EULER EQUATIONS, NONLINEAR GMM, AND OTHER ADVENTURES

ABSTRACT. Here we analyze the Hansen-Singleton model of an optimizing agent and from the Euler equations derive a nonlinear moment condition model that we then use to estimate agent's preference parameters. This thrilling exercise leads us to consider the nonlinear GMM. We provide sufficient conditions for consistency and asymptotic normality of GMM and provide the specification test for validity of the moment conditions (J-test). We also consider the "continuously updated" version of GMM, and outline the Anderson-Rubin approach to inference under weak or partial identification. As a bonus material, we end up learning technical tools such as the extremum consistency lemma and uniform law of large numbers.

1. EULER EQUATIONS AND CONDITIONAL MOMENT RESTRICTIONS

Hansen and Singleton [10] provided an econometric framework for estimating preferences of an optimizing agent and for testing validity of the model. The representative agent maximizes the expected utility

$$\max \mathrm{E} \sum_{t=0}^{\infty} \beta^t U(C_t),$$

over random consumption streams that obey the sequence of budget constraints:

$$C_t + \sum_{j=1}^{N} P_{j,t} Q_{j,t} = \sum_{j=1}^{N} P_{j,t} Q_{j,t-1} + W_t, \quad t = 0, 1, \ldots,$$

where C_t is the consumption at time t, $P_{j,t}$ is the price of security j at time t, $Q_{j,t}$ is the amount of security j at time t, N is the number of securities, and W_t is the labor income at time t.

The optimum is characterized by the first order conditions, which can be expressed as:

$$P_{j,t} = \mathrm{E}_t \left[\beta \frac{U'(C_{t+1})}{U'(C_t)} P_{j,t+1} \right], \quad j = 1, \ldots, J,$$

where E_t is the expectation conditional on the information available at time t. We can interpret this equation as a pricing equation that says that the price of the security j, $P_{j,t}$,

is the expectation of its value tomorrow, $P_{j,t+1}$, times the stochastic discount factor,

$$\beta \frac{U'(C_{t+1})}{U'(C_t)}.$$

Thus, if we observe the consumption process of an optimizing agent, we can use the resulting stochastic discount factor to price securities. This is called the consumption-based capital asset pricing model (CCAPM). Furthermore, if we observed the consumption and price processes, we can use the moment restrictions implied by the optimizing behavior to learn about agent's preferences.

Define the total return $R_{j,t+1} := P_{j,t+1}/P_{j,t}$, which we can assume to be stationary. Let's represent $\mathrm{E}_t[\cdot]$ as the conditional expectation $\mathrm{E}[\cdot \mid Z_t]$, where Z_t are the variables that represent the information available at time t. We can rewrite the optimality condition as

$$\mathrm{E}\left[R_{j,t+1}\beta \frac{U'(C_{t+1})}{U'(C_t)} - 1 \mid Z_t \right] = 0, \quad j = 1, \ldots, J.$$

In order to set up estimation we proceed to specify the parametric form of the utility function. Hansen and Singleton use the power utility $U(x) = x^{1-\alpha}/(1-\alpha)$, which exhibits constant relative risk aversion. The parameter α determines both the risk aversion coefficient and the intertemporal elasticity of substitution for consumption. Another attractive possibility is the Epstein-Zinn [5] utility function that has two different parameters determining the intertemporal elasticity of substitution and risk aversion. With the power utility specification

$$\mathrm{E}\left[R_{j,t+1}\beta c_{t+1}^{-\alpha} - 1 \mid Z_t \right] = 0, \quad j = 1, \ldots, N,$$

where $c_{t+1} = C_{t+1}/C_t$ is the total rate of consumption growth.

In order to transit to GMM estimation we need to do some logistical work. First we define

$$\rho(Y_t, \theta) := [\rho_j(Y_t, \theta)]_{j=1}^N := [R_{j,t+1} b c_{t+1}^{-a} - 1]_{j=1}^N, \quad \theta := (a, b)',$$

where Y_t is a vector comprised of c_{t+1} and $R_{j,t+1}$ for $j = 1, \ldots, N$. Letting $\theta_0 := (\alpha, \beta)$, we have the conditional moment restriction:

$$\mathrm{E}[\rho(Y_t, \theta_0) \mid Z_t] = 0. \tag{1.1}$$

We can convert this conditional moment restriction into unconditional moment restrictions. Let $B(Z_t) := [B_1(Z_t), \ldots, B_r(Z_t)]'$ denote a vector of transformations of Z_t. Then

$$\rho(Y_t, \theta_0) \perp B(Z_t),$$

since by the law of iterated expectations:

$$\mathrm{E}\rho(Y_t, \theta_0) \otimes B(Z_t) = \mathrm{E}\underbrace{\mathrm{E}[\rho(Y_t, \theta_0) \mid Z_t]}_{=0} \otimes B(Z_t) = 0. \tag{1.2}$$

We call Z_t the raw instruments and $B(Z_t)$ the technical instruments. This step is important and we explain this step as well as the choice of transformations in detail in the next section. This means that we can now apply the GMM approach with the score function

$$g(X_t, \theta) = \rho(Y_t, \theta) \otimes B(Z_t), \quad X_t := [Y_t', B(Z_t)']'.$$

We also need to specify a weighting matrix A and its estimator. Under the rational expectation hypothesis, the stream of scores $\{g(X_t, \theta_0)\}_{t=-\infty}^{\infty}$ is an uncorrelated sequence, since the past values of X_t belong to the information set. Indeed by the law of iterated expectations,

$$\mathrm{E} g(X_t, \theta_0) g(X_{t-k}, \theta_0)' = \mathrm{E} \underbrace{\mathrm{E}[g(X_t, \theta_0) \mid X_{t-k}]}_{=0} g(X_{t-k}, \theta_0)' = 0.$$

Econometricians call such unforecastable sequences as martingale-difference sequences.

Hence using this hypothesis and imposing the assumption that $\{X_t\}_{t=-\infty}^{\infty}$ are identically distributed, we conclude that the optimal weighting matrix is $A := \Omega^{-1}$, where

$$\Omega := \lim_{n \to \infty} \mathrm{Var}(\sqrt{n} \mathbb{E}_n g(X_t, \theta_0)) = \mathrm{E}[g(X, \theta_0) g(X, \theta_0)'], \tag{1.3}$$

where $\{X_t\}_{t=1}^{n}$ denotes the available sample. Using this observation Hansen and Singleton have proceeded to specify the GMM weighting matrix as $\hat{A} = \hat{\Omega}^{-1}$ for

$$\hat{\Omega} := \widehat{\mathrm{Var}}(\sqrt{n} \mathbb{E}_n g(X_t, \theta_0)) = \mathbb{E}_n[g(X_t, \tilde{\theta}) g(X_t, \tilde{\theta})'], \tag{1.4}$$

where $\tilde{\theta}$ is a preliminary estimator of θ_0. The resulting GMM estimator $\hat{\theta}$ then obeys the general properties outlined in L3 under plausible regularity conditions.

Hansen and Singleton applied their model and estimation method to the monthly aggregate U.S. consumption, using a U.S. market index (DJI) and treasury bonds (a riskless asset) as securities. We mention their choices of technical instruments below in the context of the broader discussion. Hansen and Singleton estimated the risk aversion parameter α to lie in an interval between 0 and 1, depending on the specification of the instrument set, with large standard error, and the discount factor of .997 with a very small standard error. They proceeded to statistically reject the assumptions of the rational expectations model of the representative consumer with the power utility, using the test based on J-statistics (the minimized value of the GMM objective function times n). We describe this test below.

