Disease model identification methods based on maximum test and performance analysis

Yaling Yin

1Information School, Xi’an University of Finance and Economics, Xi’an 710100, China.

Corresponding author’s e-mail: zhanglong@eurasia.edu

Abstract. Combinatorial explosion and computational burden are always the challenges for genome-wide association study. In order to reduce the computation cost, many multi-stage methods were put forward to identify the disease models. However, one-way and two-way disease models always can be detected to leave out some SNPs for non-significance. And these SNPs are combined with other SNPs to get higher disease models. In this paper, three test statistics, Max G-test, Max Entropy Difference and Max Relative Entropy, had been presented for the first stage to detection disease models with main effect and without main effect. Five testing methods were used for examining multiply simulation datasets and real dataset. Results were revealed that Max Entropy Difference test is the best method of recognition in five filtering methods with main-effect and max-statistic test is just right method to identify model without main-effect. Results also were showed that five statistics can get interest power for two-ways on simulation datasets and real dataset. We believe that these statistics can find strong and weak SNPs for next step in computationally and statistically.

1. Introduction
A large body of genetic studies shows that the mapping relationship from genotype to phenotype is very complex [1-4]. And as one of main reasons in genome wide case-control analysis studies(GWAS), gene-gene association is as a pinpoint to uncover this complex relationship. In a vast amount literature, gene-gene association including gene-gene interactions [5-8] was defined interaction between multiple gene loci. Many methods especially statistical and machine learning methods, were proposed to detect the gene-gene associations. But computational burden, missing heritability, no prior biological knowledge and linkage disequilibrium are not uniquely estimable, so many approaches have been identified by multi-step filtering method.

Marchini[9] used FIM to find two loci interaction with three-step filter, is one of the classic interaction methods. But FIM and its improved methods have used logistic regression framework to find the interaction models. Especially, in the first filter stage, many SNPs in high order model cannot be searched for lower main effect. MDR [10] is also a classic method that can relatively quickly detect the interaction model but it just does two-way SNPs interaction and had a bad effect on the other simulated datasets. Another classic method is BEAM [11]. BEAM and its subsequent methods [12,13] always used Bayesian inference to detect the two-way interaction models, but their time complexity is very poor and SNP counts can't be too large for 10e5. Both methods need a pre-screening process to reduce the number of SNPs in order to analyse the large data sets.

MECPM [14] is one of the only method that can detect high-order interaction models. And in this method, the authors used maximum entropy to judge the interaction models to candidate SNPs set.
BIC was also used to evaluate the interaction models. But MECPM also have their disadvantage and authors divide SNP into many blocks. There are 500 SNPs in one block, and searching space for detection cannot also be too larger. Team [15] and other classic methods [16-21] can only detect two-way interaction, and cannot detect higher-order interaction models. And we think that detection two and more than two-way interaction models can help to uncover complex causing of diseases. For detection high-order interaction models and increasing search space for SNP counts, we should use multiply filter step method. However classic filter methods often lost a lot of information [21], for example, in one-way filter step, many SNPs can be remove for without or weak main effect. But these SNPs may be so strong joint effect with other SNPs. A similar situation also happened on the two-way or third-way interaction models. So we think that methods for the first filter step is also one of the very important factors to improve efficient. We conclude many classic methods that find chi-square test, entropy and likelihood ratio are key evaluation criteria. We summarize these evaluation criteria, and used maximum relative entropy (MAXRE), maximum entropy difference (MAXDE), maximum likelihood test (MAXGtest), maximum statistical method (MAXtest) [22-24], chi-test to evaluation the SNPs datasets. In simulation datasets and real dataset, we can see that MAXED is better in some simulate datasets for worse joint effect and MAXtest can get better effect in main effect datasets.

2. Methods

2.1. Notation and model

The underlying assumption of genome-wide association analysis is that there are significant differences in the control samples and case disease samples. Indicates the genotype counts for cases by and for controls by, the three genotype counts across the recessive, additive and dominant model denoted by Table 1. For a Single Nucleotide Polymorphism(SNP), it's genotype is, which and is express no genetic mutation and genetic mutation, respectively. For case/control samples and are expressed each genotype counts in case/control respectively. N is sample counts. Table 1 is the contingency table of three models for one SNP. And based on table 1, we expressed the three statistics and Specifies that is equal to 0.5.

