Xu *Applied Informatics* (2015) 2:4
DOI 10.1186/s40535-015-0007-5

Applied Informatics
a SpringerOpen Journal

RESEARCH Open Access

Bi-linear matrix-variate analyses, integrative hypothesis tests, and case-control studies

Lei Xu[1,2]

Correspondence:
lxu@cse.cuhk.edu.hk
[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, The Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University, SEIEE Building 3, 800 Dongchuan Road, Minhang District, 200240 Shanghai, China

Abstract

We pursue a threefold purpose in this paper. First, we suggest a Kullback-Leibler formulation for developing a statistics and making discriminative projection for case-control studies, based on which existing typical methods are revisited and then further extended to matrix-variate counterparts. Second, we propose a bi-linear matrix form, based on which multivariate discriminative analysis and logistic, Cox, and linear mixed regression are extended into their matrix-variate counterparts. Third, we systematically address the necessity, feasibility, and methodology of integrative hypothesis tests (IHT) from the complementarity of model-based test and boundary-based test (BBT) in the data (D)-space, statistics (S)-space, and probability (P)-space. We elaborate four IHT components (modelling, comparison, classification, and assurance) and summarise four IHT types in the D-space. Then, we extend the existing efforts on multivariate tests to BBTs in the S-space. Particularly, we extend the classic univariate one-tail z-test to the multivariate ones, which is then applied to a multivariate sample-pairing delta (SPD) test for detecting a collective inclining dominance. Also, we propose a SPD discriminative analysis that extends this SPD test. Moreover, we propose a multivariate bi-test that tests the classic null and also a null about the inference reliability due to test space complexity, including a further development of Fisher combination. Finally, we suggest possible applications for gene expression biomarkers and exome-sequencing-based joint single-nucleotide variant (SNV) detection.

Keywords: Kullback divergence; Discriminative projection; Logistic, Cox, and linear mixed regressions; Bi-linear form; Boundary-based test; Integrative hypothesis test; Bayesian Ying Yang; Statistics integration; Dependence decoupling; Bi-test; Test reliability; Controlling testing complexity; Inclining dominance; Gene expression; Joint SNVs detection

Background

Typically, multivariate statistical analysis and related machine-learning studies consider a basic sampling unit in a vector $\boldsymbol{x}_t$. Though an entire data set may be regarded as given in a format of matrix that consists of $\boldsymbol{x}_1, \cdots, \mathbf{x}_N$ as the columns, each statistics is computed from an assembly of vector samples and featured by vector inner product as a basic modelling unit.

Nowadays, not only rapid developments of data acquisition techniques (DePristo et al. 2011; Koboldt et al. 2013) demand that data with a matrix X_t as shown in Figure 1 as a basic sampling unit be considered, but also ever-increasing computing ability makes such a demand possible. One typical field that longs for such demands is featured

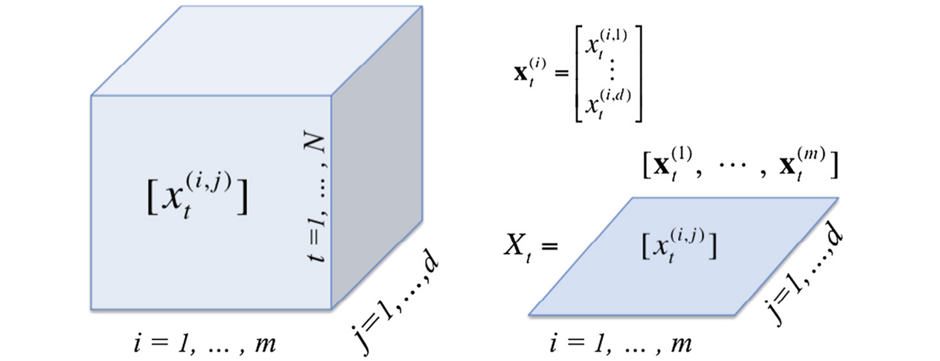
$[x_t^{(i,j)}]$
$t = 1, \ldots, N$
$i = 1, \ldots, m$
$j = 1, \ldots, d$
$\mathbf{x}_t^{(i)} = \begin{bmatrix} x_t^{(i,1)} \\ \vdots \\ x_t^{(i,d)} \end{bmatrix}$
$[\mathbf{x}_t^{(1)}, \cdots, \mathbf{x}_t^{(m)}]$
$X_t =$
$[x_t^{(i,j)}]$
$j = 1, \ldots, d$
$i = 1, \ldots, m$

Figure 1 A set of matrix-variate samples.

by image-based tasks, of which a basic sampling unit is naturally a matrix though traditional studies consider sample vectors to simplify computation. However, this simplification will miss some useful structural information, e.g. considering the rows of X_t as independent and identically distributed (i.i.d.) samples will miss the dependence cross rows. Also, recent efforts on big-data analyses eagerly demand statistical approaches for matrix-variate-based data analysis.

Another field that demands matrix-variate-based analyses is computational biology or particularly computational genomics. Typically, expression profiles of basic units (e.g. gene, miRNA, lncRNA) are analysed via vector samples (e.g. via rows or columns of expression matrix) (Simon et al. 2003). Advanced studies also examine expression profiles under different conditions (Ji et al. 2009; Persson et al. 2011) and across different time points (Bar-Joseph et al. 2012) and thus demand that sampling units in matrix format or even a high-dimensional array are considered. In a genome-wide association study or exome-sequencing analysis (DePristo et al. 2011; Gibson 2012; Purcell et al. 2007), though a majority of methods is still featured by vector-variate analysis, there are already some efforts made on matrix-variate-based data analysis.

