Gene Set Analysis II: self contained

When we want to consider a group of *p* genes together

- Test whether the mean vector μ differs from a reference level
- Test whether the mean vector differs between two groups

Hotelling's ***T*-squared statistic**

- When we had one random variable $X \sim N(\mu, \sigma^2)$ $\bar{X} \sim N(\mu, \sigma^2/n)$
  - With n iid observations, we have $z = \sqrt{n}(\bar{x} - \mu)/\sigma$
  - If we have to estimate the variance, we have $t = \sqrt{n}(\bar{x} - \mu)/s$
- The Multivariate version: X is multivariate normal with mean vector μ of length *p* and covariance matrix $\sum_{(pxp)}$
  - If we know ∑ $n(\bar{x} - \mu)'\Sigma^{-1}(\bar{x} - \mu) \sim \chi^2_p$
  - IF we have to estimate ∑ $S = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'$

We get the multivariate version in quadratic form

$$t^2 = n(\bar{x} - \mu)'\Sigma^{-1}(\bar{x} - \mu)$$

Hotelling's two-sample *T*-squared statistic

- A pooled estimate of the covariance

$$S = \frac{1}{n_x + n_y - 2}\{\sum_{i=1}^{n_x} (x_i - \bar{x})(x_i - \bar{x})' + \sum_{i=1}^{n_y} (y_i - \bar{y})(y_i - \bar{y})'\}$$

$$t^2 = \frac{(\bar{x} - \bar{y})' S^{-1} (\bar{x} - \bar{y})}{1/n_x + 1/n_y} \sim T^2(p, n_x + n_y - 2)$$

- If p<<n (a few genes, lots of samples), this is classical multivariate analysis.
- If p<n-1 this is OK
- If p>=n-1 you run into singular S

Remedy for S when p>n

- Strong assumption on S
  - Making S Diagonal
  - Regularize S by adding a small constant to diagonal
  - Diagonal S with shrinkage (recall shrunken centroid)
- Dimension reduction

- Decompose $S=UDU'$
  - Where D is diagonal with the eigenvalues
  - U is an orthogonal matrix
- Compute transformed data

  $Y^*=D^{-1/2}U'Y$

By keeping only the first few principal components, we deal with a lower dimension diagonal matrix for within-group covariance.

A multivariate approach for integrating genome-wide expression data and biological knowledge. Kong et al 2006, Bioinformatics

How do you generalize this?

- How did we generalize from a two sample t-test, to testing for a contrast after adjusting for other covariates(confounders)?
- The two sample t-test is a special case of simple linear regression, where the regressor takes only values 0 and 1

$$Y=\mu + \delta x$$

- If you had to adjust for other covariates, you would do multiple regression

$$Y = \mu + \delta x_1 + x_2^{'}\beta + \varepsilon$$

- you'd be comparing two models

$$Y = \mu + \delta x_1 + x_2^{'}\beta + \varepsilon \qquad\qquad Y = \mu + x_2^{'}\beta + \varepsilon$$

And see if the reduction in RSS is worthwhile (by adding the x_1 in your model)

The multivariate version--GlobalANCOVA

- Example of p genes measured in n samples, with d covariates, one you are interested in and two you want to adjust for

Table 1. Design matrix for a simple two-group setting with adjustment for sex (1—male; 0—female) and location (1—colon; 0—rectum)

Samples	S1	S2	S3	S4	S5	S6	S7	S8
Gene (*i*) specific mean	1	1	1	1	1	1	1	1
Group	0	0	0	0	1	1	1	1
Sex	1	1	0	0	0	0	1	1
Localization	1	0	1	0	1	0	1	0

- N=8; d=3

Linear model for each gene i

$$\tilde{x}^{(i)} = (x_1^i, \ldots, x_n^i)^t.$$

$$\tilde{x}^{(i)} = \tilde{m}^{(i)} + \tilde{\xi}^{(i)} = \begin{pmatrix} 1 & c_{11} & \cdots & c_{1d} \\ \vdots & \vdots & & \vdots \\ 1 & c_{n1} & \cdots & c_{nd} \end{pmatrix} \cdot \begin{pmatrix} \beta_{i0} \\ \vdots \\ \beta_{id} \end{pmatrix} + \begin{pmatrix} \xi_1^i \\ \vdots \\ \xi_n^i \end{pmatrix}$$

$$= C \cdot \tilde{\beta}_i^t + \tilde{\xi}^{(i)}$$

Full model(FM) versus Reduced Model (RM)

$$[C_0, C_1]$$

$$E(\tilde{x}_{\mathrm{FM}}^{(i)}) = [C_0, C_1](\tilde{\beta}_{i,0}, \tilde{\beta}_{i,1})^t$$

$$C_0$$

$$E(\tilde{x}_{\mathrm{RM}}^{(i)}) = C_0 \tilde{\beta}_{i,0}$$

Stack all genes together

$$\tilde{X} = \begin{pmatrix} \tilde{x}^{(1)} \\ \vdots \\ \tilde{x}^{(p)} \end{pmatrix} = \begin{pmatrix} C & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & C \end{pmatrix} \cdot \begin{pmatrix} \tilde{\beta}_1^t \\ \vdots \\ \tilde{\beta}_p^t \end{pmatrix} + \begin{pmatrix} \tilde{\xi}^{(1)} \\ \vdots \\ \tilde{\xi}^{(p)} \end{pmatrix} = \tilde{C} \cdot \tilde{\beta} + \Xi,$$

