T. Christensen

1 Structural Estimation

Over the course of the next five lectures we shall discuss estimation and inference methods for "structural" econometric models. Structural models attempt to formally model of the behavior of economic agents. Once we estimate deep policy-invariant parameters that govern agents' behavior, we may then perform policy experiments (or *counterfactuals*) to predict how agents' behavior and economic outcomes would change if we changed policy settings or other features of the economic environment. This ex-ante approach to policy evaluation is needed when experimentation with alternative policies is prohibitively costly or is otherwise infeasible. Next term we shall study a complementary set of "reduced form" methods for ex-post policy evaluation.

2 Some Motivating Examples

Rust (1987). Time is discrete, indexed by $t \in \mathbb{N}_0 := \{0, 1, 2, \ldots\}$. At each date t, an individual chooses an action a from a finite set $\mathcal{A} = \{0, 1, 2, \ldots, A\}$ to solve

$$V_\theta(s_t) = \mathrm{E}\left[\max_{a\in\mathcal{A}} \left(u(s_t, a; \theta_1) + \varepsilon_{t,a} + \beta \mathrm{E}[V_\theta(s_{t+1})|s_t, a; \theta_2]\right)\right], \tag{1}$$

where θ_1 is a vector of parameters indexing utilities, s_t is an observable (to the econometrician) Markov state taking values in a finite state-space $\mathcal{S} = \{0, 1, 2, \ldots, S\}$ with transition distribution $f(s_{t+1}|s_t, a; \theta_2)$ which is known up to the parameter vector θ_2, $\varepsilon_t = (\varepsilon_{t,0}, \varepsilon_{t,1}, \ldots, \varepsilon_{t,A})$ is a vector of unobservable (to the econometrician) utility shocks which are typically assumed to be IID Gumbel,[1] and $\beta \in [0, 1)$ is a discount parameter. The outer expectation is an expectation over ε_t holding s_t fixed while the inner expectation is over s_{t+1} conditional on s_t, a. We wish to estimate model parameters $\theta = (\theta_1, \theta_2)$ so that we may perform policy experiments.

We begin by noting two simplifying implications of the Gumbel assumption. First, using a closed-form expression for the expectation of the maximum of independent Gumbel random variables, we have

$$V_\theta(s_t) = \gamma + \log\left(\sum_{a'\in\mathcal{A}} \mathrm{e}^{u(s_t,a';\theta_1)+\beta \mathrm{E}[V_\theta(s_{t+1})|s_t,a';\theta_2]}\right), \tag{2}$$

[1] The Gumbel distribution has cdf $F(x) = \exp(-\exp(-x))$.

where $\gamma \approx 0.52277$ is the Euler-Mascheroni constant. The conditional choice probability (CCP) of choosing a in state s_t is also available in closed-form:

$$P(a|s_t;\theta) = \Pr\left(a = \arg\max_{a' \in \mathcal{A}} \left(u(s_t, a';\theta_1) + \varepsilon_t(a') + \beta \mathrm{E}[V_\theta(s_{t+1})|s_t, a';\theta_2]\right) \middle| s_t\right) \tag{3}$$

$$= \frac{e^{u(s_t,a;\theta_1)+\beta \mathrm{E}[V_\theta(s_{t+1})|s_t,a;\theta_2]}}{\sum_{a' \in \mathcal{A}} e^{u(s_t,a';\theta_1)+\beta \mathrm{E}[V_\theta(s_{t+1})|s_t,a';\theta_2]}} . \tag{4}$$

Note that when the agent is myopic ($\beta = 0$) the model reduces to a standard multinomial logit model and the choice probabilities have the usual multinomial logit form for a static problem. Finally, note that according to the model, (s_t, a_t) is Markovian and the transition distribution factorizes as

$$\Pr(a_t, s_t|a_{t-1}, s_{t-1}) = \Pr(a_t|s_t) \Pr(s_t|a_{t-1}, s_{t-1}) \tag{5}$$

$$= P(a_t|s_t;\theta) f(s_t|s_{t-1}, a_{t-1};\theta_2) \tag{6}$$

which allows us to write the log-likelihood as the sum of the log-likelihood of the CCPs and the log-likelihood for the law of motion of the observable component of the state s_t.

Two sampling schemes are typically used. The first is to assume that we observe a single time-series of the state s_t and the agent's action a_t for a sequence of n periods, so our data are $(s_0, a_0), \ldots, (s_n, a_n)$. In this case, we have the likelihood function

$$Q_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} \left(\log P(a_t|s_t;\theta) + \log f(s_t|s_{t-1}, a_{t-1};\theta_2)\right) . \tag{7}$$

Alternatively, we might observe a panel of data on independent, identical agents, $(s_{it}, a_{it})_{t=0}^{T}$ for $i = 1, \ldots, n$ and fixed T, in which case we have the log-likelihood

$$Q_n(\theta) = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} \left(\log P(a_{it}|s_{it};\theta) + \log f(s_{it}|s_{it-1}, a_{it-1};\theta_2)\right) . \tag{8}$$

These likelihoods are "approximate" since we condition on (s_0, a_0). To work out the full likelihood we could infer the stationary distribution of (s_0, a_0) and then add on the date-0 contribution. Ignoring the date-0 contribution will not matter in large samples for time-series data, but could bias estimates for panel data with fixed T.

To compute $Q_n(\theta)$ we proceed in two steps. First, solve the recursion (2) for V_θ. Then, given V_θ, form the likelihood using expression (4). We maximize Q_n with respect to θ to obtain the *maximum likelihood estimate* (MLE) $\hat{\theta}$ of θ. This estimation method is sometimes referred to as Rust's nested fixed point algorithm. There is an active literature on estimating single and multiple-agent dynamic discrete choice models and dynamic discrete games.

