GMM and SMM

Some useful references:

1. Hansen, L. 1982. "Large Sample Properties of Generalized Method of Moments Estimators", *Econometrica,* 50, p. 1029-54.
2. Lee, B.S. and B. Ingram. 1991 "Simulation estimation of time series models", *Journal of Econometrics*, 47, p. 197-205.
3. Newey, W. & D. McFadden, Daniel. 1986. "Large sample estimation and hypothesis testing," Handbook of Econometrics, in: R. F. Engle & D. McFadden (ed.), Handbook of Econometrics, edition 1, volume 4, chapter 36, pages 2111-2245 Elsevier (see especially Sections 1, 2.2.3, and 2.2.4)
4. Cochrane, J. 2001. Asset Pricing, Princeton University Press.
5. Wiki on GMM: http://en.wikipedia.org/wiki/Generalized_method_of_moments

In general, one can think of calibration as a special case of Simulated Method of Moments with weighting matrix $W = I$ on a just identified model and SMM as a case of GMM. More generally, SMM with second stage optimal weighting matrix $W^* = S^{-1}$ so that the W^* downweights moments with lots of noise-to-signal and gives you standard errors (which can help inform you about identification).

1 OLS as Method of Moments

- The first model we usually see in econometrics is a linear one where the true model is assumed to be $y_t = \beta x_t + u_t$ with $E[x_t u_t] = 0$, $E[u_t] = 0$, and demeaned data.
- We try to estimate the parameter vector β that maps the model βx_t to the data of interest y_t.
- An implication of $E[x_t u_t] = 0$ is that

$$E[x_t (y_t - \beta x_t)] = 0. \tag{1}$$

- The sample analogue of the moment condition (1) is

$$\frac{1}{T}\sum_{t=1}^{T} x_t \left(y_t - \widehat{\beta}_T^{MM} x_t\right) = 0 \tag{2}$$

yielding

$$\widehat{\beta}_T^{MM} = \frac{\sum_{t=1}^{T} x_t y_t}{\sum_{t=1}^{T} x_t x_t}. \tag{3}$$

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

$$\widehat{\beta}_T^{OLS} = \arg\min_{\beta} \sum_{t=1}^{T} (y_t - \beta x_t)^2. \tag{4}$$

The first order condition is

$$-2 \sum_{t=1}^{T} (y_t - \widehat{\beta}_T^{OLS} x_t) x_t = 0. \tag{5}$$

But this is identical to the moment condition in (1) so the two methods yield the same "OLS" estimate.

- Further, notice that Generalized Least Squares is simply a more general moment condition than (1) given by

$$E[x_t (y_t - \beta x_t) / \sigma^2(x_t)] = 0. \tag{6}$$

That is, instead of weighting everything equally as in OLS (which is BLUE 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.

2 Generalized Method of Moments

- This section is based on Hansen (1982, Ecta) and Hansen and Singleton (1982, Ecta).

2.1 Example

- Consider Lucas' (1978) representative agent asset pricing model with preferences $U(c_t) = \frac{c_t^{1-\psi}-1}{1-\psi}$ (see Hansen and Singleton (1982)). The problem there is to solve

$$\begin{aligned} & \max_{\{c_t, s_{t+1}\}_{t=0}^{\infty}} E_0 \sum_{t=0}^{\infty} \beta^t U(c_t) \\ s.t.\ c_t + p_t s_{t+1} &= (y_t + p_t) s_t \end{aligned}$$

with market clearing conditions $c_t = y_t$ and $s_{t+1} = 1$.

- The first order necessary conditions are given by

$$\begin{aligned} p_t c_t^{-\psi} &= E_t \beta c_{t+1}^{-\psi} (p_{t+1} + y_{t+1}) \\ &\iff E_t \left[\beta \left(\frac{c_t}{c_{t+1}} \right)^{\psi} \left(\frac{p_{t+1} + y_{t+1}}{p_t} \right) - 1 \right] = 0 \end{aligned} \tag{7}$$

- We can rewrite (7) in terms of errors

$$u_{t+1}(x_{t+1}, b_0) \equiv \beta \left(\frac{c_t}{c_{t+1}} \right)^{\psi} \left(\frac{p_{t+1} + y_{t+1}}{p_t} \right) - 1$$

  where b_0 stands in for the true parameters (β, ψ), x_{t+1} is a $k \times 1$ vector of variables observed by agents (and the econometrician) as of $t+1$ (e.g. $\{c_n, y_n, p_n\}_{n=0}^{t+1}$) and we assume that $u_{t+1}(x_{t+1}, b_0)$ has finite second moments (this is necessary for stationarity).

