ and $\mathbb{E}[r_{tj}^2] = c_j^2 \sigma_\epsilon^2$. Then,

$$\sum_{j=1}^{\infty} \sqrt{\mathbb{E}[r_{tj}^2]} = \sum_{j=1}^{\infty} |c_j| \sigma_\epsilon.$$

So, now we additionally need that $\{c_t\}$ be absolutely summable, meaning that $\sum_{j=0}^{\infty} |c_j| < \infty$.

There are several important comments to make. First, absolute summability of the coefficients in the lag polynomial is a stronger condition than square summability. It is simple to see this because

$$\left(\sum_{j=0}^{\infty} |c_j|\right)^2 = \sum_{j=0}^{\infty} c_j^2 + \sum_{i \neq j} |c_i||c_j| \geq \sum_{j=0}^{\infty} c_j^2.$$

Second, it was crucial that ϵ_t was a martingale difference sequence as Gordin's conditions are about conditional expectations, not projections. Together, this implies that Gordin's conditions are not satisfied by the Wold decomposition. The Wold decomposition applies to any stationary stochastic process and delivers a moving-average form of Wold innovations, where the Wold innovations are serially uncorrelated and the coefficients in the lag polynomial are square summable.

Example 1.15. *Consider $Y_t = X_t\beta + U_t$ and define $Z_t = X_t U_t$. Assume that $\mathbb{E}[U_t|X_t] = 0$ and so, $\mathbb{E}[Z_t] = 0$. Then, assuming that Z_t satisfies Gordin's conditions,*

$$\sqrt{T}\left(\hat{\beta} - \beta\right) \xrightarrow{d} N(0, \Sigma_X^{-1} \Omega \Sigma_X^{-1}),$$

where

$$\Sigma_X = \mathbb{E}[X_t^2], \quad \Omega = 2\pi S_Z(0) = \sum_{j=-\infty}^{\infty} \Gamma_j.$$

1.8 Auto-regressions, lag-length selection and information criteria

Consider an autoregression

$$X_t = a_1 X_{t-1} + \ldots + a_{p_0} X_{t-p_0} + u_t.$$

In general, the lag-length p is unknown and may, in fact, be infinite, $p = \infty$. These pose particular challenges and we will now discuss each case in turn.

First, suppose that the time series is generated by an infinitely long autoregression and $p_0 = \infty$. Clearly, in finite samples, we can not estimate an infinitely long autoregression and we must truncate p somewhere. Moreover, we would like to choose this truncation rule as a function of the size of the data. In other words, when T is small, we do not want to include many lags but as T grows large, we want to add more and more lags. How do we choose this truncation rule in a way that will deliver good asymptotic properties of the resulting estimator? This problem was studied by Berk (1974).[3] Berk (1974) focuses on the problem of constructing an estimator of the spectral density of a time series $S_X(\omega)$ by using a sequence of auto-regressive approximations. We derived earlier that for an auto-regression, $a_p(L)X_t = \epsilon_t$, its spectral density is $S_{X,p}(\omega) = \frac{\sigma_\epsilon^2/2\pi}{\|a_p(e^{i\omega})\|^2}$. As p grows large, perhaps $S_{X,p}(\omega)$ provide a better and better approximation of the population spectrum, $S_X(\omega)$. This is turns out to be true provided that p is chosen correctly as a function of the sample size T.

Theorem 1.3 (Theorem 1 in Berk (1974))**.** *Suppose that $a(L)X_t = \epsilon_t$, where $a(z)^{-1} \neq 0$ for all $|z| < 1$. Assume that*

1. *$a(e^{i\omega}) \neq 0$ for $-\pi \leq \omega \leq \pi$,*
2. *$\mathbb{E}\left[\epsilon_t^4\right] < \infty$,*
3. *$p_T^3/T \to 0$ and $p_T \to \infty$ as $T \to \infty$,*
4. *$\sqrt{p_T}\sum_{j=p_T+1}^{\infty} |a_j| \to 0$ as $T \to \infty$.*

Then,

$$\sup_{\omega} |\hat{S}_{X,p_T}(\omega) - S_X(\omega)| \xrightarrow{p} 0,$$
$$\sum_{j=0}^{\infty} |\hat{a}_j - a_j| \xrightarrow{p} 0.$$

Condition (3) implies that you must set p_T to grow at a slower rate than $T^{1/3}$. If this and the other conditions are satisfied, then the resulting approximation to the spectral density is *uniformly* consistent. Moreover, Berk (1974) additionally shows that this approximation converges in distribution to a normal distribution centered at the population spectrum pointwise.

Next, suppose that the time series is generated from a *finite* order autoregression, but $p_0 \leq \bar{p}$ is unknown. How do we select the correct lag-length? A common strategy is to use an **information criterion**. We will begin by heuristically deriving the **Akaike Information Criterion** (AIC) as motivation and then, discuss information criteria more generally.

[3] This is an early application of a now common idea in statistics and econometrics. Intuitively, we approximate an underlying population model with a sequence of increasingly, complex parametric models. This is a common technique in non-parametrics.

Suppose that we are interested in constructing a one-step ahead forecast of the time series X_t using an autoregression. Assume that X_t follows an auto-regression of order p and we wish to construct an order p autoregression

$$\begin{aligned} X_t &= \beta_1 X_{t-1} + \ldots + \beta_p X_{t-p} + \epsilon_t \\ &= \beta' \underline{X}_t + \epsilon_t, \end{aligned}$$

where $\underline{X}_t = (X_{t-1}, \ldots, X_{t-p})'$. Our loss function is simply squared-loss

$$L(\hat{\beta}) = \left(X_{T+1} - \hat{\beta}' \underline{X}_T\right)^2,$$

and so, the risk function is **one-step ahead mean-squared prediction error**,

$$MSPE(\hat{\beta}) = \mathbb{E}\left[\left(X_{T+1} - \hat{\beta}' \underline{X}_T\right)^2\right].$$

We can re-write the one-step ahead mean-square prediction error as

$$\begin{aligned} MSPE(\hat{\beta}) &= \mathbb{E}\left[\left(\epsilon_{T+1} - (\hat{\beta} - \beta)' \underline{X}_T\right)^2\right] \\ &= \mathbb{E}\left[\epsilon_{T+1}^2\right] - \mathbb{E}\left[(\beta - \hat{\beta})' \underline{X}_T \underline{X}_T' (\beta - \hat{\beta})\right]. \end{aligned}$$

We know that as T grows large, $\hat{\beta} - \beta \approx N(0, \Sigma_{XX}^{-1} \frac{\sigma_\epsilon^2}{T}$, where $\Sigma_{XX}^{-1} = \mathbb{E}\left[\underline{X}_T \underline{X}_T'\right]$. Now, simply assume that the estimation error $\hat{\beta} - \beta$ is independent of $\underline{X}_T$. Intuitively, think of this as a situation where the data used to estimate $\hat{\beta}$ is sufficiently far in the past that it is approximately independent of the data we are using to evaluate its predictions.[4] Then, under this assumption, it follows that we can re-write the one-step ahead mean-square prediction error as

$$MSPE(\hat{\beta}) \approx \sigma_\epsilon^2 - \frac{\sigma_\epsilon^2}{T} \mathbb{E}\left[Z'Z\right],$$

where $Z \sim N(0, I_p)$ and so, $Z'Z \sim \chi_p^2$. Therefore, we arrive ta

$$MSPE(\hat{\beta}) \approx (1 + \frac{p}{T}) \sigma_\epsilon^2.$$

We still cannot operationalize this expression for estimation as σ_ϵ^2 is unknown. Let's replace it with an estimate, $\hat{\sigma}_\epsilon^2 = \frac{T}{T-p} \frac{SSR}{T}$, where SSR is the usual sum of squared residuals and $\frac{T}{T-p}$ is the usual degrees of freedom correction. Substituting this in, we can write,

$$\begin{aligned} &MSPE(\hat{\beta}) \approx (1 + \frac{p}{T}) \frac{SSR/T}{1 - p/T}, \\ &\log\left(MSPE(\hat{\beta})\right) \overset{(1}{\approx} \log(SSR/T) + \log\left(\frac{1 + p/T}{1 - p/T}\right) \approx \log(SSR/T) + 2\frac{p}{T}, \end{aligned}$$

[4] If this were a cross-sectional setting, we would be thinking of this prediction exercise as occurring on an "hold-out dataset" that is independent of the "training dataset."

where (1) follows after a first-order Taylor approximation of $\log(1+z) - \log(1-z)$ around 1. This final expression is the **Akaike Information Criterion** (AIC),

$$AIC(p) = \log(SSR_p/T) + 2\frac{p}{T},$$

where SSR_p is the sum-of-squared residuals produced by an $AR(p)$. It is constructed via this heuristic calculation that finds a simple expression for the one-step-ahead mean-square prediction error. The AIC then selects the optimal lag-length p^* by minimizing $AIC(p)$ over a pre-specified grid $p = 0, \ldots, \bar{p}$.

More generally, an **information criterion** takes the form

$$IC(p) = \log(SSR_p/T) + pg(T),$$

where $g(T)$ is some function of the sample size. Think of $pg(T)$ as a penalty function that is increasing in the number of lags included. We select the lag length $\hat{p}$ that minimizes the information criterion in-sample

$$\hat{p} = \min_{p=0,\ldots,\bar{p}} IC(p).$$

Intuitively, adding more lags will always reduce the sum of squared residuals in sample but it may be fitting noise. This introduces a bias-variance tradeoff and the penalty $pg(T)$ tries to capture this tradeoff in a reduced-form manner. We just saw that the AIC sets

$$g(T) = \frac{2}{T}.$$

Another common choice is the **Bayes information criteria** (BIC), which sets

$$g(T) = \frac{\log(T)}{T}.$$

Of course, the obvious question now is: How do we selection $g(T)$? If we wish to select an information criterion that produces consistent model selection, meaning that we select the correct lag length with probability converging to one as $T \to \infty$, then there is a simple answer.

Theorem 1.4. *Assume that $g(T) \to 0$ and $Tg(T) \to \infty$ and $X_t \sim AR(p_0)$ with $p_0 \leq \bar{p}$. Then,*

$$\mathbb{P}\{\hat{p} = p_0\} \xrightarrow{p} 1m$$

as $T \to \infty$.

Proof. We provide a heuristic sketch of the proof. First, we show that $\mathbb{P}\{\hat{p} < p_0\} \to 0$ as $T \to \infty$. We

have that

$$\begin{aligned}\mathbb{P}\left\{\hat{p} < p_0\right\} &\leq \mathbb{P}\left\{\min_{p=0,\ldots,p_0-1} IC(p) < IC(p_0)\right\} \\ &= \mathbb{P}\left\{\min_{p=0,\ldots,p_0-1} \log(SSR_p/T) + pg(T) < \log(SSR_{p_0}/T) + p_0 g(T)\right\} \\ &= \mathbb{P}\left\{\min_{p=0,\ldots,p_0-1} \log(SSR_p/SSR_{p_0}) + (p-p_0)g(T) < 0\right\},\end{aligned}$$

where $SSR_p/SSR_{p_0} \to pc > 1$ and $(p-p_0)g(T) \to 0$. So, we conclude that this probability will also go to zero as $T \to \infty$. Next, we show that $\mathbb{P}\left\{\hat{p} > p_0\right\} \to 0$ as $T \to \infty$. By the same algebra, we arrive

$$\begin{aligned}\mathbb{P}\left\{\hat{p} > p_0\right\} &\leq \mathbb{P}\left\{\min_{p=p_0+1,\ldots,\bar{p}} \log(SSR_p/SSR_{p_0}) + (p-p_0)g(T) < 0\right\}, \\ &= \mathbb{P}\left\{\min_{p=p_0+1,\ldots,\bar{p}} 2T\log(SSR_p/SSR_{p_0}) + 2(p-p_0)Tg(T) < 0\right\},\end{aligned}$$

where $2T\log(SSR_p/SSR_{p_0}) \approx \chi^2_{p-p_0}$ because it is the log of the likelihood ratio test statistic for the null hypothesis that the $p_0+1,\ldots,p$ lags have zero coefficients. Therefore, the first-term is bounded in probability, while the second term diverges to infinity. So, we conclude that the probability of $\hat{p} > p_0$ will also go to zero. □

That is, any choice of $g(T)$ that grows smaller at a slower rate that T^{-1} will provide consistent model selection. This immediately implies that BIC provides consistent model selection but AIC will not. It can be shown that AIC will tend to select models that are too large, meaning that $\hat{p}^{AIC} > p_0$ with non-zero probability in large samples.

2 HAC/HAR Inference

HAC is an acronym for "**heteroskedasticity auto-correlation consistent**" and HAR is an acronym for **heteroskedasticity auto-correlation robust**. These describe two different approaches to constructing standard errors in time series settings. As we saw earlier, in order to construct confidence intervals and standard errors, we need to estimate the long-run variance of a time-series

$$\Omega = \sum_{j=-\infty}^{\infty} \Gamma_j.$$

It is difficult to estimate this object well in finite-samples as it depends on infinitely many auto-covariances. Heteroskedasticity auto-correlation consistent (HAC) inference focuses on constructing a consistent estimate of the long-run variance Ω, whereas heteroskedasticity auto-correlation robust (HAR) inference will allow $\hat{\Omega}$ to be inconsistent in the hopes of delivering better finite-sample performance. Recall that in classical inference, we wish to control the **size** of our hypothesis tests and then, among all tests of the same size, we wish to select the test that maximizes **power**. In this setting, we will not be able to control size exactly and so, there will be a **size-power tradeoff**.

The canonical HAC estimator is the **Newey-West estimator**

$$\hat{\Omega}^{NW} = \sum_{j=-s}^{s} (1 - |j/s|)\hat{\Gamma}_j,$$

proposed in Newey and West (1987). This only uses finitely many estimated auto-covariances and weights them with the kernel, $(1 - |j/s|)$. This kernel is known as the *triangular* or *bartlett* kernel. It is chosen to ensure that the resulting estimate $\hat{\Omega}$ is positive semi-definite. Alternative choices for the kernel will deliver alternative estimators. That is,

$$\hat{\Omega}^{SC} = \sum_{j=-s}^{s} k(|j/s|)\hat{\Gamma}_j,$$

and we will refer to these as **time domain** or **covariance domain** estimators. For a time domain estimator, the researcher must specify a kernel k and a cut-off rule as a function of the sample size $s(T)$. Alternatively, we could consider estimating Ω in the **frequency domain**. A frequency domain estimator takes the form

$$\hat{\Omega}^{WP} = \frac{1}{M} \sum_{j=1}^{M} K(j/M) \|d_Z(\omega_j)\|^2,$$

where $\omega_j = 2\pi j/T$, $d_Z(\omega) = \sqrt{1/T} \sum_{t=1}^{T} Z_t e^{-i\omega t}$ and

$$\|d_Z(\omega_j)\|^2 = d_Z(\omega_j)\overline{d_Z(\omega_j)}'$$

is the **periodogram** of Z_t at frequency ω_j. $K(\cdot)$ is a kernel in the frequency domain and M is a truncation parameter. For a frequency domain estimator, the researcher must specify a kernel $K(\cdot)$ and and a cutoff rule as a function of the sample size $M(T)$. How do we do chooses these objects to control size and maximize power?

There are two key ideas that have come out of this literature on HAC/HAR inference. First, selecting a larger truncation parameter $S(T)$ in the time domain or equivalently, a smaller bandwidth $M(T)$ in the frequency domain improves the size of tests. Why is this? Intuitively, there is a bias-variance tradeoff in estimating the long-run variance. For the time-domain estimator, including few estimated auto-covariances leads to a lower variance estimator with high bias. Choosing too small of a truncation parameter $S(T)$ corresponds to choosing a low variance, high bias estimator of the long-run variance and this will lead to standard errors that tend to be too small. But, pushing back in the other direction, is that including many auto-covariances will lead to a higher variance estimator of the long-run variance. This leads to the second key idea, which is that this problem of a high variance estimator of the long-run variance can be solved by using non-standard critical values. These are known as **fixed-b critical values**. We will discuss these more later on.

Finally, throughout this section, we provide heuristic derivations to focus on intuition. Many of the ideas and results in this section are stated formally in Lazarus et al. (2018) and Lazarus, Lewis and Stock (2018).

2.1 Null rejection rate expansion

We'll assume **exact normality** for now. We wish to compute

$$\mathbb{P}_{H_0}\left\{t^2 > c\right\} = \mathbb{P}_{H_0}\left\{\frac{(\hat{\beta}-\beta_0)^2}{\hat{\Omega}/\hat{\sigma}_x^4} > c\right\}.$$

What's the null hypothesis we are testing? Consider the model

$$y_t = x_t\beta + u_t,$$

and we wish to test the null hypothesis, $H_0 : \beta = \beta_0$. For simplicity, we assume that we are in the scalar case with $x_t \in \mathbb{R}$ and $x_t \overset{as}{=} 1$, u_t normally distributed. In other words, we wish to test whether the mean of y_t equals some constant. The model becomes

$$y_t = \beta + u_t, \quad u_t \sim \text{stationary normal}.$$

We estimate Ω using a frequency domain estimator with

$$\hat{\Omega} = \frac{1}{M}\sum_{j=1}^{M} K(j/M)\left\|d_z(\frac{2\pi j}{T})\right\|^2,$$
$$d_z(\omega_j) = \sqrt{\frac{1}{T}}\sum_{t=1}^{T} z_t e^{-i\omega_j t}, \quad \text{where } z_t = u_t.$$

Under the assumption of exact normality, note that

$$d_z(\omega_j) \sim \text{complex normal}(0, V),$$

where as $T \to \infty$, $V \to 2\pi S_z(\omega_j)$. This follows from our calculations in Section 1.4.1. So, we have that

$$\left\|d_z(\omega_j)\right\|^2 \sim \frac{\chi_2^2}{2}\cdot 2\pi S_z(\omega_j).$$

Moreover, for $j \neq k$, $\left\|d_z(\omega_j)\right\|^2, \|d_z(\omega_k)\|^2$ are independent. This again follows from our calculations in Section 1.4.1.

Note that we can re-write the rejection probability as

$$\begin{aligned}\mathbb{P}_{H_0}\left\{(\hat{\beta}-\beta_0)^2/\hat{\Omega} > c\right\} &= \mathbb{P}_{H_0}\left\{\frac{(\hat{\beta}-\beta_0)^2}{\Omega} > c\frac{\hat{\Omega}}{\Omega}\right\} \\ &= \mathbb{E}\left[\mathbb{P}_{H_0}\left\{\frac{(\hat{\beta}-\beta_0)^2}{\Omega} > c\frac{\hat{\Omega}}{\Omega}\,|\,\hat{\Omega}\right\}\right],\end{aligned}$$

where the second equality follows from iterated expectations. Under the null hypothesis and by our

assumption of exact normality, $\hat{\beta} - \beta_0 \overset{H_0}{\sim} N(0, \Omega)$. Therefore, we can write this as

$$\begin{aligned}
\mathbb{P}_{H_0}\left\{(\hat{\beta}-\beta_0)^2/\hat{\Omega} > c\right\} &= \mathbb{E}\left[\mathbb{P}_{H_0}\left\{\left(\frac{\hat{\beta}-\beta_0}{\sqrt{\Omega}}\right)^2 > c\frac{\hat{\Omega}}{\Omega} \mid \hat{\Omega}\right\}\right] \\
&= \mathbb{E}\left[\mathbb{P}_{H_0}\left\{Z^2 > c\frac{\hat{\Omega}}{\Omega} \mid \hat{\Omega}\right\}\right], \quad \text{where } Z \sim N(0,1), Z^2 \sim \chi^2_1. \\
&= \mathbb{E}\left[1 - G(c\frac{\hat{\Omega}}{\Omega})\right],
\end{aligned}$$

where $G(\cdot)$ is the cdf of a χ^2_1. Now, apply a taylor expansion of $G(c\frac{\hat{\Omega}}{\Omega})$ around c. We have that

$$\begin{aligned}
\mathbb{P}_{H_0}\left\{t^2 > c\right\} &= \mathbb{E}\left[1 - G(c) - c\left(\frac{\hat{\Omega}-\Omega}{\Omega}\right)G'(c) - \frac{c^2}{2}\left(\frac{\hat{\Omega}-\Omega}{\Omega}\right)^2 G''(c) - \ldots\right] \\
&\approx 1 - G(c) - c\frac{\mathbb{E}\left[\hat{\Omega}-\Omega\right]}{\Omega}G'(c) - \frac{c^2}{2}\mathbb{E}\left[\left(\frac{\hat{\Omega}-\Omega}{\Omega}\right)^2\right]G''(c),
\end{aligned}$$

where c is chosen such that $1 - G(c) = \alpha$ and we ignore higher-order terms. Therefore, the size of the test is not equal to α. Instead, it also depends on:

1. The **bias** of our estimator $\hat{\Omega}$: $\mathbb{E}\left[\hat{\Omega} - \Omega\right]$
2. The **mean square error** of our estimator $\hat{\Omega}$: $\mathbb{E}\left[(\hat{\Omega} - \Omega)^2\right]$.

As a result, there is a bias-variance component to the size distortion. We now provide expressions for the bias and mean-square error of this estimator.

First, we compute $\mathbb{E}\left[\hat{\Omega}\right]$. Notice that

$$\begin{aligned}
\mathbb{E}\left[\hat{\Omega}\right] &= \mathbb{E}\left[\frac{1}{M}\sum_{j=1}^{M} K(j/M)\left\|d_z(\frac{2\pi j}{T})\right\|^2\right] \\
&= \frac{1}{M}\sum_{j=1}^{M} K(j/M)\left[2\pi S_z(\frac{2\pi j}{T})\right] \\
&= \frac{1}{M}\sum_{j=1}^{M} K(j/M)2\pi\left[S_z(0) + S_z'(0)\frac{2\pi j}{T} + \frac{1}{2}\left(\frac{2\pi j}{T}\right)^2 S_z''(0) + \ldots\right],
\end{aligned}$$

where the second equality used the fact that $\left\|d_z(\omega_j)\right\|^2 \sim \frac{\chi^2_2}{2} \cdot 2\pi S_z(\omega_j)$ and the final equality took a taylor expansion of the spectrum around $\omega = 0$ (again, ignoring higher order terms). Recall that $S_z(\omega)$ is symmetric around $\omega = 0$. Therefore, if it is continuously differentiable, then $S_z'(0) = 0$. This simplifies the expression above to

$$\mathbb{E}\left[\hat{\Omega}\right] = \frac{1}{M}\sum_{j=1}^{M} K(j/M)\Omega + \frac{1}{2M}\sum_{j=1}^{M} 2\pi K(j/M)\left(\frac{2\pi j}{T}\right)^2 S_z''(0).$$

In general, the kernel is normalized such that $\frac{1}{M}\sum_{j=1}^{M} K(j/M) = 1$. Therefore, we further simplify this to

$$\begin{aligned}
\mathbb{E}\left[\hat{\Omega}\right] &= \Omega + \frac{1}{2M}\sum_{j=1}^{M} 2\pi K(j/M)\left(\frac{2\pi j}{T}\right)^2 S_z''(0) \\
&= \Omega + \frac{1}{2}\left(\frac{M}{T}\right)^2 \frac{1}{M}\sum_{j=1}^{M} K(j/M)\left(\frac{j}{M}\right)^2 4\pi^2 \frac{2\pi S_z''(0)}{2\pi S_z(0)}\Omega \\
&= \Omega + \left(\frac{M}{T}\right)^2 \frac{1}{M}\sum_{j=1}^{M} K(j/M)\left(\frac{j}{M}\right)^2 2\pi^2 \frac{S_z''(0)}{S_z(0)}\Omega,
\end{aligned}$$

where $S_z''(0)/S_z(0)$ is the *normalized curvature* of the spectrum at $\omega = 0$. The more curvature in the spectrum at $\omega = 0$, the worse the bias of our estimator $\hat{\Omega}$. Moreover, we have that

$$\frac{1}{M}\sum_{j=1}^{M} K(j/M)(j/M)^2 \approx \int_0^1 u^2 K(u)du$$

for M large. Since we're thinking of this as providing an asymptotic approximation to a limit experiment in which $T \to \infty$, we substitute this expression in and re-arrange to arrive at

$$\mathbb{E}\left[\frac{\hat{\Omega} - \Omega}{\Omega}\right] = -\left(2\pi^2 \int_0^1 u^2 K(u)du,\right) \cdot \lambda \cdot \left(\frac{M}{T}\right)^2,$$

where

$$\lambda = -\frac{S_z''(0)}{S_z(0)}.$$

Note that when $\lambda > 0$, the estimator is downward biased and when $\lambda < 0$, the estimator upward biased.

