Estimating Deep Parameters: GMM and SMM

1 Parameterizing a Model

- Calibration
  - Choose parameters from micro or other related macro studies (e.g. coefficient of relative risk aversion is 2).
  - SMM with weighting matrix $W = I$ (no standard errors, but yields consistent estimates of the true parameters under null that the model is correct).
  - Should add a robustness section where you perturb parameter values and see what happens to model.
- Estimation
  - GMM (have lots of data)
  - SMM with optimal weighting matrix $W^* = S^{-1}$ so that the W^* downweights moments with lots of noise-to-signal (gives you standard errors). This would be ideal.

2 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

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

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

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

yielding

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

- 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

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

The first order condition is

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

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.

3 Generalized Method of Moments

- This section is based on Hansen (1982, Ecta) and Hansen and Singleton (1982, Ecta).[1]
- 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

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

[1] Another good reference is http://en.wikipedia.org/wiki/Generalized_method_of_moments

- 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). In general, $u_{t+1}(x_{t+1}, b_0)$ may be an $m \times 1$ vector and b_0 may be an $\ell \times 1$ vector.

- We are interested in estimating the parameter vector b from the moments $E_t \left[u_{t+1}(x_{t+1}, b) \right] = 0$. Before proceeding, recall the following facts about identification:

  - If $\ell < m$, the model is said to be overidentified since there are more moments than parameters to estimate. In this case, which will typically be true for RBC models, we can test the model's predictions using auxiliary moments.
  - If $\ell = m$, the model is said to be just identified.
  - If $\ell > m$, the model is said to be underidentified, in which case there's not enough information to be able to estimate the parameters (go back to the drawing board).

- In the asset pricing case above, we have $\ell = 2$ and $m = 1$, 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. $m = 2$), but probably would have introduced another parameter.

- One way to deal with this underidentification issue is to let z_t denote a $q \times 1$ vector of variables in the agent's (and econometrician's) information set at time t with finite second moments. Then, since

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

  where $\otimes$ is the Kroenecker product. You can interpret (8) as an orthogonality condition. The implication follows since z_t is in the econometrician's information set and by the law of iterated expectations. 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. $z_t = [1 \; c_t/c_{t-1}]'$.

- Letting $f(x_{t+1}, z_t, b) \equiv u_{t+1}(x_{t+1}, b) \otimes z_t$ (an (mq) vector), we can define the moment $g(b) = E \left[f(x_{t+1}, z_t, b) \right]$.By (8), $g(b_0) = 0$. That is, at the true parameter vector the orthogonality condition is satisfied and now we have some more moments (i.e. mq rather than m).

- The basic idea of GMM is to replace the theoretical expected value $g(b)$ with its empirical analogue, the sample mean. Specifically, we have available sample information $\{(x_1, z_0), ..., (x_T, z_{T-1})\}$ and parameters b. If the model underlying (8) is true, then the sample analogue of $g(b)$,

$$g_T(b) = \frac{1}{T}\sum_{t=1}^{T} f(x_{t+1}, z_t, b) \equiv E_T\left[f(x_{t+1}, z_t, b)\right], \tag{9}$$

  should be close to zero evaluated at $b = b_0$ for large values of T :

$$g_T(b_0) = 0. \tag{10}$$

- Given this fact, it is reasonable to estimate b_0 by selecting b_T from a parameter space that makes g_T in (9) "close" to zero (this is the analogue of (1)). Alternatively (as in the analogue of (4)), assuming that f is continuous with respect to the parameter vector, GMM amounts to choosing b_T as

$$b_T = \arg\min_b J_T(b) \tag{11}$$

  where $J_T(b) \equiv g_T'(b) W_T g_T(b)$ and W_T is an arbitrary weighting $(mq) \times (mq)$ matrix that may depend on the data.

- Hansen 1982 (Theorem 2.1 or Theorem 2.2 depending on the assumptions one wants to make) proves that this estimator b_T exists and converges almost surely to b_0.

- Hansen 1982 (Theorem 3.1) also establishes asymptotic normality of the estimator.

- 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^* is given by the inverse of the asymptotic variance covariance matrix of the errors:

$$S = \sum_{j=-\infty}^{\infty} E\left[f(x_t, z_{t-1}, b_0) f(x_{t-j}, z_{t-j-1}, b_0)'\right]. \tag{12}$$

  S is also known as the spectral density matrix at frequency zero.[2]

- If the errors are serially uncorrelated, then a consistent estimate of S is given by:

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

[2] Optimality here means maximizing the asymptotic statistical information in the sample about a model, given the choice of moments g_T. It does not necessarily mean achieving the lowest J_T.

where $\widehat{b}_T$ is any consistent estimate of b_0.[3]

- Why does this type of weighting function make sense? Some sample moments will have much more variance than others. Hence it would seem like a good idea to pay less attention to errors from those moments (the inverse downweights those high variance moments).

- 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 $\widehat{S}_T$ from the first stage to weight $W_T = \widehat{S}_T^{-1}$ a second stage minimization of $g_T'(b)\widehat{S}_T^{-1}g_T(b)$.