- Then we are interested in estimating the parameter vector b_0 from the moments

$$E_t \left[u_{t+1}(x_{t+1}, b_0) \right] = 0.$$

2.2 Estimation

- In general, suppose there are n necessary conditions of the model:

$$E_t \left[u_{t+1}(x_{t+1}, b_0) \right] = 0 \tag{8}$$

  where

  - u_{t+1} is an $n \times 1$ vector of "errors"[1]
  - x_{t+1} is a $k \times 1$ vector of data
  - b is an $\ell \times 1$ vector of parameters where b_0 stands for the true parameter vector

- Recall the following order conditions necessary (but not sufficient) for identification:

  - If $\ell < n$, the model is said to be overidentified.
  - If $\ell = n$, the model is said to be just identified.
  - If $\ell > n$, the model is said to be underidentified.

- In the asset pricing case above, we have $\ell = 2$ and $n = 1$, so it seems like we should go back to the drawing board. If there had been a labor/leisure choice, then that foc would provide another moment condition (i.e. $n = 2$), but probably would have introduced another parameter.

- One way to deal with the case where $\ell > n$ is to add more "equations". If z_t is a $q \times 1$ vector of variables in the econometrician's (and agent's) information set, then from (8) and the law of iterated expectations we know

$$E_t \left[u_{t+1}(x_{t+1}, b_0) \otimes z_t \right] = 0 \otimes z_t = 0 \Longrightarrow E \left[u_{t+1}(x_{t+1}, b_0) \otimes z_t \right] = 0 \tag{9}$$

[1] In the SMM context, the "errors" will be the difference between model moments and data moments.

is an $nq \times 1$ vector, which may satisfy the identification order condition $\ell \leq nq$ where $\otimes$ is the Kroenecker product. Loosely speaking, one way to interpret the z_t is as a vector of instrumental variables.

- For example, in the Hansen and Singleton (1982) paper, they include past consumption growth in z_t (i.e. $z_t = [1 \ c_t/c_{t-1}]'$). This is similar to using a lagged dependent variable as an instrument provided the true errors are not autocorrelated.

- Letting $f(x_{t+1}, z_t, b) \equiv u_{t+1}(x_{t+1}, b) \otimes z_t$, define the $nq \times 1$ moment vector

$$g(b) \equiv E[f(x_{t+1}, z_t, b)] \tag{10}$$

(i.e. the unconditional average error). By (9), $g(b_0) = 0$. This is the analogue of the OLS condition (1).

- The sample analogue of (10) is the $nq \times 1$ vector

$$g_T(b) \equiv \frac{1}{T} \sum_{t=1}^{T} f(x_{t+1}, z_t, b) \tag{11}$$

The basic idea of GMM is that as $T \to \infty$, (9) implies $g_T(b_0) = 0$.This is the analogue of the OLS condition (2).

- Assuming that $g_T(b)$ is continuous in b, the GMM estimate of b solves

$$b_T = \arg\min_b J_T(b) \tag{12}$$

where $J_T(b) \equiv g_T'(b) W_T g_T(b)$ (which is $(1 \times nq)(nq \times nq)(nq \times 1)$) is a weighted sum of squared errors and W_T is an arbitrary weighting $(nq) \times (nq)$ matrix that can depend on the data. This is the analogue of the OLS condition (4).

