De-Shuang Huang
M. Michael Gromiha
Kyungsook Han
Abir Hussain (Eds.)

LNAI 10956

Intelligent Computing Methodologies

14th International Conference, ICIC 2018
Wuhan, China, August 15–18, 2018
Proceedings, Part III

Springer

Integration of Data-Space and Statistics-Space Boundary-Based Test to Control the False Positive Rate

Jin-Xiong Lv and Shikui Tu(✉)

Department of Computer Science and Engineering, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
{lvjinxiong,tushikui}@sjtu.edu.cn

Abstract. Many multivariate statistical methods have been applied to detect the difference between case and control population. However, it is difficult to control the false positive rate, especially under small sample size. Traditional family-wise error rate or false discovery rate adjusts the p values based on the distribution or ranks of p value in the same multiple testing. In this paper, we investigated the performance of integrating the Data-space boundary-based test (BBT) and Statistics-space BBT to control the false positive rate, under a previous proposed framework called Integrative Hypothesis Tests (IHT). The classification accuracy rate by Data-space BBT provides valuable information complementary to the p value from Statistics-space BBT. The simulation results demonstrated that the integration effectively controls the false positive rate even for small-sample-size cases. Experiments on the real-world dataset of bipolar disorder also validated the effectiveness of the integration.

Keywords: Integrative Hypothesis Test · Boundary-based test
False positive rate · Multivariate statistical method · Joint-SNVs analysis
Bipolar disorder

1 Introduction

Many statistical methods have been proposed in the literature to detect the difference between case and control population in the fields of social psychology, biology, and economics. One focuses on the difference between case and control population in many fields, such as social psychology, biology and economics. There are many statistical methods to detect the difference. Those methods can be divided into two groups, one for univariate and the other for multivariate methods which play a crucial role in genome-wide association study (GWAS). Recently, the multivariate methods are applied on GWAS for detection power improvement, in which single-nucleotide variants (SNVs) located in the same biological unit are collapsed into one computational unit [1–4]. However, all of them are suffering from false positive rate, especially with small sample size.

The traditional measures to control the false positive rate have two main streams. The first stream is named after family-wise error rate (FWER), which is defined as the

D.-S. Huang et al. (Eds.): ICIC 2018, LNAI 10956, pp. 628–638, 2018.
https://doi.org/10.1007/978-3-319-95957-3_65

probability of making one or more false rejections among all of the hypotheses, such as Bonferroni correction [5] and Holm correction [6]. The other stream aims to control the fraction of false rejection under the threshold α, including Benjamin-Hochberg correction [7] and Q value [8]. But they only change the scale of original p values and fix the threshold to reject less hypotheses without any other complementary information.

Recently, Integrative Hypothesis Tests (IHT) was previously proposed in [9, 10] to consider discriminating analysis and testing of case-control problems, jointly from two perspectives. One is model based such as two-sample test or model comparison to detect the difference between two populations, while the other is boundary based such as classification or model prediction about the performance of the distinguishing boundary. As preliminarily discussed in [11], the tasks of model comparison and classification were complementary to each other in nature, and it was better to jointly optimizing their performances. The advantage of IHT was demonstrated in [12] on a COPD-Lung cancer study, by combing p value (from a two-sample test) and misclassification rate in a 2D scatter plot, with a bootstrapped procedure to enhance the reliability of the ranks by the IHT. This motivated us to further investigate IHT, and following [10] Boundary-based test (BBT) was considered and empirically investigated in this paper.

Boundary-based test (BBT) is to test whether a separable plane is existed between the case population and control population [10]. It can be classified into two categories, Data-space BBT and Statistics-space BBT. Data-space BBT indicates that we seek for the separating plane in the original data space, which leads to meet traditional classification problem in the machine learning. While the Statistics-space BBT intends to ascertain the boundary between rejection region and acceptance region after calculating the statistics from the original data space. Furthermore, the Statistics-space BBT has achieved higher detection power than other multivariate methods in joint-SNVs analysis [13–15]. In order to reduce the background disturbance, the posteriori of p value (pp value) is introduced in the Statistics-space BBT [10].

In this paper, we investigate IHT by integrating Data-space BBT and Statistics-space BBT to control the false positive rate in the multivariate case. The effectiveness of the integration was validated on both synthetic datasets and real-world datasets. Results also empirically demonstrated that Data-space BBT and Statistics-space BBT were complementary to each other in controlling false positive rate. The pp value is helpful in controlling the false positive rate on the synthetic experiments.