In the rest of this paper, we start at providing a background and review on the related topics and methods, including the following:

- Two-sample test and Hotelling statistics.
- Logistic regression, Wald test, and Rao's score.
- Discriminative analyses and integrative hypothesis tests (IHT).
- Cox model and linear mixed model

Then, we pursuit a threefold purpose as follows:

(1) A Kullback-Leibler-divergence-based formulation for developing statistics and discriminative criterion for the case-control studies, based on which existing typical methods are revisited and extended to their matrix-variate counterparts.

(2) A bi-linear matrix form, based on which discriminative analysis, logistic regression, Cox model, and linear mixed model are extended into their matrix-variate counterparts.

(3) A systematic investigation of the necessity, feasibility, and implementing methods of IHT from the perspective of model-based test (MBT) versus boundary-based test (BBT) in the three levels of space, namely the data sample space (D-space), the statistics space (S-space), and probability space (P-space).

More specifically, the above third one consists of the following:

- The complementarity of MBT versus BBT in the D-space, the basic IHT components (modelling, comparison, classification, and assurance), and four types of IHT.
- Bayesian Ying Yang (BYY)-harmony-learning-based IHT formulation for coordinately optimising the performances of task A, task B, and task C in the D-space.
- The MBT vs BBT perspective in the S-space, especially extensions of the existing efforts on the integration of multiple statistics to the S-space BBT, with the help of dependence decoupling.
- A S-space BBT-based extension of univariate one-tail z-test for testing the null of multivariate zero mean, which is then applied to multivariate sample-pairing delta (SPD) test for detecting a collective inclining dominance.
- A SPD discriminative analysis that not only improves the multivariate SPD test but also further extends it to matrix-variate ones.
- A multivariate bi-test on both the classic null and also a null about test reliability by controlling the testing complexity, including a further development of the Fisher combination.

Finally, we discuss several possible IHT applications for expression-profile-based biomarker finding and exome-sequencing-based joint single-nucleotide variant (SNV) detection.

Hypothesis tests for case-control studies

Most efforts in computational genomics and generally computational biology involve case-control studies. For a case-control study, we are given two populations of vector-variate samples $X_\omega = \{\mathbf{x}_{t,\omega}, t = 1, \cdots, N_\omega\}, \omega = 0, 1$, where the one with $\omega = 1$ is called the case population while the one with $\omega = 0$ is called the control population. The task of a hypothesis test is examining a rejection of the following null assumption:

$$H_0 : \text{there is statistically no difference between two populations of samples,} \tag{1}$$

for which a statistics is computed from the samples to test the opposite assumption H_1 that there is a significant difference between the two populations.

A typical example is testing whether H_0 breaks on two populations of samples from a multivariate Gaussian distribution $G(\mathbf{x}|\mathbf{c}, \Sigma)$ with the mean vector c and the covariance matrix Σ, with help from the following Hotelling statistics (Hotelling 1931):

$$T^2 = \frac{N_0 N_1}{N}(\mathbf{c}_1 - \mathbf{c}_0)^T \Sigma^{-1}(\mathbf{c}_1 - \mathbf{c}_0), \tag{2}$$

where $N = N_0 + N_1$, and $\mathbf{c}_1, \mathbf{c}_0$ are the mean vectors of the case and control populations, respectively. Also, the covariance matrix is assumed to be $\Sigma = \Sigma_0 = \Sigma_1$.

Generally, we evaluate the difference between two populations based on population modelling by a parametric model $q(x|\theta)$, that is, firstly modelling each population of samples and then evaluating the overall difference between two resulted models. The performance is measured by the p value that describes the false alarm probability of judging that H_0 by Equation (1) significantly breaks. Such efforts are usually referred as model-based tests or sometimes called model comparison or class comparison (Simon et al. 2003).

Another typical example is logistic regression. Rewriting the above two populations of samples into a set of paired samples $\{\boldsymbol{x_t}, \omega_t\}, t = 1, \cdots, N$ with $\omega_t = 1$ and $\omega_t = 0$ indicating the sample $\boldsymbol{x_t}$ from the case and control population, respectively. We let ω_t be regressed by $\boldsymbol{x_t}$ in the following conditional probability:

$$\begin{aligned} &p(\omega_t|\boldsymbol{x_t}, \theta) = s(\zeta_t)^{\omega_t}[1 - s(\zeta_t)]^{1-\omega_t}, \\ &\zeta_t = y_t + c,\ y_t = \mathbf{w}^T\mathbf{x}_t,\ s(r) = \frac{1}{1+e^{-r}}. \end{aligned} \tag{3}$$

All the unknowns in a notation θ are estimated by maximising the following likelihood:

$$L = \prod_{t=1}^{N} p(\omega_t|\boldsymbol{x_t}, \theta), \tag{4}$$

which cannot be analytically solved due to the nonlinearity of $s(r)$ and are usually handled by a gradient-based iterative algorithm (Hosmer et al. 2013). The test of the null assumption by Equation (1) becomes testing the null assumption:

$$H_0:\ \mathbf{w} = \mathbf{0}, \tag{5}$$

where $\mathbf{w}$ is a subset of θ. It is typically made by either the Wald test or the Score test (Engle 1984), both of which are computed from one or both of the following statistics:

$$\Delta(\mathbf{w}) = \frac{\partial \ln L}{\partial \mathbf{w}},\ I(\mathbf{w}) = -\frac{\partial^2 \ln L}{\partial \mathbf{w} \partial \mathbf{w}}, \tag{6}$$

where $\Delta(\mathbf{w})$ is called the score vector, and $I(\mathbf{w})$ is called the Fisher information matrix.

