Statistical Methods for Omics Data

Agus Salim, Ph.D

Big Data Analytics: Application to Modern Genetics
Universitas Jember, 23 Dec 2015
A big problem with small sample

- It all begins with a small sample size....
- Omics technology and data generation are expensive; most researchers can only afford small sample size (less than 10 is common).
- Why do we need to worry?
- For a start, there is often not enough variability in small sample.
Underestimation of SD: Example

- We simulate data for 10000 genes, each with $n = 3$ subjects
- Data for each gene is Normally-distributed with $\mu = 0, \sigma = 1$
- This is the boxplot of log-SD across 10000 genes. What can you see?
Now, see what happen when we increase the number of subjects $n$

There are much less extreme SD (especially very small ones) when $n$ is large.
Underestimation of SD: Impact

- Extremely small SD can give rise to falsely large test-statistic
- Example: with one-sample t-test, we reject $H_0$ if test-stat $\geq c$, where $c$ is a percentile from theoretical distribution.

$$\frac{\sqrt{n}(\bar{x} - \mu_0)}{SD_x}$$

- Hence, small SD will lead to increasing chance of rejecting $H_0$ unnecessarily (increasing FDR).
One way to stabilize the SD would be by putting prior distribution on the (function of) variance parameters.

Let us denote $\sigma^2_g$ as the variance for gene $g$, Smyth (2003) assumed prior information on $\frac{1}{\sigma^2}$ using scaled $\chi^2$ distribution,

$$\frac{1}{\sigma^2_g} \sim \frac{1}{d_0 s_0^2} \chi^2_{d_0}$$
So that the posterior mean of $\sigma_g^2$ is given by

$$\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}$$

where $s_g^2$ is the sample variance for gene $g$ and $d_g$ is the associated degree of freedom.

We can see that if $d_0 = 0$ then this 'moderated' variance will be the same as the sample variance. However when $d_0 \neq 0$, the variance estimate will be moderated, especially if sample variance is small relative to $s_0^2$. 
First, we need to estimate $d_0$ and $s_0^2$ empirically from the data; Smyth (2003) developed an Empirical Bayesian approach for this where the hyper-parameters are estimated using Method of Moment (MoM).

Given the estimates of $d_0$ and $s_0^2$, the moderated t-test statistic is given by,

$$\frac{\sqrt{n}(\bar{x}_g - \mu_0)}{\sqrt{\tilde{s}^2_g}}$$

We will try this approach during the next Lab session.
Problem with small sample size: Regression

- Let us assume we have \( n \) subjects and \( p \) genes and the data are stored in \( X(n \times p) \) matrix, \( n \ll p \) and the outcome of interest is \( y(n \times 1) \).

- Suppose we want to regress \( y \) with gene expression data as covariate,

\[
y = X\beta + \epsilon
\]

- Usually, we would use OLS to estimate \( \beta \),

\[
\hat{\beta}_{OLS} = (X'X)^{-1}X'y
\]

- But since \( n \ll p \), we cannot obtain the estimate as \( (X'X) \) is not of full rank and the inverse does not exist.
Suppose that $y$ is binary with two class (0/1) and we want to classify subjects based on their gene expression data (very popular!)

Usually, we will perform either logistic regression or linear discriminant analysis (LDA)

Let’s proceed with LDA this time

Remember, we want to be able to classify well, so looking to find linear function of $X$, $Xa$ that maximizes the between-class variance $B$ relative to the within-class variance $W$.

$$\max \frac{a'Ba}{a'Wa}$$
It can be shown that the linear combination vector $a$ is given by the first eigenvector of $W^{-1}B$ matrix.

But if $n \ll p$, $W^{-1}$ does not exist!
Penalized Regression for small sample size

- When we do regression, we minimize residual sum of square (RSS) to estimate $\beta$

$$RSS(\beta) = (y - X\beta)'(y - X\beta)$$

- For reasons outlined before, this minimization is not possible when $n \ll p$, as the number of parameters to be estimated will be more than the sample size.

- However, we can seek to minimize the penalized RSS instead,

$$pRSS(\beta) = (y - X\beta)'(y - X\beta) + \lambda \sum_{i=1}^{p} | \beta_j |$$
Penalized Regression for small sample size

- It turns out that the pRSS above is equivalent to assuming prior distribution on $\beta_j$
- Park and Casella (2008) showed that the prior distribution for $\beta_j$ in this case is a Laplace distribution with parameter $\lambda$.

$$\beta_j \sim \text{Laplace}(\lambda)$$
Penalized Regression for small sample size

- The intuitive idea: for large $\lambda$, some of the $\beta_j$ will be forced to zero, hence reducing the number of parameters to be estimated.

- If small size is moderate, other prior (e.g., Normal) can also be used which will lead to ridge regression

$$pRSS(\beta) = (y - X\beta)'(y - X\beta) + \lambda \sum_{i=1}^{p} \beta_j^2$$

- The optimal $\lambda$ is usually selected via cross-validation
Other issues: Overfitting

- Model built using small sample is more prone to overfitting and contain higher amount of optimism when used for making prediction.
- Cross-validation must always be carried out to gauge the amount of optimism.
Other issues: normality assumption & permutation-based tests

- With small sample, normality assumption is not going to be valid as Central Limit Theorem (CLT) approximation will be off.
- If (computationally) feasible, permutation-based tests must be carried out.
- The idea behind permutation-based is very simple:
  1. Compute the test-statistic using the observed dataset
  2. Permute the data structure as if $H_0$ is true
  3. Compute the test-statistic for this permuted dataset
  4. Repeat 2-3 $B$ number of times ($B$ should be at least 10000)
  5. Compute the p-value as the proportion of test-statistics from permuted datasets that are at least as large as the observed test-statistic
Some Useful References