Next, we calculate the variance of $\hat{\Omega}$. We have that

$$\begin{aligned}
V\left(\hat{\Omega}\right) &= V\left(\frac{1}{M}\sum_{j=1}^{M} K(j/M)\left\|d_z(\frac{2\pi j}{T})\right\|^2\right) \\
&= \frac{1}{M^2}\sum_{j=1}^{M} K^2(j/M) V\left(\left\|d_z(\frac{2\pi j}{T})\right\|^2\right),
\end{aligned}$$

where used the fact that $d_z(\omega_j)$ are uncorrelated across j. Now, recall that under the exact normality assumption, we know that $\left\|d_z(\frac{2\pi j}{T})\right\|^2 \sim \frac{1}{2}\chi_2^2 \cdot 2\pi \cdot S_z(\frac{2\pi j}{T})$. Therefore, it's immediate that

$$V\left(\left\|d_z(\frac{2\pi j}{T})\right\|^2\right) = S_z(\frac{2\pi j}{T})^2 \cdot 4\pi^2.$$

Substituting in, we arrive at

$$V\left(\hat{\Omega}\right) = \frac{1}{M^2}\sum_{j=1}^{M} K^2(j/M)\cdot 4\pi^2 \cdot S_z(\frac{2\pi j}{T})^2.$$

As before, we know taylor expand $S_z(\omega)$ around $\omega = 0$. We drop the 1st and 2nd order terms as we they will be squared and therefore, higher order. So, we end with

$$\begin{aligned} V\left(\hat{\Omega}\right) &= \frac{1}{M^2}\sum_{j=1}^{M} K^2(j/M)\cdot \left(2\pi S_z(0)\right)^2 \\ &= \frac{1}{M^2}\sum_{j=1}^{M} K^2(j/M)\cdot \Omega^2 \\ &= \frac{1}{M}\left(\int_0^1 K^2(u)du\right)\Omega^2. \end{aligned}$$

Then, we use our expression for the bias and the variance to rewrite the normalized MSE as

$$\begin{aligned} \mathbb{E}\left[\left(\frac{\hat{\Omega}-\Omega}{\Omega}\right)^2\right] &= \frac{1}{\Omega^2}V\left(\hat{\Omega}\right) + \frac{1}{\Omega^2}Bias(\hat{\Omega})^2 \\ &= \frac{1}{M}\left(\int_0^1 K^2(u)du\right), \end{aligned}$$

where we ignored the bias squared term because it is multiplied by $(M/T)^4$. For large T, this will be higher order.

Now, we put these results to together. Recall that

$$\mathbb{P}_{H_0}\left\{t^2 > c\right\} = \alpha - c\mathbb{E}\left[\frac{\hat{\Omega}-\Omega}{\Omega}\right]G'(c) - \frac{c^2}{2}\mathbb{E}\left[\left(\frac{\hat{\Omega}-\Omega}{\Omega}\right)^2\right]G''(c).$$

Plugging in, we therefore have that the rejection probability under the null is

$$\mathbb{P}_{H_0}\left\{t^2 > c\right\} - \alpha = c\left\{2\pi\left(\int_0^1 u^2K(u)du\right)\cdot\lambda\cdot(M/T)^2\right\}G'(c) - \frac{c^2}{2}\left\{\frac{1}{M}\int_0^1 K^2(u)du\right\}G''(c).$$

This is an expression for the **size distortion**. Notice that the curvature of the spectral density only enters into the bias term and the variance term does not depend on the underlying process z_t. The portion of the size distortion that is due to the variance of our estimator only depends on the choice of kernel $K(\cdot)$, the bandwidth M and the cdf of a χ^2_1 random variable. Therefore, we can hopefully eliminate this term from the size distortion by simply using a different critical value.

2.2 Adjusted critical values

Based on our reasoning above, we consider using adjusted critical values. Let c_m be the adjusted critical value. From our calculations earlier, we have that the null rejection probability at this adjusted

critical value is approximately

$$\mathbb{P}_{H_0}\left\{t^2 > c_m\right\} = 1 - G(c_m) - c_m\mathbb{E}\left[\frac{\hat{\Omega}-\Omega}{\Omega}\right]G'(c_m) - \frac{1}{2}c_m^2\mathbb{E}\left[\left(\frac{\hat{\Omega}-\Omega}{\Omega}\right)^2\right]G''(c_m).$$

Can we set c_m such that we will eliminate the last term? To do so, we expand the null rejection probability, which is a function of c_m, around the original critical value c. In this calculation, we are implicitly assuming that $c_m - c$ is small, and so we drop higher order terms that depend on it. We have

$$\mathbb{P}_{H_0}\left\{t^2 > c_m\right\} \approx 1 - G(c) - (c_m - c)G'(c) - c\mathbb{E}\left[\frac{\hat{\Omega}-\Omega}{\Omega}\right]G'(c) - \frac{1}{2}c^2\mathbb{E}\left[\left(\frac{\hat{\Omega}-\Omega}{\Omega}\right)^2\right]G''(c).$$

We set c_m such that $(c_m - c)G'(c)$ cancels the $G''(c)$ term. So, we set

$$\begin{aligned} c_m &= c + \frac{1}{2}c^2\mathbb{E}\left[\left(\frac{\hat{\Omega}-\Omega}{\Omega}\right)^2\right]\left(-\frac{G''(c)}{G'(c)}\right) \\ &= c + \left(-\frac{c^2G''(c)}{2G'(c)}\right)\left(\frac{1}{M}\int_0^1 K^2(u)du\right), \end{aligned}$$

where we plugged in our expression for $\mathbb{E}\left[\left(\frac{\hat{\Omega}-\Omega}{\Omega}\right)^2\right]$ from earlier. At this choice of adjusted critical value, we have that the null rejection probability becomes

$$\mathbb{P}_{H_0}\left\{t^2 > c_m\right\} = \alpha - c_m\mathbb{E}\left[\frac{\hat{\Omega}-\Omega}{\Omega}\right]G'(c).$$

These are **fixed-b critical values**. The **size distortion** at fixed-b critical values are

$$\begin{aligned} \Delta^S &= \mathbb{P}_{H_0}\left\{t^2 > c_m\right\} - \alpha \\ &= c\mathbb{E}\left[\frac{\hat{\Omega}-\Omega}{\Omega}\right]G'(c) \\ &= 2\pi^2 \cdot \lambda \cdot c \cdot \left(\int_0^1 u^2K(u)du\right) \cdot \left(\frac{M}{T}\right)^2 \cdot G'(c). \end{aligned}$$

The size distortion still depends on the curvature of the spectrum at $\omega = 0$.

Remark 2.1. *Throughout these calculations, we have been implicitly assuming that $M \to \infty$ and $M/T \to 0$. This is why we have been ignoring higher-order terms. But rather than working asymptotically, we've been making progress with the assumption of exact normality.*

2.3 Fixed-b critical values

We've been studying the test statistic of the form

$$\begin{aligned} t &= \frac{\hat{\beta} - \beta_0}{\sqrt{\hat{\Omega}}} \\ &= \frac{(\hat{\beta} - \beta_0)/\sqrt{\Omega}}{\left(\frac{1}{M}\sum_{j=1}^{M} K(j/M) \left\|d_z(2\pi j/T)\right\|^2\right)^{1/2}/\sqrt{\Omega}}, \end{aligned}$$

where recall that $\left\|d_z(2\pi j/T)\right\|^2 \sim \chi_2^2/2$ i.i.d. across j and $(\hat{\beta} - \beta_0)/\sqrt{\Omega} \sim N(0,1)$. So, this test statistic approximately behaves like

$$t \sim \frac{Z}{\left(\frac{1}{M}\sum_{j=1}^{M} K(j/M)\xi_j^2\right)^{1/2}},$$

where $Z \sim N(0,1)$ and $\xi_j^2 \sim \chi_2^2/2$ i.i.d. across j. This is known as the **fixed-b asymptotic distribution** of the t-statistic. Note that we are treating M as fixed here. Jansson (2004) and Sun, Phillips and Jin (2008) provide the formal derivation of this asymptotic distribution – the key is to be careful such that $M \to \infty$ and $M/T \to 0$ at the correct rates.

Remark 2.2. *Consider the special case in which $K(u) = 1$. Then, the denominator becomes*

$$\frac{1}{M}\sum_{j=1}^{M} \xi_j^2 \sim \chi_{2M}^2/2M.$$

Therefore,

$$t \sim \frac{z}{\sqrt{\chi_{2M}^2/2M}} \sim t_{2M}.$$

*This special case corresponds to using the **equal-weighted periodogram** to estimate $\hat{\Omega}$.*

2.4 Size-power tradeoff

Recall 1. *Recall that*

$$\mathbb{P}_\delta\left\{(Z+\delta)^2 \leq x\right\} = G_\delta(x)$$

*is a non-central χ_1^2 with **non-centrality parameter** δ^2.*

The **power** of the test is

$$\mathbb{P}_{\delta^2}\left\{t^2 > c\right\}.$$

That is, it is the probability of correctly rejecting the null hypothesis. We want to consider $H_0 : \beta = \beta_0$

and $H_a : \beta \neq \beta_0$. Suppose that Ω is known for now. Then,

$$\begin{aligned}
\mathbb{P}\left\{(\hat{\beta}-\beta_0)^2/\Omega > c\right\} &= \mathbb{P}\left\{\frac{((\hat{\beta}-\beta_1)+(\beta_1-\beta_0))^2}{\Omega} > c\right\} \\
&= \mathbb{P}\left\{\left(\frac{\hat{\beta}-\beta_1}{\sqrt{\Omega}}+\frac{\beta_1-\beta_0}{\sqrt{\Omega}}\right)^2 > c\right\} \\
&= \mathbb{P}\left\{(Z+\delta)^2 > c\right\},
\end{aligned}$$

where $\delta = (\beta_1 - \beta_0)/\sqrt{\Omega}$. δ is the departure from the null in standardized units. We can plot this rejection probability as a function of δ and this will deliver the **power function** of the test.

Now, fix some δ^2. We begin with a **power expansion**, which is analogous to the size expansion that we compute earlier. First, we have that

$$\begin{aligned}
\mathbb{P}_{\delta^2}\left\{t^2 > c\right\} &= \mathbb{P}_{\delta^2}\left\{\left(\frac{(\hat{\beta}-\beta_1)+(\beta_1-\beta_0)}{\sqrt{\hat{\Omega}}}\right)^2 > c\right\} \\
&= \mathbb{P}_{\delta^2}\left\{(z+\delta)^2 > c\frac{\hat{\Omega}}{\Omega}\right\} \\
&= \mathbb{E}\left[1 - G_{\delta^2}(c\frac{\hat{\Omega}}{\Omega})\right],
\end{aligned}$$

where we followed the same steps as the size expansion so far. We now take a taylor expansion around $\hat{\Omega} = \Omega$. We have

$$\mathbb{P}_{\delta^2}\left\{t^2 > c\right\} = 1 - G_{\delta^2}(c) - c\mathbb{E}\left[\frac{\hat{\Omega}-\Omega}{\Omega}\right]G'_{\delta^2}(c) - \frac{1}{2}c^2\mathbb{E}\left[\left(\frac{\hat{\Omega}-\Omega}{\Omega}\right)^2\right]G''_{\delta^2}(c),$$

where we have that $1 - G_{\delta^2}(c)$ is the power of the test if Ω we known – it is the power of the **oracle test**. The additional terms illustrate that you lose power due to estimation of Ω with $\hat{\Omega}$.

Now, what is the power of the test if we instead used fixed-b critical values? We'll consider

$$\mathbb{P}_{\delta^2}\left\{t^2 > c_m\right\}$$

and taylor expand it around c. Note that $c_m \to c$ at rate $1/M$. Moreover, the terms that depend on $G'(c_m), G''(c_m)$ will converge to $G'(c), G''(c)$ at the same rate. We ignore higher order terms that are multiplied by $1/M$ as they will converge at rate $1/M^2$. So, we get that

$$\mathbb{P}_{\delta^2}\left\{t^2 > c_m\right\} = 1 - G_{\delta^2}(c) - (c_m - c)G'_{\delta^2}(c) - c\mathbb{E}\left[\frac{\hat{\Omega}-\Omega}{\Omega}\right]G'_{\delta^2}(c) - \frac{1}{2}c^2\mathbb{E}\left[\left(\frac{\hat{\Omega}-\Omega}{\Omega}\right)^2\right]G''_{\delta^2}(c).$$

We focus on the term

$$(c_m - c)G'_{\delta^2}(c) + \frac{1}{2}c^2\mathbb{E}\left[\left(\frac{\hat{\Omega}-\Omega}{\Omega}\right)^2\right]G''_{\delta^2}(c) = \left(-\frac{c^2G''_0(c)}{G'_0(c)}\right)\left(\frac{1}{M}\int_0^1 K^2(u)du\right)G'_{\delta^2}(c)$$
$$+\frac{1}{2}c^2\left(\frac{1}{M}\int_0^1 K^2(u)du\right)G''_{\delta^2}(c)$$
$$=\left(\frac{1}{M}\int_0^1 K^2(u)du\right)\left(\frac{1}{2}c^2G''_{\delta^2}(c) - c^2\frac{G''_0(c)}{G'_0(c)}G'_{\delta^2}(c)\right).$$

As notation, let $a_{\delta^2} = \frac{1}{2}c^2G''_{\delta^2}(c) - c^2\frac{G''_0(c)}{G'_0(c)}G'_{\delta^2}(c)$. Note that because δ^2 is known (we get to choose the alternative under consideration), this is just a number. So, we have that the power of the test using the fixed-b critical values becomes

$$\mathbb{P}_{\delta^2}\left\{t^2 > c_m\right\} = (1 - G_{\delta^2}(c)) - cG'_{\delta^2}(c)\cdot 2\pi\cdot\left(\int_0^1 u^2K(u)du\right)\cdot\lambda\cdot\left(\frac{M}{T}\right)^2 - a_{\delta^2}\left(\int_0^1 K(u)^2du\right),$$

where $1 - G_{\delta^2}(c)$ is the oracle power of the test at δ^2, the second term is a distortion introduced from estimating $\hat{\Omega}$ and the third term is a distortion that only depends on the chosen M, $K(\cdot)$ and the critical value.

So, at the fixed-b critical values, we need to **trade off** between the power loss and the size distortion that we saw earlier. The optimal testing approach will select the test with the higher power when selecting among tests of the same size. Suppose that we have two estimators of $\hat{\Omega}$ that use the same kernel but different bandwidths M. They will have different size distortions because of the different choices of M. To make an apples-to-apples comparison, we need to force them to have the same size. To do so, we introduce an additional adjustment or a "size-adjusted critical value."

Let $c_{M,T}$ be our adjustment to the fixed-b critical values. We have that

$$\mathbb{P}_{H_0}\left\{t^2 > c_{M,T}\right\} = \alpha - (c_{M,T} - c)G'_0(c) - c\mathbb{E}\left[\frac{\hat{\Omega}-\Omega}{\Omega}\right]G'_0(c).$$

We choose $c_{M,T}$ such that the last two terms are zero. We get that

$$c_{M,T} = c_M + c\mathbb{E}\left[\frac{\hat{\Omega}-\Omega}{\Omega}\right].$$

This is a **size-adjusted critical value**. However, we don't know the finite sample bias. The power using the adjusted critical values is

$$\mathbb{P}_{\delta^2}\left\{t^2 > c_{M,T}\right\} = 1 - G_{\delta^2}(c) - a_{\delta^2}\int_0^1\frac{K^2(u)}{M}du.$$

Now, define

$$\Delta_s = \mathbb{P}_{H_0}\left\{t^2 > c_m\right\} - \alpha,$$
$$\Delta_p = 1 - G_{\delta^2}(c) - \mathbb{P}_{\delta^2}\left\{t^2 > c_{M,T}\right\}.$$

We have expressions for these objects. Plugging in our results from earlier, we get that

$$\Delta_s = cG_0'(c)\left(2\pi^2 \int_0^1 u^2 K(u)du\right)\left(\frac{M}{T}\right)^2 \lambda,$$
$$\Delta_p = a_{\delta^2}\left(\int_0^1 K^2(u)du/M\right).$$

Consider the objective

$$\sqrt{\Delta_s}\Delta_p = \frac{\sqrt{\lambda}}{T}\left(\sqrt{cG_0'(c)}a_{\delta^2}\sqrt{2\pi^2}\right)\left(\sqrt{\int_0^1 u^2K(u)du}\int_0^1 K^2(u)du\right)$$
$$\geq \bar{a}_{\delta^2}\frac{\sqrt{\lambda}}{T}\left(\min_{K(\cdot)}\sqrt{\int_0^1 u^2K(u)du}\int_0^1 K^2(u)du\right),$$

Why do we choose this? This kills off the terms depending on M. The minimization is over all kernels satisfying $K(u) \geq 0$ and $\int_0^1 K(u)du = 1$. We can show that the optimal solution is the **quadratic spectral** or the **epanechnikov kernel** with

$$K(u) = \frac{3}{2}(1-u^2).$$

We still haven't solved for the optimal bandwidth M. We can show that $M = \phi T^{2/3}$, which is chosen such that $O(\Delta_s) = O(\Delta_p)$ i.e. the size distortion and power loss are of the same order.

3 Structural Vector Autoregressions

Structural vector autoregressions (SVARs) are one of the main tools for estimating dynamic causal effects in time series. To fix ideas, consider the following example from monetary policy. The FOMC meets every 6 weeks and there is a surprise component to each decision. That is, there is a departure from the expected path of the Federal Funds rate and we think of this deviation as a random treatment. Let ϵ_t^r be the monetary policy shock and we wish to understand its effects. In particular, we wish to estimate the causal effect of the monetary policy shock on some outcome Y h-periods ahead:

$$\mathbb{E}_t\left[Y_{t+h} \,|\, \epsilon_t^r = 1\right] - \mathbb{E}_t\left[Y_{t+h} \,|\, \epsilon_t^r = 0\right] = \Theta_{y,r,h}.$$

We refer to this as a **dynamic causal effect**.[5] We refer to $\{\Theta_{y,r,h} : h \geq 0\}$ as an **impulse response function**.

More generally, this is an exciting, active area of research in econometrics. The core question is:

[5]Note that we are assuming that everything here is stationary and linear.

How can we take ideas that have been developed in applied microeconometrics on causal identification to the time series setting? We'll begin by discussing the traditional, SVAR approach to these questions and then take a step back and return to this fundamental question.

3.1 Structural moving average

With the assumption of stationarity and linearity, we write

$$\begin{aligned} \underset{n\times 1}{Y_t} &= \underset{}{\Theta(L)} \underset{m\times 1}{\epsilon_t} \\ &= \begin{pmatrix} \underset{n\times 1}{\Theta_1(L)} & \underset{n\times(m-1)}{\Theta_{\cdot}(L)} \end{pmatrix} \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{\cdot,t} \end{pmatrix}, \end{aligned}$$

where ϵ_t is some unforecastable error.[6] The dynamics of Y_t are expressed in terms of the shock of interest, measurement error and other shocks. In general, $m \geq n$ as there will be more shocks/disturbances/measuremer error than observable series. The object of interest is the lag polynomial $\Theta_1(L)$.

We assume that

$$\mathbb{E}\left[\epsilon_t \epsilon_t'\right] = \Sigma_\epsilon = \text{diag}\{\sigma^2_{\epsilon_1}, \ldots, \sigma^2_{\epsilon_m}\}.$$

That is, the shocks are mutually uncorrelated within a period. We also assume that

$$\mathbb{E}\left[\epsilon_t \,|\, \epsilon_{t-1}, \epsilon_{t-2}, \ldots\right] = 0.$$

That is, the shocks are unpredictable over time.

We now introduce some useful objects for summarizing the contribution of a shock to the behavior of a time series. First, we define a **historical decomposition**, which describes the movement in the observed time series that is attributable to a particular shock.

Definition 3.1 (Historical decomposition)**.** *The **historical decomposition** describes the movement in Y_t that is attributable to $\epsilon_{1,t}$. It is, $\Theta_1(L)\epsilon_{1,t}$.*

In other words, the historical decomposition describes how the observed time series would have behaved had only the shock $\epsilon_{1,t}$ occurred. The forecast error of Y_{t+h} made using a forecast based on the available information up to time t is

$$Y_{t+h} - Y_{t+h|t}, \quad \mathbb{E}\left[Y_{t+h} \,|\, \epsilon_t, \epsilon_{t-1}, \ldots\right] = Y_{t+h|t}.$$

This error is given by

$$\begin{aligned} Y_{t+h} - Y_{t+h|t} &= \Theta_0 \epsilon_{t+h} + \Theta_1 \epsilon_{t+h-1} + \ldots + \Theta_{h-1}\epsilon_{t+1} \\ &= \Theta_{1,0}\epsilon_{1,t+h} + \Theta_{1,1}\epsilon_{1,t+h-1} + \ldots + \Theta_{1,h-1}\epsilon_{1,t+1} \\ &\quad + \Theta_{\cdot,0}\epsilon_{\cdot,t+h} + \ldots + \Theta_{\cdot,h-1}\epsilon_{\cdot,t+1}. \end{aligned}$$

[6] We can think of this as just an application of the Wold Decomposition.

The **forecast error variance decomposition** describes what fraction of the variation in the observed time series is attributable to the shock of interest.

Definition 3.2 (Forecast error variance decomposition)**.** *The **forecast error variance decomposition** (FEVD) is contribution of error $\epsilon_{1,t}$ to the mean-square in a forecast of Y_{t+h} using the information available up to time t. This is just*

$$FEVD_{1,h} = \frac{\left(\theta_{1,0}^2 + \ldots, \theta_{1,h-1}^2\right)\sigma_{\epsilon_1}^2}{V(Y_{t+h} - Y_{t+h|t})}.$$

To understand the challenges of estimating/identifying the impulse response function, consider the following thought experiment. Suppose that we observed the shock of interest directly. How would we estimate the impulse response function? In this case, the problem is simple. We would just directly regress the outcome of interest on the observed shocks

$$Y_t = \Theta_1(L)\epsilon_{1,t} + u_t, \tag{2}$$

where $u_t = \Theta_\cdot(L)\epsilon_{\cdot,t}$. The coefficients on the shocks would give us the impulse response function. This approach has, in fact, been taken in the literature as discussed in Example 3.1.

Example 3.1. *A branch of the literature on the macroeconomic effects of monetary policy pursues this approach. The idea is simple. First, we construct a credible measure of the monetary policy shock, $\hat{\epsilon}_{1,t}$. Second, we then estimate the regression in Equation (2), plugging in our estimate of the monetary policy shock.*

One common approach to measure the monetary policy shock is to examine changes in futures markets on the Federal Funds rate around small windows of an FOMC announcement and argue that these changes identify the monetary policy shock. For example, see Rudebusch (1998); Kuttner (2001); Cochrane and Piazessi (2002); Bernanke and Kuttner (2005). One common criticism of this approach is that the information that is revealed during a window around FOMC announcements is only a partial component of the monetary policy shock as information about monetary policy is typically revealed between meetings as well.

However, in general, the shocks ϵ_t are *not* observed. So, the problem of estimating/identifying the impulse response function has two parts. First, how do we construct estimates of the shocks? Second, given our estimates of the shocks, how do we estimate the impulse response function? The SVAR approach that we will develop next solves both of the problems *jointly*.

