Improving on ANOVA Under Strong Sparsity

Ery Arias-Castro, Emmanuel J. Cand'es and Yaniv Plan
University of California, San Diego, Stanford University and California Institute of Technology
earіasca@ucsd.edu, candes@stanford.edu, plan@acm.caltech.edu

Abstract

The basic task of testing for the significance of a subset of regression coefficients in a linear model goes back at least to the work of Fisher in the context of agricultural trials where he introduced the analysis of variance (ANOVA), which is still the most popular method by far. Assuming a standard linear model with i.i.d. Gaussian errors, ANOVA is indeed (essentially) optimal in a minimax sense when no information on the regression coefficients is available.

We study this problem under the assumption that the coefficient vector is sparse, a common situation in modern high-dimensional settings. With p denoting its dimension, assume the coefficient vector has order $p^{1-\beta}$ non-zero coordinates, where $0 < \beta < 1$. Under moderate sparsity, with $0 < \beta < 1/2$, and some conditions on the design matrix, we show that ANOVA remains essentially optimal. This is no longer the case under strong sparsity, with $1/2 < \beta < 1$, where we show that some method resembling the innovated Higher Criticism of Hall and Jin outperforms ANOVA. This is true for a variety of designs, including the classical (balanced) multi-way designs and more modern '$n < p$' designs arising in signal processing such as in basis pursuit or in compressive sensing.