Detecting and Estimating Sparse Mixtures

Jiashun Jin

Statistics Department
Stanford University

Abstract

Sparse Mixture Models have important applications in many areas, such as Signal and Image Processing, Genomics, Covert Communication, etc. In my talk, I will consider the problems of detecting and estimating sparse mixtuers.

Detection *Higher Criticism* is a statistic inspired by a multiple comparisons concept mentioned in passing by Tukey (1976) (but as a term, *Higher Criticism* is invented by a German historian Johann Eichhorn (1787)). We are able to show that the resulting *Higher Criticism statistic* is effective at resolving a very subtle testing problem: testing whether n normal means are all zero versus the alternative that a small fraction is nonzero; the subtlety of this 'sparse normal means' testing problem can be seen from work of Ingster (1999) and Jin(2002), who studied such problems in great detail. In their studies, they identified an interesting range of cases where the small fraction of nonzero means is so small that the alternative hypothesis exhibits little noticeable effect on the distribution on the p-values either for the bulk of the tests or for the few most highly significant tests. In this range, when the amplitude of nonzero means is calibrated with the fraction of nonzero means, the likelihood ratio test for a precisely-specified alternative would still succeed in separating the two hypotheses. We show that the higher criticism is successful throughout the same region of amplitude vs. sparsity where the likelihood ratio test would succeed. Since it does not require a specification of the alternative, this shows that Higher Criticism is in a sense optimally adaptive to unknown sparsity and size of the non-null effects. While our theoretical work is largely asymptotic, we provide simulations in finite samples. We also show Higher Criticism works very well over a range of non-Gaussian cases.

Estimation *False Discovery Rate* (FDR) control is a recent innovation in multiple hypothesis testing, in which one seeks to ensure that at most a certain fraction of the rejected null hypotheses correspond to false rejections (i.e. false discoveries). The FDR principle also can be used in highly multivariate estimation problems, where it has recently been shown to provide an asymptotically minimax solution to the problem of estimating a sparse mean vector in the presence of Gaussian white noise. In effect, FDR provides an effective method of setting a threshold for separating signal from noise when the signal is sparse and the noise is Gaussian.

In this talk we consider the application of FDR thresholding to non-Gaussian settings, in hopes of learning whether the good asymptotic properties of FDR thresholding as an estimation tool hold more broadly than just at the standard Gaussian model. We study sparse exponential model and sparse Poisson model, which are important models for non-Gaussian data, and have applications in many areas as well, such as Astronomy and Positron Emission Tomography (PET) etc. We show that the FDR principle also provide an asymptotically minimax solution to the problem of estimating a sparse mean vector even in the presence of exponential/Poisson noise, and in effect FDR provides an effective method of setting a threshold for separating signal from noise when the signal is sparse and the noise is exponential/Poisson.

We compare our results with work in the Gaussian setting by *Abramovich, Benjamini, Donoho, Johnstone* (2000).

Joint work with David L. Donoho.

References

[1] The Antiquity of the Books of Moses, *http://www.heraldmag.org/bookstore/booklet_antiquity.htm.*

[2] Jin, J. Detection Boundary for Sparse Mixtures, in preparation.

[3] Donoho, D. and Jin, J. (2002). Higher Criticism for Detecting Sparse Heterogeneous Mixtures. *Technical Report* , Statistics Department, Stanford University.

[4] Abramovich, F. and Benjamini, Y. and Donoho, D. and Johnstone, I. (2000). Adapting to Unkown Sparsity by Controlling the False Discovery Rate, *Technical Report* , Statistics Department, Stanford University.

[5] Donoho, D. and Jin, J. Asymptotic Minimaxity of False Discovery Rate for Sparse Exponential Data, in preparation.

[6] Donoho, D. and Jin, J. Asymptotic Minimaxity of False Discovery Rate Thresholding for Sparse Poisson Data, in preparation.