Integrative Hypothesis Test and A5 Formulation: Sample Pairing Delta, Case Control Study, and Boundary Based Statistics

Lei Xu

Department of Computer Science and Engineering
Chinese University of Hong Kong
lxu@cse.cuhk.edu.hk

Abstract. This paper continues the previous preliminary study on integrative hypothesis test (IHT) (Xu, LNCS7751, 2013). First, the coverage of IHT studies are elaborated from four aspects. Then, the previous preliminary A5 formulation for IHT is developed into one that integrates multiple individual tasks of discriminative analysis to improve hypothesis test with enhanced reliability. Next, a sample-pairing-delta based nonparametric statistics is proposed and its application to case control study is addressed. Moreover, a parametric separating boundary is further embedded into hypothesis test with statistics based on how far samples are away from the boundary, under which sample classification and hypothesis testing are coordinately implemented.

Keywords: Integrative hypothesis test, A5 formulation, False discovery rate, Accumulated reliability, Sample pairing delta, case control study, SNP analysis, boundary based statistics.

1 Introduction

Informally, the term *integrative hypothesis testing* (IHT) is used in [1] for discriminative analysis to integrate the evidences associated with structured variables on which one *hypothesis* is supported and to integrate the outcomes of many different hypothesis tests in order to reach a final conclusion. Moreover, taking analyses of gene expression and exome sequencing as examples, one so-called Geno-Pheno A5 analyzer is proposed in [1] to apply an A5 formulation of the problem solving paradigm [2], resulting in a general procedure for IHT implementation.

Many applications in bioinformatics and other tasks of big data analysis involve discriminative analyses on samples from different populations. Without losing much generality, this paper focuses on such tasks of two populations, which is featured by subtasks viewed from three complementary perspectives. One aims at classifying samples into their corresponding populations by a discriminant rule, which is usually named classification or decision and widely encountered in the literatures of pattern recognition and machine learning. Being popular in the literature of bioinformatics and medical informatics, the second is made under the name of *hypothesis test,* which evaluates whether two populations of samples are significantly different according to

C. Sun et al. (Eds.): IScIDE 2013, LNCS 8261, pp. 887–902, 2013.

some discriminative statistics. Both the two subtasks are made on a finite set of samples and thus highly depend on another subtask called *feature or variable selection*. The selected features or variables form the domain on which each of the two subtasks is performed. The two subtasks are related but implemented subject to different performance measures that are not monotonically related. Thus, one best set of selected features for one subtask may not be necessarily one best set for the other.

This paper continues the IHT study made in [1], starting at elaborating the coverage of IHT studies from the following aspects:

- How to integrate information associated with multiple features for testing an overall *hypothesis*. Typical topics include how to develop an overall statistics, e.g., a general choice is suggested by Eq.(24) in [1] with samples of structured features expressed in matrix format, and how to compute statistics efficiently, e.g., as partly discussed in Sect.4.1 of [1], as well as how to estimate the p-value and q-value.
- How to integrate the outcomes of multiple individual hypothesis tests to reach a overall conclusion, typically how to get a combined statistics from the statistics of individual hypothesis tests.
- How to coordinately perform *classification*, *hypothesis test,* and *feature selection*. E.g., can we have a same performance measure for both classification and hypothesis test? or otherwise how to trade off them satisfactorily.
- How to integrate multiple individual performances on either or both of classification and hypothesis test to get an overall good coordination between classification and hypothesis test.

Efforts on the first two aspects have also been partly discussed in [1] and studied also by many others [4-8]. The last two aspects are just preliminarily and incompletely involved in [1] and will be the focuses of this paper. In the next section, a preliminary A5 formulation in [1] is developed into a general formulation that integrates individual tasks of discriminative analysis by circularly implementing five basic actions. Then, it is shown in Sect.3 that this A5 formulation improves hypothesis testing with reliability enhanced (e.g., q-value reduced considerably).

Being different from conventional parametric and nonparametric statistics that firstly comes up an overall structure for each population and then detects significant difference between the summarized overall descriptions, we propose to firstly detect difference between paired samples and check whether these differences can be summarized into a significant one. Sect.4 proposes a sample pairing delta based nonparametric statistics and discusses its application to case control study and especially joint SNPs analyses. Finally, Sect.5 further proposes a boundary based statistics for coordinately implementing classification and hypothesis testing.

2 Integrative Hypothesis Testing and A5 Formulation

The A5 formulation comes from a general problem solving paradigm, refined from the mechanisms embedded in Hough Transform (HT), Randomized Hough Transform (RHT) [2,3] and Multi-sets-learning [16]. As illustrated in Fig.1, taking line detection within one image as an example, the HT detection is featured by a circular flow featured by five basic *mechanisms* or *actions*. First, it starts from picking one pixel from image, which is an instance of the action named *acquisition* (shortly denoted as A-1) for sampling evidence or data from the world in observation. Second, the HT

maps the pixel picked into a line in its parameter space $\theta=\{a,b\}$, which is an instance of the action *allocation & assumption* (A-2) that allocates information contained in the picked pixel, featured by a distributed allocation of evidence along a line that represents a set of candidate assumptions in the parameter space. Third, the HT quantizes a window of the parameter space into a lattice on which every cell is placed with an accumulator. We add one score to those accumulators located on the candidate assumptions provided by A-2, which is an example of the action *accumulation & amalgamation* (A-3) for integrating evidences about these candidate assumptions. Next, we inspect the scores of all the accumulators and detect those, that pass some threshold or become local maxima, as the final candidate conclusions on detected lines, which is an instance of the fourth action *apex-seeking & assessment* (A-4) that decides one or a set of final candidates with their corresponding scores either locating at peaks or becoming bigger than a threshold. Finally, the HT tests whether each of final candidates can be regarded as a detected line. In general, the job is named *affirmation* (A-5) that assesses whether each of candidate conclusions should be either discarded or identified as a final conclusion.