The Wald test considers the following:

$$s = I^{0.5}(\hat{\boldsymbol{w}})\mathbf{w},\ \hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}} L, \tag{7}$$

as a testing statistics that has an asymptotic normal distribution under the null assumption.

While the Rao's score (or simply the score test and often known as the Lagrange multiplier test) considers:

$$s = \Delta^T(\hat{\boldsymbol{w}})I^{-1}(\hat{\boldsymbol{w}})\Delta(\hat{\boldsymbol{w}}), \tag{8}$$

as a testing statistics that has an asymptotic distribution of χ^2_k, where k is the number of constraints imposed by the null hypothesis. It degenerates to χ^2_1 when $\boldsymbol{w}$ consists of only one parameter.

This logistic regression examines the difference between two populations via firstly building up a hyperplane boundary and then tests Equation (5) that directly aims at whether the boundary depends on variables in consideration.

Discriminative analyses and integrative tests

Other than directly aiming at the boundary, a different aspect of logistic regression is that we can use $p(\omega_t|\boldsymbol{x_t}, \theta)$ by Equation (3) to classify each sample by:

$$\omega_t = \arg\max_{\omega} p(\omega|\boldsymbol{x_t}, \theta). \tag{9}$$

Equivalently, the same result comes from the hyperplane boundary $\zeta_t = 0$ with ζ_t given in Equation (3) such that samples are classified into its two sides. The outcome is the following decomposition:

$$X_1 = X_1^{(1)} \cup X_1^{(0)},\ X_0 = X_0^{(1)} \cup X_0^{(0)}. \tag{10}$$

That is, the case set X_1 is separated into a subset $X_1^{(1)}$ with unchanged labels and a subset $X_1^{(0)}$ of samples that are relabelled as control samples, and similarly, the control set X_0 into $X_0^{(0)}$ with unchanged labels and $X_0^{(1)}$ relabelled as case samples.

Actually, seeking a hyperplane boundary is the goal of linear discriminative analyses (LDA). One classic example is the Fisher discriminative analysis (FDA). For separating samples of two populations, the FDA seeks a projection $y_t = \mathbf{w}^T \mathbf{x}_t$ to map each vector $\boldsymbol{x}_t$ into a univariate y_t such that:

$$\max_{\mathbf{w}} J_y(\mathbf{w}),\ J_y(\mathbf{w}) = \frac{\left(c_0^y - c_1^y\right)^2}{\alpha_0 {\sigma_0^y}^2 + \alpha_1 {\sigma_1^y}^2}, \tag{11}$$

where for $\omega = 0, 1$ we have

$$\begin{aligned} \alpha_\omega = \frac{N_\omega}{N},\ c_\omega^y = \frac{\sum_{t=1}^{N_\omega} y_{t,\omega}}{N_\omega}&, \\ y_{t,\omega} = \mathbf{w}^T \mathbf{x}_{t,\omega}&, \\ {\sigma_\omega^y}^2 = \frac{\sum_{t=1}^{N_\omega} \left(y_{t,\omega} - c_y^\phi\right)^2}{N_\omega}&. \end{aligned} \tag{12}$$

On the one-dimensional y_t, it follows from Equation (2) that $T^2 = \frac{N_0 N_1}{N} J_y$ and that FDA is equivalent to seeking a direction $\boldsymbol{w}$ along which two populations differ mostly.

On a small size of samples, the resulted $\boldsymbol{w}$ by FDA may suffer the well-known overfitting problem, for which efforts have been made on learning a linear boundary in the literature of machine learning. One classical method is the support vector machine (SVM) (Suykens and Vandewalle 1999; Suykens et al. 2002).

Widely adopted in the studies of pattern classification and machine learning, the performance of discriminative analyses is typically measured by the misclassification rate of Equation (10), featuring the separation or overlap of two populations around the boundary and reflecting the confusing chance incurred by a decision or prediction (sometimes called class prediction (Simon et al. 2003)).

The performance of discriminative analyses may also be measured by T^2 that considers the separation of two populations of $y_t = \mathbf{w}^T \mathbf{x}_t$. Monotonically varying with T^2, the p value may be obtained by a univariate t-test. Here, the performance is measured by only considering the salient difference between two populations along the normal direction of the boundary, instead of considering the overall difference in the entire space as addressed after Equation (2).

Alternatively, see Equation (31) in (Xu 2013a), the performance of discriminative analyses may be also measured by a statistics that jointly considers the separating boundary and its outcome by Equation (10).

Since there are different choices for evaluating the difference between two populations, we are motivated to examine whether they can be integrated for a better evaluation. The name of IHT was previously advocated in (Xu 2013a, 2013b) for a joint consideration of the misclassification rate and the p value about the overall difference. This paper will further proceed along this direction.