3.2 Sims (1980), Short-Run Restrictions and the Cholesky decomposition

The standard approach for both identifying the shock of interest and estimating dynamic causal effects relies on VARs and SVARs. The identifying assumptions underlying this approach are subtle and to illustrate these ideas, we'll focus on the analysis in Sims (1980), a seminal paper that pioneered these methods. Again, it's useful to keep the two fundamental questions in mind as we discuss this approach: (1) How do we measure the shocks of interest? (2) How do we estimate the impulse response function? We encourage you to try to think about how this approach answers both of these questions.

Consider a **vector auto-regression** (VAR)

$$\underset{n\times n}{A(L)}\ \underset{n\times 1}{Y_t} = \underset{n\times 1}{\eta_t},$$

where η_t is referred to as the **residuals, innovations** or **Wold decomposition errors**. The associated **vector moving average** is

$$\begin{aligned}\underset{n\times 1}{Y_t} &= \underset{n\times n}{A(L)^{-1}}\ \underset{n\times 1}{\eta_t} \\ &= C(L)\eta_t, \quad C(L) = I + C_1 L + C_2 L^2 + \dots\end{aligned}$$

Our object of interest is the lag polynomial in the **structural moving average** $\Theta(L)$ with

$$Y(L) = \underset{n\times m}{\Theta(L)}\ \underset{m\times 1}{\epsilon_t}, \quad \Theta(L) = \Theta_0 + \Theta_1 L + \dots$$

We assume that

1. $n = m$
2. $\text{span}(\eta_t) = \text{span}(\epsilon_t)$

Together, this is known as the **invertibility assumption**. It implies that the structural shocks ϵ_t lie in the linear space spanned by the reduced-form innovations. This is a *strong* assumption. It implies that

$$\eta_t = \Theta_0 \epsilon_t, \quad \Theta_0^{-1} \text{ exists.} \tag{3}$$

Under these assumptions, we now show that there is a structural vector auto-regression representation of the SVMA.

Proposition 3.1 (Existence of SVAR representation)**.** *Assume that Y_t is an $n \times 1$, second-order stationary, linearly regulator process. Assume that it has the structural moving average representation*

$$\underbrace{Y_t}_{n\times 1} = \underbrace{\Theta(L)}_{n\times m}\underbrace{\epsilon_t}_{m\times 1},$$

where $\mathbb{E}[\epsilon_t \epsilon_s'] = 0$ for all $s \neq t$ and $\mathbb{E}[\epsilon_t \epsilon_t'] = \text{diag}\{\sigma_1^2, \dots, \sigma_m^2\}$. Additionally, assume that Y_t is inverible meaning that

$$\text{span}(\eta_t) = \text{span}(\epsilon_t),$$

*where $\eta_t = Y_t - \text{Proj}\{Y_t \mid Y_{t-1}, Y_{t-2}, \dots\}$. Then, Y_t has a **structural vector autoregression representation***

$$A(L)Y_t = \Theta_0 \epsilon_t,$$

where $\Theta(L) = A(L)^{-1}\Theta_0$ and $A(L)$ is the projection coefficient of Y_t onto its past values, meaning that $\text{Proj}\{Y_t \mid Y_{t-1}, Y_{t-2}, \dots\} = A_1 Y_{t-1} + A_2 Y_{t-2} + \dots$ and $A(L) = I - A_1 L - A_2 L^2 - \dots$

Proof. Invertibility implies that $n = m$ and

$$\eta_t = H\epsilon_t, \quad H^{-1} \text{ exists.}$$

Therefore, we can re-write the SVMA form as

$$Y_t = \Theta(L)\epsilon_t = \Theta(L)H^{-1}\eta_t.$$

Recall that η_t are the Wold innovations and so, this must be equal to the Wold representation because the Wold representation is unique. Therefore, we have that

$$A(L)^{-1} = \Theta(L)H^{-1} \implies \Theta(L) = A(L)^{-1}H,$$

where $A(L)^{-1} = \left(I - A_1L - A_2L^2 - \ldots\right)^{-1} = I + B_1L + B_2L^2 + \ldots$ for some matrices $B_1, B_2, \ldots$. We then have that

$$\Theta_0 + \Theta_1 L + \ldots = (I + B_1L + B_2L + \ldots)H.$$

We immediately see that $H = \Theta_0$. Therefore,

$$\Theta(L) = A(L)^{-1}\Theta_0$$

and

$$\begin{aligned} A(L)Y_t &= \eta_t \\ &= H\epsilon_t = \Theta_0\epsilon_t. \end{aligned}$$

□

With the assumption of invertibility and the existence of the SVAR representation, the question of identification is relatively straightforward to answer. From Equation (3), it's immediate that

$$\begin{aligned} \Sigma_\eta &= \Theta_0\Sigma_\epsilon\Theta_0', \\ Y_t &= C(L)\eta_t \\ &= C(L) \cdot \Theta_0\epsilon_t. \end{aligned}$$

Therefore, we only need to identify Θ_0 to identify the impulse response function. Why? $C(L) = A(L)^{-1}$ can be estimated from the reduced-form VAR. The question then becomes: How do we identify Θ_0? The first equation in the previous display tells us how to do so. We just count unknowns and equations! From Σ_η we have $\frac{n(n+1)}{2}$ known parameters. But, Θ_0 has n^2 unknown parameters and Σ_ϵ has n unknown parameters. Since the structural shocks ϵ_t are unobserved, we need to adopt a scale normalization. There are two common choices: (1) **Unit variance normalization** with $\Sigma_\epsilon = I_n$ or (2) **Unit effect normalization** with $\Theta_{0,jj} = 1$ for all j. With either of these normalization's, there are now only n^2 unknown parameters but we still only have $\frac{n(n+1)}{2}$ equations. Therefore, we are *under-identified*

and need more restrictions on the system in order identify Θ_0, Σ_ϵ.

Sims (1980) introduce a **short-run timing assumption** to solve this problem of under-identification. In particular, we adopt the unit variance normalization and assume that Θ_0 is lower triangular with

$$\begin{pmatrix} \Theta_{1,1} & 0 & \dots & 0 \\ \Theta_{2,1} & \Theta_{2,2} & \dots & 0 \\ \vdots & & \ddots & 0 \\ \Theta_{n,1} & \dots & \Theta_{n,n-1} & \Theta_{n,n} \end{pmatrix} \begin{pmatrix} \eta_{1,t} \\ \vdots \\ \eta_{n,t} \end{pmatrix} = \begin{pmatrix} \epsilon_{1,t} \\ \vdots \\ \epsilon_{n,t} \end{pmatrix}.$$

This means that $\eta_{1,t}$ is exogenous and just a linear function of the first structural shock. This means that we can factor

$$\Sigma_\eta = \Theta_0 \Sigma_\epsilon \Theta_0',$$

using a **Cholesky decomposition**. We have that

$$\Sigma_\eta^{1/2} = \Theta_0 \Sigma_\epsilon^{1/2} = \Theta_0.$$

Therefore, $\Theta_0 = Chol(\Sigma_\eta)$.

As mentioned, this is referred to as a timing assumption. Why? It restricts which series are allowed to react *within the period* to particular structural shocks.

Remark 3.1. *Clearly, invertibility is the **key** assumption required to get the SVAR approach off the ground. At first, it appears quite strange but there's a natural omitted variables bias interpretation to invertibility. This intuition is developed in detail in Stock and Watson (2018).*

Suppose the the observed data is generated by a VAR, $A(L)Y_t = \eta_t$, the invertibility assumption is satisfied and that the shocks ϵ_t are observed. Consider a linear, one-step ahead forecast of the observed time series Y_t based on its own lags and the observed shocks

$$\begin{aligned} \text{Proj}\left\{Y_t \mid Y_{t-1}, Y_{t-2}, \dots, \epsilon_{t-1}, \epsilon_{t-2}, \dots\right\} &= \text{Proj}\left\{Y_t \mid Y_{t-1}, Y_{t-2}, \dots, \epsilon_{t-1}, \epsilon_{t-2}, \dots\right\} \\ &\overset{(1)}{=} \text{Proj}\left\{Y_t \mid \eta_{t-1}, \eta_{t-2}, \dots, \epsilon_{t-1}, \epsilon_{t-2}, \dots\right\} \\ &\overset{(2)}{=} \text{Proj}\left\{Y_t \mid \eta_{t-1}, \eta_{t-2}, \dots\right\} \\ &= \text{Proj}\left\{Y_t \mid Y_{t-1}, Y_{t-2}, \dots\right\}, \end{aligned}$$

where (1) follows because Y_t follows a VAR and (2) follows by the assumption of invertibility. In other words, if the invertibility assumption is satisfied, the shocks ϵ_t do not provide further information in a linear, one-step ahead forecast conditional on lagged values of the observed time series!

Example 3.2. *Corporate fuel economy standards There was a proposal in August 2018 to modify the path of corporate fuel economy standards. We wish to understand how this proposal would have affected car prices and car sales. We'll discuss several methods to see how the ideas we have discussed so far play out.*

- ***Method 1: Distributed lag model*** *Suppose that*

$$Y_t = \beta(L)X_t + u_t,$$

where $Y_t = \Delta \log(sales_t)$ and $X_t = \Delta \log(prices_t)$. Under what conditions is β identified? We need the errors u_t to be ***strictly exogenous****. That is,*

$$\mathbb{E}\left[u_t \mid X_t, X_{t-1}, \ldots\right] = 0,$$

where u_t is a demand disturbance. Think of this as a no feedback condition.

- ***Method 2: Autoregressive distributed lag model*** *We now model*

$$\alpha(L)u_t = \tilde{u}_t.$$

Then, we can write

$$\begin{aligned}\alpha(L)Y_t &= \alpha(L)\beta(L)X_t + \alpha(L)u_t \\ \tilde{Y}_t &= \beta(L)\tilde{X}_t + \tilde{u}_t.\end{aligned}$$

The idea is that after this transformation, the result error $\tilde{u}_t$ is now uncorrelated over time. The condition for identification is now strict exogeneity on the transformed variables:

$$\mathbb{E}\left[\tilde{u}_t \mid \tilde{X}_t, \tilde{X}_{t-1}, \ldots\right] = 0.$$

We can re-write this further. We have that

$$\begin{aligned}\mathbb{E}\left[\tilde{u}_t \mid \tilde{X}_t, \tilde{X}_{t-1}, \ldots\right] &= \mathbb{E}\left[\alpha(L)u_t \mid \tilde{X}_t, \tilde{X}_{t-1}, \ldots\right], \\ &= \sum_{j=0}^{p} \mathbb{E}\left[u_{t-j} \mid \tilde{X}_t, \tilde{X}_{t-1}, \ldots\right],\end{aligned}$$

where assumed $\alpha(L)$ was a lag-p polynomial. Therefore, a sufficient condition is that

$$\mathbb{E}\left[u_{t-j} \mid X_t, X_{t-1}, \ldots\right] = 0$$

for all $j = 0, \ldots, p$. This is known as ***lead-lag exogeneity****.*

- ***Method 3: Vector autoregression*** *Now let Y_t be a vector. We have the VAR*

$$A(L)Y_t = \eta t$$

and the underlying SVMA

$$Y_t = \Theta(L)\epsilon_t.$$

We can then write reduced-form moving average from the VAR

$$\begin{aligned} Y_t &= A(L)^{-1}\eta_t \\ &= c(L)\eta_t \\ &= (I + C_1 L + C_2 L^2 + \ldots)\eta_t. \end{aligned}$$

If we make the invertibility assumption, then it follows that $\eta_t = \Theta_0 \epsilon_t$ with $\epsilon_t = \Theta_0^{-1}\eta_t$. We can reduce the identification condition to a contemporaneous exogeneity condition.

In our example, think of

$$Y_t = \begin{pmatrix} \Delta \log(GDP_t) \\ \log(E_t) \\ \Delta \log(P_t) \\ \Delta \log(S_t) \end{pmatrix}.$$

The assumption is that

$$\begin{aligned} &\Delta \log(S_t) = b_{sy}\Delta \log(GDP_t) + b_{se}\log(E_t) + b_{sp}\Delta \log(P_t) + \lambda_1 Y_{t-1} + \lambda_2 Y_{t-2} + \ldots + \epsilon_t^s, \\ &\mathbb{E}\left[\epsilon_t^s \mid \Delta \log(GDP_t), \log(E_t), \Delta \log(P_t)\right] = 0. \end{aligned}$$

That is, demand shocks only respond to previous shocks, not contemporaneous shocks. This type of assumption places a cholesky structure on the SVAR system if we extend it to each series. Recall that this meant that Θ_0 is lower triangular and therefore, so is Θ_0^{-1}. So, we have that

$$\begin{pmatrix} b_{1,1} & 0 & \ldots & 0 \\ b_{2,1} & b_{2,2} & \ldots & 0 \\ \vdots & & \ddots & 0 \end{pmatrix} \eta_t = \epsilon_t.$$

Therefore, from the last row, we obtain that

$$b_{4,4}\eta_{4,t} = -b_{1,4}\eta_{1,t} - b_{2,4}\eta_{2,t} - b_{3,4}\eta_{3,t} + \epsilon_{4,t}$$

and so, we have that $\epsilon_{4,t}$ is the residual from $Proj(\eta_{4,t}|\eta_{1,t}, \eta_{2,t}, \eta_{3,t})$.

Remark 3.2 (SVAR Assumptions)**.** *Based on this discussion, we see that there are several assumptions required in the SVAR approach to identifying dynamic causal effects. These are*

1. *Linear regularity*
2. *Second order stationarity*
3. *Shocks as primitives*
4. *Additivity/linearity of $\Theta(L)$*
5. *Invertibility*

While the assumption of second order stationarity is strong, we can often make it more plausible by explicitly modeling sources of non-stationarity (e.g. break dates). The assumption of shocks as the underlying primitives that drive observed movements/co-movements in macroeconomic data is another key assumption and has a long history in macroeconomics.[7] *The assumption that the underlying impulse response functions are linear is also strong. But, given the small sample sizes in time series data, it is often difficult to precisely fit non-linear models. Finally, as we discussed in Remark 3.1, invertibility is perhaps the strongest assumption yet it is necessary in the SVAR approach. It is an open (and exciting) question about how to relax this assumption and still make progress.*

3.3 Long-run restrictions

We now explore an alternative strategy for imposing additional restrictions on an SVAR system to solve the under-identification problem. This was developed originally in Blanchard and Quah (1989). For simplicity, consider

$$Y_t = \begin{pmatrix} GDP_t \\ Unemp_t \end{pmatrix}, \quad \epsilon_t = \begin{pmatrix} \epsilon_t^s \\ \epsilon_t^d \end{pmatrix}.$$

Assume that $Y_t = \Theta(L)\epsilon_t$. The identifying restriction is that

$$\Theta_{12}(1) = 0.$$

That is, ϵ_t^d has no long-run effect on $Unemp_t$ – demand shocks do not matter in the long-run. So, we have that

$$\Theta(1) = \Theta_0 A(1)^{-1}$$

and this restriction is equivalent to

$$\left[\Theta_0 A(1)^{-1}\right]_{12} = 0.$$

There's an easy algorithm to solve for all the objects of intest. Recall that

$$\Omega = A(1)^{-1}\Sigma_\eta (A(1)^{-1})',$$

where $\eta_t = \Theta_0 \epsilon_t$, $\Sigma_\eta = \Theta_0 \Sigma_\epsilon \Theta_0'$. Substituting this in, we ahve that

$$\Omega = A(1)^{-1}\Theta_0 \Sigma_\epsilon \Theta_0' (A(1)^{-1})'.$$

With the unit variance normalization and $\Sigma_\epsilon = I$, this becomes

$$\Omega = A(1)^{-1}\Theta_0 \Theta_0' (A(1)^{-1})'$$

[7] See Ramey (2016) for an extensive discussion of this assumption. Rambachan and Shephard (2019) argue that for certain dynamic causal effects of interest, we may not need to assume that the primitives of interest are the causal effects of "shocks" to make progress.

and the long-run restriction imposes that Ω be lower triangular. Therefore,

$$Chol(\Omega) = A(1)^{-1}\Theta_0$$

and so,

$$\Theta_0 = A(1)Chol(A(1)^{-1}\Sigma_\eta(A(1)^{-1})').$$

3.4 Identification by heteroskedasticity

So, we have that $\eta_t = \Theta_0\epsilon_t$ and so, $\Sigma_\eta = \Theta_0\Sigma_\epsilon\Theta_0'$. Recall that Σ_η has $n(n+1)/2$ parameters since it is a covariance matrix, Θ_0 has n^2 parameters and Σ_ϵ has n parameters. We then have $n^2 + n - n = n^2$ total unknown parameters, where we subtracted off n due to the unit variance or unit effect normalization. So, as discussed earlier, we need additional restrictions to solve this problem of under-identification.

Now, suppose that there are two variance regimes of the shocks of interest. In regime 1,

$$\Sigma_{\eta,1} = \Theta_0\Sigma_{\epsilon,1}\Theta_0'$$

and in regime 2,

$$\Sigma_{\eta,2} = \Theta_0\Sigma_{\epsilon,2}\Theta_0'.$$

That is, the variance of the shocks changed but the impulse response coefficients did not. Now, we have $n(n+1)/2$ known parameters in $\Sigma_{\eta,1}$ and $n(n+1)/2$ known parameters in $\Sigma_{\eta,2}$ and so, in total we have $n^2 + n$ known parameters. Suppose we adopt the unit variance normalization in regime 1. Then, we have $n^2 + n$ unknown parameters – n^2 parameters in Θ_0 and n parameters in $\Sigma_{\epsilon,2}$. Therefore, we are identified! We can set this up as an instrument, where the instrument is the regime change.

3.5 Sign restrictions

Once again, assume invertibility. We have that

$$\begin{aligned} A(L)Y_t &= \eta_t, \\ Y_t &= \Theta(L)\epsilon_t, \\ \eta_t &= \Theta_0\epsilon_t. \end{aligned}$$

Suppose that some theory tells us that

$$R = \{\Theta_{h,11} \geq 0 : h = 0, 1, 2\}.$$

This is a **sign restriction**. For example, we can at least say that a monetary policy shock will raise the FFR over the next 2 quarters. Alternatively, we could specify

$$R = \{\Theta_{h,21} \leq 0 : h = 0, 1, \ldots, 4\}.$$

Maybe we can say that a monetary policy shock will lower employment over the next several quarters. Can we use these types of restrictions to make progress. In these cases, we will only be able to **set identify** the IRF. We will focus on one particular strategy for doing inference on the identified set of impulse response functions but this is part of an active area of research in macroeconometrics and the much broader literature on partial identification.

We have that

$$\begin{aligned} Y_t &= A(L)^{-1}\eta_t \\ &= A(L)^{-1}\Sigma_\eta^{1/2}\Sigma_\eta^{-1/2}\Theta_0\epsilon_t, \\ &= (A(L)^{-1}\Sigma_\eta^{1/2})(\Sigma_\eta^{-1/2}\Theta_0)\epsilon_t. \end{aligned}$$

We impose the unit variance assumption and so, $\mathbb{E}\left[\epsilon_t\epsilon_t'\right] = I$. Define $Q = \Sigma_\eta^{1/2}\Theta_0$ and note that

$$\mathbb{E}\left[\Sigma_\eta^{-1/2}\eta_t\eta_t'\Sigma_\eta^{-1/2}\right] = I, \quad \text{where } \Sigma_\eta^{1/2} = Chol(\Sigma_\eta).$$

So, we have that

$$\mathbb{E}\left[Q\epsilon_t\epsilon_t Q'\right] = \mathbb{E}\left[QQ'\right] = I,$$

where we used $\epsilon_t = \Theta_0^{-1}\eta_t$. Therefore, Q is orthonormal with $QQ' = I$ and $\Theta(L) = A(L)^{-1}\Sigma_\eta^{1/2}Q$. How do we construct the set of impulse response functions that are consistent with R. Uhlig (2005) proposes the following algorithm:

1. Estimate $\hat{A}(L), \hat{\Sigma}_\eta$.
2. Sample $\tilde{A}(L), \hat{\Sigma}_\eta$ from posterior of $A(L), \Sigma_\eta$.
3. Sample $\tilde{Q}$ from prior distribution over the space of orthonormal matrices.
4. Construct
$$\tilde{\Theta}(L) = \tilde{A}(L)^{-1}\tilde{\Sigma}_\eta^{-1/2}\tilde{Q}.$$
Keep if $\tilde{\Theta}(L) \in R$. Repeat this many times.
5. Compute the mean of accepted $\tilde{\Theta}(L)$.

This approach has a general problem. Sampling over $\tilde{Q}$ is tricky because $\tilde{Q}$ is not identified in the data. So the estimated IRF and the resulting inference will be influenced by the prior over Q.

Example 3.3. *This example is discussed in Baumeister and Hamilton (2015). Consider*

$$\begin{pmatrix} Y_{1,t} \\ Y_{2,t} \end{pmatrix} = \begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{pmatrix} \begin{pmatrix} Y_{1,t-1} \\ Y_{2,t-1} \end{pmatrix} + \begin{pmatrix} \eta_{1,t} \\ \eta_{2,t} \end{pmatrix},$$

where $\alpha_1, \alpha_2 > 0$ *and* $\mathbb{E}\left[\eta_t\eta_t'\right] = I$. *The sign restriction is*

$$R = \{\theta_{h,11} \geq 0 \, for \, h = 0, \ldots, 6 \quad \& \quad \theta_{h,22} \geq 0 \, for \, h = 0, \ldots, 6\}.$$

What is the identified set? We have that

$$Y_t = \begin{pmatrix} (1-\alpha_1 L)^{-1} & 0 \\ 0 & (1-\alpha_2 L)^{-1} \end{pmatrix} \Sigma_\eta^{1/2} Q \epsilon_t.$$

Assume for simplicity that $\Sigma_\eta = I$*. We can write this as*

$$\begin{aligned} Y_t &= \begin{pmatrix} (1-\alpha_1 L)^{-1} & 0 \\ 0 & (1-\alpha_2 L)^{-1} \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \epsilon_t \\ &= \begin{pmatrix} (1-\alpha_1 L)^{-1}\cos\theta & -(1-\alpha_1 L)^{-1}\sin\theta \\ (1-\alpha_2 L)^{-1}\sin\theta & (1-\alpha_2 L)^{-1}\cos\theta \end{pmatrix} \epsilon_t \\ &= \Theta(L)\epsilon_t. \end{aligned}$$

So, we have that

$$\Theta_{h,11} = \alpha_1^h \cos\theta, \quad \Theta_{h,22} = \alpha_2^h \sin\theta.$$

For $\Theta(L) \in R$*, we need that* $\cos\theta \geq 0$ *and* $\sin\theta \geq 0$*. This requires that* $0 \leq \theta \leq \pi/2$*. We have that the identified set is given by*

$$0 \leq \theta_{h,21} \leq \alpha_2^h.$$

We can implement Uhlig's algorithm analytically. We draw

$$\tilde{\theta}_{h,21} = \alpha_2^h \sin\theta \quad \theta \sim U[0,\pi].$$

Then,

$$\begin{aligned} \mathbb{E}\left[\tilde{\theta}_{h,21}\right] &= \mathbb{E}\left[\alpha_2^h \sin\theta\right] \\ &= 0.637\alpha_2^h \\ Median(\tilde{\theta}_{h,21}) &= 0.707\alpha_2^h \\ \mathbb{P}\left\{\tilde{\theta}_{h,21}\right\} &= \frac{2}{\pi}\sin^{-1}(\alpha_{21}^{-h}x) \end{aligned}$$

This raises many questions. How are we computing the mean of something that is not identified! It is based on the implicit prior. Baumeister and Hamilton (2015) make this point – the uninformative prior for Q *is dogmatic over the space of IRFs and this can produce strange behavior in the resulting estimator.*

3.6 Local projections

For now, we will continue to maintain the assumption of invertibility. Assume that the observed $n \times 1$ dimensional time series Y_t is represented by the structural vector moving average

$$\underset{n\times 1}{Y_t} = \Theta(L) \underset{m\times 1}{\epsilon_t}.$$

Recall that the assumption of invertibility implies that $n = m$ and $Y_t - \text{Proj}\{Y_t \mid Y_{t-1}, \ldots\} = \eta_t = \Theta_0 \epsilon_t$, where Θ_0^{-1} exists. Additionally recall that ϵ_t are interpreted as structural shocks, meaning that $\mathbb{E}[\epsilon_t \epsilon_t'] = \text{diag}\{\sigma_{\epsilon_1}^2, \ldots, \sigma_{\epsilon_n}^2\}$ and $\mathbb{E}[\epsilon_t \epsilon_s'] = 0$ for $s \neq t$.

