RUTGERS
School of Arts and Sciences

Department of
Statistics and Biostatistics

DEPARTMENT OF STATISTICS AND BIOSTATISTICS

Lucas Janson
Department of Statistics
Harvard University

Using Knockoff to find important variables with statistical guarantees

March 21, 2018
3:20 – 4:20pm
Light refreshments will be served
110 Frelinghuysen Road
Hill Center, Room 552

Abstract: Many contemporary large-scale applications, from genomics to advertising, involve linking a response of interest to a large set of potential explanatory variables in a nonlinear fashion, such as when the response is binary. Although this modeling problem has been extensively studied, it remains unclear how to effectively select important variables while controlling the fraction of false discoveries, even in high-dimensional logistic regression, not to mention general high-dimensional nonlinear models. To address such a practical problem, we propose a new framework of model-X knockoffs, which reads from a different perspective the knockoff procedure (Barber and Candès, 2015) originally designed for controlling the false discovery rate in linear models. Model-X knockoffs can deal with arbitrary (and unknown) conditional models and any dimensions, including when the number of explanatory variables p exceeds the sample size n. Our approach requires the design matrix be random (independent and identically distributed rows) with a known distribution for the explanatory variables, although we show preliminary evidence that our procedure is robust to unknown/estimated distributions. As we require no knowledge/assumptions about the conditional distribution of the response, we effectively shift the burden of knowledge from the response to the explanatory variables, in contrast to the canonical model-based approach which assumes a parametric model for the response but very little about the explanatory variables. To our knowledge, no other procedure solves the controlled variable selection problem in such generality, but in the restricted settings where competitors exist, we demonstrate the superior power of knockoffs through simulations. Finally, we apply our procedure to data from a case-control study of Crohn's disease in the United Kingdom, making twice as many discoveries as the original analysis of the same data.

Bio: I am an Assistant Professor in the Department of Statistics at Harvard University, where I develop methodology for high-dimensional inference, robust machine learning, and autonomous robotic motion planning. I am very interested in applying my work to real data, with examples so far including genetics, climate science, and health care. Prior to Harvard, I was a Ph.D. student in Stanford University's Statistics Department, where I was advised by Emmanuel Candès. In 2011, I received a B.S. in mathematics (honors thesis advised by Bala Rajaratnam) and physics, as well as a M.S. in statistics, from Stanford University.