This paper is organized in the following way. In Sect. 2 we briefly introduced IHT, as well as BBT, and then focused on the integration of Data-space BBT and Statistics-space BBT. An intuitive method for integration of them was studied. In Sect. 3, we conducted simulation experiments. In Sect. 4, we applied the intuitive integration on a SNV dataset of bipolar disorder and performed a literature search to validate the effectiveness of the integration.

2 Integrative Hypothesis Tests and Boundary-Based Test

2.1 Brief Introduction of IHT

Integrative Hypothesis Tests (IHT) was proposed in [9, 10] to consider discriminating analysis and testing of case-control problems, jointly from two perspectives, i.e., model-based perspective and boundary-based perspective, which involves four tasks as described in Table 1 of [10]. From the model-based perspective, we utilize parametric models to describe the case population and control population, and then measure the difference between two populations. From the boundary-based perspective, we detect the existence of boundary for two populations. In the following part, we focus on the task B comparison and task C classification of IHT to control the false positive rate. The task B offers p value to measure the difference and task C provides misclassification rate. Correspondingly, the Data-space BBT and Statistics-space BBT finish task B and task C, and both of them will be described in the following subsections.

Table 1. Annotation for top-20 genes of bipolar disorder

Class	Gene_symbol	Description
A	RYR3	Ryanodine Receptor 3
	NPAS3	Neuronal PAS Domain Protein 3
	WWOX	WW Domain Containing Oxidoreductase
	DLG2	Discs Large MAGUK Scaffold Protein 2
	DPP10	Dipeptidyl Peptidase Like 10
	CDH13	Cadherin 13
B	SHISA6	Shisa Family Member 6
	LRP1B	LDL Receptor Related Protein 1B
	ASTN2	Astrotactin 2
	PTPRD	Protein Tyrosine Phosphatase, Receptor Type D
	LRRC4C	Leucine Rich Repeat Containing 4C
	PRKCA	Protein Kinase C Alpha
	FHIT	Fragile Histidine Triad
	GALNT13	Polypeptide N-Acetylgalactosaminyltransferase 13
	PARK2	Parkin RBR E3 Ubiquitin Protein Ligase
C	THSD4	Thrombospondin Type 1 Domain Containing 4
	FAM155A	Family With Sequence Similarity 155 Member A
	ZNF664-FAM101A	Filamin-Interacting Protein FAM101A
	PRKCE	Protein Kinase C Epsilon
	USH2A	Usherin

As an early application of IHT, Jiang et al. takes the task B and task C into consideration to identify miRNAs biomarkers for the differentiation of lung cancer and Chronic Obstructive Pulmonary Disease (COPD) [12]. As illustrated in Fig. 1 in [12], a p value indicating difference of the two distributions and a misclassification rate

indicating a separating boundary were combined in a 2D scatter plot for an IHT rank on the features. In order to improve the reliability when the sample size is small and missing value is existed, the bootstrapping method was proposed.

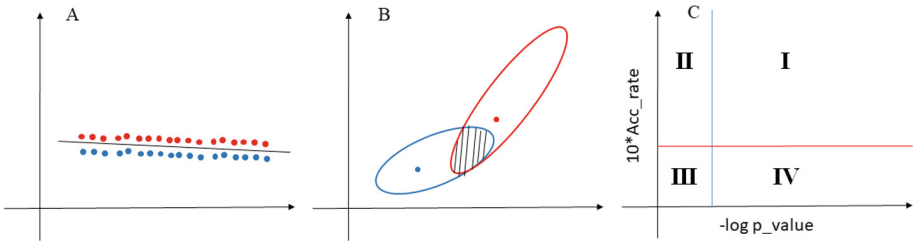

Fig. 1. Examples for explanation of integration for Data-space BBT and Statistics-space BBT.

