

STANFORD UNIVERSITY
DEPARTMENT OF STATISTICS
STATISTICS SEMINAR

4:15 p.m., Tuesday, January 28, 2003
Sequoia Hall Room 200
(Cookies at 3:45 in 1st Floor Lounge)

Jiashun Jin
Stanford University

Asymptotic Minimality of False Discovery Rate Thresholding for Sparse nonGaussian Data

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 – exponential and Poisson, 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 consider a vector $X_i, i = 1, \dots, n$ whose coordinates are independent exponential/poisson with individual means λ_i . The vector λ is thought to be sparse, with most coordinates 1 and a small fraction significantly larger than 1. This models a situation where most coordinates are simply ‘noise’, but a small fraction of the coordinates contain ‘signal’.

For exponential data, we develop an estimation theory working with $\log(\lambda_i)$ as the estimand, and use the per-coordinate mean-squared error in recovering $\log(\lambda_i)$ to measure risk. We consider minimax estimation over parameter spaces defined by constraints on the per-coordinate ℓ^p norm of $\log(\lambda_i)$: $Ave_i \log^p(\lambda_i) \leq \eta^p$. Members of such spaces are vectors (λ_i) which are sparsely heterogeneous.

For Poisson data, we use the per-coordinate weighted mean-squared error $\frac{(\hat{\lambda}_i - \lambda_i)^2}{\lambda_i}$ to measure risk. We consider minimax estimation over parameter spaces defined by con-

straints on the per-coordinate ℓ^p norm of $(\lambda_i - 1)$: $Ave_i(\lambda_i - 1)^p \leq \eta^p$. Members of such spaces are vectors (λ_i) which are sparsely heterogeneous.

For Poisson data, we use the per-coordinate weighted mean-squared error $\frac{(\hat{\lambda}_i - \lambda_i)^2}{\lambda_i}$ to measure risk. We consider minimax estimation over parameter spaces defined by constraints on the per-coordinate ℓ^p norm of $(\lambda_i - 1)$: $Ave_i(\lambda_i - 1)^p \leq \eta^p$. Members of such spaces are vectors (λ_i) which are sparsely heterogeneous.

We find that, for large n and small η , the FDR thresholding is nearly minimax for both the exponential and the Poisson model, increasingly so as η decreases.

We find that the FDR control parameter $0 < q < 1$ plays an important role, for both the exponential and the Poisson models: when $q \leq \frac{1}{2}$, the FDR estimator is nearly minimax, while choosing a fixed $q > \frac{1}{2}$ prevents near minimaxity. We compare our results with work in the Gaussian setting by *Abramovich, Benjamini, Donoho, Johnstone* (2000).

Joint with David L. Donoho