- In the just identified case, the weighting matrix does not matter. To see this, consider the following $\ell = n = 2$ case:

$$\min_{b_1, b_2} w_1 g_1(b_1, b_2)^2 + w_2 g_2(b_1, b_2)^2$$

The foc are:

$$\begin{aligned} b_1 &: \ 2w_1 g_1(b) \nabla_{b_1} g_1 + 2w_2 g_2(b) \nabla_{b_1} g_2 = 0 \\ b_2 &: \ 2w_1 g_1(b) \nabla_{b_2} g_1 + 2w_2 g_2(b) \nabla_{b_2} g_2 = 0 \end{aligned}$$

Since there are 2 equations in 2 unknowns, one would think that $b(W)$. However, in this case b is not a function of W. Rewriting the 2 foc in matrix notation

$$[g_1(b) \ g_2(b)] \begin{bmatrix} w_1 \nabla_{b_1} g_1 & w_1 \nabla_{b_2} g_1 \\ w_2 \nabla_{b_1} g_2 & w_2 \nabla_{b_2} g_2 \end{bmatrix} = [0 \ 0]$$

Then provided the 2×2 matrix is invertible, we have

$$[g_1(b)g_2(b)] = [0\ 0] \begin{bmatrix} w_1\nabla_{b_1}g_1 & w_1\nabla_{b_2}g_1 \\ w_2\nabla_{b_1}g_2 & w_2\nabla_{b_2}g_2 \end{bmatrix}^{-1} = [0\ 0].$$

2.3 Consistency

- Under certain conditions, Hansen 1982 (Theorem 2.1) proves that this estimator b_T exists and converges in probability to b_0.
- It is essential for consistency that the limit $J_\infty(b)$ have a unique maximum at the true parameter value b_0.This condition is related to identification; the distribution of the data at b_0 is different than that at any other possible parameter value.
- The conditions are:
  - $W_T \to W$ in probability, where W is a positive semi-definite matrix
  - $g(b) = 0$ (an $nq \times 1$ vector) only for $b = b_0$.
  - $b_0 \in B$ (a compact set)
  - $f(x, z, b)$ is continuous at each b
  - $E[\sup_b \|f(x, z, b)\|] < \infty$.
- The second condition (known as **Global Identification**) is hard to verify. A simpler necessary but not sufficient condition is known as **Local Identification**. If $g(b)$ is continuously differentiable in a neighborhood of b_0, then the Jacobian matrix $\nabla_b g(b)$ (which is $(nq \times \ell)$) must have full column rank (i.e. there are ℓ linearly independent columns).
- Hansen 1982 (Theorem 3.1) establishes asymptotic normality of the estimator.

2.4 Efficiency

- While the above result shows that the GMM estimator is consistent for arbitrary weighting matrices (e.g. $W = I$), it is not necessarily efficient. Hansen (1982, Theorem 3.2) shows that the statistically optimal weighting matrix $W^* = S^{-1}$ where the asymptotic variance covariance matrix is:

$$S = \sum_{j=-\infty}^{\infty} E\left[f(x_t, z_{t-1}, b_0) f(x_{t-j}, z_{t-j-1}, b_0)'\right] \tag{13}$$

- Why does this weighting matrix make sense? Some moments will have more variance than others. This downweights errors from high variance moments (i.e. those with low signal to noise). See page 194 of Cochrane (2001).

- Hansen (1982, Theorem 3.2) shows that the asymptotic distribution of the estimator when $W^* = S^{-1}$ is given by

$$\sqrt{T}(b_T - b_0) \to N(0, \left[\nabla_b g(b_0)' S^{-1} \nabla_b g(b_0)\right]^{-1}) \tag{14}$$

  where $\nabla_b g(b_0)' S^{-1} \nabla_b g(b_0)$ is an $(\ell \times nq)(nq \times nq)(nq \times \ell)$ matrix.

- The problem is that we do not know S^{-1} nor $g(b)$. If the errors are serially uncorrelated, then a consistent estimate of the asymptotic var-covar matrix S is given by

$$\widehat{S}_T = \frac{1}{T} \sum_{t=1}^{T} f(x_{t+1}, z_t, \widehat{b}_T) f(x_{t+1}, z_t, \widehat{b}_T)'$$

  where b_T is a consistent estimate of b_0.[2] In this case, the distribution of the estimator is given by

$$\sqrt{T}(b_T - b_0) \to N(0, \left[\nabla_b g_T(b_T)' S_T^{-1} \nabla_b g_T(b_T)\right]^{-1}).$$

- Notice that the precision of the estimates is related to $\nabla_b g_T(b_T)$. If the objective is very sensitive to changes in the parameters (i.e. $\nabla_b g_T(b_T)$ is high), then there will be a low variance of the estimate (since $\nabla_b g_T(b_T)' S_T^{-1} \nabla_b g_T(b_T)$ is inverted). If the objective is not very sensitive to changes in the parameters (i.e. $\nabla_b g_T(b_T)$ is low), it will produce a high variance for the estimates. See Figure SMMsurface. Simply put, this suggests that if you find big standard errors, it is 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** in the previous subsection. If $\nabla_b g_T(b_T) = 0$ for many values of the parameter space, then the Jacobian matrix will not have full column rank and the parameters of the model are not well identified.