Hansen and Singleton (1982). Consider a representative household that chooses its consumption and investment plan to maximize lifetime utility

$$\sum_{t=0}^{\infty} \delta^t \frac{C_t^{1-\gamma} - 1}{1-\gamma} \tag{9}$$

subject to a budget constraint. Standard arguments deliver the Euler equation

$$\mathrm{E}\left[\delta \left(\frac{C_{t+1}}{C_t} \right)^{-\gamma} R_{i,t+1} \middle| \mathcal{I}_t \right] = 1 \,, \tag{10}$$

where $\mathcal{I}_t$ is the information set of the agent at date t and $R_{i,t+1}$ is the gross return on asset i from date t to date $t+1$.

We will use (10) to estimate the preference parameters δ and γ from time-series data on consumption growth C_{t+1}/C_t, a vector of asset returns $R_{t+1} = (R_{1,t+1}, \ldots, R_{l,t+1})'$, and a vector of conditioning variables z_t that belong to $\mathcal{I}_t$ (i.e., $\sigma(z_t) \subseteq \mathcal{I}_t$). Then by the law of iterated expectations and (10), at the true values of (δ, γ) we have

$$\begin{aligned}
\mathrm{E}\left[\left(\delta \left(\frac{C_{t+1}}{C_t} \right)^{-\gamma} R_{t+1} - 1 \right) \otimes z_t \right] &= \mathrm{E}\left[\mathrm{E}\left[\left(\delta \left(\frac{C_{t+1}}{C_t} \right)^{-\gamma} R_{t+1} - 1 \right) \otimes z_t \middle| \mathcal{I}_t \right] \right] && (11) \\
&= \mathrm{E}\left[\mathrm{E}\left[\left(\delta \left(\frac{C_{t+1}}{C_t} \right)^{-\gamma} R_{t+1} - 1 \right) \middle| \mathcal{I}_t \right] \otimes z_t \right] && (12) \\
&= 0 \,, && (13)
\end{aligned}$$

where 1 denotes a conformable vector of ones. Let $\theta = (\delta, \gamma)$ and

$$g(X_t; \theta) = \left(\delta \left(\frac{C_{t+1}}{C_t} \right)^{-\gamma} R_{t+1} - 1 \right) \otimes z_t \,, \tag{14}$$

where $X_t = (C_{t+1}/C_t, R_{t+1}', z_t')'$. We know from (13) that the population moment condition

$$\mathrm{E}[g(X_t; \theta)] = 0 \tag{15}$$

must hold at the true value of θ. The idea of the *generalized method of moments* (GMM) is that we choose an estimate $\hat{\theta}$ of θ such that the sample average of $g(X_t; \theta)$ is as close to zero as possible. Let $\widehat{W}$ be a positive definite symmetric matrix. The sample criterion function is

$$Q_n(\theta) = -\frac{1}{2} \left(\frac{1}{n} \sum_{t=1}^{n} g(X_t; \theta) \right)' \widehat{W} \left(\frac{1}{n} \sum_{t=1}^{n} g(X_t; \theta) \right) . \tag{16}$$

Maximizing Q_n with respect to θ yields the GMM estimate $\hat{\theta}$.

Benhabib, Bisin, and Luo (2019). To identify what drives wealth dynamics in the US, the authors build a heterogeneous agent life-cycle model which exhibits various wealth accumulation factors. Solving the model leads to a stochastic process for wealth across generations and other attributes of heterogeneity. Their description of their empirical framework is as follows:

> We estimate the parameters of the model described in the previous section using a [Simulated Method of Moments (SMM)] estimator: i) we fix (or externally calibrate) several parameters of the model; ii) we select some relevant moments of the wealth process as target in the estimation; and iii) we estimate the remaining parameters by matching the targeted moments generated by the stationary distribution induced by the model and those in the data...
>
> We target as moments: the bottom 20%, 20–39%, 40–59%, 60–79%, 80–89%, 90–94%, 95–99%, and the top 1% wealth percentiles; and the diagonal of the social mobility Markov chain transition matrix defined over the same percentile ranges as states. iii) We estimate: the preference parameters μ, A; and a parameterization of the stochastic process for r ...
>
> In total, therefore, we target 15 moments and we estimate 12 parameters...
>
> We use bootstrapping to generate the standard errors for the statistics related to the return process, e.g. its mean, standard deviation, and autocorrelation coefficient. The procedure is standard. We take the parameter values for generating the return process as given, i.e. the values for the five Markov states and the diagonal matrix of the transition matrix (hence the whole Markov transition matrix), then generate the return process a sufficiently large number of times. We then calculate the mean, standard errors and the autocorrelation coefficient directly using these series of the return processes.

Abstractly, we can write their estimation procedure as maximizing

$$Q_n(\theta) = -\frac{1}{2}(g_n - \gamma_m(\theta))'\widehat{W}(g_n - \gamma_m(\theta)) \tag{17}$$

where g_n are the moments being targeted, $\gamma_m(\theta)$ are the counterparts that are simulated from the model (averaged over m simulations) at parameter value θ, and $\widehat{W}$ is a weight matrix. The resulting estimator $\hat{\theta}$ is the *simulated method of moments* (SMM) estimator. Unlike GMM estimation where the only source of randomness is sampling variation, here there is a second source of randomness that comes from using simulation to compute the model-implied moments. Both sources of randomness play a role in determining the large-sample properties of simulation-based estimators.