Cox regression and linear mixed model

Survival analyses consider the relation of the observed time y_t that a subject t passes before some event occurs to one or more covariates in $\boldsymbol{x}_t$ that may be associated with y_t. The Cox model for survival analysis (Cox and Oakes 1984) describes the hazard ratio as follows:

$$h_r(t) = e^{y_t},\ y_t = \mathbf{w}^T \mathbf{x}_t, \tag{13}$$

which shares the common part $y_t = \mathbf{w}^T \mathbf{x}_t$ with Equation (3). The difference is that $\boldsymbol{w}$ is estimated via maximising the following partial likelihood $L(\mathbf{w})$:

$$\max_{\mathbf{w}} L,\ L(\mathbf{w}) = \prod_{t:\omega_t=1} \frac{e^{\mathbf{w}^T \mathbf{x}_t}}{\sum_{\tau:y_\tau > y_t} e^{\mathbf{w}^T \mathbf{x}_\tau}}. \tag{14}$$

Again, we can test H_0 by Equation (5) with the Wald test by Equation (7) or Rao's score test by Equation (8), with help getting $\Delta(\mathbf{w}), I(\mathbf{w})$ still by Equation (6) but with L given by the above partial likelihood $L(\mathbf{w})$.

Actually, the core part $y_t = \mathbf{w}^T \mathbf{x}_t$ of Equations (3) and (13) is also the core part of the classic multivariate linear regression $y_t = \mathbf{w}^T \mathbf{x}_t + \mathbf{e}_t$ with $\boldsymbol{w}$ estimated by minimising $\sum_t \mathbf{e}_t^2$.

Denoting $\boldsymbol{y} = \left[y_1, \cdots, y_N\right]^T$, $\boldsymbol{e} = [e_1, \cdots, e_N]^T$, and $X = [\mathbf{x}_1, \cdots, \mathbf{x}_N]^T$, we may rewrite $y_t = \mathbf{w}^T \mathbf{x}_t + \mathbf{e}_t$ into $\boldsymbol{y} = X\mathbf{w} + \mathbf{e}$ as a degenerated case of the following linear mixed model (Demidenko 2013) :

$$\begin{aligned} \mathbf{y} &= X\mathbf{w} + Z\mathbf{f} + \mathbf{e}, \\ \mathbf{f} \sim G(\mathbf{f}|0, K),\ \mathbf{e} &\sim G(\mathbf{e}|0, R), \end{aligned} \tag{15}$$

where Z is a design matrix and $\boldsymbol{f}$ is a random effect vector. We may use the existing methods to estimate $\boldsymbol{w}, K, R$ (Demidenko 2013) and then test $\boldsymbol{w} = 0$ via the Wald test by Equation (7) or Rao's score test by Equation (8) but with the likelihood L replaced by:

$$L = G(\mathbf{y} - X\mathbf{w}|\mathbf{0}, ZKZ^T + R). \tag{16}$$

Moreover, an $N \times 1$ vector $\boldsymbol{y}$ may be further extended to a $N \times m$ matrix with one dependent variable extended to m-dependent variables. Accordingly, $\boldsymbol{w}, \mathbf{f}, \mathbf{e}$ are extended to $d \times m$ matrices. As a result, we have:

$$Y = XW + Z\mathbf{F} + \mathbf{E}, \tag{17}$$

where $\boldsymbol{F} = [\mathbf{f}_1, \cdots, \mathbf{f}_m]$, and $\boldsymbol{E} = [\mathbf{e}_1, \cdots, \mathbf{e}_m]$. One typical case is that $\mathbf{f}_1, \cdots, \mathbf{f}_m$ are mutually i.i.d. with each $\mathbf{f}_i \sim G(\mathbf{f}_i|0, K)$. Also, $\boldsymbol{e}_1, \cdots, \boldsymbol{e}_m$ are i.i.d. with each $\boldsymbol{e}_i \sim G(\mathbf{e}_i|0, R)$.

From inner product to bi-linear form

In many studies of multivariate statistical analysis and machine learning, a basic sampling unit is a vector $\boldsymbol{x}_t = \left[x_t^{(1)}, \cdots, x_t^{(d)}\right]^T$, and the basic computing operation is the inner product $\boldsymbol{w}^T \mathbf{x}_t$ that is linear with respect to the elements of $\boldsymbol{x}_t$ and also of $\boldsymbol{w}$. Though $\boldsymbol{w}^T \mathbf{x}_t$ becomes XW in Equation (17), it actually consists of a set of vector inner products in parallel.

Efforts have been made in (Xu 2013a, 2013b) to extend this inner product to get a matrix-variate discriminative analysis. Considering that a basic sampling unit is a matrix X_t as shown in Figure 1, the inner product is extended into a bi-linear form:

$$\begin{aligned} y_t &= \mathbf{w}^T X_t \mathbf{v} = \sum_{i=1}^{m}\sum_{j=1}^{d} w^{(i)} v^{(j)} x_t^{(i,j)} \\ &= \sum_{i=1}^{m} w^{(i)} \left(\mathbf{v}^T \mathbf{x}_t^{(i)} \right) = \mathbf{w}^T \mathbf{x}_t^{v},\ \mathbf{x}_t^{v} = X_t \mathbf{v}, \end{aligned} \tag{18}$$

which is quadratic with respect to $w^{(i)}$ and $v^{(j)}$ but still linear with respect to the elements of X_t and is featured by two consecutive layers of inner products. Similarly, we may also have $\boldsymbol{w}^T X_t \mathbf{v} = \mathbf{v}^T \mathbf{x}_t^{w}$ and $\boldsymbol{x}_t^{w} = X_t^T \mathbf{w}$. We call such a matrix-variate-based basic-computing operation a bi-linear form. This bi-linear form leads us to matrix-variate LDA and factor analyses in (Xu 2013a, 2013b). Also, using matrix normal distribution, the implementations are made by the Bayesian Ying Yang harmony learning (Xu 1995, 2015).