Suppose we are interested in the impulse response functions associated with the first shock $\epsilon_{1,t}$. For some $h \geq 0$, using the structural vector moving average form, we can write

$$\begin{aligned} Y_{t+h} &= \Theta(L)\epsilon_{t+h} \\ &= \Theta_0 \epsilon_{t+h} + \Theta_1 \epsilon_{t+h-1} + \ldots + \Theta_{h-1}\epsilon_{t+1} + \Theta_h \epsilon_t + \Theta_{h+1}\epsilon_{t-1} + \ldots \\ &= \Theta_h \epsilon_t + \Theta_{h+1}\epsilon_{t-1} + \ldots + u_{t+h}^{(h)}, \end{aligned}$$

where $u_{t+h}^{(h)} = \Theta_0 \epsilon_{t+h} + \Theta_1 \epsilon_{t+h-1} + \ldots$ only depends on future shocks. Define $\epsilon_{\cdot,t} = (\epsilon_{2,t}, \ldots, \epsilon_{n,t})$ to be the vector all shocks except for the first shock and similarly define $\Theta_{h,\cdot}$ to be the $n \times (n-1)$ matrix that contains all columns of Θ_h except for the first column. With this notation, further rewrite Y_{t+h} as

$$Y_{t+h} = \Theta_{h,1}\epsilon_{1,t} + \Theta_{h,\cdot}\epsilon_{\cdot,t} + \Theta_{h+1}\epsilon_{t-1} + \ldots + u_{t+h}^{(h)}.$$

Now notice that if $\epsilon_{1,t}$ were observed, we could simply run estimate this regression equation directly. However, $\epsilon_{1,t}$ is not observed. So what can we do?

To build intuition, let's consider the simple case in which $\eta_{1,t} = \epsilon_{1,t}$. This corresponds to assuming that Θ_0 is upper triangular and ordering $\epsilon_{1,t}$ last ("cholesky ordered first"). Under this restriction, we can write

$$\begin{aligned} Y_{t+h} &\overset{(1)}{=} \Theta_{h,1}\epsilon_{1,t} + \{\epsilon_{\cdot,t-1}, \epsilon_{\cdot,t-2}, \cdots\} + u_{t+h}^{(h)}, \\ &\overset{(2)}{=} \Theta_{h,1}\eta_{1,t} + \{\eta_{\cdot,t-1}, \eta_{\cdot,t-2}, \cdots\} + u_{t+h}^{(h)}, \\ &\overset{(3)}{=} \Theta_{h,1}(Y_{1,t} - \text{Proj}\{Y_{1,t} \mid Y_{t-1}, \ldots\}) + \{Y_{\cdot,t-1}, Y_{\cdot,t-2}, \cdots\} + u_{t+h}^{(h)} \\ &= \Theta_{h,1}Y_{1,t} + \{Y_{\cdot,t-1}, Y_{\cdot,t-2}, \cdots\} + u_{t+h}^{(h)}, \end{aligned}$$

where (1) folds $\Theta_{h,\cdot}\epsilon_{\cdot,t}$ into the error $u_{t+h}^{(h)}$ and introduces the notation $\{\cdot\}$ to refer to some arbitrary linear combination of the elements in the bracket, (2) uses the assumption $\epsilon_{1,t} = \eta_{1,t}$ and uses invertibility to rewrite the arbitrary linear combination of past shocks as a linear combination of innovations and (3) applies the definition of the innovations to rewrite the arbitrary linear combination of innovations as just lagged values of the observed time series. This is now written completely in terms of observables! In other words, we can identify the impulse response coefficients $\Theta_{h,1}$ by simply directly regressing Y_{t+h} on $Y_{1,t}$ at a variety of horizons, controlling for lagged values of the observed time series. This is the **local projections** approach in its simplest form. Notice that it still *crucially* relies on the assumption of invertibility – intuitively, if the invertibility did not hold, then controlling for lagged values of the observed time series would *not* be sufficient to control for the unobserved lagged shocks.[8]

[8] Put it another way, without the assumption of invertibility, the final regression would suffer from omitted variables bias. This intuition will come back when we discuss LP-IV.

The same idea works if instead we assume that

$$\epsilon_{1,t} = \eta_{1,t} - \text{Proj}\left\{\eta_{1,t} \,|\, \eta_{\cdot,t}\right\},$$

meaning that Θ_0 is upper triangular and we ordered $\epsilon_{1,t}$ first ("cholesky ordered last"). In this case, we have that

$$\begin{aligned}
Y_{t+h} &= \Theta_{h,1}\epsilon_{1,t} + \{\epsilon_{\cdot,t-1}, \epsilon_{\cdot,t-2}, \cdots\} + u_{t+h}^{(h)}, \\
Y_{t+h} &= \Theta_{h,1}\left(\eta_{1,t} - \text{Proj}\left\{\eta_{1,t} \,|\, \eta_{\cdot,t}\right\}\right) + \{\eta_{\cdot,t-1}, \eta_{\cdot,t-2}, \cdots\} + u_{t+h}^{(h)}, \\
&= \Theta_{h,1}\eta_{1,t} + \{\eta_{\cdot,t}\} + \{\eta_{\cdot,t-1}, \eta_{\cdot,t-2}, \cdots\} + u_{t+h}^{(h)}, \\
&= \Theta_{h,1}Y_{1,t} + \{Y_{\cdot,t}, Y_{t-1}, \ldots\} + u_{t+h}^{(h)}.
\end{aligned}$$

All that changes is we now simply need to additionally control for the other contemporaneous values of the observed times $Y_{\cdot,t}$. Once again, it is important to re-iterate what exactly the assumption of invertibility is doing here. Under invertibility, it is sufficient to *linearly control* for the observed time series to identify the full impulse response function from running this regression at a variety of horizons.

Before we begin discussing instrumental variables, we make a few remarks about local projections.

Remark 3.3. *First, local projections tends to be a relatively inefficient estimator. That is, while it identifies the impulse response function under invertibility, it tends to be quite noisy. Moreover, if the true DGP is generated by a VAR, then the local projections estimator is "leaving information on the table" by not exploiting the VAR structure in constructing the impulse responses.*

Second, it is often claimed that local projections can be easily generalized to handle non-linearities. For example, it may appear that we can simply additionally add non-linear functions of the lagged time series in this regression. However, this would, in general, break the underlying identification result. Why? As we saw, the key to this argument is ***invertibility*** *but as we have stated, invertibility is fundamentally a* ***linear concept****. In other words, it is unclear how to generalize the definition of invertibility to handle non-linearities and without this missing piece, it is unclear what the resulting "non-linear local projection" actually delivers.*

3.7 SVAR-IV

We now introduce an additional identification strategy for identifying impulse reponse functions. We will continue to assume invertibility and assume that Y_t is generated by the reduced for VAR with

$$\begin{aligned}
A(L)Y_t &= \eta_t, \\
\eta_t &= \Theta_0\epsilon_t,
\end{aligned}$$

where Θ_0^{-1} exists. The objects of interest are the impulse responses to the first shock, $\epsilon_{1,t}$. Begin by re-writing the expression for η_t as

$$\eta_t = \Theta_{0,1}\epsilon_{1,t} + \Theta_{0,\cdot}\epsilon_{\cdot,t}.$$

Suppose that we observe an **instrument** Z_t for the shock $\epsilon_{1,t}$, meaning that Z_t satisfies

1. **Relevance**: $\mathbb{E}\left[Z_t \epsilon_{1,t}\right] = \alpha \neq 0$,
2. **Contemporaneous exogeneity**: $\mathbb{E}\left[Z_t \epsilon_{\cdot,t}\right] = 0$.

These are the "usual" IV conditions, except they apply to the underlying structural shocks. Notice that with these conditions,

$$\mathbb{E}\left[\eta_t Z_t\right] = \Theta_{0,1}\alpha.$$

Moreover, under the unit-effect normalization, $\Theta_{0,1,1} = 1$, and so

$$\mathbb{E}\left[\eta_{1,t} Z_t\right] = \alpha.$$

Therefore, the impulse responses to the first shock are identified by the Wald ratio

$$\frac{\mathbb{E}\left[\eta_{j,t} Z_t\right]}{\mathbb{E}\left[\eta_{1,t} Z_t\right]} = \Theta_{0,j,1}.$$

This suggests a simple strategy for identifying the impulse responses to the first shock. We simply use Z_t as an instrument for $\eta_{1,t}$ in the regression

$$\eta_{j,t} = \Theta_{0,j,1}\eta_{1,t} + v_t,$$

where v_t is some error correlated with $\eta_{1,t}$.

Why is this called SVAR-IV? Notice that we are still assuming that Y_t is generated by a reduced-form VAR and only using the instrument to identify the relevant elements of the matrix contemporaneous coefficients, Θ_0. To obtain the full impulse response function, we would then use the underlying VAR as before.

3.8 LP-IV

This will generalize the local projections method to incorporate an instrumental variable. Moreover, this will be the first method that we consider which will no longer impose invertibility. We assume that Y_t still follows a SVMA, but do not require that the SVMA be invertible. This means that

$$\underset{n\times 1}{Y_{t+h}} = \Theta(L)\underset{m\times 1}{\epsilon_{t+h}},$$

where $m \geq n$. Using the SVMA form to re-write the observed time series as

$$Y_{t+h} = \theta_{h,1}\epsilon_{1,t} + \left\{\epsilon_{t+h}, \ldots, \epsilon_{t+1}, \epsilon_{\cdot,t}, \epsilon_{t-1}, \ldots\right\},$$

where as before, $\{\cdot\}$ refers to an arbitrary linear combination. Imposing the unit effect normalization, $\Theta_{0,1,1} = 1$, we see that

$$\begin{aligned} Y_{1,t} &= \Theta_{0,1,1}\epsilon_{1,t} + \{\epsilon_{\cdot,t}, \epsilon_{t-1}, \ldots\}, \\ &= \epsilon_{1,t} + \{\epsilon_{\cdot,t}, \epsilon_{t-1}, \ldots\}. \end{aligned}$$

Re-arranging, we can substitute in for $\epsilon_{1,t}$ to write

$$Y_{t+h} = \theta_{h,1} Y_{1,t} + \{\epsilon_{t+h}, \ldots, \epsilon_{t+1}, \epsilon_{\cdot,t}, \epsilon_{t-1}, \ldots\}.$$

Now, suppose that we have access to an **instrument** Z_t that satisfies

1. **Relevance**: $\mathbb{E}[Z_t \epsilon_{1,t}] = \alpha \neq 0$,
2. **Contemporaneous exogeneity**: $\mathbb{E}[Z_t \epsilon_{\cdot,t}] = 0$
3. **Lead-lag exogeneity**: $\mathbb{E}[Z_t \epsilon_{t+k}] = 0$ for all $k = \pm 1, \pm 2, \ldots$

From a similar argument as in SVAR-IV, we can immediately see that the Wald ratio delivers the impulse response coefficient $\Theta_{h,j,1}$

$$\Theta_{h,j,1} = \frac{\mathbb{E}\left[Z_t Y_{j,t+h}\right]}{\mathbb{E}\left[Z_t Y_{1,t}\right]}.$$

In other words, we can estimate the h-period ahead impulse response coefficient by simply using Z_t as an instrument for $Y_{1,t}$ in the regression of $Y_{j,t+h}$ on $Y_{1,t}$.

Notice that we are additionally imposing that the instrument be unconfounded with future and lagged shocks – this is the additional lead-lag exogeneity condition. In contrast, we only had to impose contemporaneous exogeneity in SVAR-IV. Why is this different? In LP-IV, we will use the instrument to identify the impulse response coefficient at *each* horizon, whereas in SVAR-IV, we only used the instrument to identify the contemporaneous coefficients. It is crucial to notice that we did not need to invoke invertibility for this identification argument to work. However, we need strong conditions on the instrument Z_t – contemporaneous exogeneity states that the instrument must be uncorrelated with all contemporaneous shocks and lead-lag exogeneity implies that it must be uncorrelated with all future and lagged shocks. Since the ϵ_t's are shocks, the condition that the instrument be uncorrelated with future shocks is not particularly strong in empirical applications. However, the condition that the instrument be uncorrelated with past shocks may bite.

To see this, let's consider LP-IV with additional controls. That is, consider the regression

$$Y_{j,t+h} = \Theta_{h,j,1} Y_{1,t} + \gamma' W_{t-1} + u_{t+h}^{(h)},$$

where W_{t-1} is a vector of lagged values of the observed time series. Define the residualized instrument as

$$Z_t^{\perp} = Z_t - \text{Proj}\{Z_t \mid W_{t-1}\}$$

and analogously define $Y_t^{\perp}$. With controls, the LP-IV conditions are

1. **Relevance**: $\mathbb{E}\left[\epsilon_{1,t} Z_t^{\perp}\right] \neq 0$
2. **Contemporaneous exogeneity**: $\mathbb{E}\left[\epsilon_{\cdot,t} Z_t\right] = 0$
3. **Lead-lag exogeneity**: $\mathbb{E}\left[\epsilon_{t+k} Z_t^{\perp}\right] = 0$ for $k = \pm 1, \pm 2, \ldots$

For sake of argument, suppose that the identification condition we were worried about is lag exogeneity and we introduce the controls W_{t-1} to possibly address this worry. What is a sufficient condition for $Z_t^{\perp}$ to satisfy lag exogeneity? A simple sufficient condition is precisely **invertibility**. Trivially, if W_{t-1} spans the space of past shocks, then $Z_t^{\perp}$ is orthogonal to all lagged shocks. This suggests that there is some sort of "no free lunch" at play – if you have an instrument that you only believe satisfies lag exogeneity after controlling for lagged values of the time series, then this identification strategy becomes "equivalent" to assuming invertibility in the first place.

4 Empirical Processes and the Functional Central Limit Theorem

Suppose that $\epsilon_t \sim WN(0, \sigma_\epsilon^2$ and $X_t = \sum_{s=1}^t \epsilon_s$. We want to approximate the distribution of some function of $X_1, \ldots, X_T$. How do we do this?

Recall 2. *Recall the following results:*

- ***CLT**: Let ϵ_t be a martingale difference sequence with variance σ_ϵ^2. Then,*

$$\xi_T = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \epsilon_t \xrightarrow{d} N(0, \sigma_\epsilon^2).$$

- ***CMT**: Let g be a continuous function. Then, if $\xi_T \xrightarrow{d} \xi$, then*

$$g(\xi_T) \xrightarrow{d} g(\xi).$$

In this section, we extend the central limit theorem to random functions. In doing so, standard brownian motion will take the place of standard normal random variables.

Definition 4.1. *Denote a **standard brownian motion** by $W(s)$ $s \in [0, 1]$. This satisfies*

1. *$W(0) = 0$,*
2. *$\forall 0 \leq t_1 < t_2 < \ldots < t_K \leq 1, W(t_2) - W(t_1), W(t_3) - W(t_2), \ldots, W(t_K) - W(t_{K-1})$ are independent with $W(t_i) - W(t_{i-1}) \sim N(0, t_i - t_{i-1})$.*
3. *Realizations of $w(s)$ are continuous with probability one.*

Example 4.1. *Let $X_t = \sum_{s=1}^t \epsilon_s$ with $\epsilon_s \sim WN(0, \sigma_\epsilon^2)$. Then, X_t is a random walk with*

$$\Delta X_t = \epsilon_t, \quad X_0 = 0.$$

We convert this a function $X_{\lfloor T\tau \rfloor}$, where $\lfloor T\tau \rfloor$ is the floor function and $\tau \in [0,1]$. We set

$$\rho_T(\tau) = X_{\lfloor T\tau \rfloor} = \sqrt{\frac{1}{T}} \sum_{s=1}^{\lfloor T\tau \rfloor} \epsilon_s.$$

This is a random function. We'll show that

$$\rho_T(\tau) \xrightarrow{d} \sigma_\epsilon W(\tau),$$

where $W(\tau)$ is a standard brownian motion.

4.1 Empirical processes, function spaces and the FCLT

Let $g(W_t, \tau) \in \mathbb{R}^s$ be a function of the random variable W_t. Define

$$\xi_T(\tau) = \sqrt{\frac{1}{T}} \sum_{t=1}^{T} \left(g(W_t, \tau) - \mathbb{E}\left[g(W_t, \tau) \right] \right).$$

This is an **empirical process**. Note that for fixed τ, we could apply a central limit theorem under our usual regularity conditions. Now, $\xi_T(\tau)$ is a function of τ and we want to study its behavior as a random function.

Example 4.2. *Consider $g(W_t, \tau) = W_t \mathbb{1}\left\{ \frac{t}{T} \leq \tau \right\}$ for $\tau \in (0,1)$. Then,*

$$\begin{aligned} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} g(W_t, \tau) &= \frac{1}{\sqrt{T}} \sum_{t=1}^{T} W_t \mathbb{1}\left\{ \frac{t}{T} \leq \tau \right\} \\ &= \frac{1}{\sqrt{T}} \sum_{t=1}^{\tau T} W_t. \end{aligned}$$

Example 4.3 (GMM)**.** *Consider $g(W_t, \tau) = h(Y_t, \tau) \otimes Z_t$. As an example, $h(Y_t, \tau) \otimes Z_t = (Y_t - \tau X_t) Z_t$. Then,*

$$\xi_T(\tau) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left(h(Y_t, \tau) \otimes Z_t - \mathbb{E}\left[h(Y_t, \tau) \otimes Z_t \right] \right).$$

Example 4.4. *Let $g(W_t, \tau) = W_t \mathbb{1}\{ t \leq \lfloor \tau T \rfloor \}$, where $\mathbb{E}[W_t] = 0$. Then,*

$$\begin{aligned} \xi_T(\tau) &= \sqrt{\frac{1}{T}} \sum_{t=1}^{T} \left(W_t \mathbb{1}\{ t \leq \lfloor \tau T \rfloor \} - \mathbb{E}\left[W_t \mathbb{1}\{ t \leq \lfloor \tau T \rfloor \} \right] \right) \\ &= \sqrt{\frac{1}{T}} \sum_{t=1}^{T} W_t \mathbb{1}\{ t \leq \lfloor \tau T \rfloor \} \\ &= \sqrt{\frac{1}{T}} \sum_{t=1}^{\lfloor \tau T \rfloor} W_t. \end{aligned}$$

Consider $0 < \tau_1 < \tau_2 < 1$. *Then,*

$$\begin{pmatrix} \xi_T(\tau_1) \\ \xi_T(\tau_2) \end{pmatrix} = \begin{pmatrix} \sqrt{1/T}\sum_{t=1}^{\tau_1 T} W_t \\ \sqrt{1/T}\sum_{t=1}^{\tau_2 T} W_t \end{pmatrix} \xrightarrow{N} (0, \Omega),$$

where

$$\Omega = \sigma_\epsilon^2 \begin{pmatrix} \tau_1 & \min(\tau_1, \tau_2) \\ \min(\tau_1, \tau_2) & \tau_2 \end{pmatrix}.$$

Let $\mathcal{C}[0,1]$ denote the space of continuous functions on $[0,1]$. Our metric on $\mathcal{C}[0,1]$ is

$$d(f,g) = \sup_{\tau \in [0,1]} |f(\tau) - g(\tau)|.$$

Therefore, we say that $f_T(\cdot)$ **converges in probability** to $f(\cdot)$ on $C[0,1]$ if for all $\delta > 0$,

$$\mathbb{P}\left\{d(f_T, f) > \delta\right\} \to 0$$

as $T \to \infty$. In $\mathbb{R}$, the CLT stated that $T^{-1/2}\sum X_t \xrightarrow{d} N(0, \sigma^2)$ for sequence of mean-zero random variables under some regularity conditions. We'll see that in $\mathcal{C}[0,1]$, the CLT will become $\xi_T(\cdot) \xrightarrow{d} W(\cdot)$ under some conditions. Finally, we note that the CMT will also extend to $\mathcal{C}[0,1]$ – if $h : \mathcal{C}[0,1] \to \mathbb{R}$ is continuous and $\xi_T(\cdot) \xrightarrow{d} \xi(\cdot)$ on $\mathcal{C}[0,1]$, then $h(\xi_T(\cdot)) \xrightarrow{d} h(\xi(\cdot))$.

Example 4.5. *Consider*

$$Y_t = \beta d(t/T) + u_t,$$

where $\frac{1}{T}\sum_{t=1}^{T} d(t/T) = 0$, $u_t = u_{t-1} + \epsilon t$ *and* $\epsilon \sim WN(0, \sigma_\epsilon^2)$. *Note that this implies that* $u_t = \sum_{s=1}^{t} \epsilon_s$.

Then, we have that

$$\sqrt{\frac{1}{T}}\left(\hat{\beta} - \beta\right) = \frac{\frac{1}{T}\sum_{t=1}^{T} d(t/T) u_t / \sqrt{T}}{\frac{1}{T}\sum_{t=1}^{T} d^2(t/T)}.$$

Define

$$\xi_T(\tau) = u_{\lfloor \tau T \rfloor}/\sqrt{T} = \frac{1}{\sqrt{T}} \sum_{s=1}^{\lfloor \tau T \rfloor} \epsilon_s, \quad \tau = t/T.$$

Then, by the definition of an integral of a step function,

$$\sqrt{\frac{1}{T}}(\hat{\beta} - \beta) = \frac{\int_0^1 d(\tau)\xi_T(\tau)d\tau}{\int_0^1 d^2(\tau)d\tau}.$$

We'll argue that

$$\frac{\int_0^1 d(\tau)\xi_T(\tau)d\tau}{\int_0^1 d^2(\tau)d\tau} \xrightarrow{d} \frac{\sigma_\epsilon^2 \int_0^1 W(\tau)d\tau}{\int_0^1 d^2(\tau)d\tau}. \tag{4}$$

How are we going to make arguments like this? Note that $\xi_T(\tau)$ is not continuous – it's a step function. Therefore, we will introduce a new function that "connects the dots" and is continuous. Define

$$\tilde{\xi}_T(\tau) = \xi_T(\tau) + \frac{\tilde{g}_{\lfloor \tau T \rfloor + 1}}{\sqrt{T}} \left(\tau T - \lfloor \tau T \rfloor \right),$$

where $\tilde{g}_t = g(W_t, \tau) - \mathbb{E}\left[g(W_t, \tau) \right]$ to simplify notation. We will use the "connect the dots" function to approximate $\xi_T(\tau)$. We'll first argue heuristically that this is a valid approximation. Suppose that $\mathbb{E}\left[\tilde{g}_t^{2+\delta} \right] < \infty$ for $\delta > 0$. By Markov's inequality,

$$\mathbb{P}\left\{ d(\tilde{\xi}_T, \xi_T) > \eta \right\} \leq \mathbb{E}\left[\frac{d(\tilde{\xi}_T, \xi_T)^{2+\delta}}{\eta^{2+\delta}} \right].$$

Now, we have that

$$\begin{aligned}
\mathbb{E}\left[d(\tilde{\xi}_T, \xi_T)^{2+\delta} \right] &= \mathbb{E}\left[\left(\sup_\tau \left| \tilde{\xi}_T(\tau) - \xi_T(\tau) \right| \right)^{2+\delta} \right] \\
&= \mathbb{E}\left[\sup_\tau \frac{\tilde{g}^{2+\delta}_{\lfloor \tau T \rfloor + 1}}{(\sqrt{T})^{2+\delta}} (\tau T - \lfloor \tau T \rfloor)^{2+\delta} \right] \\
&\leq \mathbb{E}\left[\max_t |\tilde{g}_t|^{2+\delta} \cdot \frac{1}{T} \cdot \frac{1}{T^{\delta/2}} \right] \\
&\leq \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[|\tilde{g}_t|^{2+\delta} \right] \cdot \frac{1}{T^{\delta/2}} \to 0.
\end{aligned}$$

So as $T \to \infty$, we have that

$$\tilde{\xi}_T(\cdot) - \xi_T(\cdot) \xrightarrow{p} 0.$$

Theorem 4.1 (Functional central limit theorem)**.** *Let $\xi_t = \xi_t(\cdot) \in \mathcal{C}[0,1]$ be a random continuous function with $\mathbb{E}\left[\xi_t(\tau) \right] = 0$ for all $\tau \in [0,1]$. Define the metric*

$$d(f, g) = \sup_{0 \leq \tau \leq 1} |f(\tau) - g(\tau)|.$$

Then,

$$\xi_t(\cdot) \xrightarrow{d} \xi^*(\cdot) \in \mathcal{C}[0,1]$$

if

1. ***Convergence of finite-dimensional distributions (FDD)**: Let $0 < \tau_1 < \ldots < \tau_k \leq 1$. Then,*

$$\begin{pmatrix} \xi_t(\tau_1) \\ \vdots \\ \xi_t(\tau_k) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \xi^*(\tau_1) \\ \vdots \\ \xi^*(\tau_k) \end{pmatrix}.$$

2. ***Tightness**: $\xi_t(\cdot)$ is **tight**, meaning that*

   a. *For each $\epsilon > 0$,*

$$\mathbb{P}\left\{\sup_{|\tau_1-\tau_2|<\delta} |\xi_t(\tau_1) - \xi_t(\tau_2)| > \epsilon\right\} \to 0$$

   as $\delta \to 0$.

   b. $\mathbb{P}\{|\xi_t(0)| > \lambda\} \to 0$ *as* $\lambda \to \infty$.