3 Preliminaries

3.1 Data

We will work throughout on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is the underlying uncertainty space, $\mathcal{F}$ is the collection of all events whose likelihood of occurrence the model will restrict, and $\mathbb{P}$ is the probability measure that assigns the likelihood of events in $\mathcal{F}$.

The data we observe are a sequence $X_1, \ldots, X_n$ where n is the sample size and each of the n observations, denoted X_t, is a random vector (i.e. a vector of random variables). The sample is a map from Ω to $\mathbb{R}^{n \times \dim(X)}$, so we may sometimes use the map $\omega \mapsto X_1(\omega), \ldots, X_n(\omega)$ to reflect this. We will always assume this map is *measurable*, in the sense that statements we will make about $X_1, \ldots, X_n$ correspond to events in $\mathcal{F}$.

3.2 Sampling Schemes

We will also allow for two sampling schemes, namely data that are *independent and identically distributed* (IID) or weakly dependent. IID data are most common in cross-sectional econometrics or micro-econometrics when the data are from cross-sectional surveys at the individual, household or firm level (though there may of course be clustering; we ignore this for now). Under IID sampling, the random variables $X_1, \ldots, X_n$ are independent and each observation X_t is a draw from a fixed probability distribution (and hence is "identically distributed"). With IID data, the strong law of large numbers (SLLN) asserts

$$\frac{1}{n} \sum_{t=1}^{n} f(X_t) \to_{a.s.} \mathrm{E}[f(X_t)] \tag{18}$$

for each function f for which $\mathrm{E}[f(X_t)]$ is finite.

As structural models often have a dynamic component, we also require a sampling scheme for time-series data. In this case we view $X_1, \ldots, X_n$ as a segment of the realization of a stochastic processes. Here we shall require that the data are *strictly stationary and ergodic*, which is a notion of "weak" dependence. There are two parts to this definition: strict stationarity requires that the distribution of $(X_{t_1+h}, \ldots, X_{t_k+h})$ be independent of h, for each $t_1, \ldots, t_k$. Ergodicity is a technical condition ensuring that time-series averages converge to spatial averages, i.e.:

$$\frac{1}{n} \sum_{t=1}^{n} f(X_t) \to_{a.s.} \mathrm{E}[f(X_t)] \tag{19}$$

for each function f for which $\mathrm{E}[f(X_t)]$ is finite. This convergence (19) is the result of what is known formally as the *Ergodic Theorem*, which generalizes the SLLN to weakly dependent data.

For instance, any stationary Markov process is ergodic under very mild conditions. This is useful as structural models with a dynamic component typically impose Markovianity for tractability.

3.3 Convergence Concepts and Mann–Wald Notation

Let $Z_1, Z_2, \ldots$ be a sequence of random vectors. Say that $Z_1, Z_2, \ldots$ *converges in probability* to Z_0 if

$$\Pr(\|Z_n - Z_0\| > \epsilon) \to 0 \quad \text{as} \quad n \to \infty \tag{20}$$

for each $\epsilon > 0$. If so, we write $Z_n \to_p Z_0$. Say that $Z_n = o_p(1)$ if $Z_n \to_p 0$. Notice that $Z_n \to_p Z_0$ if and only if $Z_n - Z_0 \to_p 0$. We use $Z_n \to_p Z_0$ and $Z_n = Z_0 + o_p(1)$ interchangeably. We can think of $o_p(\cdot)$ as being like a stochastic version of the $o(\cdot)$ notation from analysis.

Say that $Z_1, Z_2, \ldots$ *converges almost surely* to Z_0 if

$$\mathbb{P}\Big(\big\{\omega : \lim_{n\to\infty} Z_n(\omega) = Z_0(\omega)\big\}\Big) = 1\,. \tag{21}$$

If so, we write $Z_n \to_{a.s.} Z_0$. Recall that almost sure convergence implies convergence in probability, i.e. $Z_n \to_{a.s.} Z_0$ implies $Z_n \to_p Z_0$. Say that $Z_n = o_{a.s.}(1)$ if $Z_n \to_{a.s.} 0$. Notice that $Z_n \to_{a.s.} z_0$ if and only if $Z_n - Z_0 \to_{a.s.} 0$. We use $Z_n \to_{a.s.} Z_0$ and $Z_n = Z_0 + o_{a.s.}(1)$ interchangeably.

Let $\alpha_1, \alpha_2, \ldots$ be a sequence of positive constants. Say that $Z_n = o_p(\alpha_n^{-1})$ (resp. $Z_n = o_{a.s.}(\alpha_n^{-1})$) if $\alpha_n Z_n = o_p(1)$ (resp. $\alpha_n Z_n = o_{a.s.}(1)$).

Say that $Z_1, Z_2, \ldots$ is *bounded in probability* or *tight* if for each $\epsilon > 0$ there exists a $M_\epsilon < \infty$ such that

$$\limsup_{n\to\infty} \Pr(\|Z_n\| > M_\epsilon) \leq \epsilon\,. \tag{22}$$

If so, we write $Z_n = O_p(1)$. Similarly, say that $Z_n = O_p(\alpha_n)$ if $\alpha_n Z_n = O_p(1)$. We can think of $O_p(\cdot)$ as being like a stochastic version of the $O(\cdot)$ notation from analysis. We have the following relations:

$$o_p(a_n) \times o_p(b_n) = o_p(a_n b_n)\,, \tag{23}$$

$$O_p(a_n) \times o_p(b_n) = o_p(a_n b_n)\,, \text{ and} \tag{24}$$

$$O_p(a_n) \times O_p(b_n) = O_p(a_n b_n)\,. \tag{25}$$

The same relations hold if we replace the o_p and O_p terms by $o_{a.s.}$ and $O_{a.s.}$ terms.