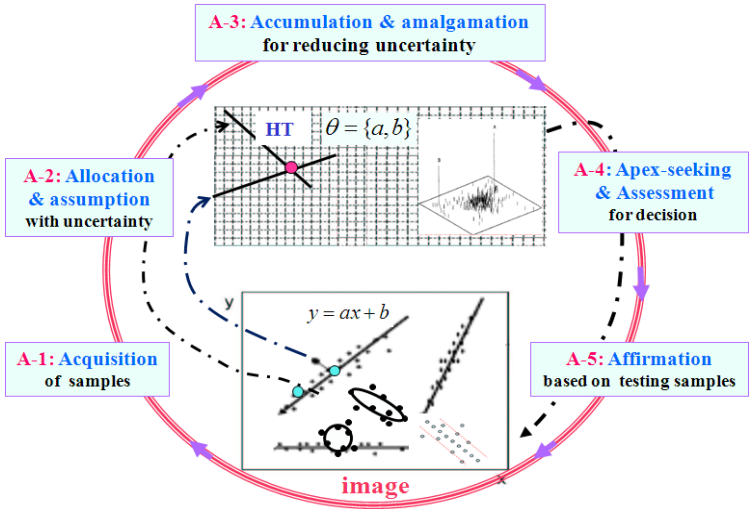

Fig. 1. A5 Formulation and One Exemplar

The tasks of object detection can be regarded as a special family of discriminant analysis that discriminates different populations of samples from geometrical shapes, while discriminant analysis usually refers to discriminate populations of samples from different statistical structures in term of either or both of *classification* and *hypothesis test*. To be specific, we consider a set of feature variables that represent the domain in observation, and then examine two populations within the domain via samples. Each sample is usually a vector or generally a structural set with each element being one specific value taken by the corresponding feature. Given one sample set $X^{(0)}$ from $P^{(0)}$ and the other sample set $X^{(1)}$ from $P^{(1)}$, we want to judge whether two populations are significantly different in term of *hypothesis test* and to classify samples into $P^{(0)}$ and $P^{(1)}$ with minimum confusion in term of *classification*. Usually, some features may be irrelevant, and some features are disturbed by noises or outliers. Instead of

considering $X^{(0)}$ and $X^{(1)}$ with all the features as a whole, both tasks are usually based on the task of finding an optimal subset of features in the name of *feature selection*.

Conventionally, the task pair of hypothesis test and feature selection is conducted separately from the task pair of classification and feature selection. In fact, two task-pairs are related but implemented under different measures that are actually not related monotonically. In sequel, we propose to use the A5 formulation to integrate multiple individual tasks of discriminative analyses with *classification of samples*, *test of hypotheses,* and *selection* of *features* performed coordinately.

Given two original sets $X^{(0)}$ and $X^{(1)}$ with each in an $M\times N$ matrix the consists of samples as rows. Let Ω to be a set of accumulation cells with each cell storing the evidence that supports the corresponding candidate assumption. Initially, Ω is empty, and gradually new cells are added in as new candidate assumptions come. Among Ω, a subset Ω_F is selected to store the final candidates. Moreover, we get $Z^{(0)}$ and $Z^{(1)}$ as a pair of testing sets on the domain of Ω_F. The samples of $Z^{(0)}$ and $Z^{(1)}$ maybe new ones from $P^{(0)}$ and $P^{(1)}$ or just randomly and partially picked from $X^{(0)}$ and $X^{(1)}$.

Schematically, the A5 formulation is featured by circularly implementing the following five *actions*.

Action A-1 (Acquisition of Evidence). Randomly picks m rows and n columns from $X^{(0)}$ and $X^{(1)}$ to form a pair of $m\times n$ matrices $Y^{(0)}$ and $Y^{(1)}$.
Remarks: Typically, the number of rows in $X^{(i)}$ is greatly larger than the number of columns in $X^{(i)}$ in real applications, which makes it unreliable to get a selection among a great number of rows merely based on a small number of columns. Randomly selecting $Y^{(i)}$ out of $X^{(i)}$ is a way of increasing an effective number of columns in a self-boosting manner, though suffering an over-optimistic risk.

Action A-2 (Allocation and Assumption). A subset of features $\{\omega\}_{=}\{\omega_j, j=1,\ldots,m\}$, corresponding to the m rows of $Y^{(i)}$, are selected by a SELECTOR with a score $s_{\{\omega\}}$, and then each ω^* is allocated a score s_{ω^*} according to its importance. Shown in Fig.2(a) is one example that is featured by a monotonic curve obtained by an ALLOCATOR.
Remarks: As an example of SELECTOR, we implement a sparse LDA learning [14] based on $Y^{(0)}$ and $Y^{(1)}$, resulting in a feature set $\{\omega\}$. We may even simply pick a set $\{\omega\}$ randomly. Usually, the allocated score $s_{\{\omega\}}$ is obtained in one of the following two ways:

Way A design an optimal classifier on the feature set $\{\omega\}$, and use its correct classification rate as the score $s_{\{\omega\}}$.

Way B make a multivariate hypothesis test (e.g., a Hotelling T^2 test [17]) based on $\{\omega\}$, and compute its corresponding q-value $q_{\{\omega\}}$ [4-6]. Then, we use 1- $q_{\{\omega\}}$ as the score $s_{\{\omega\}}$.

One rough treatment for ALLOCATOR is simply letting $s_\omega = s_{\{\omega\}}/\#\{\omega\}$ for each ω in $\{\omega\}$, where $\#A$ denotes the number of elements in the set A. A better choice is measuring the importance of a feature by its role in one sequential reduction. That is, each time a least important ω^* is removed with its score s_{ω^*} obtained as illustrated in Fig.2(a), with help of the following special treatments:

- if there are more than one feature ω^* with a same score s_{ω^*}, discard the one with the biggest p-value or misclassification rate obtained merely on this feature.
- if $s_{\omega^*}\le 0$, we discard the least important pair of features by examining all the possible pairs in $\{\omega\}$, and so on so forth.

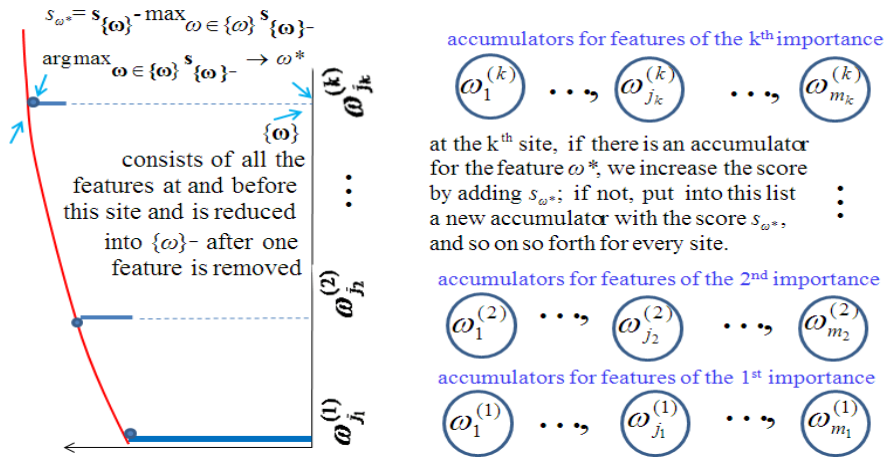

Fig. 2. Importance **of features and accumulation of evidences** (a) features are sequentially ordered by importance with the one of the 1st importance at the bottom and the curve indicates the accumulated score from the bottom up; (b) each of the ordered features is associated with a list of accumulators and each accumulator is associated with a different feature. Each list is empty at beginning and gradually added with accumulators as above described.