2. From Conditional Restrictions to Unconditional Restrictions

Here we describe in detail the transition from the conditional to unconditional restrictions. As in the previous section we consider vector-valued structural residual function $\rho(Y, \theta)$ that takes values in $\mathbb{R}^N$ and θ is the parameter.

Suppose that the conditional moment restriction holds at $\theta = \theta_0$:

$$\mathrm{E}[\rho(Y, \theta_0) \mid Z] = 0. \tag{2.1}$$

Consider $B(Z) := [B_1(Z), \ldots, B_r(Z)]'$ a vector of transformations of Z. Suppose that $\mathrm{E}\rho_j^2(Y, \theta_0) < \infty$ and $\mathrm{E}B_k^2(Z) < \infty$ for each j and k. Then the structural residual function $\rho(Y, \theta_0)$ is orthogonal to any such $B(Z)$:

$$\rho(Y, \theta_0) \perp B(Z)$$

that is,

$$\mathrm{E}[\rho(Y, \theta_0) \otimes B(Z)] = \mathrm{E}[\overbrace{\mathrm{E}[\rho(Y, \theta_0) \mid Z]}^{=0} \otimes B(Z)] = 0. \tag{2.2}$$

This follows from the law of iterated expectation, using the fact that the left side of the preceding display is finite:

$$\mathrm{E}|\rho_j(Y, \theta_0) B_k(Z)| \leq \sqrt{\mathrm{E}\rho_j^2(Y, \theta_0)} \sqrt{\mathrm{E}B_k^2(Z)} < \infty,$$

where the inequality holds by the Cauchy-Schwarz inequality.

This motivates using Z as well as transformations of Z as technical instruments $B(Z)$. Suppose that $\mathrm{E}[\rho_j(Y, \theta) \mid Z] \neq 0$ with positive probability whenever $\theta \neq \theta_0$. This is the identification assumption for the conditional moment restriction. Given this assumption, we want to choose $B(Z)$ in order to capture deviations from zero of the following regression function:

$$f(Z) = \mathrm{E}[\rho_j(Y, \theta) | Z]$$

whenever $\theta \neq \theta_0$. If we can choose $B(Z)$ that are correlated with $f(Z)$, i.e. such that $\mathrm{E}f(Z)B(Z) \neq 0$ then we can tell apart false parameter values θ from the true one θ_0. This is akin to shopping for a set of good instruments that give us a strong "first stage". There are theoretical and practical considerations to take into account:

Consideration 1. Approximation theory provides us with dictionaries of series terms

$$B(Z) = [B_1(Z), \ldots, B_r(Z)]'$$

that can approximate any function $f(Z)$ with $\mathrm{E}f^2(Z) < \infty$ in the mean square error sense:

$$\min_{\gamma \in \mathbb{R}^r} \mathrm{E}(f(Z) - \gamma' B(Z))^2 \to 0, \quad r \to \infty. \tag{2.3}$$

This means that we can use these dictionaries as technical instruments $B(Z)$ that will be correlated with $f(Z)$, that is $\mathrm{E}f(Z)B(Z) \neq 0$, at least when r is substantial. Examples of dictionaries with the property (2.3) include power transformations, cosine transformations, wavelets, among others (see 14.381 lecture notes and e.g. Newey [12] and [13]). For instance, when Z is a scalar transformed to take values in $[0, 1]$, we can use either of

- the polynomial dictionary of size r: $B(Z) = (1, Z, Z^2, ..., Z^{r-1})'$;
- the cosine dictionary of size r: $B(Z) = (1, \cos(\pi Z), \ldots, \cos(\pi(r-1)Z))'$;
- the linear spline dictionary of size r: $B(Z) = (1, Z, (Z-k_1)_+, \ldots, (Z-k_{r-2})_+)'$;

where $k_1, \ldots k_{r-2}$ denote the knots of the spline (a mesh over $[0,1]$), and $(z)_+ = \max(z, 0)$. When Z is vector, with each component transformed to take values in $[0,1]$, we can consider dictionaries with respect to each component and then take all interactions. See 14.381 notes on regression as well as Newey reference on dictionaries of series terms used. All of the dictionaries mentioned have the approximation property (2.3), though practical performance does obviously depend on the nature of the problem.[1]

Consideration 2. Our formal asymptotic theory given below requires the dimension m of the score

$$g(X, \theta) = \rho(Y, \theta) \otimes B(Z),$$

to be fixed as $n \to \infty$, which requires r to be fixed. There are rigorous asymptotic results by Newey [12] that allow for the growth of $m = N \times r$, such that $m^2/n \to 0$, while retaining the validity of consistency and asymptotic normality results outlined in L3. The main point we should keep in mind for the canonical form of GMM that we study is the following:

The number of technical instruments r should be relatively small compared to n.

If the set of technical instruments is large, and we can't figure out using economic or other reasoning which instruments are the most important ones to keep, we can also employ variable selection methods. We shall come back to this point later.

Let's now go back to Hansen and Singleton's econometric model. The situation there is a lot more complicated: Z_t in principle could consist of infinite history, so there are many possibilities. Among these, they used $B(Z_t)$ consisting of c_t and $R_{j,t}$ as well as several lags c_{t-k} and $R_{j,t-k}$. This makes sense since very distant lags would be poor predictors for the current variables. It is interesting that Hansen and Singlton didn't consider any nonlinear transformations of any of these variables. We could only speculate as to potential reasons: one reason perhaps was that just using enough lags was sufficient for statistical rejection of the model; or maybe they were motivated by the following point:

Consideration 3. There is another view on selection of moments/technical instruments in GMM framework. If we assume that the model is wrong, then we can interpret our goal of estimating θ_0 as finding a model θ_0 within Θ that describes well certain finite collection of moments ("properties of the real world"), in the sense of minimizing $g(\theta)'Ag(\theta)$. This

[1] For instance, if $z \mapsto f(z)$ behaves like a wave, then cosine transformations will tend to do the best job, if $z \mapsto f(z)$ looks like a polynomial, then polynomials tend to the best, if $z \mapsto f(z)$ exhibits spikes, then b-splines tend to perform best. The 14.381 notes provide concrete approximation examples.

motivates us to use economic reasoning in selecting the most important moments ("properties we want to explain") to put in the GMM estimation. In this case GMM estimation becomes like a formal calibration often used by macro-economists, because they don't take their models as literal descriptions of the real world. This type of consideration is useful to keep in mind though harder to use as a guide in empirical work. For example, we may want to "calibrate" the Hansen-Singleton model to "explain" the risk premium.

3. Asymptotic Properties of Nonlinear GMM and the J- Test

3.1. **Statement and Assumptions.** Here is our formal result concerning the properties of the nonlinear GMM.

We begin with general conditions, which are of high-level nature, but show some key ingredients that are sufficient for the result.

