STATS 531 Homework 7

Due Sunday 4/5

You have two weeks for this homework. You are advised not to leave it to the last moment, in case you encounter logistic difficulties with greatlakes. Sometimes, there can be a queue for the GPU, so allow time for this.

Please submit your homework report to Canvas as both a Quarto (qmd) file and a pdf file produced by it. You are welcome to collaborate with other members of your final project group, and you are also welcome to post issues to the class GitHub site to share advice or ask questions. You should run your own code, and as usual you should report on all sources and give proper acknowledgement of the extent of their contributions. Proper acknowledgement involves listing sources at the end of the report and citing the sources at all appropriate points during the report. It is expected that your solution to Question 7.2 will involve borrowing from code provided in the notes. Relevant material is also available online, if you need extra hints, but you may learn more by starting from the code in the notes. In addition, you may consult an AI for coding advice. Your report should document issues that arose and explain the work you put into your solution.

Question 7.1. Introduction to JAX on the greatlakes cluster

The greatlakes cluster is a collection of high-performance Linux machines operated by University of Michigan. Each machine has 36 CPU cores, and some machines have NVIDIA GPUs. Linux clusters and NVIDIA GPUs running CUDA are the standard platform for computationally intensive statistics and data science, so learning how to work on greatlakes is worthwhile, if this is new to you. This question may be easy if you are already familiar with greatlakes.

It is possible to access greatlakes via an on-demand web interface. However, for larger tasks it is better to submit batch jobs from a greatlakes terminal, and that is that we practice here.

Read the greatlakes notes on the course website and work through the example to run the code in the file test.py on greatlakes, on both CPU and GPU hardware.

(a) Report on any issues you had to overcome to run the test code as a batch job on greatlakes. Did everything go smoothly, or were there problems you had to overcome?

(b) Have you used a Linux cluster before?

(c) Compare the run times reported by test.py for greatlakes CPU and GPU, and your laptop. How do you interpret these results?

Question 7.2. Likelihood maximization for the SEIR model.

We consider an SEIR model for the Consett measles epidemic, which is the same model and data used for Homework 6. Write a report presenting the following steps.

You will need to tailor the intensity of your search to the computational resources at your disposal. Choose the number of starting points, number of particles employed, and the number of IF2 iterations appropriately for the size and speed of your machine. You should test your code on smaller tasks before moving to larger numbers of particles and search iterations. Do this by setting up a run level variable (see Chapter 14 for examples) so that your qmd file compiled with `RL=0` runs in seconds on your laptop, `RL=1` runs in 5 to 10 minutes on your laptop, and `RL=2` runs in under 30 minutes on a greatlakes GPU (counting run time, not counting time waiting for the job to queue which depends on how busy the cluster is). Your qmd file should collect and report the run time separately for each long calculation.

It is okay for this homework if the Monte Carlo error is larger than you would like. You should aim to get as much accuracy as possible given the constraints on run time.

You are advised to develop `RL=0` first, on your laptop. When `RL=1` is working on your laptop, try running this on greatlakes. Once you know how long `RL=1` takes on a GPU, you can magnify the computational effort appropriately for `RL=2`.

(a) Conduct a local search and then a global search using the multi-stage, multi-start approach demonstrated in the notes.

(b) How does the maximized likelihood for the SEIR model compare with what we obtained for the SIR model?

(c) How do the parameter estimates differ between SIR and SEIR?

(d) Calculate and plot a profile likelihood over the reporting rate for the SEIR model. Construct a 95% confidence interval for the reporting rate, and discuss how this profile compares with the SIR profile in Chapter 15.

Acknowledgements

Question 7.2 derives from material in a short course on Simulation-based Inference for Epidemiological Dynamics using Pypomp, which in turn derives from a previous version of this course using R-pomp.