A Posterior Look at NRC Rankings*

Yihui Xie

December 2, 2011

1 Introduction

The National Research Council released the Data-Based Assessment of Research-Doctorate Programs on September 28, 2010. We have extracted a subset of the data concentrating on the rankings of statistics departments; it is available[1] in the **cranvas** package at `https://github.com/ggobi/cranvas` as the data frame `nrcstat`. The original dataset has 61 observations (i.e., departments) and 68 variables, e.g, number of publications, citations, average GRE scores of students and number of students, etc. The goal of this analysis is to identify important factors that affect the rankings of departments. Note there are two types of rankings: S (survey-based) rankings and R (regression-based) rankings, and for each type, 5% and 95% percentiles are published; we average these two percentiles to get our response variables. I focus on R rankings in this report since they are new compared to previous years. I have not checked the details of their regressions, and this is why I call this report a "posterior" look – I'm going to use regularized regressions to explore the results of their regressions.

I cleaned up the dataset and the number of variables that come into the analysis is 46; to get a feel about the data, Figure 1 is a scatter plot of the R rankings vs GRE scores: it is not surprising that higher GRE scores tend to lead to better rankings. As a side note, always bear in mind that lower ranks are better.

```
qplot(Ave.GRE.Scores, R.Rankings, data = nrc) + geom_smooth()
```

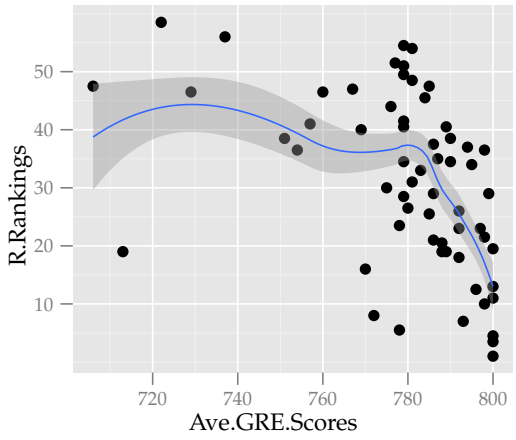

Figure 1: R rankings vs GRE scores: higher scores, better rankings (N.B. a smaller rank means better!)

*the word "posterior" here has nothing to do with Bayesian yet

[1] it may be subject to changes in the future, but the version that I used in this report is always available at `https://github.com/ggobi/cranvas/blob/2c34d81c29369b29c281206c9733fbc7c19509b4/data/nrcstat.rda` thanks to the version control tool GIT

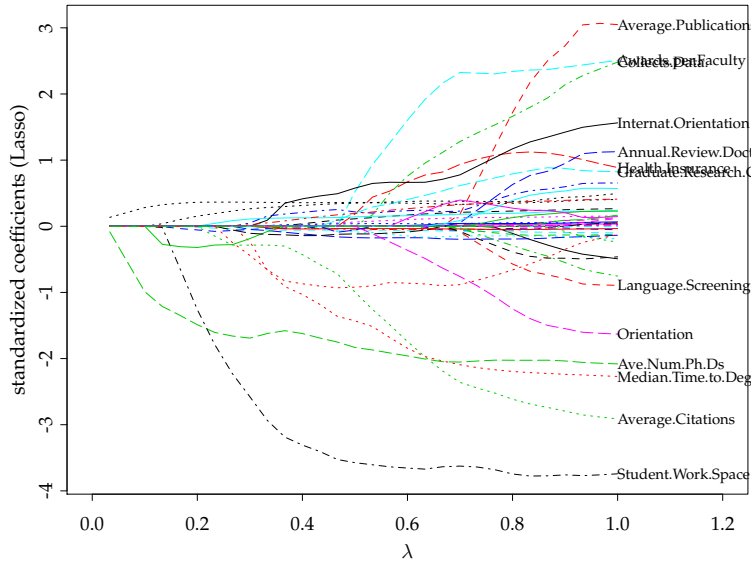

Figure 2: The Lasso estimates vs λ: statistics departments must give students more work space!

We can throw all the variables into an ordinary linear regression model and some indicators seem to be highly significant, e.g, the average completion (degree) ratio in 6 years or less (positive coefficient – to get better ranks, departments should not let students graduate easily?), and the number of PhD students (negative coefficient – more students, better ranks?). Given the large number of variables, however, the interpretation of regression estimates can be a non-trivial problem, where Lasso can come into play (Tibshirani, 1996).

2 Lasso and Bayesian Lasso

Recall the goal of the Lasso is to minimize

$$(\tilde{y} - X\beta)'(\tilde{y} - X\beta) \quad s.t. \sum_{j=1}^{p} |\beta_j| \leq \lambda$$

and different λ's will lead to different estimates of β's. The maximum value for λ is the sum of absolute values of coefficients from OLS (values greater than that will lead to OLS estimates), so in practice we may specify λ to be a value between 0 and 1, which is actually a ratio of the current constraint to the maximum one, and the **lasso2** package (Lokhorst *et al.*, 2011) uses this notation. We chose λ values from $i/30$ for $i = 1, 2, \cdots, 30$ and the ordinary Lasso gives us the results in Figure 2 (variable labels in the middle are ommitted to make the plot clearer).