Condition G: (a) The parameter space Θ is a compact subset of $\mathbb{R}^d$, and the true value θ_0 lies in the interior of Θ. (b) The moment function $\theta \mapsto g(\theta)$ identifies θ_0: $g(\theta) = 0$ if and only if $\theta = \theta_0$. (c) The empirical moment map $\theta \mapsto \hat{g}(\theta)$ converges uniformly in probability to the population moment map $\theta \mapsto g(\theta)$, namely $\sup_{\theta\in\Theta} \|\hat{g}(\theta) - g(\theta)\| \to_P 0$. (d) The estimator of the weighting matrix is consistent for a positive definite matrix, $\hat{A} \to_P A > 0$. (e) The empirical Jacobian $\theta \mapsto \hat{G}(\theta) = (\partial/\partial\theta')\hat{g}(\theta)$ is continuous and is uniformly consistent for the continuous population Jacobian matrix, $\theta \mapsto G(\theta) = (\partial/\partial\theta')g(\theta)$, namely $\sup_{\theta\in\Theta} \|\hat{G}(\theta) - G(\theta)\| \to_P 0$. (f) The minimal eigenvalue of $G'G$, where $G = G(\theta_0)$, is bounded away from zero. (g) The empirical moment function evaluated at the true parameter value obeys a central limit theorem:

$$\sqrt{n}\hat{g}(\theta_0) \overset{a}{\sim} N(0, \Omega).$$

In this condition only the hatted quantities depend on n and others do not.

These general conditions are quite plausible, as we discuss below. They imply the following useful formal result that justifies our Assertion 1 in L3.

Theorem 1 (Consistency and Asymptotic Normality of GMM)**.** *Under condition G the GMM estimator [8]*

$$\hat{\theta} \in \arg\min_{\theta\in\Theta} \hat{g}(\theta)'\hat{A}\hat{g}(\theta)$$

exists (yay!), and is consistent and asymptotically normal, namely

$$\sqrt{n}(\hat{\theta} - \theta_0) \overset{a}{\sim} (G'AG)^{-1}G'A\epsilon, \quad \epsilon \sim N(0, \Omega).$$

An important consequence of this result concerns the J-statistic, which is useful for testing the validity of the moment restrictions.

Theorem 2 (J-statistic)**.** *Under condition G, the minimized value of the optimally weighted GMM criterion function times the sample size is asymptotically distributed as χ^2 variable with $m - d$ degrees of freedom, namely*

$$J = n\hat{g}(\hat{\theta})'\hat{\Omega}^{-1}\hat{g}(\hat{\theta}) \overset{a}{\sim} \chi^2(m-d),$$

provided $\hat{\Omega} \to_P \Omega$.

We leave the proof of this theorem as an exercise for the theoretically minded reader. Note that the limit distribution is quite intuitive: the χ^2 distribution arises from using the quadratic form and the degrees of freedom takes the dimension m of the vector of the quadratic forms and subtracts off the dimension d of the parameter vector over which we minimized this quadratic form. This is obviously not a proof, but a useful mnemonic device to remember the result.

When $m > d$, we can use the J-statistic to test the null hypothesis of validity of the moment conditions, where the null is $H_o : \exists\theta \in \Theta : g(\theta) = 0$ and the alternative is $H_a : \nexists\theta \in \Theta : g(\theta) = 0$. Large values of the statistic J provide evidence against the null hypothesis. The J-test statistically rejects the null at the significance level p if $J > (1-p)$-quantile of $\chi^2(m-d)$.

In the context of IV analysis, as in L3 or as in Hansen-Singleton example, the J-test allows us to test the "validity of the exclusion or exogeneity restrictions," namely the hypothesis that the structural residuals are uncorrelated to the technical instruments, i.e. (1.2). The rejection of the null could be interpreted as statistical evidence against the assumption of certain instruments being excluded/exogenous *as well as* the validity of any other modeling assumptions. For example, Hansen and Singleton interpret their rejection as the statistical rejection of the representative consumer with power utility and rational expectations. The rejection is silent about which features of the model are not right.

We should be careful with the interpretation of the statistical rejection of economic models: Statistical rejection of the model does not necessarily mean that the model is not suitable for purposes of economic analysis. (See also Consideration 3 in Section 2). For example, models based on rational expectations are widely used in macro-economics; the well-known Black-Scholes [2] option pricing formula is widely used in financial economics, despite the fact that benchmark versions of these models are statistically rejected by a specification test such as the J-test.

Concise (and hence wrong) economic models can provide a coherent way to ask and (partly) answer economic questions about the real world. Statistical rejection of such models does not mean that these models necessarily give bad answers to economic questions. However, statistical tests could be useful in selecting amongst competing economic models.

3.2. **Discussion of Conditions.** We note that the stated condition G are merely sufficient for the validity of the general statement given in Assertion 1. Econometricians have provided much more general conditions than G. We focus on G here due to their concreteness and wide applicability. We discuss each of the conditions in detail.

The most important practical matter is to verify that indeed the true parameter value θ_0 is such that

$$g(\theta_0) = 0.$$

This involves careful economic thinking, e.g., as in Hansen and Singleton (1982). Nonlinearities sometimes make it analytically more difficult to verify this condition. If $g(\theta_0) \neq 0$ your GMM estimator will be consistent for the wrong parameter value, that is, inconsistent for θ_0.

Another matter is that the nonlinear case is technically more demanding, though this is less relevant for practice, because econometricians worked hard on developing plausible regularity conditions. Let's discuss these conditions systematically.

First, we impose the compactness condition, which we did not impose in the linear case. Compactness is useful in justifying that a minimum exists and in verifying uniform convergence. It can be dropped if $\hat{g}$ is a gradient of convex function, which was trivially true in the linear case. We impose that the true parameter θ_0 lies in the interior in order to perform linearization by Taylor expansion of the first order conditions, see the proof; in the linear case, the linearity held trivially.

Second, we impose the "global identification" condition

$$g(\theta) = 0 \text{ if and only if } \theta = \theta_0, \tag{3.1}$$

and the "local identification" condition that the minimal eigenvalue of $G'G$ is greater than zero, which is the same as saying that

$$\text{rank}(G) = \text{full}. \tag{3.2}$$

In the linear case the two were equivalent, and here they are not in general. We can have global identification but have deficient rank for G: for example, $g(\theta) = \theta^2 = 0$ if and only if $\theta = \theta_0 := 0$ but $G(\theta) = 2\theta = 0$ at $\theta = \theta_0$. On the other hand, (3.2) does not imply (3.1):

for example, $g(\theta) = \theta^2 - 1 = 0$ if either $\theta = 1$ or -1, and if $\theta_0 = 1$, then $G(\theta) = 2\theta = 2$ at $\theta = \theta_0$.

Note, however, that the full rank condition (3.2) implies local uniqueness of θ_0: namely, $g(\theta) = 0$ if and only if $\theta = \theta_0$ for all $\theta \in \mathcal{N}$, where $\mathcal{N}$ is an open neighborhood of θ_0 (this follows from the implicit function theorem, see Appendix). For this reason, econometricians say that the full rank condition implies "local identification".

To go further, if $G(\theta)$ has full rank at each $\theta \in \Theta$ this does imply together with compactness of Θ that the number of solutions to $g(\theta) = 0$ is finite (see Appendix). In principle, we could adjust our estimation results to handle the case with a finite number of solutions, but such cases seem to be rare in practice. It is interesting to also note that there are further strong sufficient conditions that force the number of solutions to 1 (these conditions follow from the global implicit function theorems of Mas-Colell [11]; their use has to be justified in specific applications; see [3, 4] for applications in econometrics).

