Weekly Report (2.21-2.28)[*]

Yanpei Cai[†]

February 28, 2025

During the past week, I engaged in reading Michael's paper and managed to get a broad map of why and how to use simulator-based statistical model with some details. I have also identified some tasks to do in the following weeks and some ways to improve my background knowledge and understanding of the paper. The following is a summary of the work that has been done during the past week and some possible measures for the plan for next week.

Summary

Recap of Bayesian Inference

Given the observed data $\boldsymbol{y}_o \in \mathbb{R}^n$ and a finite number of unknown parameters $\boldsymbol{\theta} \in \mathbb{R}^d$, Bayesian inference is the process of updating our understanding about $\boldsymbol{\theta}$ using the Bayes' rule, as

$$p_{\boldsymbol{\theta}|\boldsymbol{y}}(\boldsymbol{\theta}|\boldsymbol{y}_o) = \frac{p_{\boldsymbol{y}|\boldsymbol{\theta}}(\boldsymbol{y}_o|\boldsymbol{\theta})p_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{p_{\boldsymbol{y}}(\boldsymbol{y}_o)}$$

where

- $p_{\boldsymbol{\theta}|\boldsymbol{y}}(\boldsymbol{\theta}|\boldsymbol{y}_o)$ is the posterior distribution, representing the updated understanding about $\boldsymbol{\theta}$ after observing $\boldsymbol{y}_o$.
- $p_{\boldsymbol{y}|\boldsymbol{\theta}}(\boldsymbol{y}_o|\boldsymbol{\theta})$ is the likelihood, describing how probable the data is given a specific $\boldsymbol{\theta}$.
- $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is the prior distribution, encoding the understanding about $\boldsymbol{\theta}$ before seeing the data.
- $p_{\boldsymbol{y}}(\boldsymbol{y}_o)$ is the marginal likelihood (or evidence), computed as

$$p_{\boldsymbol{y}}(\boldsymbol{y}_o) = \int p_{\boldsymbol{y}|\boldsymbol{\theta}}(\boldsymbol{y}_o|\boldsymbol{\theta})p_{\boldsymbol{\theta}}(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

The likelihood function is defined as $\mathcal{L}_{\boldsymbol{\theta}|\boldsymbol{y}}(\boldsymbol{\theta}|\boldsymbol{y}_o) = p_{\boldsymbol{y}|\boldsymbol{\theta}}(\boldsymbol{y}_o|\boldsymbol{\theta}) = \mathbb{P}(\boldsymbol{y} = \boldsymbol{y}_o|\boldsymbol{\theta})$.

In summary, Bayesian inference allows for continuous updating of understanding as new data become available, making it a feasible framework for probabilistic reasoning.

[†] 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.

Simulator-based Statistical Models

A statistical model refers to a set of random variables that are defined on the same domain. It is defined via a family of probability density (mass) functions. A simulator-based statistical model is a parameterised stochastic data mechanism that specifies how data are generated. Given the observed data $\boldsymbol{y}_o \in \mathbb{R}^n$, a simulator-based statistical model with a finite number of unknown parameters $\boldsymbol{\theta} \in \mathbb{R}^d$ is a family of probability density (mass) functions $\{p_{\boldsymbol{y}|\boldsymbol{\theta}}\}_{\boldsymbol{\theta}}$ which enables direct sampling of data $\boldsymbol{y}_{\boldsymbol{\theta}} \sim p_{\boldsymbol{y}|\boldsymbol{\theta}}$ in a straightforward way. The prior $p_{\boldsymbol{\theta}}$ is explicitly chosen.

Notice that for any fixed $\boldsymbol{\theta}$ the output of the simulator is a random variable, but the probability density (mass) function $p_{\boldsymbol{y}|\boldsymbol{\theta}}$ cannot be analytically written in a closed form, and so the likelihood function $\mathcal{L}_{\boldsymbol{\theta}|\boldsymbol{y}}(\boldsymbol{\theta}|\boldsymbol{y}_o) = p_{\boldsymbol{y}|\boldsymbol{\theta}}(\boldsymbol{y}_o|\boldsymbol{\theta})$ is not explicitly known.

However, the probability density (mass) function $p_{\boldsymbol{y}|\boldsymbol{\theta}}$ is defined implicitly by the simulator (more details to be added). To compute the likelihood function, one need to be computed is the probability that the simulated data resemble the observed data. Instead of explicitly computing this, the simulator allows for explicitly testing whether the discrepancy between the simulated data $\boldsymbol{y}_{\boldsymbol{\theta}}$ and the observed data $\boldsymbol{y}_o$ is zero or small enough. This property yields a set of methods used to estimate parameters of a model without explicitly knowing the likelihood function, which is called the likelihood-free inference.

Likelihood-free Inference of Simulator-based Statistical Models

As per introduced in the last section, likelihood-free inference (LFI) refers to a set of statistical methods used to estimate parameters of a model without explicitly computing the likelihood function. This is particularly useful when the likelihood is intractable (p.s. in agent-based models, the likelihood function too complex to compute analytically). Instead of explicitly evaluating the likelihood, LFI methods use simulations to compare observed data with simulated data generated from the model.

General Strategy

- **Simulation-Based Approach**: Since the likelihood function is intractable, LFI generates synthetic data from the model using different parameter values.
- **Comparison to Observed Data**: The simulated data is compared to the observed data. Discrepancy is obtained directly/indirectly (e.g. distance metrics, summary statistics, etc).
- **Parameter Estimation**: The best-fitting parameters are inferred by implementing different metrics, such as rejection sampling, approximate Bayesian Computation (ABC), Kernel Density Estimation, etc.

Exact Inference v.s. Approximate Inference

- **Exact Inference**: Posterior distribution is obtained by the retained simulated data via rejection sampling. **NOTE:** Only valid for discrete random variables. (more details to be added)
- **Approximate Inference**:

  - Reduce the dimensionality of the data to some features or summary statistics.
  - Allow the discrepancy between the simulated data and the observed data to be small enough instead of to be exactly zero. **REASON:** The probability of the discrepancy to be zero can becomes smaller and smaller as the dimension of the data increases, the posterior is therefore not well-described by the accepted samples.

Examples

- **Approximate Bayesian Computation (ABC)**: e.g. Rejection ABC (more details to be added)
- **Synthetic Likelihood**: Synthetic likelihood function is obtained using a Gaussian approximation on summary statistics (e.g. if the summary statistics is obtained via averaging then the likelihood function is approximated Gaussian via CLT). It explicitly estimates a likelihood but in a simplified manner.
- **Kernal Density Estimation**: (more details to be added)

Discussion

There are few difficulties of generalising likelihood-free inference:

- Assessment of the discrepancy between the simulated data and the observed data.
- Less efficiency due to the simulations for large datasets.
- Lack of knowledge about the relation between discrepancy and model parameters

Plans for Next Week

Some possible measures proposed by Valerio, Ben, and myself:

- Continue to read the paper, dive into more details of each method, reproduce the mathematics.
- It might be worth to have a meeting with Michael directly to discuss some details.
- Also, it might be worth to discuss with Ogy and Guillermo whenever get stuck.
- Book: Bayesian Statistics the Fun Way
- Course: Probabilistic Modelling and Reasoning
- **!!!:** Never get too stressed out! Enjoy the PhD journey!