| Dominant model | Recessive model | Additive model | Total |
|----------------|-----------------|---------------|-------|
| (A/a)+(a/a)    | A/A             | (A/A)+(1-μ)A/a| (A/A)+(1-μ)A/a |
| Case           | p_1+ p_2       | p_1           | p_1+p_2 | p_1+(1-μ)p_2 | μp_2+p_1 |
| Control        | q_1+ q_2       | q_1           | q_1+q_2 | q_1+(1-μ)q_2 | μq_2+q_1 |
| Total          | n_1+ n_2       | n_1           | n_1+n_2 | n_1+(1-μ)n_2 | μn_2+n_1 |

2.2. MAXtest and scores

To obtain the Max-test, we used the Chi-square test to describe three models. Through test we found that Cochran-Armitage Trend test, Pearson’s test and Fisher test are all with similar results for detection disease models, and it is consistent with Gang Zhang’s conclusions [25]. So we use Pearson’s test to describe the models. Pearson’s test, denoted \( \chi^2 \), is given by

\[
\chi^2 = \sum_{i=1}^{2} \left( \frac{(p_i-n_iP/N)^2}{n_iP/N} + \sum_{j=1}^{2} \left( \frac{(q_i-n_iQ/N)^2}{n_iQ/N} \right) \right)
\]

Our null hypothesis in disease model recognition is that the model is related to disease, and under, asymptotically follows a Chi-square distribution with 2 Degree Freedom(DF). For the retrospective case-control study, in many logistic regression methods, each SNP can be coded two SNPs, dominant SNP and recessive SNP. We code each SNP to three SNPs, dominant SNP, recessive SNP and additive SNP. We can get three Chi-square significance to express the recessive, additive and dominant model’s significance and get the maximum value or the minimal p values as the significant
value. So we can get each SNP p values according formula 2 and used the maximum values as this SNP's significance.

Formula 2 and 3 were used in simulated datasets and real dataset.

\[ \chi^2_{\text{max}} = \max(\chi^2_D, \chi^2_A, \chi^2_R) = \max(\chi^2(0), \chi^2(1/2), \chi^2(1)) \]  

(2)

2.3. MAXGtest

The G-test is also called likelihood test and can be used to evaluate the statistical significance for a given SNP [26]. Compute as formula 3-4. In paper [26], the author used the G-test to test the significant of each SNP. we computed the significance of three SNPs with the G-test and used the maximum value as the effect of SNPs.

\[ G_{\text{test}} = \sum_{i=0}^{2} p_i \log \frac{p_i N}{n_i P} + \sum_{i=0}^{2} q_i \log \frac{q_i N}{n_i Q} \]

\[ G_{\text{max}} = \max(G_D, G_R, G_A) \]  

(3)  

(4)

2.4. MAXRE

Relative entropy is often used to measure the difference between two distributions. And in paper 15, the authors used relative entropy to detect the difference between SNP interaction models in case/control samples. And we used RE to compare dominant SNP, recessive SNP and additive SNP and adopt the maximum difference value as the significance of this SNP. We computed with formula 5-6.

\[ RE_{\text{test}} = \sum_{i=0}^{2} p_i \log \frac{p_i Q}{q_i} \]

\[ RE_{\text{max}} = \max(RE_D, RE_R, RE_A) \]  

(5)  

(6)

2.5. MAXED

Entropy is used to measure the distribution of a variable degree of stability. We adopt an entropy difference between case and control samples to compute the significance of three models. In the results, we used the maximum value to express the relevance between this snp and the phenotype. We computed with formula 7-8.

\[ ED_{\text{test}} = \text{abs}(\sum p_i \log p_i - \sum q_i \log q_i) \]

\[ ED_{\text{max}} = \max(ED_D, ED_R, ED_A) \]  

(7)  

(8)

3. Experimental results and analysis

As we have mentioned in the methods section, multi-stage detection methods will cause information loss about without main-affection SNPs or redundancy about main-affection SNPs. To illustrate this problem and improve multi-stage method efficiency, we used two types of simulated datasets and one real data set to test five statistics.