Third, having $\text{mineig}(G'G)$ being bounded away from zero is important for our asymptotic results in Theorem 1 to provide good approximation to the finite-sample behavior of the estimator. Indeed, from the statement and proof we need the separation,

$$\text{mineig}(G'AG) \geq \text{mineig}(G'G)\text{mineig}(A) > 0,$$

so that the noisy version $\text{mineig}(\hat{G}'\hat{A}\hat{G})$ is well-separated from zero. If it is well-separated, then we are in the "strongly identified" case and the normal approximation is good; otherwise are in the "weakly identified" case and need to use alternative approaches to inference. Thus, in some sense, $\text{mineig}(G'G)$ measures how well θ_0 is identified, generalizing the notion of the "first stage" to the nonlinear case.

Fourth, relative to the linear case, we see the emergence of the uniform convergence requirements, $\sup_{\theta\in\Theta} \|\hat{g}(\theta) - g(\theta)\| \to_P 0$ and $\sup_{\theta\in\Theta} \|\hat{G}(\theta) - G(\theta)\| \to_P 0$. In the linear case these conditions hold automatically. We need these conditions to establish consistency and asymptotic normality; see the proof.

Let's focus on the first requirement, $\sup_{\theta\in\Theta} \|\hat{g}(\theta) - g(\theta)\| \to_P 0$. This condition is stronger than the pointwise convergence condition $\hat{g}(\theta) \to_P g(\theta)$ holding for each θ. Indeed, even deterministic uniform convergence does not in general follow from the pointwise convergence. Thus, we can not hope to obtain the uniform convergence in probability merely through the applications of the usual (pointwise) laws of large numbers. Instead, in order to verify the uniform convergence conditions, we need to employ the uniform law of large numbers – Lemma 4 stated in the Appendix – which implies the following result.

Lemma 1 (Sufficient Condition for the Uniform Convergence)**.** *Suppose that Θ is compact, that $\theta \mapsto g(X, \theta)$ is continuously differentiable with derivative $\theta \mapsto G(X, \theta)$, that the*

envelopes of these maps have bounded expectations, namely $\mathrm{E}\sup_{\theta\in\Theta}\|g(X,\theta)\| < \infty$ *and* $\mathrm{E}\sup_{\theta\in\Theta}\|G(X,\theta)\| < \infty$*, and that* $\{X_i\}$ *are i.i.d. or stationary strongly mixing series. Then* $\sup_{\theta\in\Theta}\|\hat{g}(\theta) - g(\theta)\| \to_P 0$ *and* $\sup_{\theta\in\Theta}\|\hat{G}(\theta) - G(\theta)\| \to_P 0$.

Finally, we need a central limit theorem for the empirical average of the scores.

Lemma 2 (Sufficient Condition for CLT via Gordin's CLT). *Consider the centered sequence* $\{g(X_i,\theta_0)\}_{i=1}^{\infty}$ *that is i.i.d. or stationary and strongly mixing with mixing coefficients obeying* $\sum_{j=1}^{\infty}\alpha_j^{\delta/(2+\delta)} < \infty$ *and suppose that* $\mathrm{E}\|g(X,\theta_0)\|^{2+\delta} < \infty$ *for* $\delta > 0$*, then*

$$\sqrt{n}\mathbb{E}_n g(X_i,\theta_0) \overset{a}{\sim} N(0,\Omega), \quad \Omega = \lim_{n\to\infty}\mathrm{Var}(\sqrt{n}\mathbb{E}_n g(X_i,\theta_0)),$$

where under i.i.d. sampling Ω *simplifies to* $\mathrm{E}g(X,\theta_0)g(X,\theta_0)'$*. Otherwise,*

$$\Omega = \Sigma_0 + \sum_{\ell=1}^{\infty}(\Sigma_\ell + \Sigma_\ell'), \;\; \Sigma_\ell = \mathrm{E}g(X_i,\theta_0)g(X_{i+\ell},\theta_0)'$$

Note that this is merely an application of the classical Gordin's CLT [7] from 14.381.[2] This result also motivates estimation of Ω that we have been considering so far in the i.i.d case; in the time series case, we need to consider the Newey-West [14] or various other estimators. For example, the Newey-West estimator of Ω,

$$\hat{\Omega} = \hat{\Sigma}_0 + \sum_{\ell=1}^{L}\omega_{\ell L}(\hat{\Sigma}_\ell + \hat{\Sigma}_\ell'), \;\; \hat{\Sigma}_\ell = \sum_{i=1}^{n-\ell} g(X_i,\tilde{\theta})g(X_{i+\ell},\tilde{\theta})/n, \;\; \omega_{\ell L} = 1 - \ell/(L+1),$$

is positive semidefinite.

3.3. **Proof of Theorem 1.** The proof contains three steps. In Step 1 we demonstrate consistency. In Step 2, we demonstrate asymptotic normality. In Step 3, we collect supporting, less-interesting calculations.

Step 1 (Consistency). This step establishes consistency. We have for

$$\hat{Q}(\theta) := \hat{g}(\theta)'\hat{A}\hat{g}(\theta), \quad Q(\theta) := g(\theta)'Ag(\theta),$$

that

$$\sup_{\theta\in\Theta}|\hat{Q}(\theta) - Q(\theta)| \to_P 0.$$

[2]The mixing coefficients are $\alpha_j = \sup_{A,B,n,m}|P_n(A\cap B) - P_n(A)P_n(B)|$, for A and B ranging over σ-fields generated respectively by $(X_t : 1 \le t \le m)$ and $(X_t : m + j \le t < \infty)$.

This follows from the assumed uniform convergence $\sup_{\theta\in\Theta}|\hat{g}(\theta)-g(\theta)|\to_P 0$ and $\hat{A}\to_P A$. Step 3(a) below supplies the details. Application of the Extremum Consistency Lemma gives that any $\hat{\theta}\in\arg\min_{\theta\in\Theta}\hat{Q}(\theta)$ obeys $\hat{\theta}\to_P\theta_0=\arg\min_{\theta\in\Theta}Q(\theta)$, where θ_0 is the unique minimum of $Q(\theta)$ by the identification condition. Note that both of the argmins exist (yay!) because continuous functions attain minimum on compact sets.

Step 2 (Normality). For $\hat{G}(\hat{\theta})=(\partial/\partial\theta')\hat{g}(\hat{\theta})$, a Taylor expansion of the first order conditions gives

$$0=\hat{G}(\hat{\theta})'\hat{A}\hat{g}(\hat{\theta})=\hat{G}(\hat{\theta})'\hat{A}\{\hat{g}(\theta_0)+\hat{G}(\bar{\theta})[\hat{\theta}-\theta_0]\},$$

where $\hat{G}(\bar{\theta})$ stands for $(\partial/\partial\theta')\hat{g}(\bar{\theta})$, which denotes the Jacobian matrix whose each row is evaluated at (a row-dependent) $\bar{\theta}$ located on the line joining θ_0 and $\hat{\theta}$.

The uniform convergence hypotheses, continuity hypotheses on $\theta\mapsto G(\theta)$ and consistency $\hat{\theta}\to_P\theta_0$ yield that

$$\|\hat{G}(\hat{\theta})-G\|\to_P 0,\quad \|\hat{G}(\bar{\theta})-G\|\to_P 0,\quad \text{where } G:=G(\theta_0).$$

Step 3(b) gives details.

Hence these calculations and the assumption $\hat{A}\to_P A$ yield by the continuous mapping theorem that

$$\hat{G}(\hat{\theta})'\hat{A}\to_P G'A,\qquad \hat{G}(\hat{\theta})'\hat{A}\hat{G}(\bar{\theta})\to_P G'AG,\ \ (\hat{G}(\hat{\theta})'\hat{A}\hat{G}(\bar{\theta}))^{-1}\to_P (G'AG)^{-1},$$

where $G'AG>0$ by $A>0$ and the full rank assumption on G.

