Xu *Appl Inform* (2016) 3:1
DOI 10.1186/s40535-015-0016-4

Applied Informatics
a SpringerOpen Journal

RESEARCH Open Access

A new multivariate test formulation: theory, implementation, and applications to genome-scale sequencing and expression

Lei Xu[1,2*]

*Correspondence: lxu@cse.cuhk.edu.hk; lxu@cs.sjtu.edu.cn
[2] Department of Computer Science and Engineering, Centre for Brain-inspired Computing and Bio-Health Informatics, The School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, SEIEE Building 3, 800 Dongchuan Road, Minhang District, 200240 Shanghai, China
Full list of author information is available at the end of the article

Abstract

A new formulation is proposed for multivariate test, consisting of not only a hierarchy of numerous tests organised in a lattice taxonomy of properties that come from different combinations of multi-variates and represent different factors associated with the rejection of null hypothesis, but also by a theory of property-oriented rejection. Located on the bottom level of this taxonomy is a conventional formulation of multivariate test, featured by a property with the weakest collegiality and a rejection with the largest p value. From one level up to the next, the dimension of rejection increases, the collegiality of properties strengthen, and the p values reduce, until the top level that is featured by a property with the strongest collegiality and a rejection with the smallest p value. Instead of traversing all the combinations in the taxonomy, an easy implementation is developed to identify distinctive properties by the best first path (BFP) in a lattice taxonomy of an appropriate number of intrinsic factors that are obtained after decoupling second-order dependence cross multivariate statistics and discarding those non-distinctive components. Even away off this BFP, if needed, a particular combination of intrinsic factors may be conveniently tested in such a taxonomy too. Moreover, further improvement is made by considering some dependence of higher than second order, with the top level p value refined into one upper bound that is obtained by directional test. Furthermore, detailed implementations are also provided for applications to genome-scale sequencing and expression, with particular emphasis on multivariate phenotype-targeted test for expression profile analyses.

Keywords: Multivariate test, Lattice taxonomy, Intrinsic factors, Property-oriented rejection, Best first path, Dependence decoupling, Directional test, Case–control study, Two-variate PT test

Background

Statistical test takes a crucial role in many tasks of machine learning or data informatics in general. Classically, a univariate test is considered in various case–control problems and particularly in finding susceptibility SNP in computational genomics. Recently, there have been an increasing demand on multivariate test for jointly detecting multiple SNPs or variables. In these studies, a basic sampling unit is a vector $\mathbf{x}_t = [\, x_t^{(1)}, \ldots, x_t^{(d)}]^T$ from a population, and the problem is testing a null hypothesis H_0 that there is a normality underlying a set of sample vectors $\mathcal{X}_N = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ from this population. Typically, a

statistics is computed from $\mathcal{X}_N$ to test whether H_0 breaks significantly. When the population is described by a parametric model, the null hypothesis H_0 is typically represented as follows

$$H_0 : \theta = \mathbf{0}, \tag{1}$$

where θ is a vector with its parameters from either a part of Θ or a function $h(\Theta)$.

Illustrated in Fig. 1 are two typical examples of multivariate test, coming from the case–control study. Given two populations of vector-variate samples $X_\omega = \{\mathbf{x}_{t,\omega}, t = 1, \ldots, N_\omega, \omega = 0, 1\}$, where one with $\omega = 1$ is called the case population while the one with $\omega = 0$ is called the control population. One important task is examining whether there is a significant difference between two populations of samples. Shown in Fig. 1a is an example of two-sample test in a two-dimensional data space. Test is made on Eq. (1) with $\theta = \mathbf{c}_1 - \mathbf{c}_0$ that has a multivariate Gaussian distribution when $\mathbf{c}_1, \mathbf{c}_0$ are estimated by sample means. Shown in Fig. 1b is an example of testing logistic regression. Test is made on Eq.(1) on the coefficient vector $\theta = \beta$ that also has a multivariate Gaussian distribution when β is a maximum likelihood estimate.

Several review papers are available on recent developments of multivariate test for detecting multiple variants in GWAS (Bansal et al. 2010; Ferguson et al. 2013). From a different perspective, a brief overview is provided with four threads on the existing methods of multivariate test according to their logical and historical traces.

One originated from using the following Hotelling's statistics (Hotelling 1931):

$$T^2 = \frac{N_0 N_1}{N}(\mathbf{c}_1 - \mathbf{c}_0)^T \Sigma^{-1}(\mathbf{c}_1 - \mathbf{c}_0), \ \Sigma = \alpha_0 \Sigma_0 + \alpha_1 \Sigma_1, \quad N = N_0 + N_1, \tag{2}$$

for the above two sample test. The dependence across multiple variates, or called the linkage disequilibrium (LD) in biology, is considered by the covariance matrix Σ. However, this Σ is poorly estimated when the sample size is small while the number of variables is usually large. Unfortunately, this is often the case. To address the problem, Dempster proposed a non-exact test based on a χ^2 approximation (Dempster 1958, 1960). Also, there are other efforts, e.g. a simplified test with an equivalent asymptotic

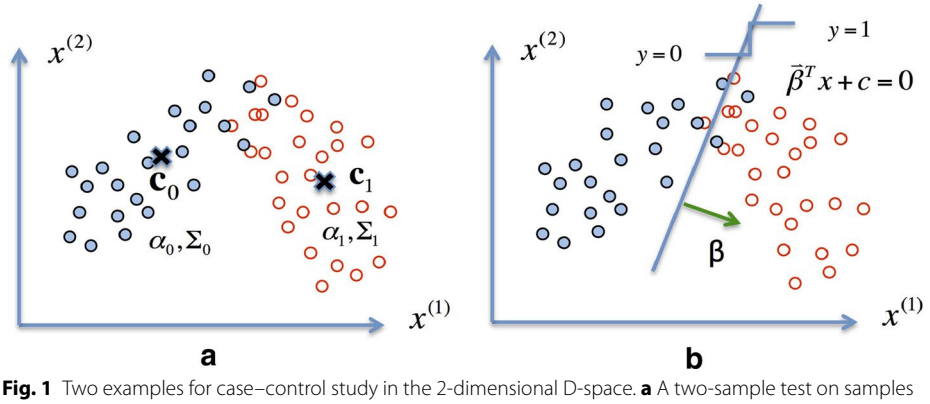


Fig. 1 Two examples for case–control study in the 2-dimensional D-space. **a** A two-sample test on samples of two populations with the mean vectors $\mathbf{c}_0, \mathbf{c}_1$ and the covariance matrices Σ_0, Σ_1, as well as the proportions α_0, α_1. **b** A test on regression coefficients of a logistic regression

power to Dempster's test (BaiZ 1996) and a generalised Hotelling's T^2 statistic that converges to a normal distribution after a suitable standardisation (Srivastava 2007).

In these methods, asymptotic approximations are used for deriving the significance levels of the test statistics under the null. However, when sample size is small and/or the data has a high missing rate, the null distributions of the test statistics may differ substantially from their asymptotic approximations. Therefore, these studies remain to be theoretical. Instead, real efforts either use the Hotelling's T^2 test directly (Fan and Knapp 2003) or impose some structure on the covariance matrix (Swanson et al. 2013), and even approximately simplify the covariance matrix into diagonal one but its ability of encoding LD information is lost (Kiezun et al. 2012).

The second thread is featured by the extensions of Wald test and Score test for jointly multiple hypotheses on single/multiple parameters, and efforts along this thread are widely encountered in studies of computational genomics. For a two sample test shown in Fig. 1a, it actually leads to the above first thread. Generally, the task is encountered in testing Eq. (1) on the coefficients $\theta = \beta$ of multivariate linear regression, multivariate logistic regression, and multivariate linear mixed model (Gudmundsson et al. 2012; Demidenko 2013; Zhou and Stephens 2014; Adhikari et al. 2015), as well as Cox regression analysis (Li and Gui 2004). Still, a use of the Fisher information matrix shares with the same problems caused by the covariance matrix. The problems will remain, though using multivariate F-test and likelihood ratio test on Eq. (1) may help to improve the Wald test and score test.

The third thread originated from Fisher's combined probability test for combining p values (Fishe 1932). However, each p value is merely a positive number that indicates the false alarm probability, already losing useful information such as an overall estimate of effect size, the direction of effects, and the dependence across effects. Progresses have been made by transforming p values into Z statistics or others on which some missing issues may be considered (Zaykin 2011), without or with help of information computed directly from datasets. Applied to rare variants, efforts made along this thread are typically referred under the term Meta-analysis (Evangelou and Ioannidis 2013).

The newest thread is featured by efforts in recent years for extending the existing GWAS from single variant to multiple rare variants. The basic idea is to let multiple variants of a unit (e.g. gene, exome, or one other biological unit) to be collapsed or summed up into a single one (Li and Leal 2008). Further developments are featured by various weighted sums via fixed weights or thresholds (Morgenthaler and Thilly 2007; Chapman and Whittaker 2008; WuM et al. 2011; Han and Pan 2010; Lee et al. 2012; Morris 2010; Price and Kryukov 2010). Turning multiple variants into a single one, the dimension of covariance matrix is thus reduced to lessen the problem of the above first thread. Most of these efforts are associated with a generalised linear regression model to test the null hypothesis that either the regression coefficients are zero or their variances are zero (Chapman and Whittaker 2008; WuM et al. 2011; Han and Pan 2010; Lee et al. 2012; Morris 2010; Price and Kryukov 2010). These studies are recently summarised under the names of burden tests and non-burden tests. Burden tests assume that all the variants in the target region have effects on the phenotype in the same direction and of similar magnitude, while non-burden tests cover various extensions beyond the assumption, for which details are further referred to one recent review (Lee et al. 2014).

In the rest of this paper, we start at discussing two limitations of the existing methods for multivariate tests, and then address the following main contexts:

1. A new multivariate test formulation that is featured not only by a hierarchy of numerous tests organised in a lattice taxonomy of properties that represent different causes and different dimensions of the null hypothesis rejection, but also by a theory of property-oriented rejection.
2. An easy implementation that identifies distinctive properties by the best first path (BFP) in a lattice taxonomy that comes from an appropriate number of intrinsic factors by decoupling second-order dependence cross multivariate statistics and discarding those non-distinctive components. Also, a particular combination that does not locate on this BFP may also be conveniently tested in such an taxonomy, if needed.
3. A further improvement is made by considering some dependence of higher than second order, with the p value of the top level refined into one upper bound obtained by a directional rejection based on a vectorial property possessed by the alarm in evaluation.

Finally, we discuss several potential applications to expression profile-based biomarker identification and exome sequencing-based joint SNV detection.

Methods

Existing methods from a vectorial statistics view: two limitations

Univariate tests are typically implemented in two complementary manners. One is illustrated in Fig. 2a. Given a significant level α, we get a boundary point and the red coloured rejection domain. For a set of statistics (e.g. those indicated by '$*****$'), every statistics $\tilde{s}$ is classified into either the rejection domain or the acceptance domain. Those falling into the rejection domain can all identify a significant rejection of H_0, featured by a worst-case false alarm probability α. Here, each statistics $\tilde{s}$ has not been provided with its accurate false alarm probability though such a probability could be much smaller than α especially when the corresponding statistics locates far away from the boundary point.