Action A-3 (Accumulation and Amalgamation). Evidences of importance are accumulated for the features $\{\omega\}$ provided by A-2, as shown in Fig.2(b).

Action A-4 (Apex-Seeking and Assessment). We pick a feature set $\Omega_F=\{\omega^{(1)},\omega^{(2)}, \ldots, \omega^{(k)}\}$ in either of the following three ways:

(1) Among the union of the 1st, 2nd,…, kth rows with the scores for a same feature added together, we pick ones with the first k largest scores as Ω_F.

(2) Within the first row (at the bottom), pick the one associated with the largest score as $\omega^{(1)}$. Similarly, $\omega^{(2)}$ ($\neq\omega^{(1)}$) is picked from the top scored one within the second row, and so on so forth. Finally, $\omega^{(k)}$ is picked from the top scored one within the k^{th} row (excluding ones that duplicate the already picked features).

(3) After getting $\omega^{(1)}$ as above, $\omega^{(2)}$ ($\neq\omega^{(1)}$) is picked from the second row such that an optimal classifier based on $\omega^{(1)}$, $\omega^{(2)}$ results in the best classification rate as one feature associated with the largest score as $\omega^{(2)}$, and so on so forth.

Action A-5 (affirmation). Make hypothesis test or classification on $Z^{(0)}$, $Z^{(1)}$ based on Ω_F.

Remarks: Some applications verify the performance of classification, while other applications verify the performance of *hypothesis test.* There are also tasks that need to verify both types of performances too. Also, we may evaluate the importance of each $\omega\in\Omega_F$ in a way similar to that shown in Fig.2(a).

The above A5 formulation will increase the reliability of the final outcomes in Ω_F by integrating analyses made on the randomly picked sample sets at A-2, which shares a common point with the methods such as cross-validation, boosting, bagging, and stacking, which are widely studied in machine learning. In the next section, a probabilistic analysis will be provided to show that the evidence accumulation by A-3 and A-4 will significantly bring down the false discovery rate of hypothesis test [4-6].

The above A5 formulation seeks consensus on the importance and stability of features, while those machine learning studies seek consensus on the classification results instead of answering which features are important in their contributions to a good performance. Even worse, a combination of multiple classifiers with each classifier using different features actually enlarge the size of feature set considerably. This is unfavorable to those real tasks (especially in bioinformatics) that need to identify which features are mainly responsible to the final outcomes.

3 False Discovery Rate and Accumulated Reliability

Some probabilistic analyses are made on understanding how the A5 formulation can considerably improve the reliability. Starting from the basic concepts in Table 1, we summarize typical probability measures about hypothesis test from a Bayesian perspective in Table 2. On the last column, π_1 is the priori that H_0 is false, and π_0 is the priori that H_0 is true, while the last row lists the probabilities that the test rejects H_0 and fails to reject H_0, respectively. The most widely used measure is the probability of Type I error, listed in the column of "reject H_0" on the same row of π_0 , which is described by the false positive probability usually called *p*-value that is controlled to be below an α level of significance.

Table 1. Basis concepts and notations in hypothesis test

	reject H_0	**fail to reject H_0**
Null H_0 is false	True positive	Type II error False negative
Null H_0 is true	Type I error False positive	True negative

(a)

	reject H_0	**fail to reject**	
H_0 false	*S*	*T*	d_1
H_0 true	*V*	*U*	d_0
	R	*W*	*d*

(b)

Instead of controlling the *p* value at the level α for each test, the familywise error rate is suggested to control the probability of making one or more Type I errors at the level α for all the hypotheses. However, this error rate is much too strict, especially when the number of hypotheses is large, and thus replaced by the false discovery rate (FDR) that is the expected proportion of false positive among all discoveries (rejected null hypotheses)[4]. As listed in a pair with the p-value in Tab.2, the q-value is actually the posteriori counterpart of the p-value from a Bayesian perspective [5,6]. With wide applications in big data analyses (e.g., genomics), FDR becomes a hot topic in the past one or two decades [4-6].

Table 2. Typical performance measures and Bayesian perspective

	Reject H_0 if ς falls in its rejection region Γ	**Fail to reject H_0** if $\varsigma \notin \Gamma$	
H_1 (H_0 is false)	$E[\frac{S}{d}] = \pi_1 p(\varsigma \in \Gamma \mid H_1)$ $= p(H_1 \mid \varsigma \in \Gamma) p(\varsigma \in \Gamma)$ Sensitivity or Power $\begin{cases} p(\varsigma \in \Gamma \mid H_1) = E[\frac{S}{d_1}] \geq 1-\beta \\ p(H_1 \mid \varsigma \in \Gamma) = E[\frac{S}{R}] \end{cases}$	$E[\frac{T}{d}] = \pi_1 p(\varsigma \notin \Gamma \mid H_1)$ $= p(H_1 \mid \varsigma \notin \Gamma) p(\varsigma \notin \Gamma)$ Type II error $\begin{cases} p(\varsigma \notin \Gamma \mid H_1) = E[\frac{T}{d_1}] \leq \beta \\ p(H_1 \mid \varsigma \notin \Gamma) = E[\frac{T}{W}] \end{cases}$	$\pi_1 = E[\frac{d_1}{d}]$
H_0 is true	$E[\frac{V}{d}] = \pi_0 p(\varsigma \in \Gamma \mid H_0)$ $= p(H_0 \mid \varsigma \in \Gamma) p(\varsigma \in \Gamma)$ Type I error $\begin{cases} p = p(\varsigma \in \Gamma \mid H_0) = E[\frac{V}{d_0}] \leq \alpha \\ q = p(H_0 \mid \varsigma \in \Gamma) = E[\frac{V}{R}] \end{cases}$	$E[\frac{U}{d}] = \pi_0 p(\varsigma \notin \Gamma \mid H_0)$ $= p(H_0 \mid \varsigma \notin \Gamma) p(\varsigma \notin \Gamma)$ Specificity $\begin{cases} p(\varsigma \notin \Gamma \mid H_0) = E[\frac{U}{d_0}] \geq 1-\alpha \\ p(H_0 \mid \varsigma \notin \Gamma) = E[\frac{U}{W}] \end{cases}$	$\pi_0 = E[\frac{d_0}{d}]$
	$E[\frac{R}{d}] = p(\varsigma \in \Gamma) =$ $\pi_1 p(\varsigma \in \Gamma \mid H_1) + \pi_0 p(\varsigma \in \Gamma \mid H_0)$	$E[\frac{W}{d}] = p(\varsigma \notin \Gamma) = 1 - p(\varsigma \in \Gamma)$	