Thus, with probability approaching one, we can solve for the deviation:

$$\sqrt{n}(\hat{\theta}-\theta_0)=-[\hat{G}(\hat{\theta})'\hat{A}\hat{G}(\bar{\theta})]^{-1}\hat{G}(\hat{\theta})'\hat{A}\sqrt{n}\hat{g}(\theta_0).$$

By the derivations given above, and since $\sqrt{n}\hat{g}(\theta_0)\overset{a}{\sim}\epsilon=N(0,\Omega)$ by assumption, we conclude by the continuous mapping theorem,

$$\sqrt{n}(\hat{\theta}-\theta_0)\overset{a}{\sim}(G'AG)^{-1}G'A\epsilon.$$

Step 3 (Boring Calculations). This step can be skipped on the first reading of the proof.

(a) It follows from the triangle inequality that $\sup_{\theta\in\Theta}\|\hat{g}(\theta)\| \leq \sup_{\theta\in\Theta}\|\hat{g}(\theta)-g(\theta)\| + \sup_{\theta\in\Theta}\|g(\theta)\| = O_P(1)$. Let $\|A\|$ be the maximum eigenvalue of A (operator norm). Then

$$\begin{aligned}\sup_{\theta\in\Theta}\left|\hat{Q}(\theta)-Q(\theta)\right| &\leq \sup_{\theta\in\Theta}\Big[\left|\hat{g}(\theta)'(\hat{A}-A)\hat{g}(\theta)\right| + \left|[\hat{g}(\theta)-g(\theta)]'A[\hat{g}(\theta)-g(\theta)]\right| \\ &\quad +2\left|\hat{g}(\theta)'A[\hat{g}(\theta)-g(\theta)]\right|\Big] \\ &\leq \sup_{\theta\in\Theta}\|\hat{g}(\theta)\|^2\|\hat{A}-A\| + \sup_{\theta\in\Theta}\|\hat{g}(\theta)-g(\theta)\|^2\|A\| \\ &\quad +2\sup_{\theta\in\Theta}\|\hat{g}(\theta)\|\|A\|\sup_{\theta\in\Theta}\|\hat{g}(\theta)-g(\theta)\| \to_P 0.\end{aligned}$$

(b) We have by the uniform convergence hypothesis that

$$\|\hat{G}(\hat{\theta})-G(\hat{\theta})\| \to_P 0, \quad \|\hat{G}(\bar{\theta})-G(\bar{\theta})\| \to_P 0,$$

where $G(\bar{\theta})$ stands for $(\partial/\partial\theta')g(\bar{\theta})$, which denotes the Jacobian matrix whose each row is evaluated at (a row-dependent) $\bar{\theta}$ located on the line joining θ_0 and $\hat{\theta}$. The continuity hypothesis on $\theta \mapsto G(\theta)$, consistency $\hat{\theta} \to_P \theta_0$, and $\bar{\theta} \to_P \theta_0$, and the continuous mapping theorem imply that

$$\|G(\hat{\theta})-G(\theta_0)\| \to_P 0, \quad \|G(\bar{\theta})-G(\theta_0)\| \to_P 0, \quad G(\theta_0)=G.$$

■

4. Continuously Updated GMM and Inference under Weak Identification

The continuously updated GMM estimator (CUE) [9] takes the form

$$\hat{\theta}^* \in \arg\min_{\theta\in\Theta}\hat{g}(\theta)'\hat{A}(\theta)\hat{g}(\theta), \quad \hat{A}(\theta)=\hat{\Omega}(\theta)^{-1}$$

where

$$\hat{\Omega}(\theta) = \widehat{\mathrm{Var}}(\sqrt{n}\hat{g}(\theta)).$$

This form is quite intuitive because it uses inverse of a variance matrix (indexed by θ) directly in the formulation of the GMM estimator. This estimator avoids iteration like the iterated GMM and instead uses a plug-in estimator for the variance matrix indexed by θ. Under i.i.d. sampling, the estimator is

$$\hat{\Omega}(\theta) = \mathbb{E}_n g(X,\theta)g(X,\theta)' - [\mathbb{E}_n g(X,\theta)][\mathbb{E}_n g(X,\theta)]'.$$

Under time series sampling, we can use the Newey-West estimator $\widehat{\mathrm{Var}}(\sqrt{n}\hat{g}(\theta))$ etc. The CUE has similar properties to GMM, and behaves better in some circumstances.[3]

[3]For example, in the single-equation linear IV model with Gaussian errors, CUE reduces to the limited information maximum likelihood estimator, which is known to have better finite-sample properties than the two-stage least squares when the number of instruments is large.

Theorem 3 (Consistency and Asymptotic Normality of CUE). *Suppose condition G holds and that the weighting matrices are uniformly consistent for some positive definite matrices*

$$\sup_{\theta\in\Theta} \|\hat{A}(\theta) - A(\theta)\| \to_P 0,$$

where the map $\theta \mapsto A(\theta) := \Omega(\theta)^{-1}$ *is continuous and* $\min_{\theta\in\Theta} \text{mineig} A(\theta) > 0$, *and where* $\Omega(\theta) := \lim_{n\to\infty} \text{Var}(\sqrt{n}\hat{g}(\theta))$. *Then the CUE estimator is first-order equivalent to the optimal GMM estimator:*

$$\sqrt{n}(\hat{\theta}^* - \hat{\theta}) \to_P 0,$$

and so it inherits the consistency and asymptotic normality properties of the GMM estimator.

The proof of this result is omitted.

The continuously updated formulation is particularly amenable to inference under weak or partial identification, following the Anderson-Rubin (AR) approach [1]. The weak identification arises when the minimal eigenvalue of $G'G$ is close to zero, relative to the sampling error. There is an analog of the F-test for weak identification that was developed by Wright [17].

We can formally capture this situation through a mental exercise, where we look at data streams $\{X_{i,n}\}_{i=1}^{\infty}$ of identically distributed random vectors with law F_n, but the law changes with n. We have the freedom to do this mental exercise, just like we have the freedom to do the mental exercises of the conventional asymptotics approximations where we let $n \to \infty$ but keep F fixed. In the "weak identification" exercise, as $n \to \infty$, F_n is such that the minimal eigenvalue of $G'G = G_n'G_n$ is zero or drifts to zero as $n \to \infty$, where $G_n = (\partial/\partial\theta')\text{E}g(X_{i,n}, \theta_0)$. Under this scenario the previous "strong identification" asymptotics breaks down. However, we can still rely on the following simple result for inference.