For example, let $X_1, \ldots, X_n$ be IID random vectors with mean μ and finite covariance matrix Σ. Then the sample mean $\bar{X}_n \to_{a.s.} \mu$ by the SLLN (and hence $\bar{X}_n \to_p \mu$). By the central limit theorem we have that $\sqrt{n}(\bar{X}_n - \mu) \to_d N(0, \Sigma)$ hence $(\bar{X}_n - \mu) = O_p(n^{-1/2})$ and $\bar{X}_n = \mu + O_p(n^{-1/2})$.

4 Extremum Estimators and Examples

Consider a class of economic models indexed by a parameter of interest, say θ. The set Θ is the *parameter space*, which is the set of plausible values of θ. We say $\hat{\theta}$ is an *extremum estimator* if it is a (measurable) function of the data $X_1, \ldots, X_n$ taking values in Θ and it satisfies

$$Q_n(\hat{\theta}) > \sup_{\theta \in \Theta} Q_n(\theta) - \eta_n \,, \tag{26}$$

where η_n is a sequence of positive random variables such that $\eta_n \to_p 0$ and $Q_n : \Theta \to \mathbb{R}$ is a *sample criterion function* which depends on both θ and the data $X_1, \ldots, X_n$. The function Q_n is a random function: different realizations of the sample $X_1, \ldots, X_n$ may produce a different sample criterion function, and therefore a different maximizer $\hat{\theta}$. We include the $-\eta_n$ term because in practice we typically use numerical methods for maximizing Q_n and therefore we are not guaranteed that the optimization routine will converge to the exact maximum.

Throughout we denote the *true* value of θ by θ_0. The true value $\theta_0 \in \Theta$ solves

$$Q(\theta_0) \geq Q(\theta) \quad \text{for all } \theta \in \Theta \text{ with } \theta \neq \theta_0 \,, \tag{27}$$

where $Q : \Theta \to \mathbb{R}$ is the population criterion function. Loosely speaking, this is the function to which Q_n would converge if we saw an infinite amount of data.

Say θ_0 is *(point) identified* if

$$Q(\theta_0) > Q(\theta) \quad \text{for all } \theta \in \Theta \text{ with } \theta \neq \theta_0 \,. \tag{28}$$

The first five lectures defined identification in terms of the full set of restrictions that the model contains about θ. The two definitions agree when Q exploits all such restrictions. Some care may need to be taken to do this. In the Hansen and Singleton (1982) example above, the model gives a conditional moment restriction against an information set $\mathcal{I}_t$ but we use only a subset of this information for estimation. It may still be that θ_0 is identified through the GMM population criterion function using only a subset of the information. But in this case the resulting estimate may not be as efficient as it could have been had we exploited the full set of model restrictions.

4.1 M-Estimators

M-estimators are based on the sample criterion function of the form

$$Q_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} m(X_t; \theta) \,. \tag{29}$$

M-estimators differ in their choice of m. With IID or SSE data the sample average $\frac{1}{n}\sum_{t=1}^{n} m(X_t;\theta)$ converges a.s. to $\mathrm{E}[m(X_t;\theta)]$ by the SLLN or Ergodic Theorem. The *population criterion function* $Q:\Theta\to\mathbb{R}$ is therefore

$$Q(\theta)=\mathrm{E}[m(X_t;\theta)]\,. \tag{30}$$

Example 4.1. Linear Regression.
Suppose $(X_t,Y_t)_{t=1}^{n}$ are IID or SSE with Y_t a scalar and X_t a vector of dimension p. Suppose that both Y_t and X_t have finite second moments. The best linear predictor for Y_t given X_t is $X_t'\theta_0$, where θ_0 solves

$$\min_{\theta\in\mathbb{R}^p}\mathrm{E}[(Y_t-X_t'\theta)^2]\,. \tag{31}$$

Equivalently, θ_0 maximizes the population criterion function

$$Q(\theta)=-\frac{1}{2}\mathrm{E}[(Y_t-X_t'\theta)^2]\,. \tag{32}$$

Expanding the quadratic, we see that θ_0 uniquely maixmizes Q whenever $\mathrm{E}[X_tX_t']$ has full rank. In regression jargon, this is the no-multicollinearity condition.

The sample criterion function is

$$Q_n(\theta)=-\frac{1}{2}\frac{1}{n}\sum_{t=1}^{n}(Y_t-X_t'\theta)^2\,. \tag{33}$$

By the usual argument, we see that the least-squares estimator

$$\hat{\theta}=\left(\frac{1}{n}\sum_{t=1}^{n}X_tX_t'\right)^{-1}\left(\frac{1}{n}\sum_{t=1}^{n}X_tY_t\right) \tag{34}$$

maximizes Q_n. This is a rare example of a situation in which we can solve for the estimator in closed form. □

Example 4.2. Maximum Likelihood (Unconditional).
Suppose $X_1,\ldots,X_n$ are IID draws from a distribution with density $f(x;\theta_0)$ with respect to a common dominating measure μ. Continuous distributions have densities with respect to Lebesgue measure; discrete distributions have densities with respect to counting measure. The joint density of $X_1,\ldots,X_n$ is

$$f(X_1,\ldots,X_n;\theta_0)=\prod_{t=1}^{n}f(X_t;\theta_0)\,. \tag{35}$$

Of course, we do not know the true θ_0. The principle of maximum likelihood asserts that we should choose θ to maximize the "probability" of having observed $X_1,\ldots,X_n$. Maximization of the

probability

$$\prod_{t=1}^{n} f(X_t;\theta) \tag{36}$$

with respect to θ is equivalent to maximizing

$$Q_n(\theta) = \frac{1}{n}\sum_{t=1}^{n} \log f(X_t;\theta) \tag{37}$$

with respect to θ. (Unconditional) maximum likelihood is therefore a M-estimator whose m function is the log density: $m(X_t;\theta) = \log f(X_t;\theta)$.