2.2 Data-Space Boundary-Based Test

The Data-space Boundary-based test aims to seek for boundary to classify the samples into case population and control population, that is, it belongs to two-class classification problem. For the simplest case, we can defined a hyperplane,

$$g(x, \mathbf{w}) = \mathbf{w}^T x + w_0 = \mathbf{w}^T (x - \mu) \tag{1}$$

where μ is the mean of population. Then, the data can be divided into two classes by,

$$\begin{cases} \text{case}, & \text{if } g(x, \mathbf{w}) > 0 \\ \text{control}, & \text{if } g(x, \mathbf{w}) \leq 0 \end{cases} \tag{2}$$

then we constructed the statistics as misclassification rate as described in the [10],

$$\mathrm{s} = \frac{\#X_1^{(0)} + \#X_0^{(1)}}{\#X^{(0)} + \#X^{(1)}} \tag{3}$$

where the $X^{(1)}$ indicates the case population, the $X^{(0)}$ is the control population, the $X_1^{(0)}$ means those that belong to control population but are classified into case population and $X_0^{(1)}$ indicates those that belong to case population but are classified into control population. The # means the number of candidates for one set. We can utilize support vector machine (SVM) [16] and fisher discrimination analysis (FDA) to obtain the boundary. Then we can apply the Eq. (3) to obtain the statistics. The smaller statistics are, the more separable two populations are.

2.3 Statistics-Space Boundary-Based Test

Different from the Data-space BBT, the Statistics-space BBT firstly computes statistics from the original data, such as the difference of means for two populations. We then

calculate p value by permutation test. It is one-class classification problem which is to ensure whether the statistic is located in rejection region or not. Then we can describe the p value,

$$p = \frac{\#X_1^{(0)}}{\#X^{(0)}} \tag{4}$$

where the $X^{(0)}$ is the statistics set located in acceptance region, the $X_1^{(0)}$ is the set that contains wrong rejections and the # means the number of candidates for one set. Because the null hypothesis of permutation test is that the sample labels are exchangeable, those located in the rejection region are the misclassification. As a result, the p value obtained by permutation test is also the misclassification rate for the one-class classification problem.

Xu has provided four key steps of Statistics-space BBT in the Table 6 of [10]. For the multivariate case, we also summaries four main ingredients. First, we determine a rejection domain $\Gamma(\tilde{\mathbf{s}})$ based on statistic $\tilde{\mathbf{s}}$ which is obtained from case-control study,

$$\Gamma(\tilde{\mathbf{s}}) = \left\{\mathbf{s} : (\mathbf{s} - \tilde{\mathbf{s}})^T \mathbf{sign}(\tilde{\mathbf{s}}) > \mathbf{0}\right\} \tag{5}$$

where $\mathbf{sign}(\mathbf{s}) = [sign[s_1], \cdots, sign[s_m]]^T$ with $sign[u] = \pi \frac{u}{|u|}$. Second, the p value can be calculated by permutation test regardless of distribution. Third, we make full use of the principle component analysis (PCA) to remove the cross-dimensional dependence for factorization of multivariate p value (see Eq. (68) in [10]). Forth, we corrected the p value into posterior version (pp value) of it to reduce the background disturbance (see Eq. (93) in [10]). In the next section, we will also show the important role that pp value played in control of the false positive rate.

2.4 Integration for Data-Space BBT and Statistics-Space BBT via IHT

From the perspective of IHT, the Data-space BBT finishes the task C and Statistics-space BBT finishes the task B, and then we integrate the p value and misclassification rate to control the false positive rate. Their complementarity is described as follows.

The Data-space BBT is to classify the samples into two classes in the original data space, case population and control population. To some extent, it takes difference both for means and covariance into consideration. However, a separable boundary is not equivalent to existence of significant difference between two populations. As Fig. 1A described, there is a separable boundary between two populations, but there is no significant difference.

The Statistics-space BBT aims to determine whether the statistics are located in rejection region or not. The mean is often regarded as the criterion to indicate the difference for case-control study, such as Hotelling's T square test for multivariate [17]. Similarly, a significant difference of means from two populations does not indicate the existence of a separable boundary as Fig. 1B showed. In the Fig. 1B, two dots indicate means of two populations and two ellipses represent two population. Although the

distance of two means is large, the shadow area is also large, which indicates that there is no separable boundary.

We have analyzed Data-space BBT and Statistics-space BBT empirically, and it is helpful to integrate both of them to obtain more accurate results, in other words, control the false positive rate. Note that the misclassification rate is replaced with accuracy rate in the 2D-scatter plot compared with the plot in the [12]. We call it p value vs. accuracy rate scatter plot.

