Mathematical Ideas and Challenges for Continuous Buhmbox

Alexandre Lamy - all2187@columbia.edu

October 23, 2017

Recall - Normal Buhmbox

To detect heterogeneity within D_A cases driven by a subgroup that is genetically similar to D_B cases, Buhmbox gets as input genotype data for D_A cases and controls and looks at D_B SNPs. So we have:

- N is the number of cases given.
- N' is the number of controls given.
- R and R' are the sample correlation matrices of the D_B SNPs for the case and control datasets.
- p_i is the risk allele frequency.
- γ_i is the OR for SNP i, i.e. $\gamma_i = \frac{(p_i^+)(1-p_i^+)}{(p_i^-)(1-p_i^-)}$ where p_i^+ is RAF in cases and p_i^- is RAF in controls.

Then we have

$$S_{BUHMBOX} = \frac{\sum_{i<j} w_{ij} y_{ij}}{\sqrt{\sum_{i<j} w_{ij}^2}}$$

with

$$Y = \sqrt{\frac{NN'}{N+N'}}(R - R')$$

and

$$w_{ij} = \frac{\sqrt{p_i(1-p_i)p_j(1-p_j)}(\gamma_i - 1)(\gamma_j - 1)}{((\gamma_i - 1)p_i + 1)((\gamma_j - 1)p_j + 1)}$$

Generalization to continuous phenotypes - no cases and controls

There is a nice extension of the Pearson Correlation Coefficient for weighted observations:
Define the weighted mean as:

$$m(x, w) = \frac{\sum_i w_i x_i}{\sum_i w_i}$$

The weighted covariance as:

$$cov(x, y, w) = \frac{\sum_i w_i (x_i - m(x, w))(y_i - m(y, w))}{\sum_i w_i}$$

Finally we have that the weighted correlation is given by:

$$corr(x, y, w) = \frac{cov(x, y, w)}{\sqrt{cov(x, x, w) cov(y, y, w)}}$$

However this confines us to using positive weights. But this might make sense as argued after last meeting (heterogeneity is not symmetric).

Now ideas/challenges on how to replace/change different parts of the Buhmbox equation:

- $R - R' : R'$ is theoretically the identity matrix (which is irrelevant in the calculation) and R could be changed to weighted correlation matrix where weights are given by some increasing function f on the phenotype. If we do want to get an R', perhaps calculate weighted correlation with flipped weights?
- $\sqrt{\frac{NN'}{N+N'}}$: This term is a little weird. Originally in the Buhmbox article they have $Y = \sqrt{N}(R - P)$ where P is the identity (null hypothesis). In this case $\sqrt{N}$ is to normalize values in Y to be $N(0, 1)$ since sample correlation has variance $\approx \frac{1-\rho^2}{N-2}$. $\sqrt{\frac{NN'}{N+N'}}$ is of similar size... Perhaps we could just use $\sqrt{N}$ or some sort of function of weights used (even though they are normalized)... Seems very hard to find variance of weighted correlation estimator.
- γ_i could estimate p_i^+ by using an increasing weight scheme and p_i^- with a decreasing weighting scheme... But what scheme? In general do we want some sort of exponential, or a tanh.