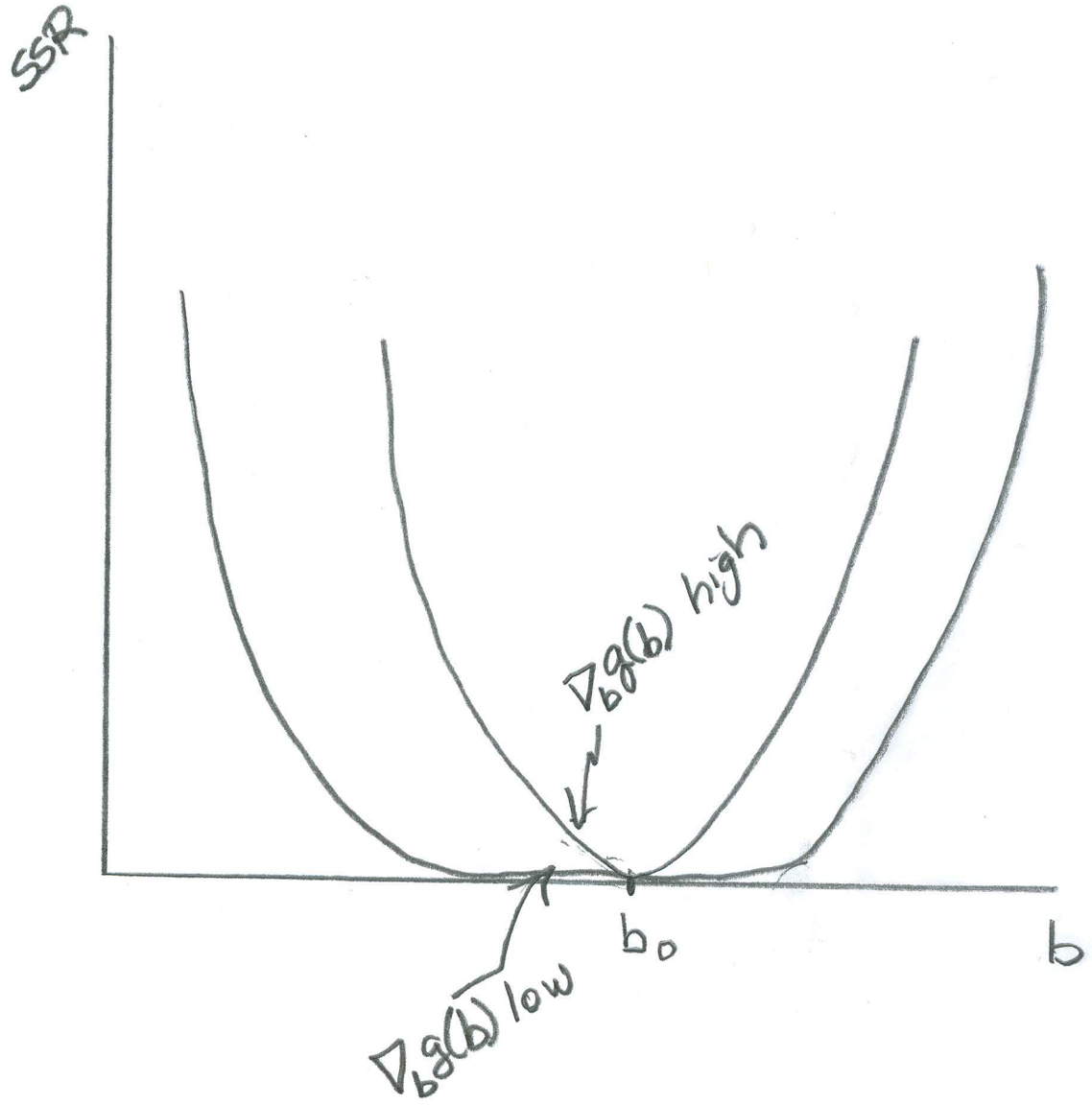
- To implement this, use a two step procedure: (i) the first stage estimate of b minimizes a quadratic form of the sample mean of errors for $W = I$, which is consistent; (ii) estimate a var-covar matrix of the residuals S_T from the first stage to form $W_T = S_T^{-1}$ in a second stage minimization of $g_T'(b) S_T^{-1} g_T(b)$. This is like the two step procedure in Generalized Least Squares.

