DISSERTATION PROPOSAL

Nandana Sengupta

Application of Machine Learning Classifiers and Regularization in Econometric Theory

Thursday, May 1, 2014
10:00 am
324 GSIA (West Wing)

The main focus in econometrics is to provide an explanation of the various observed outcomes. Structural econometricians obtain reliable estimates of parameters that describe an economic system to provide an understanding of the underlying processes that determine equilibrium outcomes. The estimation process is based on conditions implied by economic theory.

On the other hand, the main focus in machine learning is to provide accurate predictions of the variables of interest. While these techniques are extremely powerful for forecasting, it can be very hard to interpret the underlying structure implied by them.

As machine learning techniques become more popular and computers become capable of storing and processing large quantities of data, there have been many recent efforts to incorporate such techniques into structural econometric models. My research aims to extend this literature by introducing the techniques of regularization and classification (from machine learning) into Generalized Method of Moments (GMM) and Propensity Score Weighting frameworks (from econometrics).

Paper 1: Regularization Paths in the Generalized Method of Moments Framework

In the GMM framework, the objective function to be minimized is a weighted sum of squares of `m' moment conditions implied by economic theory. The derivative of the objective function with respect to the vector of parameters provides a system of `k equations in k unknowns' that is used to obtain parameter estimates. However if this matrix is nearly singular at the true parameter values, then the solution of the system of equations is highly unstable. This results in high standard errors of the parameter estimates. This is analogous to the problem of multicollinearity in linear regression. In the linear regression framework the problem is somewhat overcome by regularization. Ridge and spectral cut-off regularization are commonly used. However, due to the highly non-linear nature of the GMM objective function, ridge and spectral cut-off are not readily generalizable to the GMM framework.

In the first chapter of my thesis, we re-interpret regularization as a set of possible solutions that lie along a path between the unconstrained minimum of the objective function and the mean of a pre-defined prior. Using this interpretation, we propose an algorithm for finding the `regularized' parameter estimates. We introduce the notion of cross-validation in GMM. We also show via simulations that our method performs very well when the system of equations is unstable. As an empirical application we employ this method on the Capital Asset Pricing Model.

Paper 2: Machine Learning Classification Techniques in Propensity Score Weighting

The basic issue in estimating the effect of a particular treatment using observational data is that the data suffers from selection bias. In other words those who receive treatment (the treatment group) are inherently different from those who don't (the control group). Heckman (in his seminal 1978 paper) shows that a naive estimate of the regression parameter on a treatment dummy (say W = 1 if an individual is treated and W = 0 if the individual is a control) suffers from an omitted variable bias. The problem arises because we only observe outcomes under a single state (either treatment or control) -- thus we have to control for factors which simultaneously affect both outcome and selection into the treatment group. Rubin and Rosenbaum (1983) pioneered the work on causal inference in the presence of selection bias. They suggest a two-step estimation procedure. In the first step the probability that an individual belongs to the treatment group is estimated. This is referred to as the individual's Propensity Score. The second step involves using the Propensity Score for pre-processing the data before estimating the Average Treatment Effect (ATE).

The use of Propensity Scores is now ubiquitous in the Causal Inference literature; however the estimation of propensity scores remains an open question. While many authors use logistic regression because of its interpretability, others argue in favor of non-parametric methods. We propose the use of machine learning classifiers (like Naive Bayes, Regression Trees and Support Vector Machines) for obtaining propensity scores. We also propose cross-validation to choose between different propensity score models. We show via theoretical arguments and simulation studies why it’s useful to consider a variety of propensity score models in the first step. Finally, I apply the method to two empirical questions. First, I propose to compare results with Dahejia and Wahba(1999) and LaLonde (1986) who estimated the average treatment effect on post-intervention earnings associated with the National Supported Work (NSW) program, a fairly small program aimed at particularly disadvantaged people in the labor market. Second, I propose to study the impact of the Janani Suraksha Yojana (JSY -- a conditional cash transfer scheme by the government of India, to incentivize women to give birth in a health facility) on maternal death rates.

Paper 3: Regularization of Covariate Balancing Propensity Scores

The third chapter combines ideas from the first two chapters. A promising new methodology to estimate propensity scores involves GMM. This method, known as Covariate Balancing Propensity Scores (CBPS), uses moment conditions from the theory of covariate balancing to estimate propensity scores via GMM (Imai and Ratkovic 2013). However the authors leave open the question of how to deal with or select from the potentially infinite moment conditions implied by theory.

The use of regularization in cases of a continuum of (or infinite) moment conditions has received substantial attention recently (Carassco et al -- 2000, 2007, 2012). There is also work on moment selection using an application of LASSO regularization, when the experimenter has information that at least a few known moments must hold (Liao 2010). In my final chapter, I propose to use these methods in the CBPS framework, thereby providing the empirical economist with a way to incorporate as much information as possible in her estimation procedure.