Condition (1) is usually simple to verify. It states that at any fixed K points, the joint distribution of $\xi_t(\cdot)$ is well behaved as $T \to \infty$. Since it describes the convergence of a random vector, we can apply our usual central limit theorem arguments for this condition. However, Condition (1) alone is not enough to guarantee functional convergence. We additionally need conditions that enable us to control how the function behaves between those points. Condition (2) does so, which we can think of tightness as a form of stochastic continuity.

Moreover, notice that the FLCT does not tell us what the limiting finite-dimensional distribution is. Instead, it states that if we know the limiting finite-dimensional distribution, then we can pass the continuous limit and construct the limiting empirical process.

To illustrate how the FCLT is used as a tool for analysis in non-standard problems, we'll now consider a series of examples.

Example 4.6. *Recall the connect the dots function from earlier:*

$$\tilde{\xi}_T(\tau) = \sqrt{\frac{1}{T}}\sum_{s=1}^{\lfloor \tau T \rfloor} \epsilon_t + \sqrt{\frac{1}{T}}\epsilon_{\lfloor \tau T \rfloor + 1}(\tau T + \lfloor \tau T \rfloor).$$

Assuming that $\mathbb{E}\left[\epsilon_t^{2+\delta}\right] < \infty$, recall that we showed that the "connect the dots" function is asymptotically negligible – i.e., $o_p(1)$. Therefore, we'll apply the FCLT to derive the limiting distribution of $\tilde{\xi}_T(\tau)$ by focusing on the limiting process of $\sqrt{\frac{1}{T}}\sum_{s=1}^{\lfloor \tau T \rfloor} \epsilon_t$.

First, consider the limiting finite-dimensional distribution of this process. Assume that $\tau_1 < \tau_2 < \ldots < \tau_k$. Via a central limit theorem (e.g., Gordin's conditions), we can show that

$$\begin{pmatrix} \xi_T(\tau_1) \\ \vdots \\ \xi_T(\tau_K) \end{pmatrix} \xrightarrow{d} N(0, V),$$

where V is some variance-covariance matrix. We have that it's diagonal elements will be equal to

$$Var\left(\sqrt{\frac{1}{T}}\sum_{s=1}^{\lfloor \tau T \rfloor} \epsilon_t\right) = \sigma_\epsilon^2 \frac{\lfloor \tau T \rfloor}{T} \to \tau\sigma_\epsilon^2.$$

Next, for $\tau < \tau'$, the off-diagonal elements will be

$$\mathbb{E}\left[\left(\sqrt{\frac{1}{T}}\sum_{s=1}^{\lfloor \tau T \rfloor}\epsilon_t\right)\left(\sqrt{\frac{1}{T}}\sum_{s=1}^{\lfloor \tau' T \rfloor}\epsilon_t\right)\right] = \sigma_\epsilon^2 \frac{\lfloor \tau T \rfloor}{T} \to \sigma_\epsilon^2 \tau.$$

Therefore, the finite-dimensional distribution converges to a multivariate normal distribution with variance-covariance matrix

$$\sigma_\epsilon^2 \begin{pmatrix} \tau_1 & & & & \\ \tau_1 & \tau_2 & & & \\ \tau_1 & \tau_2 & \tau_3 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \tau_1 & \tau_2 & \tau_3 & \dots & \tau_k \end{pmatrix}$$

*We can verify that tightness holds. Therefore, this converges to a **gaussian process** and given the variance-covariance matrix, it converges to a **brownian motion** as the variance-covariance matrix equals the covariance kernel of a brownian motion. This means that*

$$\xi_T(\cdot) \xrightarrow{d} \sigma_\epsilon W(\cdot),$$

where $W(\cdot)$ is a standard Brownian motion.

Example 4.7. *Now, consider the function of the empirical process*

$$\frac{1}{T}\sum_{t=1}^{T}\xi_T(t/T).$$

Following the "connect the dots" approximation, we approximate this function of the empirical process $\xi_T(\cdot)$ with the connect the dots function $\tilde{\xi}_T(\cdot)$. Since we showed that the approximation error associated with this approximation is asymptotically negligible, we will ignore it. Therefore, we have that

$$\frac{1}{T}\sum_{t=1}^{T}\xi_T(t/T) = \int_0^1 \xi_T(\tau)d\tau,$$

up to $o_p(1)$ terms and we applied the definition of the integral of a step function to replace the sum with the integral. Now, define the function $h : C[0,1] \to \mathbb{R}$, where $h(f) = \int_0^1 f(\tau)d\tau$. This is a continuous function, and so we apply the continuous mapping theorem to show that

$$\frac{1}{T}\sum_{t=1}^{T}\xi_T(t/T) \xrightarrow{d} \sigma_\epsilon \int_0^1 \xi_T(\tau)d\tau.$$

Example 4.8. *Consider a random walk $X_t = X_{t-1} + \epsilon_t$, where $X_1 = \epsilon_1$, and so $X_t = \sum_{s=1}^{t}\epsilon_s$. Notice that if*

we just consider the usual sample average

$$\frac{1}{T}\sum_{t=1}^{T} X_t^2 = \frac{1}{T}\sum_{t=1}^{T}\left(\sum_{s=1}^{t}\epsilon_s\right)^2,$$

this will blow up and does not converge as the variance of the random walk is growing over time. However, if we instead divide by $\frac{1}{T^2}$, the resulting sum is bounded in probability and we can use the FCLT to analyze its limiting distribution. We have that

$$\begin{aligned}\frac{1}{T^2}\sum_{t=1}^{T} X_t^2 &= \frac{1}{T}\sum_{t=1}^{T}\left(X_t/\sqrt{T}\right)^2 \\ &= \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{\sqrt{T}}\sum_{s=1}^{t}\epsilon_s\right)^2,\end{aligned}$$

where $\frac{1}{\sqrt{T}}\sum_{s=1}^{t}\epsilon_s = \xi_T(t/T)$ is the rolling sum process. Therefore, we have that

$$\frac{1}{T^2}\sum_{t=1}^{T} X_t^2 = \frac{1}{T}\sum_{t=1}^{T}\xi_T^2(t/T).$$

Applying the connect the dots transformation and replacing the summation with an integral as before, we arrive at

$$\begin{aligned}\frac{1}{T^2}\sum_{t=1}^{T} X_t^2 &= \frac{1}{T}\sum_{t=1}^{T}\xi_T^2(t/T) \\ &= \int_0^1 \xi_T^2(\tau)d\tau \xrightarrow{d} \sigma_\epsilon^2\int_0^1 W^2(\tau)d\tau,\end{aligned}$$

where we again ignored the approximation errors from the connect the dots function and applied the FLCT and the continuous mapping theorem.

Example 4.9. *Consider $y_t = \beta_0 + \beta_1 D_t + u_t$, where $u_t = u_{t-1} + \epsilon_t$, $\epsilon_t \sim WN(0, \sigma_\epsilon^2)$. Therefore, u_t is a random walk and a martingale difference sequence. Let*

$$\tilde{D}_t = D_t - \bar{D} = D_t - \frac{1}{T}\sum_{t=1}^{T} D_t.$$

Then, by Frisch-Waug-Lovell, we have that

$$\hat{\beta} - \beta_1 = \frac{\sum_{t=1}^{T}\tilde{D}_t\tilde{u}_t}{\sum_{t=1}^{T}\tilde{D}_t^2}.$$

We model $\tilde{D}_t$ as a random step function in τ with $d(\tau) = \tilde{D}_{\lfloor \tau T \rfloor}$. Here, as usual, there will be an additional "connect the dots" piece so that we actually approximate this with a continuous function. So, we have that

$$\frac{1}{T}\sum_{t=1}^{T}\tilde{D}_t^2 = \frac{1}{T}\sum_{t=1}^{T} d(t/T)^2 = \int_0^1 d(\tau)^2 d\tau.$$

How should we deal with the numerator $\frac{1}{T}\sum_{t=1}^{T}\tilde{D}_t\tilde{u}_t$? It is not stationary! So, we can't just use our HAC/HAR tricks. Instead, we'll apply the FCLT. We have that

$$u_t = \sum_{s=1}^{t}\epsilon_s$$
$$\sqrt{\frac{1}{T}}u_t = \sqrt{\frac{1}{T}}\sum_{s=1}^{t}\epsilon_t.$$

We'll model this as

$$\sqrt{\frac{1}{T}}u_t = \xi_T(t/T), \quad \xi_T(\tau) = \sqrt{\frac{1}{T}}\sum_{s=1}^{\lfloor \tau T \rfloor}\epsilon_s.$$

Then, we have that

$$\sqrt{\frac{1}{T}}(\hat{\beta}_1 - \beta_1) = \frac{\frac{1}{T}\sum_{t=1}^{T}\tilde{D}_t\tilde{u}_t/\sqrt{T}}{\frac{1}{T}\sum_{t=1}^{T}\tilde{D}_t^2},$$

where

$$\begin{aligned}\tilde{u}_t/\sqrt{T} &= \sqrt{\frac{1}{T}}(u_t - \bar{u}) \\ &= \xi_T(t/T) - \frac{1}{T}\sum_{s=1}^{T}\xi_T(s/T) \\ &= \xi_T(t/T) - \int_0^1 \xi_T(\tau)d\tau.\end{aligned}$$

Therefore, we have that the numerator can be written as

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\tilde{D}_t\tilde{u}_t/\sqrt{T} &= \frac{1}{T}\sum_{t=1}^{T}d(t/T)\left(\xi_T(t/T) - \int_0^1\xi_T(\tau)d\tau\right) \\ &= \int_0^1 d(\tau)\xi_T(\tau)d\tau - \left(\int_0^1 d(\tau)d\tau\right)\left(\int_0^1\xi_T(\tau)d\tau\right), \\ &= \int_0^1 d(\tau)\xi_T(\tau)d\tau.\end{aligned}$$

Again, here we are skipping a step involving a "connect the dots" approximation that we argue will be asymptotically negligible. This is necessary to ensure that the empirical process is continuous. Using the FCLT, we can show that

$$\xi_T(\cdot) \xrightarrow{d} \sigma_\epsilon W(\cdot).$$

Then, by the continuous mapping theorem,

$$\frac{1}{T}\sum_{t=1}^{T}d(t/T)\tilde{u}_t/\sqrt{T} \xrightarrow{d} \sigma_\epsilon\int_0^1 d(\tau)W(\tau)d\tau.$$

Therefore, we have that

$$\sqrt{1/T}(\hat{\beta} - \beta) \xrightarrow{d} \frac{\sigma_\epsilon \int_0^1 d(\tau) W(\tau) d\tau}{\int_0^1 d(\tau)^2 d\tau}.$$

We can use this limiting distribution to perform hypothesis tests. In particular, it means that

$$\sqrt{1/T}(\hat{\beta} - \beta) \overset{approx}{\sim} N(0, V),$$

where

$$\begin{aligned} V\left(\sigma_\epsilon \int_0^1 d(\tau) W(\tau) d\tau\right) &= \sigma_\epsilon^2 \int_s \int_\tau d(s) d(\tau) \mathbb{E}\left[W(s) W(\tau)\right] d\tau ds \\ &= \sigma_\epsilon^2 \int_s \int_\tau d(s) d(\tau) \min(s, \tau) d\tau ds, \end{aligned}$$

which follows from properties of the covariance kernel of a brownian motion. Therefore,

$$V = \frac{\sigma_\epsilon^2 \int_s \int_\tau d(s) d(\tau) \min(s, \tau) d\tau ds}{\left(\int_0^1 d(s)^2 ds\right)^2}.$$

We end this example by making two comments. First, we used the FCLT because the error was not stationary. Second, notice that in deriving the asymptotic distribution, we divided by $\sqrt{1/T}$ and not $\sqrt{T}$. Why? Suppose that $D_t = 1$. Then, we're estimating the mean of a brownian motion. Recall that $V(W_t) = O(t)$ and so, to stabilize this asymptotically, we need to divide by $\sqrt{T}$.

4.2 FCLT for dependent increments

Now suppose that $\Delta y_t = u_t$, where u_t is serially correlated. That is, $u_t = c(L)\epsilon_t$. We can extend our FCLT tools to cover this case.

4.2.1 Beveridge-Nelson decomposition

We now show that we can decompose a time series into a random trend and stochastic component.

Proposition 4.1. *Consider a time series $\{u_t\}$ with $u_t = c(L)\epsilon_t$. We can write*

$$u_t == c(1)\epsilon_t + c^*(L)\Delta\epsilon_t,$$

where $c_i^* = -\sum_{j=i+1}^{\infty} c_j$. That is, we can write y_t as the sum of a random walk component and a weakly stationary process.

Proof. We have that

$$\begin{aligned}
u_t &= C(L)\epsilon_t \\
&= c_0\epsilon_t + c_1\epsilon_{t-1} + c_2\epsilon_{t-2} + \dots \\
&= C(1)\epsilon_t - c_1\epsilon_t - c_2\epsilon_{t-2} - \dots + c_1\epsilon_{t-1} + c_2\epsilon_{t-2} + \dots \\
&= C(1)\epsilon_t - c_1\Delta\epsilon_t - c_2(\epsilon_t - \epsilon_{t-2}) - c_3(\epsilon_t - \epsilon_{t-3}) - \dots \\
&= C(1)\epsilon_t - c_1\Delta\epsilon_t - c_2(\Delta\epsilon_t + \Delta\epsilon_{t-1}) - c_3(\Delta\epsilon_t + \Delta\epsilon_{t-1} + \Delta\epsilon_{t-2}) - \dots \\
&= C(1)\epsilon_t - (c_1 + c_2 + c_3 + \dots)\Delta\epsilon_t - (c_2 + c_3 + \dots)\Delta\epsilon_{t-1} - \dots
\end{aligned}$$

□

Example 4.10. *Let $\xi_T(\tau) = \frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor \tau T \rfloor} X_t$, where $X_t = c(L)\epsilon_t$. Then, using the Beveridge-Nelson decomposition*

$$\begin{aligned}
\xi_T(\tau) &= \frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor \tau T \rfloor}\left(C(1)\epsilon_t + C^*(L)\Delta\epsilon_t\right), \\
&= C(1)\frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor \tau T \rfloor}\epsilon_T + \frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor \tau T \rfloor} C^*(L)\Delta\epsilon_t.
\end{aligned}$$

The second term is a telescoping sum. This simplifies down to

$$\xi_T(\tau) = C(1)\frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor \tau T \rfloor}\epsilon_T + \frac{1}{\sqrt{T}}\left(C^*(L)\epsilon_{\lfloor \tau T \rfloor} - C^*(1)\epsilon_1\right).$$

The second term is $o_p(1)$, and so applying the FCLT, we see that

$$\xi_T(\tau) \xrightarrow{d} \underbrace{C(1)\sigma_\epsilon}_{\Omega^{1/2}} W(\cdot),$$

where $W(\cdot)$ is a brownian motion as before and now Ω is the long-run variance of the process.

4.3 Break Tests

In this section, we discuss a classic application of the FCLT to time series: break tests. The idea is that we have some observed time series and wish to test whether there is a "break" in its behavior at some point in time. For example, suppose the observed time series consist of several macroeconomic aggregates such as GDP, unemployment and inflation. We may wish to test whether there is a break in the variance of these time series at some time in the 1980s to test whether there exists a "great moderation" in the observed data. The literature on break tests is extremely well-developed.

Consider the time series regression

$$Y_t = \beta_1' X_t \mathbb{1}\{t \leq r\} + \beta_2' X_t \mathbb{1}\{t > r\} + u_t,$$

where (X_t, u_t) are stationary. We wish to test the null hypothesis, $H_0 : \beta_1 = \beta_2$ – that is, we wish to test whether there is a break in the coefficient on X_t at some point in the sample period. The challenge

is that the break date r is unknown and under the null hypothesis of no break, the break date r is not identified. We will use the empirical process toolkit we developed to deal with these challenges.

To build intuition, suppose that the break date r is known. Then, to test whether there is a break, we would simply estimate the regression

$$Y_t = \beta_1' X_t \mathbb{1}\{t \leq r\} + \beta_2' X_t \mathbb{1}\{t > r\} + u_t,$$

and construct the F-statistic to test whether $\beta_1 = \beta_2$. Let $F_T(r/T)$ denote this test statistic. Under the usual regularity conditions, the limiting distribution of the F-statistic will be a chi-square distribution with known degrees of freedom. Since r is unknown, we could simply repeat this test at all possible values of r and apply a Bonferroni correction to ensure proper size. However, this approach will be very conservative (underpowered). Instead, a natural thought would be to use the *maximum F-statistic* as our test statistic of $H_0 : \beta_1 = \beta_2$, r unknown. The maximum F-statistic is commonly referred to as the *sup-wald statistic* and it is given by

$$\max_{r=1,\ldots,T} F_T(r/T).$$

While this is an intuitive choice, what is its limiting distribution? How do we select critical values for this test statistic? It turns out that the limiting distribution of the sup-wald statistic is relatively simple and elegant. We will now derive it.

Once again, fix r. Then, under usual regularity conditions, we know that

$$\begin{pmatrix} \sqrt{T}(\hat{\beta}_1 - \beta_1) \\ \sqrt{T}(\hat{\beta}_2 - \beta_2) \end{pmatrix} \xrightarrow{d} N(0, \Sigma).$$

By construction, $\hat{\beta}_1$ only depends on the first r terms and $\hat{\beta}_2$ only depends on the $r+1, \ldots, T$ terms. We have that

$$\sqrt{T}(\hat{\beta}_1 - \beta_1) = \left(\frac{1}{T} \sum_{t=1}^{T} X_t X_t' \mathbb{1}\{t \leq r\} \right)^{-1} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t u_t \mathbb{1}\{t \leq r\} \right)$$

and define

$$V_T(r/T) \equiv \frac{1}{T} \sum_{t=1}^{\lfloor rT \rfloor} X_t X_t',$$
$$\xi_T(r/T) = \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor rT \rfloor} X_t u_t.$$

Throughout this calculation, there is a "connect the dots" step that we skip since we've argued earlier that the approximation error associated with this step will be asymptotically negligible. So, we have that

$$\sqrt{T}(\hat{\beta}_1 - \beta_1) = V_T(r/T)^{-1} \xi_T(r/T).$$

Similarly, we have that

$$\sqrt{T}(\hat{\beta}_2 - \beta_2) = (V_T(1) - V_T(r/T))^{-1}(\xi_T(1) - \xi_T(r/T)).$$

Our estimate of the variance-covariance matrix $\hat{\Sigma}$ is given by

$$\hat{\Sigma} \approx \begin{pmatrix} V_T(r/T)^{-1}\hat{\Omega}(r/T)V_T(r/T)^{-1} & 0 \\ 0 & (V_T(1) - V_T(r/T))^{-1}\left(\hat{\Omega}(1) - \hat{\Omega}(r/T)\right)(V_T(1) - V_T(r/T))^{-1} \quad 0 \end{pmatrix}.$$

Why are the off-diagonal terms approximately zero? We assumed that (X_t, u_t) are stationary. If the auto-covariances are absolutely summable, then as $T \to \infty$, the two non-overlapping blocks will be approximately uncorrelated. This is a handwavy argument that would need to be formalized but take it as given. Now, consider the F-statistic that tests whether $\beta_1 = \beta_2$. This is given by

$$\begin{aligned} F_T(r/T) &= \left(\sqrt{T}(\hat{\beta}_1 - \hat{\beta}_2)\right)' \left(\hat{\Sigma}_{1,1} + \hat{\Sigma}_{2,2}\right) \left(\sqrt{T}(\hat{\beta}_1 - \hat{\beta}_2)\right) \\ &= \left(V_T(r/T)^{-1}\xi_T(r/T) - (V_T(1) - V_T(r/T))^{-1}(\xi_T(1) - \xi_T(r/T))\right)' \cdot \\ &\left(V_T(r/T)^{-1}\hat{\Omega}(r/T)V_T(r/T)^{-1} + (V_T(1) - V_T(r/T))^{-1}\left(\hat{\Omega}(1) - \hat{\Omega}(r/T)\right)(V_T(1) - V_T(r/T))^{-1}\right)^{-1} \cdot \\ &\left(V_T(r/T)^{-1}\xi_T(r/T) - (V_T(1) - V_T(r/T))^{-1}(\xi_T(1) - \xi_T(r/T))\right), \end{aligned}$$

which is a true monstrosity. It turns out that this will simplify enormously. Define $\lambda = r/T$. Assume that

$$\xi_T(\cdot) \xrightarrow{d} \Omega^{1/2}W(\cdot),$$

where $W(\cdot)$ is a brownian motion. This requires placing some high-level regularity conditions on the process $X_t u_t$. Next, consider the process

$$\begin{aligned} V_T(\lambda) &= \frac{1}{T}\sum_{t=1}^{\lfloor \lambda T \rfloor} X_t X_t' \\ &= \frac{1}{T}\sum_{t=1}^{\lfloor \lambda T \rfloor} \left(X_t X_t' - V\right) + \lambda V. \end{aligned}$$

We can show that this first term converges to zero uniformly. In particular, we have that

$$\frac{1}{\sqrt{T}}\frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor \lambda T \rfloor} \left(X_t X_t' - V\right),$$

where under some regularity conditions, $\frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor \lambda T \rfloor}\left(X_t X_t' - V\right) = O_p(1)$. Then, under these conditions, $\frac{1}{\sqrt{T}}\frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor \lambda T \rfloor}\left(X_t X_t' - V\right) \xrightarrow{p} 0$ uniformly. It then follows that

$$V_T(\lambda) - \lambda V \xrightarrow{p} 0.$$

Next, we make similarly high-level assumptions to show that

$$\hat{\Omega}(r/T) \xrightarrow{p} \lambda\Omega \text{ uniformly.}$$

Under these conditions, we have that

$$\begin{aligned}
&F_T(r/T) \xrightarrow{d} \left(\lambda^{-1}V^{-1}\xi(\lambda) - (1-\lambda)^{-1}V^{-1}(\xi(1)-\xi(\lambda))\right)' \cdot \\
&\left(\lambda^{-1}V^{-1}\Omega V^{-1} + (1-\lambda)^{-1}V^{-1}\Omega V^{-1}\right)^{-1} \cdot \\
&\left(\lambda^{-1}V^{-1}\xi(\lambda) - (1-\lambda)^{-1}V^{-1}(\xi(1)-\xi(\lambda))\right), \\
&= \left(\lambda^{-1}\xi(\lambda) - (1-\lambda)^{-1}(\xi(1)-\xi(\lambda))\right)' \cdot \left(\lambda^{-1}\Omega + (1-\lambda)^{-1}\Omega\right)^{-1} \cdot \left(\lambda^{-1}\xi(\lambda) - (1-\lambda)^{-1}(\xi(1)-\xi(\lambda))\right) \\
&= \left(\frac{1}{\lambda(1-\lambda)}\right)^{-1} \left(\lambda^{-1}\Omega^{1/2}\xi(\lambda) - (1-\lambda)^{-1}\Omega^{1/2}(\xi(1)-\xi(\lambda))\right)' \left(\lambda^{-1}\Omega^{1/2}\xi(\lambda) - (1-\lambda)^{-1}\Omega^{1/2}(\xi(1)-\xi(\lambda))\right) \\
&= \left(\frac{1}{\lambda(1-\lambda)}\right)^{-1} \left(\lambda^{-1}W(\lambda) - (1-\lambda)^{-1}(W(1)-W(\lambda))\right)' \left(\lambda^{-1}W(\lambda) - (1-\lambda)^{-1}(W(1)-W(\lambda))\right) \\
&= \frac{(W(\lambda)-\lambda W(1))'(W(\lambda)-\lambda W(1))}{\lambda(1-\lambda)}.
\end{aligned}$$

Define

$$B(\lambda) = W(\lambda) - \lambda W(1).$$

This is a **brownian bridge process**. We therefore showed that the F-statistic converges to the inner-product of a Brownian bridge

$$F_T(\cdot) \xrightarrow{d} F^*(\cdot) = \frac{B(\lambda)'B(\lambda)}{\lambda(1-\lambda)}.$$

The brownian bridge process is a chi-squared process, meaning that its finite-dimensional distribution is a chi-squared distribution. With this, we now have the limiting distribution of the sup-wald statistic. We have that

$$\sup_{r=1,\ldots,T} F_T(r/T) \xrightarrow{d} \sup_{\lambda\in[0,1]} F^*(\lambda).$$

We can use this to compute critical values via simulation.

