Advanced Statistical Programming Camp
January 2015

Monday, January 26th – Friday, January 30th
Morning Session: 9:30am – 10:30am
Afternoon Session: 1:30pm – 2:30pm
Location: Sherrerd Hall 101

Instructor:	Hubert Jin
Office:	Corwin 029
Email:	`hubertj@princeton.edu`
Office Hours:	3:30pm – 4:30pm (during camp)

Description. The Advanced Statistical Programming Camp builds on the introductory Statistical Programming Camp by expanding the computing toolsets of researchers. The camp provides tools which can help analyze big datasets (e.g., voter files across many states, micro-level international trade data, large federal personnel databases) and employ computationally intensive methods (e.g., Monte Carlo simulations, Bayesian Markov chain Monte Carlo, cross-validation, or the bootstrap).

We begin by introducing some low-cost strategies for improving performance in R. To help process large data and improve the speed of computation, we then cover parallel execution of R code on both personal machines and on remote high performance computing systems available at Princeton. Lastly, we cover basic C++ and the use of Rcpp to produce tightly integrated and fast compiled code.

Prerequisites. Although anyone can sit in on the camp, the following is assumed as a starting point:

- familiarity with the R programming language up through the level covered in the Politics Statistical Programming Camp (`http://goo.gl/VkPtI8`)
- practical experience using R
- access to a computer that has been setup according to provided instructions (see below for details)

Structure. Because a lot of material will be covered over the course of a week, this camp is very much an immersion. We will meet each day from Monday through Friday during both a morning session and an afternoon session. These sessions will be on the *interactive* and *hands-on* end of the spectrum, so bringing a laptop to the camp is strongly recommended (and necessary for getting the most out of the sessions). Prior to some sessions, handouts introducing material with demonstrations will be distributed. Then, during the session, we will work through applying these tools to actual computing problems in Political Science.

Discussion Board. This camp will be using the Piazza discussion board (https://piazza.com/) to facilitate discussions and questions throughout the Advanced Statistical Programming Camp. Piazza provides an interactive environment in which to both ask questions and answer those of others. To join the Advanced Statistical Programming Camp Piazza site, click on "Search Your Classes" from the Piazza homepage. After specifying Princeton University as your school, search for "Advanced Statistical Programming Camp". You will then be prompted to enter your princeton.edu email address to confirm your registration. Piazza can also be accessed from within Blackboard by going to the ASPC organization page and clicking on the link to "Piazza Q&A". In addition, all class announcements will be made through Piazza. Blackboard will still be used for hosting all class materials and for submitting assignments.

Some tips and tricks for Piazza include:

- Piazza has apps available for the iOS and Android platforms. The apps are free downloads and provide complete access to all of Piazza's messageboard features.
- To insert LaTeX-formatted text in a post, place a double dollar sign ("$$") on both ends of the relevant text, or click the "*fx*" button in the Details toolbar above your post.
- To add formatted R code to a post, click the "pre" button in the Details toolbar above your post. A grey text box will open up where you can paste code from R. You can classify a post using pre-selected tags, or you can generate your own by prepending a hash (#) to your chosen label. Posts can then be sorted by these tags using the search bar in the left-hand column.
- We encourage you to mark helpful contributions (particularly those from classmates) using the "Thanks!" button at the bottom of each post.

Machine Setup. Instructions detailing pre-camp setup instructions are available on Blackboard. They mostly require registration for access to high-performance computing resources through Princeton University's Research Computing department.

Materials. No outside materials are necessary for this course. However, if your own work follows along the trajectory that we take in this course, the following *optional* supporting resources will likely be helpful.

- The R Inferno (http://www.burns-stat.com/documents/books/the-r-inferno/). This is a conversational treatment of some straightforward aspects of R that you can (but won't want to) go years without ever knowing. There is a free PDF available for download on the website.
- The R Manuals (http://cran.r-project.org/manuals.html). What do you mean you haven't read them yet?
- Hadley Wickham's notes on Rcpp (http://adv-r.had.co.nz/Rcpp.html).
- The Rcpp Gallery (http://gallery.rcpp.org/) which demonstrates various functionality in short, self-contained snippets.
- The Armadillo Library (think linear algebra in C++) API Documentation (http://arma.sourceforge.net/) which documents the available functionality in **RcppArmadillo**.

Topics

Day	Session	Topic	Details
Monday (1/26)	AM	Introduction	syllabus, logistics, setup of computers/accounts
		Demonstration of Methods	overview of upcoming topics to come
	PM	Monitoring Performance in R	timing individual functions
		Simple Performance Improvements	vectorization, built-in functions
		Convenience Wrappers	`apply()`, `tapply()`, and `sapply()`
		A New Looping Construct	**foreach**
		Shared Memory Parallelization	**parallel**, **doParallel**
		Parallel RNG	**doRNG**
Tuesday (1/27)	AM	TIGRESS Familiarity	`ssh`, `rsync`
		Basics of Submitting Jobs on TIGRESS Systems	`sbatch`, SLURM files
		Distributed Memory Parallelization	**doMPI**
	PM	Seamless Development and Execution on Local and Remote Systems	passing arguments, environmental variables
		More on Submitting Jobs on TIGRESS Systems	job arrays
Wednesday (1/28)	AM	Basic C++ through Rcpp	**Rcpp**, `sourceCpp()`
	PM	More Basic C++ through Rcpp	
Thursday (1/29)	AM	Matrix Algebra in C++ with RcppArmadillo	**RcppArmadillo**
	PM	C++–level Parallelization with OpenMP	multi-thread loops
Friday (1/30)	AM	Using R Packages to Organize R and C++ Code	`compileAttributes`
		Putting it All Together	

Examples

Day	Session	Details
Monday (1/26)	AM	euclidean pairwise distance (US Counties)
	PM	parametric and non-parametric bootstrap (voter turnout)
		cross-validation (civil war onset)
Tuesday (1/27)	AM	Monte Carlo analysis of omitted variable bias in linear and non-linear models
	PM	Monte Carlo analysis of coverage probabilities in non-linear models (international alliances)
Wednesday (1/28)	AM	surface of earth pairwise distance (US Counties)
	PM	Bayesian Probit regression via Gibbs Sampler (voter turnout)
Thursday (1/29)	AM	sparse and dense linear regression (linear regression of dyadic trade flows with country fixed effects)
	PM	surface of earth pairwise distance (US Counties)
		parallel Probit regression via EM (voter turnout, document classification)
Friday (1/30)	AM	Bayesian Probit regression via Gibbs and EM (voter turnout)