GC-15

Parallelizing Random Lasso

James Matthew Hamilton | Daniel Karasek | Sams Khan | Jason Wein | Seung Myeong Choo | Devyn Wilkins

Problem

Big Data is not always big enough.

- One could have terabytes of data, but still not enough samples to do any meaningful research on using traditional regression analysis. **How is that possible?**
- Think: **1,000,000 (Features)** x **200 (Samples)**
- Ratio of features to samples can impact results. The optimal dataset for traditional regression analysis will have **many samples** and **few features** comparatively, unlike above.

Solution

- Just get more samples… **NO**
- Depending on your data it could be costly to get more samples or nearly impossible
- Develop better mathematical models… **YES**

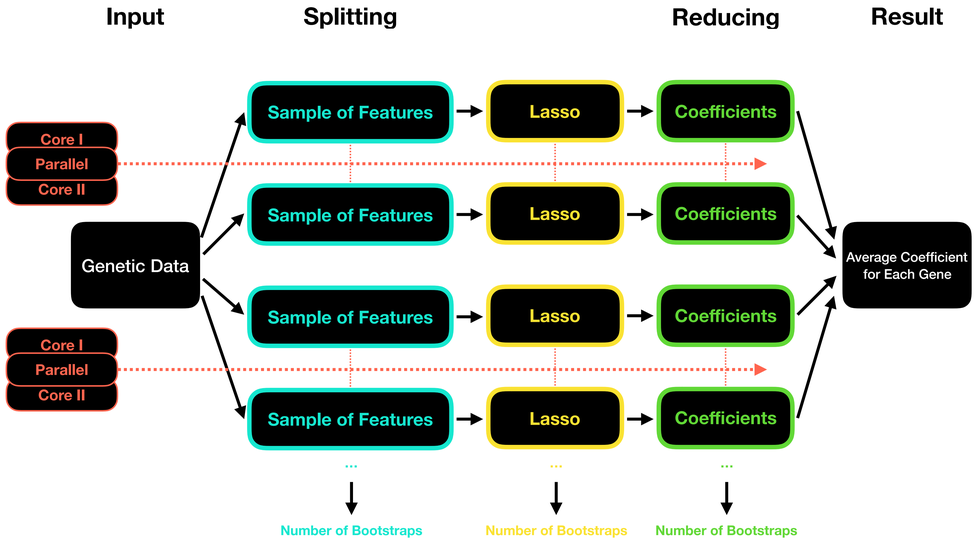
What is Random Lasso?

- Random lasso is the state-of-the-art method of regression analysis for multidimensional datasets with comparatively high features and low samples
- Genetic datasets benefit highly, but Random Lasso can be applied to any datasets that meets the following condition: a sample size that is significantly less than the number of features, i.e. **samples << features.**
- **Lasso performs poorly if samples << features**, so **we make samples == features**

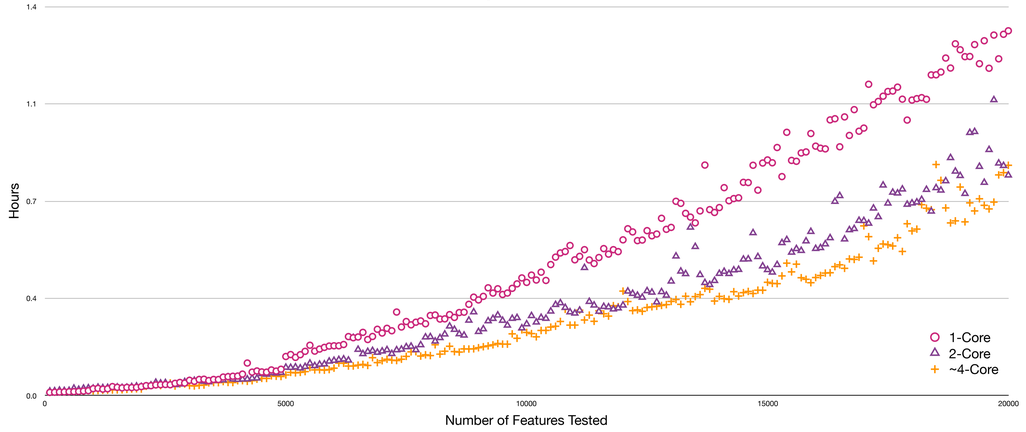
Why Random Lasso?

Common Dataset: Genetic

There exists no package for regression analysis using random Lasso. Due to the complexity of Random Lasso, researchers often opt for more juvenile methods of regression analysis that are less accurate, but pre-made in a package. We designed an open source package for Random Lasso.

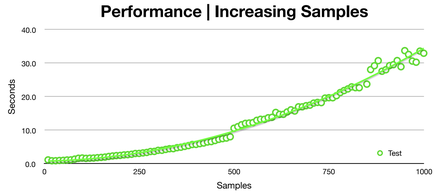


Testing Performance of Multicore | Scaled to Estimate Hours
MacBook Pro | 2.5 GHz Intel Core i5



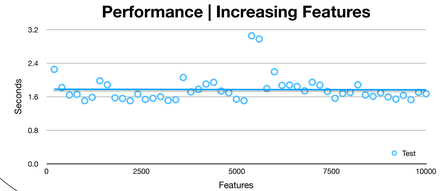
Note: The number of samples was always one-tenth of the features in this testing.

Performance | Increasing Samples



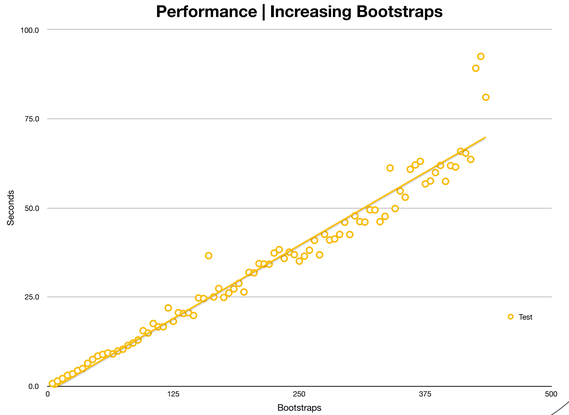
Features: 2000 | Bootstraps: 10

Performance | Increasing Features



Samples: 100 | Bootstraps: 100

Performance | Increasing Bootstraps



Samples: 100 | Features: 50

Parallelizing

- Runtime for any meaningful research using Random Lasso can be hours or days.
- Parallel programming takes advantage of multiple cores within a computer to run tasks congruently, but this is only possible if some tasks are independent of each other.
- Random Lasso has highly-independent tasking, making the prospect of parallel programming attractive.

Our Package

RandomLasso(x, y, bootstraps, alpha = c(1, 1), verbose = TRUE, cores = 2)

x: Matrix of independent data.

y: Matrix of dependent data.

bootstraps: Number of random bootstraps taken.

alpha: 0 = ridge regression, 0.5 = elastic-net, 1 = lasso

verbose: suppresses all printing and time estimating functions.

cores: by default, the system will try to detect number of cores.

Contact

James Matthew Hamilton | jhamil86@students.kennesaw.edu

Daniel Karasek | karasekdaniel@gmail.com

Sams Khan | skhan34@students.kennesaw.edu

Jason Wein | jwein2@students.kennesaw.edu

Seung Myeong Choo | schoo1@students.kennesaw.edu

Devyn Wilkins | dwilki22@students.kennesaw.edu

KENNESAW STATE UNIVERSITY