Alternatively, we may implement a test as illustrated in Fig. 2b. Judging whether a statistics $\tilde{s}$ identifies a significant rejection of H_0, we directly use $\tilde{s}$ as the boundary point to get the red coloured rejection domain, and then estimate the p value associated with $\tilde{s}$,

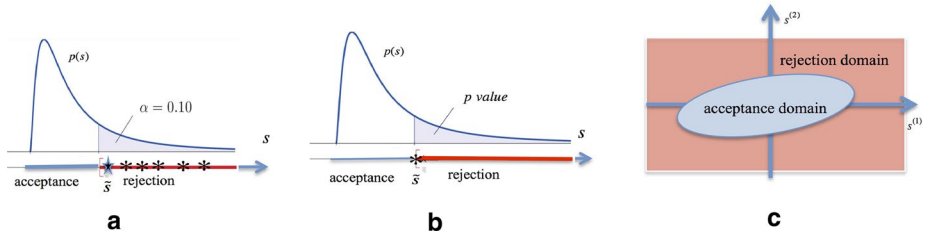


Fig. 2 A vectorial statistics view on the existing multivariate tests. **a** Univariate tests implemented for identifying whether a statistics $\tilde{s}$ leads to the rejection of H_0 in a significant level α; **b** Univariate tests implemented for computing the p value, i.e. the probability that those false alarms disturb or confuse the judgement on $\tilde{s}$; **c** In the two-dimensional space of **s**, the acceptance domain is an ellipse (*grey coloured*) specified by Eq. (3), corresponding to the acceptance range of (**a**, **b**), while the rest of the entire space is the rejection domain

i.e. the probability that those false alarms (statistics of samples that come under H_0) fall in the rejection domain. Instead of deciding whether a statistics $\tilde{s}$ leads to a significant rejection of H_0, the rejection domain is actually the domain that incurs for disturbance from false alarms. The p value indicates the probability that these false alarms disturb or confuse our judgement on $\tilde{s}$, based on which we can re-judge whether this rejection of H_0 is significant. In the studies of genome-wide sequencing and expression, what we encounter are actually tests implemented in such a manner. In this paper, we also adopt this manner.

Extended to multivariate tests, the Hotelling test, Wald test, and Score test are all featured by two key points. The first is computing a scalar statistics s in a quadratic form as follows:

$$s = \mathbf{s}^T \Sigma_{\mathbf{s}}^{-1} \mathbf{s}, \tag{3}$$

where $\mathbf{s}$ is computed from samples in the data (D)-space, and $\Sigma_{\mathbf{s}}$ is the covariance matrix of $\mathbf{s}$. The second is evaluating whether s falls in the rejection domain based on a probabilistic model (e.g. a F-distribution) on the axis of s, as illustrated in Fig. 2b. The probability of s falling in the rejection domain (red coloured) represents the false alarm probability (i.e. the p value) of rejecting H_0 while H_0 actually holds. In fact, the existing multivariate tests are mostly implemented in such a manner because it facilitates to evaluate the p value with help of probability distribution of a scalar statistics s.

Also, we may understand the Hotelling test, Wald test, and Score test from a perspective of vectorial statistics in a multi-dimensional S-space (Xu 2015).

For a two sample test illustrated in Fig. 1a, we consider the following vector

$$\mathbf{s} = \hat{\mathbf{c}}_1 - \hat{\mathbf{c}}_0, \tag{4}$$

for testing Eq. (1) with $\theta = \mathbf{s}$, where $\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_0$ are the sample means of the case samples and the control samples, respectively.

For a regression coefficient test illustrated in Fig. 1(b), we consider the vector

$$\mathbf{s} = \beta, \tag{5}$$

for testing Eq. (1) with $\theta = \mathbf{s}$, where β is a vector consisting of regression coefficients, and $\beta, \hat{c}$ are estimated by the maximum likelihood principle.

In both the above cases, the vector $\mathbf{s}$ has a multivariate Gaussian distribution

$$p(\mathbf{s}) = G(\mathbf{s}|\mathbf{0}, \Sigma_{\mathbf{s}}), \tag{6}$$

where $\Sigma_{\mathbf{s}}$ is the covariance matrix of $\mathbf{s}$, and $G(x|\mu, \Sigma)$ denotes a Gaussian distribution with a mean vector μ and a covariance matrix Σ.

For testing Eq. (1) with $\theta = \mathbf{s}$, $\mathbf{s}$ locates around the origin under H_0 and the acceptance domain is an ellipse (grey coloured) centred at the origin, as illustrated in Fig. 2c, while the rejection domain is the entire space outside of the ellipse, corresponding to the red coloured range in Fig. 2b. The probability of the scalar s falling in the rejection range (red coloured) in Fig. 2b is actually equivalent to the probability of the vector falling in rejection domain in Fig. 2c. In other words, what we observed here is actually a degenerated scenario of making multivariate tests in a multi-dimensional S-space, suffering from at least the following two limitations:

- H_0 by Eq. (1) means that all the dimensions of **s** are zero and we reject H_0 as long as at least one of these dimensions is rejected to be zero. In other words, differentiation is considered in a lumped sense without considering the roles of different dimensions and their combinations, as illustrated in Fig. 3.
- Rejection is made according to how far **s** is away from the origin (possibly weighted by the orientation of **s**, e.g. see Eq. (3)), but without taking the direction of **s** in consideration. In many applications, direction does take its role. In GWAS study, multiple SNPs' joint effect is reflected in the direction of vectorial statistics, as addressed in Ref. (Bansal et al. 2010) and particularly in its Figure 2.

In the next three subsections, the first limitation will be tackled by considering various situations featured by differentiations associated with different dimensions and their combinations. Then, the second limitation is further tackled in the subsequent subsection, featured by directional tests.

Lattice taxonomy of tests with different dimensions of rejection

Taking the two-dimensional S-space illustrated in Fig. 3 as an example, we observe that Eq. (1) with $\theta = \mathbf{s} = [s^{(1)}, s^{(2)}]^T$ becomes

$$H_0 : s^{(1)} = 0 \,\&\, s^{(2)} = 0. \tag{7}$$

The acceptance domain should be an area near the origin as illustrated in Fig. 3a. Also, the area has a rectangular shape if two dimensions are independent.

As illustrated in Fig. 3a, the rejection domain considered in the existing studies is the complement of the acceptance domain to the entire two-dimensional S-space, lumping up three rejection domains illustrated in Fig. 3b–d. Covering the largest area, it represents the rejection of H_0 via either $\neg(s^{(1)} = 0)$ or $\neg(s^{(2)} = 0)$ in the weakest collegiality. On the other hand, the collegiality that **s** falls in the rejection domain shown in Fig. 3d is

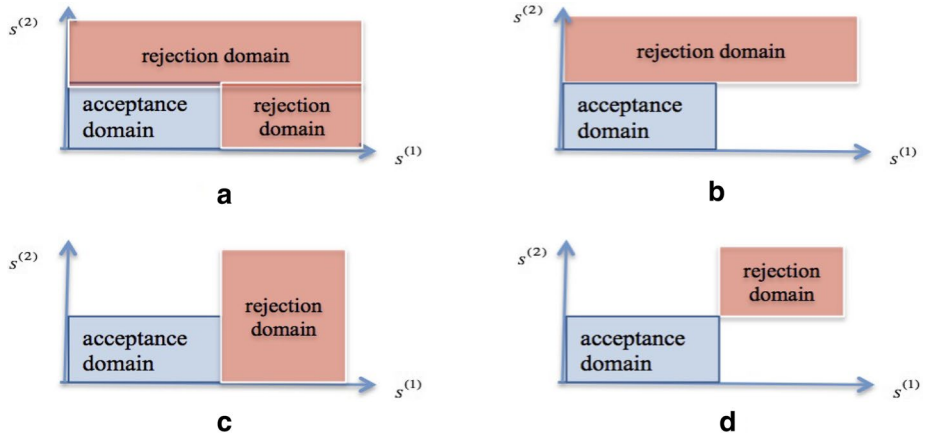


Fig. 3 One H_0 test in four different rejections. **a** The existing method considers a rejection in the weakest collegiality. **b** One dimension of rejection that considers $\neg(s^{(2)} = 0)$, where $\neg(s = 0)$ denotes 'the rejection of $s = 0$'. **c** The other dimension of rejection that considers $\neg(s^{(1)} = 0)$. **d** Two dimensions of rejection that considers $\neg(s^{(1)} = 0) \,\&\, \neg(s^{(2)} = 0)$ jointly, i.e. a rejection in the strongest collegiality

strongest, requiring both $\neg(s^{(1)}=0)$ and $\neg(s^{(2)}=0)$, which is suitable for the cases that we make a rejection in one most conservative way.

It is interesting to further consider the false alarm probability (i.e. the p value) of rejecting H_0 upon observing that $\mathbf{s}$ comes under H_0 but falls in the rejection domain in each of the four situations. With p_a, p_b, p_c, p_d denoting the p values for the cases in Fig. 3a–d, respectively, we have

$$\begin{aligned} &p_a = p_b + p_c - p_d, \\ &p_d \le p_b,\ p_d \le p_c,\ p_b \le p_a,\ p_c \le p_a, \\ &p_a = P[\neg(s^{(1)}=0)\ or\ \neg(s^{(2)}=0)|H_0], p_b = P[\neg(s^{(1)}=0)|H_0], \\ &p_c = P[\neg(s^{(2)}=0)|H_0],\ p_d = P[\neg(s^{(1)}=0)\ \&\neg(s^{(2)}=0)|H_0]. \end{aligned} \tag{8}$$

We have $p_d \le p_b$ from $P[\neg(s^{(1)}=0)\&\neg(s^{(2)}=0)|H_0] = P[\neg(s^{(1)}=0)|H_0]$ $P[\neg(s^{(2)}=0)|\neg(s^{(1)}=0), H_0]$ and $P[\neg(s^{(2)}=0)|\neg(s^{(1)}=0), H_0] \le 1$. Similarly, we also get $p_d \le p_c$. Moreover, it follows from $p_a = p_b + p_c - p_d$ we have $p_b \le p_a$ for $p_c \le p_d$ and $p_c \le p_a$ for $p_b \le p_d$.

In the three-dimensional S-space not only we have $\neg(s^{(1)}=0)\ or\ \neg(s^{(2)}=0)\ or\ \neg(s^{(3)}=0)$ in the weakest collegiality and $\neg(s^{(1)}=0)\&\neg(s^{(2)}=0)\&\neg(s^{(3)}=0)$ in the strongest collegiality but also we have

- Four choices of $[\neg(s^{(1)}=0)\&\neg(s^{(2)}=0)]\ or\ \neg(s^{(3)}=0)$, $\neg(s^{(1)}=0)\ or\ [\neg(s^{(2)}=0)\ \&\neg(s^{(3)}=0)]$, $[\neg(s^{(1)}=0)\ or\ (\neg(s^{(2)}=0)]\&\neg(s^{(3)}=0)$, and $\neg(s^{(1)}=0)\&$ $[\neg(s^{(2)}=0)\ or\ \neg(s^{(3)}=0)]$;
- Three choices of $\neg(s^{(1)}=0)\ or\ \neg(s^{2)}=0)$, $\neg(s^{(2)}=0)\ or\ \neg(s^{(3)}=0)$, and $\neg(s^{(1)}=0)\ or\ \neg(s^{3)}=0)$;
- Three choices of $\neg(s^{(1)}=0)\&\neg(s^{2)}=0)$, $\neg(s^{(2)}=0)\&\neg(s^{(3)}=0)$, and $\neg(s^{(1)}=0)\&$ $\neg(s^{3)}=0)$;
- Three choices of $\neg(s^{(1)}=0)$, $\neg(s^{(2)}=0)$, and $\neg(s^{(3)}=0)$.

