Empirical Analysis II-2.
Winter 2019

(Lars Hansen)

Simon Sangmin Oh

University of Chicago

Note:

These supplementary notes are meant to be used in conjunction with the lecture notes provided by Prof. Hansen.

Contents

1 Decision Theory

1.1 Bayesian Decision Rules

We consider the problem of a decision maker having to make a decision after observing a realization of a random vector, the distribution of which is not completely known.

1.1.1 Basic Setup

Consider the following setup:

- ▷ A random vector $Y \in \mathcal{Y}$ described by a probability density $\psi(y|\theta)$ relative to a measure τ over $\mathcal{Y}$. The likelihood function $\psi(y|\theta)$ can be thought of as a *statistical model*.
- ▷ $\theta \in \Theta$ is an unknown parameter that affects the probability distribution of Y.
- ▷ Decision rule $D : \mathcal{Y} \to \mathcal{D}$ (Borel measurable).
- ▷ Preferences represented by a utility function $U(d, \theta, y)$, where $y \in \mathcal{Y}$ denotes a realisation of Y, and $d \in \mathcal{D}$ is the decision made by the DM having observed y (it can depend on y).
  - * The θ term may show up in the utility function because the θs, in a dynamic context, inform the forecast of the consequences of my action.

I want to adjust this utility for risk. In other words, to consider the U before observing y, we can compute the expectation and define the *risk function* as

$$\overline{U}(D|\theta) = \int_{\mathcal{Y}} U[D(y), \theta, y] \psi(y|\theta) \tau(dy).$$

$\overline{U}$ is referred to as a *risk function* (in statistical decision theory) and it depends on both the decision rule D and the parameter θ. Integration of U over y conditional on θ adjusts for the risk in y. This is analogous to the "frequentist risk" discussed in Prof. Uhlig's class.

Example 1.1. *(Model selection problem).* Suppose that $\Theta = \{\theta_1, \theta_2\}$, where each value of θ corresponds to a statical model (i.e. a different probability density over $\mathcal{Y}$). A decision-maker wishes to select a model and has the following utility function:

$$U(d, \theta, y) = \begin{cases} d & \text{if } \theta = \theta_1 \\ 1 - d & \text{if } \theta = \theta_2 \end{cases},$$

where $d \in \mathcal{D} = [0, 1]$ is the probability of choosing θ_1 and $1 - d$ is the probability of choosing θ_2 (this specification allows for randomised decisions).

We focus on threshold decision rules. Partition $\mathcal{Y}$ into two disjoint sets $\mathcal{Y}_1$ and $\mathcal{Y}_2$, and consider the non-randomised decision rule:

$$D(y) = \begin{cases} 1 & \text{if } y \in \mathcal{Y}_1 \\ 0 & \text{if } y \in \mathcal{Y}_2 \end{cases}.$$

The risk function then for an arbitrary partitioning is given as:

$$\begin{aligned}
\overline{U}\left(D|\theta_1\right) &= \int_{\mathcal{Y}} U\left[D\left(y\right), \theta_1, y\right] \psi\left(y|\theta_1\right) \tau\left(dy\right) \\
&= \int_{\mathcal{Y}_1} U\left[1, \theta_1, y\right] \psi\left(y|\theta_1\right) \tau\left(dy\right) + \int_{\mathcal{Y}_2} U\left[0, \theta_1, y\right] \psi\left(y|\theta_1\right) \tau\left(dy\right) \\
&= \int_{\mathcal{Y}_1} \psi\left(y|\theta_1\right) \tau\left(dy\right) = \mathbb{P}\left(y \in \mathcal{Y}_1|\theta_1\right), \\
\overline{U}\left(D|\theta_2\right) &= \int_{\mathcal{Y}_1} U\left[1-1, \theta_2, y\right] \psi\left(y|\theta_2\right) \tau\left(dy\right) + \int_{\mathcal{Y}_2} U\left[1-0, \theta_2, y\right] \psi\left(y|\theta_2\right) \tau\left(dy\right) \\
&= \int_{\mathcal{Y}_2} \psi\left(y|\theta_2\right) \tau\left(dy\right) = \mathbb{P}\left(y \in \mathcal{Y}_2|\theta_2\right).
\end{aligned}$$

$\overline{U}\left(D|\theta_i\right)$ gives the probability that the decision rule $D\left(y\right)$ "correctly" chooses the model θ_i, while $1 - \overline{U}\left(D|\theta_i\right)$ is the probability of choosing the incorrect model.

1.1.2 Ex-ante vs. Ex-post Problems

Let a probability measure π over Ω denote the *prior distribution* that summarises the DM's ignorance about θ.

Definition 1.1. *(Ex ante decision problem)* A Bayesian DM maximises his expected utility *using his prior beliefs* with respect to his *decision rule* D; i.e.

$$\max_{D} \int_{\Theta} \overline{U}\left(D|\theta\right) \pi\left(d\theta\right)$$

The solution to the problem is the optimal decision rule $D^*\left(y\right)$.

Definition 1.2. *(Ex post decision problem)* For all $y \in \mathcal{Y}$,

$$\max_{d \in \mathcal{D}} \int_{\Theta} U\left(d, \theta, y\right) \bar{\pi}\left(d\theta|y\right) := \mathbb{E}\left[U\left(d, \theta, y\right)|y\right]$$

The maximiser d^* for each $y \in \mathcal{Y}$ induces the decision rule $\bar{D}^*\left(y\right)$.

The two problems differ in the following respects:

1. Observation of $y \in \mathcal{Y}$
   - ▷ The ex-ante problem gets its name because the agent needs to derive the decision rule before he observes $y \in \mathcal{Y}$.
   - ▷ The ex-post problem gets its name because the agent now chooses the decision rule after observing $y \in \mathcal{Y}$.
2. U vs. $\bar{U}$
   - ▷ In the ex-ante problem, the only probability measure he can use is his prior, and he will choose a decision rule that maximizes his expected *adjusted* utility ($\bar{U}$) with respect to the prior.
     - * Since y is uncertain, we need to adjust U for the different values of y and obtain $\bar{U}$ first.
   - ▷ In the ex-post problem, the agent will use the updated posterior belief that maximizes his expected utility (U) with respect to the posterior.
     - * Here we already observe y, so we can just use U in our maximization problem.

3. Decomposing ex-ante problems

   ▷ One can break the ex-ante decision problem into two subproblems: (1) Given an observation of Y, estimate θ and (2) Given an estimate of θ, choose a decision rule.

Notice that if the objective function of the ex ante decision problem is finite when evaluated at D^*, we have then $D^* \equiv \bar{D}^*$. This result may not hold, however, in an experiment setting where your decisions affect outcomes and the data you generate.

Example 1.2. *(Model Selection Revisited)* What does the Bayesian decision rule look like? We start by computing the posterior probabilities:

$$\bar{\pi}\left(d\theta_1|y\right) = \frac{\psi\left(y|\theta_1\right)\pi_1}{\psi\left(y|\theta_1\right)\pi_1 + \psi\left(y|\theta_2\right)\pi_2},$$
$$\bar{\pi}\left(d\theta_2|y\right) = \frac{\psi\left(y|\theta_2\right)\pi_2}{\psi\left(y|\theta_1\right)\pi_1 + \psi\left(y|\theta_2\right)\pi_2}.$$

Using these probabilities, for any $y \in \mathcal{Y}$, the ex-post decision problem is then:

$$\begin{aligned}\int_\Theta U\left(d,\theta,y\right)\bar{\pi}\left(d\theta|y\right) &= \int_{\{\theta_1,\theta_2\}} U\left(d,\theta,y\right)\bar{\pi}\left(d\theta|y\right) \\ &= U\left(d,\theta_1,y\right)\bar{\pi}\left(d\theta_1|y\right) + U\left(d,\theta_2,y\right)\bar{\pi}\left(d\theta_2|y\right) \\ &= d\bar{\pi}\left(d\theta_1|y\right) + \left(1-d\right)\bar{\pi}\left(d\theta_2|y\right).\end{aligned}$$

and thus:

$$\max_{d\in\mathcal{D}} \quad d\bar{\pi}\left(d\theta_1|y\right) + \left(1-d\right)\bar{\pi}\left(d\theta_2|y\right)$$

The solution to this problem is then

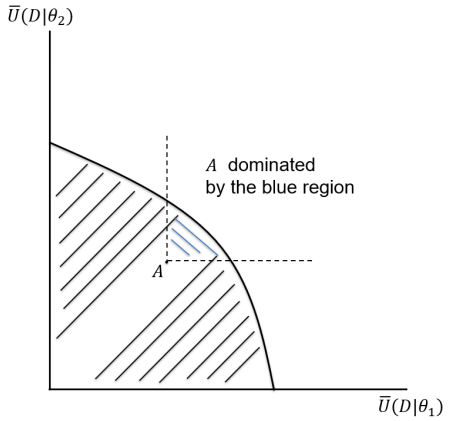
$$d^*\left(y\right) = \begin{cases} 1 & \bar{\pi}\left(d\theta_1|y\right) > \bar{\pi}\left(d\theta_2|y\right) \\ [0,1] & \bar{\pi}\left(d\theta_1|y\right) = \bar{\pi}\left(d\theta_2|y\right),\ \forall y \in \mathcal{Y}. \\ 0 & \bar{\pi}\left(d\theta_1|y\right) < \bar{\pi}\left(d\theta_2|y\right) \end{cases}$$

Assuming that $d^*\left(y\right) = 0$ if $\bar{\pi}\left(d\theta_1|y\right) = \bar{\pi}\left(d\theta_2|y\right)$, we can write a *threshold rule* that is a solution to the ex post problem:

$$d^*\left(y\right) = \begin{cases} 1 & \bar{\pi}\left(d\theta_1|y\right) > \bar{\pi}\left(d\theta_2|y\right) \\ 0 & \bar{\pi}\left(d\theta_1|y\right) \leq \bar{\pi}\left(d\theta_2|y\right) \end{cases},\ \forall y \in \mathcal{Y}.$$

Thus, the Bayesian solution of the model selection problem is to select the model for which the posterior probability is highest, and this generates a partition of the set $\mathcal{Y}$. Furthermore, it can be shown that this will lie on the boundary of the convex set we drew earlier.

Figure 1: Inadmissibility



1.2 Partial Orderings and Admissibility

Earlier, we saw that the Bayesian decision rule D^* depends on the prior distribution π. However, even before specifying a prior, the utility function $\overline{U}(D|\theta)$ allows us to identify some decision rules that cannot be dominated.

Definition 1.3. Consider decision rules $D_1 : \mathcal{Y} \to \mathcal{D}$ and $D_2 : \mathcal{Y} \to \mathcal{D}$. Under a (partial) ordering $\succsim$, D_2 is preferred overD_1, denoted $D_2 \succsim D_1$, if and only if

$$\overline{U}(D_2|\theta) \geq \overline{U}(D_1|\theta), \ \forall \theta \in \Theta.$$

The ordering is partial because $\succsim$ ranks some, but not all, pairs of decision rules. For example, it might be that for some $\theta_1, \theta_2 \in \Theta$, we have

$$\overline{U}(D_2|\theta_2) > \overline{U}(D_1|\theta_2), \ \overline{U}(D_1|\theta_1) < \overline{U}(D_2|\theta_1).$$

We restrict our attention to such undominated decision rules, called *admissible* decision rules. We rule out decision rules that gives strictly lower utility under all priors.

Definition 1.4. *(Admissibility)* A decision rule D is *admissible* if there exists no decision rule $\tilde{D}$ for which $\tilde{D} \succsim D$ with $\overline{U}\left(\tilde{D}, \tilde{\theta}\right) > \overline{U}\left(D, \tilde{\theta}\right)$ for some $\tilde{\theta} \in \Theta$.

Remark 1.1. If you want to get an admissible decision rule, you solve the Bayesian problem. You repeat the process for a set of priors, and you get a whole frontier of admissible decision rules as outlined earlier. This still doesn't address the question of where the prior comes from, but at least we have a boundary.

1.3 Max-min Decision Rules

These rules are designed to address the uncertainty regarding prior selection. Specifically, a max-min DM wishes to construct a robust rule (with respect to his uncertainty about the prior). He does so by exploring the consequences of a decision rule across alternative priors by formulating a max-min problem that expresses the DM's aversion to his uncertainty about the prior. Note that the chosen prior depends upon the DM's utility function.

Here, we assume that the DM has *multiple* prior probability distributions over θ. Specifically, let C denote a positive convex function of possible prior probability distributions π over θ. The function C reflects the DM's *ambiguity* about the prior π. This is separate from the function U which expresses DM's risk aversion (aversion to uncertainty over the outcome y and θ).

Definition 1.5. *(Max-min ex ante problem)*

$$\max_{D} \min_{\pi} \quad \left[\int_{\Theta} \overline{U}\left(D|\theta\right) \pi\left(d\theta\right) + C\left(\pi\right) \right].$$

Given a decision rule, solve the minimization problem. Rank the decision rules based on the outcomes of the minimization problem and choose the best one. Note that you can formulate it as max-max. We're just implicitly assuming that the agent is risk-averse and is hence interested in problems of the above sort.

The following are examples of a commonly used C.

1. Denote Π as the set of possible priors and consider

$$C\left(\pi\right) = \begin{cases} 0 & \pi \in \Pi \\ \infty & \pi \notin \Pi \end{cases}.$$

   This ensures that π being chosen by the max-min DM belongs in the set Π.

2. Suppose that Θ has n elements and that π^o denotes a "reference" prior. Define

$$C\left(\pi\right) = \kappa \sum_{j=1}^{n} \pi_j \left(\log \pi_j - \log \pi_j^o\right)$$

   which is a convex function in the probabilities.

   - ▷ The summation term is referred to as the *relative entropy* of the π distribution with respect to the π^o distribution. Entropy is the expected value of the log-likelihood ratio, where the expectation is taken with respect to π.
   - ▷ The function C is convex and attains minimum at $\pi = \pi^o$. The parameter κ is the penalisation factor—a higher value means greater penalty from deviating from the baseline prior distribution π^o.
   - ▷ Setting $\kappa = \infty$ reduces the max-min ex ante problem to the Bayesian (ex ante) problem with a (unique) prior π^o.
   - ▷ To verify that $C(\pi) \geq 0$, notice that $r \log r \geq r - 1$ since the convex function always lies above the gradient line. Then:

$$C\left(\pi\right) = \kappa \sum_{j=1}^{n} \frac{\pi_j}{\pi_j^\circ} \left(\log \frac{\pi_j}{\pi_j^\circ}\right) \pi_j^\circ \geq \kappa \sum_{j=1}^{n} \left(\frac{\pi_j}{\pi_j^\circ} - 1\right) \pi_j^\circ = 0$$

1.3.1 Example: Exponential Tilting

Let us solve the max-min ex ante problem assuming the entropy as the C function (choice #2 from above)

1. We first solve the inner minimization problem, fixing D, where you are choosing a probability distribution π:

$$\min_{\pi} \quad \sum_{j=1}^{n} \pi_j \left[\overline{U}\left(D|\theta_j\right) + \kappa \left(\log \pi_j - \log \pi_j^o\right)\right] \quad s.t. \quad \sum_{j=1}^{n} \pi_j = 1,\ \pi_j \geq 0,\ \forall j, .$$

▷ The Lagrangian is given by

$$\mathcal{L} = \sum_{j=1}^{n} \pi_j \left[\overline{U}\left(D|\theta_j\right) + \kappa \left(\log \pi_j - \log \pi_j^o \right) \right] + \lambda \left(1 - \sum_{j=1}^{n} \pi_j \right).$$

▷ The first-order condition with respect to π_j is given by

$$\overline{U}\left(D|\theta_j\right) + \kappa \left(\log \pi_j^* - \log \pi_j^o \right) + \kappa - \lambda = 0,$$

which can be rearranged to

$$\begin{aligned} \log \pi_j^* &= \log \pi_j^o - \frac{1}{\kappa} \overline{U}\left(D|\theta_j\right) + \left(\frac{\lambda}{\kappa} - 1 \right) \\ \Rightarrow \pi_j^* &= \pi_j^o \exp\left(-\frac{1}{\kappa} \overline{U}\left(D|\theta_j\right) \right) \exp\left(\frac{\lambda}{\kappa} - 1 \right). \end{aligned}$$

▷ Since π_j^* is a probability, they must sum to one:

$$\pi_j^* = \frac{\pi_j^o \exp\left(-\frac{1}{\kappa} \overline{U}\left(D|\theta_j\right) \right)}{\sum_{i=1}^{n} \pi_i^o \exp\left(-\frac{1}{\kappa} \overline{U}\left(D|\theta_i\right) \right)}$$

▷ *Interpretation:* π^* is the optimal prior. Since $\exp\left(-x\right)$ is decreasing in x, π_j^* (exponentially) *tilts* probabilities away from reference prior π_j° towards parameter values θ_j with lower conditional expected utilities. A higher κ tends to equalise probabilities across possible values of θ_j (if $\kappa = \infty$, we have $\pi_j^* = \pi_j^o / \sum_{i=1}^{n} \pi_j^o$). If $\overline{U}$ is small, I will twist the original reference probability more towards the parameter that gives rise to that $\bar{U}$.

2. Substituting into the objective function, we obtain the minimised objective function:

$$\begin{aligned} &\sum_{j=1}^{n} \pi_j \left[\overline{U}\left(D|\theta\right) + \kappa \left(\log \pi_j - \log \pi_j^o \right) \right] \\ =& \sum_{j=1}^{n} \pi_j \left[\overline{U}\left(D|\theta\right) + \kappa \left(\log \frac{\pi_j^o \exp\left(-\frac{1}{\kappa} \overline{U}\left(D|\theta_j\right) \right)}{\sum_{i=1}^{n} \pi_i^o \exp\left(-\frac{1}{\kappa} \overline{U}\left(D|\theta_i\right) \right)} - \log \pi_j^o \right) \right] \\ =& \sum_{j=1}^{n} \pi_j \left[\overline{U}\left(D|\theta\right) + \kappa \left(-\frac{1}{\kappa} \overline{U}\left(D|\theta_j\right) - \log \left(\sum_{i=1}^{n} \pi_i^o \exp\left(-\frac{1}{\kappa} \overline{U}\left(D|\theta_i\right) \right) \right) \right) \right] \\ =& -\kappa \log \left(\sum_{i=1}^{n} \pi_i^o \exp\left(-\frac{1}{\kappa} \overline{U}\left(D|\theta_i\right) \right) \right) \sum_{j=1}^{n} \pi_j \\ =& -\kappa \log \left(\sum_{i=1}^{n} \pi_i^o \exp\left(-\frac{1}{\kappa} \overline{U}\left(D|\theta_i\right) \right) \right). \end{aligned}$$

So you're making this adjustment based on the decision problem. Think of this as an "extra curvature adjustments."

3. Notice that our objective function satisfies:

$$\begin{aligned} &\sum_{j=1}^{n} \pi_j \left[\overline{U}\left(D|\theta\right) + \kappa \left(\log \pi_j - \log \pi_j^o \right) \right] \\ \geq& \sum_{j=1}^{n} \pi_j^* \overline{U}\left(D|\theta\right) + \kappa \sum_{j=1}^{n} \pi_j^* \left(\log \pi_j^* - \log \pi_j^o \right) \end{aligned}$$

▷ Now suppose

$$\kappa \sum_{j=1}^{n} \pi_j \left(\log \pi_j - \log \pi_j^o \right) \leq \kappa \sum_{j=1}^{n} \pi_j^* \left(\log \pi_j^* - \log \pi_j^o \right)$$

Then it is necessarily the case that

$$\sum_{j=1}^{n} \pi_j \overline{U} \left(D|\theta \right) \geq \sum_{j=1}^{n} \pi_j^* \overline{U} \left(D|\theta \right)$$

This means that as long as you're looking at $\pi_j's$ that is less than or equal to the optimal π_j^* for some discrepancy measure (the first inequality), you are guaranteed a higher expected utility (the second inequality). This is a form of a "robustness bound" since you're guaranteed a certain level of expected utility if the first condition is satisfied.

1.3.2 Min-max Decision Rules

Whenever you see a max-min problem, you can transform into a two-person zero-sum game. Specifically, we will explore the consequences of changing the order of the minimization and the maximization problem.

Definition 1.6. (*Min-max ex post problem*)

$$\min_{\pi \in \Pi} \max_{D} \quad \int_{\Theta} \overline{U} \left(D|\theta \right) \pi \left(d\theta \right) + C \left(\pi \right).$$

Now we're maximizing for a given π and then later go over all possible values of π. Observe that $C \left(\pi \right)$ appears additively separably in the objective function. Thus, we can first solve

$$\max_{D} \quad \int_{\Theta} \overline{U} \left(D|\theta \right) \pi \left(d\theta \right)$$

and reinsert the maximised objective into the min-max problem.

▷ This problem is exactly the Bayesian ex ante problem, which we know we can solve via the ex post problem.

▷ Denote the Bayesian decision rule as D^* and assuming that the solution to the minimisation problem, $\hat{\pi}$, exists, we can interpret $\hat{\pi}$ as a "worst-case prior" for the Bayesian decision rule D^*.

▷ Moreover, since D^* is a Bayesian rule, the max-min decision rule (which coincides with the min-max decision rule under the assumption here) cannot be dominated except possibly on sets to which the worst-case prior assigns zero probability measure.

Remark 1.2. This brings us back to prior choice. Savage would have recommended a form of "economic thinking" or "introspection" in choosing a prior, whereas the above result implies a more systematic way of obtaining a "robust prior choice."

1.3.3 Example: Model Selection via Min-Max

Suppose $\Theta = \{\theta_1, \theta_2\}$ and the density of y conditional on θ_1, θ_2 are $\psi \left(y|\theta_1 \right)$ and $\psi \left(y|\theta_2 \right)$. Utilities are

$$U \left(d, y, \theta_1 \right) = d, \qquad U \left(d, y, \theta_2 \right) = 1 - d \text{ where } d \in [0, 1]$$

Now a Bayesian with prior (π_1, π_2) where π_1 is the prior for model 1 to be the correct model and π_2 is the prior for model 2 to be the correct model with $\pi_1 + \pi_2 = 1$.