3.1. Simulated test

3.1.1 Simulated dataset1. We used simulated datasets [14,27] including sample 100,250,500,1000,2000 and models. Through 1000 replications of simulations of 2000 samples under the 5 susceptibility SNPs, we compared the statistical power of five statistics. The false-positive rate(FPR) and true positive rate(TPR) were counted at different cutoffs of the number of selected SNPs and then averaged over the 100 replications of simulation. The receiver operating characteristic (ROC) curve are presented in Fig.1. The ROC curve indicated that MAXtest is more powerful method but MAXED is better method that detected without or weak main effect SNPs than other methods. Through a variety of ways to test MAXED, MAXRE and MAXGtest, we used other three curves. Precision ratio and recall ration(PR) curve can be showed, positive predictive value (PP curve), false discovery rate
(FDR). PR, PP, FDR will close or equal to one when the method can perfectly identify the susceptible models. Figure 2 is AUC areas for datasets for MAXED and MAXtest. Figure 3 is answer ration for filter half of SNPs and we also got that results of MAXED and MAXtest better that other methods. From figure 3 we can see that it is possible to get all answers combination MAXtest and MAXED. So these two methods are very efficient ways and an effectively improve the ability of detection based on first step filter.

Figure 1. Comparison of average accuracy and robustness between MAXtest, chi-squares test, MAXED, MAXRE, MAXGtest on simulate datasets. ROC curve: true positive rate against false-positive rates. (b) PR curves: precision versus recall values. (c) PP curves: sensitivity ration versus specified ratio. (d) FDR curves: error discovery rates against true positive rates.

3.1.2 Simulate datasets. Based on four commonly used two-locus epistasis models [16] with different parameter setting, we used simulated data sets. Five statistic was evaluated on simulated data sets. There are four parameter settings as in previous studies [27]. These parameters are disease prevalence p(D), genetic heterogeneity, sample counts N and the minor allele frequencies MAF. The prevalence and the genetic heterogeneity are controlled by the parameters alpha and theta. When specified the p(D) and the heritability alpha and theta can be computed based on equations [16]. There is two ways simulate datasets and with/without main effect. For with main effect sample counts are 800 and 1600, SNP counts is 1000 and 12 datasets respectively. And for without main effect, samples counts are 200, 400, 800, 1600, each of 25 datasets in maf is 0.2. And for without main effect, samples counts are 200, 400, 800, 1600, each of 25 datasets in maf is 0.4.
Figure 4 is average ROC curve for with main effect and we can also see that MAXtest is the best method. Figure 5 is average ROC curve for without main effect and we can see that each method is bad but MAXED are slightly better than others.

3.2. Result and Analysis
We used five methods to analyze two simulate datasets. When there are more than one models, MAXtest is best method and MAXED is second for without main effect and MAXED is the best method for with main effect. Then combination MAXED and MAXtest is very efficient approach to get a better detection effect.

3.3. AMD dataset
We used five methods to detection AMD dataset [28] for single way model, SNP count is 103611, case sample is 96 and control sample is 50. We collected papers in PUBMED about AMD disease, get the two answer SNP sets with diseases [11,12,13,20,29,30]. The first answer set is formed as long as there appeared in a paper, and the second answer set is formed that they appeared in two or more papers. At the different p value, we designed figure 6 to detail detection effect with five methods.
From figure 6, we also can see that MAXED is getting better with the decrease of p-value. So we can detect the SNP for epitatic with MAXED and MAXtest for the first filter step.

4. Discussion

Currently, gene-gene association is still one of the core issues to identify genome-wide association studies, but in the identification and detection of pathogenic model still exists the problem of computational complexity, multi-stage filtration process, although to a computing revolution has brought increased recognition error rate, this paper multi-stage filtration process stage a variety of methods were compared, the results of simulated and real data show that the maximum difference between the statistical methods and the maximum entropy method to form the first stage of the filter can be in the low false positive based on the recognition that there is no main effect of combining sites, to improve the accuracy of subsequent research provides a supporting role, the follow-up research study provides a multi-stage process of experimental and theoretical basis for.

Acknowledgments

we are grateful to the anonymous reviewer whose suggestions and comments contributed to the significant improvement of this paper. This work was supposed by the Scientific Research Support projects of Xi'an University of Finance and Economics (Grant No. 17FCJH12) and the Education and Teaching reform project of Shaanxi Computer Education Association (Grant No. 19-12).