- Testing Overidentifying Restrictions: Hansen (1982, Lemma 4.2) shows

$$T g_T'(b_T) S_T^{-1} g_T(b_T) \to \chi^2(nq - \ell). \tag{15}$$

[2] For the case in which $f(x_{t+1}, z_t, b_0)$ is serially correlated you can use the Newey-West (1987) correction.

Figure SMM Surface



That is, the minimized value of the objective function is distributed as a chi-squared r.v. with degrees of freedom equal to the #moments-#parameters. This is known as the J-test. If the restrictions are not too "restrictive", the objective (i.e. J_T in (12)) should not be statistically different from zero.

- You often want to compare one model to another. If one model can be expressed as a special or "restricted" case of the other "unrestricted" model, we can conduct something like a likelihood ratio test. If we use the same S matrix (usually that of the unrestricted model), then J_T(restricted) must rise. If the restricted model is really true, it should not rise too much. Thus

$$TJ_T(\text{restricted}) - TJ_T(\text{unrestricted}) \sim \chi^2(\#\text{ of restrictions})$$

This is Newey-West's (1987, IER) D-test.

2.5 Using Prespecified Weighting Matrices

- 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.

- For example, if $g_T = [g_T^1\ g_T^2]'$, $W = I$, but $\frac{\partial g_T}{\partial b} = [1\ 10]$, so that the second moment is 10 times more sensitive to the parameter value than the first moment, then GMM with fixed weighting matrix sets $1 \times g_T^1 + 10 \times g_T^2 = 0$. The second moment condition will be 10 times closer to zero than the first. If you really want GMM to pay equal attention, then you can fix the d matrix directly.

- Using a prespecified weighting matrix does not mean that you are ignoring correlation of the errors in the distribution theory. The S matrix will show up in all the standard errors and test statistics.

- For example,

$$\sqrt{T}(b_T - b_0) \to N(0, [d'Wd]^{-1}\, d'WSWd\, [d'Wd]^{-1})$$

which reduces to (14) if $W = S^{-1}$. The same goes for the χ^2 overidentifying tests in (15).

3 Simulated Method of Moments

- This section is based on Lee and Ingram (1991, Journal of Econometrics). One way to think about SMM is that it is the statistical approach to do calibration.

- Let $\{x_t\}_{t=1}^T$ be a realization of an $k \times 1$ vector valued stationary and ergodic stochastic process generating the observed data (e.g. detrended GDP).

- Let $\{y_t(b)\}_{t=1}^T$ 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 b is an $\ell \times 1$ vector of parameters. In general we may take H simulations of length T.

- Let $M_T(x)$ be an $n \times 1$ vector of data moments (e.g. standard deviation of detrended GDP) and $M_N(y(b))$ be a $n \times 1$ vector of model moments of the simulated data where $N = H \cdot T$.

- Assume that $M_T(x) \overset{a.s.}{\to} \mu(x)$ as $T \to \infty$ and that $M_N(y(b)) \overset{a.s.}{\to} \mu(y(b))$ as $N \to \infty$ where $\mu(x)$ and $\mu(y(b))$ are the population moments.

- Furthermore, under the null that the model is correct at the true parameter vector b_0, then $\mu(x) = \mu(y(b_0))$. If you understand this equality you understand everything you need to know about economics. It says there is a link between data and theory.

- In summary, x_t is observed data (which we may not even have), y_t is simulated data, $M_T(x)$ is observed moments (which we will assume we have), $M_N(y(b))$ is simulated moments, and the reason we can use the model to say something about the data we don't have is that if the model is true, then the asymptotic moments have to be equal at the true parameter values (i.e. $\mu(x) = \mu(y(b_0))$).

