

STAT36900: Selected Topics in High Dimensional Inference
Lecture Note Jan 30, 2008
Class Jan 22, Jan 29

1 Overview

Four topics: large-scale multiple testing, variable selection, random matrices, and nonparametric estimation are inter-connected to each other. One way to see this is to investigate the Stein's n normal means problem, where we have a n -Gaussian vector

$$X = \mu + \epsilon, \quad \epsilon \sim N(0, \Sigma),$$

where the mean vector is unknown, with some entries being 0 and some are not.

- For large-scale multiple testing, the goal is to tell which entries are 0 and which are not.
- If we write $\Sigma = LL'$ for some matrix L (e.g. the Cholesky factorization), then the model can be viewed as

$$X^* = U\mu + \epsilon^*,$$

where $X^* = L^{-1}X$, $U = L^{-1}$, and $\epsilon^* = L^{-1}\epsilon \sim N(0, I_n)$. The large-scale multiple testing problem is then reduced into a variable selection problem.

- The topic of random matrices is also related, for in many settings we have m samples, and we hope to estimate Σ . This study of the empirical covariance matrix is thus of interest.
- The area of nonparametric estimation is related through the following setting. Suppose we observe a function $Y(t)$ over the interval $[0, 1]$,

$$Y(t) = f(t) + W(t),$$

where $f(t)$ is the function we are interested in estimating, and $W(t)$ is the Wiener process. Let $X_1, X_2, \dots, X_n$ be the first n Fourier/wavelet transform coefficients, then $X_i = \mu_i + \epsilon_i$, with μ_i being the i -th coefficient corresponding to f , and ϵ_i the i -th coefficient corresponding to $W(t)$. Then the problem of estimating f is reduced to estimating the mean vector μ in the n -normal means setting.

2 Large-scale multiple testing

The set up is that we have n pairs of component hypotheses,

$$H_{0i} \quad \text{vs.} \quad H_{1i}, \quad 1 \leq i \leq n,$$

Associated with each pair of them is a test statistic,

$$X_i, \quad 1 \leq i \leq n,$$

which can be a summary test statistics (i.e. t-test), a p -value, or a transform coefficient. We are interested in testing which hypotheses are true and which are not.

We list two examples.

- A nearly black image with 512 by 512 pixels. Each pixel contains the counts of accumulated photons. Most of the pixels are related to background only, and $X_i \sim \text{Poisson}(\lambda_0)$. Some of them are related to some bright sources, so $X_i \sim \text{Poisson}(\lambda_i)$, with $\lambda_i > \lambda_0$.

- Gene microarray. Two groups of gene microarray related to m_1 normal patients and m_2 cancer patients, each of which measures n the expression level of n genes. The measurement is on the log-odds ratio of the red over green. Roughly say, for the first group, the measurements are

$$X_{ij} = \epsilon_{ij}, \quad 1 \leq i \leq n, \quad 1 \leq j \leq m_1,$$

and for the second group, the measurements are

$$Y_{ij} = \mu_i + \eta_{ij}, \quad 1 \leq i \leq n, \quad 1 \leq j \leq m_2.$$

It is thought that only a small proportion of μ_i are nonzero (so that the corresponding genes are differentially expressed). Usually, a two-sample t-test is used for each gene, with either a pooled variance or a un-pooled variance. Denote the i -th t-score by t_i . Approximately say, when the i -th gene is not differentially expressed, $t_i \sim t_{m_1+m_2-2}$, and otherwise another distribution.

3 False Discovery Rate Thresholding

The False Discovery Rate (FDR)-thresholding is proposed by Benjamini and Hochberg in their JRSS b paper, 1995, which is also called the Sime's procedure. Suppose for each hypothesis test we have obtained a p-value,

$$p_i, \quad 1 \leq i \leq n.$$

The procedure contains 3 simple steps.

- Pick a parameter $q \in (0, 1)$ which Benjamini and Hochberg call the *FDR control parameter*.
- Sort the p -values in the ascending order,

$$p_{(1)} < p_{(2)} < \dots < p_{(n)}.$$

Determine the index k_{FDR}^q which is the largest k such that

$$p_{(k)} \leq qk/n.$$

- Reject H_{0i} if and only if

$$p_i \leq p_{(k_{FDR}^q)}.$$

The main result of FDR-thresholding procedure is that it controls the false discovery rate. The false discovery rate is the following random ratio,

$$FDR_q = \frac{\#\{\text{False Discoveries}\}}{\#\{\text{Total Discoveries}\}},$$

where we call a discover whenever a null hypothesis is rejected. The ratio is set to 0 whenever the denominator is 0. By saying the procedure controls the FDR, we mean that

$$E[FDR_q] \leq q.$$

This is the following theorem.

Theorem 3.1 *Suppose p_i 's are independent and each of them has a continuous distribution. When H_{0i} is true, $p_i \sim U(0, 1)$. Then*

$$E[FDR_q] = (1 - \epsilon_n)q \leq q.$$

For a proof see BH 95 paper in JRSS b or the paper by Ferreira and Zwiderman 06 paper in Ann. Statist..

4 Estimating the proportion of nonnull effects