The Bayesian perspective can be extended to all the rest concepts listed in Tab.2, that is, we may have the posteriori counterparts of Type II error, sensitivity, and specificity. It follows from $p(H_1|\varsigma \in \Gamma) = 1-p(H_0|\varsigma \in \Gamma)$ and $p(\varsigma \in \Gamma | H_1) = 1-p(\varsigma \notin \Gamma | H_1)$ that these measures are related, which provides not only insights but also alternative way to compute the q-value. When we get some structure about H_1, it is more feasible to estimate $p(\varsigma \in \Gamma | H_1)$ and thus its posteriori counterparts. Also, the q-value may be estimated from the likelihood ratio positive (LR+), which has been studied in diagnostic testing of evidence-based medicine. It follows from Tab.2 that we have

$$q = p(H_0 \mid \varsigma \in \Gamma) = \frac{\pi_0 p(\varsigma \in \Gamma \mid H_0)}{\pi_1 p(\varsigma \in \Gamma \mid H_1) + \pi_0 p(\varsigma \in \Gamma \mid H_0)} = \frac{1}{LR_+ \pi_1 / \pi_0 + 1}, \quad (1)$$

$$LR_+ = \frac{\text{sensitivity}}{1 - \text{specificity}} = \frac{p(\varsigma \in \Gamma \mid H_1)}{p(\varsigma \in \Gamma \mid H_0)} \text{ or } LR_+ = \frac{\pi_0}{\pi_1} \frac{1}{q} - 1,$$

from which we observe that a bigger LR_+ value actually means a smaller q value.

Next, we explain that the A5 formulation brings down the q-value significantly by A-3 and A-4, and enhances the performances of hypothesis test. For simplicity, we assume that each time every accumulator adds by a score 1 at A-3 and that the test per circling is independent from each other. Though this assumption is actually not true

because samples are randomly picked from the same $X^{(0)}$, $X^{(1)}$ at A-1, we may get some insights by this rough approximation.

We observe a particular cell $\omega_f \in \Omega$ that just reached a threshold τ after M times of circling during the A5 implementation. Precisely, the statistics in consideration is the set $\{\omega_f\}$ of all the cells in Ω with its region of rejection as follows:

$$\Gamma^{\tau} = \bigcup_{\omega_f \in \Omega} \Gamma^{\tau}_{\omega_f}, \quad \Gamma^{\tau}_{\omega_f} = [\tau, \tau+1, \ldots, +\infty), \tag{2}$$

which means that there is at least one $\omega_f \in \Gamma^{\tau}_{\omega f}$.

It follows from the last row of Tab.2 that the probability for $\{\omega_f\}$ to fall in its rejection region is given as follows:

$$P(\{\omega_f\} \in \Gamma^{\tau}) = C_M^{\tau} p^{\tau} (\varsigma \in \Gamma)[1 - p(\varsigma \in \Gamma)]^{M-\tau}, \tag{3}$$

and its corresponding P-value is correspondingly given as follows:

$$P = P(\{\omega_f\} \in \Gamma^{\tau} \mid H_0) = C_M^{\tau} p^{\tau} (1-p)^{M-\tau} < C_M^{\tau} p^{\tau}, \quad p = p(X \in \Gamma \mid H_0). \tag{4}$$

In other words, the score accumulation at A-3 can bring down the P-value when $C_M^{\tau} p^{\tau-1} < 1$, which is possible as long as the threshold τ becomes large enough.

Moreover, it is promising to observe the Q-value by the Bayes posteriori:

$$Q = P(H_0 \mid \{\omega_f\} \in \Gamma^{\tau}) = \frac{P(H_0) P(\{\omega_f\} \in \Gamma^{\tau} \mid H_0)}{P(\{\omega_f\} \in \Gamma^{\tau})} = \frac{\pi_0^M C_M^{\tau} p^{\tau} (1-p)^{M-\tau}}{C_M^{\tau} p^{\tau} (\varsigma \in \Gamma)[1 - p(\varsigma \in \Gamma)]^{M-\tau}} \tag{5}$$

$$= q^{\tau} p(H_0 \mid X \notin \Gamma)^{M-\tau} < q^{\tau}, \quad \text{with } P(H_0) = \pi_0^M \text{ and } q = p(H_0 \mid X \in \Gamma).$$

That is, score accumulation indeed brings down the Q-value simply as long as $\tau > 1$.

Such accumulated reliability increases as the threshold τ becomes larger. However, there is no free lunch. It follows from Eq.(3) that the probability of rejecting H_0 also reduces considerably, which leads to either the risk of missing significant features (i.e., detecting power reduced) for a fixed number M or a large computing cost for keeping A5 circling until finding enough significant features.

4 Sample Pairing Delta and Case Control Study

Hypothesis test focuses at evaluating a deviation from the hull assumption:

$$H_0: \text{ there is no difference between } P^{(0)} \text{ and } P^{(1)}. \tag{6}$$

Taking multivariate Gaussian population as an example, we consider

$$X^{(\ell)} = [x_1^{(\ell)}, \ldots, x_{N_\ell}^{(\ell)}], \quad \ell = 0,1, \quad x_t^{(\ell)} \, from \, G(x \mid \mu^{(\ell)}, \Sigma). \tag{7}$$

Under the null hypothesis, we get the following maximum likelihood estimation:

$$\mu = \alpha^{(0)} \mu^{(0)} + \alpha^{(1)} \mu^{(1)}, \quad \Sigma = \alpha^{(0)} \Sigma^{(0)} + \alpha^{(1)} \Sigma^{(1)}, \quad \alpha^{(\ell)} = N_\ell / N, \; N = \sum_\ell N_\ell,$$

$$\mu^{(\ell)} = \frac{1}{N_\ell} \sum_{t=1}^{N_\ell} x_t^{(\ell)}, \; \Sigma^{(\ell)} = \frac{1}{N_\ell - 1} \sum_{t=1}^{N_\ell} (x_t - \mu^{(\ell)})(x_t - \mu^{(\ell)})^T. \tag{8}$$

Typically, the deviation from H_0 is measured by the Hotelling T^2 statistics [17]:

$$\begin{aligned} T^2(\theta_0,\theta_1) &= \alpha_0\alpha_1(\mu^{(0)}-\mu^{(1)})^{\mathrm{T}}\Sigma^{-1}(\mu^{(0)}-\mu^{(1)}) \\ &= \frac{\alpha_0\alpha_1}{\alpha_0^2+\alpha_1^2}[(\mu^{(1)}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mu^{(1)}-\mu)+(\mu^{(0)}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mu^{(0)}-\mu)],\ \theta_i=\{\alpha_i,\mu^{(i)},\Sigma^{(i)}\}, \end{aligned} \tag{9}$$

which measures the deviation of $P^{(0)}$, $P^{(1)}$ from $G(x|\mu,\Sigma)$ as a reference about H_0.