1. **The Inner Maximization Problem:**

   ▷ Taking the ex-post approach:

   $$\max_d U(d, y, \theta_1)\,\bar{\pi}_1(y) + U(d, y, \theta_2)\,\bar{\pi}_2(y)$$

   which is equivalent to

   $$\max_d \{\bar{\pi}_1(y) - \bar{\pi}_2(y)\}\,d + \bar{\pi}_2(y)$$

   which yields the decision rule:

   $$D^*(y) = \begin{cases} 1 & \text{if } \bar{\pi}_1(y) > \bar{\pi}_2(y) \\ 0 & \text{if } \bar{\pi}_1(y) \leq \bar{\pi}_2(y) \end{cases}$$

   ▷ Now we need to find the posterior probabilities:

   $$\bar{\pi}_1(y) = \zeta(y)\,\pi_1\psi(y|\theta_1)$$
   $$\bar{\pi}_2(y) = \zeta(y)\,\pi_2\psi(y|\theta_2)$$

   where $\zeta(y)$ is a constant that could potentially depend on y. Then

   $$\begin{aligned} &\bar{\pi}_1(y) > \bar{\pi}_2(y) \\ \Leftrightarrow &\pi_1\psi(y|\theta_1) > \pi_2\psi(y|\theta_2) \\ \Leftrightarrow &\log\pi_1 + \log\psi(y|\theta_1) > \log\pi_2 + \log\psi(y|\theta_2) \end{aligned}$$

   Denoting $r \equiv \pi_1/\pi_2$ and $\phi(y) \equiv \psi(y|\theta_2)/\psi(y|\theta_1)$ we can rewrite the decision rule as

   $$D^*(y) = \begin{cases} 1 & \text{if } \phi(y) < r \\ 0 & \text{if } \phi(y) > r \end{cases}$$

   ▷ Note that r is a constant that is equal to

   $$r = \frac{\pi_1}{1-\pi_1}$$

   because this is a Bayesian's problem who is given the prior (π_1, π_2). There is no min-max or max-min problem here – a nice, simple Bayesian's problem.

   ▷ We further know that for every $r \in [0, +\infty]$, there must exist a prior (π_1, π_2) such that

   $$D^r(y) = \arg\max_{D(y)} \pi_1\bar{U}[D|\theta_1] + \pi_2\bar{U}[D|\theta_2]$$

   which is the ex-ante problem for the Bayesian.

   ▷ Define z to be the likelihood ratio:

   $$z = \phi(y) = \frac{\psi(y|\theta_2)}{\psi(y|\theta_1)}.$$

   We can write $\overline{U}(D|\theta_1)$ and $\overline{U}(D|\theta_2)$ equivalently as

   $$\overline{U}(D|\theta_1) = \int_{\phi(y)<r} \psi(y|\theta_1)\,\tau(dy),$$
   $$\overline{U}(D|\theta_2) = \int_{\phi(y)\geq r} \phi(y)\,\psi(y|\theta_1)\,\tau(dy).$$

▷ Suppose that $\psi(y|\theta_1)\tau(dy)$ in $\mathcal{Y}$ implies a density υ over the positive real numbers in z space, then

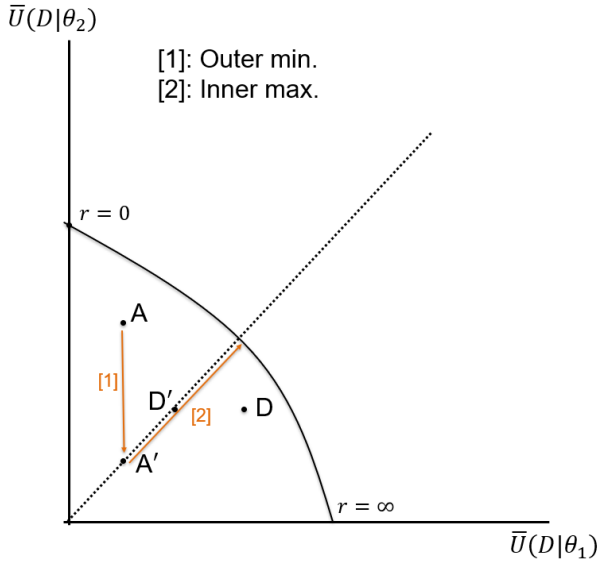
$$\overline{U}(D|\theta_1) = \int_0^r \upsilon(z)\,dz,$$
$$\overline{U}(D|\theta_2) = \int_r^\infty z\upsilon(z)\,dz.$$

Observe that

$$\frac{\partial \overline{U}(D|\theta_1)}{\partial r} = \upsilon(r) > 0, \qquad \frac{\partial \overline{U}(D|\theta_2)}{\partial r} = -r\upsilon(r) < 0,$$
$$\frac{\partial^2 \overline{U}(D|\theta_1)}{\partial r^2} = \upsilon'(r), \qquad \frac{\partial^2 \overline{U}(D|\theta_1)}{\partial r^2} = -r\upsilon'(r) - \upsilon(r).$$

Recall that a minimising DM takes as given the decision rule, which is characterised by the threshold in this case

2. **Justification of Priors that Satisfy $\overline{U}(D|\theta_1) = \overline{U}(D|\theta_2)$:**



▷ Interpretation of the graph:

* Each axis is frequentist risk, and each point represents a decision rule. There's nothing Bayesian here – no prior.
* For each decision rule, compute $\bar{U}(D|\theta_1)$ which is the expected utility assuming model 1 is correct, and $\bar{U}(D|\theta_2)$ which is the expected utility assuming model 2 is correct.

▷ We're going to argue that points off the 45-degree line cannot be the solution tohe min-max problem. Suppose that, for a given r (and therefore π_1 and π_2), it is the case that $\overline{U}(D|\theta_1) < \overline{U}(D|\theta_2)$. This is point D in the graph.

* Then, the minimising DM (currently solving the outer problem) will put all prior probability on θ_1 so that minimised objective is given by $\overline{U}(D|\theta_1)$.
* Because $\overline{U}(D|\theta_1)$ is increasing in r, the maximising DM can increase utility by increasing r. If he keeps increasing r, then at some point, we will have $\overline{U}(D|\theta_1) > \overline{U}(D|\theta_2)$ so that the minimising DM will now wants to put all prior probability on θ_2 so that the minimised objective is given by $\overline{U}(D|\theta_2)$. But since $\overline{U}(D|\theta_2)$ is decreasing in r, the maximising DM can increase utility by decreasing r.

* These incentives leads us to search for a value of r such that

$$\overline{U}\left(D|\theta_1\right) = \overline{U}\left(D|\theta_2\right).$$

which correspond to points A' and D'. At this value of r, the minimising DM who chooses probabilities has no way to influence the objective.

* For the maximising DM, causing r to deviate from this value would also reduce either side of the equality and therefore diminish the minimised objective function. Thus, the condition above characterises the solution to the max-min problem.

▷ As our final solution, the agent will choose the intersection of the frontier and the 45-degree line. This follows directly from admissibility of Bayesian estimators.

▷ Another way to see above is to consider the equivalent max-min DM problem.

* First, notice that the solution to the inner minimisation problem for the max-min DM is

$$\min\left\{\overline{U}\left(D|\theta_1\right), \overline{U}\left(D|\theta_2\right)\right\}.$$

* Let D^* be such that $\overline{U}\left(D^*|\theta_1\right) = \overline{U}\left(D^*|\theta_2\right)$. For any $D \neq D^*$, it must be that

$$\overline{U}\left(D^*|\theta_1\right) > \overline{U}\left(D|\theta_1\right) \text{ or } \overline{U}\left(D^*|\theta_2\right) > \overline{U}\left(D|\theta_2\right).$$

* That is,

$$\min\left\{\overline{U}\left(D^*|\theta_1\right), \overline{U}\left(D^*|\theta_2\right)\right\} > \min\left\{\overline{U}\left(D|\theta_1\right), \overline{U}\left(D|\theta_2\right)\right\}$$

so that D^* solves the max-min DM's maximisation problem, since we want to maximize the min.

3. **Concavity of the Curve**: To simplify notation, let

$$\bar{u}_1(r) = \overline{U}\left(D|\theta_1\right), \ \bar{u}_2(r) = \overline{U}\left(D|\theta_2\right).$$

We would like to plot the curve $\left(\bar{u}_1(r), \bar{u}_2(r)\right)$ traced out by changing r over the positive real numbers. In other words, we want to plot

$$\bar{u}_2(r) = f\left(\bar{u}_1(r)\right).$$

The slope is given by

$$\begin{aligned} \bar{u}_2'(r) &= f'\left(\bar{u}_1(r)\right)\bar{u}_1'(r) \\ \Rightarrow f'\left(\bar{u}_1(r)\right) &= \frac{-r\upsilon(r)}{\upsilon(r)} = -r < 0, \end{aligned}$$

We can also show that it is concave:

$$\begin{aligned} \bar{u}_2''(r) &= f''\left(\bar{u}_1(r)\right)\left(\bar{u}_1'(r)\right)^2 + f'\left(\bar{u}_1(r)\right)\bar{u}_1''(r) \\ \Rightarrow -r\upsilon'(r) - \upsilon(r) &= f''\left(\bar{u}_1(r)\right)\left(\upsilon(r)\right)^2 - r\upsilon'(r) \\ \Rightarrow f''\left(\bar{u}_1(r)\right) &= -\frac{1}{\upsilon(r)} < 0. \end{aligned}$$

Hence, the curve is concave.

1.3.4 A Modified Example

The curve examined above is symmetric around the 45 degree line given our utility function. What happesn if we have the following?:

$$U(d,\theta,y)=\begin{cases}1.5d & \text{if } \theta=\theta_1\\ 1-d & \text{if } \theta=\theta_2\end{cases}.$$

▷ In this case, the intersection with the x-axis moves to the right, and the intersection with the y-axis remains the same.

▷ To see this, note that the ex post problem is given by

$$\max_{d\in[0,1]} \quad 1.5d\bar{\pi}_1+(1-d)\,\bar{\pi}_2.$$

▷ The optimal decision rule (assuming the same tie-break rule):

$$d^*(y)=\begin{cases}1 & \text{if } 1.5\bar{\pi}(d\theta_1|y)>\bar{\pi}(d\theta_2|y)\\ 0 & \text{if } 1.5\bar{\pi}(d\theta_1|y)\leq\bar{\pi}(d\theta_2|y)\end{cases},\ \forall y\in\mathcal{Y}.$$

▷ To obtain the max-min solution, we need:

$$\begin{aligned}\bar{U}(d,\theta_1,y)=1.5\bar{\pi}(d\theta_1|y)=\bar{\pi}(d\theta_2|y)=U(d,\theta_2,y)\\ \Rightarrow\frac{\psi(y|\theta_2)}{\psi(y|\theta_1)}=\frac{2}{3}\frac{\pi_1}{1-\pi_1}=r.\end{aligned}$$

▷ We are still going to have the same 45-degree line, but the frontier will look different (larger intercept on the x-axis).

1.4 Risk vs. Uncertainty

Knight (1921) distinguished between risk and uncertainty. $\psi(y|\theta)$ can be thought of as ex-ante risk; curvature in U can be thought of as risk aversion. Assigning a prior π over θ can be thought of as uncertainty. Below, we discuss one tractable way to capture this notion of uncertainty.

The LHS represents one-stage reduced lottery; the RHS represents a two-stage lottery:

$$\phi^*(y^*|d,y)=\int\underbrace{\phi^*(y^*|d,y,\theta)}_{=\text{ risk}}\underbrace{\bar{\pi}^*(\theta|y)}_{=\text{ ambiguity}}$$

1. Choose y^* (next state) given θ
2. Choose θ given current $\varphi(y^*|y,D,\theta)$

This allows us to separate out the attitude towards risk and the attitude towards ambiguity.

Remark 1.3. In finance and macro, you end up having to appeal to having a high risk aversion. The decomposition above allows us to explore a different channel to match the empirical evidence.

Remark 1.4. You can never be obejctive in evaluating empirical evidence. I think that in doing empirical work, you need to make sure where the inputs are coming from. Robustness is one way to show this.

2 Stochastic Processes

2.1 Setup

▷ Invariant event: An event $\Lambda \in \mathfrak{F}$ is *invariant* if $\Lambda = \mathbb{S}^{-1}(\Lambda)$. Ω and $\emptyset$ are both invariant events. If Λ is an invariant event, then

$$\mathbb{S}^t(\omega) \in \Lambda,\ \forall \omega \in \Lambda,\ \forall t = 0, 1, 2 \ldots;$$

i.e. all sample points stay in the event Λ over time.

▷ Measure-preserving: The transformation $\mathbb{S}$ is measure-preserving if

$$\mathbb{P}(\Lambda) = \mathbb{P}\left\{\mathbb{S}^{-1}(\Lambda)\right\}$$

for all $\Lambda \in \mathfrak{F}$ where Λ is an event and $\mathfrak{F}$ is a collection of events.

▷ Conditional expectation: Let $\{\Lambda_j\}_{j=1}^{\infty}$ denote a countable partition of Ω,[1] where each Λ_j is an invariant event, and $\mathfrak{I}$ denote the collection of invariant events (a σ-algebra). The expectation conditional on a invariant set is

$$\mathbb{E}[X|\Lambda_i] = \frac{\int_{\Lambda_j} X(\omega)\, d\mathbb{P}(\omega)}{\mathbb{P}(\Lambda_j)}$$

so you have a different conditional expectation for the corresponding invariant set. The expectation conditional on the collection of all invariant sets $\mathfrak{J}$ is

$$\mathbb{E}[X|\mathfrak{J}](\omega) = \frac{\int_{\Lambda_j} X(\omega)\, d\mathbb{P}(\omega)}{\mathbb{P}(\Lambda_j)} \text{ if } \omega \in \Lambda_j.$$

The conditional expectation is constant for $\omega \in \Lambda_j$ but can vary across Λ_j's.

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

2.2 Law of Large Numbers

▷ Law of Large Numbers (*Birkhoff*): Suppose $\mathbb{S}$ is **measure-preserving**.

1. (Almost-sure convergence) For any X such that $\mathbb{E}[|X|] < \infty$,

$$\frac{1}{T}\sum_{t=1}^{T} X_t(\omega) \xrightarrow{\text{p}} \mathbb{E}[X|\mathfrak{J}](\omega).$$

2. (Mean-square convergence) For any X such that $\mathbb{E}[|X|^2] < \infty$,

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

▷ Law of Large Numbers (Standard): (Standard Law of Large Numbers) Suppose $\mathbb{S}$ is **ergodic** (i.e. for all $\Lambda \in \mathfrak{F}$, $\mathbb{P}(\Lambda) = 0, 1$)

[1] That is, $\Lambda_j \cap \Lambda_k = \emptyset, \forall j \neq k, \bigcup_{j=1}^{\infty} \Lambda_j = \Omega$ and each Λ_j is nonempty.

1. (*Almost-sure convergence*) For any X such that $\mathbb{E}[|X|] < \infty$,

$$\frac{1}{T}\sum_{t=1}^{T} X_t(\omega) \xrightarrow{\text{p}} \mathbb{E}[X].$$

2. (*Mean-square convergence*) For any X such that $\mathbb{E}[|X|^2] < \infty$,

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

▷ Once certain properties of $\mathbb{P}$ and $\mathbb{S}$ are satisfied, I can use any measurement function and it will be stationary.

▷ Why do we use $\mathbb{S}$ instead of $\mathbb{S}^{-1}$? Consider the example where $\mathbb{S}(\omega_1) = \omega_2, \mathbb{S}(\omega_2) = \omega_2$. Then

$$\mathbb{S}^{-1}(\{\omega_2\}) = \{\omega_1, \omega_2\} \Rightarrow \{\omega_2\} \neq \mathbb{S}^{-1}(\{\omega_2\})$$
$$\mathbb{S}(\{\omega_2\}) = \{\omega_2\} \Rightarrow \{\omega_2\} = \mathbb{S}(\{\omega_2\})$$

so $\{\omega_2\}$ is not invariant but using $\mathbb{S}$ instead of $\mathbb{S}^{-1}$ would seem like it satisfies the provided constraint.

2.3 Limiting Empirical Measures

▷ Limiting empirical measure: Given an event $\Lambda \in \mathfrak{F}$ and a measure-preserving transformation $\mathbb{S}$ for almost all $\omega \in \Lambda_j$, we define the limiting empirical measure Qr_j as

$$Qr_j(\Lambda)(\omega) \coloneqq \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}_\Lambda\left(\mathbb{S}^t(\omega)\right) = \frac{P(\Lambda \cap \Lambda_j)}{P(\Lambda_j)}$$

Alternately, write

$$\mathbb{P}(\Lambda) = \sum_{j=0}^{\infty} \mathbb{P}(\Lambda \cap \Lambda_j) = \sum_{j=0}^{\infty} \left(\frac{\mathbb{P}(\Lambda \cap \Lambda_j)}{\mathbb{P}(\Lambda_j)}\right) \mathbb{P}(\Lambda_j) = \sum_{j=0}^{\infty} Qr_j(\Lambda)\,\mathbb{P}(\Lambda_j)$$

Qr_j can be thought of as ergodic building blocks. They obey the law of large numbers in the usual sense.

▷ $Qr_j(\Lambda_j) = 1$ and $Qr_j(\Lambda_k) = 0, j \neq k$.

▷ If $\mathbb{P}$ and $\mathbb{S}$ is measure-preserving, $(Qr_j, \mathbb{S})$ is measure-preserving and ergodic.

2.4 Ergodic Decomposition

▷ Start with Qr_j s and pick a prior distribution π_js and construct

$$\mathbb{P}(\Lambda) = \sum_{j=0}^{\infty} Qr_j(\Lambda)\,\pi_j$$

and $\mathbb{P}$ is measure-preserving.

2.5 Example

▷ Consider the VAR(1) of the form:

$$X_{t+1} = AX_t + BW_{t+1}, \quad W_{t+1} \sim_{iid} N(0,1)$$

and A is stable i.e. eigenvalues have absolute values strictly less than 1.

* *Mean:* Taking the expectation on both sides, the moments have to satisfy:

$$\mathbb{E}[X_{t+1}] = A\mathbb{E}[X_t]$$

and for stationarity, it must be that $\mathbb{E}[X_t] = \mathbf{0}$.

* *Covariance:* Denoting it as Σ_t, we have

$$\Sigma_{t+1} = A\Sigma_t A' + BB'$$

Solving this using guess-and-verify:

$$\Sigma = \Sigma_{j=0}^{\infty} A^j BB' \left(A'\right)^j$$

which is well-defined since A is stable.

* *Stationarity*: We have stationarity:

$$X_t \sim N(0, \Sigma)$$

* *Ergodic*: We have ergodicity here since we are conditioning on the fact that we know A and B.

▷ Consider:

$$X_t = \begin{bmatrix} X_t^1 \\ X_t^2 \end{bmatrix}, \quad A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}$$

so X_t^2 is invariant for all t and suppose A_{11} is stable.

* *Mean*: Taking the expectation, we have

$$\mathbb{E}\left[X_{t+1}^1\right] = A_{11}\mathbb{E}\left[X_t^1\right] + A_{12}\mathbb{E}\left[X_t^2\right] \Rightarrow (I - A_{11})\,\mathbb{E}\left[X_t^1\right] = A_{12}\mathbb{E}\left[X_t^2\right]$$

which yields

$$\mathbb{E}\left[X_t^1\right] = (I - A_{11})^{-1} A_{12}\mathbb{E}\left[X_t^2\right]$$

* Takeaway: X_t^2 is invariant, so depending on how you specify X_t^2, you get a different behavior for the mean of the process. X_t^2 is the invariant event; so conditioning on X_t^2 implies choosing different invariant sets.
* It's only ergodic only if I announce that X_t^2 is some specified number. The ergodic building block is precisely this specification of X_t^2 that allows us to achieve ergodicity.

3 Markov Processes

3.1 Finite-state Markov Chain

▷ Denote $\mathbb{P}$ as a the transition matrix and p_{ij} as the transition probability and the state i is represented by a vector of zeros with only one at position i, u_i . Denote X_t as the realization coordinate vectors.

▷ Consider a function of X_t, $f_{n\times 1}$:

$$\mathbb{E}\left[f\left(X_{t+1}\right)|X_t = u\right] = u'\mathbb{P}f$$

or

$$\mathbb{E}\left[f\left(X_{t+1}\right)|X_t\right] = \mathbb{P}f$$

▷ Suppose we can find a solution to $\mathbb{P}f = f$ i.e.

$$\mathbb{E}\left[f\left(X_{t+1}\right)|X_t\right] = f\left(X_t\right)$$

then $\{X_t\}$ is stationary.

* To see this, it suffices to show that given

$$f\left(X_{t+1}\right) = \underbrace{\mathbb{E}\left[f\left(X_{t+1}\right)|X_t\right]}_{=f(X_t)} + \epsilon_{t+1}$$

the variance of ϵ_{t+1} is equal to zero i.e. $\mathbb{E}\left[\epsilon_{t+1}^2\right] = 0$

* Taking the expectation in the following manner:

$$\begin{aligned}
&\mathbb{E}\left[\left(f\left(X_{t+1}\right) - \mathbb{E}\left[f\left(X_{t+1}\right)|X_t\right]\right)^2\right] \\
=&\mathbb{E}\left[f\left(X_{t+1}\right)^2\right] - 2\mathbb{E}\left[f\left(X_{t+1}\right)\mathbb{E}\left[f\left(X_{t+1}\right)|X_t\right]\right] + \mathbb{E}\left[\mathbb{E}\left[f\left(X_{t+1}\right)|X_t\right]^2\right] \\
=&\mathbb{E}\left[f\left(X_{t+1}\right)^2\right] - 2\mathbb{E}\left[f\left(X_{t+1}\right)f\left(X_t\right)\right] + \mathbb{E}\left[f\left(X_t\right)^2\right] \\
=&\mathbb{E}\left[f\left(X_{t+1}\right)^2\right] - 2\mathbb{E}\left[\mathbb{E}\left[f\left(X_{t+1}\right)f\left(X_t\right)|X_t\right]\right] + \mathbb{E}\left[f\left(X_t\right)^2\right] \\
=&\mathbb{E}\left[f\left(X_{t+1}\right)^2\right] - 2\mathbb{E}\left[f\left(X_t\right)\underbrace{\mathbb{E}\left[f\left(X_{t+1}\right)|X_t\right]}_{:=f(X_t)}\right] + \mathbb{E}\left[f\left(X_t\right)^2\right] \\
=&\mathbf{0}
\end{aligned}$$

4 Additive Functionals

4.1 Setup

▷ Martingale: $\{Y_t\}$ is a martingale if $\mathbb{E}[Y_{t+1}|\mathcal{F}_t] = Y_t$. If $Y_{t+1} - Y_t := X_{t+1}$, then $\mathbb{E}[X_{t+1}|\mathcal{F}_t] = 0$.