As the Fig. 1C described, the horizontal ordinate represents negative natural logarithm of the p value and the vertical ordinate indicates accuracy rate which takes the place of misclassification rate for convenience. What's more, we multiply the accuracy rate by ten to make the scale as the same as negative natural logarithm of the p value. The scatter plot can be divided into four regions. The hypotheses located in the region I will be rejected via integration of Data-space BBT and Statistics-space BBT, while the hypotheses that belong to the rest three regions will be accepted. If we only consider the p value, the hypotheses in region I and IV are rejected. When we take the accuracy rate into consideration, those located in region I and II are rejected.

3 Simulation Framework and Corresponding Results

3.1 Simulation Framework for Type I Error

We intend to investigate the effect on the false positive rate of joint-SNVs analysis in GWAS when we integrate Data-space BBT and Statistics-space BBT via IHT. In order to mimic the pattern of the real-world data, the SNV data located in 2.5 Mb of chromosome 5 from the 1000 genomes projects was chosen and the number of the SNVs is 12,455. To remove the influence of ethnicity, we choose the Chinese Beijing (CHB) and the sample size is 97. The length of the computational unit is set to 15 kb because the length of gene is from 10 kb to 15 kb. The simulation datasets were generated by the null model,

$$y = 0.5X_1 + 0.5X_2 + \varepsilon \tag{6}$$

where y is the phenotype, X_1 is a continuous covariate generated from a standard normal distribution, X_2 is a bi-value covariate which can take 0 or 1 with a probability 0.5, and ε follows a standard normal distribution. It is of note that the generated phenotype is not associated with the genotype data when the null hypothesis holds. Then we transform the continuous phenotype into dichotomous value via the following model,

$$\text{logit}\, P(y = 1) = \alpha_0 + 0.5X_1 + 0.5X_2 \tag{7}$$

where the logit means the logit function and α_0 is the prevalence. The α_0 is set to 0.01. The sample size is set to 100 vs. 100, 500 vs. 500 and 1000 vs. 1000. And there are 1000 replicates for them in the same parameter settings. The Data-space BBT adopted the linear SVM to seek for separable hyperplane and Statistics-space BBT followed the procedure described in the [10] to obtain joint p values.

3.2 Simulation Results

The simulation results are shown in the Fig. 2. Each row represents different sample size, 100 vs.100, 500 vs.500 and 1000 vs.1000 in order. First column described the relationship between accuracy rate and *p* value and second column is for the relationship between accuracy rate and *pp* value. The blue line indicates the type I error threshold α ($\alpha = 0.05$) and the red line means the threshold for accuracy rate (99% for 1000 vs. 1000, 95% for 500 vs. 500 and 60% for 100 vs. 100). The scatter plots are divided into four regions by the two lines as the Fig. 1C described.