Other than parametric statistics, the Kolmogorov-Smirnov test and the Mann–Whitney U test are two general nonparametric methods to test whether samples come from the same distribution [15]. The former uses the maximal distance between cumulative frequency distributions of these two samples as the statistics, while the latter takes the difference between mean ranks of these two samples as the statistics.

Here, we propose another nonparametric statistics as follows:

$$\begin{aligned} &\zeta_A = D(X^{(0)}\|X^{(1)}), \\ &D(A\|B)=\frac{1}{\#A\ \#B}\sum_{x\in A}\sum_{y\in B}\delta^T(x,y)\delta(x,y),\quad x=[a_1,...,a_d],\ y=[b_1,...,b_d], \end{aligned} \tag{10}$$

where $\delta(x,y)$ indicates the variation by pairing the sample x with y, and is called sample-pairing-delta. One special case considers element by element independently:

$$\delta(x,y)=[\delta_1(a_1,b_1),\cdots,\delta_d(a_d,b_d)]^T,$$

where $\delta_j(a_j,b_j)$ measures the variation from a_j to b_j. One example is that elements are homogenous up to unknown scales w_j's, that is, we have

$$\delta_j(a_j,b_j)=w_j\delta(a_j,b_j). \tag{11}$$

In general, each element of $\delta(x,y)$ depends on vector x and vector y. Typically, we may consider that each element comes from $\delta(a_j,b_j)$, $j=1,\ldots,d$ by a linear map

$$\delta(x,y)=W[\delta(a_1,b_1),\cdots,\delta(a_d,b_d)]^T, \tag{12}$$

where W is a transformation matrix. Particularly, when $\delta(x,y)=x-y$ and $W=\Sigma^{-0.5}$, we get the following Mahalanobis distance:

$$\delta^T(x,y)\delta(x,y)=(x-y)^T\Sigma^{-1}(x-y). \tag{13}$$

This sample-pairing-delta $\delta(x,y)$ based statistics ζ_A provides a practical facility too. In real applications, there are frequently samples that have elements with missing values. Traditionally, we either discard such a sample or fill the missing values by certain estimation of the missing value based on other samples. On one hand, discarding the sample makes the small sample size problem become more serious, as encountered in bioinformatics. On the other hand, it is difficult to estimate a missing value well and usually a rough estimation leads to a bad performance.

We believe that a best policy is letting the affect of missing value to other parts of computation to be as less as possible. E.g., in the SNP analysis [13], $\delta_j(a_j,b_j)$ measures the extent of a variation from a_j into b_j. Typically, a SNP variation occurs rarely and thus if it is missed in detection we can conservatively regard that there is no variation. That is, we may simply consider a rectified scale in Eq.(11) by

$$w_j \begin{cases} = 0, & \text{at least one of } a_j, b_j \text{ is missing,} \\ = 0, & \text{when } a_j = b_j, \\ \neq 0, & \text{otherwise.} \end{cases} \tag{14}$$

The simplest case is $w_j=1$, or the degenerated case $W=I$ in Eq.(12). In general, W in Eq.(12) may consider the dependence cross elements by

$$W = C_v^{-0.5}, \quad C_v = C_v(X^{(0)} \cup X^{(1)} \| X^{(0)} \cup X^{(1)}), \tag{15}$$

$$C_v(A \| B) = \frac{1}{\#A\#B} \sum_{x \in A} \sum_{y \in B} [\delta_1(a_1, b_1), \cdots, \delta_d(a_d, b_d)][\delta_1(a_1, b_1), \cdots, \delta_d(a_d, b_d)]^T \text{ at } W=I,$$

where the matrix C_v measures co-variations across elements. Assuming independence cross elements, W could be given as follows

$$W = diag[w_1, \cdots, w_d] = diag[C_v^{-0.5}], \text{ or even simply } \quad W = (d / \sqrt{Trace[C_v]})I. \tag{16}$$

In implementation, we usually do not know the distribution of ζ_A and thus cannot estimate the p-value in a standard way. Still, we may approximately estimate the p-value by a Monte Carlo simulation under the null H_0 by Eq.(6), e.g., we make the following permutation test:

$$p_{value} = \frac{1}{\#\Pi} \{1 + \sum_{\pi \in \Pi} I[D(X_\pi^{(0)} \| X_\pi^{(1)}) > D(X^{(0)} \| X^{(1)})]\}, \tag{17}$$

where $I(a>b)=1$ if $a>b$, otherwise $I(a>b)=0$, and Π is the set of all the possible permutations, with each $\pi \in \Pi$ turning out $X_\pi^{(0)}$, $X_\pi^{(1)}$ [11].

Next, we proceed to address other favorable features of the statistics ζ_A by Eq.(10) in comparison with typical studies of hypothesis test.

First, in a standard way, a statistics and its distribution are derived under the null H_0 by Eq.(6), after which the value of this statistics is computed from samples and a p-value is estimated according to the distribution to test whether this H_0 breaks significantly. However, it is challenging to estimate the distribution of this statistics and to compute the p-value according to this distribution, which makes the standard approaches of hypothesis test limited to only a few commonly assumed distributions. In contrast, the statistics ζ_A by Eq.(10) makes hypothesis test performed in an alternative way that directly measures the difference between $P^{(0)}$ and $P^{(1)}$ regardless the null H_0, and compute the p-value by a Monte Carlo simulation under the null H_0 without estimating the distribution of ζ_A.

A parametric statistics may also be obtained in a similar way. Firstly in [10] and subsequently in [1], the following Kullback–Leibler (KL) divergence is suggested

$$\varsigma_{KL} = KL(p(X \mid \Theta^{(1)}) \| p(X \mid \Theta^{(0)})), \ \mathrm{KL(p\| q)} = \int p(x) \ln[p(x)/q(x)] dx, \tag{18}$$

as a general formulation of statistics that aims at the difference between $P^{(0)}$ and $P^{(1)}$, where $\Theta^{(1)}$ is an estimation of Θ based on $X^{(1)}$ while $\Theta^{(0)}$ is an estimation of Θ on $X^{(0)}$. By the way, getting $\Theta^{(0)}$ estimated on both $X^{(0)}$ and $X^{(1)}$, ζ_{KL} also leads to statistics in a standard sense. E.g., let q, p given by $G(x|\mu^{(1)}, \Sigma)$, $G(x|\mu^{(0)}, \Sigma)$ in Eq.(18), we are lead to the Hotelling statistics by Eq.(9), with details referred to [1] (especially p870).

