The Knockoff Filter for Controlling the False Discovery Rate

Emmanuel Candes, Stanford University

In many fields of science, we observe a response variable together with a large number of potential explanatory variables, and would like to be able to discover which variables are truly associated with the response. At the same time, we need to know that the false discovery rate (FDR)---the expected fraction of false discoveries among all discoveries---is not too high, in order to assure the scientist that most of the discoveries are indeed true and replicable. This talk introduces the knockoff filter, a new variable selection procedure controlling the FDR in the statistical linear model whenever there are at least as many observations as variables. This method achieves exact FDR control in finite sample settings no matter the design or covariates, the number of variables in the model, and the amplitudes of the unknown regression coefficients, and does not require any knowledge of the noise level. As the name suggests, the method operates by manufacturing knockoff variables that are cheap---their construction does not require any new data---and are designed to mimic the correlation structure found within the existing variables, in a way that allows for accurate FDR control, beyond what is possible with permutation-based methods. The method of knockoffs is very general and flexible, and can work with a broad class of test statistics. We test the method in combination with statistics from the Lasso for sparse regression, and obtain empirical results showing that the resulting method has far more power than existing selection rules when the proportion of null variables is high. We also apply the knockoff filter to HIV data with the goal of identifying those mutations associated with a form of resistance to treatment plans.

This is joint work with Rina Foygel Barber.