Theorem 4 (Weak Identification Robust Inference). *Suppose empirical moments are asymptotically normally distributed, namely* $\sqrt{n}(\hat{g}(\theta) - g(\theta)) \overset{\text{a}}{\sim} N(0, \Omega(\theta))$, *for each* $\theta \in \Theta$. *Assume that* $\theta \mapsto \hat{A}(\theta)$ *obeys* $\hat{A}(\theta) \to_P \Omega(\theta)^{-1}$ *for each* $\theta \in \Theta$. *Then for any* θ_0 *such that* $g(\theta_0) = 0$, *we have that*

$$W(\theta_0) = n\hat{g}(\theta_0)'\hat{A}(\theta_0)\hat{g}(\theta_0) \overset{\text{a}}{\sim} \chi^2(m).$$

The result follows trivially from the assumptions and the continuous mapping theorem. The result generalizes the previous result in L2 on the weak-identification robust inference to the general case. Note also that the result applies when $g(\theta) = 0$ has multiple solutions

(in which case the true parameter value is said to be partially identified). Consequently, the confidence region

$$CR_{1-p} = \{\theta \in \Theta : W(\theta) \leq c_{1-p}\},$$

where c_{1-p} is $(1-p)$-quantile of $\chi^2(m)$, contains θ_0 with asymptotic probability $1-p$:

$$\mathrm{P}\left(\theta_0 \in CR_{1-p}\right) = \mathrm{P}(W(\theta_0) \leq c_{1-p}) \to \mathrm{P}(\chi^2(m) \leq c_{1-p}) = 1-p, \quad n \to \infty.$$

We can approximate this confidence region in practice by specifying a grid of parameter values and then collecting all parameter values with the AR statistic less than the critical value. In the exactly identified case, when $m = d$, this approach gives a good way to perform inference in weakly identified. In the over-identified case, when $m > d$, this approach could be improved by employing other statistics; we refer to the work by Andrews and Mikusheva "Conditional Inference with a Functional Nuisance Parameter" for what appears to be the state-of-the art approach at the moment.

5. A GMM Analysis of the Consumption CAPM

Here we revisit Hansen-Singleton's analysis using the 1959-2015 U.S. monthly data on aggregate per-capita consumption of non-durable goods, the S&P 500 stock index, and the 1-year maturity U.S. treasury bonds. We have deflated all the returns using a non-durable consumption deflactor. We then constructed the series of $Y_t = (c_{t+1}, R_{1,t+1}, R_{2,t+1})'$ representing the total consumption return, the total return on the bonds, and the total return on the stock index. In this data the raw instruments Z_t could consist of the lags $Y_{t-1}, Y_{t-2}, ...$ In what follows we consider using only one lag for simplicity.

We then proceeded to estimate Hansen-Singleton's model of a representative consumer holding rational expectations and having the power utility, as described in Section 1. We used the iterated GMM and CUE estimator and considered the following two basic scenarios for the technical instrument set:

(1) the first lag only; i.e.

$$B(Z_t) = (1, c_t, R_{1,t}, R_{2,t}),'$$

(2) the first lag and all the squares and interactions in the first lag; i.e.

$$B(Z_t) = (1, c_t, R_{1,t}, R_{2,t}, c_t^2, R_{1,t}^2, R_{2,t}^2, c_t R_{1,t}, c_t R_{2,t}, R_{1,t} R_{2,t})'.$$

We also use the fact that the rational expectation assumption implies that scores are not correlated, yielding the variance simplification (1.3) and we use the corresponding variance estimator (1.4) in obtaining the optimal weights and computing the standard errors.[4]

[4]The results do not change qualitatively if we use Newey-West estimator that takes into account the potential correlation of the scores, which arises when we depart from the rational expectation assumption.

In both cases, the J-tests statistically rejects the model, which we can interpret as the model being unable to explain even very few key moments of the real data. Moreover, in this case GMM and CUE estimators yield very different estimates of the preference parameters, which is consistent with the statistical misspecification of the model. Note that under the misspecification the choice of the weighting matrix determines the pseudo-true values for which the GMM and CUE estimators are consistent and under misspecification GMM and CUE estimators converge to such different pseudo-true values and thus are no longer asymptotically equivalent.[5]

TABLE 1. Estimation Results for the Consumption CAPM

	GMM-1	CUE-1	GMM-2	CUE-2
estimated α	0.114	4.220	0.097	2.585
std. error	(0.037)	(0.548)	(0.033)	(0.277)
estimated β	0.998	1.003	0.998	1.001
std. error	(0.000)	(0.001)	(0.000)	(0.001)
J-statistic	213.008	52.882	245.100	74.802
p-value	0.000	0.000	0.000	0.000

6. GMM UNDER MISSPECIFICATION*

Misspecification arises as a consequence of failure of various modeling assumptions. By misspecification we mean here that there is no θ_0 such that $g(\theta_0) = 0$. However, we can define the pseudo-true of the parameter as the value that minimizes the distance between moment functions and zero

$$\theta_0 = \arg\min_{\theta\in\Theta} g(\theta)' A g(\theta).$$

In structural equation modeling, this is interpreted as trying to find a model that best describes properties of the real world encoded by moments. Under misspecification the choice of A affects the definition of the pseudo-true value and so the choice of A becomes very important and should be driven by economic reasoning, as opposed to the statistical reasoning. In application, we would want to chose A so that we give more weight to the moments we want to explain, and this is an application-specific matter. See, e.g., the paper by Jaganathan and Hansen,[6] where they study the problem of choosing the best economic model for stochastic discount factors, and use the inverse of variance covariance matrix of total returns as the weighting matrix A, and they gave an economic rationale for such weighting method.

[5] We did not study the GMM estimation under misspecification, but the basic points made here follow immediately from the Extremum Consistency Lemma.

[6] Hansen, Lars Peter, and Ravi Jagannathan. "Assessing specification errors in stochastic discount factor models." The Journal of Finance 52.2 (1997): 557-590.

Notes

Lars Hansen introduced the Generalized Method of Moments (GMM) and studied its properties in [8]. He first applied GMM to the Consumption CAPM model in [10], together with Kenneth Singleton. GMM can be seen as a generalization of the Method of Moments of Karl Pearson and the Method of Estimating Equations of Vidyadhar Prabakhar Godambe. The Continuously Updated GMM estimator was introduced in [9].

Appendix A. Tools

The appendix contains technical material. You only need to know the statements of the results and know how to apply them. If you plan to be an econometrician, it is a good idea to also learn the proofs.

A.1. **Tool 1: Extremum Consistency Lemma.** Consider the extremum estimator

$$\hat{\theta} \in \arg\min_{\theta \in \Theta} \widehat{Q}(\theta).$$

We assume that the estimator $\hat{\theta}$ exists throughout to simplify statements. This holds true, for example, if $\theta \mapsto \widehat{Q}(\theta)$ is continuous as a map from Θ to $\mathbb{R}$ almost surely, and Θ is compact.

Lemma 3 (Extremum Consistency). *We assume the $\hat{\theta}$ exists. Suppose (i) $\sup_{\theta \in \Theta} |\hat{Q}(\theta) - Q(\theta)| \to_P 0$, (ii) $Q(\theta) > Q(\theta_0)$ for all $\theta \neq \theta_0$, (iii) Θ is compact and $\theta \mapsto Q(\theta)$ is continuous. Then $\hat{\theta} \to_P \theta_0$.*

For intuition it is helpful to draw a picture that describes the theorem.

Proof. The proof has three steps: 1) we need to show $Q(\hat{\theta}) \to_P Q(\theta_0)$ using the assumed uniform convergence of $\hat{Q}$ to Q, and 2) we need to show that this implies that $\hat{\theta}$ must be close to θ_0 using continuity of Q, compactness of Θ, and the fact that θ_0 is the unique minimizer, 3) we then try to understand what happened in steps 1 and 2.