Second, typical existing statistics, derived either parametrically (e.g., the Hotelling T^2 statistics [17] and those frequency table based tests used for SNP analyses) or non-parametrically (e.g., the Kolmogorov-Smirnov test and the U test), shares a common feature. That is, the overall structure or description of $P^{(i)}$ is firstly summarized from its own samples, respectively for $i=0,1$, and then, the summarized overall structure of $P^{(1)}$ is compared with the summarized overall structure of $P^{(0)}$ to detect whether there is a significant difference. In contrast, the statistics ζ_A by Eq.(10) starts to detect delta $\delta(x,y)$ (i.e., difference) between paired samples x,y and then summarize these deltas for detecting a significant difference between $X^{(0)}$ and $X^{(1)}$.

Such a sample-pairing-delta based statistics ζ_A may have some asymptotic relation to a standard *hypothesis test* method. For the example by Eq.(13), it can be analytically shown that as the sample size $N^{(0)}+N^{(1)}$ becomes large enough we have

$$\zeta_A - d \to T^2(\theta_0, \theta_1) \text{ as } N^{(1)} + N^{(0)} \to \infty, \tag{19}$$

where $T^2(\theta_0,\theta_1)$ is the Hotelling statistics by Eq.(9), i.e., the difference between means prevails. In other words, ζ_A by Eq.(10) differs from the standard hypothesis test especially for detecting population difference on a finite size of samples.

Third, typical statistics such as the Hotelling T^2 statistics and those frequency table based statistics actually test a deviation from the null H_0 by Eq.(6) via testing the differences between mean values subject to certain constraints. That is, the statistics becomes zero when there is no difference between mean values. In contrast, ζ_A by Eq.(10) will be still a nonzero value $\zeta_0 \neq 0$ in such a case of no difference. For some distributions, e.g., $G(x|\mu^{(1)},\Sigma)=G(x|\mu^{(0)},\Sigma)$, this $\zeta_0 \neq 0$ tends to asymptotically a constant as shown in Eq.(19) and thus does not contain any useful information about difference between $P^{(0)}$ and $P^{(1)}$. Even so, it has no bad effect on the permutation test by Eq.(17) because inequality will not be affected by adding a constant in both the sides. For two populations $P^{(0)}$ and $P^{(1)}$ with no difference between mean values but still some higher order difference, ζ_A by Eq.(10) may detect some information about this difference. In other word, this ζ_A is more powerful than those mean-value based statistics for detecting a deviation from the null H_0 by Eq.(6).

The last but not least, we consider the case-control studies that are widely encountered in real applications of hypothesis test. The samples $X^{(0)}$ of $P^{(0)}$ come from a normal population as benchmark or called control. There is usually a reference point about the normal by which we calibrate each feature to represent its deviation from the corresponding reference. For control samples, deviations from the reference are usually isotropic random noises with zero mean. On the other hand, samples $X^{(1)}$ of $P^{(1)}$ come from one abnormal population or called case, with a systematic deviation from the reference. The task of case-control studies aims at detecting this systematic deviation. Typical statistics, especially those based on the difference between mean values, consider $P^{(0)}$ and $P^{(1)}$ in a same distribution form $p(X|\Theta)$ but different in the values that Θ takes. With help of Eq.(18), we may further proceed to consider $P^{(0)}$ and $P^{(1)}$ in different distribution forms.

Moreover, extensions may also be made to consider that the reference may not necessarily represent the normal and that deviations of control samples from the reference may not necessarily isotropic noises. Taking GWAS analysis [13] as an example, a case sample takes a symbol η and a control sample takes a symbol ξ at one SNP site. At this site, both η and ξ may take the value 0 for *SS* as a reference, the value 1 for the variation *Ss,* and the value 2 for the variation *ss.*

Generally speaking, deviations from the reference can be classified into two types, one by control samples and one by case samples. It is insightful to watch which type dominates, for which we further look into the following two scenarios:

(1) $P(\eta>\xi)$, i.e., the case samples deviate more badly than control samples do. E.g., for the SNP analysis, a disease is likely caused by certain variations. We may measure the difference between $X^{(0)}$ and $X^{(1)}$ in this scenario by

$$\zeta_A^{1>0}=\sum_j w_j\beta_j^{(0)}\beta_j^{(1)}D_j^{1>0},\ \ D_j^{1>0}=\frac{1}{N_j^{(0)}N_j^{(1)}}\sum_{\xi\in X_j^{(0)}}\sum_{\eta\in X_j^{(1)}}P(\eta>\xi)\delta^2(\xi,\eta),$$

$$\beta_j^{(i)}=N_j^{(i)}/\sum_j N_j^{(i)},\ \ w_j=1/D_{X_j^{(0)}\cup X_j^{(1)}},\ \ D_A=\frac{0.5}{\#A\#A}\sum_{\xi\in A}\sum_{\eta\in A}P(\xi\neq\eta)\delta^2(\xi,\eta), \tag{20}$$

$$N_j^{(i)}=\#X_j^{(i)}\ \text{(excluding missing elements)},\quad X_j^{(i)}\ \text{from the } j\text{th column of } X^{(i)}.$$

The vectors x,y in Eq.(10) correspond to one pair of samples from one computational unit (e.g., a gene), with the jth element pair ξ and η of x,y representing the jth site.

(2) $P(\eta<\xi)$, i.e., the control samples deviate more badly than case samples. E.g., the normal population experienced variations to adapt an environmental change, while a disease is likely caused due to a lack of such variations. We may similarly get $\zeta_A^{0>1}$ simply with $\eta>\xi$ in Eq.(20) replaced by $\eta<\xi$ while $1>0$ in Eq.(20) replaced by $0<1$.

Moreover, we may also get

$$D(X^{(0)}\|X^{(1)})=\zeta_A^{1>0}+\zeta_A^{0>1}=\sum_j w_j\beta_j^{(0)}\beta_j^{(1)}D_j,$$
$$D_j=\frac{1}{N_j^{(0)}N_j^{(1)}}\sum_{\xi\in X_j^{(0)}}\sum_{\eta\in X_j^{(1)}}P(\eta\neq\xi)\delta^2(\xi,\eta). \tag{21}$$

Finally, we may put each of the above $\zeta_A^{0>0}$, $\zeta_A^{0>1}$ and $D(X^{(0)}\|X^{(1)})$ into Eq.(17) to implement permutation test.