▷ Additive Functional: A process $\{Y_t\}$ is said to be an *additive functional* if it can be represented as

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

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

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

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

▷ Additive Martingale: An additive functional $\{Y_t\}$ is said to be an *additive Martingale* if

$$\mathbb{E}[\kappa(X_t, W_{t+1})|X_t] = 0.$$

* A linear combination of additive functionals is an additive functional.

▷ Multiplicative Functional: Let $\{Y_t\}$ be an additive functional that is described by $Y_{t+1} - Y_t = \kappa(X_t, W_{t+1})$. Then, we say that

$$\{M_t\} = \{\exp[Y_t]\}$$

is a *multiplicative functional* parameterised by κ.

▷ Multiplicative Martingale: A multiplicative functional $\{M_t\}$ is said to be an *multiplicative Martingale* if

$$\mathbb{E}[M_{t+1}|\mathfrak{F}_t] = M_t.$$

4.2 Central Limit Theory

Billingsley (1961) proved the central limit theory result related to additive martingales without appealing to iid. Gordin (1969) then extended Billingsley's result to allow for temporally dependent increments.

Theorem 4.1. *(Billingsley) Let $\{Y_t\}_{t=0}^{\infty}$ be an additive martingale process whose increments $Y_{t+1} - Y_t$ are stationary, ergodic, martingale differences:*

$$\mathbb{E}[Y_{t+1} - Y_t|\mathfrak{F}_t] = 0.$$

Then,[a]

$$\frac{1}{\sqrt{t}} Y_t \xrightarrow{\text{d}} N\left(0, \mathbb{E}\left[(Y_{t+1} - Y_t)^2\right]\right).$$

Corollary 4.1. *(Gordin) Suppose that $\{Y_t\}$ is an additive functional, that $\mathbb{T}^m$ is a strong contraction on $\mathcal{N}$ for some m, and that $\mathbb{E}[(\kappa(X_t, W_{t+1}))^2] < \infty$. (basically the assumptions for martingale decomposition to work) and that $\nu = 0$. Then*

$$\frac{1}{\sqrt{t}} Y_t \xrightarrow{\text{d}} N\left(0, \sigma^2\right),$$

where

$$\sigma^2 = \mathbb{E}\left[(\kappa_a(X_t, W_{t+1}))^2\right].$$

[a] Ergodicity can be dispensed with if we replace the variance by $\mathbb{E}\left[(Y_{t+1} - Y_t)^2 |\mathfrak{I}\right]$.

The variance formula

$$\sigma^2 = \lim_{t\to\infty} \frac{1}{t}\text{Var}\left[Y_t\right] = \mathbb{E}\left[\left(\kappa_a\left(X_t, W_{t+1}\right)\right)^2\right]$$

shows how to take into account the temporal dependence of the increments $(Y_{t+1} - Y_t)$ when computing the "long-run" volatility of the level Y_t. All that matters is the martingale component, and not the stationary $g\left(X_t\right)$ component.

▷ Observation #1: In general,

$$\mathbb{E}\left[m_t^2\right] \neq Var\left[x_t\right]$$

▷ Observation #2: we have

$$Var\left[\frac{1}{\sqrt{t}}\sum_{i=1}^{t}\left(x_i - \eta\right)\right] \to Var\left[m_t\right]$$

* In practice, how should we compute the LHS? If we have a time series, we only obtain one observation of the term inside the bracket.
* Alternatively, to estimate $Var\left[m_t\right]$, we can compute the spectrual density a frequency zero of tthe process $\{X_t - \eta\}$

▷ Observation #3: Setting $y_0 = 0$, we have

$$Y_t = \sum_{i=1}^{t} X_i$$

then

$$\frac{1}{t}Y_t \to \eta$$
$$\frac{1}{\sqrt{t}}\left(Y_t - t\eta\right) \to N\left(0, \mathbb{E}\left[m_t^2\right]\right)$$

or alternatively:

$$\sqrt{t}\left(\frac{1}{t}Y_t - \eta\right) \to N\left(0, \mathbb{E}\left[m_t^2\right]\right)$$

4.2.1 Cointegration

Consider

$$Y_t = r_1 Y_t^1 + r_2 Y_t^2$$

where $Y_t^1 = \eta_1\left\{m_t^1\right\}$ and $Y_t^2 = \eta_2\left\{m_t^2\right\}$. Then

$$m = r_1 m_1 + r_2 m_2$$

▷ Y^1 and Y^2 are cointegrated if there exists r_1, r_2 different from zero such that

$$r_1 m_t^1 + r_2 m_t^2 = 0$$

and then $[r_1, r_2]^T$ is the cointegrating vector. $[1, -1]^T$ was the cointegrating vector in the permanent income example in-class.

▷ Thinking about co-integration is like thinking about balanced growth path.

4.3 Martingale Decomposition Galore

4.3.1 Martingale decomposition: Simple Version

Example: Suppose you have $Y_{t+1} - Y_t = X_{t+1}$ where $\{X_t\}$ is stationary and ergodic. Then $Y_t = Y_0 + \sum_{j=1}^{t} X_j$.

▷ Denoting $\nu = \mathbb{E}[X_t]$, define $\bar{X}_t = \sum_{j=0}^{\infty} \mathbb{E}[X_{t+j} - \nu | \mathcal{F}_t]$. We want this to be well-defined. (not obvious yet)

* Define

$$\bar{X}_t = \sum_{j=0}^{\infty} \mathbb{E}[X_{t+j} - \nu | \mathcal{F}_t], \quad \tilde{X}_t = \sum_{j=1}^{\infty} \mathbb{E}[X_{t+j} - \nu | \mathcal{F}_t]$$

▷ Observation 1: $\bar{X}_t - \tilde{X}_t = X_t - \nu$

▷ Observation 2: $\mathbb{E}\left[\bar{X}_{t+1} | \mathcal{F}_t\right] = \tilde{X}_t$. This can be shown by law of iterated expectations:

$$\mathbb{E}\left[\bar{X}_{t+1} | \mathcal{F}_t\right] = \mathbb{E}\left[\sum_{j=0}^{\infty} \mathbb{E}[X_{t+1+j} - \nu | \mathcal{F}_{t+1}] | \mathcal{F}_t\right]$$

$$= \sum_{j=0}^{\infty} \mathbb{E}\left[\mathbb{E}[X_{t+1+j} - \nu | \mathcal{F}_{t+1}] | \mathcal{F}_t\right]$$

$$= \sum_{j=1}^{\infty} \mathbb{E}\left[\mathbb{E}[X_{t+j} - \nu | \mathcal{F}_{t+1}] | \mathcal{F}_t\right]$$

Since $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$, we have

$$= \sum_{j=1}^{\infty} \mathbb{E}[X_{t+j} - \nu | \mathcal{F}_t] := \tilde{X}_t$$

▷ Observation 3:

$$\begin{aligned} X_{t+1} &= X_{t+1} - \nu + \nu = \bar{X}_{t+1} - \tilde{X}_{t+1} + \nu \\ &= \bar{X}_{t+1} - \tilde{X}_t + \left(\tilde{X}_t - \tilde{X}_{t+1}\right) + \nu \\ &= M_{t+1} + \left(\tilde{X}_t - \tilde{X}_{t+1}\right) + \nu \end{aligned}$$

where $\mathbb{E}[M_{t+1} | \mathcal{F}_t] = 0$ since

$$\begin{aligned} \mathbb{E}[M_{t+1} | \mathcal{F}_t] &= \mathbb{E}\left[\bar{X}_{t+1} - \tilde{X}_t | \mathcal{F}_t\right] \\ &= \mathbb{E}\left[\bar{X}_{t+1} | \mathcal{F}_t\right] - \mathbb{E}\left[\tilde{X}_t | \mathcal{F}_t\right] \\ &= \tilde{X}_t - \tilde{X}_t = 0 \end{aligned}$$

▷ Observation 4: Using the result from observation 3, we have

$$\begin{aligned} Y_t &= Y_0 + \sum_{j=1}^{t} X_j \\ &= Y_0 + \sum_{j=1}^{t} \left\{ \nu + M_j + \left(\tilde{X}_{j-1} - \tilde{X}_j \right) \right\} \\ &= Y_0 + \nu t + \sum_{j=1}^{t} M_j + \tilde{X}_0 - \tilde{X}_t \end{aligned}$$

where $Y_0 + \tilde{X}_0$ is invariant; νt is the time trend; $\sum_j M_j$ is a martingale (the whole thing); and $\tilde{X}_t$ is stationary.

▷ Note that M_j s are uncorrelated, which allows us to write:

$$Var\left[\sum_{j=1}^{t} M_j \right] = t\mathbb{E}\left[M_t^2 \right]$$

which is growing linear in time. To see why it's uncorrelated, consider $k > j$ then

$$\mathbb{E}\left[M_j M_k \right] = \mathbb{E}\left[\mathbb{E}\left[M_k M_j | \mathcal{F}_{k-1} \right] \right] = \mathbb{E}\left[M_j \mathbb{E}\left[M_k | \mathcal{F}_{k-1} \right] \right] = 0$$

4.3.2 Martingale decomposition: Markov's Version

Denote $\{X_t\}$ to be Markov process and $Y_{t+1} - Y_t = \phi(X_t)$ where $\phi(X_t)$ has finite second moment.

1. $\mathbb{T}$ as a contraction:

   ▷ Since we can characterize a Markov process with a transition matrix, write

   $$\mathbb{T}\phi(X) = \mathbb{E}\left[\phi(X_{t+1}) | X_t = X \right]$$

   which maps a function into a function. Note that we can go back and forth between the transition probabilities and the conditional expectations.

   ▷ Consider writing

   $$\phi(X_{t+1}) = \mathbb{E}\left[\phi(X_{t+1}) | X_t \right] + \epsilon_{t+1}$$

   which, by the least-squares theory, implies

   $$\mathbb{E}\left[\phi(X_{t+1})^2 \right] \geq \mathbb{E}\left[\left\{ \mathbb{T}\phi(X_t) \right\}^2 \right]$$

   and thus $\|\mathbb{T}\phi\| \leq \|\phi\|$ which implies that $\mathbb{T}$ is a "weak" contraction. (We can't make it a strong contraction because the conditional expectation can be a constant).

   ▷ Define $\mathcal{Z}_t = \left\{ \phi \in \mathcal{L}^2 | E\left[\phi(X) \right] = \mathbf{0} \right\}$ and assume a restriction that $\mathbb{T}$ is a strong contraction on $\mathcal{Z}_t$ i.e. $\|\mathbb{T}\phi\| \leq \lambda\|\phi\|, \lambda \in (0,1)$. This is not always satisfied, so we have to assume this.

   ▷ Suppose $\sum_{j=0}^{\infty} \mathbb{E}\left[\phi(X_{t+j}) | X_t \right] = 0$ is mean zero. Then:

   $$\sum_{j=0}^{\infty} \mathbb{E}\left[\phi(X_{t+j}) | X_t = X \right] = \sum_{j=0}^{\infty} \mathbb{T}^j \phi(X)$$

since LIE gives us:

$$\begin{aligned}\mathbb{E}\left[\phi\left(X_{t+2}\right)|X_t\right] &= \mathbb{E}\left[\mathbb{E}\left[\phi\left(X_{t+2}\right)|X_{t+1},X_t\right]|X_t\right] \quad (\because)\ \{X_t\}\subseteq\{X_{t+1},X_t\}\\ &= \mathbb{E}\left[\mathbb{E}\left[\phi\left(X_{t+2}\right)|X_{t+1}\right]|X_t\right] \quad (\because)\ \text{markov property}\\ &= \mathbb{T}\left[\mathbb{E}\left[\phi\left(X_{t+2}\right)|X_{t+1}\right]\right]\\ &= \mathbb{T}\left[\mathbb{T}\phi\left(X_{t+1}\right)\right](X)\\ &= \mathbb{T}^2\phi\left(X\right)\end{aligned}$$

and the infinite sum above is well-defined and a strong contraction.

▷ This allows us to write:

$$(I-\mathbb{T})^{-1}=\sum_{j=0}^{\infty}\mathbb{T}^j$$

4.3.3 Martingale decomposition: VAR Example & Permanent Shocks

Setup is $Y_{t+1}-Y_t=DX_t+FW_{t+1}+\nu$ where $X_{t+1}=AX_t+BW_{t+1}$, A is stable, and $W_{t+1}\sim N\left(O,I\right)$. Furthermore, let

$$H_{t+1}=DX_t+FW_{t+1}$$

Then

$$\bar{H}_t:=\mathbb{E}\left[\left\{\sum_{i=0}^{\infty}H_{t+j}\right\}|\mathcal{F}_t\right]=\cdots=DX_{t-1}+FW_T+D\left[\left(I-A\right)^{-1}X_t\right]$$

Computing the error term:

$$\begin{aligned}\bar{H}_{t+1}-\mathbb{E}\left[\bar{H}_{t+1}|\mathcal{F}_t\right] &= DX_t+FW_{t+1}+D\left[\left(I-A\right)^{-1}X_{t+1}\right]-DX_t-D\left[\left(I-A\right)^{-1}AX_t\right]\\ &= FW_{t+1}+D\left[\left(I-A\right)^{-1}BW_{t+1}\right]\end{aligned}$$

Going back to the VAR setup:

$$\left[X_{t+1}\right]_{n\times1}=A_{n\times n}\left[X_t\right]_{n\times1}+B_{n\times k}\left[W_{t+1}\right]_{m\times1}$$

where

$$B=\begin{bmatrix}\left[B_1\right]_{n\times m}\\ O\end{bmatrix}$$

We want to infer BB'. To do this, let Q such that $QQ'=I$. then

$$B_1W_{t+1}=B_1QQ'W_{t+1}=\tilde{B}_1\tilde{W}_{t+1}\Rightarrow\tilde{B}_1=B_1Q,\tilde{W}_{t+1}=QW_{t+1}\sim N\left(O,I\right)$$

Blanchard-Quah decomposition uses a permanent shock (in their paper this is the supply-side) to identify the shock of interest.

4.3.4 The Algorithm

We summarise the algorithm for extracting martingales below.

Algorithm 1. *(Extracting martingales from additive functionals) The additive functional of interest is:*

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

▷ *Step 1: Construct κ_2.*

$$\begin{aligned}\kappa_2(X_t, W_{t+1}) &\coloneqq \kappa(X_t, W_{t+1}) - \mathbb{E}[\kappa(X_t, W_{t+1}) | X_t] \\ &\equiv \kappa(X_t, W_{t+1}) - \bar{f}(X_t)\end{aligned}$$

▷ *Step 2: Construct $f(x)$.*

$$\begin{aligned}f(x) &\coloneqq \mathbb{E}[\kappa(X_t, W_{t+1}) | X_t = x] - \mathbb{E}[\kappa(X_t, W_{t+1})] \\ &\equiv \bar{f}(x) - \nu.\end{aligned}$$

▷ *Step 3: Construct κ_1 (assuming that $\mathbb{T}$ is ρ-mixing).*

$$\begin{aligned}\kappa_1(X_t, W_{t+1}) &= g(X_{t+1}) - g(X_t) + f(X_t), \\ g(X_t) &\coloneqq (\mathbb{I} - \mathbb{T})^{-1} f(X_t) = \sum_{j=0}^{\infty} \mathbb{T}^j f(X_t).\end{aligned}$$

▷ *Step 4: Decompose κ.*

$$\kappa(X_t, W_{t+1}) = \kappa_1(X_t, W_{t+1}) + \kappa_2(X_t, W_{t+1}) + g(X_t) - g(X_{t+1}) + \nu.$$

▷ *Step 5: Substitute into Y_t.*

$$Y_t = \underbrace{Y_0 + g(X_0)}_{\textit{Invariant}} + \underbrace{t\nu}_{\textit{Time trend}} - \underbrace{g(X_t)}_{\textit{Stationary}} + \underbrace{\sum_{j=1}^{t} \kappa_a(X_{j-1}, W_j)}_{\textit{Martingale}}.$$

4.4 TA Session: Permanent vs Transient Shocks

We illustrate them using the VAR example:

$$\begin{aligned}X_{t+1} &= AX_t + BW_{t+1} \\ Y_{t+1} - Y_t &= \nu + D^T X_t + F^T W_{t+1}\end{aligned}$$

where A is stable and $X_0 = 0$. X_{t+1} is a $k \times 1$ vector; Y_t is a scalar.

▷ In class, we saw the following decomposition:

$$Y_t = Y_0 + \nu t + \underbrace{\sum_{j=1}^{t} \left(F^T + D^T (I - A)^{-1} B\right) W_j}_{\text{= martingale component}} - D^T (I - A)^{-1} X_t$$

▷ Using the lag operator, rewrite the process as

$$(I - AL) X_t = BW_t$$

and plug this back into the decomposition:

$$Y_t = Y_0 + \nu t + \sum_{j=1}^{t} \left(F^T + D^T (I - A)^{-1} B \right) W_j - D^T (I - A)^{-1} \underbrace{(I - AL)^{-1} BW_t}_{(1)}$$

Note that (1) is

$$\left(I + AL + A^2 L^2 + \cdots \right) BW_t = BW_t + ABW_{t-1} + A^2 BW_{t-2} + \cdots$$

so we have

$$Y_t = Y_0 + \nu t + \sum_{j=1}^{t} \left(F^T + D^T (I - A)^{-1} B \right) W_j - D^T (I - A)^{-1} \sum_{j=0}^{t-1} A^j BW_{t-j}$$

▷ Assume $W_1 = 0$ and $W_t = 0, \forall t \geq 2$. Then

$$Y_t = Y_0 + \nu t + \underbrace{\left(F^T + D^T (I - A)^{-1} B \right) W_1}_{\text{permanent part}} - \underbrace{D^T (I - A)^{-1} A^{t-1} BW_1}_{\text{transient part}}$$

The transiency comes from the decay that comes from A^{t-1}.

* Permanent part is just a linear combination of W_1.
* If W_1 is *orthogonal* to $F^T + D^T (I - A)^{-1} B$, we call it the *transitory* shock.
* If W_1 is *parallel* to $F^T + D^T (I - A)^{-1} B$, we call it the *permanent* shock.

▷ For a transitory impulse response, the impulse should just decay exponentially.

▷ For all shocks that are not transitory, the impulse response will converge to

▷ The permanent shock can be found by fixing the magnitude of the impulse response and finding the W_1 that gives the largest steady-state.

* Even if the shock is entirely parallel, you still have a convergence over time to the steady state.

4.5 TA Session: Small Shock Approximation a la Lombardo and Uhlig (2018)

4.5.1 Setup

Consider the following setup:

$$X_{t+1} = AX_t + BW_{t+1}$$
$$y_{t+1} - y_t = \nu + D^T X_t + F^T W_{t+1}$$

where $y_t := \log Y_t$. We will illustrate how the previous technique can be applied in a macroeconomic framework.

▷ Technology: transfer 1 unit time t good to $\exp(\rho)$ units of time $t + 1$ good. The $\{Y_t\}$s are the fruits.

▷ The feasibility constraint:

$$K_{t+1} + C_t = \exp(\rho) K_t + Y_t$$

This is equivalent to Prof. Stokey's formulation:

$$K_{t+1} + C_t = A_t K_t + (1 - \delta) K_t$$

▷ The RA maximizes

$$\mathbb{E}\left[\sum_{j=0}^{\infty} \exp(-\delta t) \log C_t\right]$$

▷ The Euler Equation:

$$U'(C_t) = \mathbb{E}\left[\exp(\rho) \exp(-\delta) u'(C_{t+1}) | X_t\right]$$

which yields

$$1 = \mathbb{E}\left[\exp(-\delta + \rho) \frac{C_t}{C_{t+1}} | X_t\right]$$

4.5.2 Solving for Response in Consumption

We want to linearize the $\mathbb{E}[\cdot]$.

▷ To solve this, define

$$\hat{K}_t = \frac{K_t}{Y_t}, \qquad \hat{C}_t = \log C_t - \log Y_t$$

and re-write the feasibility constraint:

$$\hat{K}_{t+1} \exp(\log Y_{t+1} - \log Y_t) + \exp\left(\hat{C}_t\right) - \exp(\rho) \hat{K}_t - 1 = 0$$

and the Euler equation:

$$\exp(-\delta + \rho) \mathbb{E}\left[\exp\left(-\left(\hat{C}_{t+1} - \hat{C}_t\right) - (\log Y_{t+1} - \log Y_t)\right) | X_t\right] - 1 = 0$$

Essentially, we are deterending to get the stationary distribution.

▷ Now consider perturbing the system by q i.e. changing the exposure of Y_ts to the stochastic component:

$$X_{t+1} = AX_t + BW_{t+1}$$
$$y_{t+1}(q) - y_t(q) = \nu + \left[D^T X_t + F^T W_{t+1}\right] q$$

This allows us to reformulate the previous variables as a function of q

$$[1] : \hat{K}_{t+1}(q) \exp(\log Y_{t+1}(q) - \log Y_t(q)) + \exp\left(\hat{C}_t(q)\right) - \exp(\rho) \hat{K}_t(q) - 1 = 0$$

$$[2] : \exp(-\delta + \rho) \mathbb{E}\left[\exp\left(-\left(\hat{C}_{t+1}(q) - \hat{C}_t(q)\right) - (\log Y_{t+1}(q) - \log Y_t(q))\right) | X_t\right] - 1 = 0$$

▷ Consider a Taylor expansion around 0:

$$\hat{C}_t(q) \approx \hat{C}_t(0) + \hat{C}'_t(q)$$
$$\hat{K}_{t+1}(q) \approx \hat{K}_{t+1}(0) + \hat{K}'_{t+1}(q)$$

The reason we do it around 0 is becaues $\hat{C}_t(0)$ results in a deterministic fruits process $\{Y_t\}$ and thus it is very easy to compute.

