

Faezeh Raei
Summer Class on Dynamic Programming
at Sharif University of Technology,
Summer 2006

References:

- Adda & Cooper, chapter 3,4,5
- Professor Stephen Donald's lecture notes available at <http://www.eco.utexas.edu/~donald/quantsem/week1t4.pdf>

1 Review of Methodology

- Building an economic model, described by a set of parameters and a choice structure. For example stochastic growth model, search models, discrete choice models, etc.
 - characterize the model by first order conditions or write it as a recursive problem.
- Solving the model for policy rules.
 - If not analytically solvable, use numerical and computational methods like value function iteration.
- Estimating the parameters of the model.
 - Having the policy rules, choose the optimal parameters to make the predictions of the model close to observed data.
 - check the goodness of fit.
- Using the model to evaluate different scenarios once confident enough that the model is a convincing representation of the economy. (next week's topic)

2 Estimating the parameters of the model

2.1 Maximum likelihood

Example 1 Coin flipping Suppose we have a not necessarily fair coin and we have observed a sample $\{x_1, x_2, x_3, \dots, x_T\}$ of coin draws where each x_i is heads or tails. Suppose we have observed N_1 tails and N_2 heads. We like to model this coin flipping as a probabilistic model where the draws are iid and probability of tails is p . By estimating the model we mean to find the parameter p such as the prediction of the model is close to that of observed data.

By the model, the likelihood of observing the sample $\{x_1, x_2, x_3, \dots\}$ is

$$L(x, p) = p^{N_1} (1 - p)^{N_2}$$

the maximum likelihood estimator is given by

$$p^* = \arg \max L$$

and to find it, take the first order conditions of L with respect to p and we get

$$p^* = \frac{N_1}{N_1 + N_2}$$

Example 2 Consider the discrete choice cake eating problem, described by the following Bellman equation, where W is the size of the cake, ρ is the shrinkage factor and ε is an iid shock to preferences:

$$V(W, \varepsilon) = \max\{\varepsilon u(W), EV(\rho W, \varepsilon')\}$$

$V(\cdot)$ represents the value of having a cake of size W given the realization of the taste shock ε . Let $\varepsilon^*(W, \theta)$ be defined by

$$\varepsilon^* u(W) = EV(\rho W, \varepsilon')$$

where θ is a vector of parameters relating to distribution of ε . If $\varepsilon > \varepsilon^*(W, \theta)$ the individual will eat the cake otherwise will wait. $\varepsilon^*(W, \theta)$ has no analytical expression but can be solved for numerically. The probability of not consuming a cake of size W in a given period is then

$$P(\varepsilon < \varepsilon^*(W, \theta)) = F(\varepsilon^*(W, \theta))$$

where F is the cumulative density of the shock ε . The likelihood of observing an individual i consuming the cake after t periods is

$$l_i(\theta) = (1 - F(\varepsilon^*(\rho^t W, \theta))) \times \prod_{l=1}^{t-1} F(\varepsilon^*(\rho^l W, \theta))$$

suppose we observe the cake eating behavior of N individuals then the likelihood of the sample is

$$L(\theta) = \prod_{i=1}^N l_i(\theta)$$

the maximization of L with respect to θ gives the estimate $\hat{\theta}$. Notice here that the estimated parameter $\hat{\theta}$ possibly depends on the number of observations however the true parameter of the model, θ does not. A question will arise as to how properties of $\hat{\theta}$ change as the sample size changes and whether it converges to any number θ^* ?

Example 3 For a full-fledged application of maximum likelihood in stochastic growth, see Kocherlakota et al(1994 Journal of Monetary Economics) Their goal

is to evaluate the contribution of technology shocks to aggregate fluctuations. They construct a model that includes shocks to the production function and stochastic depreciation of capital.

Properties of maximum likelihood estimator

The maximum likelihood approach requires an assumption on the distribution of random variables. Denote by $f(x_t, \theta)$ the probability of observing x_t given a parameter θ . The estimation method is designed to maximize the likelihood of observing a sequence of data $\{x_1, x_2, \dots, x_n\}$. Assuming iid shocks, the likelihood of the entire sample is

$$L_n(X, \theta) = \prod_{t=1}^n f(x_t, \theta)$$

it is easier to maximize the log of the likelihood

$$Q_n(X, \theta) = \frac{1}{n} \sum_{t=1}^n \log f(x_t, \theta)$$

Notice the functional limit of Q_n , Q is given by

$$Q(\theta) = E(\log f(x, \theta))$$

we are interested in properties like consistency and asymptotic normality.