Generally, in the n-dimensional space of vectorial statistics, testing H_0 by Eq. (1) involves testing various types of rejections featured by subsets of n different dimensions and their & and *or* connected combinations. Although an exhaustive search of all the possible types of rejections will be very tedious, we may still make a rather systematical investigation that organises major types of rejections in a partial order structure, namely two cascaded taxonomies as illustrated in Fig. 4a.

In such a way, a multivariate test is not just a single test as usually considered in the existing studies. Examining whether H_0 breaks in a lumping way is just one extreme (i.e. the bottom case) that puts the most loose requirement on making a rejection of H_0, featured by a rejection with the biggest p value and weakest collegiality. Actually, multivariate testing consists of tests in different levels of collegiality and different types, examining a total number of $2\sum_{i=2}^{n}\binom{n}{i}$ different combinations of these dimensions that may cause a significant rejection of H_0. The collegiality enhances from the bottom up towards the middle level that considers each of n different dimensions individually. From the middle level up, the collegiality further enhances level by level, until the top that

(See figure on next page.)

Fig. 4 Tests with different dimensions of rejection in two taxonomies. **a** Two cascaded taxonomies of all the possible combinations of three dimensions. The bottom three levels belong to the taxonomy of those combinations resulted from the operator OR, while the top three levels belong to the taxonomy of those combinations resulted from the operator &, i.e. AND. The bottom is the one considered by the existing method, featured by a rejection of the weakest collegiality. The collegiality gradually enhances from one level up to the next level, reaching the top featured by a rejection of the strongest collegiality. **b** Folding Figure (**a**) along the central horizontal line into two layers, one rejection with a strong collegiality featured by & on the 1st layer is paired with its counterpart rejection with a weak collegiality featured by OR on the 2nd layer. **c** The outcome of all the tests organised in this taxonomy. Listed on the right side of each combination are the resulted $\underline{p}_j^{\omega}$ and $\overline{p}_j^{\omega}$. As to be further explained in Fig. 6a, each of $\underline{p}_j^{\omega}$ and $\overline{p}_j^{\omega}$ is evidenced by two estimating values that are obtained with and without considering some higher order dependence among different dimensions. Listed in the *left-most column* are $\underline{p}_j, \overline{p}_j$, i.e. the minimum values of each level given by Eq. (12)

represents another extreme featured by the smallest p value for a rejection in the strongest collegiality.

Such a scenario may be intuitively understood from the problem of identifying a sickness of a bio-body or system with a number of intrinsic factors. The normality of the body requires that every factor runs normally, which corresponds to Eq. (1). The body falls ill if either one or more of these factors become abnormal, which corresponds to the bottom case illustrated in Fig. 4a. A wrong diagnosis of one factor's normality may lead to a wrong diagnosis that the body gets sick, that is, the chance of a false alarm is high. The other extreme is the top case illustrated in Fig. 4a, corresponding to that all the intrinsic factors jointly become abnormal. We get a wrong diagnosis that the body falls in this specific type of sick, only when the diagnosis of every factor's abnormality is wrong, that is, the chance of a false alarm is low. Moreover, there could be various types of sickness to be identified, associated with different combinations of abnormal factors and with different chances of false alarming.

Even compactly, we may fold the two cascaded taxonomies along the central horizontal line in Fig. 4a such that two layers are arranged as illustrated in Fig. 4b. At the j-th level, each combination is indexed by a j-tuple as follows:

$$\omega = \{i_1, i_2, \cdots, i_j\}, \text{ i.e. there are j indices picked out of } \{1, 2, \cdots, n\} \tag{9}$$

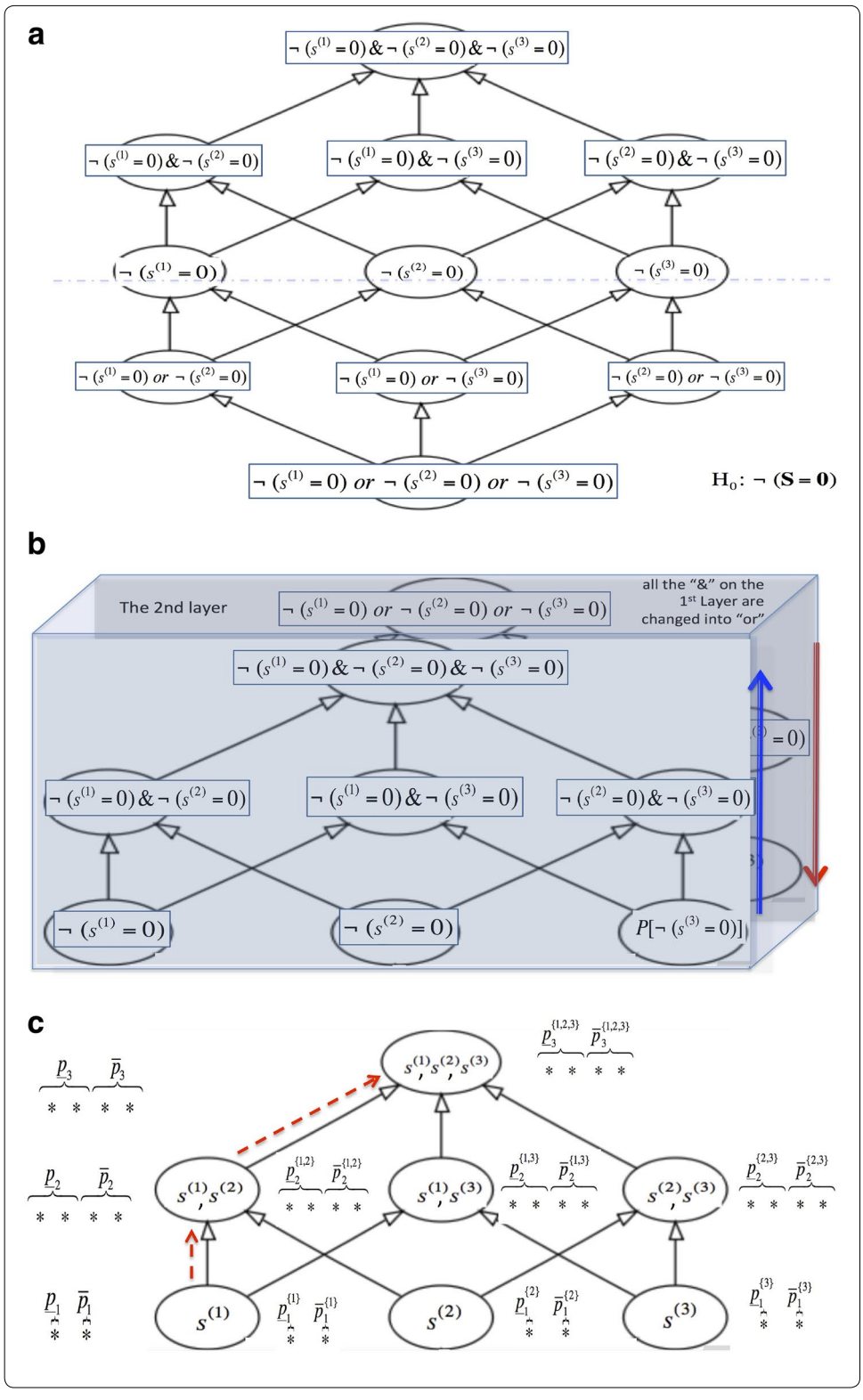
and a set Ω_j consists of of $\binom{n}{j}$ different j-tuples for a given j, where each $\omega \in \Omega_j$ is associated with the following two paired rejections:

$$\begin{aligned} &\text{On the 1st layer} : \underline{R}^{\omega} = \neg(s^{(i_1)} = 0)\ \&\ \neg(s^{(i_2)} = 0)\ \&\ \cdots \&\ \neg(s^{(i_j)} = 0), \\ &\text{On the 2nd layer} : \overline{R}^{\omega} = \neg(s^{(i_1)} = 0)\ or\ \neg(s^{(i_2)} = 0)\ or\ \cdots\ or\ \neg(s^{(i_j)} = 0). \end{aligned} \tag{10}$$

Accordingly, we have the false alarm probabilities (i.e. the p values):

$$\begin{aligned} &\underline{p}_j^{\omega} < \overline{p}_j^{\omega}, \\ &\overline{p}_j^{\omega} = P[\overline{R}^{\omega}|H_0] = P[\neg(s^{(i_1)} = 0)\ or\ \neg(s^{(i_2)} = 0)\ or\ \cdots\ or\ \neg(s^{(i_j)} = 0)|H_0], \\ &\underline{p}_j^{\omega} = P[\underline{R}^{\omega}|H_0] = P[\neg(s^{(i_1)} = 0)\&\neg(s^{(i_2)} = 0)\&\cdots\&\neg(s^{(i_j)} = 0)|H_0], \end{aligned} \tag{11}$$

which act as a pair of indicators for examining the role of the combination $s^{(i_1)}, s^{(i_2)}, \cdots, s^{(i_j)}$ in a significant rejection of H_0 by Eq. (1), where $\overline{p}_j^{\omega}$ indicates whether a significant differentiation is associated with one of the j dimensions individually, i.e.


a
$\neg(s^{(1)}=0)\&\neg(s^{(2)}=0)\&\neg(s^{(3)}=0)$
$\neg(s^{(1)}=0)\&\neg(s^{(2)}=0)$
$\neg(s^{(1)}=0)\&\neg(s^{(3)}=0)$
$\neg(s^{(2)}=0)\&\neg(s^{(3)}=0)$
$\neg(s^{(1)}=0)$
$\neg(s^{(2)}=0)$
$\neg(s^{(3)}=0)$
$\neg(s^{(1)}=0)\ or\ \neg(s^{(2)}=0)$
$\neg(s^{(1)}=0)\ or\ \neg(s^{(3)}=0)$
$\neg(s^{(2)}=0)\ or\ \neg(s^{(3)}=0)$
$\neg(s^{(1)}=0)\ or\ \neg(s^{(2)}=0)\ or\ \neg(s^{(3)}=0)$
H_0: $\neg(\mathbf{S}=\mathbf{0})$
b
The 2nd layer
all the "&" on the 1st Layer are changed into "or"
$P[\neg(s^{(3)}=0)]$
c
$s^{(1)},s^{(2)},s^{(3)}$
$s^{(1)},s^{(2)}$
$s^{(1)},s^{(3)}$
$s^{(2)},s^{(3)}$
$s^{(1)}$
$s^{(2)}$
$s^{(3)}$

indicating whether there is a simple or uni-factor significant differentiation; while $\underline{p}_j^{\omega}$ indicates whether a significant differentiation is associated with a joint effect of the j dimensions, i.e. indicating whether there is a deep or multi-factor significant differentiation.

Moreover, we may get the following pair of the p values:

$$\underline{p}_j = \underline{p}_j^{\omega^*},\ \overline{p}_j = \overline{p}_j^{\omega^*},\ \underline{\omega}^* = arg\ \min_{\omega\in\Omega_j} \underline{p}_j^{\omega}, \tag{12}$$

where $\underline{\omega}_j^*$ is the most distinctive combination and $\underline{p}_j$ indicates how distinctive a combination on the j-th level could be in the best sense. This level may not be distinctive enough if $\underline{p}_j$ is not small enough. On the other hand, there could be more than one combinations to be identified as distinctive enough, especially when $\overline{p}_j$ is very small.

