Quantitative Economics with Heterogeneity
A Guidebook

Dean Corbae **Fatih Guvenen**

Preliminary and Incomplete. Comments Welcome.

© by Fatih Guvenen. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Contents

I MODEL SPECIFICATION 3

1 Heterogeneity and Aggregation 7
1.1 Aggregation 7
1.2 Empirical Evidence on Perfect Risk Sharing 13

2 Technology 25
2.1 The CES Functional Form 25

3 Preferences 33
3.1 General Properties of Preferences 33
3.2 Benchmark Consumption Levels 40
3.3 Epstein-Zin Preferences 46
3.4 Preferences over Consumption and Leisure 53
3.5 Preferences with Home Production 55
3.6 Preferences for Households 56
3.7 Time Inconsistent Preferences 57
3.8 Literature Notes and Further Reading 59

4 Choice of Functional Forms 61
4.1 Balanced Growth 61

5 Three Key Parameters in Macroeconomics 67
5.1 Elasticities 67
5.2 Risk Aversion 67
5.3 Elasticity of Intertemporal Substitution[1] 78
5.4 Frisch Elasticity 80
5.5 Heterogeneity and Aggregate Parameters 82

6 Other Key Modeling Decisions 85
6.1 Decision Problem: Planning horizon and Decision Frequency 85
6.2 From Bachelor(ette)s to Families 89
6.3 Endogenizing Borrowing Constraints 89
6.4 Modeling Cross-Sectional Heterogeneity 90
6.5 Fixed Costs and S-s Models 92
6.6 Consumption Floor or Minimum Income 93

---

[1] This subsection borrows heavily from my paper, Guvenen (2006).

6.7 Stochastic Driving Forces 93
6.8 Closing the Model 98

II NUMERICAL TOOLS 99

7 Function Interpolation and Approximation 105
7.1 Basic Idea 106
7.2 Polynomial Interpolation 106
7.3 Spline Interpolation 108
7.4 Grid Spacing 117
7.5 Multidimensional Spline Interpolation 119
7.6 A Trick to Reduce the Curvature 119
7.7 Taking Stock 122
7.8 Exercises 124

8 Integration in Dynamic Programming 127
8.1 Main Ideas 128
8.2 Romberg Integration 129
8.3 Gaussian Quadrature 133
8.4 Automatic Integrators 136
8.5 Benchmarking Integration Methods 139
8.6 Which Integration Method to Use? 144
8.7 Taking Stock 146
8.8 Discretizing the Shock Space 146
8.9 Integration vs. Discretization: Which Method to Use? 149

9 Miscellaneous Tools 151
9.1 Root Finding 151
9.2 Numerical Differentiation 160
9.3 Taking Stock 161
9.4 Exercises 161

10 Local Optimization 165
10.1 One-Dimensional Optimization 166
10.2 Multi-Dimensional Optimization 167
10.3 Maximizing the Bellman Objective 170
10.4 Constrained Optimization 176
10.5 Exercises 177

III THE DYNAMIC TOOLKIT 179

11 Dynamic Programming: A Review of Theory 183
11.1 Monotone Mappings 184
11.2 Main Theoretical Results 188
11.3 Examples 193

11.4 The Euler Equation 195
11.5 Fancier Euler Equations 199
11.6 Dynamic Programming: An Alternative Formulation 203
11.7 Further Exercises (From Sergio Salgado) 204

12 Dynamic Programming: Numerical Methods 213
12.1 Value Function Iteration 213
12.2 Four Practical Issues 214
12.3 Accelerating VFI: Two Useful Tricks 222
12.4 Endogenous Grid Method 227
12.5 Euler Equation Method 235
12.6 Collocation Method 235
12.7 Taking Stock 236

13 Transition Operators: A Review of Theory and Numerical Implementation 237
13.1 An Illustrative Example 237
13.2 A Review of Theory 245
13.3 Numerical Implementation 248

14 Putting The Algorithms to Work 251
14.1 The Problem and Its Parameterization 251
14.2 Ensuring Accuracy 257
14.3 Checking Accuracy: An Application 261
14.4 Benchmarking VFI Algorithms 266
14.5 Checking Accuracy via Euler Equation Errors 271
14.6 Tips: Programming Practices 273

15 Challenging Problems 275
15.1 Non-Concave Value Functions 275
15.2 Consumption-Savings Problem with Epstein-Zin Utility 278
15.3 Firm's Optimal Investment Problem with Epstein-Zin Utility of Owners 279
15.4 EGM with Kinks 279
15.5 Flat objectives 279
15.6 Curse of Dimensionality 279
15.7 Timing of Events Within A Period 280
15.8 State Variable Choice 280
15.9 Timing Choice 281

IV GENERAL EQUILIBRIUM 285

16 Preliminaries 289
16.1 Recursive Competitive Equilibrium 289
16.2 Ergodicity and Stationarity 299
16.3 Reading 303

17 General Equilibrium without Aggregate Shocks **305**
17.1 The Aiyagari Model 306
17.2 Solution Methods for the Aiyagari Framework [30% Finished] 310
17.3 An Overlapping Generations Model 311
17.4 Transitions 316
17.5 Production vs. Endowment Economy Models 324

18 Power Law Framework **343**
18.1 Power Law Models of Inequality 347
18.2 A Full-Blown General Equilibrium Model with Power Laws 348
18.3 Further Reading 360
18.4 Exercises 361

19 General Equilibrium with Aggregate Shocks: The Krusell-Smith Method **363**
19.1 A Model of Heterogeneous Households 364
19.2 Krusell-Smith Algorithm 367
19.3 Details of Implementation 374
19.4 Clearing Two Asset Markets 378
19.5 Approximate Aggregation: Why the Mean is Sufficient 379
19.6 A Model of Heterogeneous Firms 381
19.7 Other Types of Models Solved with Krusell-Smith Method 387
19.8 Discussion 389
19.9 Bibliographic Notes 392
19.10 Exercises 392

20 General Equilibrium with Aggregate Shocks: Beyond Krusell-Smith **395**
20.1 Linearization-Based Approaches: Reiter's Method 395
20.2 Sequence Space Jacobian Methods 401
20.3 Models with A Small Number of Agents 401
20.4 Solving for the Recursive Competitive Equilibrium 403
20.5 Thick Tails 411
20.6 Taking Stock 411

V CALIBRATION and ESTIMATION **413**

21 On The Methodology of Research in Economics **417**
21.1 Example: A Consumption-Saving Model with Bayesian Learning . . . 418

22 Method of Moments Estimators **423**
22.1 Basic Idea 423
22.2 Generalized Method of Moments 424
22.3 Minimum Distance Estimation 431
22.4 Method of Simulated Moments (MSM) 436
22.5 Indirect Inference Estimation 449

22.6 Example: Lifecycle Model with Kalman Filtering 452
22.7 Smoothing the Objective 456
22.8 Economic Applications 458
22.9 Literature and Further Reading 458
22.10 Taking Stock 459

23 Basic Issues in Calibration 461
23.1 Calibration as Estimation: The Mechanics 462
23.2 External vs. Internal Consistency 463
23.3 Time Aggregation and Preference Parameters 463
23.4 What Data to Use? 464

24 Global Optimization 467
24.1 Multistart Algorithms 469
24.2 Quasi-Random Numbers (or Low-Discrepancy Sequences) 470
24.3 The TikTak Algorithm 474
24.4 Parallelizing the Algorithm 478
24.5 NLOPT 479
24.6 Benchmarking Optimization Algorithms 479
24.7 Further Reading 489
24.8 Practical Advice 489

25 Empirical Methods 491
25.1 Summary statistics can be misleading 491
25.2 Measurement Error in Survey Data 491
25.3 Some Useful Probability Theory 494
25.4 Conditioning and Endogeneity 494
25.5 Demand Analysis 497
25.6 Time, Cohort, and Age Effects 498

26 Measurement 505
26.1 Measuring Inequality 505
26.2 Higher-Order Moments: A Primer 507
26.3 Finite Mixture Models [Out of Place] 513
26.4 Risk Aversion with Higher-Order Moments 515

27 What Can We Learn From A Structural Model? 519
27.1 Welfare Analysis 519
27.2 Impulse Response Functions 524
27.3 Policy Experiments 524
27.4 Counterfactuals, Decompositions 524
27.5 The Methodology of Economics 525

28 Inequality: Substantive Questions 529
28.1 Consumption Inequality 529
28.2 Wealth Inequality 531
28.3 Wage and Earnings Inequality 534

29 Last But Not Least... 539
29.1 Topics Omitted Due to Space Constraints 539

VI Appendices 541

A Computer Routines: The Starter Package 543

B Mathematics Review 545
B.1 Sobol' Sequence 545
B.2 Kalman Filtering 548

C Derivations 549
C.1 549

D MSM Example 551

E US Micro Data Sources: A Primer 553
E.1 PSID: (Georgios) 553

Chapter 22

Method of Moments Estimators

The first four parts of this book dealt with the specification and numerical solution of economic models, including those that feature significant heterogeneity. Once we apply the tools we learned so far and obtain an accurate numerical solution, the next step in the research process is to assign empirically plausible values to model parameters before we can use it to study whatever economic question we set out to answer in the project. There are two approaches to accomplishing this assignment of parameters—calibration and estimation—which started out as conceptually distinct and very different strategies that have largely converged over time. The statistical foundation for those simulation-based estimation methods is built upon the Generalized Method of Moments (GMM). With this in mind, we start this chapter with a brief description of some essential points of GMM used in simulation based estimation methods, then turn our main focus upon the latter.]

22.1 Basic Idea

Your first undergraduate econometrics class started off with ordinary least squares (OLS). OLS is simply a specific case of a method of moments estimator. Specifically, the true data generation process is assumed to be the function $y_n = \beta x_n + u_n$ with

$$\mathbb{E}[x_n u_n] = 0, \mathbb{E}[u_n] = 0, \tag{22.1}$$

and demeaned data. The idea is to estimate the parameter vector β that maps the linear model βx_n to the data of interest y_n.

An implication of $\mathbb{E}[x_n u_n] = 0$ is that

$$\mathbb{E}[x_n (y_n - \beta x_n)] = 0. \tag{22.2}$$

That is, a function of the "errors" matching model to data must be zero on average.

The sample analogue of the moment condition (22.2) is

$$\frac{1}{N}\sum_{t=1}^{N} x_n \left(y_n - \widehat{\beta}_N^{MM} x_n\right) = 0 \tag{22.3}$$

yielding

$$\widehat{\beta}_N^{MM} = \frac{\sum_{n=1}^{N} x_n y_n}{\sum_{n=1}^{N} x_n x_n}. \tag{22.4}$$

An alternative way to obtain $\widehat{\beta}_N$ is to choose β that minimizes the sum of squared deviations of the data y_n from the model βx_n or

$$\widehat{\beta}_N^{OLS} = \arg\min_{\beta} \sum_{n=1}^{N} (y_n - \beta x_n)^2. \tag{22.5}$$

The first order condition is

$$-2\sum_{n=1}^{N} (y_n - \widehat{\beta}_N^{OLS} x_n) x_n = 0. \tag{22.6}$$

But this is identical to the moment condition in (22.3) so the tw22.5o methods yield the same "OLS" estimate.

Further, notice that Generalized Least Squares is simply a more general moment condition than (22.2) given by

$$\mathbb{E}[x_n \left(y_n - \beta x_n\right) / \sigma^2(x_n)] = 0. \tag{22.7}$$

That is, instead of weighting everything equally as in OLS (which is the best linear unbiased estimator if the explanatory variables have equal variance), it upweights moments inversely related to variation in the explanatory variables. Specifically, information from a given perturbation of variables with little variation is more informative than a given perturbation of variables with a lot of variation. You will see how this weighting is applied in the more general method of moments estimator to which we now turn.

To conclude, what's the basic idea? Minimizing the sum of squared errors of model from data as in (22.5) yields the same estimate of model parameters as in the moment condition (22.2). Both GMM and the Method of Simulated Moments (MSM) will take a similar approach of minimizing the weighted sum of squared errors of the model from the data.

22.2 Generalized Method of Moments

A standard (nontextbook) reference for Generalized Method of Moments (GMM) is Hansen (1982b). Let $n = 1, 2, ..., N$ index an individual unit (be it a household, a firm, time, some combination of those, etc.) and $k = 1, 2, ..., K$ index variables. Let θ be an $J \times 1$ vector of structural parameters that we are interested in estimating.

Let $f(X,\theta) : \mathbb{R}^{\mathrm{K}} \times \mathbb{R}^{\mathrm{J}} \to \mathbb{R}^{\mathrm{L}}$ be an $L \times 1$ vector of errors between model and data (e.g. $y_n - \beta x_n$) and assume that at the true parameter value θ_0 we have the moment condition

$$\mathbb{E}\left[f(X,\theta_0)\right] = 0. \tag{22.1}$$

Define

$$\mathbf{m}_N(X,\theta) := \frac{1}{N}\sum_{n=1}^{N} f(X_n,\theta) \tag{22.2}$$

to be the empirical counterpart of the population moment.

Definition 22.1. Let W_N be a $L \times L$ symmetric positive definite matrix, possibly dependent on the sample (i.e., stochastic) The GMM estimator of θ is

$$\widehat{\theta}_N(W_N) = \arg\min_{\theta}\left[\underset{1\times L}{\mathbf{m_N}(X,\theta)'}\underset{L\times L}{W_N}\underset{L\times 1}{\mathbf{m_N}(X,\theta)}\right]. \tag{22.3}$$

Note that (22.3) is the analogue of (22.5) and that $\widehat{\theta}_N$ is a random vector since it depends on X, which itself is random.