Basic assumptions for consistency

- – Θ is compact, where $\theta \in \Theta$. (technical assumption)
- $Q_n(\theta)$ converges in probability uniformly to function Q (uniform law of large numbers)
- $Q(\theta)$ is continuous
- $Q(\theta)$ is maximized at the true parameter θ^* . (identification)

with these four conditions one can show that $\widehat{\theta}_n \rightarrow \theta^*$. The idea is that $\widehat{\theta}_n$ is the maximizer of $Q_n(\theta)$ over Θ . $Q_n(\theta)$ gets close to $Q(\theta)$ which is maximized at θ^* . so it must be that $\widehat{\theta}_n$ gets close to θ^* .

Basic assumptions for asymptotic normality

under assumptions

- – $\widehat{\theta}_n$ is consistent
- θ^* is in the interior of Θ
- $Q_n(\cdot)$ is twice continuously differentiable
- $\sqrt{n} \nabla_{\theta} Q_n(\theta^*) \rightarrow Z \sim N(0, A)$ central limit theorem
- $\nabla_{\theta\theta}^2 Q_n(\theta) \rightarrow \nabla_{\theta\theta}^2 Q(\theta) = B$

one can show that

$$\sqrt{n}(\widehat{\theta}_n - \theta^*) \rightarrow^d N(0, A^{-1})$$

the maximum likelihood estimator is asymptotically normal with mean zero and a variance equal to A^{-1}/n .

For of a proof of these claims see professor Donald's notes pp:26-28.

2.2 Generalized method of moments

An alternative method to estimate the parameters is to make a moment of data as close as possible to moments predicted by data. This can be called moment calibration versus the maximum likelihood method that was more of a path calibration.

Example 4 Again consider the coin flipping sample. Given a draw $\{x_1, x_2, \dots, x_T\}$ let's look at a data moment like the fraction of tails $\mu = \frac{N_1}{N_1 + N_2}$. The fraction of tails predicted by the model is p . The method of moments estimation tries to minimize the distance between the moment from data and that one from the model. Here we need to solve

$$\min_p \left(p - \frac{N_1}{N_1 + N_2} \right)^2$$

which obviously leads to $\widehat{p} = \frac{N_1}{N_1 + N_2}$, the same as MLE estimator.

The particular moment we chose was the fraction of tails. Often in the data there are many moments to choose from. The econometric theory does not come up with a clear answer of which moments to choose, however the moments should be informative about the parameters to be estimated. This means those moments should depend on the parameters in such a way that slight variations in parameters result in different values for the moments.

In general, let μ be a $m \times 1$ vector of moments calculated from data and $\mu(\theta)$ be corresponding moments from the model and θ be a $k \times 1$ vector of parameters to be estimated. If $k = m$ we say model is just identified. If $k < m$ the model is overidentified and if $k > m$ the model is underidentified in which case the estimation can not be implemented as there are too many unknown parameters. So if $k \leq m$, the GMM estimator $\widehat{\theta}$ comes from

$$\min_{\theta} (\mu(\theta) - \mu)' W^{-1} (\mu(\theta) - \mu)$$

in this form W^{-1} is a weighting matrix. Its choice is important for obtaining efficient estimators when model is overidentified.

Example 5 Orthogonality Conditions In many models the first order conditions take the form

$$E(h(\theta^*, x_t)) = 0$$

where h can be any function. These restrictions are called orthogonality conditions. The expectation in this condition, can be approximated with a a sample

average. The orthogonality condition says this average should be close to zero. Denote the sample average of $h(\theta, x_t)$ by

$$g(\theta) = \frac{1}{T} \sum_{t=1}^T h(\theta, x_t)$$

then an estimate of θ can be found from

$$\hat{\theta} = \arg \min_{\theta} g(\theta)' W_T^{-1} g(\theta)$$

where W^{-1} is a weighting matrix which might depend on data hence the subscript T .

Example 6 sometimes the moments has the form

$$E_t(h(\theta^*, x_t)) = 0$$

i.e. the value of the random variable $h(\theta^*, x_t)$ can not be predicted from information at time t . This implies that for every random variable z_t that belongs to the information set at time t , we have $cov(z_t, h(\theta^*, x_t)) = 0$ or $E(z_t \cdot h(\theta^*, x_t)) = 0$. This is also a form of orthogonality condition and can be used as described in example 5.