The gap $\overline{p}_j - \underline{p}_j$ provides an information on a possible variety of significant differentiation on this level. The collegiality of rejection increases from the bottom up as indicated by the blue arrow on the 1st layer but reduces as indicated by the red arrow on the 2nd layer. Thus, $\underline{p}_j$ decreases, $\overline{p}_j$ increases from the bottom up, and thus the gap $\overline{p}_j - \underline{p}_j$ increases as j increases.

Illustrated in Fig. 4c is the outcome of all the tests organised in the taxonomy, from which we get a roadmap about how each dimension or a combination of dimensions contributes to a significant rejection of the null hypothesis. Comparing $\underline{p}_{j-1}^{\omega} - \underline{p}_j^{\omega}$ and $\overline{p}_{j-1}^{\omega} - \overline{p}_j^{\omega}$, as well as comparing the p values between two consecutive levels, we may understand the incremental role taken by each dimension.

However, what addressed above is still conceptual. In addition to the second limitation addressed at the end of the previous subsection, there are three coupled problems to be solved, listed as follows:

Problem 1	How to effectively compute $\underline{p}_j^{\omega}, \overline{p}_j^{\omega}$ in Eq. (12) ?
Problem 2	It is usually infeasible to enumerate all the $\sum_{i=2}^{n}\binom{n}{i}$ different combinations. How to effectively make such enumeration ?
Problem 3	There is dependence and redundancy among dimensions of **s**, which affects seriously the above two problems. How to remove the dependence and redundancy and to select appropriate number of dimensions ?

Testing implementation: independence case and latent independence

The above problems become easier to handle when dimensions of **s** are mutually independent, i.e. we have

$$p(\mathbf{s}) = \prod_i p(s^{(i)}). \tag{13}$$

In such a case, the acceptance domain will be a rectangular or hyper-cubic domain as illustrated in Fig. 3 and rejection can be made dimension by dimension independently. Thus, $\underline{p}_j^{\omega}, \overline{p}_j^{\omega}$ in Eq. (12) can be computed as follows:

$$
\begin{aligned}
\underline{p}^{\omega}_{I,j} &= P[\underline{R}^{\omega}|H_0] = P[\neg(s^{(i_1)}=0)\&\neg(s^{(i_2)}=0)\&\cdots\&\neg(s^{(i_j)}=0)|H_0] \\
&= P[\neg(s^{(i_1)}=0)|H_0]P(\neg(s^{(i_2)}=0)|H_0]\ldots P(\neg(s^{(i_j)}=0)|H_0] \\
&= \underline{p}^{\{i_1\}}_{1}\underline{p}^{\{i_2\}}_{1}\cdots\underline{p}^{\{i_j\}}_{1}, \\
\overline{p}^{\omega}_{I,j} &= P[\overline{R}^{\omega}|H_0] = P[\neg(s^{(i_1)}=0)\ or\ \neg(s^{(i_2)}=0)\ or\ \cdots\ or\ \neg(s^{(i_j)}=0)|H_0] \\
&= \sum_i \underline{p}^{\{i\}}_{1} - \sum_{i\neq j}\underline{p}^{\{i\}}_{1}\underline{p}^{\{j\}}_{1} + \cdots + (-1)^{j-1}\prod_i \underline{p}^{\{i\}}_{1}, \\
\underline{p}^{\{i\}}_{1} &= P[\neg(s^{(i)}=0)|H_0] = P[s^{(i)}\in\Gamma_1(\tilde{s}^{(i)})|H_0],
\end{aligned}
\tag{14}
$$

from which Problem 1 is simply solved, where $\Gamma_1(\tilde{s}^{(i)})$ is a univariate rejection domain that has either one tail or two tails as follows:

$$
\begin{aligned}
&\text{One tail}\ :\ \Gamma_1(\tilde{s}^{(i)}) = \{s^{(i)} : (s^{(i)}-\tilde{s}^{(i)})sign(\tilde{s}^{(i)}) > 0\}, \\
&\text{Two tail}\ :\ \Gamma_1(\tilde{s}^{(i)}) = \{s^{(i)} : s^{(i)}sign(s^{(i)}) - \tilde{s}^{(i)}sign(\tilde{s}^{(i)}) > 0\}.
\end{aligned}
\tag{15}
$$

For Problem 2, we may sort $\underline{p}^{\{i\}}_{1} = P[\neg(s^{(i)}=0)|H_0], i=1,2,\ldots,n$ into an ascending order and simply get the best combination of the j-th level by picking the first j ones to compute Eq. (12), that is, we have

$$
\begin{aligned}
&\underline{p}^{\{1\}}_{1} \le \underline{p}^{\{2\}}_{1} \le \cdots \le \underline{p}^{\{n\}}_{1}, \\
&\underline{p}_j = \underline{p}^{\{1,2,\ldots,j\}}_{j} = \underline{p}^{\{1\}}_{1}\underline{p}^{\{2\}}_{1}\ldots\underline{p}^{\{j\}}_{1},\ \omega=\{1,2,\ldots,j\}, \\
&\overline{p}_j = \overline{p}^{\{1,2,\ldots,j\}}_{j},\ \text{given by Eq. (14)}.
\end{aligned}
\tag{16}
$$

That is, the left-most column on Fig. 4c can be easily obtained. In many applications, it is unnecessary to find out all the distinctive combinations. Instead, getting the best combination of each level is already enough. Moreover, we may also obtain $\underline{p}^{\omega}_{j}$ by Eq. (14) whenever it needs to examine a particular combination $\omega=\{i_1,i_2,\ldots,i_j\}$. Even if we need to find out all the distinctive combinations, this ascending order of $\underline{p}^{\{i\}}_{1}$ also provides some hints to get an effective search.

Due to the assumption by Eq. (13) and the ascending order of $\underline{p}^{\{i\}}_{1}$, Problem 3 is simplified into how to select an appropriate number k^* of dimensions. In the idealistic case that some $s^{(i)}$ can be regarded as "do-not-care" featured by $\underline{p}^{\{i\}}_{1} = P[\neg(s^{(i)}=0)|H_0] = 1$, it follows from Eq. (14) that we observe that

$$
\underline{p}_1 > \underline{p}_2 > \cdots > \underline{p}_{k^*} = \cdots = \underline{p}_n,
\tag{17}
$$

by which we can determine one appropriate k^*. However, when statistics is computed from a small size of samples, $\underline{p}^{\{i\}}_{1}$ will be a small unknown number $1 > \delta^{(i)} > 0$ even when $s^{(i)}$ is regarded as "do-not-care", which leads to $\underline{p}_1 > \underline{p}_2 > \cdots > \underline{p}_n$. Thus, it is difficult to determine k^* by Eq. (17).

One simplest way to detect and discard those "do-not-care" dimensions is checking

$$
\underline{p}^{\{i\}}_{1} = P[\neg(s^{(i)}=0)|H_0] \ge d_i,
\tag{18}
$$

where d_i is a filtering threshold determined after a statistical analysis.

However, the above solution is too rough. Not only it is not easy to determine d_i, but also it does not consider the joint effect of different dimensions. Returning to the thinking line of Eq. (17), we seek to correct each $\underline{p}_j$ by a normalisation term, see Eqn. (93) in Ref. (Xu 2015a).

Given a set of samples, a new set π of samples is obtained by permutation, and there is a set Π of different choices of π. Similar to getting $\underline{p}_j$ by Eq. (17) on the original set of samples, we get $\underline{p}_j^{\pi}$ on the samples of π and then make the following p value estimated in the probability space (shortly pp value)

$$pp_j = P[p_j^{\pi} < p_j | H_0] = \frac{\sum_{\pi \in \Pi_\Gamma} p_j^{\pi}}{\sum_{\pi \in \Pi} p_j^{\pi}}, \ \Pi_\Gamma = \{\pi : p_j^{\pi} < p_j, \ \forall \pi \in \Pi\}, \tag{19}$$

which represents the probability that the false alarm rate on randomly permuted samples is smaller than the one on the original samples, i.e. the probability that the rejection associated with p_j is really a false alarm.

The above mentioned unknown $\delta^{(i)}$ for some dimension of "do-not-care" may be approximately regarded as unchanged over different permutations. Hence, its effect to both the denominator and the numerator will be cancelled out by Eq. (19), and we get closer to Eq. (17) after p_j is replaced by pp_j.

Let $J(k) = \ln pp_k$, we may find an appropriate k^* of distinctive dimensions by

$$k^* = arg \min_k J(k). \tag{20}$$

In a summary, the outcome of tests is simply the best first path (red coloured) of the length k^*, instead of traversing all the tests on the taxonomy illustrated in Fig. 4c. This path provides the k^* most distinctive dimensions and the order of their importances. From the bottom up, if a subpath $\underline{p}_j$ of the length j corresponds to a significant rejection of H_0, any extensions of this subpath also give more significant rejections of H_0.

One essential problem is that the assumption by Eq. (13) is difficult to be satisfied and thus the resulted $\underline{p}_j^{\omega}$ by Eq. (14) is usually too optimistic. Instead, we may further consider the latent independence as illustrated in Fig. 5a. That is, we observe a latent coordinate wherein components are mutual independent subject to an additive noise e that is independent of $\mathbf{s}_u$ and typically Gaussian with a spherical covariance matrix. In the new coordinate, we can get not only the effect of noise e in consideration but also Eq. (13) satisfied at least conceptually. Implementation may just follow those addressed between Eqs. (13) and (20), simply with each appearance of $\mathbf{s}$ replaced by $\mathbf{s}_u$.

When each component of $\mathbf{s}_u$ comes from a non-Gaussian univariate, the latent model is called non-Gaussian factor analysis (NFA) (Xu 2003, 2009; Tu and Xu 2014) and the mapping from $\mathbf{s}$ to $\mathbf{s}_u$ is featured by a distribution with a non-linear regression as illustrated in Fig. 5c. When each component of $\mathbf{s}_u$ is a Gaussian univariate, the latent model becomes the classical factor analysis (FA) and the mapping from $\mathbf{s}$ to $\mathbf{s}_u$ is a distribution with a linear regression as illustrated in Fig. 5b. Particularly, a FA model may be called either FA-b with an additional constraint $AA^T = I$ or FA-a for a conventional setting. For the maximum likelihood learning, FA-a and FA-b are equivalent. However,

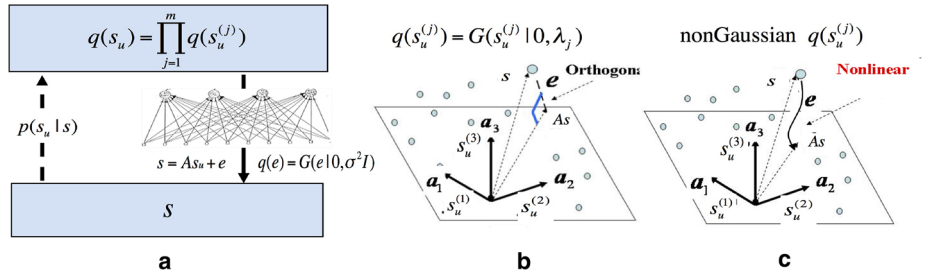


Fig. 5 The linear model for latent independence. **a** ***s*** comes from a number m of independent intrinsic factors of $\mathbf{s}_u$ via a linear transformation plus an additive noise e. **b** intrinsic factors are Gaussian. **c** intrinsic factors are non-Gaussian

FA-b becomes more favourable for determining m. Readers are referred to Sect.2.2 in Xu (2011) and Tu and Xu (2011) for further studies on FA-b versus FA-a.