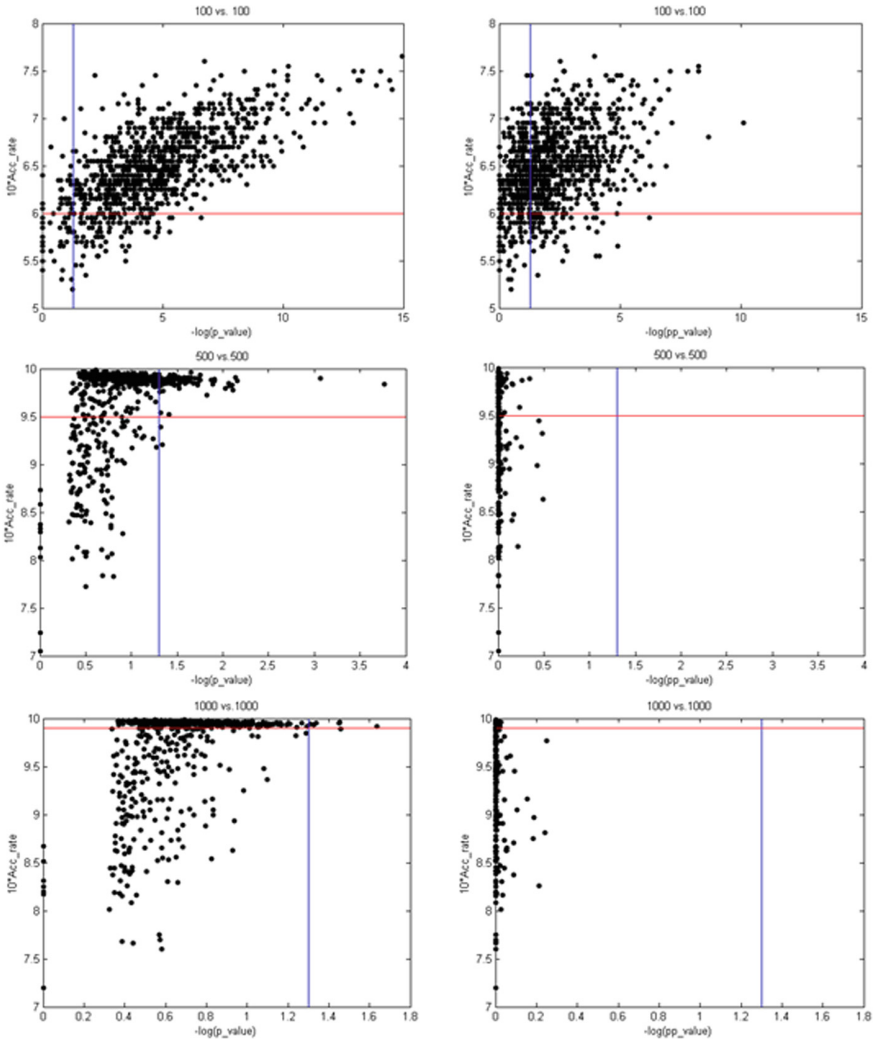

Fig. 2. Scatter plot *p* value vs. accuracy rate for simulation datasets

Altogether, there are four points to be mentioned. First, p value vs. accuracy rate scatter plot reduces the false positive rate under different sample size as the number of replicates decreasing in the region I. Second, for the cases of 500 vs. 500 and 1000 vs. 1000, the accuracy rate has no influence on controlling false positive rate when we use the pp value, while for the p value, the accuracy rate helps with reducing the number of replicates in the region I. Thus, the pp value has an effect on control of the false positive rate. Third, with the sample size increasing, false positive rate decreases. If we intend to decrease the false positive rate even further, the threshold of accuracy rate needs to be raised. Forth, if the sample size is large enough, the pp value can control the false positive rate efficiently.

4 Results for Bipolar Disorder SNV Dataset

4.1 Basic Information for Real-World Dataset and Data Pre-processing

Bipolar disorder is responsible for the loss of more disability-adjusted life-years and leads to high risk of suicide and self-harm [18–20]. The dataset of bipolar disorder (GSE71443) comes from GEO (Gene Expression Omnibus) database. The sample size is 65 bipolar disorder patients vs. 74 controls and the number of probes is 906600. There is no missing value to impute. We regarded the Hardy-Weinberg's equilibrium p value as the measure for quality control. If any of the three Hardy-Weinberg's equilibrium p values for case population, control population and the total population, is smaller than 1E-4, the SNVs would be filtered out. As a result, there are high-quality 722130 probes.

Then, we annotated the SNVs using the Annovar software [21]. And the unit was defined as the body of gene (exclude the upstream, downstream, intergenic and so on). Finally, we obtained 13896 genes. As the small sample size, we selected the first 20 SNVs in ascending order of their single locus p values to represent these genes, and then calculated the joint p values for them via Statistics-space BBT. All of the joint p values are corrected into pp values and the joint p values indicate the pp values. The Data-space BBT is also performed for the 20 dimensional data to calculate the misclassification rate via linear SVM.

4.2 Results for the Bipolar Disorder

After we obtained the p value and accuracy rate, p value vs. accuracy rate scatter plot of 13896 genes is shown in the Fig. 3 and the top-20 genes in the ascending order of their joint p values are marked in red. In addition, we also showed the relationship between two correction p values and original p values and found that both of them only change the scale of original p values and fix the threshold to reject less hypotheses without any other complementary information. We then conducted the literature search for them and they were described in the Table 1 (see details in the Table 5 in [15]). They are classified into three categories. Class A indicates that there is direct association with the bipolar disorder. Class B contains the genes related to the other psychiatric disorders or brain disorders, while the genes showed that there is no relationship to our best

knowledge of the literature were collected in class C. Next, we will utilize the scatter plot to exclude some genes and we hope that the genes belonging to class A can be reserved as much as possible.