Step 1. By the uniform convergence,

$$\hat{Q}(\hat{\theta}) - Q(\hat{\theta}) \to_P 0 \text{ and } \hat{Q}(\theta_0) - Q(\theta_0) \to_P 0.$$

Also, by $\hat{Q}(\hat{\theta})$ and $Q(\theta_0)$ being minima,

$$Q(\theta_0) \leq Q(\hat{\theta}) \text{ and } \hat{Q}(\hat{\theta}) \leq \hat{Q}(\theta_0).$$

Therefore

$$\begin{aligned} Q(\theta_0) &\leq Q(\hat{\theta}) = \hat{Q}(\hat{\theta}) + [Q(\hat{\theta}) - \hat{Q}(\hat{\theta})] \leq \hat{Q}(\theta_0) + [Q(\hat{\theta}) - \hat{Q}(\hat{\theta})] \\ &= Q(\theta_0) + \underbrace{\hat{Q}(\theta_0) - Q(\theta_0) + Q(\hat{\theta}) - \hat{Q}(\hat{\theta})}_{o_P(1) \text{ by the uniform convergence}}, \end{aligned}$$

implying that $Q(\theta_0) \leq Q(\hat{\theta}) \leq Q(\theta_0) + o_P(1)$. It follows that $Q(\hat{\theta}) \to_P Q(\theta_0)$.

Step 2. By compactness of Θ and continuity of $Q(\theta)$, for any open subset $\mathcal{N}$ of Θ containing θ_0, we have that

$$\inf_{\theta \notin \mathcal{N}} Q(\theta) > Q(\theta_0).$$

Indeed, $\inf_{\theta \notin \mathcal{N}} Q(\theta) = Q(\theta^*)$ for some $\theta^* \in \Theta \setminus \mathcal{N}$. By identification, $Q(\theta^*) > Q(\theta_0)$.

But, by $Q(\hat{\theta}) \to_P Q(\theta_0)$, we have

$$Q(\hat{\theta}) < \inf_{\theta \notin \mathcal{N}} Q(\theta),$$

with probability approaching one for all large n, and hence $\hat{\theta} \in \mathcal{N}$ with probability approaching one for all large n. ■

A.2. **Tool 2: The Uniform Law of Large Numbers.** The following is a very useful uniform law of large numbers.

Lemma 4 (Uniform Law of Large Numbers). *Assume that $(X_i)_{i=1}^{\infty}$ is an i.i.d. or stationary strongly mixing sequence. Consider a map $(z, \theta) \mapsto m(z, \theta)$. Define $\hat{M}(\theta) := \mathbb{E}_n[m(X, \theta)]$. Suppose that (i) $\theta \mapsto m(X, \theta)$ is continuous at each $\theta \in \Theta$ with probability one; (ii) Θ is compact, and (iii) $\mathrm{E}[\sup_{\theta \in \Theta} |m(X, \theta)|] < \infty$. Then $\theta \mapsto M(\theta) := \mathrm{E}[m(X, \theta)]$ is continuous on Θ and*

$$\sup_{\theta \in \Theta} |\hat{M}(\theta) - M(\theta)| \to_P 0.$$

The proof applies more generally to any stationary ergodic process.

Proof. We prove that for any $\epsilon > 0$, $\sup_{\theta \in \Theta} \hat{M}(\theta) - M(\theta) \leq \epsilon$ with probability approaching one. A similar argument shows that we also have that with probability approaching one $\sup_{\theta \in \Theta} M(\theta) - \hat{M}(\theta) \leq \epsilon$, which implies the claim.

First, we show that for every $\varepsilon > 0$ there is a finite cover of Θ by balls $(U_{\theta(j)})_{j=1}^{p}$ such that

$$\max_{j \leq p} \mathrm{E} D_{U_{\theta(j)}}(X) \leq \varepsilon, \quad D_U(X) := \sup_{\bar{\theta} \in U} [m(X, \bar{\theta}) - \mathrm{E} m(X, \bar{\theta})].$$

Indeed for each $\theta \in \Theta$, let $U_{l,\theta} \searrow \{\theta\}$ be a decreasing sequence of open balls in Θ with diameter converging to zero. The continuity hypothesis and the dominated convergence theorem imply that $\theta \mapsto \mathrm{E}m(X,\theta)$ is continuous at any given θ, and therefore yield that $D_{U_{l,\theta}}(X) \searrow (m(X,\theta) - \mathrm{E}m(X,\theta))$. By the dominated convergence theorem, $\mathrm{E}D_{U_{l,\theta}}(X) \searrow 0$. Hence for each θ, we can find $l_0(\theta)$ such that for $l > l_0(\theta)$ we have $\mathrm{E}D_{U_{l,\theta}}(X) < \varepsilon$. By compactness of Θ the open cover $\{U_{l,\theta} : l > l_0(\theta), \theta \in \Theta\}$ has a finite subcover $(U_{\theta(j)})_{j=1}^p$ for which the claim holds.

Second, we have by the (ordinary) law of large numbers for i.i.d. or strongly mixing data that

$$\sup_{\theta\in\Theta} \hat{M}(\theta) - M(\theta) \leq \max_{j\leq p} \mathbb{E}_n D_{U_{\theta(j)}}(X_i) \to_P \max_{j\leq p} \mathrm{E}D_{U_{\theta(j)}}(X_i) \leq \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary the claim follows. ■

A.3. **Tool 3: Uniqueness of Solutions to System of Equations⋆.** This is more advanced material given for reference purposes. Here we take the opportunity to discuss the questions of uniqueness of solutions to system of nonlinear equations $g(\theta) = 0$, where $\theta \mapsto g(\theta)$ is a C^1 mapping from an open neighborhood of $\Theta \subset \mathbb{R}^d$ to $\mathbb{R}^d$. Denote the Jacobian map by $\theta \mapsto G(\theta)$.

Lemma 5 (Local Uniqueness)**.** *(1) Suppose that $g(\theta_0) = 0$ and $G = G(\theta_0)$ has full rank, then there exists an open neighborhood $\mathcal{N}$ of θ_0 such that $g(\theta) \neq 0$ for all $\theta \in \mathcal{N} \setminus \{\theta_0\}$. (2) Moreover, if Θ is compact and $G(\theta) = (\partial/\partial\theta')g(\theta)$ has full (column) rank for each θ, the number of solutions to $g(\theta) = 0$ is finite.*

Proof. (1) By the Implicit Function Theorem there exists an open neighborhood $\mathcal{N}$ of θ_0 such that g is injective between $\mathcal{N}$ and the image set $g(\mathcal{N}) := \{g(\theta) : \theta \in \mathcal{N}\}$.

(2) By the Implicit Function Theorem, we can cover Θ with a collection of open sets $\mathcal{N}_k$ such that g is injective between $\mathcal{N}_k$ and the image set $g(\mathcal{N}_k)$. By compactness there is a finite sub cover $\{\mathcal{N}_k\}$ over Θ. Each of the sets $\mathcal{N}_k$ can have at most one solution θ_k to $g(\theta) = 0$. ■

Lemma 6 (Global Uniqueness via Quasi-Positive Definiteness)**.** *Suppose that Θ is a convex bounded set of full dimension, $g(\theta_0) = 0$, and $G(\theta) + G(\theta)' > 0$ for all $\theta \in \Theta$. Then $g(\theta) = 0$ has a unique solution at θ_0.*

Proof. By Gale and Nikaido's Global Implicit Function Theorem [6], such g is injective between Θ and $g(\Theta)$. ■

This result is useful when g is a gradient of a convex function, as for example, in probit/logit estimation to be discussed later. In that case the Jacobian is symmetric $G(\theta) = G(\theta)'$ and positive definite under mild assumptions.