In implementation, we need to determine not only A, σ^2, and the parameters (if any) in each univariate distribution $q(s_u^{(i)})$, but also the distribution $p(\mathbf{s}_u|\mathbf{s})$ and the number m, which is computationally difficult. The BYY harmony learning provides a tool for this purpose, and readers are referred to Xu (2015b) for a recent summary, together with Algorithm 4 for FA, Algorithms 6 and 7 for binary FA, and Algorithm 8 for non-Gaussian FA.

Approximately, we may consider to get only the second-order independence by a linear orthogonal project $\mathbf{s} \to \mathbf{s}_u$ as follows

$$\begin{aligned} &\mathbf{s}_u = U^T\mathbf{s}, \ \text{with} \ U = [\mathbf{u}_1, \ldots, \mathbf{u}_m], \ \mathbf{s}_u = [s_u^{(1)}, \ldots, s_u^{(m)}]^T, \\ &\Sigma_{\mathbf{s}} U = U\Lambda, \ \Lambda = diag[\lambda_1, \ldots, \lambda_m], \end{aligned} \tag{21}$$

where $\Sigma_{\mathbf{s}}$ is the covariance matrix of $\mathbf{s}$, and $\mathbf{u}_1, \ldots, \mathbf{u}_m$ are its eigenvectors that correspond to the non-zero eigenvalues $\lambda_j, j = 1, \ldots, m$. The resulted elements $s_u^{(1)}, \ldots, s_u^{(m)}$ become mutually independent in a sense of the second-order statistics, i.e. its covariance matrix is a diagonal matrix Λ.

Generally, we may estimate $\Sigma_{\mathbf{s}}$ from a set $\{\tilde{\mathbf{s}}^{\pi}\}$ (including $\tilde{\mathbf{s}}$), with each $\tilde{\mathbf{s}}^{\pi}$ obtained by a set of samples that comes from a permutation π of the original samples sets. Specifically, we let $\Sigma_{\mathbf{s}} = \Sigma_{\pi}$ with Σ_{π} given by Eqn. (69) in Ref. Xu (2015a).

Also, we may get $\Sigma_{\mathbf{s}}$ by the Fisher information matrix (i.e. Eqn. (6) in Ref. Xu 2015a) for a regression coefficient test. For a two sample test, we may use Σ in Eq. (2) as $\Sigma_{\mathbf{s}}$ under the assumption that not only samples are i.i.d. but also case population and control population are uncorrelated.

As illustrated in Fig. 6a, b, the Cartesian coordinate is rotated into the one spanned by the eigenvectors $\mathbf{u}_j, j = 1, \ldots, m$. After the rotation, the acceptance domain is an ellipse and further becomes a sphere after the normalisation by $\Lambda^{-0.5}$. Approximately, we may implement tests in the new coordinate following those addressed from Eqs.(13) to (19), simply with each appearance of $\mathbf{s}$ replaced by $\mathbf{s}_u$.

Moreover, we select one appropriate number k^* by Eq. (20). Alternatively, we may also compare the use of $J(k) = \ln pp_k$ with a model selection criterion, e.g. Eqn. (29) in Xu (2011), given as follows:

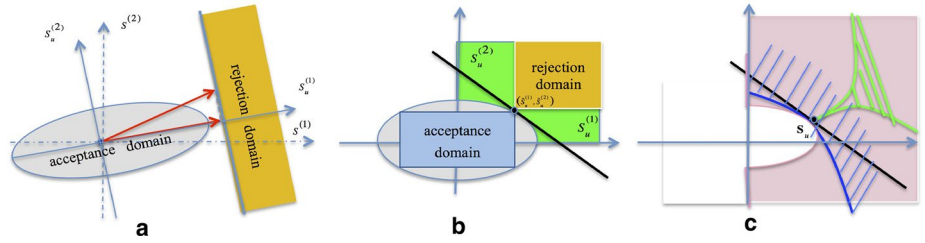


Fig. 6 Getting the second-order independence. **a** The coordinate system is rotated to the one spanned by the eigenvectors. A rejection domain is specified by a combination of these eigenvectors. **b** In the new coordinate, we approximately regard that Eq. (13) is satisfied. **c** Some examples of the boundary-based rejection domains

$$J(k) = \sum_{j=1}^{k} \ln(\lambda_j - \sigma^2) + m \ln \sigma^2 + k \ln(2\pi e),\ \sigma^2 = \frac{1}{m-k} \sum_{j=k+1}^{m} \lambda_j. \tag{22}$$

As illustrated in Fig. 6b, the p value $\overline{p}^{\omega}_{I,j}$ by Eq. (14) corresponds to the rejection domain that covers everywhere outside of the blue coloured acceptance domain and thus serves as a highest bound in a pessimistic sense. Similarly, the p value $\underline{p}^{\omega}_{I,j}$ by Eq. (14) corresponds to the rejection domain illustrated in Fig. 6b by the orange coloured box and thus serves as a lowest bound in an optimistic sense. $\overline{p}^{\omega}_{I,j}, \underline{p}^{\omega}_{I,j}$ work well only when Eq. (13) is satisfied, which unfortunately does not hold usually. In the sequel, we further add in $\overline{p}_{\neg e,j}$, $\underline{p}_{h,j}$ to partially tackle this problem.

Higher order independence and property-oriented test

With help of the second-order independence by Eq. (21), we get a rejection domain that covers everywhere outside of the ellipse illustrated in Fig. 6b. Its difference from the blue coloured box is illustrated by the grey area that reflects the influence of higher order dependence. Keeping this influence may not only simplify the computation of $\overline{p}^{\omega}_{I,j}$ by Eq. (14) but also enhance reliability because the problem of removing higher order dependence becomes more difficult especially based on merely a small size of samples.

Thus, we simply consider the elliptic acceptance domain illustrated in Fig. 6b by the following statistics:

$$\begin{aligned} t^{\omega}_{\neg e,j} &= \sum_{\ell=1}^{j} \frac{s_u^{(i_\ell)\,2}}{\lambda_{i_\ell}} = \mathbf{s}_u^T \Lambda_j^{-1} \mathbf{s}_u, \quad \text{for } \omega = \{i_1, i_2, \ldots, i_j\}, \\ \Lambda_j &= diag[\lambda_{i_1}, \cdots, \lambda_{i_j}],\ \mathbf{s}_u = [s_u^{(i_1)}, \ldots, s_u^{(i_j)}]^T, \end{aligned} \tag{23}$$

which has a univariate distribution (e.g. approximately an F-distribution) and is actually a simplified and truncated version of Eq. (3) that takes a key role in the Hotelling test, Wald test, and Score test.

Considering the ellipse $1 = \sum_{\ell=1}^{j} \frac{s_u^{(i_\ell)\,2}}{t^{\omega}_{\neg e,i_\ell} \lambda_{i_\ell}}$ that passes through $\tilde{\mathbf{s}}_u$ we get the corresponding value of t_j as follows

$$\tilde{t}^{\omega}_{\neg e,j} = \sum_{\ell=1}^{j} \frac{\tilde{s}_u^{(i_\ell)\,2}}{\lambda_{i_\ell}}, \tag{24}$$

which is featured by an elliptic equation $t^{\omega}_{\neg e,j} = \tilde{t}^{\omega}_{\neg e,j}$ with its inner side defining an elliptic acceptance domain as illustrated in Fig. 6b and its outer side defining the following rejection domain:

$$\Gamma^{\omega}_{\neg e,j}(\tilde{\mathbf{s}}_u) = \Gamma(B_{\tilde{\mathbf{s}}_u}(\mathbf{s}_u))_{B_{\tilde{\mathbf{s}}_u}(\mathbf{s})=t^{\omega}_{\neg e,j}-\tilde{t}^{\omega}_{\neg e,j}}, \tag{25}$$

which is an example of the following family of rejection domains

$$\Gamma(B_{\tilde{\mathbf{s}}_u}(\mathbf{s}_u)) = \{\mathbf{s}_u : B(\mathbf{s}_u|\tilde{\mathbf{s}}_u) > 0\}, \tag{26}$$

featured by a boundary equation $B(\mathbf{s}_u|\tilde{\mathbf{s}}_u) = 0$. Its positive side $B(\mathbf{s}_u|\tilde{\mathbf{s}}_u) > 0$ defines the rejection domain, while its negative side defines the acceptance domain.

It follows from Eq. (25) that we get a lower bound of $\overline{p}^{\omega}_{I,j}$ as follows

$$\overline{p}^{\omega}_{\neg e,j} = P[\mathbf{s}_u \in \Gamma^{\omega}_{\neg e,j}(\tilde{\mathbf{s}}_u)|H_0] < \overline{p}^{\omega}_{I,j}. \tag{27}$$

On the other hand, the rejection domain that corresponds $\underline{p}^{\omega}_{I,j}$ by Eq. (14) differs from $\Gamma^{\omega}_{\neg e,j}(\tilde{\mathbf{s}}_u)$ in the green area that features higher order independence. To enhance the reliability of $\underline{p}^{\omega}_{I,j}$, we prefer to ignore some higher order independence since it is unreliable to estimate, though we still need to consider some higher order independence to improve the testing power.

A trade-off solution is considering a rejection domain given by a half-space illustrated in Fig. 6b in one of the following Choice (a) and Choice (b):

$$\begin{aligned}
&\Gamma^{\omega}_{h,j}(\tilde{\mathbf{s}}_u) = \Gamma(B_{\tilde{\mathbf{s}}_u}(\mathbf{s}_u)) \; with \; B_{\tilde{\mathbf{s}}_u}(\mathbf{s}) = \overrightarrow{\beta}^T S_j^{-0.5}(\mathbf{s}_u - \tilde{\mathbf{s}}_u),\\
&\overrightarrow{\beta} = \begin{cases} S_j^{-0.5}\tilde{\mathbf{s}}_u, & \text{Choice (a)},\\ sign(S_j^{-0.5}\mathbf{s}_u), & \text{Choice (b)}, \end{cases} \; S_j = \begin{cases} I, & \text{Case (1)},\\ \Lambda_j, & \text{Case (2)}, \end{cases}\\
&sign(\mathbf{x}) = [sign(x^{(1)}), \ldots, sign(x^{(j)})]^T, \; sign(\xi) = \begin{cases} 1, & if \; \xi > 0,\\ -1, & if \; \xi < 0. \end{cases}
\end{aligned} \tag{28}$$

which comes from a blue coloured linear boundary that passes through $\tilde{\mathbf{s}}_u$ and covers merely a part of green area while ignoring those of even higher order independence. When S is given by Case (1), we observe that Choice (a) and Choice (b) here are actually Choice (b) and Choice (c) addressed by Eqn. (70) and Figure 5 in Ref. Xu (2015a).

It follows from Eq. (28) that we get the following upper bound of $\overline{p}^{\omega}_{I,j}$:

$$\underline{p}^{\omega}_{h,j} = P[\mathbf{s}_u \in \Gamma^{\omega}_{h,j}(\tilde{\mathbf{s}}_u)] > \underline{p}^{\omega}_{I,j}. \tag{29}$$

Together with Eq. (27), we get the following quad

$$\overline{p}^{\omega}_{I,j} > \overline{p}^{\omega}_{\neg e,j} > \underline{p}^{\omega}_{h,j} > \underline{p}^{\omega}_{I,j} \tag{30}$$

as the values of each quad $** \; **$ listed with each combination in Fig. 4c.