Example 7 In the standard intertemporal of model of consumption with stochastic income and no borrowing constraints, the first order condition gives

$$u'(c_t) = \beta R E_t u'(c_{t+1})$$

one can use this restriction to form $h(\theta, c_t, c_{t+1}) = u'(c_t) - \beta R u'(c_{t+1})$ where θ is parametrizing the utility function. On average $h(\theta, c_t, c_{t+1})$ should be equal to zero at the true value of the parameter. The Euler equation actually brings more information than we have so far. As explained in example 6, for every random variable z_t , at information set of time t , we can form a moment $g(\theta, c_t, c_{t+1}) = z_t \cdot [u'(c_t) - \beta R u'(c_{t+1})]$ and $g(\theta, c_t, c_{t+1})$ should on average be zero. So one can estimate θ by solving

$$\min_{\theta} \left(\frac{1}{T} \sum_{t=1}^T z_t \cdot [u'(c_t) - \beta R u'(c_{t+1})] \right)^2$$

if we have more than one z_t variable, say z_t and ω_t then we can exploit as many orthogonality conditions. In this case consider the moment (2×1 vector) defined by

$$h(\theta, c_t, c_{t+1}) = \begin{pmatrix} z_t \cdot [u'(c_t) - \beta R u'(c_{t+1})] \\ \omega_t \cdot [u'(c_t) - \beta R u'(c_{t+1})] \end{pmatrix}$$

and the estimator comes from

$$\min_{\theta} h(\theta, c_t, c_{t+1})' W h(\theta, c_t, c_{t+1})$$

where W is a 2×2 weighting matrix. The extra moments in the later example, give rise to a test of the model, as we will see in what follows.

Example 8 For a detailed application of this method see the following papers

- Kydland & Prescott (1982) *Econometrica*, Time to Build and Aggregate Fluctuations
- King, Plosser & Rebelo (1988) *Journal of Monetary Economics*, Production, Growth, Business Cycles I

Properties of GMM estimator

Let $\hat{\theta}_T$ be the GMM estimator, that is the solution to

$$\min_{\theta} g(\theta)' W_T^{-1} g(\theta)$$

where

$$g(\theta) = \frac{1}{T} \sum_{t=1}^T h(\theta, x_t)$$

then under regularity conditions, (see pages 22-30 of Dr. Donald's notes):

- $\hat{\theta}_T$ is a consistent estimator of θ
- The GMM estimator is asymptotically normal:

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \rightarrow^d N(0, \Sigma)$$

where $\Sigma = (DW_{\infty}^{-1}D')^{-1}$ and where

$$D' = p \lim_T \left\{ \frac{\partial g(\theta, X_T)}{\partial \theta'}(\theta_0) \right\}$$

the empirical counterpart of D' is

$$\widehat{D}'_T = \frac{\partial g(\theta, X_T)}{\partial \theta'}(\hat{\theta}_T)$$

this means that asymptotically one can treat GMM estimator $\hat{\theta}_T$ as $N(\theta_0, \frac{\Sigma}{T})$.

Optimal Weighting Matrix

The choice of weighting matrix has no effect on the convergence of the estimator to the true value but a one chosen properly can reduce the asymptotic variance of the estimator. Optimal weight can be shown to be estimated by

$$W_T = \frac{1}{T} \sum_{t=1}^T h(\theta, x_t) h(\theta, x_t)'$$

this choice minimizes the asymptotic variance of the GMM estimator.

Overidentifying Restrictions

If the number of restrictions m is greater than the number of parameters k , then the model is overidentified. One only needs k moments to estimate the k parameters. The remaining restrictions can be used to evaluate the model. Under the null hypothesis that the model is the true one, the additional moments should be close to zero at the true parameter. This forms the basis of a specification test

$$Tg(\hat{\theta}_T)W_T^{-1}g(\hat{\theta}_T) \rightarrow \chi^2(m - k)$$

one has to evaluate T times the value of the objective with the critical value of the chi-square.

Comparison with maximum likelihood

The choice of the method depends on the problem and data set. Path calibration methods like MLE use all the information in the data set and hence are usually more efficient. The drawback is that one has to specify the entire model including the distribution of the shocks. Sometimes for tractability one has to assume shocks are normal and this imposes too much restriction on the model.