We now verify the identification condition (28). The population criterion function is

$$Q(\theta) = \mathrm{E}[\log f(X_t;\theta)]\,. \tag{38}$$

Suppose that the model is correctly specified. That is, $X_1,\ldots,X_n$ are IID draws from a distribution with density $f(X_t;\theta_0)$. This means when we take expectations over functions of X we integrating respect to $f(\,\cdot\,;\theta_0)$. Then

$$\begin{aligned} Q(\theta) &= \mathrm{E}[\log f(X_t;\theta)] - \mathrm{E}[\log f(X_t;\theta_0)] + \mathrm{E}[\log f(X_t;\theta_0)] && (39)\\ &= -\mathrm{E}\Big[\log\Big(\frac{f(X_t;\theta_0)}{f(X_t;\theta)}\Big)\Big] + \mathrm{E}[\log f(X_t;\theta_0)] && (40)\\ &= -K(f(\,\cdot\,;\theta_0)\|f(\,\cdot\,;\theta)) + \mathrm{E}[\log f(X_t;\theta_0)]\,. && (41)\end{aligned}$$

where

$$K(f\|g) = \int f\log(f/g)\,d\mu \tag{42}$$

is the *Kullback-Leibler divergence* (or *relative entropy*) of g from f. This divergence is a measure of the discrepancy between densities. It has the properties that (i) $K(f\|g) \geq 0$ and (ii) $K(f\|g) = 0$ if and only if $f = g$ holds f-almost everywhere. To see why property (i) holds, by convexity of $-\log(x)$ and Jensen's inequality, we have

$$K(f\|g) = \int -\log(g/f) f\,d\mu \geq -\log\left(\int (g/f) f\,d\mu\right) = -\log\left(\int g\,d\mu\right) = -\log(1) = 0.$$

For property (ii), we note that if $f = g$ holds f-almost everywhere then $\log(f/g) = 0$ holds f-almost everywhere and hence

$$K(f\|g) = \int \log(f/g) f\,d\mu = \int 0\,d\mu = 0\,.$$

Conversely, suppose $f \neq g$ with positive f-measure. Then by Jensen's inequality, which holds strictly here because $-\log x$ is strictly convex, we have

$$K(f\|g) = \int -\log(g/f) f\,d\mu > -\log\left(\int g\,d\mu\right) = 0.$$

Consider (41). As $\mathrm{E}[\log f(X_t;\theta_0)]$ does not vary with θ, we see that maximizing Q is equivalent to minimizing $K(f(\,\cdot\,;\theta_0)\|f(\,\cdot\,;\theta))$. By properties (i) and (ii), a sufficient condition for the identification condition (28) is therefore

$$\Pr(f(X_t;\theta) \neq f(X_t;\theta_0)) > 0 \quad \text{for all} \quad \theta \neq \theta_0\,, \tag{43}$$

where the probability is with respect to the true distribution (under θ_0). This is a type of "rank" condition that says different θ must induce noticeably different distributions. We can usually verify this condition directly from the model. □

Example 4.3. Maximum Likelihood (Markov Processes).
Suppose that $Y_0, \ldots, Y_n$ are generated by a SSE Markov process with transition density $f(Y_t|Y_{t-1};\theta_0)$. Once again, "density" means with respect to a dominating measure μ. The Rust (1987) model falls into this framework with $Y_t = (s_t, a_t)$ and some special structure on the transition density.

Suppose that the functional form f is known (i.e., computable) but θ_0 is unknown. Let $X_t = (Y_t, Y_{t-1})$ for $t = 1, \ldots, n$. If θ was the true parameter, the joint density of $X_1, \ldots, X_n$ would be

$$f(X_1, \ldots, X_n;\theta) = \prod_{t=1}^{n} f(Y_t|Y_{t-1};\theta) \times f_0(Y_0;\theta) \tag{44}$$

where f_0 is the unconditional (or stationary) density of Y_t. The average log likelihood is

$$\frac{1}{n}\sum_{t=1}^{n} \log f(Y_t|Y_{t-1};\theta) + \frac{1}{n}\log f(Y_0;\theta)\,. \tag{45}$$

As n gets large the second term becomes negligible. Therefore, maximizing the average log likelihood is approximately the same as maximizing

$$Q_n(\theta) = \frac{1}{n}\sum_{t=1}^{n} \log f(Y_t|Y_{t-1};\theta)\,. \tag{46}$$

Hence we have a M-estimator with $m(X_t;\theta) = \log f(Y_t|Y_{t-1};\theta)$. The population criterion function is now

$$Q(\theta) = \mathrm{E}[\log f(Y_t|Y_{t-1};\theta)]\,, \tag{47}$$

where the expectation is with respect to the joint (stationary) distribution of (Y_t, Y_{t-1}).

We may use similar arguments to write the maximum likelihood for higher-order Markov processes (i.e. Markov processes for which the distribution of Y_t depends on $Y_{t-1}, \ldots, Y_{t-k}$ but not also on Y_{t-j} for any $j < k$) as M-estimators.