22.2.1 Large Sample Properties

The next set of results on consistency and efficiency of the GMM estimator draws on Theorems 2.1 and 3.2 in Hansen (1982b) which depend on a set of assumptions which crucially includes what is known as a **Global Identification** condition:[1]

$$\text{The } L \times 1 \text{ vector } g(\theta) \equiv \mathbb{E}\left[f(X,\theta)\right] = 0 \text{ only for } \theta = \theta_0. \tag{22.4}$$

Unfortunately, the global identification condition (22.4) is hard to verify. A simpler necessary but not sufficient condition is known as **Local Identification** or a rank condition. If $g(\theta)$ is continuously differentiable in a neighborhood of θ_0, then

$$\text{The L}\times\text{J Jacobian matrix } G \equiv \nabla_\theta g(\theta) \text{ must have full column rank.} \tag{22.5}$$

That is, the Jacobian matrix should have J linearly independent columns. Intuitively, this condition is not satisfied if a small change in a given parameter in one of the columns has zero impact in minimizing any of the L errors in the objective.

Theorem 22.2. *(Asymptotic distribution of the GMM estimator):*

A (Consistency) Under the above conditions, $\text{plim}_{N\to\infty}\widehat{\theta}_N = \theta_0$.

B (Asymptotic Normality) Under some additional higher order moment regularity conditions

$$\sqrt{N}(\widehat{\theta}_N(\mathbf{W_N}) - \theta_0) \to \mathcal{N}(0, \left(\mathbf{G'WG}\right)^{-1}\left(\mathbf{G'WSWG}\right)\left(\mathbf{G'WG}\right)^{-1}) \tag{22.6}$$

---

[1] The other conditions include: (i) plim$plimW_{N\to\infty} = W$ where W is a positive semi-definite matrix; (ii) $\theta_0 \in \Theta$ a compact subset of $\mathbb{R}^j$; (iii) $f(X,\theta)$ is continuous at each θ; and (iv) $[\mathbb{E}\sup_\theta \|f(X,\theta)\|] < \infty$.

where

$$\mathbf{S} := \mathbb{E}\left(m_\infty(X,\theta_0)\, m_\infty(X,\theta_0)'\right) \tag{22.7}$$

is the $L \times L$ asymptotic variance-covariance matrix.

$\boxed{\text{C}}$ (Optimal weighting matrix) A lower bound for the asymptotic variance of the GMM estimator indexed by $\mathbf{W}_N$ is given by $\left(\mathbf{G}'\mathbf{S}^{-1}\mathbf{G}\right)^{-1}$ and is achieved if $\mathbf{W}_N$ is such that

$$\mathbf{W} = \text{plim}_{N\to\infty}\mathbf{W}_N = \mathbf{S}^{-1}. \tag{22.8}$$

Several points are important to note. First, *in theory*, consistency is assured no matter what weighting matrix one chooses. However, as we will discuss in subsection 22.3.2, the choice of weighting matrix can be crucial to your estimation in small samples. Second, with regard to efficiency of the estimator, (22.8) implies that errors which come from high variance moments should be downweighted (to be discussed in more detail in subsection 22.3.2). Finally, there is a simple necessary "pre-test" related to the local identification or rank condition in (22.5). Specifically, the **Order Condition** for identification requires that $L \geq J$ (see Table E.1).

TABLE I – Order Conditions

Order	Condition
overidentified if	L>J
just identified if	L=J
underidentified if	L<J

22.2.2 How to Generate Moment Conditions?

As just discussed, the order condition $L \geq J$ is necessary for identification. We turn to examples here.

Example 1: Hansen and Singleton (1982)

Consider a consumer-investor who solves the following lifetime maximization problem over time t:

$$\max_{S^i_{t+1}} \mathbb{E}_0\left(\sum_{t=0}^{\infty}\beta^t U(C_t)\right)$$

$$\text{s.t} \quad C_t + \sum_{i=1}^{I} P^i_t S^i_{t+1} = \sum_{i=1}^{I} R^i_t S^i_t + W_t,$$

where R_t^i is the total return on asset i. For example, for a discount bond $R_t^i = 1$, and for a stock $R_t^i = P_t^i + D_t^i$. Let $x_t^i \equiv \frac{R_{t+1}^i}{P_t^i}$ and $U(C) = \frac{C^{1-\alpha}}{1-\alpha}$. The FOCs are

$$\mathbb{E}_t\left(\beta\left(\frac{C_{t+1}}{C_t}\right)^{-\alpha} x_t^i - 1\right) = 0 \qquad \text{for } i = 1, .., I \tag{22.9}$$

Define the $I \times 1$ vector of model implied errors from FOCs:

$$m(C_t, x_t; \alpha, \beta) \equiv \begin{bmatrix} \beta\left(\frac{C_{t+1}}{C_t}\right)^{-\alpha} x_t^1 - 1 \\ \beta\left(\frac{C_{t+1}}{C_t}\right)^{-\alpha} x_t^i - 1 \\ \vdots \\ \vdots \\ \beta\left(\frac{C_{t+1}}{C_t}\right)^{-\alpha} x_t^I - 1 \end{bmatrix}$$

As discussed in Section 22.2, equation (22.9) defines the moment conditions in (22.1).’

With $I \geq J = 2$, the order condition is satisfied and we can identify the true values α_0 and β_0 and obtain small sample estimates:

$$\left(\widehat{\alpha}_T, \widehat{\beta}_T\right) = \arg\min_{\alpha,\beta}\left[\left(\frac{1}{T}\sum_{t=0}^{T} m(C_t, x_t; \alpha, \beta)\right)' W \left(\frac{1}{T}\sum_{t=0}^{T} m(C_t, x_t; \alpha, \beta)\right)\right]$$

A major benefit of GMM (over Maximum Likelihood (ML) for example) is that we do not need to be specific about the budget constraint beyond the assets we use in the moment conditions (despite the way Hansen and Singleton (1982) present it in their paper). What if we have more parameters than assets (i.e. $L < J$)? This is where the main contribution of Hansen (1982b) and Hall (1978) lies (and why we have switched from n to t to make clear the importance of the time dimension to the fix).

Let us revisit the model and assume there is only a single risk-free asset in which case we have an underidentified model ($1 = L < J = 2$):

$$\mathbb{E}_t\left(\beta\left(\frac{C_{t+1}}{C_t}\right)^{-\alpha} \frac{1}{P_t} - 1\right) = 0.$$

A well-known property of conditional expectation is:

$$\mathbb{E}(f(X,\theta)|Z) = 0 \Rightarrow \mathbb{E}(f(X,\theta)h(Z)) = 0 \tag{22.10}$$

for every measurable function of Z. So, in principle, we can generate infinitely many moment conditions from just one. In practice, we can generate many by using functions of past (the reason why we switched notation to time t) publicly known information.

In principle, any function of pre-determined variables would work. For example:

$$m(C_t, x_t; \alpha, \beta) \equiv \begin{bmatrix} \left(\beta\left(\frac{C_{t+1}}{C_t}\right)^{-\alpha}\frac{1}{P_t}-1\right)C_{t-1} \\ \left(\beta\left(\frac{C_{t+1}}{C_t}\right)^{-\alpha}\frac{1}{P_t}-1\right)C_{t-2}^2 \\ \left(\beta\left(\frac{C_{t+1}}{C_t}\right)^{-\alpha}\frac{1}{P_t}-1\right)P_{t-2}^3\log C_{t-3} \\ \vdots \\ \left(\beta\left(\frac{C_{t+1}}{C_t}\right)^{-\alpha}\frac{1}{P_t}-1\right) \end{bmatrix}$$

Caution: In practice, *weak instruments* are a big problem: i.e., lagged variables that have very low correlation with the underlying moments (see, e.g., Nelson and Startz (1990)).

Example 2: Hall and Mishkin (1982)

Income process (permanent-plus-transitory model):

$$\begin{aligned} Y_t &= Y_t^P + \eta_t \\ Y_t^P &= Y_{t-1}^P + \epsilon_t \end{aligned} \tag{22.11}$$

Quadratic Preferences:

$$\max E_t\left[-\frac{1}{2}\sum_{\tau=0}^{T-t}(1+\delta)^{-\tau}\left(C^*-C_{t+\tau}\right)^2\right]$$
$$s.t$$
$$\sum_{\tau=0}^{T-t}(1+r)^{-\tau}\left(Y_{t+\tau}-C_{t+\tau}\right)+A_t=0$$

FOC: $E_t\left[(1+\delta)^{-\tau}\left(C^*-C_{t+\tau}\right)\right]=(1+r)^{-\tau}\left(C^*-C_t\right)$

Assume $\delta = r$, and we get the Euler equation (i.e., consumption is a martingale):

$$E_t\left[C_{t+\tau}\right] = C_t \tag{22.12}$$

Take the conditional expectation of budget constraint and substitute $E_t\left[C_{t+\tau}\right]=C_t$ (22.12):

$$\sum_{\tau=0}^{T-t}(1+r)^{-\tau}\left(E_tY_{t+\tau}-C_t\right)+C_t=0$$

Define: $H_t = E_t\left(\sum_{\tau=0}^{T-t}(1+r)^{-\tau}Y_{t+\tau}\right)$ and $\gamma_t = \frac{1}{\left(\sum_{\tau=0}^{T-t}(1+r)^{-\tau}\right)}$ and we get the consumption function:

$$C_t = \gamma_t \left(H_t + A_t \right) \tag{22.13}$$

- ➢ In most, empirical applications we do not have data on wealth, so A_t creates a problem.
- ➢ First-difference & use specification of income to get the Permanent Income Hypothesis (PIH) Model:
$$\Delta C_t = \epsilon_t + \gamma_t \eta_t \tag{22.14}$$
- ➢ **Caution:** This is not simply the Euler equation. It requires the derivation of the consumption function (which means you need to take a stand on budget constraint, the income process, etc.
- ➢ Quadratic utility implies certainty equivalence... so no precautionary savings in eq (22.14).
- ➢ How to use this in empirical work—we do not observe η_t and ϵ_t in the data?
- ➢ Plus: consumption is measured with error: $C_t = C_t^{**} + \nu_t$

Empirical Implementation

- ➢ Use covariances:

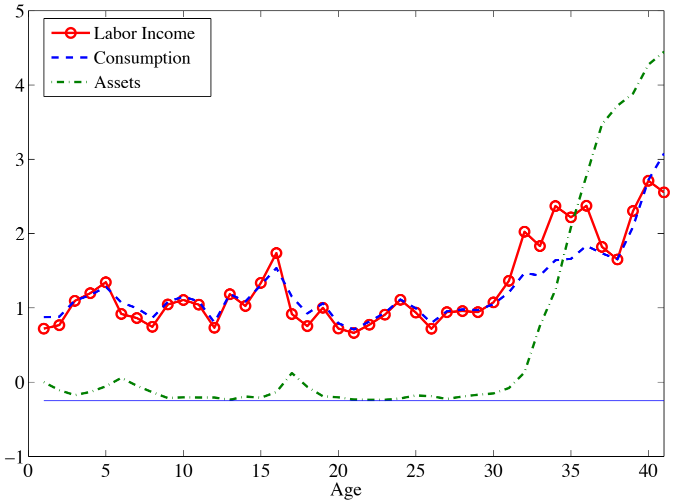
$$\begin{aligned} cov\left(\Delta Y_t, \Delta Y_{t-1}\right) &= -\sigma_\eta^2 \quad \leftarrow \text{reveals persistent shock variance} \\ C_0 &= cov\left(\Delta Y_t, \Delta C_t\right) = \sigma_\epsilon^2 + \beta \sigma_\eta^2 \\ C_1 &= cov\left(\Delta Y_{t+1}, \Delta C_t\right) = -\beta \sigma_\eta^2 \\ cov\left(\Delta C_t, \Delta C_{t-1}\right) &= -\sigma_\nu^2 \end{aligned}$$

- ➢ These moments are “dynamic” meaning they require panel data on the same individuals for 3 periods.
- ➢ Blundell and Preston (1998): used a similar set of moments that are “static” or need cross-sectional data only. For $T - t$ large and r small:

$$\begin{aligned} \Delta \text{var}_{kt}\left(C\right) &= \sigma_{\eta,t}^2 \\ \Delta \text{cov}_{kt}\left(Y_t, C_t\right) &= \sigma_{\eta,t}^2 \\ \Delta \text{var}_{kt}\left(Y_t\right) - \Delta \text{var}_{kt}\left(C_t\right) &= \sigma_{\varepsilon,t}^2 - \sigma_{\varepsilon,t-1}^2 \end{aligned}$$

Despite the appealing simplicity of GMM, the Euler equation derived above cannot identify many parameters of interest. And some other times we cannot even derive an Euler equation. For example, suppose we want to identify the properties of the income process by observing income and consumption data. The standard method since Hall and Mishkin (1982) is to derive structural equations explicitly and estimate

FIGURE 22.2.1 – Binding Constraints Biases Inference. True $\rho = 0.5$



them. This requires combining the Euler equation with the budget constraint and solving the model fully in closed form.