▷ Note that the individual components above are processes, not numbers. Similarly to the macro class, we want to do

$$\hat{C}_t = C\left(X_t, \hat{K}_t\right), \quad \hat{K}_{t+1} = K\left(X_t, \hat{K}_t\right)$$

where X_t is the exogenous state and K_t is the endogenous state.

Now define a new function

$$F_1\left(\hat{K}_{t+1}(q), \hat{C}_t(q), \hat{K}_t(q), \Delta y_{t+1}(q)\right) \equiv [1]$$

where $\Delta y_{t+1}(q) = y_{t+1}(q) - y_t(q)$

▷ Since F_1 is equal to zero for all q, we have

$$F_1(q) \approx F_1|_{q=0} + q\frac{\partial F_1}{\partial q}|_{q=0} \approx 0$$

as well. Similar argument holds for $F_2 = 0$:

$$F_2(q) \approx F_2|_{q=0} + q\frac{\partial F_2}{\partial q}|_{q=0} \approx 0$$

▷ Obtaining $F_1|_{q=0}$: Making the assumption that $\delta = \rho - \nu$, the economy has a steady state of:

$$\hat{C}_t(0) = 0, \qquad \hat{K}_{t+1}(0) = 0$$

which yields

$$\log C_t - \log Y_t = 0 \Rightarrow C_t = Y_t$$
$$\hat{K}_{t+1}(0) = \frac{K_{t+1}}{Y_{t+1}} = 0$$

so you consume fruit everyday and save nothing.

▷ Obtaining $\partial F_1/\partial q$:

$$\frac{\partial F_1}{\partial q}|_{q=0} = \hat{C}_t'(0)\frac{\partial F_1}{\partial \hat{C}_t}|_{q=0} + \hat{K}_{t+1}'(0)\frac{\partial F_1}{\partial \hat{K}_{t+1}}|_{q=0} + \hat{K}_t'(0)\frac{\partial F_1}{\partial \hat{K}_t}|_{q=0} + \Delta y_{t+1}'(0)\frac{\partial F_1}{\partial \Delta y_{t+1}}|_{q=0} = 0$$

* Note that we already know

$$\frac{\partial F_1}{\partial \hat{C}_t}\left(\hat{K}_{t+1}(q), \hat{C}_t(q), \hat{K}_t(q), \Delta y_{t+1}(q)\right)|_{q=0}$$

since we've computed the relevant quantities evaluated at zero in the previous step. The similar argument follows for the other derivatives.

* Deriving analogously for F_2, we have two linear equations of $\hat{K}_{t+1}'(0)$ and $\hat{C}_t'(0)$.
* The term with $\Delta y_{t+1}'$ is good since

$$\frac{\partial F_1}{\partial \Delta y_{t+1}}|_{q=0} = 0$$

and the term vanishes.

* The term with $\hat{K}_t'(0)$ is also good since the derivative is simply $\exp(\rho)$.

Going through a similar process with F_2, we obtain:

$$[3] : \hat{K}^1_{t+1} \exp(\nu) + \hat{C}^1_t - \exp(\rho)\hat{K}'_t = 0$$
$$[4] : \mathbb{E}\left[\left(\hat{C}^1_{t+1} - \hat{C}^1_t\right) + \Delta y'_{t+1}\right] = 0$$

To solve this, guess and verify:

$$\hat{C}^1_t := C'_t(q=0) = MX_t + \Gamma_K \hat{K}^1_t$$

- ▷ Plug this into $[3]$ and express $\hat{K}_{t+1}$ as a function of $\hat{K}_t$ and $\hat{X}_t$.
- ▷ Plug this into $[4]$ and replace X_{t+1} as a function of X_t and solve for M and Γ_K.

The resulting solution is

$$\hat{C}'_t = \lambda D'(I - \lambda A)^{-1} X_t + \{\exp(\rho)\}$$
$$\hat{K}'_{t+1} = \hat{K}'_t - \exp(-\nu)\lambda D'(I - \lambda A)^{-1} X_t$$

where $\lambda = \exp(v - \rho)$. We also obtain

$$\hat{C}^1_{t+1} - \hat{C}^1_t = -D^T X_T + \lambda D^T (I - \lambda A)^{-1} B W_{t+1}$$

This allows us to compute the log consumption progress:

$$\log C_{t+1} - \log C_t = -D^T X_t + \lambda D^T (1 - \lambda A)^{-1} B W_{t+1} + \nu + D^T X_t + F^T W_{t+1}$$
$$= \left(\lambda D^T (I - \lambda A)^{-1} B + F^T\right) W_{t+1} + \nu$$

This is the function that Professor Hansen plotted in class. He provided two plots – one is the permanent shock to the $\log Y_t$ process and one is the transitory shock to the $\log Y_t$ process.

- ▷ If λ is really close to one, then the permanent shock which is parallel to

  $$D^T (I - \lambda A)^{-1} B + F^T$$

  will be almost parallel to

  $$\lambda D^T (I - \lambda A)^{-1} B + F^T$$

  in which case the permanent shock will have a big impact on consumption.
- ▷ There is no D which is why you see the impulse response as a constant. This is constrast with the general response with the long-term convergence (adjustment takes time due to the transitory part, which was $D[\cdots]$). If $D = 0$, then it will be just a straight line.

4.5.3 Comparison with Log-linearization

If you want to attain the second order in log-linearization, you have to deal with

$$\hat{X}_{t+1} = a\hat{X}_t + b\hat{X}^2_t + \cdots$$

which fucks with stationarity. But for this methodology, even if you go to second order:

$$\hat{X}_{t+1} = \hat{X}_{t+1}(0) + [\cdots]\hat{X}_t{}'(0) + [\cdots]\hat{X}''_t(0)$$

you are taking the square of the derivative so it's okay.

5 Likelihood Processes (Ch. 8)

5.1 Likelihood Constructions

Let $\{W_t\}$ be a process of shocks satisfying $\mathbb{E}[W_{t+1}|\mathcal{F}_t] = 0$. Let $\{X_t\}$ be a Markov process such that $X_{t+1} = \phi(X_t, W_{t+1})$. We also have data that we observe, denoted as Y_t, such that

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

We make the following assumptions:

- ▷ There exists a function χ such that $W_{t+1} = \chi(X_t, Y_{t+1} - Y_t)$
- ▷ $Y_{t+1} - Y_t$ has a density $\psi(\cdot|x)$ with respect to measure τ conditioned on $X_t = x$.

The example below illustrates how to construct χ and ψ.

Example 5.1. (8.1.4 of Lecture Notes) Suppose $X_{t+1} = AX_t + BW_{t+1}$ and $Y_{t+1} - Y_t = DX_t + FW_{t+1}$ where A is a stable matrix, $\{W_{t+1}\}_{t=0}^{\infty}$ is an i.i.d. sequence with mean 0 and covariance I. Also assume F is non-singular. Then applying F^{-1} to both sides:

$$W_{t+1} = F^{-1}(Y_{t+1} - Y_t) - F^{-1}DX_t := \chi(X_t, Y_{t+1} - Y_t)$$

And the density is then

$$Y_{t+1} - Y_t \sim N\left(DX_t, FF'\right) := \psi(\cdot|x)$$

The likelihood function conditioned on X_0:

$$L_t = \prod_{j=1}^{t} \psi(Y_j - Y_{j-1}|X_{j-1})$$

$$\Rightarrow \ell_t = \sum_{j=1}^{t} \log\left(\psi(Y_j - Y_{j-1}|X_{j-1})\right)$$

and $\log L_t := \ell_t$ are stationary increments. Both of these are stochastic processes.

5.2 Likelihood Ratios

We show that the likelihood ratio process is a positive martingale.

1. Construct $L_t(\theta)$ and $L_t(\theta_0)$ recursively as above
2. Decompose the likelihood in the following manner:

$$\frac{L_{t+1}(\theta)}{L_{t+1}(\theta_0)} = \frac{L_t(\theta)}{L_t(\theta_0)} \cdot \left[\frac{\psi(Y_{t+1} - Y_t|X_t, \theta)}{\psi(Y_{t+1} - Y_t|X_t, \theta_0)}\right]$$

3. Taking the expectation of the LHS:

$$\begin{aligned}
\mathbb{E}\left[\frac{L_{t+1}(\theta)}{L_{t+1}(\theta_0)}|X_t = x, \theta_0\right] &= \frac{L_t(\theta)}{L_t(\theta_0)}\mathbb{E}\left[\frac{\psi(Y_{t+1} - Y_t|X_t, \theta)}{\psi(Y_{t+1} - Y_t|X_t, \theta_0)}|X_t = x, \theta_0\right] \\
&= \frac{L_t(\theta)}{L_t(\theta_0)} \int_Y \frac{\psi(y^*|x, \theta)}{\psi(y^*|x, \theta_0)} \psi(y^*|x, \theta_0)\, \tau(dy) \\
&= \frac{L_t(\theta)}{L_t(\theta_0)} \int_Y \psi(y^*|x, \theta)\, \tau(dy) \\
&= \frac{L_t(\theta)}{L_t(\theta_0)}
\end{aligned}$$

and thus we verify that the likelihood ratio process is a multiplicative (positive) martingale under the θ_0 probability model.

Next we show that the log likelihood ratio process is a super-martingale. This is easy since from Jensen's inequality, we have

$$\mathbb{E}\left[\log \frac{L_{t+1}(\theta)}{L_{t+1}(\theta)} | X_t = x, \theta_0\right] \leq \log\left[\frac{L_t(\theta)}{L_t(\theta_0)}\right]$$

Now using the above result, we have define $v(\theta)$ as

$$\begin{aligned} v(\theta) &:= \mathbb{E}\left[\log \psi(Y_{t+1} - Y_t | X_t, \theta) - \log \psi(Y_{t+1} - Y_t | X_t, \theta_0) | \theta_0\right] \\ &= \mathbb{E}\left[\log \frac{L_{t+1}(\theta)}{L_{t+1}(\theta_0)} - \log \frac{L_t(\theta)}{L_t(\theta_0)} | \theta_0\right] \\ &= \mathbb{E}\left[\underbrace{\mathbb{E}\left[\log \frac{L_{t+1}(\theta)}{L_{t+1}(\theta_0)} - \log \frac{L_t(\theta)}{L_t(\theta_0)} | X_t = x, \theta_0\right]}_{\leq 0} | \theta_0\right] \leq 0 \\ \Rightarrow v(\theta) &= \mathbb{E}\left[\log \psi(Y_{t+1} - Y_t | X_t, \theta) - \log \psi(Y_{t+1} - Y_t | X_t, \theta_0) | \theta_0\right] \leq 0 \end{aligned}$$

since the conditional expectation is ≤ 0. Typically, it is < 0. This leads to

$$\frac{1}{t}\left[\log L_t(\theta) - \log L_t(\theta_0)\right] \to v(\theta) < 0$$

So the log likelihood process will tend to infinity, implying that the ratio of likelihood tends to zero.

▷ First, we can show that $v(\theta_0) \geq v(\theta), \forall \theta \in \Theta$.

* Starting with

$$\mathbb{E}\left[\log \frac{L_{t+1}(\theta)}{L_{t+1}(\theta)} | X_t = x, \theta_0\right] \leq \log\left[\frac{L_t(\theta)}{L_t(\theta_0)}\right]$$

* Then:

$$\begin{aligned} 0 &\geq \mathbb{E}_{\theta_o}\left[\left(\log L_t(\theta) - \log L_{t-1}(\theta)\right) - \left(\log L_t(\theta_o) - \log L_{t-1}(\theta_o)\right) | \mathfrak{F}_{t-1}\right] \\ &= \mathbb{E}_{\theta_o}\left[\log L_t(\theta) - \log L_{t-1}(\theta) | \mathfrak{F}_{t-1}\right] - \mathbb{E}\left[\log L_t(\theta_o) - \log L_{t-1}(\theta_o) | \mathfrak{F}_{t-1}\right] \\ &= \mathbb{E}_{\theta_o}\left[g(X_t|\theta) - g(X_{t-1}|\theta) + \nu(\theta) + \kappa_a(X_{t-1}, W_t|\theta) | \mathfrak{F}_{t-1}\right] \\ &\quad - \mathbb{E}_{\theta_o}\left[g(X_t|\theta_o) - g(X_{t-1}|\theta_o) + \nu(\theta_o) + \kappa_a(X_{t-1}, W_t|\theta_o) | \mathfrak{F}_{t-1}\right] \\ [*] &= \nu(\theta) - \nu(\theta_o) + \mathbb{E}_{\theta_o}\left[g(X_t|\theta) - g(X_t|\theta_o) | \mathfrak{F}_{t-1}\right] \\ \Rightarrow 0 &\geq \nu(\theta) - \nu(\theta_o), \ \forall \theta \in \Theta, \end{aligned}$$

where in step $[*]$, we the fact that conditional on $\mathfrak{F}_{t-1}$, $g(X_{t-1}|\theta) = g(X_{t-1}|\theta_o)$, and in the last step, we take unconditional expectation of both sides and use the fact that the unconditional expectation of g's are zero. Finally, recall that the original inequality were strict unless the two log-likelihoods are equal; i.e. inequality above holds with equality if and only if $\theta = \theta_o$ (with probability one).

▷ Second, the above result naturally implies that the solution to

$$\max_{\theta \in \Theta} v(\theta)$$

is $\theta = \theta_0$.

▷ Third, the solution to the above problem also solves the problem below:

$$\max_{\theta \in \Theta} \hat{v}_t(\theta) = \max_{\theta \in \Theta} \lim_{t \to \infty} \frac{1}{t} \log L_t(\theta)$$

5.3 Score Processes

Suppose you want to solve the following problem conditionally, not unconditionally:

$$\max_{\theta} \int_{y} \log\left(\psi\left(y^{*}|x,\theta\right)\right) \psi\left(y^{*}|x,\theta_{0}\right) \tau\left(dy^{*}\right) := \mathbb{E}_{\theta_{0}}\left[\log\left(\psi\left(y^{*}|x,\theta\right)\right)\right]$$

▷ Since we had this inequality from before::

$$\mathbb{E}\left[\log \frac{L_{t+1}\left(\theta\right)}{L_{t+1}\left(\theta\right)} | X_{t} = x, \theta_{0}\right] \leq \log\left[\frac{L_{t}\left(\theta\right)}{L_{t}\left(\theta_{0}\right)}\right]$$

then the parameter θ_0 necessarily solves the above maximization problem since

$$\mathbb{E}_{\theta_{0}}\left[\log\left(\psi\left(y^{*}|x,\theta\right)\right)\right] \leq \log\left(\mathbb{E}_{\theta_{0}}\left[\psi\left(y^{*}|x,\theta\right)\right]\right)$$

▷ Suppose that we can differentiate under the integral sign and get to the first-order condition (for interior solutions)

$$\begin{aligned} 0 &= \int_{\mathcal{Y}} \left[\frac{d}{d\theta} \log \psi\left(y^{*}|x,\theta_{o}\right)\right] \psi\left(y^{*}|x,\theta_{o}\right) \tau\left(dy^{*}\right) \\ &= \mathbb{E}_{\theta_{o}}\left[\frac{d}{d\theta} \log \psi\left(y^{*}|x,\theta_{o}\right) | X_{t}, \theta_{o}\right]. \end{aligned}$$

Define the score process $\{S_t\}$ as

$$S_{t} = \left.\frac{d}{d\theta} \log L_{t}\left(\theta\right)\right|_{\theta=\theta_{o}} = \sum_{j=1}^{t} \frac{d}{d\theta} \log \psi\left(Y_{j} - Y_{j-1}|X_{j-1},\theta_{o}\right) + \frac{d}{d\theta} \log L_{0}\left(\theta_{o}\right).$$

▷ Using this definition, note that

$$S_{t+1} - S_{t} = \frac{d}{d\theta} \log \psi\left(Y_{t+1} - Y_{t}|X_{t},\theta_{o}\right).$$

thus implying

$$\mathbb{E}_{\theta_{o}}\left[S_{t+1} - S_{t}|X_{t},\theta_{o}\right] = 0.$$

Thus, $\{S_t\}$ is a martingale.

▷ Property #1:

$$\frac{1}{\sqrt{t}} S_{t} \xrightarrow{\mathrm{d}} N\left(0, \mathbf{V}\right),$$

where $\mathbf{V}$ is Fisher information matrix

$$\mathbf{V} = \mathbb{E}\left[\left(S_{t+1} - S_{t}\right)\left(S_{t+1} - S_{t}\right)^{T} |\theta_{0}\right] = \mathbb{E}_{\theta_{o}}\left[\left(\frac{d}{d\theta} \log \psi\left(Y_{j} - Y_{j-1}|X_{j-1},\theta_{o}\right)\right)^{2}\right].$$

* To see this, it helps to write

$$\begin{aligned} S_{t} &= \left(S_{t} - S_{t-1}\right) + \left(S_{t-1} - S_{t-2}\right) + \cdots + \left(S_{1} - S_{0}\right) + S_{0} \\ &= S_{0} + \sum_{j=1}^{t} \left(S_{j} - S_{j-1}\right) = S_{0} + \sum_{j=1}^{t} \frac{d}{d\theta} \log \psi\left(Y_{j} - Y_{j-1}|X_{j-1},\theta_{o}\right) \end{aligned}$$

$$\Rightarrow \frac{1}{\sqrt{t}} S_{t} = \frac{1}{\sqrt{t}} S_{0} + \sqrt{t} \frac{1}{t} \sum_{j=1}^{t} \frac{d}{d\theta} \log \psi\left(Y_{j} - Y_{j-1}|X_{j-1},\theta_{o}\right).$$

▷ Property #2: With some additional regularity conditions, the following result is typical:[2]

$$\sqrt{t}\left(\theta_t^{MLE} - \theta_o\right) \xrightarrow{\text{d}} N\left(0, \mathbf{V}^{-1}\right),$$

where θ_t^{MLE} maximises the log-likelihood function $\log L_t(\theta)$.

5.4 Nuisance Parameters

Suppose there is a vector θ of unknown parameters but we are interested in information about only one of the components of the parameter vector. Call the first component of θ the "parameter of interest" θ_o, and other components the "nuisance parameters" $\vartheta_0 := \tilde{\theta}_0$.

1. Write the multivariate score process as

$$\{\mathbf{S}_{t+1}\}_{t=0}^{\infty} = \left\{ \begin{bmatrix} S_{t+1} \\ \tilde{\mathbf{S}}_{t+1} \end{bmatrix} \right\}_{t=0}^{\infty},$$

where $\{S_{t+1}\}$ is the partial derivative of the log-likelihood with respect to θ and $\left\{\tilde{\mathbf{S}}_{t+1}\right\}$ is the partial derivative with respect to $\boldsymbol{\vartheta}_o$.

2. Run a population regression

$$S_{t+1} - S_t = \boldsymbol{\beta}'\left(\tilde{\mathbf{S}}_{t+1} - \tilde{\mathbf{S}}_t\right) + U_{t+1}$$

$$\Rightarrow \begin{bmatrix} S_{t+1} - S_t \\ \tilde{\mathbf{S}}_{t+1} - \tilde{\mathbf{S}}_t \end{bmatrix} = \begin{bmatrix} 1 & \boldsymbol{\beta}' \\ \mathbf{0} & \mathbf{1} \end{bmatrix} \begin{bmatrix} U_{t+1} \\ \tilde{\mathbf{S}}_{t+1} - \tilde{\mathbf{S}}_t \end{bmatrix}.$$

[2] From Bonhomme's class. $\hat{\theta}_t$ satisfies the first-order condition:

$$\frac{1}{\sqrt{t}} S_t = \sqrt{t}\frac{1}{t}\sum_{j=1}^{t} \frac{d}{d\theta} \log\psi\left(Y_j - Y_{j-1} | X_{j-1}, \hat{\theta}_t\right) = 0.$$

Taylor expansion around θ_o yields

$$0 \approx \sum_{j=1}^{t} \frac{d}{d\theta} \log\psi\left(Y_j - Y_{j-1} | X_{j-1}, \theta_o\right) + \sum_{j=1}^{t} \frac{d}{d\theta d\theta'} \log\psi\left(Y_j - Y_{j-1} | X_{j-1}, \tilde{\theta}_t\right)\left(\hat{\theta}_t - \theta_o\right),$$

where $\tilde{\theta}_t$ lies in between $\hat{\theta}_t$ and θ_o (component by component). Rearranging yields

$$\hat{\theta}_t - \theta_o \approx -\left[\sum_{j=1}^{t} \frac{d}{d\theta d\theta'} \log\psi\left(Y_j - Y_{j-1} | X_{j-1}, \tilde{\theta}_t\right)\right]^{-1} \left(\sum_{j=1}^{t} \frac{d}{d\theta} \log\psi\left(Y_j - Y_{j-1} | X_{j-1}, \theta_o\right)\right)$$

$$\Rightarrow \sqrt{t}\left(\hat{\theta}_t - \theta_o\right) \approx -\left[\frac{1}{t}\sum_{j=1}^{t} \frac{d}{d\theta d\theta'} \log\psi\left(Y_j - Y_{j-1} | X_{j-1}, \tilde{\theta}_t\right)\right]^{-1} \left(\sqrt{t}\frac{1}{t}\sum_{j=1}^{t} \frac{d}{d\theta} \log\psi\left(Y_j - Y_{j-1} | X_{j-1}, \theta_o\right)\right)$$

$$= -\left[\frac{1}{t}\sum_{j=1}^{t} \frac{d}{d\theta d\theta'} \log\psi\left(Y_j - Y_{j-1} | X_{j-1}, \tilde{\theta}_t\right)\right]^{-1} \frac{1}{\sqrt{t}} S_t.$$

The information matrix identity (see end of this section) gives us that

$$-\mathbb{E}_{\theta_o}\left[\frac{d}{d\theta d\theta'} \log\psi\left(Y_j - Y_{j-1} | X_{j-1}, \theta_o\right)\right] = \mathbf{V}.$$

So that, in fact

$$\sqrt{t}\left(\hat{\theta}_t - \theta_o\right) \xrightarrow{\text{d}} N\left(0, \mathbf{V}^{-1}\mathbf{V}\mathbf{V}^{-1}\right) \overset{\text{d}}{=} N\left(0, \mathbf{V}^{-1}\right).$$

3. Construct the matrix

$$\begin{aligned}\hat{\mathbf{V}} &= \mathbb{E}\left[\begin{bmatrix} S_{t+1}-S_t \\ \tilde{\mathbf{S}}_{t+1}-\tilde{\mathbf{S}}_t \end{bmatrix}\begin{bmatrix} S_{t+1}-S_t & \tilde{\mathbf{S}}_{t+1}-\tilde{\mathbf{S}}_t \end{bmatrix}\right] \\ &= \underbrace{\begin{bmatrix} 1 & \boldsymbol{\beta}' \\ \mathbf{0} & \mathbf{1} \end{bmatrix}}_{=\mathbf{A}} \underbrace{\begin{bmatrix} \mathbb{E}\left[U_{t+1}^2\right] & \mathbf{0} \\ \mathbf{0} & \mathbb{E}\left[\left(\tilde{\mathbf{S}}_{t+1}-\tilde{\mathbf{S}}_t\right)\left(\tilde{\mathbf{S}}_{t+1}-\tilde{\mathbf{S}}_t\right)'\right] \end{bmatrix}}_{=\mathbf{B}} \underbrace{\begin{bmatrix} 1 & \mathbf{0} \\ \boldsymbol{\beta} & \mathbf{1} \end{bmatrix}}_{=\mathbf{A}'}\end{aligned}$$

then

$$\hat{\mathbf{V}}^{-1} = \begin{bmatrix} 1 & -\boldsymbol{\beta}' \\ \mathbf{0} & \mathbf{1} \end{bmatrix}\begin{bmatrix} \frac{1}{\mathbb{E}\left[U_{t+1}^2\right]} & \mathbf{0} \\ \mathbf{0} & \left(\mathbb{E}\left[\left(\tilde{\mathbf{S}}_{t+1}-\tilde{\mathbf{S}}_t\right)\left(\tilde{\mathbf{S}}_{t+1}-\tilde{\mathbf{S}}_t\right)'\right]\right)^{-1} \end{bmatrix}\begin{bmatrix} 1 & \mathbf{0} \\ -\boldsymbol{\beta} & \mathbf{1} \end{bmatrix}$$

and the (1,1) element, which is our main point of interest is given by

$$\hat{v}_{11}^{-1} = \frac{1}{\mathbb{E}_{\theta_0}\left[U_{t+1}^2\right]}.$$

which is referred to as the Firsher Information for estimating θ_0 with $\tilde{\theta_0}$ unknowns.

