Talk on 2011-09-09

# 1 Warfarin Data

## 1.1 Data description

For this data set, the sample size is 304. Two phenotypes are measured, labeled as ydose and yinr, respectively. Also, some covariant are recorded: sex, age, height and weight. The genotypes for two genes are provided. Gene A contains 7 SNPs and gene B contains a 4-allelic marker.

## 1.2 Data Analysis

Two Trait-Gene Similarity regression models with 3 tests are performed for this data set. The first model contains the Similarity between 2 genes and the similarity for the interaction, as described below:

$$Z_{ij} = b^A S_{ij}^A + b^B S_{ij}^B + b^{AB} S_{ij}^{AB} + e_{ij}$$

where $Z_{ij}$ is the normalized trait similarity between individual $i$ and $j$, $S_{ij}^A$ and $S_{ij}^B$ record the gene similarity between individual $i$ and $j$ for gene A and gene B respectively. $S_{ij}^{AB}$ describes the genetic similarity of the interaction between gene A and gene B. $b^A$, $b^B$ and $b^{AB}$ are the effects of $S_{ij}^A$, $S_{ij}^B$ and $S_{ij}^{AB}$, respectively.

Q: what is "normalized trait similarity"?
A: $Z_{ij} = (Y_i - \mu_i)^T (Y_j - \mu_j)$ where $\mu_i = E(Y_i | X_i, G_i) = X_i \gamma$

Since the data is a little different from the usual 2-allelic marker data. The genetic similarity matrix for a gene is computed in the following way:

Define $M$ is number of the SNPs the gene contains. For $ith$ individual's $mth$ marker, two SNPs are recorded, denoted as $G_{i1}^m$ and $G_{i2}^m$, then we have the genetic similarity $s_{ij}^m$ between individual $i$ and $j$ for the $mth$ marker in a particular gene:

$$s_{ij}^m = \begin{cases} 1, & G_{i1}^m = G_{i2}^m = G_{j1}^m = G_{j2}^m \\ 0, & G_{i1}^m \neq G_{j1}^m \text{ and } G_{i1}^m \neq G_{j2}^m \text{ and } G_{i2}^m \neq G_{j1}^m \text{ and } G_{i2}^m \neq G_{j2}^m \\ 0.5, & o.w \end{cases}$$

and the similarity of the gene $S_{ij}$ is

$$S_{ij} = \frac{1}{M} \sum_{m=1}^{M} s_{ij}^m$$

and $S_{ij}^{AB} = S_{ij}^A \times S_{ij}^B$.

Q: I'm puzzled with why it will become more complicated than biallelic case....
I assume that you are using UNWEIGHTED maxIBS. Then the score is simple, if the $G_i = G_j$, then score "1", if $G_i$ and $G_j$ only have one allele the same, the score "0.5", otherwise "0". The rule is the same as before.... (If you put weight by allele frequency, that will be more complicated...)
A: No, actually I am using Unweighted "Typical IBS". I think the reason the 4-allelic case seems to be more complicated than biallelic is that for biallelic there are only 3 possibilities while for 4-allelic case there are 10 possibilities which make "only have one allele the same" hard to define.

There are two tests under this model. One the joint test under the hypothesis: $b^A = b^B = b^{AB} = 0$, the other one is the interaction test under the hypothesis $b^{AB} = 0$.

Other model for this data set is shown below:

$$Z_{ij} = b^A S_{ij}^A + b^B S_{ij}^B + e_{ij}$$

In this model, we just consider the main effect of the gene similarity. The goal of this model is to test $b^A = 0$ with the adjustment of Gene B.

Table 1: Our method result

| | | | One Gene Test | |
|---|---|---|---|---|
| Traits | Joint test | Epi test | Gene A | Gene B |
| ydose | $3.15 \times 10^{-21}$ | 1 | $2.6 \times 10^{-22}$ | $3.38 \times 10^{-8}$ |
| yinr | 0.41 | 0.54 | 0.46 | 0.59 |

Table 2: PCA and PLS result

| | PCA | | PLS | |
|---|---|---|---|---|
| Traits | Joint test | Epi test | Joint test | Epi test |
| ydose | $2.2 \times 10^{-16}$ | 0.6495 | $2.2 \times 10^{-16}$ | 0.312 |
| yinr | 0.3118 | 0.47 | 0.2113 | 0.117 |

This table shows that for either trait, the epi test is not significant. However, trait ydose has a very significant in joint test, and One gene test for both gene A and gene B. Can we use it as an application of One gene test?

## 2 Debug

There is a weird problem I met when I debug the code for Epi test. Although now I directly use the similarity matrix $S_A$ and $S_B$ to estimate $\tau_A$ and $\tau_B$ through EM algorithm, which should get rid of the haplotype $H$ and the variance matrix $R$ for the haplotype effect, I still use the way you used in your one gene Trait-Similarity regression model to compute a fake $H$ and $R$ to facilitate the computation for $V^{-1}$. The $H_B$ and $R_B$ from $S_B$ can not fully satisfy the restriction that $S_B = H_B R_B H_B^T$, I think that is caused by the numerical calculation error. The weird part is, if I directly use $S_B$ to estimate $tau_B$ in the loop, $\tau_B$ can not converge, however, if I compute $S_B^* = H_B R_B H_B^T$ and replace $S_B$ by $S_B^*$ in the loop, everything is fine. I don't know whether it is big question since $S_B^*$ is quite close to $S_B$. But it is better to point out.

Q1: Does this problem only occur with GxG test or also for one G test conditional on another G effect? And this only occurs with this dataset or also occurred in simulation?
A1: No, Both for the One Gene test and the simulation, everything is fine.
Q2: Can you show me how you get $\tau_B$ directly? Wouldn't the dimension too big to calculate $V^{-1}$?
A2:
Q3: To resolve this, you may want to do a small scale of simulation to make sure the type I error and estimate of tau B are ok when you use fake H and fake R.
A3: I will double check that.