References

[1] Moore, Jason. H, and Scott M. Williams. (2009) Epistasis and its implications for personal genetics. The American Journal of Human Genetics. 85(2):309-320.
[2] Moore, Jason. H, Folkert. W. Asselbergs, and Scott. M. Williams. (2010) Bioinformatics challenges for genome-wide association studies. Bioinformatics. 26 (4):445-455.
[3] Gilbert, Diamond, Diane, and Jason. H. Moore. (2011) Analysis of Gene-Gene Interactions. Current Protocols in Human Genetics. pp.1-14.
[4] Li .Pei, et al . (2014). An overview of SNP interactions in genome-wide association studies. Briefings in Functional Genomics. 14 (2):143-155.
[5] Wang. Xuefeng, Robert.C. Elston, and Xiaofeng. Zhu. (2010) The meaning of interaction. Human Heredity. 70(4):269-277.
[6] Wang, Xuefeng, Robert. C. Elston, and Xiaofeng. Zhu. (2011) Statistical interaction in human genetics: how should we model it if we are looking for biological interaction? Nature Reviews Genetics. 12(1):74-74.
[7] Kempthorne. O. (1968) The correlation between relatives on the supposition of Mendelian inheritance. American journal of human genetics.20(2):402.
[8] Moore, J. H. (2005) A global view of epistasis. Nature genetics. **37**(1):pp.13-14.

[9] Marchini, J, Donnelly, P, and Cardon, L. R. (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. Nature genetics. **37**(4):413-417.

[10] Ritchie, Maryln. D, et al. (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. The American Journal of Human Genetics. **69**(1):138-147.

[11] Zhang, Y, Liu, J. S. (2007). Bayesian inference of epistatic interactions in case-control studies. Nature genetics.. **39**(9):1167-1173.

[12] Tang, W, Wu, X, and Jiang, R. (2009) Epistatic module detection for case-control studies: a Bayesian model with a Gibbs sampling strategy. PLOS genetics. **5**(5).

[13] Jiang, Rui, Tang, W, Wu, X. (2009) A random forest approach to the detection of epistatic interactions in case-control studies. BMC bioinformatics. **10**(1).

[14] Miller, D. J, Zhang, Y and Yu, G. (2009) An algorithm for learning maximum entropy probability models of disease risk that efficiently searches and sparsely encodes multilocus genomic interactions. Bioinformatics. **25**(19):2478-2485.

[15] Zhang, X, Huang, S and Zou, F. (2010) TEAM: efficient two-locus epistasis tests in human genome-wide association study. Bioinformatics. **26**(12):217-227.

[16] Wan, X, Yang, C and Yang, Q. (2010) BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. The American Journal of Human Genetics. **87**(3):325-340.

[17] Wang, Y, Liu, X and Robbins, K. (2010) AntEpiSeeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm. BMC Research Notes. **3**(1).

[18] Wan, X, Yang, C and Yang Q. (2010) Predictive rule inference for epistatic interaction detection in genome-wide association studies. Bioinformatics. **26**(1):30-37.

[19] Xie, M, Li, J and Jiang, T. (2012) Detecting genome-wide epistases based on the clustering of relatively frequent items. Bioinformatics. **28**(1):5-12.

[20] Shang, J, Zhang, J and Sun Y. (2014) EpiMiner: A three-stage co-information based method for detecting and visualizing epistatic interactions. Digital Signal Processing. **24**(1-3)

[21] Yang, Can, et al. (2009) SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. Bioinformatics. **25**(4):504-511.

[22] Li, Q, Yu, K and Li, Z. (2008) MAX-rank: a simple and robust genome-wide scan for case-control association studies. Human Genetics. **123**(6):617-623.

[23] González, J. R, Carrasco, J. L and Dudbridge, F. (2008) Maximizing association statistics over genetic models. Genetic Epidemiology. **32**(3):246-254.

[24] Freidlin, B, Zheng, G and Li Z. (2002) Trend tests for case-control studies of genetic markers: power, sample size and robustness. Human Heredity. **53**(3):146-152.

[25] Zheng, Gang, Jungnam. Joo and Yaning. Yang (2009) Pearson's test, trend test, and MAX are all trend tests with different types of scores. Annals of human genetics. **73**(2):133-140.

[26] Cordell, H. J, Clayton, D. G. (2002) A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. The American Journal of Human Genetics. **70**(1):124-141.

[27] Yuan, X, Zhang, J and Wang, Y. (2011) Simulating linkage disequilibrium structures in a human population for SNP association studies. Biochemical Genetics. **49**(5):395-409.

[28] Klein, R. J, Zeiss. C and Chew, E. Y. (2005) Complement factor H polymorphism in age-related macular degeneration. Science. **308**(5720):385-389.

[29] Liao, Z, Zeng, Q, Liao.B and Li.X. (2014) A Novel Two-Stage Approach for Epistasis Detection in Genome-Wide Case-Control Studies. Biochemical Genetics. **52**(9):403-414.

[30] Ik-Soo, Huh, Sohee Oh and Taesung Park. (2011) A chi-square test for detecting multiple joint genetic variants in genome-wide association studies. Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on, IEEE. pp.708-713.