4. Notice that

$$\mathbb{E}\left[\left(S_{t+1}-S_t\right)^2\right] = Var\left[S_{t+1}-S_t\right] \geq Var\left[U_{t+1}\right] = \mathbb{E}\left[U_{t+1}^2\right] = \hat{v}_{11},$$

where $\hat{v}_{11}$ is information for θ_o

   ▷ Since the left-hand side is the information for θ_o when we know the nuisance parameters, we see that the likelihood function contains more information about θ when $\boldsymbol{\vartheta}$ is known to be $\boldsymbol{\vartheta}_o$ than when θ and $\boldsymbol{\vartheta}$ are unknown.

The punchline is that the more nuisance parameters you have, the more risk of losing information. We only get a nice CLT for additive martingales but not for multiplicative martingales.

6 Learning (Ch. 9)

6.1 Learning about discrete states

Suppose that $\{\mathbf{X}_t\}$ evolves as an n-state Markov process with (a known) transition matrix Π. Crucially, we assume that $\mathbf{X}_t$ is unknown. However, there is a vector of signals denoted by $\mathbf{Y}_{t+1} - \mathbf{Y}_t$ with density $\psi_i(\mathbf{y}^*)$ if the state i is realized (i.e. if $\mathbf{X}_t$ is the ith coordinate vector). We want to compute the probability that the state is i given the signal history; i.e. we want to compute the vector of conditional probabilities:

$$\mathbf{Q}_t = \begin{bmatrix} q_t^1 \\ q_t^2 \\ \vdots \\ q_t^n \end{bmatrix} = \mathbb{E}\left[\mathbf{X}_t | \mathbf{Y}^t, \mathbf{Q}_0\right] = \mathbb{P}\left[\mathbf{X}_t | \mathbf{Y}^t, \mathbf{Q}_0\right],$$

where $\mathbf{Q}_0$ is the vector of initial probabilities and q_t^i gives the probability that the current state is the ith state. We will do so recursively; i.e. we will derive the expression for

$$\mathbf{Q}_{t+1} = \mathbb{P}\left(\mathbf{X}_{t+1} | \, \mathbf{Q}_t, \mathbf{Y}^{t+1}, \mathbf{Q}_0\right).$$

Using Bayes' rule,[3] we can write $\mathbf{Q}_{t+1}$ as

$$\begin{aligned} \mathbf{Q}_{t+1} &= \mathbb{P}\left(\mathbf{X}_{t+1} | \, \mathbf{Q}_t, \mathbf{Y}_{t+1} - \mathbf{Y}_t, \mathbf{Y}^t, \mathbf{Q}_0\right) \\ &= \frac{\mathbb{P}\left(\mathbf{X}_{t+1}, \mathbf{Y}_{t+1} - \mathbf{Y}_t | \, \mathbf{Q}_t, \mathbf{Y}^t, \mathbf{Q}_0\right)}{\mathbb{P}\left(\mathbf{Y}_{t+1} - \mathbf{Y}_t | \, \mathbf{Q}_t, \mathbf{Y}^t, \mathbf{Q}_0\right)} \\ &= \frac{\mathbb{P}\left(\mathbf{X}_{t+1}, \mathbf{Y}_{t+1} - \mathbf{Y}_t | \, \mathbf{Q}_t\right)}{\mathbb{P}\left(\mathbf{Y}_{t+1} - \mathbf{Y}_t | \, \mathbf{Q}_t\right)}, \end{aligned}$$

where $\mathbf{Q}_0$ is given. Our goal is therefore to derive expressions for the numerator and the denominator of the expression above.

Step 1: Find the joint distribution for $(\mathbf{X}_{t+1}, \mathbf{Y}_{t+1} - \mathbf{Y}_t) | \mathbf{X}_t$. (Further past values of $\mathbf{X}_t$ is not relevant due to the Markovian structure). We begin by assuming that the state is known. The probability that tomorrow's state is the jth coordinate vector is given by π_{ij} for all $j = 1, 2, \ldots, n$; i.e. $(\pi_{i1}, \pi_{i2}, \ldots, \pi_{in})' = \Pi' \mathbf{X}_t$. So,

$$\mathbb{P}\left(\mathbf{X}_{t+1} | \mathbf{X}_t = \mathbf{x}_i\right) = \Pi' \mathbf{X}_t = \begin{bmatrix} \pi_{i1} \\ \pi_{i2} \\ \vdots \\ \pi_{in} \end{bmatrix}$$

gives the probability distribution over the next-period state.

Let $\text{vec}\left[\psi_i(\mathbf{y}^*)\right]$ denote the column vector of $\psi_i(\mathbf{y}^*)$'s. Given $\mathbf{X}_t$ as the ith coordinate vector, the density for $\mathbf{Y}_{t+1} - \mathbf{Y}_t$ is given by

$$\psi_i(\mathbf{y}^*) = \mathbf{X}_t' \text{vec}\left[\psi_i(\mathbf{y}^*)\right].$$

Since $\mathbf{X}_{t+1}$ and $\mathbf{Y}_{t+1} - \mathbf{Y}_t$ are independent conditional on $\mathbf{X}_t$ by assumption, the joint density conditioned on $\mathbf{X}_t$ is given by the product of the two probabilities; i.e.

$$\begin{aligned} \mathbb{P}\left(\mathbf{X}_{t+1}, \mathbf{Y}_{t+1} - \mathbf{Y}_t | \mathbf{X}_t\right) &= \mathbb{P}\left(\mathbf{X}_{t+1} | \mathbf{X}_t\right) \mathbb{P}\left(\mathbf{Y}_{t+1} - \mathbf{Y}_t | \mathbf{X}_t\right) \\ &= \Pi' \mathbf{X}_t \mathbf{X}_t' \text{vec}\left[\psi_i(\mathbf{y}^*)\right] \end{aligned}$$

[3] Recall that $\mathbb{P}(A|B) = \mathbb{P}(A, B) / \mathbb{P}(B)$. Here, let $A = \mathbf{X}_{t+1}$ and $B = \mathbf{Y}_{t+1} - \mathbf{Y}_t$ and condition everything on $\mathbf{Q}_t$.

Step 2: Find the joint distribution of $(\mathbf{X}_{t+1}, \mathbf{Y}_{t+1} - \mathbf{Y}_t)|\mathbf{Q}_t$. In step 1, we assumed that the state is known. But the premise here is that the states are hidden and so, in fact, we can only condition on $\mathbf{Q}_t$, which is the vector of probabilities of the current state. We obtain the joint distribution of $(\mathbf{X}_{t+1}, \mathbf{Y}_{t+1} - \mathbf{Y}_t)|\mathbf{Q}_t$ by taking an "average" of (**??**) (conditioned on $\mathbf{Y}^t$ and $\mathbf{Q}_0$); i.e.

$$\mathbb{P}(\mathbf{X}_{t+1}, \mathbf{Y}_{t+1} - \mathbf{Y}_t|\mathbf{Q}_t) = \begin{bmatrix} q_t^1\pi_{11}\psi_1(\mathbf{y}^*) + q_t^2\pi_{21}\psi_2(\mathbf{y}^*) + \cdots + q_t^n\pi_{n1}\psi_n(\mathbf{y}^*) \\ q_t^1\pi_{12}\psi_1(\mathbf{y}^*) + q_t^2\pi_{22}\psi_2(\mathbf{y}^*) + \cdots + q_t^n\pi_{n2}\psi_n(\mathbf{y}^*) \\ \vdots \\ q_t^1\pi_{1n}\psi_1(\mathbf{y}^*) + q_t^2\pi_{2b}\psi_2(\mathbf{y}^*) + \cdots + q_t^n\pi_{nn}\psi_n(\mathbf{y}^*) \end{bmatrix}_{n\times 1}.$$

We wish to write this more succinctly.

First, observe that, if $\mathbf{X}_t$ is the ith coordinate vector, then $\mathbf{X}_t\mathbf{X}_t'$ is an $n \times n$ matrix with 1 on the ith diagonal and zero everywhere else. Since $\mathbf{Q}_t$ is the vector of (conditional) probabilities that the current state is the 1st, 2nd, ..., nth state,

$$\begin{aligned}\mathbb{E}\left[\mathbf{X}_t\mathbf{X}_t'|\mathbf{Q}_t\right] &= q_t^1\text{diag}\left[\mathbf{e}_1\right] + q_t^2\text{diag}\left[\mathbf{e}_2\right] + \cdots + q_t^n\text{diag}\left[\mathbf{e}_n\right] \\ &= \text{diag}\left[q_t^1\mathbf{e}_1\right] + \text{diag}\left[q_t^2\mathbf{e}_2\right] + \cdots + \text{diag}\left[q_t^n\mathbf{e}_n\right] \\ &= \text{diag}\left[\mathbf{Q}_t\right],\end{aligned}$$

where $\mathbf{e}_i$ is the ith coordinate vector and where $\text{diag}\left[\mathbf{Q}_t\right]$ is a diagonal matrix with the entries of $\mathbf{Q}_t$ on the diagonal. The joint distribution that we want can therefore be expressed as

$$\begin{aligned}\mathbb{P}\left(\mathbf{X}_{t+1}, \mathbf{Y}_{t+1} - \mathbf{Y}_t|\mathbf{Q}_t, \mathbf{Y}^t, \mathbf{Q}_0\right) &= \mathbb{E}\left[\Pi'\mathbf{X}_t\mathbf{X}_t'\text{vec}\left[\psi_i(\mathbf{y}^*)\right]|\mathbf{Q}_t\right] \\ &= \Pi'\mathbb{E}\left[\mathbf{X}_t\mathbf{X}_t'|\mathbf{Q}_t\right]\text{vec}\left[\psi_i(\mathbf{y}^*)\right] \\ &= \Pi'\text{diag}\left[\mathbf{Q}_t\right]\text{vec}\left[\psi_i(\mathbf{y}^*)\right],\end{aligned}$$

Thus, $\mathbf{Q}_t$ encodes the information in history of signals that is relevant for this calculation. Observe that, conditioned on $\mathbf{Q}_t$, the distributions for $\mathbf{X}_{t+1}$ and $\mathbf{Y}_{t+1} - \mathbf{Y}_t$ are not independent.[4]

Step 3: Find the implied distribution for $\mathbf{Y}_{t+1} - \mathbf{Y}_t|\mathbf{Q}_t$. By law of total probabilities:

$$\mathbb{P}(\mathbf{Y}_{t+1} - \mathbf{Y}_t|\mathbf{Q}_t) = \sum_{i=1}^n \mathbb{P}(\mathbf{X}_{t+1} = \mathbf{e}_i, \mathbf{Y}_{t+1} - \mathbf{Y}_t|\mathbf{Q}_t).$$

That is, to obtain the distribution of $\mathbf{Y}_{t+1} - \mathbf{Y}_t|\mathbf{Q}_t$, we need to sum across the rows (or elements) of (**??**), which is equivalent to summing over the "hidden" states:

$$\begin{aligned}\mathbb{P}(\mathbf{Y}_{t+1} - \mathbf{Y}_t|\mathbf{Q}_t) &= q_t^1(\pi_{11} + \pi_{12} + \cdots + \pi_{1n})\psi_1(\mathbf{y}^*) + q_t^2(\pi_{21} + \pi_{22} + \cdots + \pi_{2n})\psi_2(\mathbf{y}^*) \\ &\quad + \cdots + q_t^n(\pi_{n1} + \pi_{n2} + \cdots + \pi_{nn})\psi_n(\mathbf{y}^*) \\ &= q_t^1\psi_1(\mathbf{y}^*) + q_t^2\psi_2(\mathbf{y}^*) + \cdots + q_t^n\psi_n(\mathbf{y}^*) \\ &= \underbrace{\mathbf{1}_n'\Pi'\text{diag}\left[\mathbf{Q}_t\right]}_{=\mathbf{Q}_t'}\text{vec}\left[\psi_i(\mathbf{y}^*)\right],\end{aligned}$$

[4] Observe that

$$\begin{aligned}\mathbb{P}(\mathbf{X}_{t+1}|\mathbf{Q}_t) &= \Pi'\mathbf{Q}_t, \\ \mathbb{P}(\mathbf{Y}_{t+1} - \mathbf{Y}_t|\mathbf{Q}_t) &= \mathbf{Q}_t'\text{vec}\left[\psi_i(\mathbf{y}^*)\right], \\ \mathbb{P}(\mathbf{X}_{t+1}|\mathbf{Q}_t)\,\mathbb{P}(\mathbf{Y}_{t+1} - \mathbf{Y}_t|\mathbf{Q}_t) &= \Pi'\mathbf{Q}_t\mathbf{Q}_t'\text{vec}\left[\psi_i(\mathbf{y}^*)\right] \\ &\neq \Pi'\text{diag}\left[\mathbf{Q}_t\right]\text{vec}\left[\psi_i(\mathbf{y}^*)\right].\end{aligned}$$

where we used the fact that

$$\mathbf{1}_n'\Pi' = \mathbf{1}_n' \Rightarrow \mathbf{1}_n'\Pi'\text{diag}\left[\mathbf{Q}_t\right] = \left[q_t^1, q_t^2, \ldots, q_t^n\right] = \mathbf{Q}_t'.$$

Observe that $\mathbf{Q}_t$ is a vector of weight used in forming a mixture distribution. Suppose, for instance, ψ_i is a normal distribution with mean $\boldsymbol{\mu}_i$ and variance-covariance matrix $\boldsymbol{\Sigma}_i$, then the distribution of $\mathbf{Y}_{t+1} - \mathbf{Y}_t$ conditioned on $\mathbf{Q}_t$ is a mixture of normals with mixing probabilities given by respective entries of $\mathbf{Q}_t$

Step 4: Compute $\mathbf{Q}_{t+1}$. Substituting the expressions we obtained into the equation for $\mathbf{Q}_{t+1}$ yields

$$\begin{aligned}\mathbf{Q}_{t+1} &= \frac{\mathbb{P}\left(\mathbf{X}_{t+1}, \mathbf{Y}_{t+1} - \mathbf{Y}_t | \mathbf{Q}_t\right)}{\mathbb{P}\left(\mathbf{Y}_{t+1} - \mathbf{Y}_t | \mathbf{Q}_t\right)} \\ &= \frac{\Pi'\text{diag}\left(\mathbf{Q}_t\right)\text{vec}\left[\psi_i\left(\mathbf{Y}_{t+1} - \mathbf{Y}_t\right)\right]}{\mathbf{Q}_t'\text{vec}\left[\psi_i\left(\mathbf{Y}_{t+1} - \mathbf{Y}_t\right)\right]}.\end{aligned}$$

Observe that steps (3) and (4) define a Markov process for $\mathbf{Q}_{t+1}$. Step (3) gives $\mathbf{Y}_{t+1} - \mathbf{Y}_t$ drawn from a (history dependent) mixture of densities ψ_i, and step (4), we construct $\mathbf{Q}_{t+1}$ as a function of $\mathbf{Q}_t$ and $\mathbf{Y}_{t+1} - \mathbf{Y}_t$ from step (3).

Algorithm 2. *(Learning about discrete states). $\{\mathbf{X}_t\}$ is a hidden n-state Markov process with (a known) transition matrix Π, There is a vector of signals denoted by $\{\mathbf{Y}_{t+1} - \mathbf{Y}_t\}$ with density $\psi_i(\mathbf{y}^*)$ if the state is i (i.e. $\mathbf{X}_t$ is the ith coordinate vector). $\mathbf{Q}_t$ denotes the probability that the state is i given the signal history, and $\mathbf{Q}_0$ is the vector of initial probabilities. $\mathbf{Y}_{t+1} - \mathbf{Y}_t | \mathbf{X}_t$ and $\mathbf{X}_{t+1} | \mathbf{X}_t$ are independent.*

▷ *Step 1: Find the joint distribution for $\left(\mathbf{X}_{t+1}, \mathbf{Y}_{t+1} - \mathbf{Y}_t\right) | \mathbf{X}_t$*

$$\begin{aligned}\mathbb{P}\left(\mathbf{X}_{t+1}, \mathbf{Y}_{t+1} - \mathbf{Y}_t | \mathbf{X}_t\right) &= \mathbb{P}\left(\mathbf{X}_{t+1} | \mathbf{X}_t\right)\mathbb{P}\left(\mathbf{Y}_{t+1} - \mathbf{Y}_t | \mathbf{X}_t\right) \\ &= \Pi'\mathbf{X}_t\mathbf{X}_t'\text{vec}\left[\psi_i\left(\mathbf{y}^*\right)\right].\end{aligned}$$

▷ *Step 2: Find the joint distribution of $\left(\mathbf{X}_{t+1}, \mathbf{Y}_{t+1} - \mathbf{Y}_t\right) | \mathbf{Q}_t$ by taking the expectation of the above term.*

$$\begin{aligned}\mathbb{P}\left(\mathbf{X}_{t+1}, \mathbf{Y}_{t+1} - \mathbf{Y}_t | \mathbf{Q}_t\right) &= \mathbb{E}\left[\Pi'\mathbf{X}_t\mathbf{X}_t'\text{vec}\left[\psi_i\left(\mathbf{y}^*\right)\right] | \mathbf{Q}_t\right] \\ &= \Pi'\text{diag}\left[\mathbf{Q}_t\right]\text{vec}\left[\psi_i\left(\mathbf{y}^*\right)\right].\end{aligned}$$

▷ *Step 3: Find the implied distribution for $\mathbf{Y}_{t+1} - \mathbf{Y}_t | \mathbf{Q}_t$.*

$$\begin{aligned}\mathbb{P}\left(\mathbf{Y}_{t+1} - \mathbf{Y}_t | \mathbf{Q}_t\right) &= \sum_{i=1}^{n}\mathbb{P}\left(\mathbf{X}_{t+1} = \mathbf{x}_i, \mathbf{Y}_{t+1} - \mathbf{Y}_t | \mathbf{Q}_t\right) \\ &= \underbrace{\mathbf{1}_n'\Pi'\text{diag}\left[\mathbf{Q}_t\right]}_{=\mathbf{Q}_t'}\text{vec}\left[\psi_i\left(\mathbf{y}^*\right)\right],\end{aligned}$$

▷ *Step 4: Compute $\mathbf{Q}_{t+1}$.*

$$\begin{aligned}\mathbf{Q}_{t+1} &= \frac{\mathbb{P}\left(\mathbf{X}_{t+1}, \mathbf{Y}_{t+1} - \mathbf{Y}_t | \mathbf{Q}_t\right)}{\mathbb{P}\left(\mathbf{Y}_{t+1} - \mathbf{Y}_t | \mathbf{Q}_t\right)} \\ &= \frac{\Pi'\text{diag}\left(\mathbf{Q}_t\right)\text{vec}\left[\psi_i\left(\mathbf{Y}_{t+1} - \mathbf{Y}_t\right)\right]}{\mathbf{Q}_t'\text{vec}\left[\psi_i\left(\mathbf{Y}_{t+1} - \mathbf{Y}_t\right)\right]}.\end{aligned}$$

7 Generalized Method of Moments

Generalised Method of Moments (GMM) refers to a class of estimators which are constructed from exploiting the sample moment counterparts of population moment conditions (sometimes known as orthogonality conditions) of the data generating model. GMM estimators have become widely used, for the following reasons.

- ▷ GMM estimators have large sample properties that are easy to characterise in ways that facilitate comparison. A family of such estimators can be studied a priori in ways that make asymptotic efficiency comparisons easy. The method also provides a natural way to construct tests which take account of both sampling and estimation error.
- ▷ In practice, researchers find it useful that GMM estimators can be constructed without specifying the full data generating process (which would be required to write down the maximum likelihood estimator.) This characteristic has been exploited in analysing partially specified economic models, in studying potentially misspecified dynamic models designed to match target moments, and in constructing stochastic discount factor models that link asset pricing to sources of macroeconomic risk.

7.1 Preliminaries

Let $\mathbf{Z}$ be an index set of $\mathcal{Z}$, where $\mathcal{Z}$ is a liner space of m-dimensional random vectors. Let Φ be a known m-dimensional function and suppose that an economic model implies

$$\mathbb{E}\left[\mathbf{Z}\cdot\Phi\left(\mathbf{Y},\boldsymbol{\beta}\right)\right]=0,\ \forall\mathbf{Z}\in\mathcal{Z},$$

where $\boldsymbol{\beta}\in\mathbb{R}^k$ is an unknown parameter and $\mathbf{Y}$ is the data. The function Φ can be non-linear.