4.4 Long-run trends

We now consider the problem of inference on long-run trends and projections. This is another application of the FLCT. For example, how do we characterize properties of a time series with a unit root?

First, assume that the time series Y_t is

$$\Delta Y_t = \mu + u_t,$$

where u_t is second-order stationary. This model implies that Y_t has a unit root and we say it is *order of integration* is one, $I(1)$. We observe the time series from $t = 1, \dots, T$ and we want to construct an h-step ahead prediction of Y_{T+h}. Our goal is to construct the predictive distribution of our forecast, $\hat{Y}_{T+h|T}$.

Begin by noticing that

$$\begin{aligned} Y_{T+h} - Y_T &= \sum_{t=T+1}^{T+h} \Delta Y_t, \\ &= \mu h + \sum_{t=T+1}^{T+h} u_t. \end{aligned}$$

When h is large and we want to make a forecast far into the future, it appears that our estimate of the mean μ will matter the most in our forecast (as it is being scaled by h). In this sense, we may be able to ignore the short-run dynamics of the innovations u_t. By the Wold Decomposition, we can write $u_t = C(L)\epsilon_t$. Therefore,

$$\begin{aligned} \sum_{t=T+1}^{T+h} u_t &= \sum_{t=T+1}^{T+h} C(L)\epsilon_t, \\ &\overset{(1)}{=} \sum_{t=T+1}^{T+h} \left(C(1)\epsilon_t + C^*(L)\Delta\epsilon_t \right), \\ &\overset{(2)}{=} C(1) \sum_{t=T+1}^{T+h} \epsilon_t + C^*(L)\epsilon_{T+h} - C^*(L)\epsilon_t, \end{aligned}$$

where (1) follows by the Beveridge-Nelson decomposition and (2) follows because the last sum telescopes. So, we see that

$$Y_{T+h} - Y_T = \mu h + C(1) \sum_{t=T+1}^{T+h} \epsilon_t + C^*(L)\epsilon_{T+h} - C^*(L)\epsilon_t.$$

The first term μh is the cumulation of the trend μ. The second term $C(1)\sum_{t=T+1}^{T+h} \epsilon_t$ is the cumulation of errors ϵ_t. It's variance is $O_p(\sqrt{h})$, and we will apply a CLT to this term. The final term $C^*(L)\epsilon_{T+h} - C^*(L)\epsilon_t$ is a remainder and it is $o_p(1)$. In other words, we have that

$$\frac{(Y_{T+h} - Y_T) - \mu h}{\sqrt{h}} = \frac{1}{\sqrt{h}} \sum_{t=T+1}^{T+h} C(1)\epsilon_t + o_p(1)$$

as h grows large. We will think of the sum as a function of h and apply a FLCT. To construct the variance of this asymptotically, we will need to construct the long run variance $\Omega = C^2(1)\sigma_\epsilon^2$. In other words, we only need to construct estimates of μ and Ω. We do not need to model the period-by-period dynamics in Y to do inference on the long-run trend.

Our estimate of the trend component μ is simply

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} \Delta Y_t = \mu + \frac{1}{T}\sum_{t=1}^{T} u_t$$

and therefore our h-step ahead prediction is

$$\hat{Y}_{T+h|T} = Y_T + \hat{\mu} h.$$

Substituting in, the error from this prediction is

$$\begin{aligned} Y_{T+h} - Y_{T+h|T} &= (\mu - \hat{\mu})h + \sum_{t=T+1}^{T+h} u_t \\ &= -hT^{-1}\sum_{t=1}^{T} u_t + \sum_{t=T+1}^{T+h} u_t. \end{aligned}$$

Assume that

$$\frac{h}{T} = \lambda.$$

Heuristically, think of λ as being about equal to 1.5, meaning that if we observed 100 years of data, we are making a 50 year ahead forecast. The fact that h scales with T makes precise the notion of a "long-run forecast." Then, we have that

$$\begin{aligned} \frac{Y_{T+\lfloor T\lambda \rfloor} - \hat{Y}_{T+\lfloor T\lambda \rfloor|T}}{\sqrt{T}} &= -\frac{h}{T}\sqrt{\frac{1}{T}}\sum_{t=1}^{T} u_t + \sqrt{\frac{1}{T}}\sum_{t=T+1}^{T+\lfloor T\lambda \rfloor} u_t \\ &= -\lambda\sqrt{\frac{1}{T}}\sum_{t=1}^{T} u_t + \sqrt{\frac{1}{T}}\sum_{t=T+1}^{T+\lfloor T\lambda \rfloor} u_t \\ &\xrightarrow{d} -\lambda\Omega^{1/2}W(1) + \Omega^{1/2}\left(W(1+\lambda) - W(1)\right), \end{aligned}$$

where applied the FCLT at the last step. For fixed λ, we have that this forecast error is distributed as $N(0, \Omega(\lambda^2 + \lambda))$. We can use this to construct prediction intervals.

5 Drifting Parameters and Local Asymptotic Power

Consider a simple gaussian location model

$$Y_t = \mu + u_t,$$

where $u_t \sim N(0, \sigma_u^2)$ i.i.d. We want to test the null and alternative

$$\begin{aligned} H_0 &: \mu = \mu_0 = 0, \\ H_1 &: \mu \neq \mu_0 = 0. \end{aligned}$$

Suppose that σ_u is known. Then, the distribution of the t-statistic $t = \frac{\bar{Y}}{\sigma_u/\sqrt{T}}$ is easy to compute. Our test is simply that we reject if the t-statistic squared is larger than some critical value, $t^2 > c$. Under the fixed alternative $\mu = \mu_1 \neq 0$, the probability of rejection is the **power** of the test. This is given by

$$\begin{aligned}
\mathbb{P}_{\mu_1}\left\{t^2 > c\right\} &= \mathbb{P}_{\mu_1}\left\{\left(\frac{\bar{Y}}{\sigma_u/\sqrt{T}}\right)^2 > c\right\} \\
&= \mathbb{P}_{\mu_1}\left\{\left(\frac{\bar{u}+\mu_1}{\sigma_u/\sqrt{T}}\right)^2 > c\right\} \\
&= \mathbb{P}_{\mu_1}\left\{\left(\frac{\sqrt{T}\bar{u}}{\sigma_u} + \frac{\sqrt{T}\mu_1}{\sigma_u}\right)^2 > c\right\} \\
&= \mathbb{P}_{\mu_1}\left\{\left(Z + \frac{\sqrt{T}\mu_1}{\sigma_u}\right)^2 > c\right\} = \mathbb{P}_{\mu_1}\left\{\chi^2_{1,T\frac{\mu_1^2}{\sigma_u^2}} > c\right\},
\end{aligned}$$

where $\chi^2_{1,T\frac{\mu_1^2}{\sigma_u^2}}$ is a non-central chi-square distribution with non-centrality parameter $\delta = T\frac{\mu_1^2}{\sigma_u^2}$.

We were able to analytically work out the power of this test under strong conditions – we relied on u_t being exactly normally distributed and i.i.d. But, this intuition will generalize because the key step was to replace a scaled sample average with a normally distributed random variable. Under general conditions, this step will be justified by a central limit theorem. So, the question is: In what sense does this result generalize and provide a good approximation for the behavior of tests in a wide variety of settings? That is, under what conditions does the following statement hold true for all possible values of μ_1

$$\mathbb{P}_{\mu_1}\left\{\left(\frac{\bar{Y}}{\sigma_u/\sqrt{T}}\right)^2 > c\right\} - \mathbb{P}_{\mu_1}\left\{\left(Z + \frac{\sqrt{T}\mu_1}{\sigma_u}\right)^2 > c\right\} \to 0$$

as $T \to \infty$?

For this to be true, we will need to ensure that the test has a non-trivial limit distribution. This is where we will introduce the idea of **drifting sequences**. In particular, notice from above that if $\sqrt{T}\mu_1$ diverges as $T \to \infty$, then the asymptotic approximation of our test will be that our test rejects with probability one. Clearly, this would not be a good approximation. To fix this problem, we consider an asymptotic approximation involving the **drifting sequence**

$$m = \frac{\sqrt{T}\mu_1}{\sigma_u} \implies \mu_1 = \mu_{1,T} = \frac{\sigma_u m}{\sqrt{T}} \quad m \text{ fixed}$$

This is also referred to as a **pitman drift**. That is, in our asymptotic approximation, we are imagining that we are testing a drifting sequence of alternatives that gets closer to the null hypothesis as $T \to \infty$.

Remark 5.1. *What's the intuition underlying this drifting sequence of parameters? In empirical applications, we are usually testing null hypotheses against alternatives that are "hard" to distinguish – meaning, we do not actually have a test with power equal to one in finite samples. In order to preserve this in our asymptotic*

approximation, we need to allow the alternative to also drift. Otherwise, for any fixed alternative, we would eventually reach a sample size that is large enough such that we would have power equal to one against it. The drifting sequence preserves the "hardness" of the problem that we are trying to solve.

Under this drifting sequence of alternatives, we want to show that

$$\sup_m \mathbb{P}_{\mu_{1,T}} \left\{ \left(\frac{\sqrt{T}\bar{Y}}{\hat{\sigma}_u} \right) > c \right\} - \mathbb{P}_m \left\{ (Z+m)^2 \right\} \to 0$$

as $T \to \infty$. Heuristically, we will need assumptions that generate (1) $\sqrt{T}\bar{u}$ is asymptotically normally distributed and (2) $\hat{\sigma}_u$ is uniformly consistent for σ_u. Then, under these assumptions, researchers argue that

$$\begin{aligned}
\mathbb{P}_{\mu_{1,T}} \left\{ \left(\frac{\sqrt{T}\bar{Y}}{\hat{\sigma}_u} \right) > c \right\} &= \mathbb{P}_{\mu_{1,T}} \left\{ \left(\frac{\sqrt{T}\bar{u} + \sqrt{T}\mu_{1,T}}{\sigma_u} \right)^2 \frac{\sigma_u^2}{\hat{\sigma}_u^2} > c \right\} \\
&= \mathbb{P}_{\mu_{1,T}} \left\{ \left(\frac{\sqrt{T}\bar{u}}{\sigma_u} + m \right)^2 \frac{\sigma_u^2}{\hat{\sigma}_u^2} > c \right\} \to \mathbb{P}_m \left\{ (Z+m)^2 < c \right\}
\end{aligned}$$

because $\frac{\sigma_u^2}{\hat{\sigma}_u^2} \xrightarrow{p} 1$ and $\frac{\sqrt{T}\bar{u}}{\sigma_u} \xrightarrow{d} N(0,1)$.

6 Weak Identification

6.1 Review of GMM

There is a moment condition

$$\mathbb{E}\left[h(Y_t, \theta_0) \otimes Z_t \right] = 0,$$

where $h(Y_t, \theta)$ is $G \times 1$ and Z_t is $m \times 1$. θ_0 is the unique value of the parameter that sets this moment equal to zero. Think of $h(Y_t, \theta)$ as the error term and Z_t as an instrument.

In the data, we consider the **sample moment function** with

$$\begin{aligned}
\phi_t(\theta) &= h(Y_t, \theta) \otimes Z_t, \\
\sqrt{1/T} \sum_{t=1}^{T} \phi_t(\theta) &= \sqrt{1/T} \sum_{t=1}^{T} h(Y_t, \theta) \otimes Z_t.
\end{aligned}$$

In the **just identified case**, $dim(\theta) = dim(Z_t)$. In the **over-identified case**, $dim(\theta) < dim(Z_t)$. Note that in the just-identified case, we can solve uniquely for the value $\hat{\theta}$ that sets the sample moment function equal to zero provided a rank condition is satisfied.

Example 6.1 (Linear IV)**.** *Consider the linear IV model*

$$Y_t = X_t'\beta + u_t,$$

where Z_t is an instrument. We assume that $\mathbb{E}\left[Z_t X_t\right] \neq 0$ and $\mathbb{E}\left[u_t Z_t\right] = 0$. Exogeneity gives us a moment

condition with $h(Y_t, \theta) = Y_t - X_t'\beta$ and so,

$$\mathbb{E}\left[(Y_t - X_t'\beta)Z_t\right] = 0.$$

When $dim(\beta) = dim(Z_t)$, we can solve this exactly and arrive at our IV estimand

$$\beta^{IV} = \frac{\mathbb{E}\left[Y_t Z_t\right]}{\mathbb{E}\left[X_t Z_t\right]}$$

and the IV estimator is just the sample analogue:

$$\hat{\beta}^{IV} = \frac{Z'Y}{Z'X}.$$

Example 6.2 (New-Keynesian Phillips Curve)**.** *Consider*

$$\pi_t = \gamma_F \mathbb{E}_t\left[\pi_{t+1}\right] + \gamma_b \pi_{t-1} + \lambda X_t + u_t.$$

The tricky part in estimating this is dealing with $\mathbb{E}_t\left[\pi_{t+1}\right]$. Why? We can re-write this as

$$\pi_t = \gamma_F \pi_{t+1} + \gamma_b \pi_{t-1} + \lambda X_t + \left(u_t - \gamma_f(\pi_{t+1} - \mathbb{E}_t\left[\pi_{t+1}\right])\right).$$

Clearly, the error is correlated with π_{t+1}. One approach in dealing with is to try to find an instrument.

In the over-identified case, the number of equations in the sample moment function exceeds the number of parameters and so, we can't solve the sample moment condition exactly. So, we then define our estimator in terms of the quadratic objective function

$$S_T(\theta) = \left[\sqrt{\frac{1}{T}}\sum_{t=1}^{T}\phi_t(\theta)\right]' W_T \left[\sqrt{\frac{1}{T}}\sum_{t=1}^{T}\phi_t(\theta)\right].$$

If W_T is positive semi-definite, then we can find a minimum. Our estimator is then

$$\hat{\theta}^{GMM} = \arg\min_{\theta} S_T(\theta).$$

Example 6.3 (Linear IV)**.** *We have that*

$$\sqrt{\frac{1}{T}}\sum_{t=1}^{T}\phi_t(\theta) = \sqrt{\frac{1}{T}}\sum_{t=1}^{T}(Y_t - X_t'\beta)Z_t$$
$$\sqrt{\frac{1}{T}}(Y - X\beta)'Z.$$

Let $W = (Z'Z)^{-1}$. Then,

$$\begin{aligned} S_T(\theta) &= \frac{1}{T}(Y - X\beta)'Z(Z'Z)^{-1}Z'(Y - X\beta) \\ &= \frac{1}{T}(\hat{Y} - \hat{X}\beta)'(\hat{Y} - \hat{X}\beta), \end{aligned}$$

where

$$\begin{aligned}\hat{X} &= Z(Z'Z)^{-1}Z'X = P_Z X \\ \hat{Y} &= P_Z Y.\end{aligned}$$

The minimizer of this is

$$\hat{\beta}^{2SLS} = (X'P_Z X)^{-1}(X'P_Z Y),$$

and so, 2SLS is equivalent to over-identified GMM with $W = (Z'Z)^{-1}$.

We'll now consider the asymptotic properties of GMM. To do so, we'll use techniques from **extremum estimation**. In particular, we will argue that $S_T(\theta) \xrightarrow{p} S^*(\theta)$ uniformly over $\sqrt{T}(\theta - \theta_0)$. That is, it converges uniformly over parameters that lie in a $1/\sqrt{T}$ neighborhood of $\theta_0 - \theta = \theta_0 + b/\sqrt{T}$. See Newey and McFadden (1994) for an in-depth treatment of the extremum estimator approach.

As notation, let

$$\begin{aligned}\phi_t(\theta) &= h(Y_t, \theta) \otimes Z_t \\ \mu(\theta) &= \mathbb{E}\left[\phi_t(\theta)\right] \\ R(\theta) &= \partial\mu/\partial\theta|_\theta \\ R(\theta_0) &= \partial\mu/\partial\theta|_{\theta=\theta_0}.\end{aligned}$$

We first state a result on the consistency of the GMM estimator.

Assumption 1 (Assumptions for consistency)**.** *Assume that*

i $\mu(\theta_0) = 0$ *uniquely at* θ_0 *and* $\mu(\theta)$ *is continuous in* θ.

ii $\frac{1}{T}\sum_{t=1}^T \phi_t(\theta) \xrightarrow{p} \mu(\theta)$ *uniformly in* θ. *As a scalar, this means*

$$\sup_{\theta\in\Theta} |\frac{1}{T}\sum_{t=1}^T \phi_t(\theta) - \mu(\theta)| \xrightarrow{p} 0.$$

iii $W_T(\theta) \xrightarrow{p} W(\theta)$ *uniformly, where* $W(\theta)$ *is positive semi-definite.*

Proposition 6.1 (Consistency of GMM estimator)**.** *Under the conditions of Assumption 1,*

$$\frac{1}{T}S_T(\theta) = \left[\frac{1}{T}\sum_{t=1}^T \phi_t(\theta)\right]' W_T \left[\frac{1}{T}\sum_{t=1}^T \phi_t(\theta)\right] \xrightarrow{p} \mu(\theta)'W(\theta)\mu(\theta)$$

and $\hat{\theta}^{GMM} \xrightarrow{p} \theta_0$.

We now state an asymptotic normality result.

Assumption 2 (Assumptions for Asymptotic Normality)**.** *Assume that*

a $\partial\mu/\partial\theta|_\theta = R(\theta)$ *is bounded and continuous.*

b $\sqrt{\frac{1}{T}}\sum_{t=1}^{T}(\phi_t(\theta) - \mu(\theta)) = \mathcal{V}_T(\theta) \xrightarrow{d} \mathcal{V}(\theta)$*, where* $\mathcal{V}(\theta)$ *is a Gaussian process and satisfies a stochastic lipschitz condition*

$$|\mathcal{V}(\theta) - \mathcal{V}(\theta')| \leq K(\theta - \theta'),$$

almost surely where $K = O_p(1)$ *uniformly over* Θ*.*

c $W_T(\theta) \xrightarrow{p} W(\theta)$ *uniformly in* θ*.*

Proposition 6.2 (Asymptotic Normality of GMM estimator)**.** *Under the conditions in Assumption 1 and Assumption 2, the GMM estimator satisfies*

$$\sqrt{T}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N(0, \Sigma(W)),$$

where

$$\begin{aligned}
\Sigma(W) &= (R'WR)^{-1}\left(R'W\Omega WR\right)(R'WR)^{-1}, \\
R &= R(\theta_0), \\
W &= W(\theta_0).
\end{aligned}$$

Proof. We provide a sketch of the argument. We focus on a $1/\sqrt{T}$ neighbhorhood of θ_0. That is, we consider values of $\theta = \theta_0 + b/\sqrt{T}$.

Consider

$$\begin{aligned}
\sqrt{1/T}\sum_{t=1}^{T}\phi_t(\theta) &= \sqrt{1/T}\sum_{t=1}^{T}(\phi_t(\theta) - \mu(\theta)) + \sqrt{T}\mu(\theta) \\
&= \mathcal{V} + \sqrt{T}\mu(\theta).
\end{aligned}$$

Substituting in $\theta = \theta_0 + b/\sqrt{T}$, we have that

$$\begin{aligned}
\sqrt{1/T}\sum_{t=1}^{T}\phi_t(\theta) &= \mathcal{V}(\theta_0 + \frac{b}{\sqrt{T}}) + \sqrt{T}\mu(\theta_0 + \frac{b}{\sqrt{T}}), \\
&= \mathcal{V}(\theta_0 + \frac{b}{\sqrt{T}}) + \sqrt{T}\mu(\theta_0) + b\mu'(\tilde{\theta}), \\
&= \mathcal{V}(\theta_0 + \frac{b}{\sqrt{T}}) + b\mu'(\tilde{\theta}), \\
&= \mathcal{V}(\theta_0) + \left(\mathcal{V}(\theta_0 + \frac{b}{\sqrt{T}}) - \mathcal{V}(\theta_0)\right) + b\mu'(\tilde{\theta}),
\end{aligned}$$

where $\tilde{\theta} \in (\theta_0, \theta_0 + b/\sqrt{T})$ and $\mathcal{V}(\theta_0 + \frac{b}{\sqrt{T}}) - \mathcal{V}(\theta_0) = O_p(\frac{1}{\sqrt{T}})$ via the stochastic lipschitz condition.

Therefore, as T grows large, we have that

$$\sqrt{1/T}\sum_{t=1}^{T}\phi_t(\theta) \xrightarrow{d} \mathcal{V}(\theta_0) + bR$$

for $\theta = \theta_0 + \frac{b}{\sqrt{T}}$. Moreover, we have that $W(\theta) \xrightarrow{p} W(\theta_0) = W$ for these values of θ as well. Putting these together, we have that

$$S_T(\theta) \xrightarrow{d} (\nu(\theta) + Rb)' W (\nu(\theta_0) + Rb)' = S^*(b).$$

This is a quadratic form (we used the Taylor expansion to expand away the non-linearity of θ in ϕ). So, we now have that

$$\max_b S^*(b) \implies R'W(\nu + Rb^*) = 0$$

and solving the FOC gives that

$$b^* = (R'WR)^{-1}R'W\nu,$$

where $\nu(\theta_0) \sim N(0, 2\pi S_{\phi(\theta_0)}(0))$. So, it follows that

$$\hat{b} = \sqrt{T}(\hat{\theta}^{GMM} - \theta_0) \xrightarrow{d} N(0, \Sigma),$$

□

Then, recall that we can derive the efficient GMM estimator which sets $W = \Omega^{-1}$.

6.2 Feasible efficient GMM

Suppose we have that $\hat{\Omega}(\theta) \xrightarrow{p} \Omega(\theta)$ and suppose that this convergence is uniform in θ. Then, locally, we have that $\hat{\Omega} \xrightarrow{p} \Omega(\theta_0)$ where $\hat{\Omega} = \Omega(\hat{\theta})$.