Start with the standard life-cycle version of the Permanent Income Certainty-Equivalent (PICE) model:

$$\max_{\{C_t, a_{t+1}\}_{t=0}^{T}} \left(-\frac{1}{2} \sum_{t=0}^{T} \beta^t (C_t - c^*)^2 \right)$$

$$\text{s.t} \quad C_t + a_{t+1} = (1 + r_t)a_t + y_t$$

$$y_t = \rho y_{t-1} + \eta_t$$

Notice that we assume: (i) quadratic utility, (ii) no borrowing limit, (iii) no valued leisure, and (iv) no retirement. You can solve this model in closed form and obtain several moment conditions that link consumption and income changes to the underlying shocks to income:

$$\text{cov}(\Delta y_t, \Delta c_t) = ...$$

$$\Delta C_t = \phi_t \eta_t,$$

where ϕ_t is the annuitization factor. [[TEXT TO BE ADDED]]

When $\rho = 1$ we have $\phi_t = 1$: $\Delta C_t = \eta_t = \Delta Y_t$. When $\rho = 0$ (iid shocks): $\Delta C_t = \phi_t \eta_t = \psi_t \Delta Y_t$, where $\phi_t > \psi_t \approx 0$. Thus, the response of consumption growth to income growth reveals persistence of income shocks.

22.2.3 Small Sample Properties

What makes a small sample? Oftentimes, we do not know. Sometimes, with a couple of hundred observations we get close to large sample properties. Sometimes 30,000 is too small. The *Journal of Business and Economic Statistics* published a special issue in July 1996 on the small sample properties of GMM. A lot of what we know (which is not much) comes from articles in that issue. Well-worth reading. Some important conclusions:

- Efficient GMM (i.e., one that uses the optimal weighting matrix) is very often inferior to using an identity weighting matrix. This is because the optimal weighting matrix relies on the fourth moment of the data, which is very difficult to estimate precisely (requires very large sample size) and thus can be severely biased in small samples.
- Wald test rejects too often.

22.2.4 Advantages and Limitations of GMM

The advantages are:

A No need to fully specify the model. As long as your model delivers sufficiently many moment conditions (L) to identify the parameters of interest (J) you are done.

B GMM encompasses many many estimation methods as special cases (ML, IV, OLS, GLS, etc etc). So if you understand GMM you will know a lot of econometrics.

The limitations are:

What if you cannot derive a moment condition from your model? Some examples include cases when frequently binding borrowing constraints prevent the Euler equations from holding as an equality, or when preferences are of the Epstein-Zin form and markets are incomplete, among others. This is where Method of Simulated Moments (MSM) will come into play.

22.3 Minimum Distance Estimation

Sometimes you don't have data X_n or have an extremely complicated set of equations describing equilibrium (e.g. discrete choice problems or problems with non-differentiabilities due to binding constraints) for which $f(X_n, \theta)$ does not take a simple form. An example of the former case is from Hopenhayn and Rogerson (1993) who build a firm dynamics model to understand the consequences of firing taxes on employment. They choose model parameters to match statistics (e.g. serial correlation in log employment and a histogram of the size distribution of firms aged 0-6 years) from the Longitudinal Research Data (LRD) file, which is not a publicly available dataset. To replicate their results, you don't need to have government special sworn status in order to create statistics using non-publicly available Census data; all you need to do is use their data statistics (say $\hat{\pi}$) in Table 1 to estimate the parameters of a general equilibrium model.

Consider a statistic of the data, $\hat{\pi}$, with a normal limit distribution: $\sqrt{N}(\hat{\pi} - \pi) \rightarrow \mathcal{N}(0, \Omega)$. Also given is a distance function $h(\cdot)$ which is continuously differentiable. Assume that there is a unique parameter vector θ_0 such that $\mathbb{E}[h(\pi, \theta_0)] = 0$ as in (22.1) where $h : \mathbb{R}^K \times \mathbb{R}^J \rightarrow \mathbb{R}^L$ and $L \geq J$. The minimum distance estimator is:

$$\hat{\theta} = \arg\min_{\theta} \left[h(\pi, \theta)' \times W \times h(\pi, \theta) \right]$$

In some problems, like the covariance matrix estimation below, GMM and minimum distance estimation (MDE) are equivalent.

22.3.1 Example: Income Dynamics

Specify a parametric income process for individual i over time, e.g.:

$$\begin{aligned} y_t^i &= \alpha^i + z_t^i + \varepsilon_t^i \\ z_t^i &= \rho z_{t-1}^i + \eta_t^i \end{aligned}$$

Derive the theoretical autocovariances of income implied by this specification:

$$\begin{aligned} \text{var}_i(y_t^i) &= \sigma_\alpha^2 + \text{var}_i(z_t^i) + \sigma_\varepsilon^2, \\ \text{var}_i(z_t^i) &= \sum_{s=1}^{t} \rho^{2t} \sigma_\eta^2, \\ \text{cov}(y_t^i, y_{t+j}^i) &= \sigma_\alpha^2 + \rho^j \text{var}_i(z_t^i). \end{aligned}$$

Stack into a covariance matrix, vectorize it, call it $C(\theta)$, where $\theta \equiv (\rho, \sigma_\alpha^2, \sigma_\eta^2, \sigma_\varepsilon^2)$. We then construct the empirical counterpart of the covariance matrix:

$$C(X) = vec \begin{bmatrix} \text{var}_i(y_1^i) & & & & \\ \vdots & \text{var}_i(y_2^i) & & & \\ \vdots & \cdots & \ddots & & \\ \text{cov}(y_1^i, y_t^i) & \cdots & \text{var}_i(y_t^i) & & \\ \vdots & & & \ddots & \\ \text{cov}(y_1^i, y_T^i) & & & & \text{var}_i(y_T^i) \end{bmatrix}$$

Finally, we choose θ to bring the theoretical covariance matrix as close to its empirical part as possible:

$$\hat{\theta} = \arg\min_{\theta} \left[(C(\theta) - C(X))' \times W \times (C(\theta) - C(X)) \right].$$

22.3.2 Weighting Matrix

For all the estimation techniques covered in this chapter—GMM, MDE, MSM, II—the most important decision you will have to make is the choice of the weighting matrix, W. Your empirical findings will often depend *critically* on your choice of W. So devote most of your energy to this choice. This is also crucial when you do calibration, which is essentially MSM.

Before turning to examples in small samples, we note two implications of the material from Section 22.2.1 for the choice of weighting matrix. First, the optimal asymptotic weighting matrix $\mathbf{W} = \mathbf{S}^{-1}$ downweights errors from high variance moments (i.e. those with low signal to noise). Further, the variance-covariance matrix of the estimates in (22.6) is given simply by $\left(\mathbf{G}'\mathbf{S^{-1}G}\right)^{-1}$. This implies that estimates drawn from high variance moments will have large standard errors, ceteris paribus. The other thing to note is that the precision of the estimates is related to $G \equiv \nabla_\theta g(\theta)$. If the objective is very sensitive to changes in the parameters (i.e. $\nabla_\theta g(\theta)$ is high), then there will be a low variance of the estimate (since $\nabla_\theta g(\theta)' S^{-1} \nabla_\theta g(\theta)$ is inverted). If the objective is not very sensitive to changes in the parameters (i.e. $\nabla_\theta g(\theta)$ is low), it will produce a high variance of the estimates. Simply put, this suggests that if you find big standard errors, it *may* be because the objective is not very sensitive to changes in the parameters so it is hard to find the true unique maximum. This is how local identification is linked to standard errors. This is related to the point on **Local Identification**; if $\nabla_\theta g(\theta) = 0$ for certain values of the parameter space, then the Jacobian matrix may not have full column rank and the parameters of the model are not well identified.

Second, *in theory*, the weighting matrix does not matter for the just identified case. To see this, consider the following $L = J = 2$ case:

$$\theta(W) = arg \min_{\theta=(\theta_1,\theta_2)} w_1 g_1(\theta_1,\theta_2)^2 + w_2 g_2(\theta_1,\theta_2)^2$$

The foc are:

$$\begin{aligned} \theta_1 &: \quad 2w_1 g_1(\theta)\nabla_{\theta_1} g_1 + 2w_2 g_2(\theta)\nabla_{\theta_1} g_2 = 0 \\ \theta_2 &: \quad 2w_1 g_1(\theta)\nabla_{\theta_2} g_1 + 2w_2 g_2(\theta)\nabla_{\theta_2} g_2 = 0 \end{aligned}$$

Since there are 2 equations in 2 unknowns, one would think that $\theta(W)$. However, in the just identified case θ is not a function of W. To see this, rewriting the 2 foc in matrix notation

$$[g_1(\theta) \quad g_2(\theta)] \begin{bmatrix} w_1\nabla_{\theta_1} g_1 & w_1\nabla_{\theta_2} g_1 \\ w_2\nabla_{\theta_1} g_2 & w_2\nabla_{\theta_2} g_2 \end{bmatrix} = [0 \ 0] \tag{22.1}$$

Then provided the 2×2 matrix is invertible, a necessary condition for identification, we have

$$[g_1(\theta) \quad g_2(\theta)] = [0 \ 0] \begin{bmatrix} w_1\nabla_{\theta_1} g_1 & w_1\nabla_{\theta_2} g_1 \\ w_2\nabla_{\theta_1} g_2 & w_2\nabla_{\theta_2} g_2 \end{bmatrix}^{-1} = [0 \ 0].$$

So what do we know about the choice of weighting matrix? The most thorough analysis of this question is in Altonji and Segal (1996), who conduct extensive Monte

Carlo simulations to compare the performance of alternative choices. This paper was written partly in response to earlier problems encountered by researchers in using the optimal weighting matrix in empirical applications (see, e.g., Abowd and Card (1989) and the 1987 working paper version of Altonji et al. (2002)). Notice that although this is a very useful paper it is specific to the context of covariance matrix estimation, a limitation that should be kept in mind when applying their findings to other contexts.

Altonji and Segal (1996) reach a few key conclusions. They consider a linear model relating sample variances and covariances to a single population parameter. Specifically, consider empirical observations D_{k_i}, where $k_i = 1, 2, ..., K$ indexes the variable for which statistics are going to be computed; and $i = 1, 2, ..., N_k$ indicates the number of observations for that variable. The sample mean and variance for variable k is computed as:

$$\overline{D}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} D_{k_i},$$

and

$$m_k = \frac{1}{(N_k - 1)} \sum_{i=1}^{N_k} \left(D_{k_i} - \overline{D}_k\right)^2, \qquad \mathbb{E}(m_k) = \mu_k.$$

Let m be a $K \times 1$ vector of stacked second order sample moments. We specify the relationship between sample and population moments as:

$$m = \mu + \varepsilon = f(\theta) + \varepsilon.$$

Altonji and Segal consider a linear model for f: $m = X\theta + \varepsilon$. One application of this simple framework is the estimation of the sample variance when a researcher has observations on m_p in 10 time periods, with varying number of observations in each period. In this case, X is a 10×1 vector of ones and θ is the population variance. The minimum distance estimator is:

$$\theta_{MDE} = \arg\min_{\theta} \left[(m - f(\theta))' \times W \times (m - f(\theta))\right].$$

One choice for W is the identity matrix, in which case the estimation reduces to a least squares regression of m on X (a vector of ones). This choice is not efficient if the elements of ε are heteroskedastic or correlated. Heteroskedasticity will arise, for example, when each moment is estimated with a different sample size, N_k. When we use the optimal weighting matrix, let us refer to the resulting estimator as "optimal minimum distance" or OMD estimator. In this context, it will correspond to the generalized least squares estimator. In particular, we would use the inverse of the covariance matrix of ε, which we denote Ω. The usual estimator for each diagonal element of this matrix is given by:

$$\omega_k \equiv \hat{\text{var}}(m_k) = \frac{N_k^2}{(N_k - 1)(N_k - 2)^2} \left[\frac{1}{N_k} \sum_{i=1}^{N_k} \left(D_{k_i} - \overline{D}_k\right)^4 - \left(\frac{1}{N_k} \sum_{i=1}^{N_k} \left(D_{k_i} - \overline{D}_k\right)^2\right)^2\right]. \tag{22.2}$$

There are similar expressions for the covariances between m_k and $m_{k'}$. Stacking these terms into the covariance matrix gives us the empirical estimate of $\hat{\Omega}$, whose inverse is the optimal weighting matrix.

Altonji and Segal show that the OMD is seriously biased for the example above when the underlying distribution generating the data has thick tails (lognormal, student-t distribution, etc.) For sufficiently small samples the bias can be as large as 60% of the true value. In particular, the bias is given by:

$$\mathbb{E}(\theta_{OMD} - \theta) = \mathbb{E}\left[\left(\sum_{k=1}^{K} \omega_k^{-1}\right) \sum_{k=1}^{K} \omega_k^{-1} \varepsilon_k\right].$$

Now if the estimated variance terms, ω_k, were uncorrelated with the errors, ε_k, the errors would average out and the bias would be zero. They show that this is not true. For example, if the distribution has positive skewness the two terms (ω_k and ε_k) are positively correlated. The problem is that the weights are constructed with the same data and thus also involve the same ε_k as the one that goes into the moments. In general, they show that the bias is negative, so the OMD estimates are downward biased (although this result has exceptions, they also construct an example where the bias goes in the other direction). Moreover, with thicker tails, the distribution has high kurtosis, which in turn means that the second moments are imprecisely estimated, as seen from the expression in (22.2). Therefore, the OMD estimator is also likely to have higher standard error (less precision) relative to EWMD estimator.

