BSTA 670 Statistical Computing: Proposal 2

John Pluta

Fall 2014

An important feature of the LASSO penalized regression function is that it can perform variable selection by setting the coefficient estimates of certain variables exactly to zero [1]. Thus LASSO is a popular tool for sparse data, where only some subset of the coefficients have some non-zero number as their true weight. However, it is important to understand to what extent LASSO estimates choose the correct covariates, and how consistently these are chosen. Work by [2] attempt to quantify this through the definition of an "Irrepresentable Condition". This condition essentially quantifies the relationship between the relevant (with non-zero beta-estimates) and irrelevant variables; the less similar they are, the more likely that LASSO will pick the correct covariates.

In this paper, the authors simulated datasets while varying both the total number of variables, and the proportion of variables that were relevant, and calculated what percentage of the simulated data met the Irrepresentable Condition for each case (Table 3 in [2]). The results show that fewer overall variables and greater sparsity imply more consistent and more accurate model selection. Similarly, Figure 2 of that paper shows the effect of the degree of the Irrepresentable condition on how frequently LASSO solutions are sign-consistent with the true model. Both results show strong evidence that this Irrepresentable condition must be met for valid results.

For my project, I will replicate the simulation that proceeds Figure 2. I will perform a similar experiment but with a dataset where $p >> n$. One of the commonly referenced highlights of LASSO is that it can accommodate this kind of data, yet the results of [2] seem to indicate that correct model selection would be unlikely in these conditions (this case was not tested in the paper). This is an important result to be able to verify or deny.

References

[1] Tibshirani, R. 1996. *Regression shrinkage and selection via the lasso.* J. R. Statist Soc. 58(1):267-288.

[2] Zhao, P., Bin, Y. 2006. *On model selection consistency of Lasso.* Journal of Machine Learning Research 7:2541-2563.