In addition to turning a multivariate test into multi-levels of tests in a taxonomy illustrated in Fig. 4, the above quad by Eq. (30) alone represents a further development of multivariate test already. The first two $\overline{p}^{\omega}_{I,j}, \overline{p}^{\omega}_{e,j}$ represent a conventional practice of making a multivariate test on a vectorial statistics $\mathbf{s}$ of the mutual independence or loosely

the second-order independence cross the components of $\mathbf{s}$. The other two $\underline{p}^{\omega}_{h,j}, \underline{p}^{\omega}_{I,j}$ represent new developments. It follows from Eq. (14) that $\underline{p}^{\omega}_{I,j}$ is featured by a probabilistic product of multiple univariate one-tailed or two-tailed tests for a vectorial statistics $\mathbf{s}$ of mutual independence, while $\underline{p}^{\omega}_{h,j}$ takes certain high order dependence in consideration.

All the above and previously addressed tests can be regarded as examples of property-oriented tests. Recalling the univariate test introduced in Fig. 2b, one key issue is estimating the probability that the false alarms disturb the judgement on a given scalar statistics $\tilde{s}$ according to a property owned by the statistics, which can be further generalised into the property sharing condition given in Table 1.

For a scalar statistics $\tilde{s}$, there is only one property $s \geq \tilde{s}$ to consider, as illustrated by the red coloured rejection range in Fig. 2b. However, for vectorial statistics $\mathbf{s}$ there are various choices to be considered. First, we may consider the properties of $\mathbf{s}$ either directly in its own Cartesian coordinate or one of subspaces spanned by different combinations of its eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_m$ given in Eq. (21), with some dependence and redundancy removed, as well as some disturbing noises discarded. Second, we may consider various properties featured by different types of combinations of the components $\mathbf{s}_u = [s_u^{(1)}, \ldots, s_u^{(m)}]^T$.

Summarised below are four types introduced previously:

Basic property a property owned by each scalar component $s^{(i)} = 0, i \in \{1, 2, \ldots, n\}$ individually, i.e. the bottom level illustrated in Fig. 4.

Logical combination a property obtained by combining several basic properties via logical connections & and *or*, i.e. other levels illustrated in Fig. 4b.

Linear boundary equation a property with its corresponding rejection domain given by Eq. (26) and featured with a linear equation $B_{\tilde{\mathbf{s}}_u}(\mathbf{s}_u) = 0$, e.g. a half-space illustrated in Fig. 6b and defined by one linear equation in Eq. (28). Actually, it includes each basic property above as its degenerated case.

Quadratic boundary equation a property with the rejection domain given by Eq. (26) and featured by a quadratic equation $B_{\tilde{\mathbf{s}}_u}(\mathbf{s}_u) = 0$, e.g. the domain outside of the ellipse illustrated in Fig. 6b and defined by Eq. (25) or Eq. (3).

There are many other properties to be considered too. The last two above can be further extended by considering $B_{\tilde{\mathbf{s}}_u}(\mathbf{s}_u) = 0$ in a higher order equation. Beyond Eq. (25), it is also possible to use an even general mathematical model $\Gamma(\tilde{\mathbf{s}})$ to express rejection domain, e.g. the two-branching curved boundary illustrated in Fig. 6c.

Naturally, we come to a question, is there any necessary condition that such a rejection domain should satisfy ?

If an alarm is able to disturb the judgement on $\tilde{s}$, one usually expects that enlarging the magnitude of this alarm should make this disturbance more stronger, which leads to the

Table 1 A theory of property-oriented rejection-based test

Key point	Description
Property sharing condition	A necessary condition for false alarms to disturb the judgment on a given statistics is sharing with the statistics' property that we consider
Alarm scale-up nature	If an alarm vector $S^{-0.5}(\mathbf{s} - \tilde{\mathbf{s}})$ falls within a rejection domain, $\gamma S^{-0.5}(\mathbf{s} - \tilde{\mathbf{s}})$ will also fall within the rejection domain for any $\gamma > 1$
Least complexity principle	A rejection domain is modelled by a smallest number of parameters that can be well determined from given samples

scale-up nature given in Table 1, based on which we may exclude many bad choices of rejection domain.

However, it is still not enough yet to fix a reasonable rejection domain. On one hand, the p value reduces as the rejection domain becomes smaller, which seemly leads us to choose a rejection domain as small as we want. On the other hand, in order to specify a rejection domain, what we can rely on are merely the known $\tilde{\mathbf{s}}_u$ and $\lambda_1, \ldots, \lambda_m$, which have already been used in Eqs. (25) and (28) for defining a linear or quadratic equation $B_{\tilde{\mathbf{s}}_u}(\mathbf{s}_u) = 0$. For a complicated rejection domain, e.g. the green coloured one shown in Fig. 6c, there are more unknowns to be specified. We need to either let an enough large part of them becoming known or get some priories that enable to fix those unknowns. In other words, we encounter the problems of unreliability and over-fitting, especially when there is a finite size of samples for us to compute Eq. (21), which thus leads to the least complexity principle given in Table 1.

The situation is similar to the problem of selecting an appropriate number k^* of dimensions, as addressed between Eqs. (16) and (19). Although it remains an open challenge to choose a rejection domain modelled by a smallest number of parameters, we are at least able to determine k^* by Eq. (20), which provides another perspective to understand the rationale of examining $\underline{p}^{\omega}_{h,j}, \underline{p}^{\omega}_{I,j}$ for all $j \leq k^*$ in the taxonomy illustrated in Fig. 4c.

Algorithm 1 Testing vectorial statistics in a lattice taxonomy

Require: estimate $\Sigma_{\mathbf{s}}$.

Decoupling get $\tilde{\mathbf{s}}_u = [\tilde{s}_u^{(1)}, \cdots, \tilde{s}_u^{(m)}]^T$ *and* $\Lambda = diag[\lambda_1, \cdots, \lambda_m]$ by Eq.(21).

Level 1: get $\underline{p}_1^{\{i\}} = P[\tilde{s}_u^{(i)} \in \Gamma_1(\tilde{s}_u^{(i)})|H_0]$ by $\Gamma_1(\tilde{s}_u^{(i)})$ given by Eq.(15) with $\tilde{s}^{(i)}$ replaced by $\tilde{s}_u^{(i)}$

Sort $\underline{p}_1^{\{1\}} \leq \underline{p}_1^{\{2\}} \leq \cdots \leq \underline{p}_1^{\{n\}}$, get $\underline{p}_j = \underline{p}_1^{\{1\}} \underline{p}_1^{\{2\}} \cdots \underline{p}_1^{\{j\}}$ for $j = 1, \cdots, m$,

and further get $pp_j, j = 1, \cdots, m$ by Eq.(19). Then, we choose an appropriate k^* by Eq.(20).

Level j: for $j = 1, \cdots, k^*$ we compute

$\overline{p}_j = \sum_i \underline{p}_1^{\{i\}} - \sum_{i \neq j} \underline{p}_1^{\{i\}} \underline{p}_1^{\{j\}} + \cdots + (-1)^{j-1} \prod_i \underline{p}_1^{\{i\}}$,

$\overline{p}_{\neg e,j} = P[\mathbf{s}_u \in \Gamma_{\neg e,j}^{\{1,2,\cdots,j\}}(\tilde{\mathbf{s}}_u)|H_0]$ with $\Gamma_{\neg e,j}^{\{1,2,\cdots,j\}}(\tilde{\mathbf{s}}_u)$ given by Eq.(25).

$\underline{p}_{h,j} = P[\mathbf{s}_u \in \Gamma_{h,j}^{\omega}(\tilde{\mathbf{s}}_u)]$ with $\Gamma_{h,j}^{\omega}(\tilde{\mathbf{s}}_u)$ given by Eq.(28). Alternatively, $\underline{p}_{h,j}^{\omega}$ may also be obtained by a univariate one-tail test on the B-score shown in Table 2.

Outputs: $\overline{p}_j,\ \overline{p}_{\neg e,j},\ \underline{p}_{h,j},\ \underline{p}_j,\ j = 1, \cdots, k^*$.

Remarks:

(a) The outcomes are featured by the best first path of the taxonomy, with each level j only considering the best combination $\underline{\omega}^* = arg\min_{\omega \in \Omega_j} \underline{p}_{I,j}^{\omega}$ without traversing all the combinations.

(b) If needed we may also get $\overline{p}_{I,j}^{\omega} > \overline{p}_{\neg e,j}^{\omega} > \underline{p}_{h,j}^{\omega} > \underline{p}_{I,j}^{\omega}$ by Eq.(14), Eq.(25), and Eq.(29) for a particular combination ω.

(c) The best first path may also be obtained by the best combination $\underline{\omega}^* = arg\min_{\omega \in \Omega_j} \underline{p}_{h,j}^{\omega}$ on each level j but in a huge cost of enumerating $\binom{k^*}{j}$ possible combinations of ω.

(d) We may also consider the extensions of $\overline{p}_{\neg e,j}, \underline{p}_{h,j}$ with $\Gamma_{\neg e,j}^{\omega}(\tilde{\mathbf{s}}_u), \Gamma_{h,j}(\tilde{\mathbf{s}}_u)$ given by Eq.(31).

(e) We may obtain $\overline{pp}_j,\ \overline{pp}_{\neg e,j},\ \underline{pp}_{h,j},\ \underline{pp}_j,\ j = 1, \cdots, k^*$ by Eq.(19) too.

Outcomes:

(1) whether H_0 is rejected significantly as a whole.
(2) the maximum number of intrinsic factors that are responsible for a significance of rejecting H_0.
(3) the minimum number of intrinsic factors that are responsible for a significance of rejecting H_0.
(4) other combinations of intrinsic factors contribute to a significance of rejecting H_0.
(5) those probes or features in the D-space that contribute critically to a significance of rejecting H_0, via a multivariate test on these features and then a comparison on whether the resulted p-value is close to the p-value obtained by the above (1) & (2).
(6) important samples marked by certain critical probes that are regarded as to be more likely the major contributors to a significance of rejecting H_0.

Moreover, the statistics obtained from random samples is probabilistic, and thus the property we consider is probabilistic too. To increase the reliability, we may use a bootstrap method.

Given a set of samples, we obtain $\tilde{\mathbf{s}}$ and determine $\Gamma^{\omega}_{\neg e,j}(\tilde{\mathbf{s}}_u)$ by Eq. (25) and $\Gamma_{h,j}(\tilde{\mathbf{s}}_u)$ by Eq. (28), Then, we get a resampling set of samples on which we obtain $\tilde{\mathbf{s}}^r$ and determine $\Gamma^{\omega}_{\neg e,j}(\tilde{\mathbf{s}}^r_u)$ and $\Gamma_{h,j}(\tilde{\mathbf{s}}^r_u)$. After getting an enough large size of $\{\tilde{\mathbf{s}}^r_u\}$, we obtain the following unions as the final choice of rejection domains

$$\Gamma^{\omega}_{\neg e,j}(\tilde{\mathbf{s}}_u) = \cup_{\tilde{\mathbf{s}}^r_u \in \{\tilde{\mathbf{s}}^r_u\}} \Gamma^{\omega}_{\neg e,j}(\tilde{\mathbf{s}}^r_u),\ \Gamma_{h,j}(\tilde{\mathbf{s}}_u) = \cup_{\tilde{\mathbf{s}}^r_u \in \{\tilde{\mathbf{s}}^r_u\}} \Gamma_{h,j}(\tilde{\mathbf{s}}^r_u), \tag{31}$$

from which we get $\overline{p}^{\omega}_{\neg e,j}$ by Eq. (27) and $\underline{p}^{\omega}_{h,j}$ by Eq. (29).

