The Empirical Validation of Agent Based Models

Lecture #3

Domenico Delli Gatti

Outline

- What is validation?
- Data Generating Processes
  - The *real-Data Generating Process (rDGP)*
  - The *model-Data Generating Process (mDGP)*
- Internal validity
- Sensitivity analysis
- Calibration
- Stylized facts

Validation

- The main modeling problem in ABMs is the *choice of parameter values, initial conditions and stochastic disturbances*
- Such a choice is related to the empirical *validation* of the model, that is the capability of the model to reproduce the empirical evidence, both at the micro, cross-sectional and macro level

DGPs

- *Given a certain configuration of parameter values, initial conditions and stochastic disturbances*, the ABM is used to produce simulated or artificial data
- Therefore the ABM provides a *model-Data Generating Process (mDGP)*
- The mDGP must be compared with the *real world-Data Generating Process (rDGP)*
- The *empirical validation* of an ABM consists in checking whether the output of mDGP resembles real world empirical observations, i.e. observations drawn from the rDGP
- If not, the choice of *parameter values, initial conditions and stochastic disturbances should be reconsidered...*

rDGP

- rDGP generates *true* microeconomic data z_{it} characterizing agent i in a population indexed by $i \in I$ over a given sample period $1 < t < T$
- These data can be visualized as arranged in a matrix. Each row is a time series concerning the i-th individual. Each column is a cross section of the data at a given point in time
- Aggregating z_{it} over I we obtain the macroeconomic data Z_t ("from the bottom up")
- We can also compute cross sectional statistics, the cross sectional mean z_t and variance V_t for example

The matrix of real data

$i \downarrow, t \rightarrow$	1	2	3
1	z_{11}	z_{12}	z_{13}
2	z_{21}	z_{22}	z_{23}
Macro variable	Z_1	Z_2	Z_3
cross-sectional statistics	z_1, V_1	z_2, V_2	z_3, V_3

mDGP

- Given the *parameters*, the *initial conditions* and the realizations of the random variables that define the *shocks (random seeds)*, running the code the ABM generates *simulated or artificial* microeconomic data x_{it} and, through aggregation, the macroeconomic data X_t. These data can be visualized as arranged in a matrix

mDGP (cont'd)

- For instance, let x_{it} be the production of consumption goods by firm i at time t. This is a cell in a matrix describing the micro-production behaviour of firms over the period $T^* < t < T$.
- The modeller can compute, for example, the time series of the cross sectional mean x_t and cross sectional variance V_t of production of consumption goods

The matrix of simulated data (for a certain run)

$i \downarrow, t \rightarrow$	1	2	3
1	x_{11}	x_{12}	x_{13}
2	x_{21}	x_{22}	x_{23}
Macro variable	X_1	X_2	X_3
cross-sectional statistics	x_1, V_1	x_2, V_2	x_3, V_3

Ergodicity

- The real system is assumed to be *ergodic, i.e to self-organize, after a transition period, into a quasi-stationary long run equilibrium*. The mDGP therefore *must be* (sufficiently) *stationary* for all $t > T^*$ (where T^* are transients).
- If it isn't, the parameter configuration should be changed (*calibration*) until the mDGP stabilizes

Internal validity

- Due to the stochastic component of the model (random seed), there will be one x_{it} and one X_t *for each run*. In other words, there will be a run for each realization of the random variables that defines the shocks
- Denote the time series of the cross sectional mean and variance *for run m* with x_t^m and $V_t^{x,m}$, $m = 1, 2, \ldots M$
- After M realization of the shock and therefore M runs we have M time series of x_t^m and $V_t^{x,m}$
- Therefore, *for each period*, we have a distribution of M cross sectional means x_t^m and variances $V_t^{x,m}$

The matrix of simulated data (run m)			
$i \downarrow, t \rightarrow$	1	2	3
1	x_{11}^m	x_{12}^m	x_{13}^m
2	x_{21}^m	x_{22}^m	x_{23}^m
Macro variable	X_1^m	X_2^m	X_3^m
cross-sectional statistics	x_1^m, V_1^m	x_2^m, V_2^m	x_3^m, V_3^m

Internal validity (cont'd)

- We can compute the moments of these distributions, for instance $\bar{x}_t :=< x_t^m >$, $\bar{V}_t :=< (x_t^m - \bar{x}_t)^2 >$
- $\bar{x}_t$ is the mean of the cross-sectional means generated by the runs *(cross-simulations mean)*; $\bar{V}_t$ is the variance of the cross-sectional means generated by the runs *(cross-simulations variance)*

Internal validity (cont'd)

- If the cross-simulations variance is "small", we can deem the result essentially independent of the random seed
- In other words, the model is *robust* wrt changes of the random seed
- In this case the model is considere *internally valid*

Sensitivity analysis

- *For each possible set of parameters and initial conditions,* there will be a specific $\bar{x}_t$ and $\bar{V}_t$
- For instance for the set A we get $\bar{x}_t^A = < x_t^{A,m} >$, $\bar{V}_t^A = < \left(x_t^{A,m} - \bar{x}_t^A \right)^2 >$
- $\bar{x}_t^A$ is the cross-simulations mean generated by the set A; $\bar{V}_t^A$ is the cross-simulations variance generated by the set A
- Therefore, *exploring the space of parameter configurations and initial conditions (sets A,B,C etc) through Monte Carlo simulations,* we generate a distribution of $\bar{x}_t$ $\left(\bar{x}_t^A, \bar{x}_t^B, \bar{x}_t^C, \ldots\right)$ which helps analyzing in depth the behaviour of the mDGP.

Sensitivity analysis (cont'd)

- One can compute the mean and the variance of the distribution of cross-simulations means generated by the different parameter sets: $\hat{x}_t = E\left(\bar{x}_t^p\right); \hat{V}_t = E\left(\bar{x}_t^p - \hat{x}_t\right)^2$
- $\hat{x}_t$ is the *cross-configurations mean*, i.e. the mean of the cross-simulations means $\bar{x}_t^A, \bar{x}_t^B, \bar{x}_t^C$ generated by the different sets of parameters
- $\hat{V}_t$ is the *cross-configurations variance*, i.e. the variance of the cross-simulations means $\bar{x}_t^A, \bar{x}_t^B, \bar{x}_t^C$ generated by the different sets of parameters

Sensitivity analysis (cont'd)

- The modeler looks for a sample of $\bar{x}_t$ fairly *stable* (low cross-configurations variance), and as *close* as possible to the corresponding statistics generated by rDGP
- This validation exercise can be thought of as a way of calibrating the model in an optimal way, i.e. minimizing the distance between the simulated and the real data.

The indirect calibration approach

1. Identification of essential *stylized facts* the modeller wants to reproduce/explain
2. Development of an ABM capable of *replicating* the stylized facts
3. Use of empirical evidence on stylized facts to *restrict the space of parameters and initial conditions* of the ABM (for example if the mDGP is non stationary)
4. Analysis of of emergence of new stylized facts

Stylized facts: Aggregate evidence

- irregular oscillatory behaviour of GDP, unemployment rate, inflation rate etc.
- co-movements: productivity is pro-cyclical, unemployment rate anti-cyclical etc.
- dynamics of leverage, bankruptcies
- Phillips curve
- Okun's law
- Beveridge curve

Stylized facts: Cross sectional evidence

- firms' size distribution is log-normal or power law
- firms' growth rate distribution is tent shaped