

Debug process 2011-10-02

The decomposition of S matrix

In previous versions of my code, I used to decompose the S matrix into H and R where $S = H^T R H$, but we apply this trick, the generated H and R cannot fully satisfy this standard, i.e., there is a difference between $S^* = H^T R H$ and S . In most cases, the difference is not big (around 10^{-14}), but in Alison's Warfarin data, the difference can range from -0.34 to 0.4 which is really huge. Based on this observation, I delete all the code related to H and R , (H and R are used in the calculation for the conditional expectation and variance, and to facilitating the computation for V^{-1}) and replace them with the code using S directly. I have double checked the results when the difference between S^* and S is small, the result using H, R are almost identical to the result using S directly, but the later one is slower, since the computation for V^{-1} is proportional to the sample size.

BTW, I have a stupid question about the S . Do I have to force $S_{i,i} = 1$ since the genotype of the same people is always the same? But when I do so, the rank of S becomes very big so that I guess I should not do so.

The EM algorithm

After the correction for the S matrix decomposition, I tried to rerun the Gain data. For previous Gain data analysis, I just ran the joint test for all possible genes combinations and the epistasis test for the genes combinations which has a significant p value for joint test. Since there are only 3 genes combinations have a significant p value, I ran the Epitasis test for 3 times. But this time I want to run the One Gene test for each gene combination so I found another problem for the EM algorithm.

In the One Gene test, if I want to test the hypothesis $\tau_A = 0$, I need to use the EM algorithm to estimate τ_B and σ first. I found out the in some cases, the estimation for τ_B can not converge, it seems that it want to converge to 0 (I have checked the estimated τ_B for each iteration for one particular case and found that $\hat{\tau}_B$ drops from 1.5 to 0.004 after 6,000 iterations, then I tried another genes combination and similar things happened.) Then I wondered whether my derivation or coding is wrong, but I checked them for three times and everything is correct. Finally, I ran the simulation again and the result shows me that the EM algorithm works perfectly for the simulated data. Even I set both τ_A and τ_B to 0, the estimations will converge at around 0.02.

I think for the One Gene test, it is similar to your $G \times E$ test, so I wonder whether you met such problem before and how you solve it. (I read the $G \times E$ paper again and I think there may be some typo in the appendix.) Also, since the simulation study

shows that if both of τ_A and τ_B are 0, the estimations are around 0.02, I can just set τ_B to zero if the estimation for τ_B is smaller than 0.01?