Finally, summarised in Algorithm 1 are the main steps of implementing tests in a lattice taxonomy.

Directional test, matrix-variate test, and phenotype-targeted test

Property-oriented multivariate tests can be divided into two categories, namely directional tests versus non-directional tests. As previously addressed in Fig. 1 and the last paragraph of the introduction section, the existing multivariate tests, and the ones with $\overline{p}^{\omega}_{I,j}, \overline{p}^{\omega}_{\neg e,j}$ as well, are mostly non-directional tests, which can be regarded as extensions of univariate two-tailed tests, featured by merely considering how far the vectorial statistics is away from the origin (possibly weighted by its orientation) but without taking its direction in consideration.

In contrast, a directional test is featured by that its rejection domain relates to certain direction. Precisely, this rejection domain at least contains a non-empty set D of unit vectors such that $\gamma_0 \mathbf{d}_0$ locates outside of the rejection domain for some $\mathbf{d}_0 \in D$ and a large enough $\gamma_0 > 0$, i.e. at least the rejection domain does not contain some directions. Directional tests can be regarded as extensions of univariate one-tailed tests to multivariate tests. In addition to the previously addressed half-space associated with $\underline{p}^{\omega}_{h,j}$ (i.e. the one illustrated by the black line), examples shown in Fig. 6c are all directional tests. The one outside the ellipse and lilac coloured can be regarded as a directional counterpart of the bottom one in Fig. 4b, associated with the largest p value. The other side of the black line contains the green coloured rejection domain with its p value smaller than $\underline{p}^{\omega}_{h,j}$. They may all be regarded as examples of the boundary-based test (BBT) (see Table 6 in Ref. Xu 2015a), with their rejection domains featured by quadratic, linear, and two-branching curved boundary, respectively. Moreover, $\underline{p}^{\omega}_{I,j}$ is featured by a probabilistic product of multiple univariate one-tailed tests, and its corresponding rejection domain is actually an orthant of the S-space along the direction of the vector $sign(\tilde{\mathbf{s}}_u)$.

Alternatively, the directional test associated with the half-space type rejection domain given in Eq. (28) can be implemented from another aspect. We may obtain $\underline{p}^{\omega}_{h,j}$ by a univariate one-tailed test on the B-score as given in Table 2. Also, we may consider other two types of projection scores in that table. One is the FDA score with the projection direction being the normal direction of the best linear separating hyperplane as shown in Fig. 1. Another measure is the misclassification rate of the case–control samples by the linear boundary. Several machine learning methods are available for learning such a linear boundary, with two examples given in Table 2.

Intuitively, directional test may also be understood by an analogy to radar detection of an intruding fly that is approaching the borderline of a country. An alarm will sound as long as a fly approaches the borderline along whatever a direction, which corresponds to the bottom case as illustrated in Fig. 4a, e.g. outside of the ellipse domain along any

Table 2 Three typical projection scores

Type	Description
B-score	We get a projection score $B_{\tilde{\mathbf{s}}_u}(\mathbf{s})$ by Eq. (28), namely, the projection of statistics $\mathbf{s}_u - \tilde{\mathbf{s}}_u$ onto a particular direction $S_j^{-0.5}\vec{\beta}$. Actually, we are lead to a univariate one-tailed test on this projection score, which may be simply implemented by either one-tailed z-test or one-tailed t-test. Also, we may estimate the univariate distribution of the score, and then compute $\underline{p}^{\omega}_{h,j}$ based on the estimated distribution
FDA score	We get $\vec{\beta}$ by making the Fisher discriminative analysis (FDA) on the control samples and the case samples, and then obtain the projection score $\vec{\beta}^T\mathbf{s}$ with $\mathbf{s}$ given by Eq. (4), where the arrow of $\vec{\beta}$ points from the control to the case, as illustrated in Fig. 1b, while the classical FDA does not care about which direction of two choices is taken as the arrow. A directional test can be made by either one-tailed z-test or one-tailed t-test, using the statistics $t_{\vec{\beta}} = \frac{\vec{\beta}^T\mathbf{s}}{\sigma},\ \sigma^2 = \alpha_0\sigma_0^2 + \alpha_1\sigma_1^2,$ (32) where σ_0^2, σ_1^2 are the sample variances of the projections of control–case samples onto $\vec{\beta}$, respectively, and α_0, α_1 are corresponding proportions We may also perform a non-directional test with the arrow of $\vec{\beta}$ ignored, by using a two-tailed *z*-test or *t*-test, which is suggested in Table 2(1) of Ref. Xu (2015a) as one example of the boundary-based two-sample test or BBT in short
Learning LDA score	We may also perform either a directional test or a non-directional test as above, but with $\vec{\beta}$ obtained by (a) Support vector machine (SVM) (Suykens 1999; Suykens et al. 2002), as suggested in Table 4(c) of Ref. Xu (2015a) (b) Sparse logistic regression (Shevade and Keerthi 2003; Koh et al. 2007)

direction in Fig. 6a, b. A misreport from anywhere of the country's border will trigger such a false alarm, that is, the chance of a false alarm is high. On the other hand, a fatal and urgent alarm will sound when a fly comes along a direction towards the capital of the country because this attack may make many fatal components of this country disordered jointly, which corresponds to the normal direction of the half-space illustrated by the black line in Fig. 6b. Only a misreport from this direction will trigger this type of alarm, that is, the chance of a false alarm is low. Moreover, there could be different alarms from different directions, corresponding to different directional tests.

Moreover, extensions can be made from a vector $\mathbf{s}$ to a matrix. As addressed in Ref. Xu (2015a), many tasks of big data analyses demand extending vector-based sampling units to sampling units in matrix format. Illustrated in Fig. 7a is a data cubic encountered in the case–control studies. Accordingly, we encounter matrix-variate tests in matrix-variate logistic regression (see the details from Eqs. (48) to (57) in Ref. Xu 2015a) and matrix-variate discriminative analysis (see the details from Eqs (33) to (43) in Ref. Xu 2015a). For the former, test is made on regression coefficients in two vectors, which is still a multivariate test. For the latter, the situation becomes quite different, and further details are addressed in the sequel.

With an n-dimensional vector $\mathbf{s}$ extended into an $n \times m$ matrix, Eqs. (1) and (4) are extended into the following matrix form

$$H_0 : \mathbf{S} = \mathbf{0},\ with\ \mathbf{S} = \hat{\mathbf{C}}_1 - \hat{\mathbf{C}}_0. \tag{33}$$

Although we may stack the columns of $\mathbf{S}$ into a long vector, the covariance matrix $\Sigma_{\mathbf{s}}$ becomes an $nm \times nm$ matrix, and its estimation from a finite number of samples becomes not feasible. Three feasible approximate methods are suggested in Table 3. The

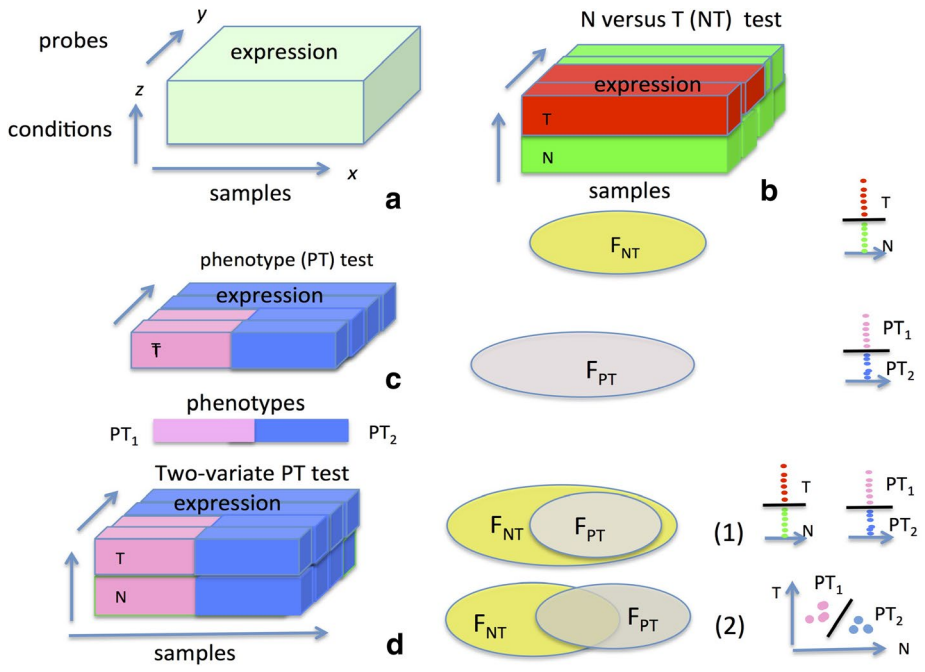


Fig. 7 A case–control study that considers matrix-variate samples. **a** Data matrices stacked along the x-axis to form a data cubic. Each element in the data cubic is a real or discrete value, called the expression at the corresponding coordinate. Each matrix sample consisting of rows obtained from different conditions is arranged along the z-axis, while each row consisting of probes or called measurements is located along the y-axis. **b** Row-wise bi-partition of the data cubic with the control representing those normal (N) conditions and the case representing those abnormal or trouble (T) conditions. Test is made per a probe, aiming at differentiation expression (DE) cross conditions (N and T), shortly called NT test. Those with the *p* value below a threshold and the fold rate of T over N bigger than a threshold are selected as a set F_{NT} of significant DE probes. On the *right side*, the scatter plot consists of dots locating on a vertical line, with two dots per sample (T in *red*, N in *green*). **c** Column-wise bi-partition of the data cubic with the case and control representing two of targeted phenotypes (e.g. cancer stages I, II, III, & IV, 3-year survival, 5-year survival, etc). Only the T layer is considered. Similar to **b** but with *T*, *N* replaced by PT_1, PT_2, the test aims at a set F_{PT} of significant DE probes of targeted phenotypes (PT_1, PT_2), shortly called PT test. **d** Both the T and N layers are jointly considered. We have two ways for getting F_{PT}. One considers two steps, firstly getting F_{NT} in the same way as (**b**) and then getting $F_{PT} \subseteq F_{NT}$ in the same way as (**c**), featured by two *lines* of scatters on its right side. The second considers T and N jointly in a 2D sample and makes a two-variate PT test to get a set F_{PT} of significant DE probes cross PT_1, PT_2. The *right side* is a 2D scatter plot with a dot representing two dots in (**b**)

Table 3 Three feasible approximate techniques for matrix-variate two-sample test

Method	Description
(a) Multivariate test per probe	If probes are independent, we make a multivariate test on each [*condition*, *sample*] matrix slice per probe. Such a multivariate test can be implemented by Algorithm 1 or one of the methods given in Table 2
(b) LDA-based multivariate test	we make the map $s_u(f) = \mathbf{s}^T\vec{\beta}$ onto $\vec{\beta}$ per probe f with $\vec{\beta}$ obtained by either FDA or learning-based methods in Table 2, and make a multivariate test on $[s_u(f_1), \cdots, s_u(f_g)]^T$ to consider multiple probes $f_1, \ldots, f_g$ jointly
(c) Bilinear MDA-based test	We make the matrix-variate discriminative analysis (MDA) (see Eq.(33) & Eq.(34) in Ref. Xu (2015a)) to obtain $\mathbf{v}$, $\vec{\beta}$, based on which we test $\vec{\beta} = \mathbf{0}$ by a multivariate test on $\mathbf{s}_v = \mathbf{v}^T\mathbf{S}$ given $\mathbf{v}$ fixed and test $\mathbf{v} = \mathbf{0}$ by a multivariate test on $\mathbf{s}_{\vec{\beta}} = \vec{\beta}^T\mathbf{S}^T$ given $\vec{\beta}$ fixed Alternatively, we may make test on the scalar statistics $s_{v\vec{\beta}} = \mathbf{v}^T\mathbf{S}\vec{\beta}$

first applies to the cases that probes are either independent or assumed to be independent. The second is firstly mapping the vector $\mathbf{s}$ into one feature by $s_u(f) = \mathbf{s}^T\mathbf{u}$ and then making multivariate test on $[s_u(f_1), \ldots, s_u(f_g)]^T$ jointly. The third is a bilinear matrix-variate two sample test.

