Statistical Inference For Ultra-High Dimensional Data

Song Xi Chen and Yingli Qin

Department of Statistics

Iowa State University

May 2009

Introduction

- Testing significance for gene-sets rather than a single gene is a latest development in genetic data analysis
- Thus, it's interesting to consider two random samples
  - $X_{11}, \cdots, X_{1n_1} \in R^p$ and $X_{21}, \cdots, X_{2n_2} \in R^p$
  - means $\mu_1 = (\mu_{11}, \cdots, \mu_{1p})'$ and $\mu_2 = (\mu_{21}, \cdots, \mu_{2p})'$
  - covariance matrices Σ_1 and Σ_2
- We consider testing a high dimensional hypothesis

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_1 : \mu_1 \neq \mu_2$$

- When $p > n_1 + n_2 - 1$, Hotelling's T^2 test is not defined

Main Results

- Start with a general model assumption

$$X_{ij} = \Gamma_i Z_{ij} + \mu_i, \text{ for } j = 1, 2, \cdots, n_i \text{ and } i = 1 \text{ and } 2$$

where $\Gamma_i \Gamma_i' = \Sigma_i$ and $Z_{ij} \sim$IID $(0, I_m)$

- Define $T_n = \frac{\sum_{i \neq j}^{n_1} X_{1i}' X_{1j}}{n_1(n_1-1)} + \frac{\sum_{i \neq j}^{n_2} X_{2i}' X_{2j}}{n_2(n_2-1)} - 2\frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} X_{1i}' X_{2j}}{n_1 n_2}$

Theorem 1 Under some mild assumptions

$$\frac{T_n - ||\mu_1 - \mu_2||^2}{\sqrt{Var(T_n)}} \xrightarrow{\text{d}} N(0,1) \quad \text{as } p \to \infty \text{ and } n \to \infty.$$

- In **Theorem 2**, we proposed a ratio consistent estimator for $Var(T_n)$ to formulate a test procedure based on the asymptotic normality

Advantages of the proposed test

- No explicit condition required for the relationship between p and sample sizes n_1, n_2
- Not necessary to assume $\Sigma_1 = \Sigma_2$
- Do not require data are normally distributed
- More powerful than other existing parametric methods

Poster Outline

- Introduction
- Main Results
- Simulation Results
- Disscussions