We now verify the identification condition (28). Suppose that the model is correctly specified. Notice that

$$Q(\theta) = -\mathrm{E}\Big[\mathrm{E}\Big[\log\Big(\frac{f(Y_t|Y_{t-1};\theta_0)}{f(Y_t|Y_{t-1};\theta)}\Big)\Big|Y_{t-1}\Big]\Big] + \mathrm{E}[\log f(Y_t|Y_{t-1};\theta_0)]\,. \tag{48}$$

In the first expression on the right-hand side, the outer expectation is with respect to the stationary distribution of Y_{t-1} implied by $f(\cdot\,|\,\cdot;\theta_0)$ while the inner expectation is taken with respect to $Y_t \sim f(\cdot\,|Y_{t-1};\theta_0)$. By similar arguments to the unconditional case, we see that θ_0 is identified if for all $\theta \neq \theta_0$ there is a set of Y_{t-1} with positive probability (under the stationary distribution) upon which

$$\Pr(f(Y_t|Y_{t-1};\theta) \neq f(Y_t|Y_{t-1};\theta_0)|Y_{t-1}) > 0, \tag{49}$$

where the Pr is taken with respect to $f(\cdot\,|Y_{t-1};\theta_0)$. In words, this says that for each $\theta \neq \theta_0$ there exist states Y_{t-1} that we visit with positive probability and in which the transition distribution $f(\cdot\,|Y_{t-1};\theta)$ is noticeably different from $f(\cdot\,|Y_{t-1};\theta_0)$.

As a simple example, consider the AR(1) model

$$Y_t = (1-\rho)\mu + \rho Y_{t-1} + u_t, \tag{50}$$

where the u_t are IID $N(0,\sigma^2)$ and $\rho \in (-1,1)$. The conditional distribution of Y_t given Y_{t-1} is

$$N((1-\rho)\mu + \rho Y_{t-1}, \sigma^2). \tag{51}$$

Each of the three parameters (μ,ρ,σ^2) is identified: varying σ^2 changes the dispersion of the distribution, varying μ changes the mean in a common way across all states, and varying ρ changes the mean in a state-dependent way.

The stationary (or unconditional) distribution of Y_t may be shown to be

$$Y_t \sim N(\mu, (1-\rho^2)^{-1}\sigma^2). \tag{52}$$

A naive estimation strategy would be to ignore the dynamics and attempt to estimate (μ,ρ,σ^2) based on the likelihood of the unconditional distribution. While μ is identified from the stationary distribution, the parameters (ρ,σ^2) are not, as we can choose different (ρ,σ^2) pairs to produce the same value of the unconditional variance $(1-\rho^2)^{-1}\sigma^2$. □

4.2 GMM

The M-estimators we just looked at are based on the idea that there is some "true" parameter θ_0 that uniquely maximizes $\mathrm{E}[m(X_t;\theta)]$ over the parameter space $\Theta \subseteq \mathbb{R}^p$. GMM is based upon the idea that that there exists a "true" parameter θ_0 in the parameter space Θ that uniquely sets a

vector of population moments to zero, i.e., $\mathrm{E}[g(X_t;\theta_0)] = 0$ where $g : \mathbb{R}^{\dim(X)} \times \Theta \to \mathbb{R}^K$ is jointly measurable in the data X_t and θ. We assume there $K \geq p := \dim(\theta)$ moment conditions.

The Hansen and Singleton (1982) example motivated a particular choice of g from the first-order condition for optimality of a consumption/investment plan. The idea of GMM applies more broadly. For instance, we may g so that it consists of the moments that we really want our model to explain. For instance, in structural models of the labor market we may want to explain quantities like mean wages, mean wage changes among employed individuals, and durations of unemployment, etc, and would choose g accordingly.

Formally, $\hat{\theta}$ is a *GMM estimator* if it is an extremum estimator in the sense of display (26) where

$$Q_n(\theta) = -\frac{1}{2} g_n(\theta)' \widehat{W} g_n(\theta) \quad \text{with} \quad g_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} g(X_t;\theta)\,, \tag{53}$$

and where $\widehat{W}$ is a $K \times K$ positive-definite symmetric weight matrix that may be a function of the data. Letting W (another positive-definite and symmetric weight matrix) denote the probability limit of W and noting that $g_n(\theta) \to_{a.s.} \mathrm{E}[g(X_t;\theta)]$ by the SLLN or Ergodic Theorem, we see that the population criterion function $Q : \Theta \to \mathbb{R}$ is

$$\begin{aligned} Q(\theta) &= -\frac{1}{2} \mathrm{E}[g(X_t;\theta)]' W \mathrm{E}[g(X_t;\theta)] &(54)\\ &= -\frac{1}{2} g(\theta)' W g(\theta) \quad \text{where} \quad g(\theta) = \mathrm{E}[g(X_t;\theta)]\,. &(55) \end{aligned}$$

We say the GMM model is *correctly specified* if there exists some θ_0 in Θ for which $\mathbb{E}[g(X_t;\theta_0)] = 0$. This is a weaker notion of correct specification than with maximum likelihood (ML). With ML the model was correctly specified if there was a value θ_0 that gave us the true distribution of the data; here the model is correctly specified if there is a θ_0 that matches the moments we care about (i.e., those we put in the g function), but not possibly the full distribution of the data.