- A two step procedure for estimating a weighting matrix should be familiar from your basic econometrics course (e.g. Generalized Least Squares).

- Standard Errors of the Estimate: Hansen (1982, Theorem 3.2) shows that if you use the optimal weighting matrix $W^* = S^{-1}$, then the efficient estimator of b

$$\sqrt{T}(b_T - b_0) \to N(0, \left[d'S^{-1}d\right]^{-1}) \tag{13}$$

  where $d \equiv \frac{\partial g_T(b_0)}{\partial b_0}$ measures how the objective changes with changes in the parameter value.

- Notice that the precision of the estimates, measured through the asymptotic variance above, is related to the sensitivity of the auxiliary parameters to movements in the structural parameters through d. If the sensitivity is low, the derivative will be near zero, which will produce a high variance for the estimates.

- Testing Overidentifying Restrictions: Hansen (1982, Lemma 4.2) shows that if you use the optimal weighting matrix $W_T = \widehat{S}_T^{-1}$ where $\widehat{S}_T$ is an estimate a var-covar matrix of the residuals of a first stage optimization with $W = I$, then

$$Tg_T'(\widehat{b})\widehat{S}_T^{-1}g_T(\widehat{b}) \to \chi^2(mq - \ell). \tag{14}$$

  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.

[3] For the case in which $f(x_{t+1}, z_t, b_0)$ is serially correlated, a consistent estimator for S is given by

$$\widehat{S}_T = \widehat{\Gamma}_0 + \sum_{v=1}^{s}\left(1 - \frac{v}{s+1}\right)(\widehat{\Gamma}_v + \widehat{\Gamma}_v')$$

where estimates of the v^{th}-order covariance matrices, $v = 0, \ldots, s$ are given by

$$\widehat{\Gamma}_v = \frac{1}{T}\sum_{t=v+1}^{T} f(x_{t+1}, z_t, \widehat{b}_T)f(x_{t+1-v}, z_{t-v}, \widehat{b}_T)'.$$

This estimator is due to Newey-West (1987).

- 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.

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

4 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 b be a $\ell \times 1$ vector of parameters.

- 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 $\{y_n(b)\}_{n=1}^N$ 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).

- Let $M_T(x)$ be a $m \times 1$ vector of data moments (e.g. standard deviation of detrended GDP) and $M_N(y(b))$ be a $m \times 1$ vector of model moments of the simulated data.

- 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_n 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 models have to be equal at the true parameter values (i.e. $\mu(x) = \mu(y(b_0))$).

- Given a symmetric $m \times m$ 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

$$\widehat{b}_{TN} = \arg\min_b [M_T(x) - M_N(y(b))]' W_T [M_T(x) - M_N(y(b))] \qquad (15)$$

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

- 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 a draw of ε_N 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 $J^1_{TN}(b)$ 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 $\widehat{b}^1_{TN}$ of b_0 is consistent, generate N 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^2_{TN} and obtain the corrected estimate $\widehat{b}^2_{TN}$.

- Once you have the optimal weighting matrix, you can generate standard errors as in (13) 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 NT 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.

4.1 An Example

- The parameters of the exogenous driving process for technology shocks Z_t in an RBC model are often estimated from the residuals of a simple regression of $\ln y_t$ on $\ln n_t$ and perhaps $\ln k_t$ (if that data is available).

- An alternative is to estimate the parameters using SMM. The simulated moments will come from the decision rules you compute via the method of undetermined coefficients for the following simple RBC model. That is, derive the linear decision rules for the planner's problem:

$$\max E\left[\sum_{t=0}^{\infty}\beta^t \ln(C_t)\right]$$

s.t.

$$C_t + K_{t+1} = Z_t K_t^{\theta} + (1-\delta)K_t$$
$$Z_t = (1-\rho_{\varepsilon}) + \rho_{\varepsilon} Z_{t-1} + \varepsilon_t$$

where ε_t is an iid random shock drawn from $N(0, \sigma_{\varepsilon}^2)$. In order to keep the estimation exercise simple, we will assume that a subset of the deep structural parameters are set to their calibrated values: $\theta = 0.36, \delta = 0.015$, and $\beta = 0.9921$.

- The exercise is to match the autocorrelation and standard deviation of real output data (i.e. ρ_y and σ_y in an AR1 of Hodrick-Prescott filtered output data) to the same moments simulated from the model in order to determine the "unobservable" ρ_{ε} and σ_{ε} parameters of the technology shock process $\{Z_t\}$.

- Call the data moment vector $M_T^d = (\rho_y, \sigma_y) = (0.8426, 0.0190)$. Call the parameter vector $b = (\rho_{\varepsilon}, \sigma_{\varepsilon})$. Call the simulated model moments $M_{TN}^m(b) = (\rho_y(b), \sigma_y(b))$.

- The idea of SMM is contained in the next three steps, which can be thought of as matlab functions. The program on my website which runs all these functions is called "SMMexample.m".

