High Performance Computing - An EBS Perspective

Ze-Yu Zhong

Wednesday 27 September 2023

High Performance Computing for Econometrics

HPC can be well suited for some scientific workloads, such as those in Econometrics. Hands on session Agenda:

- Overview + Prerequisites
- Login Node Setup
- IV Estimator
- Bash Script

Recommended that people try to code/follow along
https://github.com/zeyuz35/hpc_workshop

Prerequisites

HPC account + Project

- Various computational projects available - MonARCH is specifically built for Monash staff/students

ssh client

- putty (Windows), terminal (macOS or Linux)

ftp client to access HPC files

- MobeXterm (Windows), Cyberduck (macOS), Nautilus/Dolphin (Linux)

Login Node Setup

1. Login to MonARCH
   - ssh user@monarch.erc.monash.edu
2. Load R modules
   - module load R
3. Link folder for R libraries (R_LIBS)
4. Install R libraries
   - R
   - install.packages("tidyverse")
5. Copy code over to cluster
   - Recommended to use git
   - cd project/user && git clone

Monte Carlo Example

IV Regression Example:

$$y = x_1\beta_1 + e_1, \quad e_1 \sim N(0,1) \tag{1.1}$$
$$x_1 = \gamma z + e_2, \quad e_2 \sim N(0,1) \tag{1.2}$$
$$cor(e_1, e_2) = \rho$$

Interested in properties of IV estimator $\widehat{\beta_1}$ across different

- Instrument strength $\gamma \in \{0, 0.25, 0.5\}$
- Endogeneity $\rho \in \{0, 0.25, 0.5\}$
- Sample size $N \in \{100, 200, 500\}$

Record $\widehat{\beta}_1$ and $se(\widehat{\beta}_1)$ for each specification, for $R = 1000$ replications

Code

1. Code up a minimum working example that runs on your local machine, e.g. for a small number of replications
2. Take note of how long, extrapolate how much time it would take to run on the HPC cluster
3. Convert the local code to something that is distributed across different HPC arrays
   - Typically, letting each array handle a different DGP specification is most straightforward
4. Prepare job script, and submit

Job Script using Bash

See documentation for basic Bash script. Parallelization:

- HPC can be well suited for parallelized/split workloads
- This is done via the ARRAY environment variable

Practical Advice

Other uses for HPC relevant for EBS:

- Rolling/expanding window estimation (use rsample to set up slices)
- Cross Validation (rsample)
- Bootstrap/Jackknife
- Access to expensive GPUs (advanced)

Do not request too many resources - this can take a long time to be allocated. Do not mess with job priority unless you have a legitimate reason - this is bad etiquette

Extra Resources

Data Fluency Workshops (free for students!):

- Introduction to Bash/Shell Scripts
- Introduction to HPC

Advanced Issues

rcpp

- Compiling rcpp code on one array and asking other arrays to use this is inconsistent - no guarantee that different arrays are of same architecture
- Solution: ask explicitly for same compute instance nodes OR compile code for each array (inefficient, but not usually not prohibitively so)