The model is *identified* if there is only one such θ_0 that sets the population moments to zero:

$$\mathbb{E}[g(X_t;\theta)] = 0 \quad \iff \quad \theta = \theta_0\,. \tag{56}$$

If the GMM model is identified, then $Q(\theta)$ is uniquely maximized at θ_0. This is a weaker notion of identification for maximum likelihood: it may be that $\mathbb{E}[g(X_t;\theta)] = 0$ for two distinct values of θ, but these values generate different distributions over observables. Hence, maximum likelihood would distinguish the two but GMM would not. In this sense, GMM is a "limited information" estimation procedure as it may not exhaust the full amount of information that the model contains about the parameters of interest.

GMM generalizes IV estimation. In the special case of an IV model

$$Y_{1t} = Y_{2t}'\theta + u_t, \quad \mathrm{E}[u_t Z_t] = 0,$$

where Y_2 is an endogenous regressor and Z is a vector of instruments, we have

$$g(X_t;\theta) = Z_t(Y_{1t} - Y_{2t}'\theta)$$

with $X_t = (Y_{1t}, Y_{2t}, Z_t)$. Here the dimension K of g is the dimension of the IV Z_t. By analogy with linear IV estimation, clearly we need the number of moments K to be at least as large as the number of parameters p for identification. If $K = p$ and the model is identified then we say it is *just identified*. If $K > p$ and the model is identified we say it is *over identified*. In this case, the additional $K - p$ restrictions are testable.

Common choices of $\widehat{W}$ are the $K \times K$ identity matrix or the two-step "optimal" weight matrix

$$\widehat{W} = \left(\frac{1}{n}\sum_{t=1}^{n} g(X_t;\tilde{\theta})g(X_t;\tilde{\theta})'\right)^{-1} \tag{57}$$

where $\tilde{\theta}$ is a consistent estimator of θ_0 (such as a GMM estimator we've computed using the identity weight matrix). Later we will discuss when this approach is optimal, and in what sense. There are also other estimators for which $\widehat{W}$ itself depends on θ. These fall into the class of "generalized empirical likelihood" estimators, which are outside the scope of this course.

4.3 SMM

A special case of GMM arises when the moment conditions are *separable* in parameters and data:

$$g(X_t;\theta) = g(X_t) - \gamma(\theta) \tag{58}$$

where $g : \mathbb{R}^{\dim(X)} \to \mathbb{R}^K$ and $\gamma : \Theta \to \mathbb{R}^K$. Although this looks a bit restrictive, this situation will arise when $\mathrm{E}[g(X_t)]$ are "targeted" population moments that we want to match and $\gamma(\theta)$ are their model-implied counterparts. As we don't observe the true population moments, we estimate them using the sample average:

$$g_n = \frac{1}{n}\sum_{t=1}^{n} g(X_t) \tag{59}$$

If we know the functional form for γ, then we can use GMM to estimate θ by maximizing

$$Q_n(\theta) = -\frac{1}{2}(g_n - \gamma(\theta))'\widehat{W}(g_n - \gamma(\theta)) \tag{60}$$

for some $K \times K$ weight matrix $\widehat{W}$. This is sometimes referred to as estimating the model by "calibration". The matrix $\widehat{W}$ may be chosen to prioritize fitting some moments more than others.

Suppose the model is sufficiently complicated that there is no closed-form expression for $\gamma(\theta)$. For example, your model might have heterogeneous agents or firms and specify the equilibrium cross-sectional distribution of various quantities as

$$f(X_t|\theta)\,. \tag{61}$$

The model-implied aggregate moments would then be

$$\gamma(\theta) = \int \gamma(X_t;\theta) f(X_t|\theta)\, d\mu(X_t)\,. \tag{62}$$

This integral might be difficult to calculate numerically, even if X is reasonably low-dimensional. The idea of SMM is to use simulation to approximate the integral in the above display. That is, generate a large number of observations $X_1^\theta, \ldots, X_m^\theta$ from $f(X|\theta)$ then set

$$\gamma_m(\theta) = \frac{1}{m}\sum_{s=1}^{m} \gamma(X_s^\theta;\theta)\,. \tag{63}$$

To generate X_s^θ, one typically simulates a vector of primitive shocks ε_s, then solves the model to obtain X_s^θ. SMM can be computationally burdensome to implement when solving the model is a non-trivial exercise. The usual caveats about numerical integration also apply.

Formally, $\hat{\theta}$ is a *SMM estimator* if it is an extremum estimator in the sense of display (26) for which the sample criterion function is

$$Q_n(\theta) = -\frac{1}{2}(g_n - \gamma_m(\theta))'\widehat{W}(g_n - \gamma_m(\theta)) \tag{64}$$

for some $K \times K$ weight matrix $\widehat{W}$. Note here there are two sources of uncertainty that must be accounted for when computing standard errors, namely (i) sampling uncertainty from the data $X_1, \ldots, X_n$ used to compute g_n and (ii) additional uncertainty introduced by replacing the true $\gamma(\theta)$ by a simulation-based estimate $\gamma_m(\theta)$. The population version of the criterion function is

$$Q(\theta) = -\frac{1}{2}(g_0 - \gamma(\theta))'W(g_0 - \gamma(\theta)) \tag{65}$$

where g_0 and W denote the probability limits of g_n and $\widehat{W}$, respectively.

Note: when implementing SMM, it is important to use the same random seed for generating $X_1^\theta, \ldots, X_m^\theta$ for each θ. Otherwise, you can get different values of $Q_n(\theta)$ for the same θ, so whatever numerical procedure you use to maximize $Q_n(\theta)$ will not converge.