There are two approaches that we want to focus in on. The first is **two-step GMM**. In step 1, we construct a preliminary estimate with $W_T = I$ and this gives $\hat{\theta^{(1)}}$. In step 2, we plug-in our first step estimate and construct $\hat{\Omega}(\hat{\theta}^{(1)}) = 2\pi S_{\phi(\hat{\theta}^{(1)})}(0)$. Then, we use

$$S_T^{(2)}(\theta) = \left[\sqrt{\frac{1}{T}}\sum_{t=1}^{T}\phi_t(\theta)\right]' \hat{\Omega}^{-1}(\hat{\theta}^{(1)}) \left[\sqrt{\frac{1}{T}}\sum_{t=1}^{T}\phi_t(\theta)\right].$$

What does two-step GMM look like in the linear IV case? Recall that

$$q_t(\theta) = (Y_t - X_t\beta)Z_t$$
$$q_t(\theta_0) = u_t Z_t.$$

Assume that $\mathbb{E}[X_t u_t X_s u_s] = 0$ for $t \neq s$ and that the errors are homoskedastic. Then, we have that

$$\Omega = \mathbb{E}\left[u_t Z_t u_t Z_t'\right] = \sigma_u^2 \Sigma_{ZZ}.$$

An efficient and feasible 2-step estimator of $\hat{\sigma}_u^2$ will just be a constant that multiplies the objective function. It won't affect the minimizer and so, we can skip the first step and use $\hat{W}_T = Z'Z/T$. We have that

$$S_T(\theta) = \frac{1}{T}(Y - X\beta)' Z \frac{Z'Z}{T} Z'(Y - X\beta).$$

Therefore, $\hat{\beta}^{2step} = \hat{\beta}^{2SLS}$.

The second is **continuously updating GMM**. We have that

$$S_T^{CUE}(\theta) = \left[\sqrt{\frac{1}{T}}\sum_{t=1}^{T}\phi_t(\theta)\right]' \hat{\Omega}^{-1}(\theta)\left[\sqrt{\frac{1}{T}}\sum_{t=1}^{T}\phi_t(\theta)\right].$$

Here every θ is the same and we directly minimize this over θ in one shot. To understand what's going on, we again will consider the linear IV model with homoskedastic, serially uncorrelated errors. We have that

$$S_T^{CUE}(\theta) = \frac{(Y - X\beta)' Z (Z'Z)^{-1} Z'(Y - X\beta)}{\hat{\sigma}^2},$$

where $\hat{\sigma}^2 = \frac{(Y-X\beta)'(Y-X\beta)}{T}$. We'll rewrite this as

$$\begin{aligned} S_T^{CUE}(\beta) &= \frac{u(\beta)' P_Z u(\beta)}{u(\beta)' u(\beta)} \\ &= \frac{u(\beta)' P_Z u(\beta)}{u(\beta)'(I - P_Z + P_Z) u(\beta)} \\ &= \frac{u(\beta)' P_Z u(\beta)}{u(\beta)' M_Z u(\beta) + u(\beta)' P_Z u(\beta)}, \end{aligned}$$

where $M_Z = I - P_Z$, $P_Z = Z(Z'Z)^{-1}Z'$. So, we have that

$$S_T^{CUE}(\beta) = \frac{u' P_Z u / u' M_Z u}{1 + u' P_Z u / u' M_Z u}.$$

Minimizing this is equivalent to minimizing $u' P_Z u / u' M_Z u$ as S_T^{CUE} is monotone transformation of this object. Therefore, we have that

$$\min_\beta S_T^{CUE}(\beta) \implies \min_\beta \frac{(Y - X\beta)' P_Z (Y - X\beta)}{(Y - X\beta)' M_Z (Y - X\beta)/(T - k)}.$$

The RHS is known as the **Anderson-Rubin statistic**. How do we interpret this? Suppose we regress $Y - X\beta_0 = \gamma Z + v_t$. The AR statistic is the homoskedastic F-statistic for testing $\gamma = 0$. In other words, continuously updating GMM is finding the value of β that minimizes the correlation between $Y - X\beta$

and Z_t. This generalizes the non-linear moment condition as well. It turns out that you can show that CUE GMM is equivalent to LIML in the linear IV case.

6.3 J statistic

Suppose you compute the AR statistic and it rejects. It could reject for two reasons: (1) Your moment condition is wrong, (2) You selected the wrong value of θ. How do we tell which has occurred? Consider

$$\phi_t(\hat{\hat{\theta}}),$$

where $\hat{\hat{\theta}}$ is the efficient GMM estimator. Then, look at the value of the quadratic objective at this efficient GMM estimator

$$\left[\sqrt{\frac{1}{T}}\sum_{t=1}^{T}\phi_t(\hat{\hat{\theta}})\right]' \hat{\Omega}^{-1}(\hat{\hat{\theta}}) \left[\sqrt{\frac{1}{T}}\sum_{t=1}^{T}\phi_t(\hat{\hat{\theta}})\right].$$

We have the following result. Under the null that $\mathbb{E}\left[\phi_t(\theta_0)\right] = 0$, then

$$S_T(\hat{\hat{\theta}}) \xrightarrow{d} \chi^2_{df},$$

where $df = dim(\phi) - dim(\theta)$. This tests the alternative that $\mathbb{E}\left[\phi_t(\theta_0)\right] = \mu(\theta_0) \neq 0$. For fixed $\mu(\theta)$, this will reject with probability converging to 1 for a fixed alternative.

6.4 Weak Identification: Building intuition with linear IV

Consider the reduced-form linear IV model

$$y_t = \gamma z_t + w_t,$$
$$x_t = \pi z_t + v_t,$$

where the structural equation of interest is

$$y_t = \beta x_t + u_t.$$

The reduced-form parameters γ, π are related to the structural parameter β by

$$\beta = \frac{\gamma}{\pi} = \frac{\mathbb{E}\left[yz\right]}{\mathbb{E}\left[xz\right]}.$$

The linear IV estimator is simply the ratio of the OLS estimates of the reduced-form parameters

$$\hat{\beta}^{IV} = \frac{\hat{\gamma}}{\hat{\pi}}.$$

To fix ideas, consider the finite-sample gaussian case in which T is fixed,

$$\begin{pmatrix} w_t \\ v_t \end{pmatrix} \sim N(0, V)$$

and z_t is fixed (i.e., we condition on the realizations of the instrument). In this case,

$$\begin{pmatrix} \hat{\gamma} \\ \hat{\pi} \end{pmatrix} \sim N(\begin{pmatrix} \gamma \\ \pi \end{pmatrix}, \Sigma).$$

Note that

$$\begin{aligned} \hat{\gamma} - \beta\hat{\pi} &= \frac{\sum_{t=1}^{T} y_t z_t - \beta \sum_{t=1}^{T} x_t z_t}{\sum_{t=1}^{T} z_t^2} \\ &= \frac{\sum_{t=1}^{T} u_t z_t}{\sum_{t=1}^{T} z_t^2}. \end{aligned}$$

Therefore,

$$\begin{aligned} \hat{\beta}^{IV} - \beta &= \frac{\hat{\gamma}}{\hat{\pi}} - \frac{\hat{\pi}\beta}{\hat{\pi}} \\ &= \frac{\hat{\gamma} - \hat{\pi}\beta}{\hat{\pi}} \end{aligned}$$

Notice that the numerator of this expression will be well-behaved. In particular,

$$\hat{\gamma} - \beta\hat{\pi} \sim N(0, \begin{pmatrix} 1 & -\beta \end{pmatrix} \Sigma \begin{pmatrix} 1 \\ -\beta \end{pmatrix}).$$

The problem is that it is being divided by another random variable $\hat{\pi}$. In other words, we can re-express the difference between the IV estimator and the structural parameter as

$$\hat{\beta}^{IV} - \beta = \frac{\hat{\gamma} - \hat{\pi}\beta}{\pi} \frac{1}{\hat{\pi}/\pi} = \frac{\hat{\gamma} - \hat{\pi}\beta}{\pi} \frac{1}{1 + \frac{\hat{\pi} - \pi}{\pi}},$$

where we can apply our usual CLT arguments to the first term but the second term is a random ratio. Moreover, the denominator of the ratio depends on the magnitude of the sampling variation in $\hat{\pi}$ relative to the magnitude of $\hat{\pi}$, $\frac{\hat{\pi} - \pi}{\pi}$. When the sampling variation in $\hat{\pi}$ of the same order of magnitude as π, then this ratio should treated as random *even asymptotically*.

In other words, we can think of the "strong instruments" assumption as assuming that this ratio converges in probability to one as the sample size grows large

$$\frac{1}{1 + \frac{\hat{\pi} - \pi}{\pi}} \xrightarrow{p} 1.$$

In this case, our usual asymptotic approximation for the distribution of the IV estimator will work

well. In the unidentified case with $\pi = 0$, we have that

$$\begin{aligned}
\hat{\beta}^{IV} - \beta &= \frac{\hat{\gamma} - \hat{\pi}\beta}{\hat{\pi}} \\
&= \frac{\sqrt{\frac{1}{T}}\sum_{t=1}^{T} u_t z_t}{\sqrt{\frac{1}{T}}\sum_{t=1}^{T} x_t z_t} \\
&= \frac{\sqrt{\frac{1}{T}}\sum_{t=1}^{T} u_t z_t}{\sqrt{\frac{1}{T}}\sum_{t=1}^{T} v_t z_t} \\
&\sim \frac{N(0, V_1)}{N(0, V_2)},
\end{aligned}$$

which is the ratio of correlated normal distributions. When $u \perp\!\!\!\perp v$, this ratio is actually Cauchy distributed.

So far, we heuristically argued that what will matter in the finite sample normal model is the distance of π from zero (i.e., whether the sampling variation of $\hat{\pi}$ is of the same order of magnitude as π). The key parameter that will govern this is $\pi^2/\Sigma_{\pi\pi}$, which is known as the *concentration parameter*. The concentration parameter can be thought as a population first-stage F-statistic for testing whether π is different from zero. The key question is: how do we extend this intuition beyond the finite-sample normal model? To do so, we need to think about a limiting experiment in which this intuition is preserved. The key step in constructing such a limiting experiment will be to consider a drifting sequence of parameter values

$$\pi_T = \frac{C}{\sqrt{T}}.$$

Where does this come from? As T grows large, the sampling variation in $\hat{\pi}$ is of the order $\frac{1}{\sqrt{T}}$. So allowing π to drift to zero at a $1/\sqrt{T}$ rate formalizes the idea that π is of the same order of magnitude as the sampling variation of $\hat{\pi}$. Under this nesting, we have that

$$\begin{aligned}
\frac{1}{1 + \frac{\hat{\pi} - \pi_T}{\pi_T}} &= \frac{1}{1 + \frac{\sqrt{T}(\hat{\pi} - \pi_T)}{\sqrt{T}\pi_T}} \\
&= \frac{1}{1 + \frac{\sqrt{T}(\hat{\pi} - \pi_T)}{C}},
\end{aligned}$$

and then apply our asymptotic approximations.

This heuristic derivation of weak identification in the linear IV model heavily relied on the linear structure. The linear structure enabled us to construct an explicit expression for the IV estimator and we worked directly with this expression. In the next section, we generalize this intuition to the general, non-linear GMM case. As we will see, we once again will introduce a drifting parameter sequence that will capture the key features of weak identification.s

6.5 GMM with Weak Identification

There is ample evidence that the asymptotic approximations for GMM that we derived in Section 6.1 are very poor in finite samples. Why? The key problem is that the mean vector $\mu(\theta) = \mathbb{E}[\phi_t(\theta)]$ is not far from zero when $\theta \neq \theta_0$ and this implies that the objective function is not locally quadratic. To formalize this idea, we will now think of $\sqrt{T}\mu(\theta) \to m(\theta)$, meaning that we consider a sequence of approximations that involve a drifting sequence of the mean vector. Notice that as T gets large, the mean vector $\mu(\theta) = \frac{m(\theta)}{\sqrt{T}}$ goes to zero for all values of θ. Under this sequence, we will show that

$$S_T(\theta) \xrightarrow{S}{}^{*} (\theta) = [\mathcal{V}(\theta) + m(\theta)]^{-1} W(\theta) [\mathcal{V}(\theta) + m(\theta)].$$

Recall our notation from Section 6.1. We defined

$$\mathcal{V}_T(\theta) = \sqrt{1/T} \sum_{t=1}^{T} (\phi_t(\theta) - \mathbb{E}[\phi_t(\theta)])$$
$$\mu(\theta) = \mathbb{E}[\phi_t(\theta)].$$

Additionally, recall our assumptions for asymptotic normality.

a) $\partial\mu/\partial\theta|_\theta = R(\theta)$ is bounded and continuous.

b) $\sqrt{\frac{1}{T}} \sum_{t=1}^{T} (\phi_t(\theta) - \mu(\theta)) = \mathcal{V}_T(\theta) \xrightarrow{d} \mathcal{V}(\theta)$, where $\mathcal{V}(\theta)$ is a Gaussian process and satisfies a stochastic lipschitz condition

$$|\mathcal{V}(\theta) - \mathcal{V}(\theta')| \leq K(\theta - \theta'),$$

almost surely where $K = O_p(1)$ uniformly over Θ.

c) $W_T(\theta) \xrightarrow{p} W(\theta)$ uniformly in θ.

We modify these assumptions. In particular, we modify assumption (i) as that imposes strong identification. We'll now assume

a') $\sqrt{T}\mu(\theta) = m(\theta)$ uniformly in θ with $m(\theta_0) = 0$ but may not be the unique root

What does this change? Under the assumption of strong identification, we now have that

$$\begin{aligned}\sqrt{T}\mu(\theta) &= \sqrt{T}\mu(\theta_0 + b/\sqrt{T}) \\ &= \sqrt{T}\left(\mu(\theta_0) + R(\tilde{\theta})b/\sqrt{T}\right) \\ &= R(\tilde{\theta})b \xrightarrow{p} Rb = 0 \cdot b = 0.\end{aligned}$$

This was the strong identification assumption. Under (a'), we now have

$$\begin{aligned}S_T(\theta) &= \left[\mathcal{V}_T(\theta) + \sqrt{T}\mu(\theta)\right]' W_T(\theta) \left[\mathcal{V}_T(\theta) + \sqrt{T}\mu(\theta)\right] \\ &\xrightarrow{d} [\mathcal{V}(\theta) + m(\theta)] W(\theta) [\mathcal{V}(\theta) + m(\theta)].\end{aligned}$$

The efficient GMM estimator sets $W(\theta) = \Omega(\theta)^{-1}$. Therefore, the efficient GMM objective function converges to

$$S_T^{\text{efficient}}(\theta) \xrightarrow{d} [\mathcal{V}(\theta) + m(\theta)]\, \Omega(\theta)\, [\mathcal{V}(\theta) + m(\theta)] \equiv S^*(\theta),$$

which is a non-central χ^2 process. Depending on the true value of the function $m(\theta)$, the non-centrality parameter may be quite complex and so, we can't get a closed form expression for $\theta^* = \arg\min_\theta S^*(\theta)$ in general. In the case of linear IV, $m(\theta)$ is linear in θ, and so we can go further.

We now make two remarks about this result.

Remark 6.1 (The unidentified case)**.** *In the unidentified case, $\mu(\theta) = 0$ for all θ. Then, the limiting distribution of the GMM objective function is*

$$S^*(\theta)\mathcal{V}(\theta)'W(\theta)\mathcal{V}(\theta).$$

For efficient GMM, this is given by

$$S^*(\theta)\mathcal{V}(\theta)'\Omega(\theta)^{-1}\mathcal{V}(\theta),$$

which is a $\chi^2_{dim(\phi)}$ process.

Remark 6.2 (Anderson-Rubin confidence intervals)**.** *The Anderson-Rubin identification robust confidence interval is an important tool in this literature. At the true θ_0, the continuous updating GMM objective converges to*

$$S_T^{CUE}(\theta_0) \xrightarrow{d} S^*(\theta_0) = \mathcal{V}'(\theta_0)\Omega^{-1}(\theta_0)\mathcal{V}(\theta_0) \sim \chi^2_{dim(\phi)}.$$

We can use this to construct a confidence interval via test inversion. Under the null hypothesis that $\mu(\theta) = 0$,

$$S_T^{CUE}(\theta) \xrightarrow{d} S^*(\theta) = \mathcal{V}'(\theta)\Omega^{-1}(\theta)\mathcal{V}(\theta) \sim \chi^2_{dim(\phi)}.$$

Therefore, an asymptotically valid test for this null hypothesis compares the value of continuous updating GMM objective function against a critical value based upon the $\chi^2_{dim(\phi)}$ distribution. By test-inversion, we can construct a confidence interval for θ_0.

Notice that this confidence interval is "identification-robust" as it did not require us to assume that there exists a unique θ_0 that satisfies the moment condition.

6.5.1 Application of linear IV

We now illustrate this result in linear IV. Consider

$$\begin{aligned} y_t &= x_t\beta_0 + u_t, \\ x_t &= z_t'\pi + v_t \end{aligned}$$

where z_t is an instrument for x_t. The moment is $\phi_t(\theta) = z_t(y_t - x_t\beta)$, and so

$$\begin{aligned}\sqrt{\frac{1}{T}}\sum_{t=1}^{T}\phi_t(\theta) &= \sqrt{\frac{1}{T}}\sum_{t=1}^{T} z_t(y_t - x_t\beta) \\ &= \sqrt{\frac{1}{T}}\sum_{t=1}^{T} z_t(u_t - x_t(\beta - \beta_0)) \\ &= \sqrt{\frac{1}{T}}\sum_{t=1}^{T} z_t u_t - \sqrt{\frac{1}{T}}\sum_{t=1}^{T} z_t x_t(\beta - \beta_0),\end{aligned}$$

where we define $\beta - \beta_0 = \theta$. Now, we add and subtract $\mathbb{E}\left[z_t x_t\right]\theta$ to get

$$\sqrt{\frac{1}{T}}\sum_{t=1}^{T}\phi_t(\theta) = \sqrt{\frac{1}{T}}\sum_{t=1}^{T} z_t u_t - \sqrt{\frac{1}{T}}\sum_{t=1}^{T}(z_t x_t - \mathbb{E}\left[z_t x_t\right])\theta - \sqrt{T}\mathbb{E}\left[z_t x_t\right]\theta.$$

Our local-to-zero assumption is that

$$\sqrt{T}\mathbb{E}\left[z_t x_t\right]\theta = m(\theta) \implies \sqrt{T}\mathbb{E}\left[z_t x_t\right] = C.$$

Next, we write

$$\begin{aligned}\sqrt{\frac{1}{T}}\sum_{t=1}^{T}\left(z_t x_t - \mathbb{E}\left[z_t x_t\right]\right) &= \sqrt{\frac{1}{T}}\sum_{t=1}^{T}\left(z_t(z_t'\pi + v_t) - \mathbb{E}\left[z_t(z_t'\pi + v_t)\right]\right) \\ &= \frac{1}{T}\sum_{t=1}^{T}\left(z_t z_t' - \Sigma_{ZZ}\right)\sqrt{T}\pi + \sqrt{\frac{1}{T}}\sum_{t=1}^{T} z_t v_t.\end{aligned}$$

By the local-to-zero assumption, notice that $\sqrt{T}\pi = \Sigma_{ZZ}^{-1}C$ and $\frac{1}{T}\sum_{t=1}^{T}\left(z_t z_t' - \sigma_{ZZ}\right)$ by a LLN. Therefore, we conclude that

$$\sqrt{\frac{1}{T}}\sum_{t=1}^{T}\left(z_t x_t - \mathbb{E}\left[z_t x_t\right]\right) = \sqrt{\frac{1}{T}}\sum_{t=1}^{T} z_t v_t + o_p(1).$$

Define

$$\begin{pmatrix}\xi_{1T} \\ \xi_{2T}\end{pmatrix} = \begin{pmatrix}\sqrt{1/T}\sum_t z_t u_t \\ \sqrt{1/T}\sum_t z_t v_t\end{pmatrix}$$

and write

$$\sqrt{\frac{1}{T}}\sum_{t=1}^{T}\phi_t(\theta) = \xi_{1T} - \xi_{2T}\theta - C\theta + o_p(1).$$

Under usual regularity conditions,

$$\begin{pmatrix}\xi_{1T} \\ \xi_{2T}\end{pmatrix} = \begin{pmatrix}\sqrt{1/T}\sum_t z_t u_t \\ \sqrt{1/T}\sum_t z_t v_t\end{pmatrix} \xrightarrow{d} \begin{pmatrix}\xi_1^* \\ \xi_2^*\end{pmatrix} \sim N(0, \Sigma_{\xi^*})$$

and we would like to apply a FCLT to the function in order to conclude that

$$\mathcal{V}_T(\theta) = \xi_{1T} - \xi_{2T}\theta \xrightarrow{d} \mathcal{V}(\theta) = \xi_1^* - \xi_2^*\theta.$$

To do so, we first show that the finite dimensional distributions of $\mathcal{V}_T(\theta)$ converge. This is straightforward. Fix $\theta_1, \ldots, \theta_k$ for some scalar k and consider the vector

$$\begin{pmatrix} \mathcal{V}_T(\theta_1) \\ \vdots \\ \mathcal{V}_T(\theta_k) \end{pmatrix} = \begin{pmatrix} 1 & -\theta_1 \\ \vdots & \vdots \\ 1 & -\theta_k \end{pmatrix} \begin{pmatrix} \xi_{1T} \\ \xi_{2T} \end{pmatrix},$$

which is asymptotically normal as the vector ξ_T obeys a central limit theorem. Next, we need to establish tightness. That is, for each $\epsilon > 0$, we wish to show that

$$\mathbb{P}\left\{ \sup_{|\theta_1 - \theta_2| < \delta} |\mathcal{V}_T(\theta_1) - \mathcal{V}_T(\theta_2)| > \epsilon \right\} \to 0 \text{ as } \delta \to 0.$$

This is similarly straightforward. Substituting in for $\mathcal{V}_T(\theta)$, we see that

$$\begin{aligned} \mathbb{P}\left\{ \sup_{|\theta_1 - \theta_2| < \delta} \|\mathcal{V}_T(\theta_1) - \mathcal{V}_T(\theta_2)\| > \epsilon \right\} &= \mathbb{P}\left\{ \sup_{|\theta_1 - \theta_2| < \delta} \| - \xi_{2T}(\theta_1 - \theta_2)\| > \epsilon \right\} \\ &= \mathbb{P}\left\{ \sup_{|\theta_1 - \theta_2| < \delta} \|\xi_{2T}\| |\theta_1 - \theta_2| > \epsilon \right\} \\ &= \mathbb{P}\left\{ \|\xi_{2T}\| > \epsilon/\delta \right\} \leq \frac{\mathbb{E}\left[\|\xi_{2T}\|^2\right]}{\epsilon^2/\delta^2} \xrightarrow{0} \end{aligned}$$

provided that $\mathbb{E}\left[\|\xi_{2T}\|^2\right]$ is finite. Tightness in this example because the stochastic process $\mathcal{V}_T(\theta)$ is linear in the parameter θ. In fact, in the case where $\mathcal{V}_T(\theta)$ is a scalar (one instrument), it is just a line with a random intercept and random slope. Therefore, we can apply an FCLT to $\mathcal{V}_T(\theta)$ and we have that

$$\mathcal{V}_T(\theta) \xrightarrow{d} \xi_1^* - \xi_2^*\theta.$$

Applying the results we derived on weak identification for the general GMM estimator, we arrive at the linear IV objective converges to

$$S_T(\theta) \xrightarrow{d} (\xi_1^* - \xi_2^*\theta - C\theta)' W(\theta) (\xi_1^* - \xi_2^*\theta - C\theta) = S^*(\theta).$$

Therefore, $\hat{\theta} \xrightarrow{d} \theta^* = \arg\min_\theta S^*(\theta)$ and we can derive this in closed form.