- Given a symmetric $n \times n$ weighting matrix W_T (which may depend on the data - hence the subscript T), Lee and Ingram show that under certain conditions the simulation estimator $\widehat{b}_{TN}$ which minimizes the weighted sum of squared errors of the model moments from the data moments - ie. the solution to

$$\widehat{b}_{TN} = \arg\min_b [M_T(x) - M_N(y(b))]' W_T [M_T(x) - M_N(y(b))] \tag{16}$$

  - is a consistent and asymptotically normal estimator of b_0 (i.e. $\lim_{T,N\to\infty} \Pr ob(|\widehat{b}_{TN} - b_0| < \varepsilon) = 1$).

- Basically, SMM is just GMM where the errors are just the difference between the data moment and the model moment $g_{TN} = M_T - M_N(y(b))$, i.e. the difference between the data moment and the model moment.

- Since the solution to this problem is essentially a special case of Hansen's (1982) GMM estimator, the conditions are from his paper: (i) x and $y(b)$ are independent; (ii) the model must be identified; and (iii) $M_N(y(b))$ must be continuous in the mean.

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

1. For any given value of b,say b^i,

   (a) simulate artificial data from the model (in the context of the growth model, this would be H draws of $\{\varepsilon_t\}_{t=1}^T$ which implies a realization of technology shocks and then via decision rules (which depend on parameters b^i) a realization of a variable of interest like real output $y(b^i)$), and

   (b) compute a moment (i.e. $M_N(y(b^i))$), and evaluate the objective function $J_{TN}(b^i) = [M_T(x) - M_N(y(b^i))]'W[M_T(x) - M_N(y(b^i))]$; and

2. choose a new value for the parameters, say b^{i+1}, for which $J_{TN}(b^{i+1}) \leq J_{TN}(b^i)$.

- A minimization routine constructs this sequence of smaller and smaller $J_{TN}(b^i)$ for you. **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.

- As before, to obtain the optimal weighting matrix, you use a two stage procedure. See Section 4.2.3 on Asymptotic Properties of Indirect Inference Estimators in Gourieroux and Monfort (1996, Simulation Based Econometric Methods, Oxford University Press) for justification of this process.

  - In the first stage, minimize $J_{TN}^1(b)$ constructed using $W = I$.
  - Now the second stage is different since above, you had a vector of residuals to construct the variance-covariance matrix. Here, you don't have enough "data" to get a var-covar matrix.
  - Since the resulting estimate $\widehat{b}_{TN}^1$ of b_0 is consistent, generate H repetitions of model moments (from T length simulated samples) analogous to the data moments in order to construct an estimate of the var-covar matrix $\widehat{S}_T$ of the "data moments".
  - Use $W_T = \widehat{S}_T^{-1}$ to construct the second stage J_{TN}^2 and obtain the corrected estimate $\widehat{b}_{TN}^2$.

- Once you have the optimal weighting matrix, you can generate standard errors as in (14) of Hansen (1982).

- Alternatively, you can run a Monte Carlo experiment to compute standard errors of the estimates. Recall that each estimate of $\widehat{b}_{TN}$ is derived for a given HT draw of the shocks ε_t to the underlying data generation process. Different draws will generate different estimates of $\widehat{b}_{TN}$. You can generate a histogram and summary statistics (mean and sd of $\widehat{b}_{TN}$) which are interpretable as the point estimate and standard error of b.

4 Simulated Annealing

- In general there is nothing to say that the objective function $J_{TN}(b)$ (where $b \in \mathbb{R}^k$ is the parameter vector) is nicely behaved (i.e. doesn't have local minima).
- In that case, the minimization routine (e.g. fminsearch in matlab), may stop before reaching the true global minimum or get stuck at a local minimum. One way to deal with that is to start fminsearch at many different initial conditions and see if it gives back the same $\widehat{b}$.
- A more systematic approach is to use simulated annealing, which is a technique used to find global maxima when the shape of the objective function is not known and the global minimum is possibly hidden among many local extreme points.
- Each step of the simulated annealing algorithm replaces the current solution by a random "nearby" solution, chosen with a probability that depends on the difference between the corresponding function values and on a global parameter Υ (called the temperature) that is gradually decreased during the process.
- Notation:
  - Let Υ^i be the temperature value at iteration i.
  - Denote b^{i*} as the best point after i temperature iterations.
  - Let $h < H$ be the number of function evaluations at a given temperature (when $h = H$ we will reduce the temperature).
  - Let b^i be the best point within a temperature Υ^i.
  - Let ϵ be the smallest number your computer takes. We use ϵ as the tolerance parameter to terminate the procedure.