Given that $\boldsymbol{\beta}$ is a $k\times 1$ vector, we require $\mathbf{Z}^j$ entries in $\mathcal{Z}$ for $j=1,2,\ldots,k$ which gives us k equations in k unknowns:

$$\mathbb{E}\left[\mathbf{Z}^j\cdot\Phi\left(\mathbf{Y},\mathbf{b}\right)\right]=0,\ j=1,2,\ldots,k$$

to use in identifying $\mathbf{b}=\boldsymbol{\beta}$. We can think of the $\mathbf{Z}^j$'s as "instruments". The question then becomes how should we choose $\mathbf{Z}^j$'s and what are the properties of the estimator for $\boldsymbol{\beta}$ obtained from this procedure?

7.1.1 Finite unconditional moment restrictions

The set $\mathcal{Z}$ can potentially be very large. Here, we collapse $\mathcal{Z}$ so that it consists of linear combinations of $r\geq k$ basis $\mathbf{Z}$'s that satisfy $\mathbb{E}\left[\mathbf{Z}\cdot\Phi\left(\mathbf{Y},\boldsymbol{\beta}\right)\right]=0,\ \forall\mathbf{Z}\in\mathcal{Z}$. Let

$$F\left(\mathbf{X},\mathbf{b}\right)_{r\times 1}=\begin{pmatrix}\mathbf{Z}^{1\prime}\\ \mathbf{Z}^{2\prime}\\ \vdots\\ \mathbf{Z}^{r\prime}\end{pmatrix}_{r\times m}\Phi\left(\mathbf{X},\mathbf{b}\right)_{m\times 1}.$$

The unconditional moment restriction is

$$\mathbb{E}\left[F\left(\mathbf{X},\boldsymbol{\beta}\right)\right]=\mathbf{0}_{r\times 1},$$

where $\mathbf{X}$ contains $\mathbf{Y}$ and the r basis $\mathbf{Z}$'s. Let $\mathcal{Z}=\mathbb{R}^r$. In this case, an estimator may be associated with an $r\times k$ selection matrix $\mathbf{A}$ or real numbers where each of the k columns are vectors in $\mathcal{Z}$. Thus, the equation used to estimate are

$$\mathbf{A}'\mathbb{E}\left[F\left(\mathbf{X},\mathbf{b}\right)\right]=\mathbf{0}_{k\times 1},$$

which are satisfied for $\mathbf{b}=\boldsymbol{\beta}$. A choice of $\mathbf{A}$ implements/indexes a particular GMM estimator.

Example 7.1. (Moment matching) Suppose that

$$\Phi(\mathbf{Y}, \mathbf{b}) = \Psi(\mathbf{Y}) - \Gamma(\mathbf{b}),$$

where

$$\mathbb{E}[\Psi(\mathbf{Y})] = \Gamma(\boldsymbol{\beta}).$$

In this formulation, $\Psi(\mathbf{Y})$ defines the moments to be matched and $\Gamma(\mathbf{b})$ give the predicted moments from a model as a function of a potential parameter vector $\mathbf{b}$. We presume that $r \geq k$ and $\mathcal{Z} = \mathbb{R}^r$. In other words, LHS can be obtained from data and the RHS can be obtained from the model.

Example 7.2. (IV) Suppose $\alpha \cdot Y_t = u_t$ (disturbances) and suppose I know that $\mathbb{E}[Z_t u_t] = 0$ where Z_t is a $r \times 1$ matrix. This is equivalent to saying:

$$\alpha' \mathbb{E}\left[Y_t Z_t'\right] = \mathbf{0}$$

In the population, α is only identified up to a scalar. So you need β to impose conditions on α. What you're identifying here is a null-space and you use β to get identification.

7.1.2 Conditional moment restrictions

Consider the following conditional moment restriction:

$$\mathbb{E}[\Phi(\mathbf{Y}, \boldsymbol{\beta}) | \mathcal{K}] = 0,$$

where $\mathcal{K}$ is some conditioning information set.

Suppose that $\Phi(\mathbf{Y}, \mathbf{b})$ for $\mathbf{b} \in \mathbb{R}^k$ has a finite second moment and $\mathcal{Z}$ contains at least m-dimensional random vectors that are bounded and $\mathcal{K}$ measurable.

By Law of Iterated Expectations, we can see that conditional moment restriction implies unconditional moment restriction:

$$\mathbb{E}[\mathbf{Z} \cdot \Phi(\mathbf{Y}, \boldsymbol{\beta})] = \mathbb{E}[\mathbf{Z} \cdot \mathbb{E}[\Phi(\mathbf{Y}, \boldsymbol{\beta}) | \mathcal{K}]] = 0$$

since $\mathbf{Z}$ is in the $\mathcal{K}$ information set. In general, the converse is not true; i.e. unconditional moment restriction does not imply conditional moment restriction.

However, if the unconditional moment restriction holds for all $\mathbf{Z} \in \mathcal{Z}$, then we have the orthogonality condition for the conditional expectation:

$$\mathbb{E}[\mathbf{Z} \cdot (\Phi(\mathbf{Y}, \boldsymbol{\beta}) - \mathbf{0})] = 0, \ \forall \mathbf{Z} \in \mathcal{Z}$$

so that we can get back to

$$\mathbb{E}[\Phi(\mathbf{Y}, \boldsymbol{\beta}) | \mathcal{K}] = 0.$$

So, in this case, the family of unconditional moment restrictions imply the conditional moment restriction.

Example 7.3. (*Stochastic discount factor*) Consider the following moment restriction

$$\mathbb{E}\left[\left(\frac{S_{t+\ell}}{S_t}\right)\mathbf{R}_{t+\ell} | \mathfrak{F}_t\right] = \mathbf{1}_n,$$

where $S_{t+\ell}/S_t$ is a stochastic discount factor over horizon ℓ, $\mathbf{R}_{t+\ell}$ is an n-dimensional vector of returns, and $\mathfrak{F}_t$ is the information available at time t. Here,

$$\Phi(\mathbf{Y}, \mathbf{b}) = \left(\frac{S_{t+\ell}}{S_t}\right)\mathbf{R}_{t+\ell} - \mathbf{1}_n,$$
$$\mathcal{K} = \mathfrak{F}_t.$$

7.2 Traditional Way of Presenting GMM (via Minimization Problem)

Given some parameter space $\mathbf{P}$, we can obtain the GMM estimator $\mathbf{b}_N$ by solving

$$\min_{\mathbf{b}\in\mathbf{P}} \quad \left(\frac{1}{\sqrt{N}}\sum_{t=1}^{N} F\left(\mathbf{X}_t,\mathbf{b}\right)\right)' \left(\mathbf{V}_N\left(\mathbf{b}\right)\right)^{-1} \left(\frac{1}{\sqrt{N}}\sum_{t=1}^{N} F\left(\mathbf{X}_t,\mathbf{b}\right)\right).$$

Notice that we are simultaneously solving for the estimator $\mathbf{b}$ as well as the variance $\mathbf{V}_N\left(\mathbf{b}\right)$.

The first-order condition is

$$\mathbf{0} = 2\left(\frac{1}{\sqrt{N}}\sum_{t=1}^{N} \frac{\partial}{\partial \mathbf{b}} F\left(\mathbf{X}_t,\mathbf{b}\right)\right)' \left(\mathbf{V}_N\left(\mathbf{b}\right)\right)^{-1} \left(\frac{1}{\sqrt{N}}\sum_{t=1}^{N} F\left(\mathbf{X}_t,\mathbf{b}\right)\right) + \text{other terms}.$$

(The other terms, such as the derivative of $\left(\mathbf{V}_N\left(\mathbf{b}\right)\right)^{-1}$, are going to be pre- and post- multiplied by $N^{-1/2}\sum_{t=1}^{N} F\left(\mathbf{X}_t,\mathbf{b}\right)$, which converges in probability to zero so we need not worry about them). Notice that

$$\frac{1}{N}\sum_{t=1}^{N} \frac{\partial}{\partial \mathbf{b}} F\left(\mathbf{X}_t,\mathbf{b}\right) \xrightarrow{\text{p}} \mathbf{D} = \mathbb{E}\left[\frac{\partial F\left(\mathbf{X},\boldsymbol{\beta}\right)}{\partial \mathbf{b}}\right]$$

So asymptotically,

$$\left(\frac{1}{N}\sum_{t=1}^{N} \frac{\partial}{\partial \mathbf{b}} F\left(\mathbf{X}_t,\mathbf{b}\right)\right)' \left(\mathbf{V}_N\left(\mathbf{b}\right)\right)^{-1} \xrightarrow{\text{p}} \mathbf{D}'\mathbf{V}^{-1} = \mathbf{A}^{*\prime}.$$

The first-order condition in the limit is

$$\mathbf{0} \approx \mathbf{A}^{*\prime}\frac{1}{\sqrt{N}}\sum_{t=1}^{N} F\left(\mathbf{X}_t,\mathbf{b}\right).$$

Thus, the estimator is asymptotically efficient.

7.3 GMM Estimation with Constant Selection

Consider $F\left(\mathbf{x},\mathbf{b}_{k\times 1}\right)_{r\times 1}$ and consider $\mathbb{E}\left[F\left(\mathbf{x},\mathbf{b}\right)|\ell\right] = 0$ (where we will drop conditioning). Then consider

$$\mathbf{A}^{i}_{r\times 1}\cdot\mathbb{E}\left[F\left(\mathbf{X}_t,\mathbf{b}\right)\right] = \mathbf{0}, \forall t = 1,...,k$$

so we have k equations yielding a unique $\mathbf{b}_N$. Stacking $\mathbf{A} = \left[\mathbf{A}^1,...,\mathbf{A}^k\right]$,we have

$$\mathbf{A}'\mathbb{E}\left[F\left(\mathbf{X}_t,\mathbf{b}\right)\right] = 0 \Leftrightarrow \mathbf{b} = \beta$$

Then we say $\mathbf{b}_N$ is a GMM estimator where

$$\mathbf{A}'\frac{1}{\sqrt{N}}\sum_{t=1}^{N} F\left(\mathbf{X}_t,\mathbf{b}_N\right) = \mathbf{0}$$

where $\mathbf{A}$ is the selection matrix. We are interested in studying a family of estimators indexed by $\mathbf{A}$.

7.3.1 Example: Euler Equations and SDF

Consider a consumption process $\{C_t\} : (C_0, C_1, ...)$. We can think of this being on an infinite-dimensional indifference curve. Now stay on same indifference curvea and consider a different point for some scalar r:

$$\left(\underline{C_0 - P_0(r)}, C_1, C_2, ..., \underline{C_\tau + r\xi_\tau}, C_{\tau+1}, \cdots\right)$$

where ξ_τ is a random variable dependent on time τ. To change time-τ consumption by that amount, we need to reduce today's consumption by $P_0(r)$.

Now define $\pi_0^\tau(S_\tau)$, the slope of the indifference curve, to be

$$\pi_0^\tau(\xi_\tau) = \frac{d}{dr} P_0(r)\,|_{r=0}$$

and this also gives us the shadow price of asset (security) that pays off ξ_τ at date τ. Using a SDF notation:

$$\pi_0^\tau(\xi_\tau) = \mathbb{E}\left[\left(\frac{S_\tau}{S_0}\right)\xi_\tau|\mathfrak{F}_0\right]$$

where $S_{t+\ell}/S_t$ is a stochastic discount factor over horizon τ.

▷ Stochastic: Since ξ_τ is random, we can't discount it at some fexed rate. The riskiness of ξ_τ has to demand some kind of compensation. The way you discount the cash flow by something that is random – you discount different states of the world differently.

So how should we get the SDF? We know that prices equal the marginal rates of substitution, so we use this insight:

$$\frac{S_\tau}{S_0} = \frac{MU_\tau^0}{MU_0^0}$$

where the RHS is the intertemporal marginal rate of substitution. To proceed, we need to define the preferences (from the perspective from time zero) as

$$\mathbb{E}\left[\sum_{t=0}^{\infty}\exp(-\delta\tau)\,U(C_t)\,|\mathfrak{F}_0\right]$$

Then $MU_\tau^0 = \exp(-\delta\tau)\,U'(C_\tau)$ and $MU_0^0 = U'(C_0)$ which gives us:

$$\frac{S_\tau}{S_0} = \exp(-\delta\tau)\,\frac{U'(C_\tau)}{U'(C_0)}$$

Thus:

$$\pi_0^\tau(\xi_\tau) = \mathbb{E}\left[\exp(-\delta\tau)\,\frac{U'(C_\tau)}{U'(C_0)}\xi_\tau|\mathfrak{F}_0\right]$$

This is called the Euler Equation since we are sitting in the optimal consumption process and throwing in this perturbation to see how consumption changes. This is an example of the conditional moment restriction.

7.3.2 Adding Distributional Assumptions

Now for pedagogical reasons, we add some distributional assumptions. Note that these conditions are not necessary for the estimation. Assume ξ_τ is the gross return i.e, an object that has price $\pi_0^\tau(\xi_\tau) = 1$ and $(\xi_\tau, S_\tau/S_0)$ are jointly lognormal. Furthermore:

$$\log S_\tau - \log S_0 = \mu_0^S + \sigma^S \cdot W_\tau, \quad W_\tau \sim N(\mathbf{0}, I)$$
$$\log \xi_\tau = \mu_0^\xi + \sigma^\xi \cdot W_\tau$$

Note that μ_0^ξ may also depend on τ.

▷ Adding the two equations, we have:

$$\log S_\tau - \log S_0 + \log \xi_\tau = \left(\mu_0^s + \mu_0^\xi\right) + \left(\sigma^S + \sigma^\xi\right) W_\tau$$

$$\Rightarrow \log \frac{S_\tau}{S_0}\xi_\tau = \left(\mu_0^s + \mu_0^\xi\right) + \left(\sigma^S + \sigma^\xi\right) W_\tau$$

▷ Therefore:

$$\log \mathbb{E}\left[\frac{S_\tau}{S_0}\xi_\tau | \mathfrak{F}_0\right] = \mu_0^S + \mu_0^\xi + \frac{1}{2}\left|\sigma^S + \sigma^\xi\right|^2$$

which must be equal to zero since $\pi_0^\tau(\xi_\tau) = 1$.

▷ Rearranging:

$$\begin{aligned} 0 &= \mu_0^s + \mu_0^\xi + \frac{1}{2}\left(\sigma^S\right)^2 + \frac{1}{2}\left(\sigma^\xi\right)^2 + \left|\sigma^S \cdot \sigma^\xi\right| \\ \Leftrightarrow -\sigma^S \cdot \sigma^\xi &= \underbrace{\left[\mu_0^\xi + \frac{1}{2}\left|\sigma^\xi\right|^2\right]}_{[1]} + \underbrace{\left[\mu_0^S + \frac{1}{2}\left|\sigma^S\right|^2\right]}_{[2]} \qquad [A] \end{aligned}$$

* [1] is equal to the log expected return, and [2] is minus of log riskless return. To get the riskless return, set $\sigma^\xi = 0$, then the (riskfree) return is μ_0^ξ. We can explicitly obtain this expression from $[A]$:

$$0 = \mu_0^\xi - \left(\mu_0^S + \frac{1}{2}\left|\sigma^S\right|^2\right) \Rightarrow \mu_0^\xi = -\left(\mu_0^S + \frac{1}{2}\left|\sigma^S\right|^2\right)$$

* Therefore, $-\sigma^S \cdot \sigma^\xi$ is the risk compensation. We can think of $-\sigma^S$ as the risk price vector. It is linear in terms of how exposed you are to the risk.

7.4 GMM Limiting Approximation

Start with

$$\frac{1}{\sqrt{N}}\sum_{t=1}^N F(\mathbf{X}_t, \boldsymbol{\beta}) \xrightarrow{\mathrm{d}} N(0, \mathbf{V}), \qquad \mathbf{V} \coloneqq \mathbb{E}\left[\mathbf{H}_{t+1}\mathbf{H}_{t+1}'\right]$$

How can we get here? Take consistency of $\mathbf{b}_N$ as given:

▷ Taylor approximation of F around $\mathbf{b}_N = \boldsymbol{\beta}$ gives

$$\mathbf{A}'\frac{1}{\sqrt{N}}\sum_{t=1}^N F(\mathbf{X}_t, \mathbf{b}_N) \approx \mathbf{A}'\frac{1}{\sqrt{N}}\sum_{t=1}^N F(\mathbf{X}_t, \boldsymbol{\beta}) + \mathbf{A}'\frac{1}{\sqrt{N}}\sum_{t=1}^N \frac{\partial F(\mathbf{X}, \boldsymbol{\beta})}{\partial \mathbf{b}'}(\mathbf{b}_N - \boldsymbol{\beta}).$$

You can also get the same result using the mean-value theorem, but this introduces a new point $\bar{\mathbf{b}}_N$ which will also converge to β.

▷ Mean-value theorem with law of large numbers implies

$$\frac{1}{N}\sum_{t=1}^N \frac{\partial F(\mathbf{X}, \boldsymbol{\beta})}{\partial \mathbf{b}'} \xrightarrow{\mathrm{p}} \mathbb{E}\left[\frac{\partial F(\mathbf{X}, \boldsymbol{\beta})}{\partial \mathbf{b}'}\right] =: \mathbf{D},$$

where $\mathbf{D}$ is an $r \times k$ vector of expectation of partial derivatives of F with respect to $\mathbf{b}$; i.e.

$$\mathbf{D} \coloneqq \mathbb{E}\begin{bmatrix} \frac{\partial F_1(\mathbf{X},\boldsymbol{\beta})}{\partial b_1} & \frac{\partial F_1(\mathbf{X},\boldsymbol{\beta})}{\partial b_2} & \cdots & \frac{\partial F_1(\mathbf{X},\boldsymbol{\beta})}{\partial b_k} \\ \frac{\partial F_2(\mathbf{X},\boldsymbol{\beta})}{\partial b_1} & \frac{\partial F_2(\mathbf{X},\boldsymbol{\beta})}{\partial b_2} & \cdots & \frac{\partial F_2(\mathbf{X},\boldsymbol{\beta})}{\partial b_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial F_r(\mathbf{X},\boldsymbol{\beta})}{\partial b_1} & \frac{\partial F_r(\mathbf{X},\boldsymbol{\beta})}{\partial b_2} & \cdots & \frac{\partial F_r(\mathbf{X},\boldsymbol{\beta})}{\partial b_k} \end{bmatrix}_{r\times k}.$$

So,

$$\underbrace{\mathbf{A}'\frac{1}{\sqrt{N}}\sum_{t=1}^{N} F\left(\mathbf{X}_t, \mathbf{b}_N\right)}_{\text{k equations stacked}} \approx \mathbf{A}'\frac{1}{\sqrt{N}}\sum_{t=1}^{N} F\left(\mathbf{X}_t, \boldsymbol{\beta}\right) + \mathbf{A}'\sqrt{N}\frac{1}{N}\sum_{t=1}^{N}\frac{\partial F\left(\mathbf{X}, \boldsymbol{\beta}\right)}{\partial \mathbf{b}'}\left(\mathbf{b}_N - \boldsymbol{\beta}\right)$$

$$\approx \mathbf{A}'\frac{1}{\sqrt{N}}\sum_{t=1}^{N} F\left(\mathbf{X}_t, \boldsymbol{\beta}\right) + \mathbf{A}'\mathbf{D}\sqrt{N}\left(\mathbf{b}_N - \boldsymbol{\beta}\right).$$

▷ Since the left-hand side equals zero given that we are considering estimators of the kind

$$\mathbf{A}'\frac{1}{\sqrt{N}}\sum_{t=1}^{N} F\left(\mathbf{X}_t, \mathbf{b}_N\right) = \mathbf{0}$$

rearranging gives

$$\mathbf{A}'\mathbf{D}\sqrt{N}\left(\mathbf{b}_N - \boldsymbol{\beta}\right) \approx \mathbf{A}'\frac{1}{\sqrt{N}}\sum_{t=1}^{N} F\left(\mathbf{X}_t, \boldsymbol{\beta}\right).$$

Assuming that $\mathbf{A}'\mathbf{D}$ is nonsingular, then

$$\underline{\sqrt{N}\left(\mathbf{b}_n - \boldsymbol{\beta}\right) \approx -\left(\mathbf{A}'\mathbf{D}\right)^{-1}\mathbf{A}'\frac{1}{\sqrt{N}}\sum_{t=1}^{N} F\left(\mathbf{X}_t, \boldsymbol{\beta}\right).}$$

The key here is that the covariance matrix depends on $\mathbf{A}$ and $\mathbf{D}$ so the natural question is bounding this covariance.

7.5 GMM Efficiency Bound

Every GMM estimator satisfies:

$$\sqrt{N}\left(\mathbf{b}_n - \boldsymbol{\beta}\right) \approx -\left(\mathbf{A}'\mathbf{D}\right)^{-1}\mathbf{A}'\frac{1}{\sqrt{N}}\sum_{t=1}^{N} F\left(\mathbf{X}_t, \boldsymbol{\beta}\right),$$

We also know from the Central Limit Theorem:

$$\frac{1}{\sqrt{N}}\sum_{t=1}^{N} F\left(\mathbf{X}_t, \boldsymbol{\beta}\right) \xrightarrow{\text{d}} N\left(0, \mathbf{V}\right),\ \mathbf{V} = \mathbb{E}\left[\mathbf{H}_{t+1}\mathbf{H}_{t+1}'\right].$$

Combining these two together, we have:

$$\sqrt{N}\left(\mathbf{b}_n - \boldsymbol{\beta}\right) \xrightarrow{\text{d}} N\left(0, Cov\left[\mathbf{A}\right]\right),$$

where $Cov\left[\mathbf{A}\right]$ denotes the asymptotic variance-covariance of the estimator (indexed by the selection matrix $\mathbf{A}$):

$$Cov\left[\mathbf{A}\right] \coloneqq \left(\mathbf{A}'\mathbf{D}\right)^{-1}\mathbf{A}'\mathbf{V}\mathbf{A}\left(\mathbf{D}'\mathbf{A}\right)^{-1}.$$

The goal is to find the greatest lower bound for $Cov\left[\mathbf{A}\right]$ provided that there exist an $\mathbf{A}$ such that $\mathbf{A}'\mathbf{D}$ is non-singular.