Moment calibration methods like GMM, use only part of the information provided by data and mostly concentrate on some moments of the data like mean or variance. This method does not necessarily require the whole specification of the model.

MLE is generally more efficient than GMM. However GMM with optimal weighting matrix is as efficient as MLE.

2.3 Simulation based methods

In general in dynamic programming models, the likelihood function or the analytical form of the moments are difficult to write out, so simulation methods are of great help. However they come at a cost, since they are time consuming. These methods are often used because the calculations of the moments are too difficult for example when multiple integrals are involved. To see this point let's look at an example

Example 9 Again consider the discrete choice cake eating problem as described in example 2, but here assume correlated taste shocks hence the Bellman equation becomes

$$V(W, \varepsilon) = \max\{\varepsilon u(W), E_{\varepsilon'|\varepsilon} V(\rho W, \varepsilon')\}$$

where the expectation depends on ε as well. Still we can define $\varepsilon^*(W, \theta)$ by

$$\varepsilon^* u(W) = E_{\varepsilon'|\varepsilon} V(\rho W, \varepsilon') \quad (1)$$

The probability of waiting t periods to consume the cake is then

$$P_t = P(\varepsilon_1 < \varepsilon^*(W), \varepsilon_2 < \varepsilon^*(\rho W), \varepsilon_3 < \varepsilon^*(\rho^2 W), \dots, \varepsilon_t > \varepsilon^*(\rho^t W))$$

in the case of iid shocks this probability could have been decomposed to a product of t terms as

$$(1 - F(\varepsilon^*(\rho^t W, \theta))) \times \prod_{l=1}^{t-1} F(\varepsilon^*(\rho^l W, \theta)) \quad \text{iid case}$$

but now it can not be decomposed. Hence for its computation one has to compute a multiple integral of order t , which conventional methods of integration can not handle. We will show how simulation methods can handle such situations.

The idea of simulation methods, is to construct some simulated data from the model, then compare properties of these simulated data with properties of real data and try to make them as close as possible.

Simulated method of moments (SMM)

This method was developed by McFadden(1989), Lee and Ingram(1991), and Duffie and Singleton (1993). The idea is to compare a function (moment) of simulated data with a function of observed data.

Let $\{x(u_t, \theta_0)\}_{t=1}^T$ be a sequence of observed data. Here θ_0 represents the true parameter that we believe has generated the observed data and u_t are the shocks involved. Denote by $\mu(x_t)$ the moment from observed data. If we want to perform GMM, we should have the analytical form of moments as $\mu(x_t)$ However suppose we can not solve the model and extract the closed form of $\mu(x(u_t, \theta))$ in the general case but we can compute μ for a sample of draws u_t^s . Now what we do is to construct S copies of simulated data by drawing shocks u_t^s . Each of these S copies have length of T and denote them

by $\{x(u_t^s, \theta)\}, t = 1, \dots, T$ and $s = 1, \dots, S$, now one computes the moment μ for each copy of sample data and averages it for all the S copies. For every parameter vector θ one has to do this procedure (actually computer does it!). The SMM estimator is defined by

$$\widehat{\theta}_{S,T}(W) = \arg \min_{\theta} \left[\sum_{t=1}^T \left(\mu(x_t) - \frac{1}{S} \sum_{s=1}^S \mu(x(u_t^s, \theta)) \right) \right]' \times \\ W_T^{-1} \times \left[\sum_{t=1}^T \left(\mu(x_t) - \frac{1}{S} \sum_{s=1}^S \mu(x(u_t^s, \theta)) \right) \right]$$

the criteria is similar to that in definition of method of moments only that here we have included a simulated approximation of the model moments.

Example 10 Applying SMM to example 9. Suppose we have a data set of T cake eaters for which we observe the duration of their cake $D_t, t = 1, \dots, T$. Given a parameter θ that describes preferences and the process of ε , we can use value function iteration to compute V and hence using (1) we can compute the thresholds $\varepsilon^*(W)$. Next we can draw a series of ε shocks, say $\{\varepsilon_t^s\}_{t=1, \dots, T}$ from its distribution, Having the threshold $\varepsilon^*(W)$, we can determine that the shock process $\{\varepsilon_t^s\}$ implies that the cake eater t when will eat the cake. This gives rise to simulated data. We can repeat this step in order to construct S data sets each containing T simulated durations. To identify the parameters of the model, we can use mean and variance of durations. Both of these moments should be calculated from the simulated and observed data sets. If we want to identify more than two parameters, we can include more moments like the fraction of cakes eaten at the end of first, second, third... periods. If we include more moments than parameters then we can perform the overidentifying test described in previous section.