To get further insight, we directly extend the vector inner product into the following matrix format:

$$y_t = \mathrm{vec}^T[O]\,\mathrm{vec}[X_t] = \sum_{i=1}^{m}\sum_{j=1}^{d} o^{(i,j)} x_t^{(i,j)}, \tag{19}$$

which is still linear with respect to the elements of X_t but unable be decomposed into two inner products, where vec$[O]$ denotes the vectorisation of a matrix O.

Comparing Equations (18) and (19), we observe that the bi-linear form can be regarded as constrained in the following structure :

$$o^{(i,j)} = w^{(i)} v^{(j)},\ or\ O = \mathbf{w}\mathbf{v}^T. \tag{20}$$

That is, the weighting along the rows of X_t is unrelated to one along the columns of X_t. It significantly reduces the number of free parameters of $o^{(i,j)}$ from md into $m + d$ for $w^{(i)}$ and $v^{(j)}$, which is favourable because we usually have a small-size N for a given sample set $\mathcal{X}_N$. However, it also suffers the limitation of being applicable only to the cases where the dependence across rows of X_t is not related to one along the columns of X_t. To extend such a limitation, further generalisations of bi-linear matrix forms will be proposed in Equation (40).

Methods

KL statistics and matrix-variate tests

Given the case and control samples $X_\omega = \{\mathbf{x}_{t,\omega}, t = 1, \cdots, N_\omega$ and $\omega = 0, 1\}$ from a parametric family $q(\mathbf{x}|\theta)$, all the unknown parts of the true value θ^* are estimated under H_0 by Equation (1), e.g. by the maximum likelihood from $X_0 \cup X_1$. Also, we estimate $\hat{\theta}$ from X_1 and test whether H_0 breaks by the following formulation (see Equation (36) in (Xu 2012a)):

$$\begin{aligned} s_{KL} &= KL(q(\mathbf{x}|\theta^*)||q(\mathbf{x}|\hat{\theta})),\ \text{with} \\ KL(p||q) &= \int p(u) \ln \frac{p(u)}{q(u)} du, \end{aligned} \tag{21}$$

from which the Hotelling T^2 statistics (Hotelling 1931) and FDA are obtained as its special cases.

Alternatively, we may also rewrite H_0 into

$$H_0 : \text{no difference between } \; q(\mathbf{x}|\theta_1) \; \text{ and } \; q(\mathbf{x}|\theta_0), \tag{22}$$

with X_1 from $q(\mathbf{x}|\theta_1)$ and X_0 from $q(\mathbf{x}|\theta_0)$. We estimate θ_1 from the case samples X_1 and θ_0 from the control samples X_0 by either the maximum likelihood or other learning principles, and test H_0 by the following case-control formula:

$$s_{KL} = KL(q(\mathbf{x}|\theta_0)||q(\mathbf{x}|\theta_1)), \tag{23}$$

which directly measures the discrepancy between the case population and control population and provides a general formulation for model-based tests. In contrast, s_{KL} by Equation (21) indirectly considers the difference of the case population from the pool of both populations under H_0.

For the special case that $q(\mathbf{x}|\theta) = G(\mathbf{x}|\mathbf{c}, \Sigma)$, s_{KL} by Equation (21) and s_{KL} by Equation (23) are equivalent with merely a slight difference of a constant scale, resulting in:

$$\begin{aligned} s_{KL} &= KL(G(x|\mathbf{c}_0, \Sigma)||G(x|\mathbf{c}_1, \Sigma)) \\ &= 0.5Tr\left[(\mathbf{c}_0 - \mathbf{c}_1)\,(\mathbf{c}_0 - \mathbf{c}_1)^T\, \Sigma^{-1}\right]. \end{aligned} \tag{24}$$

It relates to the Hotelling statistics by Equation (2) via $T^2 = 2\frac{N_0N_1}{N_0+N_1}s_{KL}$, i.e. the Hotelling statistics is covered as a special case of the general formulation by Equation (21).

The equivalence no longer exists when we consider other examples of $q(\mathbf{x}|\theta_1)$ and $q(\mathbf{x}|\theta_0)$. Because the case population reflects an abnormal situation and thus has a distribution that is quite different from the control population; $q(\mathbf{x}|\theta_1)$ may come from a parametric family that is different from the one of $q(\mathbf{x}|\theta_0)$. For an example, we may consider a Gaussian for the control samples while a mixture of two Gaussians for the case samples.

In addition to testing $\boldsymbol{c}_0 = \mathbf{c}_1$ as considered by the Hotelling statistics, we may use s_{KL} by Equation (23) to develop statistics for other null hypotheses of the type $\theta_0^s = \theta_1^s$. For examples, θ_i^s could be a covariance Σ_i.

Generally, we may use s_{KL} by Equation (21) to develop a statistics for testing a general relation given by a vector equation $\boldsymbol{h}(\theta) = 0$ that consists of one or several joint equations, for which we estimate θ_0 from samples of $X_0 \cup X_1$ subject to the constraint $\boldsymbol{h}(\theta) = 0$ and estimate θ_1 from only the case samples X_1 without the constraint. The above type $\theta_0^s = \theta_1^s$ is a special case $h(\theta) = \theta_0^s - \theta_1^s = 0$. Also, the equality may be extended to several subsets $\{\theta_i^s\}$ that are equal to each other, with each θ_i^s to be either of the mean vector $\boldsymbol{c}_i$ or a covariance Σ_i. Even the simplest case $\theta^s = 0$, $\theta^s \subseteq \theta$ has been widely studied. For examples, θ^s could be the variances for the variance analyses or $\boldsymbol{w} = 0$ in Equation (5) for logistic regression and Cox regression.