1. Establish that post-multiplying the weighting matrix with a nonsingular matrix does not alter the asymptotic variance-covariance matrix of the GMM estimator.

   ▷ This is intuitive because premultiplying the estimation equation

   $$\mathbf{A}'\frac{1}{N}\sum_{t=1}^{N} F_t\left(\mathbf{X}_t, \mathbf{b}_N\right) = \mathbf{0}.$$

   will not change the optimal $\mathbf{A}$.

2. Find $\tilde{\mathbf{A}}$ such that $\mathbf{A}'\mathbf{V}\tilde{\mathbf{A}} = \mathbf{A}^{'}\mathbf{D}$ for all $\mathbf{A}$. It turns out to be $\tilde{\mathbf{A}} = \mathbf{V}^{-1}\mathbf{D}$.

   ▷ This is a step to help us figure out $\mathbf{A}^*$ in the next step.

3. Above result allows us to normalize $\mathbf{A}$ to $\mathbf{A}^*$ such that $\left\langle \mathbf{A}^* | \tilde{\mathbf{A}} \right\rangle = \mathbf{A}^*\mathbf{D} = \mathbf{I}$ can be done without loss of efficiency, and without affecting the GMM estimators.

4. Denote $\mathbf{A}^* = \tilde{\mathbf{A}}\left(\mathbf{D}'\tilde{\mathbf{A}}\right)^{-1} = \mathbf{V}^{-1}\mathbf{D}\left(\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\right)^{-1}$ and show that $Cov\left[\mathbf{A}^*\right]$ is the GMM efficiency bound, i.e.

   $$Cov\left[\mathbf{A}\right] \geq Cov\left[\mathbf{A}^*\right] = Cov\left[\tilde{\mathbf{A}}\right]$$

   ▷ First, observe that $\mathbf{A}'\mathbf{V}\mathbf{A}^* = \mathbf{A}'\mathbf{D}\left(\mathbf{D}'\mathbf{V}^{-\mathbf{1}}\mathbf{D}\right)^{-1} = \left(\mathbf{D}\mathbf{V}^{-1}\mathbf{D}\right)^{-1}$ provided $\mathbf{A}'\mathbf{D} = \mathbf{I}$.

   ▷ Second, note that

   $$\begin{aligned} Cov\left[\mathbf{A}\right] &= \left(\mathbf{A}'\mathbf{V}\tilde{\mathbf{A}}\right)^{-1}\mathbf{A}'\mathbf{V}\mathbf{A}\left(\tilde{\mathbf{A}}'\mathbf{V}\mathbf{A}\right)^{-1} \\ &= \mathbf{I}^{-1}\mathbf{A}'\mathbf{V}\mathbf{A}\mathbf{I} = \mathbf{A}'\mathbf{V}\mathbf{A} \end{aligned}$$

   ▷ Then

   $$\begin{aligned} \left(\mathbf{A}-\mathbf{A}^*\right)'\mathbf{V}\left(\mathbf{A}-\mathbf{A}^*\right) &= \mathbf{A}'\mathbf{V}\mathbf{A} - \mathbf{A}^*\mathbf{V}\mathbf{A} - \mathbf{A}'\mathbf{V}\mathbf{A}^* + \mathbf{A}^{*'}\mathbf{V}\mathbf{A}^* \\ &= \mathbf{A}'\mathbf{V}\mathbf{A} - \mathbf{A}^*\mathbf{V}\mathbf{A}^* \\ &= Cov\left[\mathbf{A}\right] - Cov\left[\mathbf{A}^*\right] \end{aligned}$$

   where the penultimate equality is from the fact that $\mathbf{A}^{*'}\mathbf{D} = \mathbf{I}$ and thus we can safely plug in $\mathbf{A}^*$ into the equation in the first bullet.

   ▷ The LHS is positive semi-definite by construction so the RHS must also be positive semi-definite.

Just remember that $\underline{\mathbf{A}^* = \mathbf{V}^{-1}\mathbf{D} \text{ is the efficient selection matrix}}$ where $\mathbf{D}_{r\times k} = \mathbb{E}\left[\partial F\left(\mathbf{X}_t, \boldsymbol{\beta}\right)/\partial \mathbf{b}'\right]$.

Example 7.4. (Imposing Restrictions) Consider $\alpha'\mathbb{E}\left[Y_t|\mathcal{F}_{t-\ell}\right] = 0, \ell \geq 1$. Suppose we have $z_{t-\ell} \in \mathcal{F}_{t-\ell}$. Then we have a matrix

$$\mathbb{E}\left[z_{t-\ell}Y_t'\right]\alpha = 0$$

where β restricts α. Regress y_t^1 onto $z_{t-\ell}$ and regress y_t^2 onto $z_{t-\ell}$.

▷ If both of them have zero predictability, we are in trouble because $\mathbb{E}\left[z_{t-\ell}Y_t'\right]\alpha$ will be all zeros.

▷ I just need one of these two regression to tell me that it's predictable.

Think of the structural model as $\alpha' y_t = u_t$ and we need $\mathbb{E}[z_{t-\ell} u_t] = 0$ for things to be working. If it has rank two, we're fucked. If it has rank one, we're good but to make it rank one, we can have:

$$\alpha = \begin{bmatrix} 1 \\ \beta \end{bmatrix} \quad vs. \qquad \alpha = \begin{bmatrix} \beta \\ 1 \end{bmatrix}$$

which yields different

$$\mathbb{E}\left[\frac{\partial F(\mathbf{X}, \boldsymbol{\beta})}{\partial \mathbf{b}'}\right] =: \mathbf{D} = \mathbb{E}\left[y_t^2 z_{t-2}\right] \quad vs. \quad \mathbb{E}\left[y_t^1 z_{t-2}\right]$$

Example 7.5. *(Serial correlation)* Suppose we have quarterly data and the moment condition as

$$\mathbb{E}\left[F(\mathbf{X}_t, \boldsymbol{\beta}) | \mathfrak{F}_{t-\ell}\right] = \mathbf{0}.$$

Furthermore, denote $G_t = F(\mathbf{X}_t, \boldsymbol{\beta})$. Let $\ell = 2$. Then,

$$\log Y_{t+2} - \log Y_t = \text{constant} + \text{parameters} + u_{t+2},$$

where $u_{t+2} = F(\mathbf{X}_t, \boldsymbol{\beta})$. We assume

$$\mathbb{E}\left[u_{t+2} | \mathfrak{F}_t\right] = \mathbb{E}\left[F(\mathbf{X}_{t+2}, \boldsymbol{\beta}) | \mathfrak{F}_t\right] = 0.$$

By law of iterated expectations,

$$\mathbb{E}\left[F(\mathbf{X}_{t+2}, \boldsymbol{\beta}) F(\mathbf{X}_t, \boldsymbol{\beta})'\right] = \mathbb{E}\left[\underbrace{\mathbb{E}\left[F(\mathbf{X}_{t+2}, \boldsymbol{\beta}) | \mathfrak{F}_t\right]}_{=0} F(\mathbf{X}_t, \boldsymbol{\beta})'\right] = 0,$$

$$\mathbb{E}\left[F(\mathbf{X}_{t+3}, \boldsymbol{\beta}) F(\mathbf{X}_t, \boldsymbol{\beta})'\right] = \mathbb{E}\left[\underbrace{\mathbb{E}\left[F(\mathbf{X}_{t+3}, \boldsymbol{\beta}) | \mathfrak{F}_{t+1}\right]}_{=0} F(\mathbf{X}_t, \boldsymbol{\beta})'\right] = 0,$$

$$\mathbb{E}\left[F(\mathbf{X}_{t-2}, \boldsymbol{\beta}) F(\mathbf{X}_t, \boldsymbol{\beta})'\right] = \mathbb{E}\left[F(\mathbf{X}_{t-2}, \boldsymbol{\beta}) \underbrace{\mathbb{E}\left[F(\mathbf{X}_t, \boldsymbol{\beta})' | \mathfrak{F}_{t-2}\right]}_{=0}\right] = 0,$$

$$\mathbb{E}\left[F(\mathbf{X}_{t-3}, \boldsymbol{\beta}) F(\mathbf{X}_t, \boldsymbol{\beta})'\right] = \mathbb{E}\left[F(\mathbf{X}_{t-3}, \boldsymbol{\beta}) \underbrace{\mathbb{E}\left[F(\mathbf{X}_t, \boldsymbol{\beta})' | \mathfrak{F}_{t-2}\right]}_{=0}\right] = 0.$$

So, any cross products with time subscript difference greater or equal to two is zero. But, note that we have not imposed

restrictions on $\mathbb{E}\left[F\left(\mathbf{X}_{t+1},\boldsymbol{\beta}\right)F\left(\mathbf{X}_t,\boldsymbol{\beta}\right)'\right]$, $\mathbb{E}\left[F\left(\mathbf{X}_t,\boldsymbol{\beta}\right)F\left(\mathbf{X}_t,\boldsymbol{\beta}\right)'\right]$, or $\mathbb{E}\left[F\left(\mathbf{X}_{t-1},\boldsymbol{\beta}\right)F\left(\mathbf{X}_t,\boldsymbol{\beta}\right)'\right]$. This means that

$$\begin{aligned}
\mathbf{V} &= \lim_{N\to\infty}\mathbb{E}\left[\left(\frac{1}{\sqrt{N}}\sum_{t=1}^{N}F\left(\mathbf{X}_t,\boldsymbol{\beta}\right)\right)\left(\frac{1}{\sqrt{N}}\sum_{t=1}^{N}F\left(\mathbf{X}_t,\boldsymbol{\beta}\right)\right)'\right]\\
&= \lim_{N\to\infty}\mathbb{E}\left[\frac{1}{N}\sum_{t=1}^{N}F\left(\mathbf{X}_t,\boldsymbol{\beta}\right)F\left(\mathbf{X}_t,\boldsymbol{\beta}\right)'\right] + \lim_{N\to\infty}\mathbb{E}\left[\frac{1}{N}\sum_{t=1}^{N-1}F\left(\mathbf{X}_t,\boldsymbol{\beta}\right)F\left(\mathbf{X}_{t+1},\boldsymbol{\beta}\right)'\right]\\
&\quad + \lim_{N\to\infty}\mathbb{E}\left[\frac{1}{N}\sum_{t=2}^{N}F\left(\mathbf{X}_t,\boldsymbol{\beta}\right)F\left(\mathbf{X}_{t-1},\boldsymbol{\beta}\right)'\right]\\
&= \lim_{N\to\infty}\mathbb{E}\left[\frac{1}{N}\sum_{t=1}^{N}F\left(\mathbf{X}_t,\boldsymbol{\beta}\right)F\left(\mathbf{X}_t,\boldsymbol{\beta}\right)'\right] + \lim_{N\to\infty}\mathbb{E}\left[\frac{N-1}{N}\frac{1}{N-1}\sum_{t=1}^{N-1}F\left(\mathbf{X}_t,\boldsymbol{\beta}\right)F\left(\mathbf{X}_{t+1},\boldsymbol{\beta}\right)'\right]\\
&\quad + \lim_{N\to\infty}\mathbb{E}\left[\frac{N-1}{N}\frac{1}{N-1}\sum_{t=2}^{N}F\left(\mathbf{X}_t,\boldsymbol{\beta}\right)F\left(\mathbf{X}_{t-1},\boldsymbol{\beta}\right)'\right].
\end{aligned}$$

Equivalently, we can visualise above as:

$$\begin{aligned}
&\mathbb{E}\left[F\left(\mathbf{X}_t,\boldsymbol{\beta}\right)\left(F\left(\mathbf{X}_j,\boldsymbol{\beta}\right)\right)'\right]\\
&= \begin{bmatrix}
\mathbb{E}\left[F_1^tF_1^j\right] & \mathbb{E}\left[F_1^tF_2^j\right] & 0 & \cdots & \cdots & 0\\
\mathbb{E}\left[F_2^tF_1^j\right] & \mathbb{E}\left[F_2^tF_2^j\right] & \mathbb{E}\left[F_2^tF_3^j\right] & \ddots & \ddots & \vdots\\
0 & \mathbb{E}\left[F_3^tF_2^j\right] & \mathbb{E}\left[F_3^tF_3^j\right] & \ddots & \ddots & \vdots\\
\vdots & \ddots & \ddots & \ddots & \ddots & 0\\
\vdots & \ddots & \ddots & \ddots & \mathbb{E}\left[F_{r-1}^tF_{r-1}^j\right] & \mathbb{E}\left[F_{r-1}^tF_r^j\right]\\
0 & \cdots & \cdots & 0 & \mathbb{E}\left[F_r^tF_{r-1}^j\right] & \mathbb{E}\left[F_r^tF_r^j\right]
\end{bmatrix}.
\end{aligned}$$

By law of large numbers,

$$\begin{aligned}
\mathbf{V} &= \mathbb{E}\left[F\left(\mathbf{X}_t,\boldsymbol{\beta}\right)F\left(\mathbf{X}_t,\boldsymbol{\beta}\right)'\right] + \lim_{N\to\infty}\frac{N-1}{N}\mathbb{E}\left[F\left(\mathbf{X}_t,\boldsymbol{\beta}\right)F\left(\mathbf{X}_{t+1},\boldsymbol{\beta}\right)'\right]\\
&\quad + \lim_{N\to\infty}\frac{N-1}{N}\mathbb{E}\left[F\left(\mathbf{X}_t,\boldsymbol{\beta}\right)F\left(\mathbf{X}_{t-1},\boldsymbol{\beta}\right)'\right]\\
&= \mathbb{E}\left[F\left(\mathbf{X}_t,\boldsymbol{\beta}\right)F\left(\mathbf{X}_t,\boldsymbol{\beta}\right)'\right] + \mathbb{E}\left[F\left(\mathbf{X}_t,\boldsymbol{\beta}\right)F\left(\mathbf{X}_{t+1},\boldsymbol{\beta}\right)'\right] + \mathbb{E}\left[F\left(\mathbf{X}_t,\boldsymbol{\beta}\right)F\left(\mathbf{X}_{t-1},\boldsymbol{\beta}\right)'\right]\\
&= \mathbb{E}\left[G_tG_t'\right] + \mathbb{E}\left[G_tG_{t+1}'\right] + \mathbb{E}\left[G_tG_{t-1}'\right]
\end{aligned}$$

and we can estimate $\mathbf{V}$ by the sample counterpart.

Note that 2SLS attains the efficiency bound.

7.6 Testing with GMM

7.6.1 Calibration & Verification

1. **Calibrate** (= estimate) using k equations:

$$\mathbb{E}\left[F_1\left(\mathbf{X}_t,\boldsymbol{\beta}\right)\right] = \mathbf{0}$$

where F_1 has k coordinates.

2. **Verify** (= test) using $r-k$ equations:

$$\mathbb{E}\left[F_2\left(\mathbf{X}_t, \boldsymbol{\beta}\right)\right] = \mathbf{0}$$

while taking into account the initial estimation.

▷ To test this, use the fact that

$$\frac{1}{\sqrt{N}}\sum_{t=1}^{N} F_2\left(X_t, \mathbf{b}_N\right) \approx B'\left[I - D\left(A'D\right)^{-1}A'\right]\frac{1}{\sqrt{N}}\sum_{t=1}^{N} F\left(X_t, \boldsymbol{\beta}\right)$$

where $A' = \begin{bmatrix} I & 0 \end{bmatrix}$ and $B' = \begin{bmatrix} 0 & I \end{bmatrix}$.

▷ Rewriting the above, partition $D' = \begin{bmatrix} D_1 & D_2 \end{bmatrix}$ to write:

$$\frac{1}{\sqrt{N}}\sum_{t=1}^{N} F_2\left(X_t, \mathbf{b}_N\right) \approx \begin{bmatrix} -D_2\left(D_1\right)^{-1} & I \end{bmatrix}\frac{1}{\sqrt{N}}\sum_{t=1}^{N} F\left(X_t, \boldsymbol{\beta}\right)$$

and use a chi-square test.

7.6.2 Testing with Efficient Estimation

1. Set up the moment restriction and use the efficient selection matrix $A' = D'V^{-1}$.

2. Denote Λ such that $V^{-1} = \Lambda'\Lambda$. Since:

$$\frac{1}{\sqrt{N}}\sum_{t=1}^{N} F_2\left(X_t, \mathbf{b}_N\right) \approx B'\left[I - D\left(A'D\right)^{-1}A'\right]\frac{1}{\sqrt{N}}\sum_{t=1}^{N} F\left(X_t, \boldsymbol{\beta}\right)$$

Multiplying each side by Λ and plugging in $A' = D'V^{-1}$:

$$\frac{1}{\sqrt{N}}\sum_{t=1}^{N} \Lambda F_2\left(X_t, \mathbf{b}_N\right) = B'\left[I - \Lambda D\left(D'\Lambda'\Lambda D\right)^{-1}D'\Lambda'\right]\frac{1}{\sqrt{N}}\sum_{t=1}^{N} \Lambda F\left(X_t, \boldsymbol{\beta}\right)$$

Denote $\Delta = \Lambda D\left(D'\Lambda'\Lambda D\right)^{-1}D'\Lambda'$ and recognize that Δ and $I-\Delta$ are both idempotent. Δ has rank k and $I-\Delta$ has rank $r-k$.

3. Then we have:

$$\left(\frac{1}{\sqrt{N}}\sum_{t=1}^{N} \Lambda F\left(\mathbf{X}_t, \boldsymbol{\beta}\right)\right)'\frac{1}{\sqrt{N}}\sum_{t=1}^{N} \Lambda F\left(\mathbf{X}_t, \boldsymbol{\beta}\right) \Rightarrow \chi^2\left(r\right)$$

which we can split into two components:

$$\left(\frac{1}{\sqrt{N}}\sum_{t=1}^{N} \Lambda F\left(\mathbf{X}_t, \boldsymbol{\beta}\right)\right)'\left(I-\Delta\right)\frac{1}{\sqrt{N}}\sum_{t=1}^{N} \Lambda F\left(\mathbf{X}_t, \boldsymbol{\beta}\right) \Rightarrow \chi^2\left(r-k\right)$$

$$\left(\frac{1}{\sqrt{N}}\sum_{t=1}^{N} \Lambda F\left(\mathbf{X}_t, \boldsymbol{\beta}\right)\right)'\Delta\frac{1}{\sqrt{N}}\sum_{t=1}^{N} \Lambda F\left(\mathbf{X}_t, \boldsymbol{\beta}\right) \Rightarrow \chi^2\left(k\right)$$

4. If you replace $\boldsymbol{\beta}$ with $\mathbf{b}_N$, you lose degrees of freedom.

7.7 Arguing for Consistency

We want to show

$$\frac{1}{\sqrt{N}}\sum_{t=1}^{N} F\left(\mathbf{X}_t, \mathbf{b}\right) \to \mathbb{E}\left[F\left(\mathbf{X}_t, \mathbf{b}\right)\right]$$

i.e. a functional version of LLN.

7.8 Application: Exchange Rates

We consider regressions of the form:

$$\log S_{t+\ell} = \beta_0 + \beta_1 \log F_t + U_{t+\ell}$$

with $\mathbb{E}\left[U_{t+\ell}|\mathfrak{F}_t\right] = 0$. It's not actually a regression since we don't have exogenous regressors i.e regressors are pre-determined. It turns out that you can still do least squares since

$$\begin{aligned}\mathbb{E}\left[U_{t+\ell}\right] &= 0\\ \mathbb{E}\left[\log F_t U_{t+\ell}\right] &= 0\end{aligned}$$

and we can use GMM to work out the OLS distribution.

Can you do better? We know that combining the properties above, we have

$$\mathbb{E}\left[Z_t U_{t+\ell}\right] = 0, \forall Z_t \in \mathfrak{F}_t$$

so I can put in other stuff. The Us will be serially correlated i.e. $U_{t+\ell}$ and $U_{t+\ell-1}$ will be correlated. You can put in lag regressors and improve efficiency. What you cannot do is put regressors going forward $(F_{t+1}, F_{t+2}, ...)$. GLS can let you go both directions.

7.9 Key GMM Equations

Baseline restriction:

$$\mathbb{E}\left[F\left(\mathbf{X}, \mathbf{b}\right)\right] = \mathbf{0}_{r\times 1} \text{ if and only if } \mathbf{b} = \beta$$

Sample counterpart of the moment condition

$$\mathbf{A}'\frac{1}{N}\sum_{t=1}^{N} F_t\left(\mathbf{X}_t, \mathbf{b}_N\right) = \mathbf{0}.$$

GMM asymptotic normality

$$\frac{1}{\sqrt{N}}\sum_{t=1}^{N} F\left(\mathbf{X}_t, \boldsymbol{\beta}\right) \sim N\left(0, \mathbf{V}\right),$$

where

$$\begin{aligned}\mathbf{V} &= \mathbb{E}\left[\mathbf{H}_{t+1}\mathbf{H}_{t+1}'\right]\\ &\approx \lim_{N\to\infty} \mathbb{E}\left[\left(\frac{1}{\sqrt{N}}\sum_{t=1}^{N} F\left(\mathbf{X}_t, \boldsymbol{\beta}\right)\right)\left(\frac{1}{\sqrt{N}}\sum_{j=1}^{N} F\left(\mathbf{X}_j, \boldsymbol{\beta}\right)\right)'\right]\end{aligned}$$

and $\mathbf{H}_{t+1}$ is the martingale component of $F\left(\mathbf{X}_t, \boldsymbol{\beta}\right)$.

Efficient selection matrix

$$\mathbf{A}^* = \mathbf{V}^{-1}\mathbf{D},$$

where $\mathbf{D}_{r\times k} = \mathbb{E}\left[\partial F\left(\mathbf{X}_t, \boldsymbol{\beta}\right)/\partial \mathbf{b}'\right]$. Note that $\mathbf{V}$ also has the interpretation of the weighting matrix.

Approximation

$$\sqrt{N}\left(\mathbf{b}_N - \boldsymbol{\beta}\right) \approx -\left(\mathbf{A}'\mathbf{D}\right)^{-1}\mathbf{A}'\frac{1}{\sqrt{N}}\sum_{t=1}^{N} F\left(\mathbf{X}_t, \boldsymbol{\beta}\right)$$

Hypothesis testing with GMM For both tests outlined below, you construct Wald tests that is distributed according to χ^2.