- Before anything, use a random number generator to simulate N repetitions of T period samples for innovations (call them e_t^n) from a uniform distribution $U(0,1)$.

1. Derive decision rules and generate moments. This is the function "smmgenerate.m" on my website.

   (a) For a given parameter vector b, calculate the saving decision rule $K_{t+1} = g(K_t, Z_t; b)$ as a function of the state variables. In this particular case, we will derive linearized decision rules $k_{t+1} = x_k(b)k_t + x_z(b)z_t$ using the method of undetermined coefficients as in Christiano (2001). It can be shown that the coefficients x_k and x_z solve the following set of equations:

$$x_z\left[x_k - (1-\rho_{\varepsilon}) - \frac{1}{\beta}\right] + \frac{\bar{Y}}{\bar{K}}(1-\rho_{\varepsilon}) + \frac{\bar{C}}{\bar{K}}(1-\beta(1-\delta))(\rho_{\varepsilon} - (1-\theta)x_z) = 0 \tag{16}$$

and

$$(1-x_k)\left[\frac{1}{\beta}-x_k\right]-\frac{\bar{C}}{\bar{K}}(1-\beta(1-\delta))(1-\theta)x_k=0. \qquad (17)$$

Note that since (17) there is no parameter to be estimated in the second equation and the only coefficient to be calculated is x_k, we can compute it outside of the estimation procedure. The other coefficient, x_z should be calculated for each set of parameter estimates.

(b) Using the above draw $\{e_t^n\}_{t=1,n=1}^{T,N}$ from $U(0,1)$, use the parameter vector b to draw $\{\varepsilon_t^n\}_{t=1,n=1}^{T,N}$ from $N(0,\sigma^2)$ (we need to do this since the parameter vector b actually includes σ^2). Given the law of motion for technology shocks and the decision rules from step (1a), use the simulated $z_t^n(b)$ process to simulate N sample paths of output. Analogous to the data moments, construct the model moments $M_{TN}^n(b)=(\sigma_y(b),\rho_y(b))$ by taking the average over the N repetitions of σ_y and ρ_y each computed from the length T series.

2. Choose b^1 in such a way as to bring the simulated moments $M_{TN}^m(b)$ constructed in step (1b) as close as possible to the data moments for a weighting matrix $W_T=I$. That is, we solve

$$\widehat{b}_{TN}^1=\arg\min_b[M_T^d-M_{TN}^m(b)]'I[M_T^d-M_{TN}^m(b)]. \qquad (18)$$

This will yield consistent but inefficient estimates of the true parameter vector. Remember, since this step embeds step 1, **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. The objective function in (18) is called "smmJ.m" which is passed the identity weighting matrix and fminsearch is applied to this function.

3. Since the resulting estimate $\widehat{b}_{TN}^1$ of b_0 is consistent, generate N repetitions of model moments from T length simulated data. Instead of averaging the moments as in step 1b, construct an estimate of the var-covar matrix $\widehat{S}_T$ from the N repetitions. The function "smmgenerate.m" can be used to calculate $\widehat{S}_T$. Then use $W_T=\widehat{S}_T^{-1}$ and solve

$$\widehat{b}_{TN}^2=\arg\min_b[M_T^d-M_{TN}^m(b)]'\widehat{S}_T^{-1}[M_T^d-M_{TN}^m(b)].$$

This is achieved by passing the weighting matrix $\widehat{S}_T^{-1}$ to the function "smmJ.m".

4. Construct standard errors. As in (13), simply compute the numerical derivative $\Delta M_{TN}^m(b)/\Delta b$ and form

$$\left(1+\frac{1}{N}\right)\left[\left(\frac{\Delta M_{TN}^m(b)}{\Delta b}\right)'\left(\widehat{S}_T^{-1}\right)^{-1}\left(\frac{\Delta M_{TN}^m(b)}{\Delta b}\right)\right]^{-1}.$$

This is achieved by the function "smmstderror.m".

- Since there is sampling error in $M_{TN}^m(b)$ (it is only the case that $M_{TN}^m(b) \overset{a.s.}{\to} \mu^m(b)$ as $TN \to \infty$ where $\mu^m(b)$ is the population model moment vector), different random draws will generate different estimates of the underlying parameters. We can get the standard errors of the parameter estimates either by using the optimal weighting matrix as in step 4 or generating them through a Monte Carlo. To conduct a Monte Carlo, run a large number of step 3 saving each estimate of $\widehat{\rho}_\varepsilon$ and $\widehat{\sigma}_\varepsilon^2$. Then generate a histogram and summary statistics (mean and sd of $\widehat{\rho}_\varepsilon$ and $\widehat{\sigma}_\varepsilon^2$) which are interpretable as the point estimates of $\widehat{\rho}_\varepsilon$ and $\widehat{\sigma}_\varepsilon^2$ and standard errors of $\widehat{\rho}_\varepsilon$ and $\widehat{\sigma}_\varepsilon^2$, respectively. This is achieved by the function "smmstderrorMC.m".

5 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).

6 Summary

Errors: $f(x_{t+1}, z_t, b)$, an mq 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.