Properties of SMM

One can ask what is the relation of simulated estimator with the actual GMM estimator, whether they converge to the same number or not? Moreover what is the relation of variances of estimation?

The answer is that fortunately when S is fixed and $T \rightarrow \infty$ the SMM estimator converges to the true parameter of the model. i.e. SMM is consistent. Moreover when $S \rightarrow \infty$ the variance of the SMM estimator is the same as the variance of the GMM estimator.

For a proof of these claims and the formula for variance of SMM see Adda & Cooper page 88 or pages 58-64 of Dr. Donald's notes.

Simulated Maximum likelihood (SML)

This is not as commonly used as SMM and its requirement for consistency are harder to satisfy.

Suppose that we write the model as $x(u_t, \theta)$, where θ is a vector of parameters and u_t are the shocks involved in the model. The distribution of u_t implies a

distribution for $x(u_t, \theta)$; call it $\phi(x_t, \theta)$. This can be used to evaluate the likelihood of observing some particular sample in data, say x_t . In many cases, the exact distribution of $x(u_t, \theta)$ is not easily determined. In this case we can evaluate the likelihood using simulations. Let $\hat{\phi}(x_t, u, \theta)$ be an unbiased simulator of $\phi(x_t, \theta)$, i.e.

$$E_u \hat{\phi}(x_t, u, \theta) = \lim_s \frac{1}{S} \sum_{s=1}^S \hat{\phi}(x_t, u, \theta) = \phi(x_t, \theta)$$

the SML is defined as

$$\hat{\theta}_{S,T} = \arg \max_{\theta} \sum_{t=1}^T \log \left(\frac{1}{S} \sum_{s=1}^S \hat{\phi}(x_t, u_t^s, \theta) \right)$$

One way to construct an unbiased estimator of $\phi(x_t, \theta)$ is called important sampling. Remember that $\phi(x_t, \theta)$ is a probability and usually involves integrals that are possibly multiple. Conventional methods of computing integrals which require constructing a grid of the domain require a very long time to compute. However Monte Carlo methods provide a faster algorithm.

Suppose you want to compute the integral

$$\int \dots \int h(x) f(x) dx$$

where x is a random variable (can be a vector) with density f and suppose we are interested in $Eh(x)$. Notice that distribution of x can may depend on some unknown parameters. Instead of making grids on domain of x and computing the integral as the sum over the grid, we can draw random numbers from distribution x , say $\{x_s\}_{s=1}^S$ and approximate the integral with

$$\frac{1}{S} \sum_{s=1}^S h(x_s)$$

by law of large numbers we know this sum converges to the desired integral, i.e. $Eh(x)$.

This method is easy to implement, however if distribution of x depend on unknown parameters, can make the estimation more complicated.

A better method when f depends on unknown parameters is to use the method of important sampling. Choose a distribution function g with the same domain as x . Then write the integral as

$$\int \dots \int \frac{h(x) f(x)}{g(x)} g(x) dx$$

now draw random numbers from distribution g and estimate the integral with

$$\frac{1}{S} \sum_{s=1}^S \frac{h(x_s) f(x_s)}{g(x_s)}$$

this way, the random number generator does not depend on unknown parameters and we have more flexibility in choosing g , it can be normal distribution, etc. The more similar g to hf the smaller will be the variance of estimation. For more details see Dr Donald's notes , pages 53-56.

Now being able to find an unbiased estimator of $\phi(x_t, \theta)$ one can perform SML.

Example 11 Applying SML to example 9. The difficulty was computing the integral

$$P_t = P(\varepsilon_1 < \varepsilon^*(W), \varepsilon_2 < \varepsilon^*(\rho W), \varepsilon_3 < \varepsilon^*(\rho^2 W), \dots, \varepsilon_t > \varepsilon^*(\rho^t W))$$

using one of above methods this integral can be computed by simulations. Then one can write the maximum likelihood and perform the estimation.

Properties of SML

Unlike SMM, SML does not have good consistency properties. It is consistent only if S, T both go to infinity. If S is fixed, even if $T \rightarrow \infty$ the estimator will be inconsistent, this is contrary to the result for SMM.

For a proof of these claims and the formula for bias and variance of SML see Adda & Cooper page 92 or pages 58-64 of Dr. Donald's notes.