$\tilde{X}$ is an np column vector

$\tilde{C}$ is an $(np)\times[(d+1)p]$

$$\begin{aligned} \tilde{R} &= (\Delta_{np} - \tilde{C}(\tilde{C}^t\tilde{C})^{-1}\tilde{C}^t)\tilde{X} = (\Delta_{np} - \tilde{H})\tilde{X} \\ \mathrm{RSS} &= \tilde{R}^t \cdot \tilde{R} = \tilde{X}^t(\Delta_{np} - \tilde{H})\tilde{X}. \end{aligned}$$

$$F_{GA} = \frac{\mathrm{RSS_{RM}} - \mathrm{RSS_{FM}}}{\mathrm{RSS_{FM}}} \cdot \frac{n-q}{f}$$

- If the random component is homoscedastic

$$\Xi \sim N(0, \sigma^2 \Delta_{np})$$

This is like your multiple regression.

follows F distribution

$$F_{GA} = \frac{\mathrm{RSS_{RM}} - \mathrm{RSS_{FM}}}{\mathrm{RSS_{FM}}} \cdot \frac{n-q}{f}$$

(f is the number of parameters of interest, i.e., the difference between full and reduced models)

For p genes, n samples and q parameters in the full model, q-1 parameters in the reduced (null) model, the degrees of freedom are

p, p(n-q)

- In general, however, the genes are correlated though the samples are usually considered as independent.
- $\tilde{\Sigma}_{\text{genes}}$ is not diagonal
- We can still compute the statistic

$$F_{GA} = \frac{\text{RSS}_{\text{RM}} - \text{RSS}_{\text{FM}}}{\text{RSS}_{\text{FM}}} \cdot \frac{n-q}{f}$$

- But how do we know its behavior under null?

Permutation based p-value

- Permute the rows of C_1
- Keep $C_{0,}$ so the covariate structure if perserved
- In computation, use the residuals from the reduced model on original data, and fit these residuals to $[C_0, C_1^b]$
- Repeat for B times, generate a null distribution of the statistic

- Pro: easy in concept, does not require asymtotic theory
- Con: same limitation we mentioned in other permutation

- What about asymptotic distribution?
  - Do we know it?
  - How well does it approximate?

$$\mathrm{RSS}_{\mathrm{effect}} = \mathrm{RSS}_{\mathrm{RM}} - \mathrm{RSS}_{\mathrm{FM}} = \sum_{i=1}^{np} \pi_i \cdot \chi^2_{1,i}$$

- A weighted sum of chi-squared variables
- The weights depend on two parts:
  - n Eigenvalues of $\tilde{H}_{\mathrm{FM}} - \tilde{H}_{\mathrm{RM}}$ where H stands for hat matrix
  - p Eigenvalues of $\tilde{\Sigma}_{\mathrm{genes}}$.
  - Same challenge again: that p may be large.
  - Regularize the estimate of $\tilde{\Sigma}_{\mathrm{genes}}$.

$$\mathrm{RSS}_{\mathrm{effect}} = \mathrm{RSS}_{\mathrm{RM}} - \mathrm{RSS}_{\mathrm{FM}} = \sum_{i=1}^{np} \pi_i \cdot \chi^2_{1,i}$$

- A weighted sum of chi-squared variables
- The weights depend on two parts:
  - n Eigenvalues of $\tilde{H}_{\mathrm{FM}} - \tilde{H}_{\mathrm{RM}}$ where H stands for hat matrix
  - p Eigenvalues of $\tilde{\Sigma}_{\mathrm{genes}}$.
  - Same challenge again: that p may be large.
  - Regularize the estimate of $\tilde{\Sigma}_{\mathrm{genes}}$.

Table 3. False positive fraction (top) and power (bottom) at $\alpha = 5\%$

Scenario	Level permutation	Level asymptotic
S1-30genes	0.055	0.062
S1-200genes	0.046	0.069
S2-30genes	0.046	0.059
S2-200genes	0.057	0.101
S4-30genes	0.049	0.073
S4-200genes	0.049	0.070
Scenario	**Power permutation**	**Power asymptotic**
S3-30genes	0.204	0.283
S3-200genes	0.121	0.225
S4-30genes	0.366	0.425
S4-200genes	0.647	0.732

Questions for you

- What would you do?
- Is there anything else you want to see before you make a decision?