Asymptotic distribution of the largest eigenvalue with application to genetic data

Chong Wu

University of Minnesota

September 30, 2016

Table of Contents

UNIVERSITY OF MINNESOTA

Gene-gene interaction

- Play an important role in many complex diseases
- Improve our understanding of the genetic regulation
- Only find few replicable human gene-gene interactions
- Reason: poor power, confounding, measurement error, etc.
- Improve power by global testing

Gene-gene interaction

- Only study binary phenotype, such as disease status
- Some notations:
    - $Y_{n\times 1}$: phenotype
    - $G_{n\times p}$: genetic marker, such as SNPs
    - $C_{n\times q}$: covariates
    - P_1, P_0: population partial correlation matrix of case ($Y = 1$) or control ($Y = 0$) conditional on C
- Global test:

$$H_0 : P_1 = P_0, \qquad H_A : P_1 \neq P_0$$

UNIVERSITY OF MINNESOTA

Gene-gene interaction

Idea:

- If $P_1 = P_0$, the largest eigenvalue of P_1 and P_0 is equal
- If we know the distribution of largest eigenvalue, we can calculate the *p*-value
- We will revisit this problem later

Population Structure

- Human originally spread many thousand years ago
- Migration and genetic drift led to genetic diversity
- Inference of population structure is an important step in genetic study

Inferring Population Structure with PCA

- PCA is the most widely used method
- Apply PCA to the genotype data and get top Principal Components (PCs)
- PCs explain difference among samples
- Top PCs often reflects genetic variation due to ancestry in the sample

Inferring Population Structure with PCA

- Before applying PCA, one often wishes to determine if the samples are from a population that has structure
- If not, PCs probably capture noise and decrease power
- We are statisticians and we use formal hypothesis testing

Inferring Population Structure with PCA

- If top PCs correspond to ”large” eigenvalues, we are expect nonrandom population structure
- Problem: how large is large

Idea

We can solve the above two problems by studying the distribution of the largest eigenvalue of a random matrix

Table of Contents

University of Minnesota

Theory

- $x_1, \ldots, x_n \sim N_p(\mu, \Sigma)$
- $X_{n \times p} = (x_1, \ldots, x_n)'$
- $A = X'X$ follows a Wishart distribution, $A \sim W_p(\Sigma, n)$
- $p = 1$, $A \sim \sigma^2 \chi^2_{(n)}$

Definition θ

Let $A \sim W_p(I, m)$ be independent of $B \sim W_p(I, n)$, where $m \geq p$. Then the largest eigenvalue θ of $(A + B)^{-1} B$ is called the *greatest root statistic* and its distribution is denoted $\theta(p, m, n)$.

Theory

- $W(p,m,n) = \log\left(\frac{\theta(p,m,n)}{1-\theta(p,m,n)}\right)$
- $(W - \mu(p,m,n))/\sigma(p,m,n) \to F_1$ (Tracy Widom)
- F_1 is asymmetric, exponential decay tail

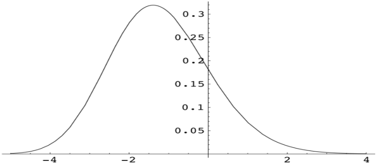

Figure: Density of the Tracy Widom distribution F_1

Theory

Theorem 1

Suppose that independent samples from two normal distribution $N_p(\mu_1, \Sigma_1)$ and $N_p(\mu_2, \Sigma_2)$ lead to covariance estimates S_i: $n_i S_i \sim W_p(n_i, \Sigma_i)$ for $i = 1, 2$. Then the largest root test of the null hypothesis $H_0 : \Sigma_1 = \Sigma_2$ is based on the largest eigenvalue θ of $(n_1 S_1 + n_2 S_2)^{-1} n_2 S_2$.

Note:

- We assume normal distribution
- $p \leq n_1$

Theory

Theorem 2 (Johnstone, 2001)

$A = X'X$ is a Wishart matrix. Let $\{\lambda_i\}_{1 \leq i \leq p}$ be the eigenvalues of A. The distribution of the largest eigenvalue λ_1 is approximately to a Tracy Widom distribution.

