Chapter 8: Correlation Questions

$$\bar{F}(x) = \frac{\#\{z_i \geq x\}}{N}$$ if

z_i are iid

$$Var(\bar{F}(x)) = \frac{F(x)(1-F(x))}{N}$$

for

$$F(x) = \mathbb{P}(z_i \geq x)$$

Chapter 7 focuses on the dependence version

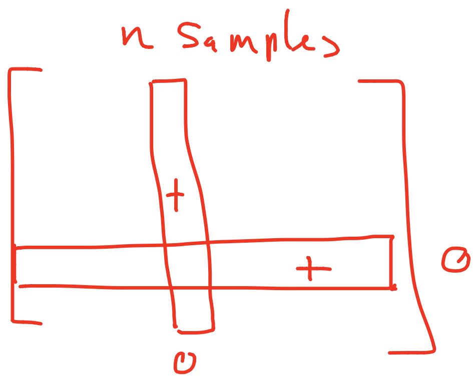
$$Var(\bar{F}(x)) \doteq \frac{\bar{F}(x)(1-\bar{F}(x))}{N} + \left\{ \frac{\hat{\alpha}\, \hat{\sigma}_0^2\, \hat{f}^{(1)}(x)}{\sqrt{2}} \right\}^2$$

For $$\alpha^2 = \frac{1}{N(N-1)} \sum_{i \neq j} \sum \rho_{ij}^2$$

average square correlation.

Row and Column Correlations

$X =$ N genes



- Assume row & column sums are zero.
- Next assume $\sum_{i=1}^{N} x_{ij}^2 = N \quad \sum_{j=1}^{n} x_{ij}^2 = n$

then the <u>row correlation</u>

$$\hat{\sigma}_{ii'} = \frac{1}{n} \sum_{j=1}^{n} x_{ij} x_{i'j}$$

next the <u>column correlation</u>

$$\hat{\Delta}_{jj'} = \frac{1}{N} \sum_{i=1}^{N} x_{ij} x_{ij'}$$

So in matrix notation

$$\hat{\Delta} = \frac{X^T X}{N} \qquad \hat{\Sigma} = \frac{X X^T}{N}$$

<u>Thrm:</u> Suppose $X = UDV^T$ $e_k = d_k^2$

then $\hat{\Delta}, \hat{\Sigma}$ have mean 0 and

empirical variance $c_2 = \sum_{k=1}^{k} \frac{e_k^2}{(Nn)^2}$

<u>pf:</u>

$$\sum \Delta_{jj'} = 1^T \underbrace{\hat{\Delta} 1}_{\text{row sums}} = 1^T 0 = 0$$

empirical variance of Δ

$$\frac{1}{n^2} \sum_{j=1}^{n} \sum_{j'=1}^{n} \Delta_{jj'}^2$$

$$= \frac{1}{n^2} \sum_{j=1}^{n} \sum_{j'=1}^{n} \Delta \odot \Delta$$

$$= \frac{1}{n^2} \operatorname{tr}(\Delta \Delta)$$

$$= \frac{1}{n^2 N^2} \operatorname{tr}\left((X^T X)^2\right)$$

$$= \frac{1}{n^2 N^2} \operatorname{tr}\left((V d^2 V^T)^2\right)$$

$$= \frac{1}{n^2 N^2} \operatorname{tr}\left((d^2)^2\right) = \frac{1}{n^2 N^2} \sum_{k=1}^{K} e_k^2$$

$$= \sum_{k=1}^{K} \left(\frac{e_k}{nN}\right)^2$$

Rmk: The standardization in some sense is decorrelating the sample-sample & gene-to gene matrix.

Cor: The main diagonal sums should be 1. So the off diagonal entries of the column correlation

$$\hat{\mu} = -\frac{1}{n-1} \qquad \hat{\alpha} = \frac{n}{n-1}\left(c_2 - \frac{1}{n-1}\right)$$

Pf: Looking at the expectation of the off diagonal.

$$\frac{1}{n(n-1)} \sum_{j \neq j'} \Delta_{jj'} = \frac{1}{n(n-1)}\left(\sum_{j \neq j'} - n\right)$$

$\Rightarrow$

$$\frac{1}{n(n-1)} \sum_{j \neq j'} \Delta_{jj'} = -\frac{1}{n-1}$$

Estimating variance

$$\alpha = \left[\frac{\sum_{i < i'} \rho_{ii'}^2}{N(N-1)\ c}\right]^{1/2}$$

$\left[\quad N(N-1)/2 \quad \right]$

$$\bar{\alpha} = \left[\frac{\sum_{i<i'} \rho_{ii'}^2}{N(N-1)/2} \right]$$

$$\hat{\alpha} = \left[\frac{n}{n-1} \left(\bar{\alpha} - \frac{1}{n-1} \right) \right]^{1/2}$$

Multivariate Normal Calculations

$$X \sim N(M, \Sigma \otimes \Delta)$$

row by row covariance

column by column covariance.