Lemma 7 (Global Uniqueness via Positive Principal Minors). *Suppose that Θ is a rectangular set of full dimension and $g(\theta_0) = 0$, and that determinant of $G(\theta)$ is positive and all other principal sub matrices of order less than d have nonnegative determinants, for all $\theta \in \Theta$. Then $g(\theta) = 0$ has a unique solution at θ_0.*

Proof. By Gale and Nikaido's Global Implicit Function Theorem [6], such g is injective between Θ and $g(\Theta)$. ∎

Lemma 8 (Global Uniqueness via Generalized Positive Principal Minors). *Suppose that Θ is a compact, convex polyhedron of full dimension and $g(\theta_0) = 0$. Suppose that for every $\theta \in \Theta$ and every linear space L spanned by a face of Θ containing θ , the determinant of the linear map from L to L formed by projecting the operator $G(\theta)$ on L has a positive sign. Then $g(\theta) = 0$ has a unique solution at θ_0.*

Proof. By Mas-Colell's Global Implicit Function Theorem [11], such g is injective between Θ and $g(\Theta)$. ∎

Lemma 9 (Global Identification via Generalized Positive-Quasi-Definiteness). *Suppose that Θ is a compact, convex set of full dimension with C^1 boundary $\partial\Theta$ and $g(\theta_0) = 0$. Suppose that for every $\theta \in \Theta$, the determinant of $G(\theta)$ is positive, and that for each $\theta \in \partial\Theta$, $v'(G(\theta) + G(\theta)')v \neq 0$ for all $v \in T_\theta : v \neq 0$, where T_θ is the tangent plane of $\partial\Theta$ at θ. Then $g(\theta) = 0$ has a unique solution at θ_0.*

Proof. By Mas-Colell's Global Implicit Function Theorem [11], such g is injective between Θ and $g(\Theta)$. ∎

The last two results require some mathematical sophistication on the part of the reader. For an application of the penultimate lemma to identification of structural quantile models, see [3, 4].

Appendix B. Problems

(1) Suppose we want to test the equality of projection parameter γ in the regression model:

$$Y = D\gamma + U, \quad U \perp D,$$

and of the structural parameter α in the structural equation model

$$Y = D\alpha + \epsilon, \quad \epsilon \perp Z.$$

Represent the joint problem of estimation of γ and α in the GMM framework by "stacking" together the two systems of moment equations that can be used to identify γ and α. Write down the resulting score function, and the properties of the GMM estimator of γ and α. State any additional assumptions you may need. From

this deduce the properties of the GMM estimator for the difference $\alpha - \gamma$ and describe its large sample properties. Construct a Wald test for the equality $\alpha = \gamma$ of the parameters. This resulting construction yields a version of the Hausman test.

(2) If you like art, try to give a graph-theoretic depiction of the structural equation model given by the system of equations (1.1). The article by Judea Pearl in *Statistics Surveys* [15] explains the art. In L2 and L3 we presented graph-theoretic depictions of the two structural equation models with instrumental variables. Graphs could be used to nicely decorate one's research article as well as reduce boredom.

(3) Provide the proof of Theorem 2. This follows from expanding $\sqrt{n}\hat{g}(\hat{\theta})$ as $\sqrt{n}\hat{g}(\theta_0) + \hat{G}(\bar{\theta})\sqrt{n}(\hat{\theta} - \theta)$ and then substituting in the first order expansion for

$$\sqrt{n}(\hat{\theta} - \theta) = -(G'\Omega^{-1}G)^{-1}G'\Omega^{-1}\sqrt{n}\hat{g}(\theta_0) + o_P(1),$$

which was obtained in the proof of Theorem 1. Finish the proof by the appeal to the continuous mapping theorem and using the properties of the normal distribution.

(4) Work with the Hansen-Singleton model. Describe how Euler equations lead to conditional moment restrictions and how conditional moment restrictions yield unconditional moment restrictions. Explain how you can set up the score functions for GMM estimation and properties of the resulting estimator. How does the rational expectation assumption affect the form of Ω? Provide primitive regularity conditions that imply condition G (that is, give (as primitive as possible) conditions on the consumption and price processes such that condition G holds). State the large sample properties of the GMM estimator.

(5) Adapt condition G to the misspecified case as in Section 6, prove consistency for the pseudo-true value and asymptotic normality the estimator. Explain the role of the extremum consistency lemma in the proof.

(6) Using the data provided GMM-based estimation and specification testing for the consumption CAPM model along the lines of Section 5, with the difference being that you should use 2 lags of financial returns as instruments. (Begin by replicating Section 5, but please don't report the replication results). The results reported there were produced in R using the gmm package. Provide detailed explanations for what you are doing; the Hansen and Singleton's article is a good example of how you could write things up.

(7) If you like challenges, provide GMM-based estimation and specification testing for the consumption CAPM model along the lines of Section 5, except for the power utility specification, try the Epstein-Zinn time non-separable utility specification.

This utility specification leads to a stochastic discount factor for the form:

$$\beta^{\lambda}(C_{t+1}/C_t)^{-\alpha\lambda}R_{0,t+1}^{\lambda-1},$$

where $R_{0,t+1}$ is the total return on the optimal portfolio (see, e.g. Stock and Wright article in *Econometrica* [16]). In this parameterization the risk aversion is $1-\lambda(1-\alpha)$ and the elasticity of the intertemporal substituion is $1/\lambda$.

References

[1] Theodore W Anderson and Herman Rubin. Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, pages 46–63, 1949.

[2] Fisher Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81:631–654, 1973.

[3] Victor Chernozhukov and Christian Hansen. An iv model of quantile treatment effects. *Econometrica*, 73(1):245–262, 2005.

[4] Victor Chernozhukov and Christian Hansen. Quantile models with endogeneity. *Annual Review of Economics*, 5:57–81, 2013.

[5] Larry G Epstein and Stanley E Zin. Substitution, risk aversion, and the temporal behavior of consumption and asset returns: A theoretical framework. *Econometrica: Journal of the Econometric Society*, pages 937–969, 1989.

[6] David Gale and Hukukane Nikaidô. The Jacobian matrix and global univalence of mappings. *Math. Ann.*, 159:81–93, 1965.

[7] M. I. Gordin. The central limit theorem for stationary processes. *Dokl. Akad. Nauk SSSR*, 188:739–741, 1969.

[8] Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.

[9] Lars Peter Hansen, John Heaton, and Amir Yaron. Finite sample properties of some alternative gmm estimators. *Journal of Business and Economic Statistics*, 14(3):262–280, 1996.

[10] Lars Peter Hansen and Kenneth J. Singleton. Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica*, 50(5):1269–1286, 1982.

[11] Andreu Mas-Colell. Homeomorphisms of compact, convex sets and the Jacobian matrix. *SIAM J. Math. Anal.*, 10(6):1105–1109, 1979.

[12] Whitney K. Newey. Efficient estimation of models with conditional moment restrictions. In *Econometrics*, volume 11 of *Handbook of Statist.*, pages 419–454. North-Holland, Amsterdam, 1993.

[13] Whitney K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79:147–168, 1997.

[14] Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987.

[15] Judea Pearl. Causal inference in statistics: an overview. *Stat. Surv.*, 3:96–146, 2009.

[16] James H. Stock and Jonathan H. Wright. Gmm with weak identification. *Econometrica*, 68:1055–1096, 2000.

[17] Jonathan H Wright. Detecting lack of identification in gmm. *Econometric Theory*, 19(02):322–330, 2003.