5 Boundary Based Statistics

The task of discriminant analysis is selecting a subset of features on which we observe either a best separation between $P^{(0)}$ and $P^{(1)}$ in term of *classification* or a significant overall difference between $P^{(0)}$ and $P^{(1)}$ in term of *hypothesis test*.

For a best separation, we consider a boundary that separates samples from $P^{(0)}$ and $P^{(1)}$ with a least number of misclassified samples, where a misclassified sample is one that comes from one population but be classified into the other. Whether two populations could be well separated relates to whether there is a significant overall difference between the populations. However, two concepts are not same. Also, classification and hypothesis test are implemented under different performance measures that are not monotonically related. The one for classification focuses on boundary separation, while the one for hypothesis test focuses on difference of the overall structures. A best performance for one may not be necessarily the best for the

other. Conventionally, classification and hypothesis test are studied separately. Here, we attempt to integrate the two subtasks by reexamining the role of discriminant boundary in getting statistics for hypothesis test.

Observing Eq.(8), μ actually acts as a boundary that separates two populations, and T^2 measures the Mahalanobis distance to this boundary from both the sides of $\mu^{(0)}$ and $\mu^{(1)}$. A boundary is also implied between x and y in Eq.(13). However, such a rough boundary only results in a rough separation of $P^{(0)}$ and $P^{(1)}$. Thus, we are motivated to embed an optimal best boundary into one statistics in order to judge whether two populations are significantly different.

One rather straightforward way is a two step implementation as follows.

Step 1: use a Bayes classifier or one alternative to separate samples into

$$X^{(0)} = X_1^{(0)} \bigcup X_0^{(0)}, \quad X^{(1)} = X_0^{(1)} \bigcup X_1^{(1)}, \text{ with a confusion matrix } \begin{bmatrix} n_{00}, & n_{01} \\ n_{10}, & n_{11} \end{bmatrix}, \tag{22}$$

where $X_i^{(j)}$ consists of samples that come from $X^{(j)}$ and classified into the population $P^{(i)}$, with $n_{ji} = \#X_i^{(j)}$.

Step 2: measure the deviation from the null hypothesis H_0 by

$$\zeta_B = \frac{n_{00}n_{11}D(X_0^{(0)} \| X_1^{(1)})}{n_{00}n_{01}D(X_0^{(0)} \| X_1^{(0)}) + n_{10}n_{11}D(X_0^{(1)} \| X_1^{(1)})}, \tag{23}$$

where $D(A\|B)$ is same as defined in Eq.(10). Alternatively, we may replace this nonparametric statistics by a parametric statistics. Considering the asymptotic approximation by Eq.(19), we may let that

$$D(X_0^{(j)} \| X_1^{(\ell)}) \text{ is replaced with } T^2(\theta_0^{(j)}, \theta_1^{(\ell)}) + d, \tag{24}$$

where $\theta_i^{(j)}$ is a maximum likelihood estimation of a Gaussian distribution based on the subset $X_i^{(j)}$ of samples. It follows from Eqn.(5) in [1] that we may also use

$$T^2(\theta_0^{(j)}, \theta_1^{(\ell)}) = \frac{n_{j0}n_{\ell 1}}{(n_{j0} + n_{\ell 1})^2} KL(X_0^{(j)} \| X_1^{(\ell)}). \tag{25}$$

Still, the above two step implementation is made under two different measures, i.e., misclassification at Step 1 and ζ_B at Step 2.

Instead, we can also use a same measure to implement two steps coordinately, for which we estimate θ to maximize a parametric $\zeta_B(\theta)$ that varies with a discriminant boundary described by a parameter θ.

Given a boundary $f(x,\theta)=0$ such that a sample x is classified into $P^{(1)}$ if $f(x,\theta)>0$ and into $P^{(0)}$ if $f(x,\theta)<0$, one example of such a parametric $\zeta_B(\theta)$ is given as follows:

$$\begin{aligned} &\zeta_B(\theta) = D_{separa}(\theta) / D_{confus}(\theta), \\ &D_{confus}(\theta) = \sum_{y \in X^{(1)},\ f(y,\theta)<0} s(d_f(y,\theta)) + \sum_{x \in X^{(0)},\ f(x,\theta)>0} s(d_f(x,\theta)), \\ &D_{separa}(\theta) = \sum_{y \in X^{(1)},\ f(y,\theta)>0} s(d_f(y,\theta)) + \sum_{x \in X^{(0)},\ f(x,\theta)<0} s(d_f(x,\theta)), \end{aligned} \tag{26}$$

where $s(r)$ is a monotonically increasing function with respect to a scalar variable r, and $d_f(x,\theta)$ denotes the shortest distance of a sample x to $f(x,\theta)=0$ by

$$d_f(x,\theta) = \sqrt{\min_{u,\ f(u,\theta)=0} \|x-u\|^2}. \tag{27}$$

When $s(r)=0$ for $r=0$ otherwise $s(r)=1$ for $r>0$, it follows from Eq.(26) that

$$\zeta_B(\theta) = (n_{11}+n_{00})/(n_{10}+n_{01}) = -1+(n_{10}+n_{01})^{-1}, \tag{28}$$

which is monotonically decreasing with respect to the misclassification error, and its maximization is equivalent to minimizing the misclassification. On the other hand, when $s(r)= r^2$, $\zeta_B(\theta)$ by Eq.(26) is conceptually a counterpart of ζ_A by Eq.(10), Eq.(15), and Eq.(16). In other words, $\zeta_B(\theta)$ by Eq.(26) covers *classification* and *hypothesis test* as two special cases, and generally trades off the natures of both.

We may emphasize one over other by different choices of $s(r)$, e.g., we consider

$$s(r) = r^{\alpha}, \quad \alpha > 0, \tag{29}$$

for which we are leaded to hypothesis test as $\alpha=2$ and to classification as $\alpha\to 0$, as well as to one intermediate case when α takes a value between $(0,2]$.

In implementation, one bottleneck is solving Eq.(27) efficiently, though we simply have

$$d_f(x,\theta) = (x-c)^T w/\|w\|, \text{ for a linear boundary } f(u,\theta)= w^T(u-c)=0. \tag{30}$$

Indirectly, $d_f(x,\theta)$ is able to be solved by the Lagrangian for a quadratic function $f(x,\theta)=0$ [12], for which one example is a Bayes classifier with the boundary below

$$f(x,\theta) = \ln[\alpha_1 G(x\,|\,\mu^{(1)},\Sigma^{(1)})] - \ln[\alpha_0 G(x\,|\,\mu^{(0)},\Sigma^{(0)})] = 0,$$

which is a quadratic equation of x and degenerates to a linear equation when $\Sigma^{(0)}= \Sigma^{(1)}$.