Next we implement the Bayesian Lasso using the Gibbs sampler proposed in Park and Casella (2008), from which our notations are borrowed. For each fixed λ, we have full conditionals

$$\begin{aligned}
\boldsymbol{\beta}|\sigma^2, D_\tau &\sim MVN(A^{-1}X'\mathbf{y},\, A^{-1}\sigma^2) \quad \text{where } A = X'X + D_\tau^{-1} \text{ and } D_\tau = diag(\tau_1, \ldots, \tau_p) \\
\sigma^2|\boldsymbol{\beta}, D_\tau &\sim IGamma((n-1)/2 + p/2,\, ((\mathbf{y} - X\boldsymbol{\beta})'(\mathbf{y} - X\boldsymbol{\beta})/2 + \boldsymbol{\beta}' D_\tau^{-1}\boldsymbol{\beta}/2)) \\
1/\tau_j^2 &\overset{ind}{\sim} IGauss(\sqrt{\lambda^2\sigma^2/\beta_j^2},\, \lambda^2)
\end{aligned}$$

which is easy to sample from. The Gibbs sampler is initialized with σ^2 and $\boldsymbol{\beta}$ by their MLE's, and we start with τ_j's, then iterate through all the above conditionals.

Figure 3 is the trace plot of Bayesian Lasso for a series of λ's between 0 and 1. For some reason, Bayesian estimates are systematically smaller than ordinary Lasso estimates, and they differ in the speed of approaching 0 as $\lambda \to 0$ as well. Student work space still seems to be important here,

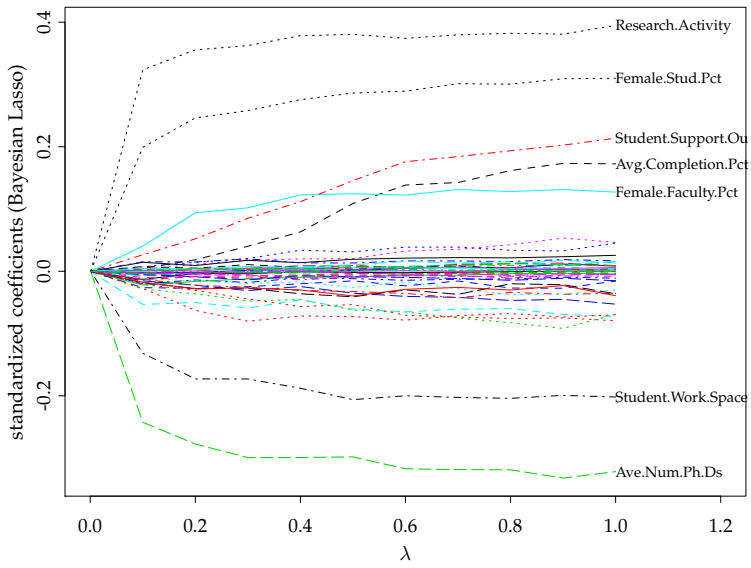

Figure 3: Bayesian Lasso estimates vs λ: statistics departments must enroll more PhD's!

although less than the number of PhD students. There are also some new factors that can lead to worse rankings in Bayesian Lasso estimates (those above 0), which I prefer not to discuss here.

3 Conclusions

In this report, we analyzed the NRC rankings (R) data of 61 statistics departments in the US. Both ordinary Lasso and Bayesian Lasso were presented; the indicator "Student work space" stood out in both trace plots. The Bayesian Lasso estimates approches to 0 more slowly than ordinary Lasso (there are not many dramatic changes in the traces in Bayesian Lasso), but they concentrate more around 0 in general; the reason is unclear to me. It is interesting to see some factors like research activity (in Bayesian Lasso) and publication (in ordinary Lasso) are negatively related to rankings (although the interpretation should be "conditional on other variables, negatively correlated"). It might worth looking at credible intervals of coefficients, but they are omitted in this report due to page limits.

This report was written with the **knitr** package; long chunks of R code (e.g., for the Gibbs sampler) to represent the real hard work were intentionally hidden but available at `https://github.com/downloads/yihui/knitr/Stat615-Report1-Yihui-Xie.Rnw`.

References

Lokhorst J, Venables B, Turlach B (2011). *lasso2: L1 constrained estimation aka 'lasso'*. R package version 1.2-12, URL `http://CRAN.R-project.org/package=lasso2`.

Park T, Casella G (2008). "The bayesian lasso." *Journal of the American Statistical Association*, **103**(482), 681–686.

Tibshirani R (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288.