$\mathbb{E}(x_{ij}) = m_{ij}$ $\quad Cov(x_{ij}, x_{i'j'}) = \sigma_{ii'} \Delta_{jj'}$

Here we are going to assume the samples are not correlated

$\Delta = I$.

<u>Thm:</u> Under the MVN model

$s^2 = \frac{n}{\cdots}$

$$\alpha_n - \frac{\quad}{n+1}\left(\bar{\alpha}^{2} - \frac{1}{n}\right)$$ is an

unbiased estimator.

$$\bar{\alpha}^2 = \frac{1}{N(N-1)} \sum_{i \neq i'} \hat{\sigma}_{ii'}^2$$

$$\mathbb{E}(\hat{\sigma}_{ii'}^2) = \mathbb{E}\left[\left(\frac{1}{n}\sum_{j=1}^{n} x_{ij}x_{i'j}\right)\left(\frac{1}{n}\sum_{j=1}^{n} x_{ij}x_{i'j}\right)\right]$$

$$= \frac{1}{n^2}\,\mathbb{E}\left\{\sum_{j=1}^{n} x_{ij}^2 x_{i'j}^2 + \sum_{j \neq j'} x_{ij}x_{i'j}x_{ij'}x_{i'j'}\right]$$

$$= \frac{1}{n^2}\left\{n\left(1+2\sigma_{ii'}\right) + n(n-1)\left(\sigma_{ii'}^2\right)\right\}$$

$$= \frac{1}{n}\left(1+(n+1)\sigma_{ii'}^2\right)$$

Rules we used above

$$\mathbb{E}(x_1^2 x_2^2) = \sigma_1^2\sigma_2^2 + 2\sigma_{12}$$

$$\mathbb{E}(X_1 X_2 X_3 X_4) = \mathbb{E}(X_1 X_2)\mathbb{E}(X_3 X_4) + \mathbb{E}(X_1 X_3)\mathbb{E}(X_2 X_4) + \mathbb{E}(X_1 X_4)\mathbb{E}(X_2 X_3)$$

$$\mathbb{E}(\bar{\alpha}^2) = \frac{1}{N(N-1)} \sum_{i \neq i'} \left\{ \frac{1}{n} \left(1 + (n+1)\theta_{ii'}\right) \right\}$$

$$= \frac{1}{n} + \frac{n+1}{n} \frac{1}{N(N-1)} \sum_{i \neq i'} \theta_{ii'}^2$$

So

$$\mathbb{E}\left(\bar{\alpha}^2 - \frac{1}{n}\right) = \frac{n+1}{n} \underbrace{\frac{1}{N(N-1)} \sum_{i \neq i'} \theta_{ii'}^2}_{\alpha^2}$$

and

$$\frac{n}{n+1}\left\{\bar{\alpha}^2 - \frac{1}{n}\right\}$$ is unbiased.

Thrm: Under the model the column covariance estimator $\hat{\Delta} = \frac{X^T X}{N}$ has

mean and covariance $\hat{\Delta} \sim \left(\Delta, \underbrace{\frac{\Delta^{(2)}}{N_{eff}}}_{\text{tensor}}\right)$

$$N_{eff} = \frac{N}{1+(N-1)\alpha^2}$$

So what is effective sample size? N_{eff}.

$\alpha = 0.2, \quad N_{eff} = 5$

So even slightly correlated genes

$$\Delta^{(2)}_{jk,\ell n} = \Delta_{j\ell} \Delta_{kn} + \Delta_{k\ell} \Delta_{\ell n}$$

Chapter 9 Enrichment Analysis

- Used to compensate the lack

of power.

- Leverage biological data (from Gene Ontology usually)
- Let S is a specific gene set category.
- denote $Z_S = \{z_i : i \in S\}$
- Testing based on KS test

$H_0: F^{(1)}(x) = F^{(2)}(x)$

test stat: $\sup_x |\hat{F}_1(x) - \hat{F}_2(x)|$

$F_1(x)$ CDF of z score on S

$F_2(x)$ ——〃—— S^c