Weekly Report (4.4-4.10)*

Yanpei Cai†

April 11, 2025

During the past week, I mainly used my time to tackle with coding while I spent a little amount of time to write a summary of maths in BOLFI. I have to say most of my time was spent on making up the oversight of my previous works, leading to a reconstruction of codes for the model which has yet completed. The followings are break down to progress and problems identified in each part with more details.

Coding

With available datasets, what can be observed is the percentage of never-smokers (N), smokers (S), and quitters (Q) in the population by the end of each year. With the agent-based model depicting the intricate dynamics of state change process for each agent, through simulations we are able to get the simulated data of each kind by the end of each step (year), which follows the methodology of BOLFI. By implementing BOLFI, what we want are inferences for the parameters which characterise the state change process as well as the social network topology.

The main problem emerged in my codes was actually caused by the oversight on the order of state change for each agent, and the insufficient proficiency of Mesa. When I was summarising the simulated data, the huge observed discrepancies reminded me to double check the model construction in my codes. I noticed that for simplicity of my codes' structure, I initially set each agent to update its state consecutively in the order of its node's **unique_id** instead of set the **SimultaneousActivation** mechanism. This actually causes a natural contagion of the cluster of nodes in the same state to nodes in other states in their neighbourhood. That explains why the contagion can be very fast, as can be seen in the wrong simulated data with moderate parameter values.

To solve this problem, I went back to the codes of **NSQ_Model** and **NSQ_Agent** to set up a more systematic formalism with specific settings. Given the short of time, this has yet completed but I am looking forward to completing the codes by next Monday.

Maths

Handwritten notes is attached after the main parts of this report.

†Artificial Intelligence and its Applications Institute, School of Informatics, The University of Edinburgh
Email Address: yanpei.cai@ed.ac.uk

*This is a summary of the works was done in the past week.

Questions Raised for This Week

- **Coding**: (first three tasks are expected to be completed by next Monday and send to Valerio for inspection)
  - **NSQ_Model** and **NSQ_Agent** settings for allowing all agents to update their states simultaneously.
  - **NSQ_Model** and **NSQ_Agent** settings in a more systematic formalism.
  - Data generating mechanism to compile with the time scale of both the UK and the US dataset.
  - **BOLFI** settings to cope with inferences for model parameters given the simulated data obtained in the last task.
- **Maths**: (in the **arXiv** version paper, **S3 Appendix. Parameter estimation from Christakis**)
  - Why does the interaction parameter **g** between SMOKER and QUITTER can be replaced by the probability of quitter over t yeras, as in (20)?
  - How does the expression in (22) are derived with applying (21) to both the interaction term and spontaneous term?
- **General Question (Physics to Complexity Science)**:

  How do statistical physics and those physical rules further assist our modeling of social systems and real-world social phenomena (e.g., heterogeneity of agents; capturing complexity versus model simplicity, etc.)?
- **General Question (Long-term Plan and Short-term Goals in PhD Journey)**:

  From my perspective, the short-term goals are coding the agent-based modelling mechanism for our smoking contagion model and implementing BOLFI to obtain better inferences for the model parameters. Can we have a discussion about the long-term plan (since you both seem to have many potential options)?

Plan for Next Week

- **Coding**:
  - The first three tasks in **Questions.Coding** by next Monday and get feedback if further problems arise.
  - Testing the **BOLFI** implementation to our smoking contagion model.
    IDEA: Firstly, obtain inference for single parameter by keep other model parameters and network generation parameters fixed.
- **Maths**:

  Try to answer the questions raised in **Questions.Math** (seek help and get feedback early next week if no progress are made during the weekend).

Question.

Given · a set (list) of observed data $\underline{y_o} = (y_o^{(1)}, y_o^{(2)}, \dots, y_o^{(n)})$ of n data points. (possibly dependent).

· a simulator-based statistical model, represented as

$\underline{\theta} = (\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(d)}) \in \mathbb{R}^d$ —input→ [SBM] —output→ $\underline{y_o}$ ↠ simulated data $\underline{y_o} \sim P_{\underline{Y}|\underline{\theta}}$

↘ $-\underline{\theta}$: finite # of parameters

$-$ SBM: a parameterised stochastic data generating mechanism.

⇒ a family of pdf(s) $\{P_{\underline{Y}|\underline{\theta}}\}_\theta$ that allows exact sampling of data.

Objectives : perform statistical inference of model parameters. (given observations).

Obstacles : generative process of the model can be very complex, so that the analytical form of the pdf $P_{\underline{Y}|\underline{\theta}}$ is often intractable, leads to inference using the likelihood function $L(\underline{\theta}) = P_{\underline{Y}|\underline{\theta}}(\underline{y_o}|\underline{\theta})$ is not available.

Solutions ⇒ Likelihood-Free Inference (LFI) Methods.

Basic Idea. since SBM allows data to be generated (via sampling from the pdf $P_{\underline{Y}|\underline{\theta}}$ of unknown analytical form, we generate data $\underline{y_s}$ with multiple values of input $\underline{\theta}$ and identify the model parameters $\underline{\theta}$ by finding values which yield simulated data $\underline{y_s}$ resemble the observed data $\underline{y_o}$.