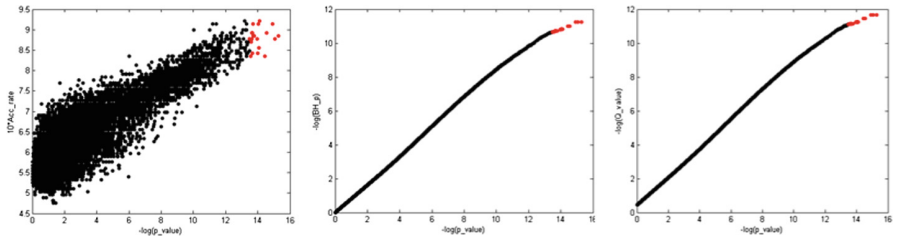

Fig. 3. Scatter plot *p* value vs. accuracy rate for all of 13896 genes

The top-20 genes were shown in the Fig. 4. The blue diamond represents genes in class A. the red circle means the genes in class B and the green star indicates genes in class C. The number of genes located in the region I is 8. Among them, the number of genes in class A decreased from 6 to 4, from 9 to 1 for class B and from 5 to 3 for class C. As a result, the remained genes are shown in the Table 2 and the scatter plot is able to control the false positive rate.

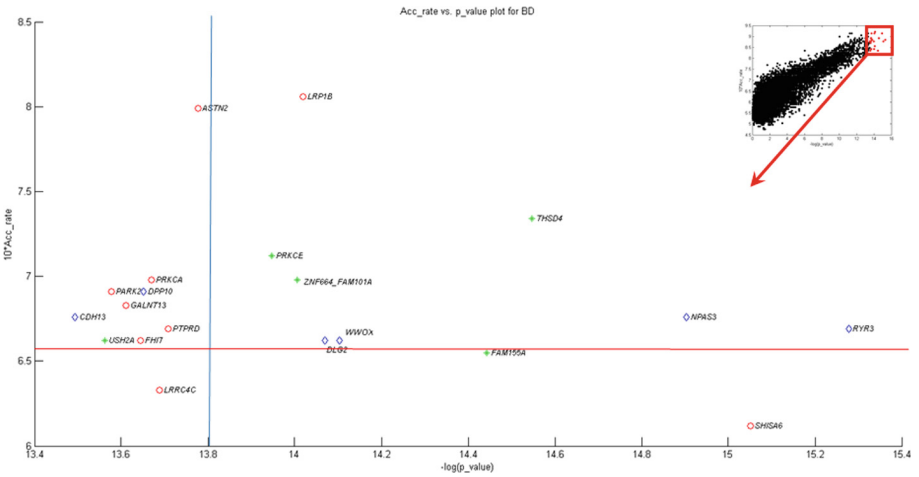

Fig. 4. Scatter plot *p* value vs. accuracy rate for top-20 genes

Table 2. Annotation for top-20 genes after integration

Class	Gene_symbol	Description
A	RYR3	Ryanodine Receptor 3
	NPAS3	Neuronal PAS Domain Protein 3
	WWOX	WW Domain Containing Oxidoreductase
	DLG2	Discs Large MAGUK Scaffold Protein 2
B	LRP1B	LDL Receptor Related Protein 1B
C	THSD4	Thrombospondin Type 1 Domain Containing 4
	ZNF664-FAM101A	Filamin-Interacting Protein FAM101A
	PRKCE	Protein Kinase C Epsilon

5 Discussion

We investigated the performance of integrating the Data-space BBT and Statistics-space BBT under the IHT. The simulation experiments were designed to further elucidate the effectiveness of the integration. The simulation results showed that the integration can help with controlling the false positive rate even for small sample size and the *pp* value can also control the false positive rate. While for the SNV datasets of bipolar disorder with small sample size, the integration also played a crucial role in controlling the false positive rate. In the future, there are three perspectives. First, the influence for the scatter plots under different classification methods is a worthwhile studying. Similarly, the influence for the scatter plots under different multivariate statistical methods also raises a mandatory research. Second, we will try to construct a statistic to replace the scatter plot via integrating the p value and accuracy rate. Third, for the scatter plot, it is essential to validate whether there is an arc line instead of the lines being parallel to the axis to control the false positive rate more efficiently. Altogether, IHT is the framework to control the false positive rate without being limited to boundary-based tests, and it is worthwhile to investigate more deeply its characteristics.