These existing case–control studies can be roughly classified into two classes featured by column-wise versus row-wise bi-partitions of the data cubic, as illustrated in Fig.7b–d, respectively. Examples of Fig. 7b can be found in most of the SNP analyses in GWAS and those gene expression studies under a single condition (e.g. from a tumour tissue only). Examples of Fig. 7c, d1 can be found in many current studies of gene expression with tumour versus its paired adjacent tissue. All of these existing efforts are featured by making univariate tests.

One widely adopted existing practice is making a NT test for getting a set F_{NT} of significant differentiation expression (DE) probes. However, the resulted biomarkers may not be optimal in the sense of differentiating phenotypes, because a differentiation expression between T versus N may not well cope with the distinctions between phenotypes. For prognosis purpose, there are also efforts that further make a PT test in Fig. 7d(1) for getting $F_{PT} \subseteq F_{NT}$. Even so, the selection of F_{PT} is merely based on either the tumour expression or the fold change of T over N, without a best use of information contained in the (T,N) pair for differentiations between phenotypes.

Proposed recently in Ref. Xu (2015a), considering both T and N jointly in a 2D vector by a two-variate PT test paves a new road for reconsidering the task. As shown in Fig. 7d(2), samples of PT_1 and PT_2 can be well separated by a line on the 2D scatter plot. Transforming the 2D scatter plot into a plot on vertical line, however, it becomes no longer possible to separate PT_1 and PT_2. In other words, even in the cases that we are unable to identify biomarkers for distinguishing PT_1 versus PT_2 in the existing ways as shown in Fig. 7b–d (1), we may still find biomarkers for distinguishing PT_1 versus PT_2 by the new method.

Discussions

Whole genome sequencing analyses

As previously addressed from Eqs. (1) to (3), multivariate tests take an important role in many tasks of whole genome sequencing analyses. Using the new methods proposed in this paper to tackle these tasks, we may expect the following features :

1. Disturbing influences of those "do-not-care" variants may be reduced such that the ability of identifying variants for significant differentiation can be considerably improved.
2. Directions of risk effect versus preventive effect with the multiple variants are taken in consideration.
3. Important samples may be identified by observing whether their critical variants have major contributions to significant differentiation, as addressed at the bottom of Algorithm 1.
4. Extensions of the joint multiple-variant sequencing analysis may be made not only to identify variants that significantly differentiate tumour versus normal but also to identify variants that significantly differentiate a pair of phenotypes.

Table 4 Two-variate TP tests : implementations and applications

Implementations	Applications
1. Make the two-variate PT test by the method given in Table 3 (a), which provides an improvement on making the two-variate PT test by Hotelling test as suggested in Table 8(1)(a) of Ref. Xu (2015a)	(a) Identify mRNA and lncRNA biomarkers for tumour vs normal in expression analysis
2. Make the FDA-based PT test by Table 3 (b), especially one-tailed z-test or one-tailed t-test by Eq. (32), which is a complementary to the FDA-based PT test listed in Table 2(1) of Ref. Xu (2015a), where a univariate two-tailed *t* test is made	(b) Identify mRNA and lncRNA biomarkers for 3-year & 5-year survival in expression analysis
3. Find probes that significantly differentiate not only phenotypes but also abnormal vs normal, as well as their common part	(c) Identify mRNA and lncRNA biomarkers for cancer grades I, II, III, & IV in expression analysis (d) For each of the above cases (a)(b)(c), we use the FDA projection $s_U(f) = \mathbf{s}^T \hat{\beta}$ to replace the original expression value for painting heatmaps and making the corresponding clustering analysis

Genome-scale expression profile of mRNA and lncRNA expression

In the existing studies on genome-scale expression profile of mRNA or/and lncRNA, there are many examples of the case–control study on the data cubic shown in Fig. 7d, with testing made in one of the ways shown in Fig. 7b, d(1). Instead, the new method shown in Fig. 7d(2) provides a better choice.

Summarised in Table 4 are some proposed implementations and applications.

Conclusions

Instead of understanding and making multivariate test in a single rejection, multivariate test actually consists of a hierarchy of numerous tests organised in a lattice taxonomy, with the bottom level in the lowest rejection collegiality (the largest p value) and the top level in the highest rejection collegiality (the smallest p value), while the ones on the intermediate levels represent different situations in which the null hypothesis is rejected and are featured by different p values. The outcomes consist of not only whether the null hypothesis is rejected significantly as a whole, but also those combinations of multiple components that are responsible for a significance of rejecting the null hypothesis, and those probes that contribute considerably to a significance of rejecting the null hypothesis. Not only detailed implementations are presented, but also several potentials are addressed on possible applications to expression profile-based biomarker identification and exome sequencing-based joint SNV detection.

Author details
[1] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, NT, Hong Kong, China. [2] Department of Computer Science and Engineering, Centre for Brain-inspired Computing and Bio-Health Informatics, The School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, SEIEE Building 3, 800 Dongchuan Road, Minhang District, 200240 Shanghai, China.

Acknowledgements
This work was partialy supported by the National Basic Research Program of China (973 Program 2009CB825404).

Competing interests
The author has no competing interests.

Received: 12 November 2015 Accepted: 24 December 2015
Published online: 13 January 2016

References

Adhikari K, Reales G, Smith AJ, Konka E, Palmen J, Quinto-Sanchez M, Acuña-Alonzo V, Jaramillo C, Arias W, Fuentes M et al (2015) A genome-wide association study identifies multiple loci for variation in human ear morphology. Nat Commun 6:7500

BaiZ D, Saranadasa H (1996) Effect of high dimension: by an example of a two sample problem. Statistica Sinica 6(2):311–329

Bansal V, Libiger O, Torkamani A, Schork NJ (2010) Statistical analysis strategies for association studies involving rare variants. Nat Rev Genetics 11(11):773–785

Chapman J, Whittaker J (2008) Analysis of multiple snps in a candidate gene or region. Genetic Epidemiol 32(6):560

Demidenko E (2013) Mixed models: theory and applications with R. probability and statistics. John Wiley and Sons, Hoboken

Dempster AP (1958) A high dimensional two sample significance test. Ann Math Stat 995–1010

DempsterA P (1960) A significance test for the separation of two highly multivariate small samples. Biometrics 16(1):41–50

Evangelou E, Ioannidis JP (2013) Meta-analysis methods for genome-wide association studies and beyond. Nat Rev Genet 14(6):379–389

Fan R, Knapp M (2003) Genome association studies of complex diseases by case-control designs. Am J Hum Genet 72(4):850–868

Ferguson J, Wheeler W, Fu Y, Prokunina-Olsson L, Zhao H, Sampson J (2013) Statistical tests for detecting associations with groups of genetic variants: generalization, evaluation, and implementation. Euro J Human Genet 21(6):680–686

Fisher RA (1932) Statistical methods for research workers, 4th edn, Oliver and Boyd, Edinburgh, pp 99–101

Gudmundsson J, Sulem P, Gudbjartsson DF, Masson G, Agnarsson BA, Benediktsdottir KR, Sigurdsson A, Magnusson OT, Gudjonsson SA, Magnusdottir DN (2012) A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. Nat Genet 44(12):1326–1329

Han F, Pan W (2010) A data-adaptive sum test for disease association with multiple common or rare variants. Human Heredity 70(1):42–54

Hotelling H (1931) The generalization of student's ratio. Ann Math Stat 2(3):360–378

Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ, Gupta N, Sklar P, Sullivan PF, Moran JL (2012) Exome sequencing and the genetic basis of complex traits. Nature genetics 44(6):623–630

Koh K, Kim SJ, Boyd SP (2007) An interior-point method for large-scale l1-regularized logistic regression. J Mach Learn Res 8(8):1519–1555

Lee S, Abecasis GR, Boehnke M, Lin X (2014) Rare-variant association analysis: Study designs and statistical tests. Am J Human Genet 95(1):5–23

Lee S, Wu MC, Lin X (2012) Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13(4):762–775

Li H, Gui J (2004) Partial cox regression analysis for high-dimensional microarray gene expression data. Bioinformatics 20(suppl 1):208–215

Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Human Genet 83(3):311–321

Morgenthaler S, Thilly WG (2007) A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (cast). Mut Res/Fund Mol Mech Mutag 615(1):28–56

MorrisA P, Zeggini E (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genetic Epidemiol 34(2):188

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR (2010) Pooled association tests for rare variants in exon-resequencing studies. Am J Human Genet 86(6):832–838

ShevadeS K, KeerthiS S (2003) A simple and efficient algorithm for gene selection using sparse logistic regression. Bioinformatics 19(17):2246–2253

SrivastavaM S (2007) Multivariate theory for analyzing high dimensional data. J Jpn Stat Soc 37(1):53–86

Suykens JA, Van Geste lT, De Brabanter J, De Moor B, Vandewalle J (2002) Least squares support vector machines. World Scientific Publishing, Singapore

SuykensJ A, Vandewalle J (1999) Least squares support vector machine classifiers. Neural Process Lett 9(3):293–300

SwansonD M, Blacker D, AlChawa T, Ludwig KU, Mangold E, Lange C (2013) Properties of permutation-based gene tests and controlling type 1 error using a summary statistic based gene test. BMC Genet 14(1):108

Tu S, Xu L (2011) An investigation of several typical model selection criteria for detecting the number of signals. Front Elect Electronic Eng China 6(2):245–255

Tu S, Xu L (2014) Learning binary factor analysis with automatic model selection. Neurocomputing 134:149–158

WuM C, Lee S, Cai T, Li Y, Boehnke M, Lin X (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89(1):82–93

Xu L (2003) Independent component analysis and extensions with noise and time: a bayesian ying-yang learning perspective. Neural Inform Process Lett Rev 1:1–52

Xu L (2009) Independent subspaces. In: Rabunal JR, Dorado J, Sierra AP (eds.) Encyclopedia of Artificial Intelligence. IGI Global Snippet, Hershey, Pennsylvania, pp 892–901

Xu L (2011) Codimensional matrix pairing perspective of byy harmony learning: hierarchy of bilinear systems, joint decomposition of data-covariance, and applications of network biology. Front Electr Electron Eng China 6:86–119. A special issue on Machine Learning and Intelligence Science: IScIDE2010 (A)

Xu L (2015a) Bi-linear matrix-variate analyses, integrative hypothesis tests, and case-control studies. Appl Inform 2(1):1–39

Xu L (2015b) Further advances on bayesian ying yang harmony learning. Appl Inform 2(5)

ZaykinD V (2011) Optimally weighted z-test is a powerful method for combining probabilities in meta-analysis. J Evol Biol 24(8):1836–1841

Zhou X, Stephens M (2014) Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat Methods 11(4):407–409