  Steps:

  1. Set $i = 1$ and $\Upsilon^{i=1}$ to some positive value. Make an initial guess for parameter values b^1. Set $b^{1*} = b^1$.
  2. Set $h = 0$ and evaluate $J_{TN}(b)$ at b^i.
  3. For each h we will try random changes to the parameters in one dimension at a time. For $n \leq N$, let $\widehat{b}_n = b^i_n + (2u-1)$ and $\widehat{b}_{-n} = b^i_{-n}$, where u is distributed uniform $[0, 1]$. Evaluate $J_{TN}(b)$ at $\widehat{b}$.
     - If $J_{TN}(\widehat{b}) < J_{TN}(b^i)$, record $\widehat{b}$ as your best point within temperature i, i.e. $b^i = \widehat{b}$. Moreover, if $J_{TN}(b^i) > J_{TN}(b^{i*})$ set $b^{i*} = b^i$.

- If $J_{TN}(\widehat{b}) > J_{TN}(b^i)$ then keep $\widehat{b}$ as the best point within a temperature only with some probability that depends on the temperature and on the distance between $J_{TN}(b^i)$ and $J_{TN}(\widehat{b})$. This step is intended to avoid local minima (we can accept a point that is worse than the optimal one so far). Set $p = \exp[(J_{TN}(\widehat{b}) - J_{TN}(b^i))/\Upsilon^i]$. Let v denote a random variable distributed uniform $[0, 1]$. If $v < p$ accept $\widehat{b}$ as your best point, i.e. $b^i = \widehat{b}$. Otherwise, keep b^i as your best point.

4. After you finish looping over n, if $h < H$, set $h = h + 1$ and return to step 3.
5. If $h = H$, we check the tolerance. If $|J_{TN}(b^{i*}) - J_{TN}(b^{i-1*})| < \epsilon$ you are done.
6. If $|J_{TN}(b^{i*}) - J_{TN}(b^{i-1*})| > \epsilon$, set $\Upsilon^{i+1} = \phi\Upsilon^i$ where $\phi \in [0, 1]$. Moreover set $b^{i+1} = b^i$, $i = i+1$ and return to step 3. The parameter ϕ is set to 0.85 in the examples provided below.
7. If $\Upsilon^i = 0$, and you never found that $|J_{TN}(b^{i*}) - J_{TN}(b^{i-1*})| < \epsilon$, restart the procedure at a different initial guess and possible at a higher initial temperature.

- On my website are 3 programs: crazy.m (just an example objective function with lots of local minima); trycrazy.m (the main code which calls the annealing algorithm); and simulan.m (a simulated annealing algorithm written by E.G. Tsionas)

5 Summary

Errors: $f(x_{t+1}, z_t, b)$, an nq vector of "errrors" (e.g. from OLS $f(x_{t+1}, z_t, b) = x_{t+1} - b \cdot x_t$).

True Model: $E[f(x_{t+1}, z_t, b_0)] = 0$ for each element of the vector.

Sample analogue: $\frac{1}{T}\sum_{t=1}^{T} f(x_{t+1}, z_t, b_T) \approx 0$.

Estimator: $b_T = \arg\min_b \left(\frac{1}{T}\sum_{t=1}^{T} f(x_{t+1}, z_t, b)\right)' W_T \left(\frac{1}{T}\sum_{t=1}^{T} f(x_{t+1}, z_t, b)\right)$.

Consistency: Theorem 2.1 or 2.2 of Hansen, $b_T \to b_0$ as $T \to \infty$.

Efficiency: Theorem 3.2 of Hansen provided $W_T = \widehat{S}_T^{-1}$ where $\widehat{S}_T = \frac{1}{T}\sum_{t=1}^{T} f(x_{t+1}, z_t, b_T) f(x_{t+1}, z_t, b_T)'$.

SMM: Let $f(x_{t+1}, z_t, b) = M_T(x) - M_N(y(b))$, i.e. the difference between the data moment and the model moment.

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.