1. Testing coefficient estimates:

$$\sqrt{N}\left(\mathbf{b}_N - \boldsymbol{\beta}\right) \approx -\left(\mathbf{A}'\mathbf{D}\right)^{-1}\mathbf{A}'\frac{1}{\sqrt{N}}\sum_{t=1}^{N} F\left(\mathbf{X}_t, \boldsymbol{\beta}\right)$$

2. Testing moment conditions:

$$\frac{1}{\sqrt{N}}\sum_{t=1}^{N} F\left(\mathbf{X}_t, \mathbf{b}_N\right) \approx \left(\mathbb{I}_r - \mathbf{D}\left(\mathbf{A}'\mathbf{D}\right)^{-1}\mathbf{A}'\right)\frac{1}{\sqrt{N}}\sum_{t=1}^{N} F\left(\mathbf{X}_t, \boldsymbol{\beta}\right)$$

   Note that premultiplying the RHS matrix is equal to zero:

$$\mathbf{A}'\left(\mathbb{I}_r - \mathbf{D}\left(\mathbf{A}'\mathbf{D}\right)^{-1}\mathbf{A}'\right) = \mathbf{0}$$

   and this matrix is adjusting for the estimation of β.

7.10 Revisiting Linear Models via GMM

7.10.1 Base line

Let $\mathbf{Y}_t \in \mathbb{R}^{k+1}$, $\boldsymbol{\beta} \in \mathbb{R}^k$, and $u_t \in \mathbb{R}$ such that

$$\mathbf{Y}_t'\begin{bmatrix} 1 \\ \boldsymbol{\beta} \end{bmatrix} = u_t.$$

Our goal is to estimate $\boldsymbol{\beta}$ consistently. Let $\mathfrak{F}_{t-\ell}$ denote the information set in period $t-\ell$ and $\mathbf{Z}_{t-\ell} \in \mathfrak{F}_{t-\ell} \subset \mathbb{R}^r$.

Moment condition We suppose that, for some ℓ,

$$\mathbb{E}\left[u_t|\mathfrak{F}_\tau\right]\begin{cases} = 0 & \text{if } \tau \le t-\ell, \\ \neq 0 & \text{if } t-\ell < \tau \le t+\ell. \end{cases}$$

If $\ell = 0$, then

$$\begin{aligned} \mathbb{E}\left[u_t|\mathfrak{F}_t\right] &= 0, \\ \mathbb{E}\left[u_t|\mathfrak{F}_\tau\right] &\neq 0,\ \forall \tau \neq t. \end{aligned}$$

If, instead, $\ell = 1$, then

$$\begin{aligned} \mathbb{E}\left[u_t|\mathfrak{F}_{t-1}\right] = \mathbb{E}\left[u_t|\mathfrak{F}_{t-2}\right] = \cdots &= 0, \\ \mathbb{E}\left[u_t|\mathfrak{F}_t\right], \mathbb{E}\left[u_t|\mathfrak{F}_{t+1}\right] &\neq 0. \end{aligned}$$

Thus, the choice of ℓ determines the correlation between the error term u_τ for $t-\ell < \tau \leq t+\ell$. The above conditional moment restriction implies that

$$\begin{aligned}
&\mathbb{E}\left[u_t|\mathfrak{F}_{t-\ell}\right] = 0 \\
\Rightarrow\ &\mathbf{Z}_{t-\ell}\mathbb{E}\left[u_t|\mathfrak{F}_{t-\ell}\right] = \mathbf{Z}_{t-\ell}0 \\
\Leftrightarrow\ &\mathbb{E}\left[\mathbf{Z}_{t-\ell}u_t|\mathfrak{F}_{t-\ell}\right] = \mathbf{0}_{r\times 1} \\
\Rightarrow\ &\mathbb{E}\left[\mathbb{E}\left[\mathbf{Z}_{t-\ell}u_t|\mathfrak{F}_{t-\ell}\right]\right] = \mathbf{0}_{r\times 1} \\
\Leftrightarrow\ &\mathbb{E}\left[\mathbf{Z}_{t-\ell}u_t\right] = \mathbf{0}_{r\times 1};
\end{aligned}$$

i.e. it implies that $\mathbf{Z}_{t-\ell}$ is orthogonal to u_t. Let

$$\mathbf{X}_t \coloneqq \begin{bmatrix} \mathbf{Y}_t \\ \mathbf{Z}_{t-\ell} \end{bmatrix}$$

and partition $\mathbf{Y}_t$ as

$$\mathbf{Y}_t = \begin{bmatrix} Y_t^1 \\ \mathbf{Y}_t^2 \end{bmatrix},$$

where Y_t^1 is a scalar and $\mathbf{Y}_t^2$ is an $k \times 1$ vector. Define

$$\begin{aligned}
F\left(\mathbf{X}, \boldsymbol{\beta}\right)_{r\times 1} &\coloneqq \mathbf{Z}_{t-\ell}u_t \\
&= \mathbf{Z}_{t-\ell}\mathbf{Y}_t' \begin{bmatrix} 1 \\ \boldsymbol{\beta} \end{bmatrix} \\
&= \mathbf{Z}_{t-\ell} \begin{bmatrix} Y_t^1 & \mathbf{Y}_t^{2\prime} \end{bmatrix} \begin{bmatrix} 1 \\ \boldsymbol{\beta} \end{bmatrix} \\
&= \mathbf{Z}_{t-\ell}\left(Y_t^1 + \mathbf{Y}_t^{2\prime}\boldsymbol{\beta}\right).
\end{aligned}$$

We can then write the orthogonality condition, (**??**), as

$$\mathbb{E}\left[F\left(\mathbf{X}, \boldsymbol{\beta}\right)\right] = \mathbf{0}_{r\times 1} = \mathbb{E}\left[\mathbf{Z}_{t-\ell}u_t\right].$$

Thus, we see that the GMM moment condition coincides with the orthogonality condition for IV. The sample analog of the moment restriction is

$$\begin{aligned}
\mathbf{0}_{r\times 1} &= \frac{1}{N}\sum_{t=1}^{N} F\left(\mathbf{X}_t, \mathbf{b}\right) \\
&= \frac{1}{N}\sum_{i=1}^{N} \mathbf{Z}_{t-\ell}\left(Y_t^1 + \mathbf{Y}_t^{2\prime}\mathbf{b}\right).
\end{aligned}$$

Derivative matrix Consider

$$\frac{\partial F\left(\mathbf{X}_t, \boldsymbol{\beta}\right)}{\partial \mathbf{b}'} = \frac{\partial \mathbf{Z}_{t-\ell}\left(Y_t^1 + \mathbf{Y}_t^{2\prime}\mathbf{b}\right)}{\partial \mathbf{b}'} = \mathbf{Z}_{t-\ell}\mathbf{Y}_t^{2\prime}.$$

Since F is linear, we realise that the derivative does not involve $\boldsymbol{\beta}$. Then,

$$\mathbf{D} \coloneqq \mathbb{E}\left[\frac{\partial F\left(\mathbf{X}_t, \boldsymbol{\beta}\right)}{\partial \mathbf{b}'}\right] = \mathbb{E}\left[\mathbf{Z}_{t-\ell}\mathbf{Y}_t^{2\prime}\right].$$

The sample analog, $\mathbf{D}_N$, is then

$$\mathbf{D}_N = \frac{1}{N}\sum_{t=1}^{N}\frac{\partial F\left(\mathbf{X}_t, \mathbf{b}\right)}{\partial \mathbf{b}'} = \frac{1}{N}\sum_{t=1}^{N}\mathbf{Z}_{t-\ell}\mathbf{Y}_t^{2\prime}.$$

Note $\mathbf{D}_N \xrightarrow{\text{p}} \mathbf{D}$.

Selection matrix We have r linear equations and k unknowns. The selection matrix, $\mathbf{A}$, thus has dimension $r \times k$.

Estimating $\boldsymbol{\beta}$ Premultiplying the sample analog of the moment restriction by the (transpose of the) selection matrix yields

$$\begin{aligned}
\mathbf{0}_{k\times 1} &= \mathbf{A}' \frac{1}{N} \sum_{t=1}^{N} F\left(\mathbf{X}_t, \mathbf{b}_N\right) \\
&= \mathbf{A}' \frac{1}{N} \sum_{i=1}^{N} \mathbf{Z}_{t-\ell} \left(Y_t^1 + \mathbf{Y}_t^{2\prime} \mathbf{b}_N\right) \\
\Leftrightarrow -\mathbf{A}' \frac{1}{N} \sum_{i=1}^{N} \mathbf{Z}_{t-\ell} \mathbf{Y}_t^{2\prime} \mathbf{b}_N &= \mathbf{A}' \frac{1}{N} \sum_{i=1}^{N} \mathbf{Z}_{t-\ell} Y_t^1 \\
\Leftrightarrow \mathbf{b}_N &= -\left(\mathbf{A}' \frac{1}{N} \sum_{i=1}^{N} \mathbf{Z}_{t-\ell} \mathbf{Y}_t^{2\prime}\right)^{-1} \left(\mathbf{A}' \frac{1}{N} \sum_{i=1}^{N} \mathbf{Z}_{t-\ell} Y_t^1\right).
\end{aligned}$$

We can see that the estimator $\mathbf{b}_N$ depends on the selection matrix $\mathbf{A}$.

7.10.2 Least squares

Given partition (**??**), we can write

$$\begin{aligned}
\mathbf{Y}_t' \begin{bmatrix} 1 \\ \boldsymbol{\beta} \end{bmatrix} = u_t &\Leftrightarrow Y_t^1 + \mathbf{Y}_t^{2\prime} \boldsymbol{\beta} = u_t \\
&\Leftrightarrow Y_t^1 = -\mathbf{Y}_t^{2\prime} \boldsymbol{\beta} + u_t,
\end{aligned}$$

which is the familiar regression form. We can then impose that

$$\mathbb{E}\left[u_t | \mathfrak{F}_t\right] = 0,$$

which implies that

$$\mathbb{E}\left[\mathbf{Y}_t u_t\right] = 0.$$

The sample analog of the moment restriction is

$$\begin{aligned}
\mathbf{0}_{r\times 1} &= \frac{1}{N} \sum_{t=1}^{N} F\left(\mathbf{X}_t, \mathbf{b}\right) \\
&= \frac{1}{N} \sum_{i=1}^{N} \mathbf{Y}_t \left(Y_t^1 + \mathbf{Y}_t^{2\prime} \mathbf{b}\right) \\
&= \frac{1}{N} \sum_{i=1}^{N} \begin{bmatrix} Y_t^1 \\ \mathbf{Y}_t^2 \end{bmatrix} \left(Y_t^1 + \mathbf{Y}_t^{2\prime} \mathbf{b}\right).
\end{aligned}$$

In this case, $r = k + 1$. Now, let

$$\mathbf{A} = \begin{bmatrix} 0 & \mathbf{0}_{1\times k} \\ \mathbf{0}_{k\times 1} & \mathbb{I}_k \end{bmatrix}_{r\times(k+1)}.$$

This selection matrix selects only $\mathbf{Y}_t^2$ from $\mathbf{Y}_t$. Then,

$$\begin{aligned}
\mathbf{A}'\mathbf{0}_{r\times 1} &= \mathbf{A}'\frac{1}{N}\sum_{i=1}^{N}\begin{bmatrix} Y_t^1 \\ \mathbf{Y}_t^2 \end{bmatrix}\left(Y_t^1 + \mathbf{Y}_t^{2\prime}\mathbf{b}_N\right) \\
\Leftrightarrow \mathbf{0}_{(k+1)\times 1} &= \frac{1}{N}\sum_{i=1}^{N}\begin{bmatrix} 0 & \mathbf{0}_{1\times k} \\ \mathbf{0}_{k\times 1} & \mathbb{I}_k \end{bmatrix}\begin{bmatrix} Y_t^1 \\ \mathbf{Y}_t^2 \end{bmatrix}\left(Y_t^1 + \mathbf{Y}_t^{2\prime}\mathbf{b}_N\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\begin{bmatrix} 0 \\ \mathbf{Y}_t^2 \end{bmatrix}\left(Y_t^1 + \mathbf{Y}_t^{2\prime}\mathbf{b}_N\right) \\
\Rightarrow \mathbf{0}_{k\times 1} &= \frac{1}{N}\sum_{i=1}^{N}\mathbf{Y}_t^2\left(Y_t^1 + \mathbf{Y}_t^{2\prime}\mathbf{b}_N\right),
\end{aligned}$$

which is the usual OLS condition. Rearranging yields that

$$\mathbf{b}_N = \left(\frac{1}{N}\sum_{i=1}^{N}\left(-\mathbf{Y}_t^2\right)\left(-\mathbf{Y}_t^2\right)'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\left(-\mathbf{Y}_t^2\right)Y_t^1\right).$$

7.10.3 Two-stage least squares

Suppose $\ell = 1$ and that

$$\mathbb{E}\left[u_t^2|\mathfrak{F}_{t-1}\right] = \sigma^2.$$

Thus, we are assuming no serial correlation and homoscedasticity. The moment restriction is, as before,

$$\mathbf{0}_{r\times 1} = \mathbb{E}\left[\mathbf{Z}_{t-1}u_t\right].$$

Efficient selection matrix Recall that the efficient selection matrix is given by

$$\mathbf{A}^* = \mathbf{V}^{-1}\mathbf{D}.$$

(we can also multiply this by any conforming nonsingular matrix.) As derived before, the derivative matrix is given by

$$\mathbf{D} \coloneqq \mathbb{E}\left[\frac{\partial F\left(\mathbf{X}_t, \boldsymbol{\beta}\right)}{\partial \mathbf{b}'}\right] = \mathbb{E}\left[\mathbf{Z}_{t-1}\mathbf{Y}_t^{2\prime}\right].$$

To obtain $\mathbf{V}$, we construct an additive martingale $\{H_\tau\}$ as

$$\begin{aligned}
H_\tau &\coloneqq \sum_{t=1}^{\tau}\mathbf{Z}_{t-1}u_t = \sum_{t=1}^{\tau}\mathbf{Z}_{t-1}\left(Y_t^1 + \mathbf{Y}_t^{2\prime}\boldsymbol{\beta}\right) \\
\Rightarrow H_\tau - H_{\tau-1} &= \mathbf{Z}_{\tau-1}u_\tau = \kappa\left(\mathbf{Z}_{\tau-1}, u_\tau\right),
\end{aligned}$$

so that $\mathbf{Z}_{\tau-1}u_\tau$ is the additive increment. Given the moment restriction, observe that

$$\mathbb{E}\left[\kappa\left(\mathbf{Z}_{\tau-1}, u_\tau\right)|\mathbf{Z}_{\tau-1}\right] = \mathbf{Z}_{\tau-1}\mathbb{E}\left[u_\tau\right] = \mathbf{0}_{r\times 1}$$

so that we verify that $\{H_\tau\}$ is indeed an additive martingale. Then, by Billingsley Central Limit Theorem,

$$\sqrt{N}\left(\frac{1}{N}\sum_{t=1}^{N}\mathbf{Z}_{t-1}u_t\right) \xrightarrow{\text{d}} N\left(0, \mathbf{V}\right),$$

where

$$\begin{aligned}\mathbf{V} &= \mathbb{E}\left[\left(\mathbf{Z}_{t-1}u_t\right)\left(\mathbf{Z}_{t-1}u_t\right)'\right] \\ &= \mathbb{E}\left[\mathbf{Z}_{t-1}\mathbb{E}\left[u_t^2|\mathfrak{F}_{t-1}\right]\mathbf{Z}_{t-1}'\right] \\ &= \sigma^2\mathbb{E}\left[\mathbf{Z}_{t-1}\mathbf{Z}_{t-1}'\right].\end{aligned}$$

Then,

$$\mathbf{A}^* = \frac{1}{\sigma^2}\mathbb{E}\left[\mathbf{Z}_{t-1}\mathbf{Z}_{t-1}'\right]^{-1}\mathbb{E}\left[\mathbf{Z}_{t-1}\mathbf{Y}_t^{2\prime}\right].$$

Recall that

$$\mathrm{Cov}\left[\mathbf{A}^*\right] = \mathrm{Cov}\left[\mathbf{A}^*\mathbf{B}\right]$$

for any nonsingular matrix $\mathbf{B}$. Letting

$$\mathbf{B} = \sigma^2\mathbb{I},$$

then

$$\tilde{\mathbf{A}}^* := \mathbf{A}^*\mathbf{B} = \mathbb{E}\left[\mathbf{Z}_{t-1}\mathbf{Z}_{t-1}'\right]^{-1}\mathbb{E}\left[\mathbf{Z}_{t-1}\mathbf{Y}_t^{2\prime}\right].$$

This gives us an efficient selection matrix. Observe that this is coefficient on $\mathbf{Z}_{t-1}'$ on the population regression of $\mathbf{Y}_t^{2\prime}$ on $\mathbf{Z}_{t-1}$. That is, the selection matrix is coefficient from the first-stage regression in 2SLS.

Relationship between GMM and IV estimators Let us consider the population regression first. The moment restriction is

$$\mathbf{0}_{r\times 1} = \mathbb{E}\left[\mathbf{Z}_{t-1}\left(Y_t^1 + \mathbf{Y}_t^{2\prime}\boldsymbol{\beta}\right)\right].$$

Premultiplying by $\tilde{\mathbf{A}}^{*\prime}$

$$\begin{aligned}\mathbf{0}_{k\times 1} &= \mathbb{E}\left[\tilde{\mathbf{A}}^{*\prime}\mathbf{Z}_{t-1}\left(Y_t^1 + \mathbf{Y}_t^{2\prime}\boldsymbol{\beta}\right)\right] \\ \Leftrightarrow \mathbb{E}\left[\tilde{\mathbf{A}}^{*\prime}\mathbf{Z}_{t-1}\mathbf{Y}_t^{2\prime}\right]\boldsymbol{\beta} &= \mathbb{E}\left[\tilde{\mathbf{A}}^{*\prime}\mathbf{Z}_{t-1}Y_t^1\right].\end{aligned}$$

Consider the best linear predictor of $\mathbf{Y}_t^{2\prime}|\mathbf{Z}_{t-1}$:

$$\mathbf{Y}_t^2 = \boldsymbol{\Pi}'\mathbf{Z}_{t-1} + \mathbf{v}_t,$$

where, as noted before, $\boldsymbol{\Pi} = \tilde{\mathbf{A}}^*$. By the property of BLP,

$$\mathbb{E}\left[\mathbf{Z}_{t-1}\mathbf{v}_t'\right] = \mathbf{0}.$$

This means that

$$\begin{aligned}\mathbb{E}\left[\tilde{\mathbf{A}}^{*\prime}\mathbf{Z}_{t-1}\left(\tilde{\mathbf{A}}^{*\prime}\mathbf{Z}_{t-1} + \mathbf{v}_t\right)'\right]\boldsymbol{\beta} &= \mathbb{E}\left[\tilde{\mathbf{A}}^{*\prime}\mathbf{Z}_{t-1}Y_t^1\right] \\ \Leftrightarrow \left(\mathbb{E}\left[\tilde{\mathbf{A}}^{*\prime}\mathbf{Z}_{t-1}\mathbf{Z}_{t-1}'\tilde{\mathbf{A}}^*\right] + \tilde{\mathbf{A}}^{*\prime}\mathbb{E}\left[\mathbf{Z}_{t-1}\mathbf{v}_t'\right]\right)\boldsymbol{\beta} &= \mathbb{E}\left[\tilde{\mathbf{A}}^{*\prime}\mathbf{Z}_{t-1}Y_t^1\right] \\ \Rightarrow \left(\tilde{\mathbf{A}}^{*\prime}\mathbb{E}\left[\mathbf{Z}_{t-1}\mathbf{Z}_{t-1}'\right]\tilde{\mathbf{A}}^*\right)\boldsymbol{\beta} &= \mathbb{E}\left[\tilde{\mathbf{A}}^{*\prime}\mathbf{Z}_{t-1}Y_t^1\right] \\ \Leftrightarrow \boldsymbol{\beta} &= \mathbb{E}\left[\tilde{\mathbf{A}}^{*\prime}\mathbf{Z}_{t-1}\mathbf{Z}_{t-1}'\tilde{\mathbf{A}}^*\right]^{-1}\mathbb{E}\left[\tilde{\mathbf{A}}^{*\prime}\mathbf{Z}_{t-1}Y_t^1\right] \\ &= \mathbb{E}\left[\tilde{\mathbf{A}}^{*\prime}\mathbf{Z}_{t-1}\mathbf{Z}_{t-1}'\tilde{\mathbf{A}}^*\right]^{-1}\mathbb{E}\left[\tilde{\mathbf{A}}^{*\prime}\mathbf{Z}_{t-1}Y_t^1\right]\end{aligned}$$

This gives us the familiar IV (population) estimator (over-identified case). Hence, in this case, the GMM estimator and the IV estimator coincide.

7.10.4 Heteroscedasticity robust estimator

Now suppose that $\mathbb{E}\left[u_t^2|\mathfrak{F}_{t-1}\right]$ is not a constant. The only difference is that we can no longer simplify $\mathbf{V}$ as we did before:

$$\begin{aligned}\mathbf{V} &= \mathbb{E}\left[\left(\mathbf{Z}_{t-1}u_t\right)\left(\mathbf{Z}_{t-1}u_t\right)'\right] \\ &= \mathbb{E}\left[u_t^2\mathbf{Z}_{t-1}\mathbf{Z}_{t-1}'\right].\end{aligned}$$

Then,

$$\mathbf{A}^* = \mathbb{E}\left[u_t^2\mathbf{Z}_{t-1}\mathbf{Z}_{t-1}'\right]^{-1}\mathbb{E}\left[\mathbf{Z}_{t-1}\mathbf{Y}_t^{2\prime}\right].$$

We can estimate $\mathbf{V}$ as

$$\hat{\mathbf{V}} = \frac{1}{N}\sum_{t=1}^{N}\hat{u}_t^2\mathbf{Z}_{t-1}\mathbf{Z}_{t-1}',$$

where $\hat{u}_t = Y_t^1 + \mathbf{Y}_t^{2\prime}\hat{\boldsymbol{\beta}}$. The TSLS estimator is no longer efficient but remains consistent.