Not only Equation (21) provides a general formulation of developing a statistics for a composite test, but also a bird view of the existing statistics for further understanding, improvements, and extensions.

Simply with each vector $\boldsymbol{x}$ replaced by a matrix X, we can extend Equations (21) and (23) to consider matrix-variate samples. Without losing generality, we focus on Equation (23) and get:

$$s_{KL} = KL(q(X|\theta_0)||q(X|\theta_1)). \tag{25}$$

We consider $q(\mathbf{x}|\theta)$ given by the following matrix normal distribution (MND) (Dutilleul 1999; Xu 2012a) :

$$N(X|C,\Omega,\Sigma)=\frac{e^{-0.5Tr\left[\Omega^{-1}(X-C)^T\Sigma^{-1}(X-C)\right]}}{(2\pi)^{0.5md}|\Sigma|^{0.5d}|\Omega|^{0.5m}}, \tag{26}$$

where a matrix Ω describes the cross-column dependence of the matrix variate X, and a matrix Σ describes the cross-row dependence of X. This matrix distribution is equivalent to a multivariate Gaussian distribution $G(\text{vec}(X)|\text{vec}(C),\Sigma\otimes\Omega)$, where $\otimes$ denotes the Kronecker product.

With each sample $X_{t,\omega}$ from $N\left(X|C^x_\omega,\Omega^x_\omega,\Sigma^x_\omega\right)$ under the assumption:

$$\Sigma^x=\Sigma^x_0=\Sigma^x_1,\ \Omega^x=\Omega^x_0=\Omega^x_1, \tag{27}$$

it follows from Equation (25) that we obtain:

$$\begin{aligned} s_{KL}&=KL\left(N\left(X|C^x_1,\Omega^x,\Sigma^x\right)||N\left(X|C^x_0,\Omega^x,\Sigma^x\right)\right)\\ &=Tr\left[\Omega^{x\ -1}\left(C^x_1-C^x_0\right)^T\Sigma^{x\ -1}\left(C^x_1-C^x_0\right)\right], \end{aligned} \tag{28}$$

as the matrix-variate counterpart of Equation (24), where parameters are typically estimated by the maximum likelihood principle (Xu, 2015).

Generally, with help of Equation (25), we may also develop statistics for distributions other than matrix normal distributions.

Model-based two-sample tests

The tests for H_0 by Equation (22) are featured by comparing the difference between two parametric models $q(\mathbf{x}|\theta_1)$ and $q(\mathbf{x}|\theta_0)$ on the entire domain of $\boldsymbol{x}$. Its basis is modelling the case population by $q(\mathbf{x}|\theta_1)$ with its parameter θ_1 estimated from X_1 and modelling the control population by $q(\mathbf{x}|\theta_0)$ with its parameter θ_0 estimated from X_0. Thus, these tests are called *model-based two-sample tests* or *model-based tests* in short wherever there is no confusion caused.

Typically, a statistics s is considered to measure the difference between two models. The bigger the value s is, the larger the difference is. We reject H_0 when s takes a large enough value s^*, while the false positive probability of this rejection is called the p value.

Usually, how to get a statistics s from samples is task-dependent. It is typically a function of the first- and second-order statistics that are random variables directly obtained from samples of populations, e.g. see the Hotelling statistics by Equation (2). Equation (23) provides a general perspective of getting such a statistics s_{KL}, covering not only the first- and second-order statistics but also ones beyond.

Actually, Equation (23) can be further generalised. Adding in the priorities α_1,α_0 for $q(\mathbf{x}|\theta_1)$ and $q(\mathbf{x}|\theta_0)$, we have:

$$\begin{aligned} KL_{10}&=KL(\alpha_1 q(\mathbf{x}|\theta_1)||\alpha_0 q(\mathbf{x}|\theta_0))\\ &\quad=\alpha_1 KL(q(\mathbf{x}|\theta_1)||q(\mathbf{x}|\theta_0))+\alpha_1\delta_\Gamma,\\ \delta_\Gamma&=\ln\frac{\alpha_1}{\alpha_0}=\ln\alpha_1-\ln\alpha_0, \end{aligned} \tag{29}$$

which describes the difference observed from the case side. From the control side, we have also:

$$KL_{01}=KL(\alpha_0 q(\mathbf{x}|\theta_0)||\alpha_1 q(\mathbf{x}|\theta_1))=\alpha_0 KL(q(\mathbf{x}|\theta_0)||q(\mathbf{x}|\theta_1))-\alpha_0\delta_\Gamma.$$

We further get their average and difference as follows:

$$\begin{aligned} KL_{\text{sum}} &= \frac{KL_{10} + KL_{01}}{2} = \int \frac{\alpha_1 q(\mathbf{x}|\theta_1) - \alpha_0 q(\mathbf{x}|\theta_0)}{2} \ln \frac{\alpha_1 q(\mathbf{x}|\theta_1)}{\alpha_0 q(\mathbf{x}|\theta_0)} d\mathbf{x}, \\ KL_{\text{dif}} &= KL_{10} - KL_{01} = \int q(\mathbf{x}|\theta) \ln \frac{\alpha_1 q(\mathbf{x}|\theta_1)}{\alpha_0 q(\mathbf{x}|\theta_0)} d\mathbf{x}, q(\mathbf{x}|\theta) = \alpha_1 q(\mathbf{x}|\theta_1) + \alpha_0 q(\mathbf{x}|\theta_0). \end{aligned} \tag{30}$$