4.4 SMD

Simulated minimum distance (SMD) estimation generalizes SMM to a setting where g_n are "sample statistics" that are not necessarily expressible as "moments". For example, g_n may be quantiles of a wealth distribution. The idea of *minimum distance* (MD) estimation is to estimate θ by minimizing the distance between g_n and its true model-implied counterpart $\gamma(\theta)$. Say $\hat{\theta}$ is a *MD estimator* if it is an extremum estimator in the sense of display (26) for which the sample criterion function is

$$Q_n(\theta) = -\frac{1}{2}(g_n - \gamma(\theta))'\widehat{W}(g_n - \gamma(\theta))\,. \tag{66}$$

for some $K \times K$ weight matrix $\widehat{W}$. When the expressions for $\gamma(\theta)$ are not available in closed form but we can simulate from the model, we can compute estimates $\gamma_m(\theta)$ based on simulation. We say $\hat{\theta}$ is a *simulated minimum distance* (SMD) estimator if the sample criterion function is

$$Q_n(\theta) = -\frac{1}{2}(g_n - \gamma_m(\theta))'\widehat{W}(g_n - \gamma_m(\theta))\,. \tag{67}$$

The population version of these criterion functions is the same as in display (65).

5 Misspecification

When the model is misspecified, we interpret the maximizer θ_0 of the population criterion function Q as the *pseudo-true* parameter. That is, θ_0 is no longer the true data-generating parameter but rather a parameter that brings the (misspecified) model "closest", in a certain sense, to the true data-generating process.

5.1 Maximum Likelihood

To fix ideas, consider the simplest case of unconditional maximum likelihood. Suppose the data $X_1, \ldots, X_n$ are IID from a distribution with true density $g(x)$. Repeating the argument from Example 4.2 above, we have

$$\begin{aligned} Q(\theta) &= \mathrm{E}[\log f(X_t;\theta)] - \mathrm{E}[\log g(X_t)] + \mathrm{E}[\log g(X_t)] && (68)\\ &= -\mathrm{E}\Big[\log\Big(\frac{g(X_t)}{f(X_t;\theta)}\Big)\Big] + \mathrm{E}[\log g(X_t)] && (69)\\ &= -K(g\|f(\,\cdot\,;\theta)) + \mathrm{E}[\log g(X_t)]\,, && (70)\end{aligned}$$

where the expectation is taken with respect to $X \sim g$. Evidently, maximizing Q is equivalent to minimizing

$$K(g \| f(\cdot\,; \theta)) \tag{71}$$

with respect to θ. The pseudo-true parameter θ_0 is therefore that which brings the model-implied distribution $f(\cdot\,;\theta)$ as close as possible to the true distribution g, where we measure closeness by Kullback–Leibler divergence.

There are infinitely many other ways of measuring "distance" between probability measures: take any non-negative convex function ϕ with $\phi(1) = 0$ and set

$$D_\phi(f \| g) = \int \phi(f/g) g \, d\mu;$$

Kullback–Leibler divergence corresponds to the special case of $\phi(x) = x \log x - x + 1$. We could in principle construct different criterion functions based on different ϕ functions or different notions of distance. Under correct specification these would all have the true parameter as their population maximizer. However, under misspecification the pseudo-true parameters may depend on the choice of criterion function. It is not clear why the parameter that minimizes Kullback–Leibler divergence, as opposed to some other divergence, is the interesting one in this case.

5.2 GMM

A GMM model is misspecified if $\mathrm{E}[g(X_t;\theta)] \neq 0$ for all $\theta \in \Theta$. The pseudo-true parameter θ_0 is still well defined as the maximizer of the population criterion function. However, here θ_0 will depend implicitly on the asymptotic weight matrix W. Different choices of $\widehat{W}$ and hence different W correspond to different estimands under misspecification. We can see this most clearly in the case of linear IV:

$$Y_{1t} = Y_{2t}'\theta + u_t, \quad \mathrm{E}[u_t Z_t] = 0.$$

Suppose that the model is over identified (i.e., the number of instruments K exceeds the number of regressors p). The model is misspecified if there is no $\theta \in \Theta$ for which $\mathrm{E}[Z_t(Y_{1t} - Y_{2t}'\theta)] = 0$.[2] The population criterion function is

$$Q(\theta) = -\frac{1}{2}\mathrm{E}[Z_t(Y_{1t} - Y_{2t}'\theta)]' W \mathrm{E}[Z_t(Y_{1t} - Y_{2t}'\theta)].$$

This is a quadratic function of θ so we can solve for the maximizer in closed-form to obtain

$$\theta_0 = \left(\mathrm{E}[Y_{2t}Z_t']W\mathrm{E}[Z_tY_{2t}']\right)^{-1}\mathrm{E}[Y_{2t}Z_t']W\mathrm{E}[Z_tY_{1t}]\,,$$

[2] If the model is just identified, we can always solve the moment condition by setting $\theta_0 = \mathrm{E}[Z_tY_{2t}']^{-1}\mathrm{E}[Z_tY_{1t}]$.

which is well defined whenever $\mathrm{E}[Y_{2t}Z_t']$ has full rank p. As $K > p$, we cannot simplify this expression further. The pseudo-true parameter θ_0 therefore depends on W. That is, $\theta_0 = \theta_0(W)$ and so different choices of weight matrix lead to different estimands.

Additional References

- Benhabib, J., A. Bisin, and M. Luo (2019). Wealth Distribution and Social Mobility in the US: A Quantitative Approach. *American Economic Review* 109(5), 1623–1647.
- Hansen, L. P. and K. Singleton (1982). Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models. *Econometrica* 50(5), 1269–1286.
- McFadden, D. (1989). A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration. *Econometrica*, 57(5), 995–1026.
- Rust, J. (1987). Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher. *Econometrica* 55(5), 999–1033.
- White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica* 50(1), 1–25.