Note:

- We assume independent normal distribution for each X_i, $1 \leq i \leq n$
- $n/p \to \gamma \geq 1$ or $n < p$ are both large

Table of Contents

Gene-gene interaction

- Only study binary phenotype, such as disease status
- Some notations:
  - $Y_{n\times 1}$: phenotype
  - $G_{n\times p}$: genetic marker, such as SNPs
  - $C_{n\times q}$: covariates
  - P_1, P_0: population partial correlation matrix of case ($Y=1$) or control ($Y=0$) conditional on C
- Global test:

$$H_0 : P_1 = P_0, \qquad H_A : P_1 \neq P_0$$

Methods (GET)

- S^0, S^1: sample partial correlation matrix whose elements contain the correlation between each pair of genetic markers conditional on the values of the covariates C
- $n = \sum_{i=1}^{n} I(Y_i = 1)$, $d = \sum_{i=1}^{n} I(Y_i = 0)$
- By theorem 1, the largest eigenvalue of $(dS^1 + (n-d)S^0)^{-1}(n-d)S^0$ follows a Tracy Widom distribution asymptotically
- Calculate the *p*-value based on the asymptotic Tracy Widom distribution

Real Data Analysis

- Analyze GWAS data from the GLAUGEN study
- Aim: characterize genetic markers and gene-environment interactions associated with primary open-angle glaucoma
- After QC, 976 cases, $1,136$ controls and $200,432$ SNPs
- For each phenotype
  - Perform filtering and only select 100 SNPs
  - $n/p \sim 20$
  - Testing SNP-SNP interactions

Real Data Analysis

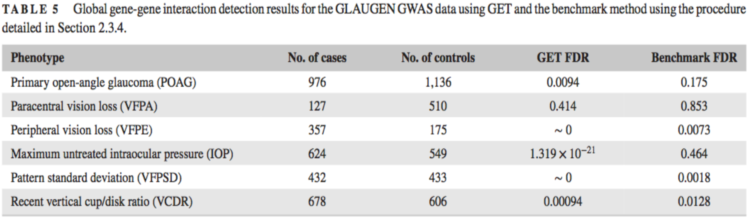

TABLE 5 Global gene-gene interaction detection results for the GLAUGEN GWAS data using GET and the benchmark method using the procedure detailed in Section 2.3.4.

Phenotype	No. of cases	No. of controls	GET FDR	Benchmark FDR
Primary open-angle glaucoma (POAG)	976	1,136	0.0094	0.175
Paracentral vision loss (VFPA)	127	510	0.414	0.853
Peripheral vision loss (VFPE)	357	175	~ 0	0.0073
Maximum untreated intraocular pressure (IOP)	624	549	1.319×10^{-21}	0.464
Pattern standard deviation (VFPSD)	432	433	~ 0	0.0018
Recent vertical cup/disk ratio (VCDR)	678	606	0.00094	0.0128

Methods (Patterson, 2006)

- By theorem 2, testing the largest eigenvalue of A is significant or not
- Linkage disequilibrium among SNPs will reduce the effective sample size, but we can adjust it

Results

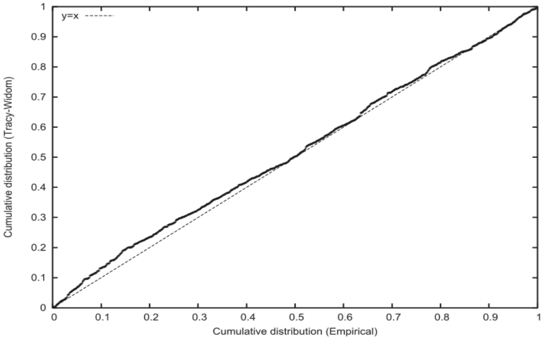

Figure: PP plot corresponding to a sample size of $n = 200$ and $p = 50,000$ markers.

Table of Contents

Normal distribution

- Genetic data (*G*) do not have the normal distribution
- For Theorem 2
    - can be applied if the the high-order moments of each cell no greater than the normal distribution (Soshnikov, 2002)
- For Theorem 1
    - We don't have any theory guarantee if *G* does not follow normal distribution

Tracy Widom approximation

- Conservative in nearly all cases
- Can be used for initial screening

Take home message

The largest eigenvalue of a matrix follows a Tracy Widom distribution asymptotically, which can be applied in genetic data analysis

Reference

- Johnstone, Iain M. ”Approximate null distribution of the largest root in multivariate analysis.” *The annals of applied statistics* 3.4 (2009): 1616.
- Johnstone, Iain M. ”On the distribution of the largest eigenvalue in principal components analysis.” *Annals of statistics* (2001): 295-327.
- Frost, H. Robert, Christopher I. Amos, and Jason H. Moore. ”A global test for genegene interactions based on random matrix theory.” *Genetic Epidemiology* (2016).
- Patterson, Nick, Alkes L. Price, and David Reich. ”Population structure and eigenanalysis.” *PLoS Genetics* 2.12 (2006): e190.