Second, Altonji-Segal examine a real life scenario—the estimation conducted by Abowd and Card (1989)—and show that their estimates of the variance from the data are downward biased by about 70% to 80% compared to the true value!

They propose an alternative way to construct a weighting matrix, by splitting the sample. In this case, half of the data set is used to construct an estimate of $\hat{\Omega}$ and the other half is used to construct the moments. They show that this approach virtually eliminates the bias of the OMD estimator. This is a clever idea whose only downside is perhaps that it results in a smaller sample for estimation. This can be seen in their Table 1, where the IWOMD has often twice the RMSE than the EWMD. The identity matrix works as well as this alternative method and does not result in halving the sample size. (See also Geweke et al. (1997) and McFadden and Ruud (1994). The latter paper emphasizes the role of creating multiple runs of the simulation so as to improve the stability of the numerical estimation procedure. See their discussion of Table 3.)

Finally, note that prespecified rather than “efficient" weighting matrices can emphasize economically interesting results, they can avoid the trap of blowing up standard errors rather than improving model errors, they can lead to estimates that are more robust to small model misspecifications. This is analogous to the fact that OLS can be preferable to GLS in some contexts.

22.4 Method of Simulated Moments (MSM)

In the previous subsection, we covered traditional estimation techniques, which work by maximizing an objective function (or minimizing errors) of real data and a vector of parameters that we are interested in. GMM and MDE were two examples of this approach. In this subsection we cover methods that took these methods one important step further by introducing "simulations" into the estimation. One of the first application of this technique appears in Pakes (1986) and its theory has been introduced in McFadden (1989) and Pakes and Pollard (1989).

As discussed in earlier subsections, GMM is an extremely useful estimation techniques in situations when we have a known function $f(X, \theta)$ of the data and a vector of parameters of interest such that condition (22.1) $\mathbb{E}\left[f(X, \theta_0)\right] = 0$ holds. But suppose that the true function f is not known or is extremely difficult to evaluate. In that case, estimating parameter values in (22.3) $\widehat{\theta}_N(W_N) = \arg\min_\theta \left[\underset{1\times L}{\mathbf{m_N}(X, \theta)'}\underset{L\times L}{W_N}\underset{L\times 1}{\mathbf{m_N}(X, \theta)}\right]$ using (22.2) $\mathbf{m}_N(X, \theta) := \frac{1}{N}\sum_{n=1}^{N} f(X_n, \theta)$ may not be well-behaved. Instead we will construct $\mathbf{m}_N(X, \theta)$ using simulation methods.

To this end, let $\{X_n\}_{n=1}^N$ be a realization of a $K \times 1$ vector valued stationary and ergodic stochastic process generating the observed data (e.g. detrended GDP) and let $M_N(X)$ be an $L \times 1$ vector of associated data moments (e.g. standard deviation of detrended GDP). Now, let $\{Y_n(\theta)\}_{n=1}^N$ be a realization of an $K \times 1$ vector valued stationary stochastic and ergodic process generating the simulated data (e.g. GDP generated by the model) where θ is a $J \times 1$ vector of parameters. Ergodicity ensures that we may take H simulations of length N and construct the $L \times 1$ model moment analogues $M_{HN}(Y(\theta))$ of simulated data.[2] Finally assume that $M_N(X) \overset{a.s.}{\to} \mu(X)$ as $N \to \infty$ and that $M_{HN}(Y(\theta)) \overset{a.s.}{\to} \mu(Y(\theta))$ as $HN \to \infty$ where $\mu(x)$ and $\mu(y(b))$ are the population moments.

The fundamental assumption in the Method of Simulated Moments (MSM) is that there is a relation between data and theory. Specifically, under the null that the model is correct at the true parameter vector θ_0, then $\mu(X) = \mu(Y(\theta_0))$. Given a symmetric $L \times L$ weighting matrix W_N (which could depend on data - hence the subscript N), Lee and Ingram (1991b) show that under certain conditions the simulation estimator $\widehat{\theta}_{HN}$ which minimizes a "minimum distance" objective function (i.e. the weighted sum of squared errors of the model moments from the data moments as in Chamberlain (1984)) - ie. the solution to

$$\widehat{\theta}_{HN} = \arg\min_\theta [M_N(X) - M_{HN}(Y(\theta))]' W_N [M_N(X) - M_{HN}(Y(\theta))] \qquad (22.1)$$

- is a consistent and asymptotically normal estimator of θ_0 (i.e. $\lim_{HN\to\infty} \Pr ob(|\widehat{\theta}_{HN} - \theta_0| < \varepsilon) = 1$).

One can think of MSM as just GMM where the errors are just the difference between the data moments and the model moments $\mathbf{m_{HN}}(X, \theta) = M_N(X) - M_{HN}(Y(\theta))$, i.e.

---

[2] The ergodic theorem allows the time average of an ergodic process to be equal to the ensemble average. This means that statistical sampling can be done at one time across a group of identical processes, or over time on a single process, without changing the measured result.

the difference between the data moment and the model moment. Since the solution to this problem is essentially a special case of the GMM estimator in Hansen (1982a), the conditions are from his paper: (i) X and $Y(\theta)$ are independent; (ii) the model must be identified; and (iii) $M_{HN}(Y(\theta))$ must be continuous in the mean. In that case, the same results about consistency and asymptotic normality hold as in (22.2.1) as $N \to \infty$.

To take an example, in the stochastic neoclassical growth model, the parameters could be the risk aversion parameter, the time discount factor, the persistence and innovation variance of total factor productivity shocks. And the moments could be the volatility of output, consumption, investment, the autocorrelation and cross-correlation patterns and so on. The moments are most easily obtained via simulation, which makes MSM convenient. Remember to make sure when picking moments that at least the order condition $J \leq L$ is satisfied.

22.4.1 Basic Algorithm

You can think of the estimation as being conducted in a sequence of two steps (or calls of functions):

$\boxed{\textbf{A}}$ For any given value of θ, say θ^i at iteration i,

  (a) simulate artificial data from the model. For example, in the context of an AR1 model, draw $\{\varepsilon_t\}_{t=1}^{T=HN}$ shocks from a Normal distribution and construct $Y_t = \theta^i Y_{t-1} + \varepsilon_t$

  (b) compute a moment (i.e. $M_T(Y(\theta^i))$), and evaluate the objective function $\mathcal{F}_{HN}(\theta^i) = [M_N(X) - M_T(Y(\theta^i))]' W_N [M_N(X) - M_T(Y(\theta^i))]$; and

$\boxed{\textbf{B}}$ choose a new value for the parameters, say θ^{i+1}, for which $\mathcal{F}_{HN}(\theta^{i+1}) \leq \mathcal{F}_{HN}(\theta^i)$.

A minimization routine such as those discussed in Chapter 10 constructs this sequence of smaller and smaller $\mathcal{F}_{HN}(\theta^i)$ for you. **Importantly, you must use the same random draw throughout each simulation** of the artificial data or else you wouldn't know whether the change in the objective function were coming from a change in the parameter or a change in the draw.

To obtain the optimal weighting matrix even if you don't have the data to construct S_N, you can in principle use a two stage procedure.[3] In the first stage (i.e. $s = 1$), minimize $\mathcal{F}_{HN}^{s=1}(\theta)$ constructed using $W = I$. Since the resulting estimate $\widehat{\theta}_{HN}^{s=1}$ of θ_0 is consistent, generate H repetitions of model moments (from N length simulated samples) analogous to the data moments in order to construct an estimate of the variance-covariance matrix $\widehat{S}_{HN}$ of "data moments". Then use $W_{HN} = \widehat{S}_{HN}^{-1}$ to construct the second stage $\mathcal{F}_{HN}^{s=2}(\theta)$ and obtain the corrected estimate $\widehat{\theta}_{HN}^{s=2}$. Once you have the optimal weighting matrix, you can generate standard errors as in (22.6) of in Hansen (1982a).

Alternatively, you can run a Monte Carlo experiment to compute standard errors of the estimates. Recall that each estimate of $\widehat{\theta}_{HN}$ is derived for a given draw of the shocks

---

[3] See Section 4.2.3 on Asymptotic Properties of Indirect Inference Estimators in Gourieroux and Monfort (2002)for justification of this process.

ε_t to the underlying data generation process. Different draws will generate different estimates of $\widehat{\theta}_{HN}$. You can generate a histogram and summary statistics (mean and standard deviation of $\widehat{\theta}_{HN}$ which are interpretable as the point estimate and standard error of θ.

22.4.2 An Example

This example is intended to illustrate the large and small sample properties of the MSM. The example is designed to see how certain moments are better able to identify the parameter we are trying to estimate for an overidentified model. It expands on Section 3 of Michaelides and Ng (2000) who undertake Monte Carlo exercises for three simulation estimators.[4] We take the true data generation process to be a first order moving average (MA(1)) process

$$X_t = \varepsilon_t - \theta_0 \varepsilon_{t-1}, \varepsilon_t \overset{i.i.d}{\sim} N(0,1) \tag{22.2}$$

with $J = 1$ parameter $\theta_0 = 0.5$ and $\varepsilon_0 = 0$.

We will take the model generation process to be

$$Y_t(\theta) = e_t - \theta e_{t-1}, \quad e_t \overset{i.i.d}{\sim} N(0,1) \tag{22.3}$$

with $J = 1$ parameter θ and $e_0 = 0$. We do not know the true parameter value θ_0 so will estimate it via the method of simulated moments.

Let M_T denote the mapping from some $K \times 1$ vector z_t (which could be true data X or simulated data $Y(\theta)$) to an $L \times 1$ moment vector. Here we take $K = 1$ and $J = 1$ and consider an overidentified model with $L = 4$ moments: mean, variance, first order autocorrelation, and second order autocorrelation given by:

$$M_T(z_t) = \begin{bmatrix} z_t \\ (z_t - \bar{z})^2 \\ (z_t - \bar{z})(z_{t-1} - \bar{z}) \\ (z_t - \bar{z})(z_{t-2} - \bar{z}) \end{bmatrix}. \tag{22.4}$$

22.4.2.1 Asymptotics

Note that we can write the population (unconditional) moment vector for the true data and the model using $M_\infty(z)$ as $\mu(X) = E[M(X)]$ and $\mu(Y(\theta)) = E[M(Y(\theta))]$ where we drop the subscript ∞ for notational brevity in this subsection.

---

[4] In a computational exercise to this chapter, we will consider an AR1 case.

For this particular $L = 4$ mapping we know the population data moments

$$\mu(X) = \begin{bmatrix} E[\varepsilon_t] - \theta_0 E[\varepsilon_{t-1}] \\ E\left[(\varepsilon_t - \theta_0\varepsilon_{t-1})^2\right] = E[\varepsilon_t^2] - 2\theta_0 E[\varepsilon_t\varepsilon_{t-1}] + \theta_0^2 E[\varepsilon_{t-1}^2] \\ E[(\varepsilon_t - \theta_0\varepsilon_{t-1})(\varepsilon_{t-1} - \theta_0\varepsilon_{t-2})] = E[\varepsilon_t\varepsilon_{t-1}] - \theta_0 E[\varepsilon_t\varepsilon_{t-2}] - \theta_0 E[\varepsilon_{t-1}^2] + \theta_0^2 E[\varepsilon_{t-1}\varepsilon_{t-2}] \\ E[(\varepsilon_t - \theta_0\varepsilon_{t-1})(\varepsilon_{t-2} - \theta_0\varepsilon_{t-3})] = E[\varepsilon_t\varepsilon_{t-2}] - \theta_0 E[\varepsilon_t\varepsilon_{t-3}] - \theta_0 E[\varepsilon_{t-1}\varepsilon_{t-2}] + \theta_0^2 E[\varepsilon_{t-1}\varepsilon_{t-3}] \end{bmatrix}$$

$$= \begin{bmatrix} 0 \\ 1 + \theta_0^2 \\ -\theta_0 \\ 0 \end{bmatrix} \tag{22.5}$$

and the associated population model moments

$$\mu(Y(\theta)) = \begin{bmatrix} 0 \\ 1 + \theta^2 \\ -\theta \\ 0 \end{bmatrix}. \tag{22.6}$$

The L moment conditions we are going to use in MSM is given by

$$m(X, \theta) = M(X) - M(Y(\theta)). \tag{22.7}$$

Then define an $L \times 1$ vector (22.4) as in Section 22.2 above

$$\begin{aligned} g(\theta) &\equiv E[m(X,\theta)] = E[M(X)] - E[M(Y(\theta))] \\ &= \mu(X) - \mu(Y(\theta)) \end{aligned}$$

Note that **Global Identification** requires $g(\theta) = 0 \iff \theta = \theta_0$. In this example, the Global Identification condition is that there is a unique solution $\theta = \theta_0$ to the following equation

$$\begin{aligned} & g(\theta) = \mu(X) - \mu(Y(\theta)) = 0 \\ \iff & \mu(X) = \begin{bmatrix} 0 \\ 1 + \theta_0^2 \\ -\theta_0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 + \theta^2 \\ -\theta \\ 0 \end{bmatrix} = \mu(Y(\theta)). \end{aligned} \tag{22.8}$$

In particular, in (22.8) the mean and second autorcorrelation do not identify θ, the variance does not uniquely identify it since $\theta = \pm\theta_0$, while the autocorrelation uniquely identifies $\theta = \theta_0$.