Acknowledgement. This work was supported by a grant from Shanghai Jiao Tong University, NO. WF220403029.

References

1. Han, F., Pan, W.: A data-adaptive sum test for disease association with multiple common or rare variants. Hum. Hered. **70**(1), 42–54 (2010)
2. Lee, S., Wu, M.C., Lin, X.: Optimal tests for rare variant effects in sequencing association studies. Biostatistics **13**(4), 762–775 (2012)
3. Wu, M.C., et al.: Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. **89**(1), 82–93 (2011)
4. Price, A.L., et al.: Pooled association tests for rare variants in exon-resequencing studies. Am. J. Hum. Genet. **86**(6), 832–838 (2010)
5. Dunn, O.J.: Multiple comparisons among means. J. Am. Stat. Assoc. **56**(293), 52–64 (1961)

6. Holm, S.: A simple sequentially rejective multiple test procedure. Scand. J. Stat. **6**, 65–70 (1979)
7. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Royal Stat. Soc. Series B (Methodol.) **57**(1), 289–300 (1995)
8. Storey, J.D.: The positive false discovery rate: a Bayesian interpretation and the q-value. Ann. Stat. **31**(6), 2013–2035 (2003)
9. Xu, L.: Integrative hypothesis test and A5 formulation: sample pairing delta, case control study, and boundary based statistics. In: Sun, C., Fang, F., Zhou, Z.-H., Yang, W., Liu, Z.-Y. (eds.) IScIDE 2013. LNCS, vol. 8261, pp. 887–902. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-42057-3_112
10. Xu, L.: Bi-linear matrix-variate analyses, integrative hypothesis tests, and case-control studies. Appl. Inform. **2**, 4 (2015). Springer, Berlin Heidelberg
11. Xu, L., Jiang, C.: Semi-blind bilinear matrix system, BYY harmony learning, and gene analysis applications. In: 2012 6th International Conference on New Trends in Information Science and Service Science and Data Mining (ISSDM). IEEE (2012)
12. Jiang, K.-M., Lu, B.-L., Xu, L.: Bootstrapped integrative hypothesis test, COPD-lung cancer differentiation, and joint miRNAs biomarkers. In: He, X., Gao, X., Zhang, Y., Zhou, Z.-H., Liu, Z.-Y., Fu, B., Hu, F., Zhang, Z. (eds.) IScIDE 2015. LNCS, vol. 9243, pp. 538–547. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23862-3_53
13. Lv, J.-X., et al.: A comparison study on multivariate methods for joint-SNVs association analysis. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE (2016)
14. Lv, J., Tu, S., Xu, L.: A comparative study of joint-SNVs analysis methods and detection of susceptibility genes for gastric cancer in Korean population. In: Sun, Y., Lu, H., Zhang, L., Yang, J., Huang, H. (eds.) IScIDE 2017. LNCS, vol. 10559, pp. 619–630. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67777-4_56
15. Lv, J.-X., et al.: Comparative studies on multivariate tests for joint-SNVs analysis and detection for bipolar disorder susceptibility genes. Int. J. Data Min. Bioinform. **17**(4), 341–358 (2017)
16. Suykens, J.A., Vandewalle, J.: Least squares support vector machine classifiers. Neural Process. Lett. **9**(3), 293–300 (1999)
17. Hotelling, H.: The generalization of student's ratio. In: Kotz, S., Johnson, N.L. (eds.) Breakthroughs in Statistics. Springer Series in Statistics (Perspectives in Statistics), pp. 54–65. Springer, New York (1992). https://doi.org/10.1007/978-1-4612-0919-5_4
18. Anderson, I.M., Haddad, P.M., Scott, J.: Bipolar disorder. Br. Med. J. BMJ (Online) **345** (2012)
19. Pompili, M., et al.: Epidemiology of suicide in bipolar disorders: a systematic review of the literature. Bipolar Disord. **15**(5), 457–490 (2013)
20. Merikangas, K.R., et al.: Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative. Arch. Gen. Psychiatry **68**(3), 241–251 (2011)
21. Wang, K., Li, M., Hakonarson, H.: ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. **38**(16), e164–e164 (2010)