General Manner. constructing approximate likelihood functions with the discrepancy Δ_θ between the observed data $\underline{y_o}$ and data $\underline{y_s}$ simulated with parameter values $\underline{\theta}$.

First Step. on a statistical level, the approximation consists of reducing the observed data $\underline{y_o}$ to some features (often means summary statistics).

* identifying summary statistics helps to reduce dimensionality and to retain sufficient information for the inference of $\underline{\theta}$.

$$L(\underline{\theta}) = P_{\underline{Y}|\underline{\theta}}(\underline{y_o}|\underline{\theta}) \Rightarrow L(\underline{\theta}) = P_{\underline{\Phi}|\underline{\theta}}(\underline{\Phi_o}|\underline{\theta}).$$

$\underline{\Phi} = (\Phi^{(1)}, \Phi^{(2)}, \dots, \Phi^{(p)}) \in \mathbb{R}^p$: summary statistics

↳ * still of unknown analytical form (inherited property from $P_{\underline{Y}|\underline{\theta}}$).

⇒ * $L(\underline{\theta})$ still needs to be approximated by some methods.

↳ Examples. - parametric approximation (Synthetic likelihood).

- non-parametric approximation (kernel density estimation, i.e. KDE).

Synthetic Likelihood (SL).

Key Ideas. - central limit theorem (CLT) supports the key multivariate normality approximation of the summary statistics, as $\underline{\Phi} \sim N(\underline{\mu}_\theta, \underline{\Sigma}_\theta)$, if the # of samples n of $\underline{z}$ is sufficiently large ⇒ this suggests (unknown)

$$p_{\underline{\Phi}|\underline{\theta}}(\underline{\phi}|\underline{\theta}) \overset{(\approx)}{=} \frac{1}{(2\pi)^{p/2}|\det \underline{\Sigma}_\theta|^{1/2}} \exp\left(-\tfrac{1}{2}(\underline{\phi}-\underline{\mu}_\theta)^T \underline{\Sigma}_\theta^{-1}(\underline{\phi}-\underline{\mu}_\theta)\right).$$

⇒ limiting likelihood approximation (with infinite computing power)

$$\hat{L}_S(\underline{\theta}) = p_{\underline{\Phi}|\underline{\theta}}(\underline{\Phi}_o|\underline{\theta})$$

$$\hat{l}_S(\underline{\theta}) = \log(\hat{L}_S(\underline{\theta})) = -\frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\det \underline{\Sigma}_\theta| - \tfrac{1}{2}(\underline{\Phi}_o-\underline{\mu}_\theta)^T \underline{\Sigma}_\theta^{-1}(\underline{\Phi}_o-\underline{\mu}_\theta)$$

- Simulator used to provide estimates of unknown mean $\underline{\mu}_\theta$ and covariance matrix $\underline{\Sigma}_\theta$ via a sample average E^N over N independently generated summary statistics, allowing the likelihood approximation to be computable.

$$\underline{\Phi}_\theta = (\underline{\Phi}_\theta^{(1)}, \underline{\Phi}_\theta^{(2)}, \dots, \underline{\Phi}_\theta^{(N)}). \quad \underline{\Phi}_\theta^{(i)} \overset{iid.}{\sim} p_{\underline{\Phi}|\underline{\theta}}, \quad \underline{\Phi}_\theta^{(i)} = (\Phi_\theta^{(i)(1)}, \Phi_\theta^{(i)(2)}, \dots, \Phi_\theta^{(i)(p)})^T$$

$$\hat{\underline{\mu}}_\theta = E^N[\underline{\Phi}_\theta] = \frac{1}{N}\sum_{i=1}^{N} \underline{\Phi}_\theta^{(i)}, \quad \hat{\underline{\Sigma}}_\theta = E^N[(\underline{\Phi}_\theta-\hat{\underline{\mu}}_\theta)(\underline{\Phi}_\theta-\hat{\underline{\mu}}_\theta)^T].$$

⇒ computable likelihood estimation (with finite computing power)

$$\hat{L}_S(\underline{\theta}) = \frac{1}{(2\pi)^{p/2}|\det \hat{\underline{\Sigma}}_\theta|^{1/2}} \exp\left(-\tfrac{1}{2}(\underline{\Phi}_o-\hat{\underline{\mu}}_\theta)^T \hat{\underline{\Sigma}}_\theta^{-1}(\underline{\Phi}_o-\hat{\underline{\mu}}_\theta)\right).$$

$$\hat{l}_S(\underline{\theta}) = \log(\hat{L}_S(\underline{\theta})) = -\frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\det \hat{\underline{\Sigma}}_\theta| - \tfrac{1}{2}(\underline{\Phi}_o-\hat{\underline{\mu}}_\theta)^T \underline{\Sigma}_\theta^{-1}(\underline{\Phi}_o-\hat{\underline{\mu}}_\theta)$$

Conceptual Diagram.