Introduction to Model Selection, Regularization, and Post-Model Selection Inference

Christian Hansen (University of Chicago)
8-9 September, 2016

Course Description:

As in many other fields, economists are increasingly making use of high-dimensional models – models with many unknown parameters that need to be inferred from the data. Such models arise naturally in modern data sets that include rich information for each unit of observation (a type of "big data") and in nonparametric applications where researchers wish to learn, rather than impose, functional forms. High-dimensional models provide a vehicle for modeling and analyzing complex phenomena and for incorporating rich sources of confounding information into economic models.

My goal in this short course is two-fold. First, I wish to provide an overview of and introduction to variable selection methods with an emphasis on penalized estimation methods. Second, I will present an introduction to recent proposals that adapt high-dimensional methods to the problem of doing valid inference about model parameters and illustrate applications of these proposals for doing inference about economically interesting parameters.

Outline:

Lecture 1: Selection and Penalized Estimation Methods

- James, G., D. Witten, T. Hastie, and R. Tibshirani (2014), An Introduction to Statistical Learning with Applications in R, Springer. [Chapters 6]
- Belloni, A. and V. Chernozhukov (2013), "Least squares after model selection in high-dimensional sparse models," *Bernoulli*, 19(2), 521-547.
- Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012), "Sparse models and methods for optimal instruments with an application to eminent domain," *Econometrica*, 81(2), 608-650.
- Belloni, A., V. Chernozhukov, C. Hansen, and D. Kozbur (2016), "Inference in high-dimensional panel models with an application to gun control," forthcoming *Journal of Business and Economic Statistics*.

- Chen, J., and Z. Chen (2008), "Extended Bayesian Information Criterion for Model Selection with Large Model Spaces," *Biometrika*, 95, 759–771.
- Fan, J. and J. Lv (2008), "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society, Series B*, 70(5), 849-911.
- Jing, B.-Y., Q.-M. Shao, and Q. Wang (2003), "Self-normalized Cramer-type large deviations for independent random variables,'' *Annals of Probability*, 31(4), 2167-2215.
- D. Kozbur (2016), "Testing-based forward model selection," arXiv:1512.02666
- H. Wang (2009), "Forward regression for ultra-high dimensional variable screening," *Journal of the American Statistical Association*, 104, 1512-1524.

Lecture 2: Inference after Model Selection or Regularization

- Belloni, A., D. Chen, V. Chernohukov, and C. Hansen (2012), "Sparse models and methods for optimal instruments with an application to eminent domain," *Econometrica*, 80(6), 2369-2430
- Belloni, A., V. Chernozhukov, and C. Hansen (2014), "High-dimensional methods and inference on structural and treatment effects," *Journal of Economic Perspectives*, 28(2), 29-50
- Belloni, A., V. Chernozhukov, and C. Hansen (2014), "Inference on treatment effects after selection amongst high-dimensional controls," *Review of Economic Studies*, 81(2), 608-650
- Belloni, A., V. Chernozhukov, and C. Hansen (2015), "Inference in high dimensional panel models with an application to gun control," forthcoming *Journal of Business and Economic Statistics*
- Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2013), "Program evaluation with high-dimensional data," forthcoming *Econometrica*
- Chernozhukov, V., C. Hansen, and M. Spindler (2015), "Post-selection and post-regularization inference in linear models with many controls and instruments," *American Economic Review*, 105(5), 486-490
- Chernozhukov, V., C. Hansen, and M. Spindler (2015), "Valid post-selection and post-regularization inference: An elementary, general approach," *Annual Review of Economics*, 7, 649-688
- Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey (2016), "Double machine learning for treatment and causal parameters," working paper.