In general we never actually have analytical expressions for $\mu(X)$ and $\mu(Y(\theta))$ so cannot obtain an estimate as above. That's why we will use MSM to estimate θ in a finite sample. Before moving to the finite sample case, it is instructive to state the asymptotic version of MSM since we can obtain an analytic expression for the variance covariance matrix S and hence the optimal weighting matrix $W^* = S^{-1}$. Let S denote the $L \times L$ asymptotic variance-covariance matrix of the L moment conditions $m(X, \theta)$

at the true parameter value $\theta = \theta_0$:

$$S = \sum_{j=-\infty}^{\infty} E[m(X_t, \theta_0)m(X_{t-j}, \theta_0)'] \tag{22.9}$$

If we make the assumption that the second moment of the true data and model are equal at the parameter $\theta = \theta_0$, then X_t and $Y_t(\theta_0)$ have the same population means and the same variance-covariance matrix in which case we can use either actual data or simulated data. In that case,

$$S_x \equiv \sum_{j=-\infty}^{\infty} E[\{M(X_t)-\mu(X)\}\{M(X_{t-j})-\mu(X)\}'] = \sum_{j=-\infty}^{\infty} E[\{M(Y_t(\theta_0))-\mu(Y(\theta_0))\}\{M(Y_{t-j}(\theta_0))-\mu(Y$$

where S_x (S_y) is the asymptotic variance-covariance matrix of $M(X)$ (or $M(Y(\theta_0))$). Note that data draws in X and simulation draws in $Y(\theta_0)$ are uncorrelated.

Let $\Gamma_j \equiv E[\{M(X_t) - \mu(X)\}\{M(X_{t-j}) - \mu(X)\}']$ denote the j-th autocovariance of $M(X_t)$. Then the asymptotic variance-covariance matrix of $M(X_t)$ is given by

$$S_x = \Gamma_0 + \sum_{j=1}^{\infty}(\Gamma_j + \Gamma_j') \tag{22.10}$$

Because we know the true DGP, we can compute Γ_js analytically:

$$\Gamma_0 = \begin{bmatrix} \sigma_x^2 & 0 & 0 & 0 \\ 0 & 2\sigma_x^4 & -2\theta_0\sigma_x^2 & 0 \\ 0 & -2\theta_0\sigma_x^2 & \sigma_x^4 + \theta_0^2 & -\theta_0\sigma_x^2 \\ 0 & 0 & -\theta_0\sigma_x^2 & \sigma_x^4 \end{bmatrix}$$

$$\Gamma_1 = \begin{bmatrix} -\theta_0 & 0 & 0 & 0 \\ 0 & 2\theta_0^2 & 0 & 0 \\ 0 & -2\theta_0\sigma_x^2 & \theta_0^2 & 0 \\ 0 & 2\theta_0^2 & -\theta_0\sigma_x^2 & \theta_0^2 \end{bmatrix}$$

$$\Gamma_j = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \forall j \geq 2.$$

where $\sigma_x^2 = 1 + \theta_0^2$ is the variance of X_t. So the asymptotic variance-covariance matrix is:

$$\begin{aligned} S_x &= \Gamma_0 + \Gamma_1 + \Gamma_1' \\ &= \begin{bmatrix} (1-\theta_0)^2 & 0 & 0 & 0 \\ 0 & 2(1 + 4\theta_0^2 + \theta_0^4) & -4\theta_0(1+\theta_0^2) & 2\theta_0^2 \\ 0 & -4\theta_0(1+\theta_0^2) & 1 + 5\theta_0^2 + \theta_0^4 & -2\theta_0(1+\theta_0^2) \\ 0 & 2\theta_0^2 & -2\theta_0(1+\theta_0^2) & 1 + 4\theta_0^2 + \theta_0^4 \end{bmatrix} \end{aligned}$$

At $\theta_0 = 0.5$,

$$S_x = \begin{bmatrix} 0.25 & 0 & 0 & 0 \\ 0 & 4.125 & -2.5 & 0.5 \\ 0 & -2.5 & 2.3125 & -1.25 \\ 0 & 0.5 & -1.25 & 2.0625 \end{bmatrix}$$

The inverse of S is the optimal weighting matrix:

$$W^* = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 1.1115 & 1.5705 & 0.6823 \\ 0 & 1.5705 & 2.8621 & 1.3539 \\ 0 & 0.6823 & 1.3539 & 1.1400 \end{bmatrix} \tag{22.11}$$

In order to compute standard errors, we need the derivative of $g(\theta)$ (an $L \times J$ matrix): [5]

$$\begin{aligned} \nabla_\theta g(\theta_0) &= -\nabla_\theta \mu(Y(\theta_0)) \\ &= -\begin{bmatrix} 0 \\ 2\theta_0 \\ -1 \\ 0 \end{bmatrix} \end{aligned} \tag{22.13}$$

This derivative is useful to see if the parameter is **Locally Identified**. To see this, take the first order approximation of $g(\theta)$ around θ_0:

$$\begin{aligned} g(\theta) &\approx g(\theta_0) + \nabla_\theta g(\theta_0)(\theta - \theta_0) \\ &= \nabla_\theta g(\theta_0)(\theta - \theta_0) \end{aligned}$$

since $g(\theta_0) = 0$. For $\theta = \theta_0$ to be the unique solution to $\nabla_\theta g(\theta_0)(\theta - \theta_0) = 0$, it must be true that

$$\text{rank}(\nabla_\theta g(\theta_0)) = J$$

From (22.13), we can see that $\text{rank}(\nabla_\theta g(\theta_0)) = 1 = J$, so the parameter is locally identified. In contrast, if $\nabla_\theta g(\theta_0)$ in (22.13) was the zero vector (which has rank $0 < J$), then the objective would not respond to changes in the parameter.

Finally, once we know these analytical results, we can write down the population analogue of the SMM objective function as

$$F(\theta) = g(\theta)' W^* g(\theta)$$

---

[5] Recall, from Theorem 3.2 of Hansen:

$$\sqrt{T}(b_T - b_0) \to N(0, \left[\nabla_b g(b_0)' S^{-1} \nabla_b g(b_0)\right]^{-1}) \tag{22.12}$$

The first order condition is

$$\begin{aligned}
&\nabla_\theta \left(g(\theta)' W^* g(\theta)\right) = 0 \\
&\iff -\begin{bmatrix} 0 & 2\theta & -1 & 0 \end{bmatrix} \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 1.1115 & 1.5705 & 0.6823 \\ 0 & 1.5705 & 2.8621 & 1.3539 \\ 0 & 0.6823 & 1.3539 & 1.1400 \end{bmatrix} g(\theta) = 0
\end{aligned}$$

where the second line results after some algebra.[6] If we evaluate $\nabla_\theta g(\theta)$ at $\theta = \theta_0$ and compute $\nabla_\theta g(\theta_0)' W^*$ then we obtain

$$\begin{bmatrix} 0 & 0.4590 & 1.2916 & 0.6715 \end{bmatrix} g(\theta) = 0$$

This means that the weight on the mean is zero. This is because the mean is not useful at all for the estimation of θ. On the other hand, the second order autocovariance gets a positive weight, even though it is not useful itself. This is because even though the second order autocovariance doesn't depend on θ, it is correlated with the variance and first order autocovariance, which is useful for the estimation of θ. If we want to make the estimator efficient, we should take the information in the second autocovariance into account.

22.4.3 Small Samples

In general, we only have a finite sample of size N data, so we must construct the $L \times 1$ vector of data moments $M_N(X)$. Given the finite sample, in general $M_N(X) \neq \mu(X)$, but $M_N(X) \overset{a.s.}{\to} \mu(X)$ as $N \to \infty$.

22.4.3.1 Sample moments for the true data

We first generate a series of random sample $\{\varepsilon_t\}_{t=1}^N$ from $N(0,1)$ and then construct a series of $\{X_t\}_{t=1}^N$ using the true DGP in (22.2) or

$$X_t = \varepsilon_t - \theta_0 \varepsilon_{t-1}$$

with $\varepsilon_0 = 0, \theta_0 = 0.5$, and $N = 200$ The 'true' data is plotted in Figure 1.

Using this data, we can compute the $L \times 1$ data moment vector by

$$\hat{M}_N(X) = \frac{1}{N} \sum_{t=1}^{N} M_N(X_t).$$

---

[6] To see this, go to the Appendix.

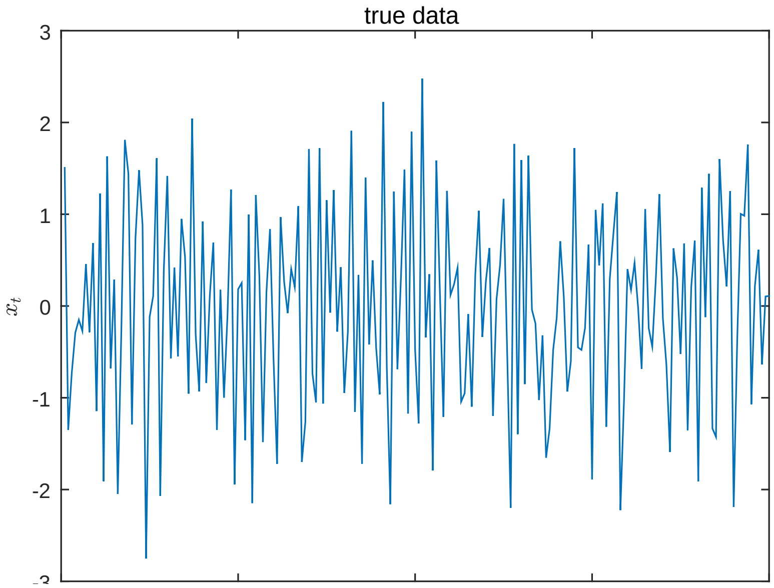


FIGURE 22.4.1 – Simulated 'true' data

For our case, the $L = 4$ data moment vector obtained from this simulation is

$$\hat{M}_{N=200}(X) = \begin{bmatrix} -0.0153 \\ 1.1874 \\ -0.4269 \\ -0.0868 \end{bmatrix}, \tag{22.14}$$

compared to the population moments given by

$$\mu(X) = \begin{bmatrix} 0 \\ 1.25 \\ -0.5 \\ 0 \end{bmatrix}.$$

We can use this data to estimate the variance-covariance matrix. To correct for possible autocorrelation in $\mathbf{m_{HN}}(X, \theta)$, we can apply the Newey-West correction to what was presented in (22.10). In particular, letting

$$\widehat{\Gamma}_{N,j} \equiv \frac{1}{N} \sum_{t=j+1}^{T} \left[M_N(X_t) - \hat{M}_N(X) \right] \left[M_N(X_{t-j}) - \hat{M}_N(X) \right]'$$

denote the j-th autocovariance of $M_N(X)$. Then the estimated sample variance-covariance matrix of $M_N(x_t)$ is given by

$$\widehat{S}_{x,N} = \widehat{\Gamma}_{N,0} + \sum_{j=1}^{\infty} \left(1 - \frac{j}{i(N)+1} \right) (\widehat{\Gamma}_{N,j} + \widehat{\Gamma}'_{N,j})$$

where $i(N)$ is the key to the Newey-West correction (here taken to be 4). The sample variance-covariance matrix is given by

$$\hat{S}_{X,N=200} = \begin{bmatrix} 0.4147 & 0.0058 & -0.0895 & -0.0244 \\ 0.0058 & 1.8946 & -0.8869 & -0.1872 \\ -0.0895 & -0.8869 & 1.2988 & -0.6078 \\ -0.0244 & -0.1872 & -0.6078 & 1.5729 \end{bmatrix}, \tag{22.15}$$

compared to the population variance-covariance matrix

$$S_x = \begin{bmatrix} 0.25 & 0 & 0 & 0 \\ 0 & 4.125 & -2.5 & 0.5 \\ 0 & -2.5 & 2.3125 & -1.25 \\ 0 & 0.5 & -1.25 & 2.0625 \end{bmatrix}.$$

22.4.3.2 SMM Estimation

We first draw a series of random sample $\{\{e_t^h\}_{t=1}^N\}_{h=1}^H$. We will use the same draw in the whole estimation process. Given parameter value θ, we can compute $\{\{Y_t^h(\theta)\}_{t=1}^N\}_{h=1}^H$ in (22.3) or

$$Y_t^h(\theta) = e_t^h - \theta e_{t-1}^h$$

where $e_0^h = 0$, $N = 200$, and $H = 10$. Then given θ, we can compute the simulated moment

$$\hat{M}_{HN}(Y(\theta)) = \frac{1}{H}\sum_{i=1}^{H}\frac{1}{N}\sum_{t=1}^{N} M_{HN}(Y_t^h(\theta)).$$

Our objective is to choose θ so that the weighted sum of squared residuals between the model moments $\hat{M}_{HN}(Y(\theta))$ and data moments $\hat{M}_N(X)$ is minimized. In the first stage, we will use the $L \times L$ identity matrix I as a weighting matrix $W = I$ and obtain a consistent estimate of the parameter θ which solves

$$\begin{aligned} \hat{\theta}_{HN}^1 &= \arg\min_{\theta} \mathcal{F}_{HN}(\theta) \\ \text{where} \quad \mathcal{F}_{HN}(\theta) &\equiv \Big[\hat{M}_N(X) - \hat{M}_N^H(Y(\theta))]'[\hat{M}_N(X) - \hat{M}_{HN}(Y(\theta))\Big]. \end{aligned}$$

Figure 2 plots the objective function $\mathcal{F}_{HN}(\theta)$ weighted by the identity matrix [[NOTE THAT SUBSEQUENT FIGURES TAKE PARAMETER VECTOR AS b INSTEAD OF θ]].