Suppose we additionally assume that the error is homoskedastic, meaning that $\Omega = \Sigma_{ZZ}\sigma_u^2$. Then, the two-step GMM objective function (which is just the 2SLS objective) will converge to

$$S_T^{\text{2-step}}(\theta) \xrightarrow{d} (\xi_1^* - \xi_2^*\theta - C\theta)' \Sigma_{ZZ}^{-1} (\xi_1^* - \xi_2^*\theta - C\theta).$$

In this special case of 2SLS, the limiting distribution of the 2SLS estimator is equivalent to

$$\begin{aligned}\theta^* &= \arg\min \left(\xi_1^* - \xi_2^*\theta - C\theta\right)' \Sigma_{ZZ}^{-1} \left(\xi_1^* - \xi_2^*\theta - C\theta\right) \\ &= \frac{(C+\xi_2^*)'\Sigma_{ZZ}^{-1}\xi_1^*}{(C+\xi_2^*)'\Sigma_{ZZ}^{-1}(C+\xi_2^*)}\end{aligned}$$

Moreover, we can rewrite this as

$$\theta^* = \frac{(\lambda + Z_v)'Z_u}{(\lambda+Z_v)'(\lambda+Z_v)},$$

where

$$\begin{aligned}&\lambda = \Sigma_{ZZ}^{-1/2}C, \\ &\begin{pmatrix}\Sigma_{ZZ}^{-1/2}\xi_1^* \\ \Sigma_{ZZ}^{-1/2}\xi_2^*\end{pmatrix} = \begin{pmatrix}Z_u \\ Z_v\end{pmatrix} \sim N(0, I \otimes \begin{pmatrix}\sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2\end{pmatrix}.\end{aligned}$$

The denominator is a non-central χ^2 distribution with non-centrality parameter $\lambda'\lambda$. When $\lambda'\lambda$ is large, the randomness in the denominator is effectively negligible and we are in the strong instruments case. When it is small, the randomness will matter a lot and we are in the weak instruments case. Therefore, the parameter λ governs the strength of identification in the linear IV setting and it is referred to as the *concentration parameter*.

We can further re-write this expression. Since Z_u, Z_v are jointly normal, define the projection of Z_u onto Z_v as

$$Z_u = \delta Z_v + \eta, \quad \eta \perp\!\!\!\perp Z_v.$$

Substituting this in, rewrite the expression for θ^* as

$$\theta^* = \delta\frac{(\lambda+Z_v)'Z_v}{(\lambda+Z_v)'(\lambda+Z_v)} + \frac{(\lambda+Z_v)'\eta}{(\lambda+Z_v)'(\lambda+Z_v)}.$$

Conditional on Z_v, this is just a normal distribution with mean zero and variance equal to $\frac{\sigma_\eta^2}{(\lambda+Z_v)'(\lambda+Z_v)}$. Unconditional on Z_v, this is a mixture of normal distributions

$$\theta^* = \int N\left(\delta\frac{(\lambda+Z_v)'Z_v}{(\lambda+Z_v)'(\lambda+Z_v)}, \frac{\sigma_\eta^2}{(\lambda+Z_v)'(\lambda+Z_v)}dF_{Z_v}\right).$$

Why is this useful? In the case where the instruments are fully irrelevant and $\pi = C = \lambda = 0$, this expressions provides a lot of intuition. In particular, in this case, the projection coefficient δ simplifies to just equal $\frac{\sigma_{uv}}{\sigma_v^2} = \frac{\sigma_{ux}}{\sigma_x^2}$, which is just the omitted variables bias term of OLS. Therefore, δ equals the probability limit of $\hat{\beta}^{OLS} - \beta_0$. Moreover, this implies that the 2SLS estimator is centered at the omitted

variables bias term with additional noise

$$\theta^* = \int N\left(\delta, \frac{\sigma_\eta^2}{Z_v' Z_v} dF_{Z_v}\right).$$

7 Filtering

Suppose the observed time series y_t is driven by some latent process. What can we learn about the latent process?

Example 7.1. *Suppose that*

$$y_t = \mu_t + \sigma_\epsilon \epsilon_t,$$
$$\mu_t = \mu_{t-1} + \sigma_\eta \eta_t.$$

*This is a **local drift problem**. We may wish to estimate and learn about the latent process μ_t.*

Example 7.2. *Suppose that*

$$\pi_t = \lambda(u_t - u_t^*) + \beta \pi_t^e + \epsilon_t,$$

where π_t is inflation, u_t is unemployment and u_t^* is the NAIRU. This is a simple model of the Phillip's curve. Suppose that

$$u_t^* = u_{t-1}^* + \eta_t,$$

meaning that we model NAIRU as a random walk. We wish to understand the dynamics of the latent process u_t^*.

Example 7.3. *Suppose that*

$$\underset{n\times 1}{y_t} = \underset{n\times r}{\lambda(L)} \underset{r\times 1}{f_t} + e_t, \quad \mathbb{E}\left[e_t e_t'\right] = \sigma_\epsilon^2 I$$

where

$$A(L) f_t = \eta_t.$$

The factors f_t are unobserved and they drive all of the observed comovements in the data. We wish to learn about the factors f_t.

7.1 General filtering problem

Let s_t denote the latent state vector, y_t is the vector of observables at time t and let $Y_t = (y_t, \ldots, y_1)$. Our goals are:

1. Estimate some parameters θ that govern the observed process,
2. Estimate the distribution $s_t | Y_t$. This is known as the **filtering problem**.

3. Estimate the distribution $s_t|Y_T$. This is known as the **smoothing problem**.

There are three key components of the **latent variable model**:

- **State density**: $f(s_t|s_{t-1}, Y_{t-1})$. This is the density of the state at time t given the history of the latent states and observed outcomes.
- **Measurement density**: $f(y_t|s_t, Y_{t-1})$. This is the density given the current state and past data.
- **Likelihood**: This is the likelihood of the observed data

$$f(Y_T; \theta) = \left(\Pi_{t=2}^{T} f(y_t|Y_{t-1}, \theta)\right) f(y_1, \theta).$$

From these objects, we derive the following equations:

- **Prediction equation**: This gives the density of the current state given the past observed data

$$f(s_t|Y_{t-1}) = \int f(s_t|s_{t-1}, Y_t) f(s_{t-1}|Y_{t-1})\, ds_{t-1}.$$

- **Likelihood equation**: This gives the density of the current observable given past observed data

$$f(y_t|Y_{t-1}) = \int f(y_t|s_t, Y_{t-1}) f(s_t|Y_{t-1})\, ds_t.$$

- **Updating equation**: This gives the density of the current state given current and past observed data

$$f(s_t|Y_t) = \frac{f(y_t|s_t, Y_{t-1}) f(s_t|Y_{t-1})}{f(y_t|Y_{t-1})}.$$

These three equations provide a recursive system. We begin at $t = 1$ and iterate forward through time. The likelihood equation will deliver us the likelihood of the observed data and we can use this to estimate θ via maximum likelihood. The updating equation will give us a route to solving the filtering problem.

There are two simple cases that we will work through analytically: (1) Linear, Gaussian case, which will deliver the **Kalman filter** and (2) The discrete, markov case.

7.2 The Kalman Filter

Now, assume that all distributions are normally distributed. All of the results follow from carefully applying properties of joint normal distributions. We have that

- **State equation**: $s_t = Ts_{t-1} + R\epsilon_t$,
- **Measurement equation**: $y_t = Zs_t + S\eta_t$,

where the innovations ϵ_t, η_t are i.i.d. jointly normal with

$$\begin{pmatrix} \epsilon_t \\ \eta_t \end{pmatrix} \sim N\left(0, \begin{pmatrix} Q & 0 \\ 0 & H \end{pmatrix}\right).$$

Then, define

$$s_{t|t-1} = \mathbb{E}\left[s_t \mid Y_{t-1}\right] = Ts_{t-1|t-1},$$

which follows from the state and measurement equations. Moreover, notice that

$$V(s_t|Y_{t-1}) = RQR'.$$

Therefore, the distribution of the current state s_t given the previous state s_{t-1} and observed data is

$$f(s_t|s_{t-1}, Y_{t-1}) = N\left(Ts_{t-1|t-1}, RQR'\right).$$

The prediction equation is then given by

$$f(s_t|Y_{t-1}) = N(s_{t|t-1}, P_{t|t-1}),$$

where

$$\begin{aligned}
s_{t|t-1} &= Ts_{t-1|t-1}, \\
P_{t|t-1} &= \mathbb{E}\left[(s_t - s_{t|t-1})(s_t - s_{t|t-1})' \mid Y_{t-1}\right] \\
&= \mathbb{E}\left[(T(s_{t-1} - s_{t-1|t-1}) + R\epsilon_t)(T(s_{t-1} - s_{t-1|t-1}) + R\epsilon_t)' \mid Y_{t-1}\right] \\
&= TP_{t-1|t-1}T' + RQR'.
\end{aligned}$$

This gives us the **prediction equation**.

Now, we turn to constructing the **likelihood**. We have that

$$f(y_t|Y_{t-1}) = N\left(y_{t|t-1}, v_t\right),$$

where

$$\begin{aligned}
\nu_t &= y_t - y_{t|t-1} \\
&= y_t - Zs_{t|t-1} \\
v_t &= V\left(\nu_t|Y_{t-1}\right) \\
&= V\left(Z(s_t - s_{t|t-1}) + S\eta_t\right) \\
&= ZP_{t|t-1}Z' + SHS'.
\end{aligned}$$

Continuing, we next derive the **updating equation**. It is

$$f(s_t|Y_t) = f(s_t|\nu_t, Y_{t-1}).$$

The easiest way to get this is to derive the joint distribution of s_t, ν_t conditional on Y_{t-1}. We have that

$$\begin{pmatrix} \nu_t \\ s_t \end{pmatrix} = \begin{pmatrix} 0 \\ s_{t|t-1} \end{pmatrix} + \begin{pmatrix} Z(s_t - s_{t-1}) + S\eta_t \\ s_t - s_{t|t-1} \end{pmatrix}.$$

Therefore,

$$\begin{pmatrix} \nu_t \\ s_t \end{pmatrix} \mid Y_{t-1} \sim N\left(\begin{pmatrix} 0 \\ s_{t|t-1} \end{pmatrix}, \begin{pmatrix} v_t & ZP_{t|t-1} \\ P_{t|t-1}Z' & P_{t|t-1} \end{pmatrix} \right)$$

and it follows immediately that

$$s_t | \nu_t, Y_{t-1} \sim N(s_{t|t}, P_{t|t}),$$

where

$$\begin{aligned} s_{t|t} &= s_{t|t-1} + P_{t|t-1}Z'v_t^{-1}\nu_t \\ P_{t|t} &= P_{t|t-1} - P_{t|t-1}Z'v_t^{-1}P_{t|t-1}. \end{aligned}$$

We can then use these expressions to compute all values recursively. The tricky part is defining the initial condition. Recall that

$$s_t = Ts_{t|t-1} + S\eta t,$$

and so, we can then have that

$$\begin{aligned} s_{1|0} &= \mathbb{E}\left[s_1\right] \\ &= T\mathbb{E}\left[s_0\right] + S\mathbb{E}\left[\eta_0\right] = 0, \\ P_{1|0} &= V\left(s_1\right) \\ &= TP_{0|0}T' + SQS' \\ P_1 &= TP_1T' + SQS' \end{aligned}$$

by stationarity. Then, we can us this to directly solve for P_1.

7.2.1 The Kalman smoother

Recall that the state and measurement equations are

$$\begin{aligned} s_t &= Ts_{t-1} + R\epsilon_t \\ y_t &= Zs_t + S\eta_t, \end{aligned}$$

where (ϵ_t, η_t) are jointly normally distributed. The **Kalman Smoother** then gives that $s_t|Y_T$. How? We do so via a backwards recursion. The Kalman Filter terminates with computing $s_{T:T}, P_{T:T}$. We can then work backwards.

Recall our notation:

$$\begin{aligned} s_{t+1} &= Ts_t + R\epsilon_{t+1}, \\ s_{t+1|T} &= Ts_{t|t}, \\ s_{t+1} - s_{t+1|T} &= T(s_t - s_{t|T}) + R\epsilon_{t+1}. \end{aligned}$$

Then, we have that

$$\begin{pmatrix} s_{t+1} \\ s_t \end{pmatrix} \mid Y_t \sim N\left(\begin{pmatrix} s_{t+1|t} \\ s_{t|t} \end{pmatrix}, \begin{pmatrix} P_{t+1|t} & TP_{t|t} \\ P_{t|t}T' & P_{t|t} \end{pmatrix} \right).$$

So, we have that

$$s_t = s_{t|t} + P_{t|t}T'P_{t+1|t}^{-1}(s_{t+1} - s_{t+1|t}).$$

Now, we can show that

$$\begin{aligned} s_{t|T} &= \mathbb{E}\left[s_t \mid Y_T\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[s_t \mid s_{t+1}, Y_T\right] \mid Y_T\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[s_t \mid s_{t+1}, Y_t\right] \mid Y_T\right] \\ &= \mathbb{E}\left[s_{t|t} + P_{t|t}T'P_{t+1|t}^{-1}(s_{t+1} - s_{t+1|t}) \mid Y_T\right] \\ &= s_{t|t} + P_{t|t}T'P_{t+1|t}^{-1}(s_{t+1|T} - s_{t+1|t}). \end{aligned}$$

We can also derive a recursive expression for the variance as well.

7.3 Markov-Switching Filter

If the latent states are discrete, then all of the integrals in the prediction, likelihood and update equations becomes sums. In this case, it is quite easy to compute. We'll consider a simple case. Suppose that

$$y_t = \mu + \beta s_t + \epsilon_t,$$

where

$$s_t = \begin{cases} 1 \text{ w.p. } p \text{ if } s_{t-1} = 1, \\ 0 \text{ w.p. } 1-p \text{ if } s_{t-1} = 1, \\ 1 \text{ w.p. } 1-q \text{ if } s_{t-1} = 0, \\ 0 \text{ w.p. } q \text{ if } s_{t-1} = 0. \end{cases}$$

We think of s_t as being a regime and it is unobserved. We'll assume that $\epsilon_t \sim N(0, \sigma_\epsilon^2)$. We then have that

1. **State density**:

$$\mathbb{P}\left\{s_t \mid s_{t-1}, Y_{t-1}\right\} = p^{s_t}(1-p)^{1-s_t}\mathbb{1}\left\{s_{t-1}=1\right\} + q^{s_t}(1-q)^{1-s_t}\mathbb{1}\left\{s_{t-1}=0\right\}.$$

2. **Measurement density**:

$$f(y_t|s_t, Y_{t-1}) = N(\mu + \beta s_t, \sigma_\epsilon^2)$$

3. **Prediction density**:

$$\mathbb{P}\left\{s_t \mid Y_{t-1}\right\} = p^{s_t}(1-p)^{1-s_t}\mathbb{P}\left\{s_{t-1}=1 \mid Y_{t-1}\right\} + q^{s_t}(1-q)^{1-s_t}\mathbb{P}\left\{s_{t-1}=0|Y_{t-1}\right\}$$

4. **Likelihood**:

$$f(y_t|Y_{t-1}) = N(\mu+\beta, \sigma_\epsilon^2)\mathbb{P}\left\{s_{t-1}=1 \mid Y_{t-1}\right\} + N(\mu, \sigma_\epsilon^2)\mathbb{P}\left\{s_{t-1}=0 \mid Y_{t-1}\right\}.$$

5. **Updating**:

$$\mathbb{P}\left\{s_t=1 \mid Y_t\right\} = \frac{N(\mu+\beta, \sigma_\epsilon^2)\mathbb{P}\left\{s_t=1|Y_{t-1}\right\}}{f(y_t|Y_{t-1}},$$
$$\mathbb{P}\left\{s_t=0 \mid Y_t\right\} = \frac{N(\mu, \sigma_\epsilon^2)\mathbb{P}\left\{s_t=0|Y_{t-1}\right\}}{f(y_t|Y_{t-1}}$$

8 Dynamic factor models

There are three ways that we can think about dynamic factor models:

1. *Data compression*: Dynamic factor models reduce the dimensionality of a large number of time series, summarizing all of the comovements into a few factors.
2. **Forecasting tool**: By summarizing a large number of time series into a few factors, we can simply use the factors to make forecasts.
3. **Structural analysis**: Dynamic factor models are often used as an input into structural vector autoregression analyses – these are referred to as *structural DFMs* or *factor-augmented VARs*.

8.1 Dynamic factor models

Suppose there are N time series, where $X_{i,t}$ is a single time series for $i = 1, \ldots, N$. The *dynamic* or *state space* form of the dynamic factor model is

$$X_{i,t} = \lambda_i(L)f_t + e_{i,t},$$
$$\psi(L)f_t = \eta_t,$$

where $\mathbb{E}\left[e_{i,t}e_{j,t}\right] = 0$, $e_{i,t}$ may be serially correlated and $\mathbb{E}\left[\eta_t e_s'\right] = 0$. We assume that f_t is $q \times 1$ and we call f_t the **dynamic factors**. Intuitively, the dynamic factors are a common component of the observed time series and this common component explains all observed comovements in the data.

Assume that $\lambda(L)$ is a p-th order polynomial. Write

$$\underbrace{X_t}_{N\times 1} = \begin{pmatrix} X_{1,t} \\ \vdots \\ X_{N,t} \end{pmatrix}, \quad \Lambda = \begin{pmatrix} \lambda_{1,0} & \dots & \lambda_{1,p} \\ \vdots & & \vdots \\ \lambda_{N,0} & \dots & \lambda_{N,p} \end{pmatrix}, \quad F_t = \begin{pmatrix} f_t \\ \vdots \\ f_{t-p} \end{pmatrix}.$$

Then, we can also write the dynamic factor model in the *static form* as

$$\begin{aligned} X_t &= \Lambda F_t + e_t, \\ \Psi(L) F_t &= G\eta_t, \end{aligned}$$

for some matrix G.

The key questions are:

1. How do we estimate Λ, F_t?
2. What are r? What is q? where r is the number of static factors and q is the number of dynamic factors.
3. How do we translate this into a structural DFM?

First, let's discuss estimation. There are two main approaches. The first approach uses the state space set-up and estimates the factors and loadings using KF/MLE. The second approach estimates F_t using principal components. Second, we discuss choosing r, q. The common approach is to use an information criterion to do so. (Stock and Watson, 2016) provides an extensive review of the literature on the estimation of dynamic factor models, which describes these different methods in detail.

8.2 Structural DFMs

Assume that all identified parameters can be estimated exactly in a DFM.

Recall 3. *Assume that $A(L)Y_t = \eta_t$ and assume invertibility. So, $\eta_t = \Theta_0 \epsilon_t$ with Θ_0^{-1} exists. Then, we have that $Y_t = A(L)^{-1}\Theta_0 \epsilon_t$, where $\Theta(L) = A(L)^{-1}\Theta_0$. In general, invertibility is a very strong assumption. It will fail if there are more structural shocks that observed series or there is measurement error in the series.*

Now, return to the DFM. We have that

$$\begin{aligned} Y_T &= \Lambda F_t + e_t \\ \Psi(L) F_t &= G\eta_t. \end{aligned}$$

We assume invertibility with

$$\eta_t = \Theta_0 \epsilon_t$$

Substituting this in, we get that

$$Y_t = \left(\Lambda \Psi(L)^{-1} G \Theta_0\right) \epsilon_t + e_t,$$

where now $\Theta(L) = \left(\Lambda\Psi(L)^{-1}G\Theta_0\right)$. We need to now make two normalizations:

1. **Unit effect normalization**: As before, we normalize $\Theta_{0,jj} = 1$ for all j.

2. **Factor normalization**: We write

$$Y_t = \Lambda RR'F_t + e_t = \Lambda F_t + e_t$$

Therefore, we can only identify the space spanned by the factor and so, we need to introduce another normalization. It is common to use the **name factor normalization**. To understand what this is, consider an example

$$\begin{pmatrix} FFR_t \\ \Delta\log(WTI_t) \\ \Delta\log(GDP_t) \\ Y_{4:n,t} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ & \text{unrestricted} & \end{pmatrix} \begin{pmatrix} F_t^{FFR} \\ F_t^{Oil} \\ F_t^{GDP} \end{pmatrix} + e_t$$

where unrestricted $= \hat{\Lambda}_{4:n}^{PC}\hat{\Lambda_{1:3}}^{-1}$. That is, we name the factors the monetary policy factor, the oil factor and some real business activity factor. We then impose that the loadings of these factors on the FFR, WTI and GDP be one.

Together, these normalizations imply that 1 unit monetary policy shock raises the FFR factor by 1 unit and a 1 unit increase in the FFR factor increases the FFR by 1 unit. With these normalizations, we are back in the SVAR world and can use our tricks from earlier to identify Θ_0.

References

Baumeister, Christiane, and James Hamilton. 2015. "Sign Restrictions, Structural Vector Autoregressions, and Useful Prior Information." *ECMA*, 83(5): 1963–1999.

Berk, Kenneth. 1974. "Consistent Autoregressive Spectral Estimates." *The Annals of Statistics*, 2(3): 489–502.

Bernanke, Ben S, and Kenneth Kuttner. 2005. "What Explains the Stock Market's Reaction to Federal Reserve Policy?" *Journal of Finance*, 60: 1221–1257.

Blanchard, Olivier J., and Danny Quah. 1989. "The Dynamic Effects of Aggregate Demand and Supply Disturbances." *American Economic Review*, 79: 655–673.

Brockwell, Peter J., and Richard A. Davis. 1991. *Time Series: Theory and Methods.* New York, USA:Springer.

Cochrane, John H., and Monika Piazessi. 2002. "The Fed and Interest Rates - A High-Frequency Identification." *American Economic Review*, 92: 90–95.

Hamilton, James D. 1994. *Time Series Analysis.* Princeton University Press.

Hamilton, James D. 2018. "Why You Should Never Use the Hodrick-Prescott Filter." *The Review of Economics and Statistics*, 100(5): 831–843.

Hayashi, Fumio. 2000. *Econometrics.* Princeton University Press.

Jansson, Michael. 2004. "The Error in Rejection Probability of Simple Autocorrelation Robust Tests." *Econometrica*, 72(3): 937–946.

Kuttner, Kenneth. 2001. "Monetary policy surprises and interest rates: evidence from the Fed funds futures market." *Journal of Monetary Economics*, 47: 523–544.

Lazarus, Eben, Daniel Lewis, and James H. Stock. 2018. "The Size-Power Tradeoff in HAR Inference."

Lazarus, Eben, Daniel Lewis, James H. Stock, and Mark W. Watson. 2018. "HAR Inference: Recommendations for Practice." *Journal of Business and Economic Statistics*, 36(4): 541–575.

Newey, Whitney, and Daniel McFadden. 1994. "Large sample estimation and hypothesis testing." In *Handbook of Econometrics*. Vol. 4, 2111–2245.

Newey, Whitney K., and Kenneth D. West. 1987. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica*, 55: 703–708.

Rambachan, Ashesh, and Neil Shephard. 2019. "A nonparametric dynamic causal model for macroeconometrics."

Ramey, Valerie A. 2016. "Macroeconomics Shocks and their Propagation." In *Handbook of Macroeconomics*. Vol. 2A, , ed. John B. Taylor and Harald Uhlig, 71–162. Amsterdam, The Netherlands:North Holland.

Rudebusch, Glenn D. 1998. "Do Measures of Monetary Policy in a VAR Make Sense?" *International Economic Review*, 39: 907–931.

Sims, Christopher A. 1980. "Macroeconomics and Reality." *Econometrica*, 48: 1–48.

Stock, James H., and Mark W. Watson. 2016. "Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics." In *Handbook of Macroeconomics*. Vol. 2A, , ed. John B. Taylor and Harald Uhlig, 415–525.

Stock, James H., and Mark W. Watson. 2018. "Identification and Estimation of Dynamic Causal Effects in Macroeconomics using External Instruments." *Economic Journal*, 128: 917âĂŞ–948.

Sun, Yixiao, Peter C. B. Phillips, and Sainan Jin. 2008. "Optimal Bandwidth Selection in Heteroskedasticity Autocorrelation Robust Testing." *Econometrica*, 76(1): 175–194.

Uhlig, Harold. 2005. "What Are the Effects of Monetary Policy on Output? Results from an Agnostic Identification Procedure." *Journal of Monetary Economics*, 52: 381–419.