Instead of considering $f(x,\theta)=0$ as a linear or quadratic equation globally in the sample space, we may model a discriminant boundary by a mixture of linear functions $f(x,\theta_j)=0$, $j=1,2,\ldots,m$ as follows:

$$\begin{aligned} \zeta_B(\theta_j) &= \textstyle\sum_j \zeta_B(\theta_j), \quad \zeta_B(\theta_j) = D_{separa}(\theta_j)/D_{confus}(\theta_j), \\ D_{confus}(\theta_j) &= \textstyle\sum_{y\in X_k^{(1)},\ f(y,\theta_j)<0} s(d_f(y,\theta_j)) + \sum_{x\in X_k^{(0)},\ f(x,\theta_j)>0} s(d_f(x,\theta_j)), \\ D_{separa}(\theta_j) &= \textstyle\sum_{y\in X_k^{(1)},\ f(y,\theta_j)>0} s(d_f(y,\theta_j)) + \sum_{x\in X_k^{(0)},\ f(x,\theta_j)<0} s(d_f(x,\theta_j)). \end{aligned} \tag{31}$$

That is, for each $f(x,\theta_j)=0$ we concentrate on considering the k nearest samples of $X^{(0)}$, $X^{(1)}$ to $f(x,\theta_j)=0$ as follows:

$$X_k^{(i)}(\theta_j) = \{k \text{ samples of } X^{(i)} \text{ with the minimum distrances to } f(x,\theta_j)=0\},\ i=0,1.$$

The overall separation of $X^{(0)}$, $X^{(1)}$ is described by combining m local liner discriminant boundaries. For a small k, the focus is on a border based misclassification with some ignorance of overall difference between populations. In other words, we get a better classification but a weak detecting power for hypothesis test. As k increases, the focus gradually switches to increasing the detecting power on the difference in overall structure, but suffering some classification accuracy. The value of k trades off between classification and hypothesis test. Also, we may control $s(r)$ by Eq.(29) for a similar role. E.g., $\alpha<0$ discounts samples away from the border.

For a classification task, samples around the border take major role on estimating the boundary structure between populations, while samples away from the border take a role of regularizing the estimation especially when there are few boundary samples.

How to decide an appropriate k in Eq.(31) remains a challenge. We start from a small k to iteratively maximize $\zeta_B(\theta)$ for learning θ, with k gradually increasing. Then, with the resulted θ, we test hypotheses on $X^{(0)}$, $X^{(1)}$ as k becomes large enough.

For a hypothesis testing task, we start from a large k and gradually reduce to a small k for an improved classification performance.

The statistics by Eq.(26) is different from ones by Eq.(9) and Eq.(10) that are made directly in the apparent domain *(shortly A-domain)*, where samples are directly observed and overall structures are compared. Also, this statistics differs from ones that is featured by statistics computed indirectly in an inner domain *(shortly I-domain)* where samples are mapped into and where misclassification or separation is measured [1,10]. Instead, ζ_B by Eq.(26) integrates both the statistics $D(\cdot\|\cdot)$ in the *A-domain* and the statistics n_{ji} in an inner decision domain to measure the differences between two populations.

Last but not least, we need further theoretical understanding on the third issue in Sect.1. Namely, we want to know whether *best classification* and *best hypothesis test* will become asymptotically equivalent as the sample size tends to infinite, subject to a moderate condition or can be achieved under a same performance measure, e.g., under ζ_B by Eq.(31) with a same value of k ?

Acknowledgment. This work was supported by a CUHK Direct grant for 2013-2014.

References

1. Xu, L.: Matrix-variate discriminative analysis, integrative hypothesis testing, and geno-pheno A5 analyzer. In: Yang, J., Fang, F., Sun, C. (eds.) IScIDE 2012. LNCS, vol. 7751, pp. 866–875. Springer, Heidelberg (2013)
2. Xu, L.: A unified perspective and new results on RHT computing, mixture based learning, and multi-learner based problem solving. Pattern Recognition 40, 2129–2153 (2007)
3. Xu, L., Oja, E.: Randomized Hough transform. In: Encyclopedia of Artificial Intelligence, pp. 1354–1361. IGI Global, Hershey (2008)
4. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. of Royal Statistical Society B 57(1), 289–300 (1995)
5. Storey, J.D.: A direct approach to false discovery rates. Journal of the Royal Statistical Society, Series B 64(3), 479–498 (2002)
6. Storey, J.D., Tibshirani, R.: Statistical significance for genome-wide studies. Proc. of the National Academy of Sciences 100(16), 9440–9445 (2003)
7. Glezko, G.V., Emmert-Streib, F.: Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets. Bioinformatics 25, 2348–2354 (2009)
8. Alves, G., Yu, Y.: Combining independent, weighted p-values: achieving computat. stability by a systematic expansion with controllable accuracy. PLoS ONE 6(8), e22647 (2011)
9. Zaykin, D.V.: Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. J. Evol. Biol. 24(8), 1836–1841 (2011)

10. Xu, L.: Semi-Blind Bilinear Matrix System, BYY Harmony Learning, and Gene Analysis Applications. In: Proc. of 6th International Conf. on New Trends in Information Science, Service Science and Data Mining, Taipei, October 23-25, pp. 661–666 (2012)
11. Good, P.I.: Permutation, Parametric and Bootstrap Tests of Hypotheses. Springer (2005)
12. Liu, Z.Y., Qiao, H., Xu, L.: Multisets mixture learning based ellipse detection. Pattern Recognition, 39731–39735 (2006)
13. Bansal, V., Libiger, O., Torkamani, A., Schork, N.J.: Statistical analysis strategies for association studies involving rare variants. Nature Reviews Genetics 11, 773–785 (2010)
14. Clemmensen, L., Hastie, T., Witten, D., Ersbøll, B.: Sparse discriminant analysis. Technometrics 53, 406–413 (2011)
15. Bagdonavicius, V., Kruopis, J., Nikulin, M.S.: Non-parametric tests for complete data. ISTE & WILEY, London & Hoboken (2011)
16. Xu, L.: Data smoothing regularization, multi-sets-learning, and problem solving strategies. Neural Networks 16, 817–825 (2003/2012)
17. Hotelling, H.: The generalization of Student's ratio. Annals of Mathematical Statistics 2(3), 360–378 (1931)