The consistent value of the estimator in this case is

$$\hat{\theta}_{HN}^1 = 0.4850.$$

In the second stage, we will use the inverse of the long-run variance-covariance matrix S as a weighting matrix to obtain the efficient SMM estimator. There are two ways to implement this.

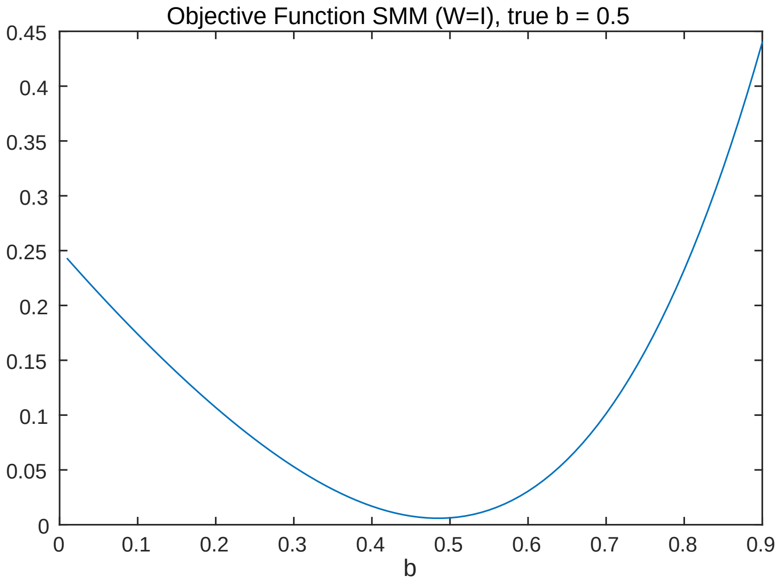


FIGURE 22.4.2 – $\mathcal{F}_{HN}(\theta)$ when $W = I$

Optimal Weighting Matrix from Data We have already computed the variance covariance matrix in (22.15). Then, the weighting matrix is the inverse of (22.15), and it is given by

$$W_N^* = \begin{bmatrix} 2.5146 & 0.2155 & 0.4282 & 0.2301 \\ 0.2155 & 1.0110 & 0.9316 & 0.4836 \\ 0.4282 & 0.9316 & 1.8197 & 0.8206 \\ 0.2301 & 0.4836 & 0.8206 & 1.0140 \end{bmatrix}$$

Figure 3 plots the objective function $\mathcal{F}_{HN}(\theta)$ when using the optimal weighting matrix W_N^*.

By minimizing this function, we can obtain the efficient SMM estimator[7]

$$\hat{\theta}^2_{HN,data} = 0.4993.$$

---

[7] If we use the actual data, we don't need $\hat{b}^1_{TH}$ to estimate the variance-covariance matrix because we only use the actual data. Hence this is actually not the second step estimator.

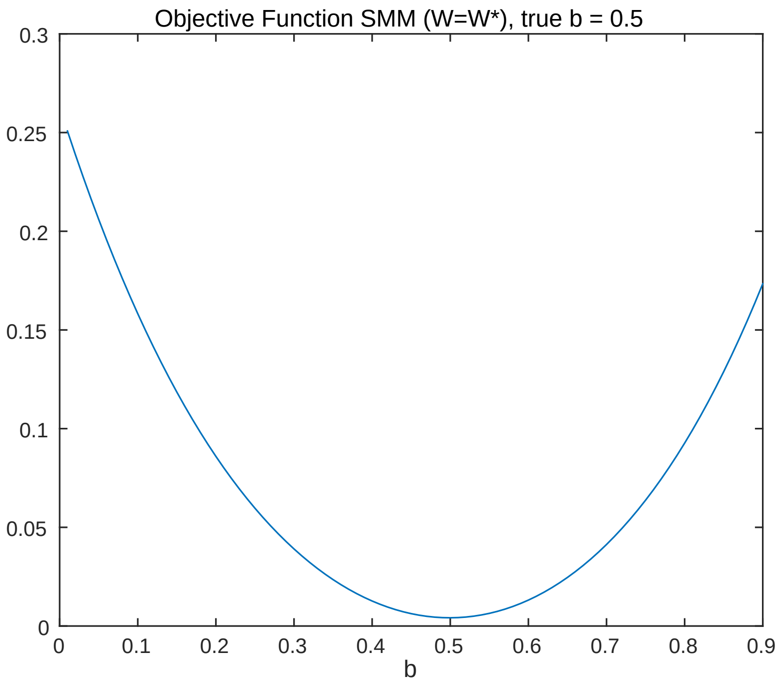


FIGURE 22.4.3 – $\varGamma_{HN}(\theta)$ when $W = W_N^*$

Optimal Weighting Matrix from Simulation Instead of the true data, we can use the simulated data under $\theta = \hat{\theta}^1_{TH}$ (which is consistent) to estimate the variance-covariance matrix. The variance-covariance matrix in this case is

$$\hat{S}_{y,HN} = \begin{bmatrix} 0.4472 & -0.0600 & 0.0459 & -0.0348 \\ -0.0600 & 3.4277 & -1.9441 & 0.3156 \\ 0.0459 & -1.9441 & 1.8748 & -0.8975 \\ -0.0348 & 0.3156 & -0.8975 & 1.7306 \end{bmatrix}.$$

Then the weighting matrix is

$$W^*_{HN} = \begin{bmatrix} 2.2441 & 0.0369 & 0.0023 & 0.0395 \\ 0.0369 & 0.8928 & 1.1272 & 0.4225 \\ 0.0023 & 1.1272 & 2.1335 & 0.9009 \\ 0.0395 & 0.4225 & 0.9009 & 0.9688 \end{bmatrix}.$$

If we use this as the weighting matrix, we obtain the second stage estimate[8]

$$\hat{\theta}^2_{HN,sim} = 0.4970.$$

---

[8] While there may be differences in the objective function when we use the optimal weighting matrix derived from the true and simulated data, in this case we did not find a large enough change so we don't plot it.

Standard Errors and J Test Once we have computed the estimator, we want to compute the standard errors of the estimator. We have

$$\sqrt{T}(\widehat{\theta}^2_{HN} - \theta_0) \to N(0, (1 + 1/H) \left[\nabla_\theta g_{HN}(\widehat{\theta}^2_{HN})' \widehat{S}^{-1}_{y,HN} \nabla_\theta g_{HN}(\widehat{\theta}^2_{HN})\right]^{-1})$$

where

$$\begin{aligned} g_{HN}(\theta) &\equiv \frac{1}{N}\sum_{t=1}^{N} \mathbf{m_{HN}}(x_t, \theta) = \frac{1}{N}\sum_{t=1}^{N} M_N(x_t) - \frac{1}{H}\sum_{h=1}^{H}\frac{1}{N}\sum_{t=1}^{N} M_{HN}(Y_t^h(\theta)) \\ &= \hat{M}_N(X) - \hat{M}_{HN}(Y(\theta)) \end{aligned}$$

So the derivative of g_N is given by

$$\begin{aligned} \nabla_\theta g_{HN}(\hat{\theta}^2_{HN}) &= -\nabla_\theta \hat{M}_{HN}(Y(\hat{\theta}^2_{HN})) \\ &= -\frac{1}{HN}\sum_{h=1}^{H}\sum_{t=1}^{N}\frac{\partial M_{HN}(Y_t^h(\hat{\theta}^2_{HN}))}{\partial\theta} \end{aligned}$$

In general we don't have an analytical formula for this derivative, so we will use the numerical derivative. Once can compute $\hat{M}_{HN}(Y(\hat{\theta}^2_{HN}))$, then compute $\hat{M}_{HN}(Y(\hat{\theta}^2_{HN} - s))$, take the difference, and divide by the step size s. The result is

$$\frac{\Delta \hat{M}_{HN}(Y(\hat{\theta}^2_{HN}))}{\Delta\theta} = \begin{bmatrix} -0.0104 \\ 0.9342 \\ -0.9330 \\ -0.0234 \end{bmatrix}$$

Again, since there is a small sample error, this is broadly consistent with the theoretical result computed in (22.13) evaluated at $\theta_0 = 0.5$ given by $[0\ 1\ \ -1\ 0]'$.

The standard error of the estimator is

$$\text{Std}(\hat{\theta}^2_{HN}) = \frac{1}{N}\left[\nabla_\theta g_{HN}(\hat{\theta}^2_{HN})' \left\{\left(1 + \frac{1}{H}\right)\hat{S}_{y,HN}\right\}^{-1} \nabla_\theta g_{HN}(\hat{\theta}^2_{HN})\right]^{-1} = 0.089.$$

Once we have estimated the parameter, we can also test if the moment condition is true or not.

$$N\frac{H}{1+H} \times [\hat{M}_N(X) - \hat{M}_{HN}(Y(\hat{\theta}^2_{HN}))]' W^*_{HN} [\hat{M}_N(X) - M_{HN}(Y(\hat{\theta}^2_{HN}))] = 0.7588.$$

The asymptotic distribution of this test statistic is $\chi(L - J)$, where L is the number of moments ($= 4$) and J is the number of parameters ($= 1$). The p value is 0.14, so we cannot reject the hypothesis that the model is true.

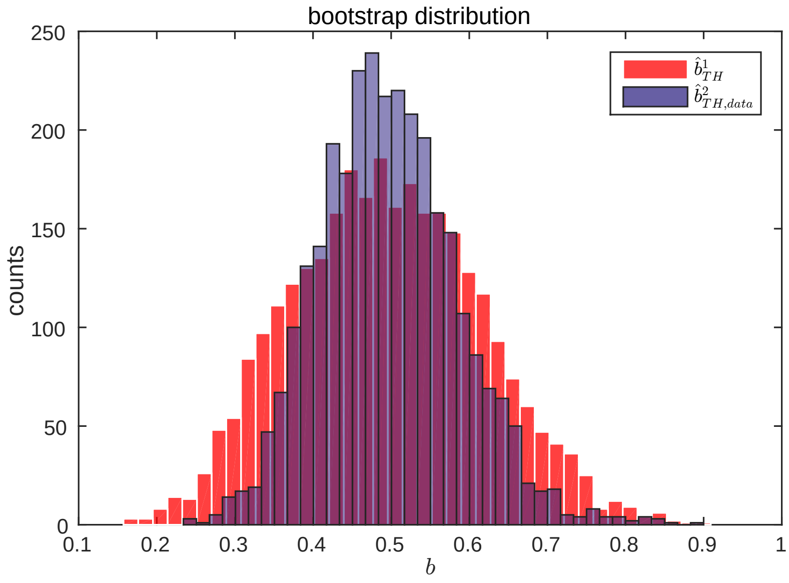


FIGURE 22.4.4 – Bootstrap distributions: histogram

22.4.4 Bootstrap

In order to see the finite sample distribution of the estimators, we can use the bootstrap method. The algorithm is as follows.

- **A** Draw ε_t and e_t^h from $N(0,1)$ for $t = 1, 2, \ldots, N$ and $h = 1, 2, \ldots, H$. Compute $(\hat{\theta}^1_{HN}, \hat{\theta}^2_{HN,data}, \hat{\theta}^2_{HN,sim})$ as described.
- **B** Repeat 1 using another seed.

Every time you do step 1, the seed needs to change (which is done automatically by matlab if you don't specify it). Otherwise you will keep getting the same estimators.

The histogram of the estimator is plotted in figure 4. As theory predicts, $\hat{\theta}^2_{HN,data}$, which is the efficient estimator, has a smaller variance than $\hat{\theta}^1_{HN}$. To make it easier for us to compare the distributions, figure 5 plots the density function of the estimators, obtained by Kernel density estimation

$$f(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{cI} \sum_{i=1}^{I} \exp\left[-\frac{1}{2}\left(\frac{x - x_i}{c}\right)^2\right]$$

where I is the number of data and c is the bandwidth. We can see that the distribution of $\theta^2_{HN,data}$ looks very similar to that of $\theta^2_{HN,sim}$. This is because the model nests the true DGP (in the sense that it is the true DGP at θ_0), so even if we use the simulated data to estimate the variance-covariance matrix, we can obtain the efficient estimator.

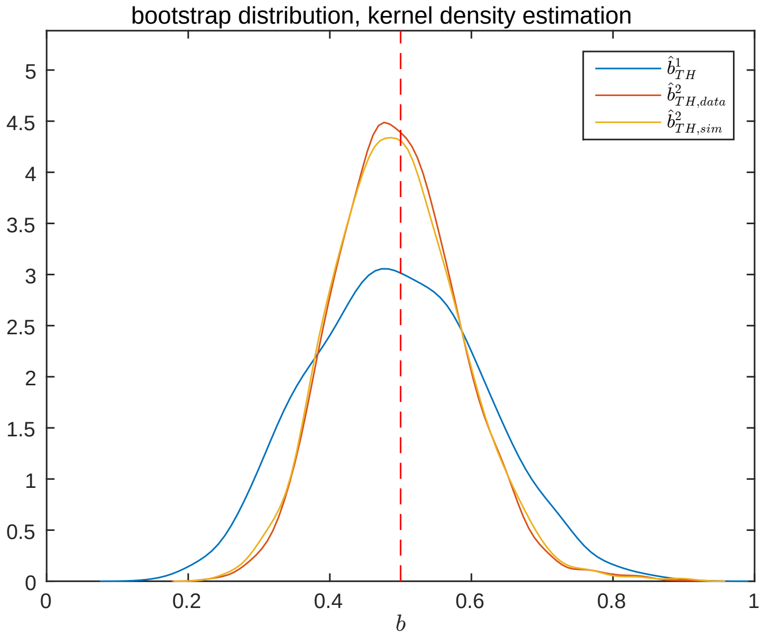


FIGURE 22.4.5 – Bootstrap distributions, approximated by the kernel density estimation.

