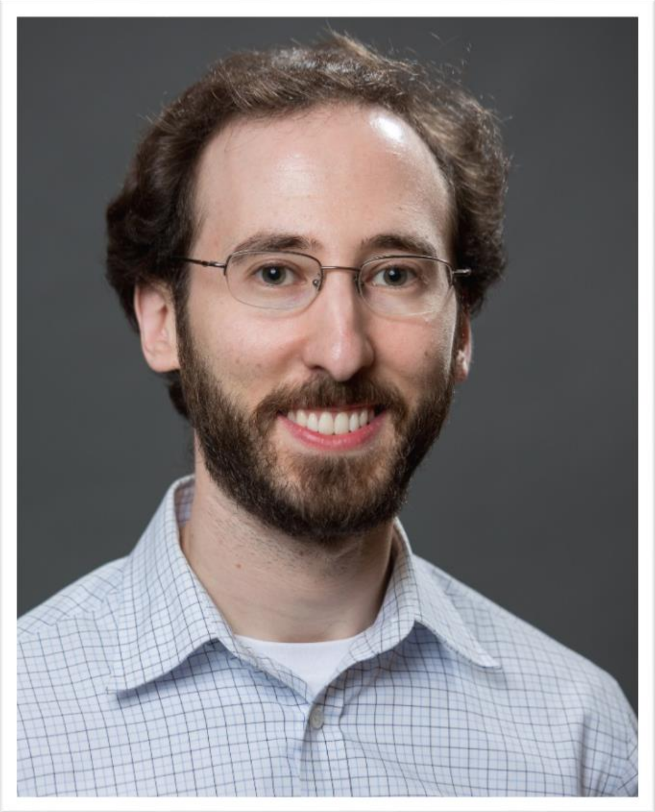
UNIVERSITY OF CALIFORNIA
UC RIVERSIDE

STATISTICS DEPARTMENT SEMINAR

Jacob Bien, Ph.D.
Assistant Professor
University of
Southern California,
Los Angeles, CA

Olmsted Hall 420
February 26th 2019
3:45-4:45pm
Reception in Olmsted 1331
at 3:15 P.M.



"HIGH-DIMENSIONAL VARIABLE SELECTION WHEN FEATURES ARE SPARSE"

Abstract

It is common in modern prediction problems for many predictor variables to be counts of rarely occurring events. This leads to design matrices in which a large number of columns are highly sparse. The challenge posed by such "rare features" has received little attention despite its prevalence in diverse areas, ranging from biology (e.g., rare species) to natural language processing (e.g., rare words). We show, both theoretically and empirically, that not explicitly accounting for the rareness of features can greatly reduce the effectiveness of an analysis. We next propose a framework for aggregating rare features into denser features in a flexible manner that creates better predictors of the response. An application to online hotel reviews demonstrates the gain in accuracy achievable by proper treatment of rare words. This is joint work with Xiaohan Yan.

Biography

Jacob Bien is an assistant professor in the Department of Data Sciences and Operations in the Marshall School of Business at the University of Southern California (USC). He received a B.S. in physics and a Ph.D. in statistics from Stanford University. Before joining USC, he was an assistant professor at Cornell University in the Department of Biological Statistics and Computational Biology and in the Department of Statistical Science. Dr. Bien's research focuses on statistical machine learning and in particular the development of novel methods that balance flexibility and interpretability for analyzing complex data. He combines ideas from convex optimization and statistics to develop methods that are of direct use to scientists and others with large datasets. Particular areas of focus include variable selection, clustering, prototype selection and the modeling of dependence in high-dimensional data. His work has been supported by the National Science Foundation, the National Institutes of Health, and the Simons Foundation. He serves as an associate editor of Biometrika and the Journal of Computational and Graphical Statistics.