22.5 Indirect Inference Estimation

In his review article, Smith (2008) gives a concise description of indirect inference:

> Like other simulation-based methods, indirect inference requires only that it be possible to simulate data from the economic model for different values of its parameters. Unlike other simulation-based methods, indirect inference uses an approximate, or auxiliary, model to form a criterion function. The auxiliary model does not need to be an accurate description of the data generating process. Instead, the auxiliary model serves as a window through which to view both the actual, observed data and the simulated data generated by the economic model: it selects aspects of the data upon which to focus the analysis.

Indirect inference is a simulation-based method for estimating the parameters of economic models, by mapping the model to the data via an "auxiliary statistical model."[9] It is most useful in estimating models for which the likelihood function (or any other criterion function to be maximized) is analytically intractable or too difficult to evaluate, as is the case here.

---

[9] It was first introduced by Smith (1990, 1993) and later extended by Gourieroux et al. (1993) and Gallant and Tauchen (1996)

The hallmark of indirect inference is the use of an “auxiliary model,” which is a statistical model that captures aspects of the data upon which to base the estimation. An important advantage over GMM is that this auxiliary model does *not* need to correspond to any valid moment condition of the economic model for the structural estimates to be consistent. This allows significant flexibility in choosing an auxiliary model: it can be any statistical model relating the model variables to each other *as long as* each structural parameter has an independent effect on at least one (reduced form) parameter of the auxiliary model. Technically, this last condition corresponds to the “full rank” requirement of the binding function (Smith (1993) and Gourieroux et al. (1993)). While the full rank condition is sufficient for the consistency of the estimates, as I discuss below, the precise specification of the auxiliary model does matter for the efficiency and small sample behavior of the estimator.

The mild set of assumptions needed on the auxiliary model allows one to incorporate many realistic features into the structural model without having to worry about whether or not one can directly derive the likelihood (or moment conditions for GMM) in the presence of these features. Estimation via auxiliary models allows one to think in terms of the *dynamic* structural relationships that characterize most economic models that are difficult to express as simple unconditional moments, as is often done with MSM. When estimating economic models it is useful to think of an auxiliary model as a reduced form of the structural model.

- Define the binding function.
- Give a formal description of things like pseudo-true values and other concepts relating to auxiliary models.

In what follows we elaborate on the key criteria that an auxiliary model should satisfy. Specifically:

- A key condition for an auxiliary model is that it should be feasible to estimate it both on real and simulated data. This can be a set of regressions, a likelihood to be maximized, etc. Generalized indirect inference allows one to slightly relax this assumption.
- it should be rich enough to provide a good descriptive statistical model of the data.
- The key condition is that it must be easy and fast to estimate.

22.5.1 General Procedure

It is useful to begin with an overview of the procedure:

Step 1. Estimate the auxiliary model using “real” data:[10]

$$\hat{\beta} = \arg\max_{\beta} \mathcal{L}(y^{US}; \beta).$$

---

[10] For example, if the auxiliary model is a multiple regression model, the likelihood would be expressed as this: $\mathcal{L}(y_{1,t} - \beta^{US}_{1,t} X_t, y_{2,t} - \beta^{US}_{2,t} X_t, ..., y_{N,t} - \beta^{US}_{N,t} X_t; \Sigma^{US})$

Step 2. Fix a parameter vector θ and use the structural model to generate M statistically independent simulated data sets, denoted with $\{y_{it}^m\}$ for $m = 1, 2, ..., M$. Using each simulated data set, obtain $\beta^M(\theta)$ by maximizing the likelihood of the auxiliary model:

$$\hat{\beta}^m(\theta) = \arg\max_{\beta} \mathcal{L}(y^m(\theta); \beta).$$

Average the resulting estimates to obtain:

$$\tilde{\beta}(\theta) = \frac{1}{M}\sum_{m=1}^{M} \hat{\beta}^m(\theta).$$

Step 3. The indirect inference estimator is obtained as the value of θ that makes $\hat{\beta}^{US}$ and $\tilde{\beta}(\theta)$ as "close" as possible. That is, using an appropriate distance function, d, define:

$$\theta^{II} = \arg\min_{\theta}(d(\hat{\beta},\tilde{\beta}(\theta))).$$

After this basic outline of the procedure, it is useful to discuss some of the key details. The first of which is: what is a good choice for the distance function?

22.5.2 Three Choices for Distance Function

The first two choices for an appropriate metric have been proposed by Smith (1990, 1993) and extended by Gourieroux et al. (1993).

1. Wald Approach[11]

The Wald objective is the familiar quadratic objective in the difference between the two parameter vectors:

$$\theta^{Wald} = \arg\min_{\theta}(\tilde{\beta}(\theta) - \hat{\beta})' \mathbf{W}(\tilde{\beta}(\theta) - \hat{\beta}),$$

where, as usual, **W** is a positive definite weighting matrix.

Examples using this approach include Low and Pistaferri (2012)..

2. Likelihood Ratio (LR) Approach

The likelihood ratio approach maximizes the likelihood of the auxiliary model using "real" data *but* evaluated using $\tilde{\beta}(\theta)$:

$$\theta^{LR} = \arg\max_{\theta} \mathcal{L}(y^{US}, \tilde{\beta}(\theta)).$$

An alternative and equivalent way to write this objective is:

---

[11] The Wald and LR approaches were first proposed in Smith (1990, 1993) and later extended by Gourieroux et al. (1993). The LM approach was first proposed by Gallant and Tauchen (1996).

$$\theta^{LR} = \arg\min_{\theta} \left[\mathcal{L}(y^{US}, \hat{\beta}) - \mathcal{L}(y^{US}, \tilde{\beta}(\theta))\right].$$

Since the first likelihood term does not contain θ, it appears as if we simply flipped the sign on the second likelihood and replaced the max operator with a min operator. While technically this is correct, the current formulation allows us to interpret the minimized objective as a goodness of fit statistic. For example, if our structural model as correctly specified and hence corresponded to the true DGP of "real" data this minimized objective would be zero. Any deviation from that tells us how well the model fits.

One advantage of the LR approach is that it does not require the estimation of a weighting matrix. This saves you the time to write the code for it (which is a non-trivial task) and also mitigates the issues that arise with the "optimal" weighting matrix in the Wald and LM approaches (as well as in GMM more generally). There is of course weighting going on, but it is different than the optimal one in this framework. In other words, the likelihood approach can be restated as the minimization of a quadratic objective with a non-optimal weighting matrix (see Guvenen and Smith (2009) for a proof). But the difference is asymptotically small. And in small sample it works much better. Notice that in this approach each moment is weighted by its contribution to the likelihood of the auxiliary model. This provides a powerful way of combining coefficients that have very different scales. Also note that in our way of doing, we are also using the information in the variance-covariance matrix.

This objective has been employed by Keane and Smith (2003) and Guvenen and Smith (2014).

3. LM approach

The LM approach, first proposed by Gallant and Tauchen (1996), minimizes a quadratic objective in the average score vector of the auxiliary model:

$$\theta^{LM} = \arg\min_{\theta}(\frac{1}{M}\sum_{m=1}^{M}\mathcal{L}_{\beta}(y^m(\theta), \beta^{US}))'\mathbf{V}(\frac{1}{M}\sum_{m=1}^{M}\mathcal{L}_{\beta}(y^m(\theta), \beta^{US})),$$

where $\mathcal{L}_\beta$ denotes the score vector (i.e., vector of partial derivatives of $\mathcal{L}$ with respect to the elements of β^{US}), and $\mathbf{V}$ is a positive definite matrix.

This approach has certain advantages over the first two. It is widely used in financial econometrics and has been implemented by Nagypal (2007) in labor economics, and by Bansal et al. (2007) in asset pricing.

22.6 Example: Lifecycle Model with Kalman Filtering

As we have seen in previous chapters, idiosyncratic earnings risk is one of the key features of the economic environment individuals face when making many economic decisions, from consumption and savings, to labor supply and fertility. In our modeling of earnings risk up to this point (e.g., in the income-fluctuation problem in Chapter 12, the Aiyagari (1994) model in Chapter 17, etc.), we have made the implicit assumption that an econometric process for earnings (such as an AR(1) plus transitory shock model) fitted to panel data on earnings *fluctuations* provides a good representation for the earnings *risk perceived* by the workers. This assumption puts the econometrician and the individuals whose data is being used on equal footing, for example, ruling out the possibility that the individual has some advance information about his future earnings before it is realized or that the individual may not be able to tell apart the persistent and transitory components of an earnings shock he receives immediately as the income process postulates.

In this section, we present a slightly simplified version of the model studied in Guvenen and Smith (2014). Consider a lifecycle model in which individuals have consumption preferences that take the CRRA form: $U(c) = C^{1-\sigma}/(1-\sigma)$ and solve a consumption-savings problem, subject to borrowing constraints.

The goal of the paper is to estimate the degree of income uncertainty faced by workers in the labor market. To that end, let y_t^i denote the log (labor) income of individual i at age t:

$$y_t^i = \underbrace{g\left(t, \text{educ}\right)}_{\text{common life cycle component}} + \underbrace{\left[\alpha^i + \beta^i t\right]}_{\text{heterog. income profiles}} + \underbrace{\left[z_t^i + \varepsilon_t^i\right]}_{\text{stochastic component}}, \tag{22.1}$$

$$z_t^i = \rho z_{t-1}^i + \eta_t^i \tag{22.2}$$

where $g(.)$ captures the common lifecycle component, which is a quadratic polynomial in age with coefficients that vary across education groups; $\left[\alpha^i + \beta^i t\right]$ account for potential *Heterogeneity* in *Income Profiles* (or HIP as in Guvenen (2007b)); z_t follows an AR(1) process, and η_t^i and ε_t^i are independent of each other and $i.i.d$ over time and across individuals.

Information Structure. At the heart of the model is a Kalman filtering problem. Individuals know α^i, observe y_t^i and ε_t^i, and must learn about $\left(\beta^i, z_t^i\right)$. Before entering the labor market, each individual observes a signal about his β^i and use the Kalman filter to update their beliefs about $\left(\beta^i, z_t^i\right)$ as new observations on labor income are realized every period. Prior beliefs are given by a bivariate normal distribution with mean $(\widehat{\beta}_{1|0}^i, \widehat{z}_{1|0}^i)$ and covariance matrix

$$\mathbf{P}_{1|0} = \begin{bmatrix} \sigma_{\beta,0}^2 & 0 \\ 0 & \sigma_{z,0}^2 \end{bmatrix}.$$

For convenience, let us express the prior standard deviation as a fraction of total heterogeneity in the population: $\sigma_{\beta,0} = \lambda\sigma_\beta$. If $\lambda = 0 \to \sigma_{\beta,0} = 0$ and $\widehat{\beta}_{1|0}^i = \beta^i$ (No prior uncertainty). If $\lambda = 1 \to \sigma_{\beta,0} = \sigma_\beta$, and $\widehat{\beta}_{1|0}^i = \overline{\beta}$ (Full prior uncertainty). For realistic parameter values learning about β^i is very slow (Guvenen (2007b)).

The Dynamic Programming Problem of an individual is given by:

$$\begin{aligned} V_t(\omega_t, \widehat{\beta}_t, \widehat{z}_t) &= \max_{a_{t+1}} \left\{ U(C_t) + \beta \mathbb{E}_t \left[V_{t+1}(\omega_{t+1}, \widehat{\beta}_{t+1}, \widehat{z}_{t+1}) \right] \right\} \\ &\text{s.t.} \qquad (22.3) \\ C_t + a_{t+1} &= \omega_t = (1+r)\,a_t + Y_t \qquad (22.4) \\ a_{t+1} &\geq \underline{a}, \\ &\text{and} \quad \text{Kalman recursions} \end{aligned}$$

and $V_{T+1} \equiv 0$.

Identification and Choosing An Auxiliary Model

A useful way to think about an auxiliary model is as a “reduced form” for the true structural model with the key property that it lends itself to fast and efficient estimation (unlike the true structural model itself). Notice that if we could derive some explicit structural relationships directly from the true structural model, it would be better to estimate them directly via GMM. As discussed above, the GMM approach has the advantage of not needing to specify the full model making it more robust to model specification (as well as often being much faster). The problem is that this is often impossible. That is where the words “reduced form” come into play. The procedure is then to *simplify the true structural model in the most minimal way possible* to obtain a version that does allow the derivation of structural relationships. These equations then provide an excellent starting point for finding a good auxiliary model. Here is how we proceed.

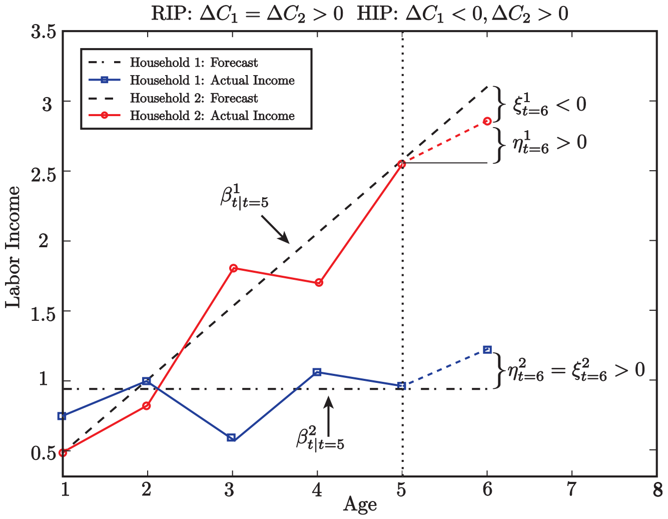
Suppose we make the following modifications to the model above: (i) instead of CRRA, assume quadratic utility, (ii) no borrowing constraints, (iii) $\varepsilon_t = 0$, and (iv) assume that y_t is the level (instead of log) of income. Under these new assumptions, the entire model can be solved in closed form. Then, the decision rule for consumption can be manipulated to obtain the following key equation:

$$\Delta C_t = \Pi_t \times \underbrace{\left(y_t^i - \left(\alpha + \widehat{\beta}_{t-1}^i t + \rho \widehat{z}_{t-1}^i \right) \right)}_{\widehat{\xi}_t} \qquad (22.5)$$

where Π_t is function of $\left(r, \rho, \lambda, P_{t|t}, R, T\right)$. Here, $\widehat{\xi}_t$ is the *perceived innovation* to income: it is the gap between income realized at t and the individual's forecasted income given his beliefs at $t-1$. Notice that $\widehat{\xi}_t$ depends on beliefs, which in turn depend on (i) prior beliefs and (ii) the entire history of income realizations. Therefore, the consumption response (i.e., change) at time t depends on *an individual's entire earnings history.*

This is very different from what is implied by the standard permanent income hypothesis (PIH) model. To see this, let us assume $\sigma_\beta^2 \equiv 0$ so that there is no heterogeneity in income growth rates and therefore $\beta^i \equiv \overline{\beta}$. With β^i known, observing y_t immediately reveals z_t, leaving no room for Bayesian learning. Therefore, the model

FIGURE 22.6.1 – Identification of Key Parameters from Consumption *Growth Rates*



reduces to the PIH and consumption growth is given by

$$\Delta C_t = \Psi_t \times \eta_t, \tag{22.6}$$

where $\Psi_t(r, \rho, R, T)$ is another known coefficient. Equation (22.6) shows the familiar condition that consumption changes proportionally to the *true innovation to income* in the current period, η_t. History plays no role in determining consumption growth today. This is in stark contrast to equation (22.5) in which it is the *innovation to beliefs* that drives consumption response.

To understand how these different implications can be exploited for identification, we discuss three scenarios. First, consider the top panel of Figure E.1, which plots the earnings histories of two individuals over 6 periods. Let us focus on the consumption response of each individual between periods five and six, and see what we can infer from them about the model parameters. We drew the figure such that both workers experience the same income shock in period 6: $\eta^1_{t=6} = \eta^2_{t=6} > 0$. In the general model with $\sigma^2_\beta > 0$, Bayesian learning is active, so the two workers enter period 5 with $\widehat{\beta}^1_{t|t=5} \gg \widehat{\beta}^2_{t|t=5} \approx 0$, given their different earnings growth histories. Even though both workers experience a positive income growth between periods five and six, the first worker is disappointed relative to his expectation, so $\widehat{\xi}^1_{t=6} < 0$ despite $\eta^2_{t=6} > 0$. Therefore, his consumption should decline according to equation (22.6). In contrast, the second worker is pleasantly surprised from any positive income growth $\widehat{\xi}^2_{t=6} > 0$, so his consumption should rise. Hence, we should observe $\Delta C^1_{t=6} < 0$ and $\Delta C^2_{t=6} > 0$ given the earnings histories in the figure.

In contrast, in the restricted model with $\sigma^2_\beta \equiv 0$, there would be no connection between past income growth and current consumption growth and both workers would

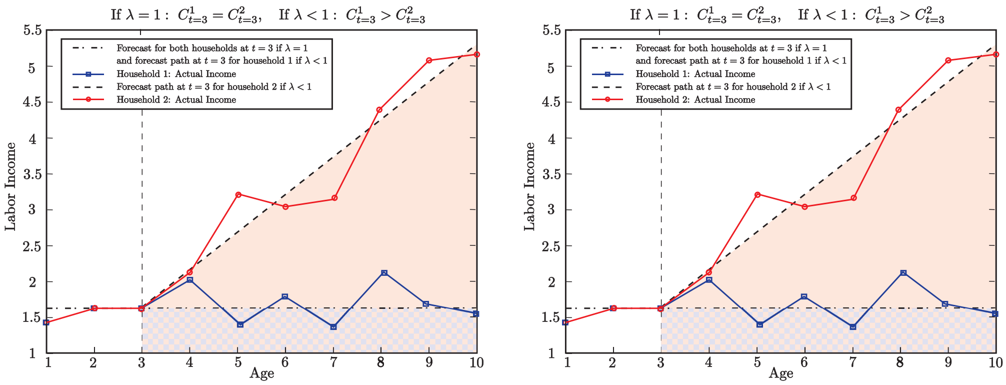


FIGURE 22.6.2 – Identification of Key Parameters from Consumption *Levels*

experience positive consumption growth: $\Delta C^1_{t=6} = \Delta C^2_{t=6} = \eta^2_{t=6} > 0$. This thought experiment implies that the nature of the dependence of consumption growth on a worker's earnings history (or lack thereof) identifies whether there is HIP and Bayesian learning or not: the joint hypothesis of $\sigma^2_\beta > 0$ and $\lambda > 0$ versus $\sigma^2_\beta \equiv 0$ (in which case λ is irrelevant).

Proposition 22.3. [Information in Consumption Growth] *In the life-cycle certainty-equivalent model with permanent shocks described above, consumption growth satisfies the following properties:*

$\boxed{\textbf{A}}$ Controlling for current income growth, on average, consumption growth will be a decreasing function of an individual's true β^i, that is, $\frac{\partial \Delta C^i_t}{\partial \beta^i} < 0$.

$\boxed{\textbf{B}}$ Although the prediction in (i) holds for all values of λ (including when $\lambda = 0$ and, hence, when there is no learning), the relationship becomes stronger as λ rises, that is, $\frac{\partial^2 \Delta C^i_t}{\partial \beta^i \partial \lambda} < 0$ for all t.

$\boxed{\textbf{C}}$ Similarly (holding everything else constant), the response of consumption growth to income growth becomes stronger as λ increases: $\frac{\partial^2 \Delta C^i_t}{\partial \Delta Y^i_t \partial \lambda} < 0$.

We can similarly work through the other two examples in Figure E.1 to prove that consumption levels and their relationships to past earnings levels identify key parameters. See Propositions 2 and 3 in Guvenen and Smith (2014) for details. **[To Be Added Here]**

22.7 Smoothing the Objective

A key consideration in doing simulation-based estimation concerns whether the objective function varies with the structural parameters in a smooth (continuous and differentiable) fashion. The general rule of thumb is that, unless we are extra careful, structural estimation almost always manages to introduce kinks and jumps into the objective function. This is a first order issue, for example, if any of the moments depend on a discrete choice made by the individuals in our model. Not surprisingly, the earliest papers on this topic (McFadden (1989), Pakes and Pollard (1989), Keane (1994), and others) focused specifically on how to simulate probabilities in a smooth fashion in discrete-choice models.

For example, in models with fixed costs, such as lumpy investment, housing, employment, etc., such discrete choices arise easily and if our moments depend on these, the objective function is likely to have jumps, making derivative based methods inapplicable or less effective. Similarly, if we choose moments, such as the percentiles of a distribution, etc. we are going to have jumps. Finally, and most commonly, some approximation and interpolation methods (e.g., linear interpolation) that are used in solving dynamic programming problems can easily introduce kinks and jumps into the numerical solution, even when the underlying true model would have smooth decision rules. So what to do? [Talk about ways to smooth these statistics]. Here we discuss ways to "smooth" the objective.

22.7.1 Generalized Indirect Inference

The indirect inference method as described so far requires us to estimate precisely the same auxiliary model on the actual data and on the simulated data. But in some applications, a natural auxiliary model that makes sense for the real data could imply an objective for the indirect inference procedure that is not smooth. A typical example is the estimation of discrete choice models. Keane and Smith (2003) propose and implement a clever generalization of the indirect inference procedure to deal with such situations. The method can be described in the following simple steps:

A Pick the natural auxiliary model for the discrete choice real-world data. Denote its likelihood as $\mathcal{L}_n(y, \beta)$ for a given sample size n.

B Specify an auxiliary model on simulated data parameterized with λ, denoted with $\tilde{\mathcal{L}}(y, \beta; \lambda)$ that has two features:

(a) in fixed samples, it is smooth: $\partial \tilde{\mathcal{L}}_n(y, \beta; \lambda)/\partial \beta$ is continuous, and

(b) as the sample size grows and we let λ go to zero, $\tilde{\mathcal{L}}_n$ converges to $\mathcal{L}_\infty$:

$$\lim_{\lambda \to 0, n \to \infty} \tilde{\mathcal{L}}_n(y, \beta; \lambda) = \mathcal{L}_\infty(y, \beta)$$

Under these assumptions, we are estimating different auxiliary models on real data and on simulated data but the difference vanishes as the sample size increases.

An example should help clarify how this works. Keane and Smith (2003) estimate a discrete choice model where the natural auxiliary model to be estimated on real data is something like a probit or a logit.

Denote the latent utility of choices $j = 1, ..., J - 1$ for individual i be denoted with $u^m_{itj}(\theta)$ in the structural model with parameter vector θ and mth simulated data set. The corresponding choices y^m_{itj} are indicators (1 meaning that alternative j is chosen by individual i, 0 means otherwise). The auxiliary model on simulated data will be taken to depend on the underlying latent utility and not the step function given by the indicators:

$$g(u_{itj}^m(\theta), \lambda) = \frac{\exp(u_{itj}^m(\theta)/\lambda)}{1 + \sum_{k=1}^{J-1} \exp(u_{itk}^m(\theta)/\lambda)}.$$

For positive values of λ away from zero, this function will be smooth and continuous in θ. As λ approaches zero however, it will converge to y_{itj}^m—which is a step function that assigns a value of 1 to alternative j that is the most preferred option and zero to all others.

A key question, as usual, is how to choose λ for a given, finite, sample size. The trade-off is clear: a large λ yields a smoother objective at the expense of bias, because $\tilde{\mathcal{L}}_n$ will differ from $\mathcal{L}_n$ by a larger extent. A smaller λ will be the opposite: smaller bias by bringing $\tilde{\mathcal{L}}_n$ and $\mathcal{L}_n$ closer together, at the expense of choppier objective. The choppiness can be mitigated by increasing M as usual, but this then slows down the estimation. Keane and Smith advocate a two step approach. In the first step, they propose to take a large value of λ and a small value of M. This makes the objective very smooth, and produces a consistent but biased estimate of θ_0—denote this $\hat{\theta}_1$. In the second step, they take a large value of M and a small value of λ. A larger M reduces the extra choppiness coming from the choice of a small λ. However, now this estimation faces all the challenges highlighted above, of being slow and difficult to maximize. Keane and Smith suggest not to fully reestimate the model with this new objective but rather to take a one Newton-Raphson step from $\hat{\theta}_1$, which is now asymptotically equivalent to the estimate maximizing the objective. In the LR approach, this step is:

$$\hat{\theta}_2 = \hat{\theta}_1 - \left(\hat{J}^T \mathcal{L}_{\beta\beta}(y, \beta(\hat{\theta}_1))\hat{J}\right)^{-1} \hat{J}^T \mathcal{L}_{\beta}(y, \beta(\hat{\theta}_1)),$$

which is asymptotically equivalent to the Generalized Indirect Inference (GII) estimate based on the LR approach. Here $\hat{J}$ is an estimate of the Jacobian of the binding function ($\beta(\theta)$) and $\mathcal{L}_{\beta\beta}$ is the Hessian of the likelihood of the auxiliary model, all evaluated at $\hat{\theta}_1$. Keane and Smith provide Monte Carlo results showing that this approach to estimating discrete choice models with GII can be quite a bit faster and easier than maximizing the choppy likelihood under indirect inference and simulate maximum likelihood.

22.8 Economic Applications

TBC

➢

22.9 Literature and Further Reading

TBC

22.10 Taking Stock

In the context of discrete choice models on cross-sectional data, McFadden (1989) first extended GMM to allow for simulated moments. Pakes and Pollard (1989) proved the asymptotic properties of such simulation estimators. Lee and Ingram (1991a) applied it to the estimation of time-series models and Duffie and Singleton (1993) followed with a (time-series) asset pricing application. Since its inception in the early 1990s, indirect inference has been embraced and found a wide range of applications in financial econometrics and has made a more modest impact on other fields of economics, despite holding much promise. Tauchen (1997) contains a survey and references of these finance applications.

For discrete choice models, Keane (1994) extended McFadden's method to a panel data and Keane and Smith (2003) developed a generalized indirect inference estimator for the same problem. Smith (2008) contains a concise overview of indirect inference.

Magnac et al. (1995) and Nagypal (2007) estimated discrete-choice models and Low and Pistaferri (2012) study the insurance value and incentive costs of disability benefits system.

If the model is overidentified, then this objective need not be zero and we can test whether the model is correct via a J-test.