For $q(\mathbf{x}|\theta) = G(\mathbf{x}|c, \Sigma)$, we have:

$$\begin{aligned} KL_{1,0} &= \alpha_1 \left(\delta_{\alpha,\Sigma} + \delta\mathbf{c}^T \Sigma_1^{-1} \delta\mathbf{c} \right), KL_{0,1} = \alpha_0 \left(-\delta_{\alpha,\Sigma} + \delta\mathbf{c}^T \Sigma_0^{-1} \delta\mathbf{c} \right), \\ KL_{\text{sum}} &= \frac{(\alpha_1 - \alpha_0)\delta_{\alpha,\Sigma} + \delta\mathbf{c}^T \Sigma_\Gamma^{-1} \delta\mathbf{c}}{2}, KL_{\text{dif}} = \delta_{\alpha,\Sigma} + \delta\mathbf{c}^T \left[\alpha_1 \Sigma_1^{-1} - \alpha_0 \Sigma_0^{-1} \right] \delta\mathbf{c}, \\ \Sigma_\Gamma^{-1} &= \alpha_0 \Sigma_0^{-1} + \alpha_1 \Sigma_1^{-1}, \delta\mathbf{c} = (\mathbf{c}_1 - \mathbf{c}_0)/\sqrt{2}, \delta_{\alpha,\Sigma} = \ln \frac{\alpha_1}{|\Sigma_1|^{0.5}} - \ln \frac{\alpha_0}{|\Sigma_0|^{0.5}}, \end{aligned} \tag{31}$$

from which we observe how an overall difference is structured from the statistics on individual differences. For KL_{sum}, the role of anti-dispersion difference $\delta_{\alpha,\Sigma}$ is cancelled while the position difference $\delta\mathbf{c}$ is averaged. For KL_{dif}, the role of $\delta_{\alpha,\Sigma}$ is summed up while the position difference $\delta\mathbf{c}$ is cancelled. In other words, the roles of KL_{sum} and KL_{dif} are complementary. According to the nature of tasks, we may use either of them separately or the both of them jointly.

The performance of examining H_0 by Equation (22) is typically evaluated via the p value, which depends on not only how p is approximately estimated but also how well $q(\mathbf{x}|\theta_0)$ models X_0 and $q(\mathbf{x}|\theta_1)$ models X_1. A poor modelling makes the resulted p unreliable. Thus, the performance evaluation should also consider its corresponding modelling error or generally the likelihood:

$$L = \ln\left[\alpha_0 q(X_0|\theta_0) + \alpha_1 q(X_1|\theta_1)\right]. \tag{32}$$

The modelling error depends not only on what type of model is used but also on an appropriate model complexity. Using a model with a big model complexity can lead to an over-optimistic result, i.e. suffering an over-fitting problem. To remedy it, we need to consider either an average of modelling errors on training and testing samples (e.g. by cross validation (Stone 1974)) or approximated generalisation error by one of the model-selection criterion (e.g. BIC (Schwarz 1978)).

Jointly, model-based two-sample tests involve two tasks, that is, the first two tasks summarised in Table 1. Task A is a typical topic of machine learning, from which those existing studies can be adopted, while task B is a typical topic of a statistical test, with its corresponding ε_B being a nonnegative measure that monotonically decreases towards zero as s tends towards a large value.

It is an open challenge to integrate ε_A and ε_B into one objective to optimise because of lacking investigations on how to combine them. A preliminary study has been made empirically with the help of the 2D scattering plots of ε_A versus ε_B as illustrated in Figure 2. Each scattering point denotes a performance pair $(\varepsilon_A, \varepsilon_B)$, associated with one miRNA on the samples for gene expression. Those points located near the origin (e.g. those in the orange colour) act as the interested candidate points.

Table 1 Four Tasks of Integrative Hypothesis Tests

Tasks	Description
Task A (modelling)	estimate θ_ω such that $q(\mathbf{x}\|\theta_\omega)$ models the corresponding population of samples, with the performance evaluated by its corresponding ε_A, e.g., the average error or generalisation error.
Task B (comparison)	develop a statistics s based on the resulted models to test H_0 by Equation (22), with the performance evaluated by its corresponding ε_B that measures the difference between two populations, e.g., the p-value.
Task C(classification)	classify each sample to either $\omega = 1$ or 0, with the performance evaluated by its corresponding ε_C, e.g., either the rate of incorrect classification by Equation (44) or alternatively the corresponding p-value obtained by a test based on a statistics by Equation (47).
Task D (assurance)	test whether a reliable separating boundary exists between the two populations of samples, with the performance evaluated by its corresponding ε_D.

Matrix-variate discriminative analysis

As addressed around Equation (11), the classic FDA seeks a projection $y_t = \mathbf{w}^T\mathbf{x}_t$ to maximize J_y. Moreover, it follows from the bi-linear form by Equation (18) that a matrix-variate discriminative analysis is obtained by:

$$\begin{aligned}\{\mathbf{w}^*, \mathbf{v}^*\} &= \arg\max_{\mathbf{w},\mathbf{v}} J(\mathbf{w},\mathbf{v}), J(\mathbf{w},\mathbf{v}) = \frac{\mathbf{w}^T\left(C_1^x - C_0^x\right)\mathbf{v}\mathbf{v}^T\left(C_1^x - C_0^x\right)^T\mathbf{w}}{\mathbf{w}^T\Sigma_{\mathbf{v}}\mathbf{w}}, \\ &= \frac{\mathbf{v}^T\left(C_1^x - C_0^x\right)^T\mathbf{w}\mathbf{w}^T\left(C_1^x - C_0^x\right)\mathbf{v}}{\mathbf{v}^T\Sigma_{\mathbf{w}}\mathbf{v}}, \Sigma_{\mathbf{v}} = \sum_{\omega=0,1}\sum_{t=1}^{N_\omega}\left(X_{t,\omega} - C_\omega^x\right)\mathbf{v}\mathbf{v}^T\left(X_{t,\omega} - C_\omega^x\right)^T, \\ \Sigma_{\mathbf{w}} &= \sum_{\omega=0,1}\sum_{t=1}^{N_\omega}\left(X_{t,\omega} - C_\omega^x\right)^T\mathbf{w}\mathbf{w}^T\left(X_{t,\omega} - C_\omega^x\right),\end{aligned} \tag{33}$$

which may be solved by iterating:

$$\begin{aligned}&\text{fix } \boldsymbol{v}, \text{ get } \ \mathbf{w}^* \propto \Sigma_{\mathbf{v}}^{-1}\left(C_1^x - C_0^x\right)\mathbf{v}, \mathbf{w} = \mathbf{w}^*/\|\mathbf{w}^*\|, \\ &\text{fix } \boldsymbol{w}, \text{ get } \ \mathbf{v}^* \propto \Sigma_{\mathbf{w}}^{-1}\left(C_1^x - C_0^x\right)^T\mathbf{w}, \mathbf{v} = \mathbf{v}^*/\|\mathbf{v}^*\|,\end{aligned} \tag{34}$$

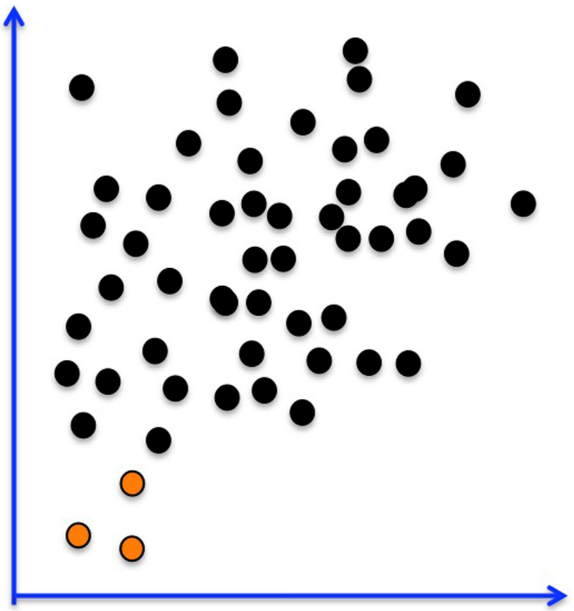

Figure 2 2D scattering plots for joint analyses.

Generally, the bi-linear form by Equation (18) may also be rewritten into the following matrix format:

$$Y_t = V^T X_t W, \tag{35}$$

with a $m \times m_s$ matrix V and a $d \times d_s$ matrix W. It degenerates back to Equation (18) when $m_s = 1, d_s = 1$. Mapping into one variable y_t may lose too much discriminative information. Instead, Equation (35) maps X_t into either of a size-reduced matrix, a column vector, or a row vector according to practical problems, e.g. from not only genomics data in genetic biology but also image or table data in various tasks of big data analyses.

With X_t replaced by Y_t, equations from Equations (25) to (29) are directly applicable. If X_t comes from an MND, Y_t comes from an MND too. Accordingly, Equation (33) becomes:

$$\begin{aligned} &\{W^*, V^*\} = \arg\max_{W,V} J(W,V), \\ &J(W,V) = Tr\left[\Omega^{y\,-1}\left(C_1^y - C_0^y\right)^T \Sigma^{y\,-1}\left(C_1^y - C_0^y\right)\right], \end{aligned} \tag{36}$$

where the parameters are given in a way similar to Equation (28). Also, its solution may be obtained by iterating:

$$\begin{aligned} &\text{Fixing} W, \text{get } V \text{ by solving } \nabla_V J(W,V) = 0, \\ &\text{Fixing} V, \text{get } W \text{ by solving } \nabla_W J(W,V) = 0. \end{aligned} \tag{37}$$

Actually, Equation (35) computes a set of the bi-linear matrix forms in parallel as follows:

$$Y_t = \left[y_t^{(k,\ell)}\right],\ y_t^{(k,\ell)} = \sum_{i=1}^{m}\sum_{j=1}^{d} w^{(i,k)} v^{(j,\ell)} x_t^{(i,j)}. \tag{38}$$

Each $y_t^{(k,\ell)}$ above and the bi-linear form by Equation (18) suffer the limitation discussed after Equation (20), which is relaxed with $v^{(j)}$ replaced by $v_i^{(j)}$ or $v^{(j,\ell)}$ replaced by $v_i^{(j,\ell)}$, i.e. adding another dimension by a subscript i.

Focusing on the former, we extend Equation (20) into:

$$\begin{aligned} &o^{(i,j)} = w^{(i)} v_i^{(j)}, \\ &v_i^{(j)} \text{is subject to a constraint, e.g. one of} \\ &\begin{cases} \sum_{j=1}^{d} v_i^{(j)} = 1, & \text{Choice (a)}, \\ \text{from a Gaussian density,} & \text{Choice (b)}, \\ \text{from a Laplace density,} & \text{Choice (c)}. \end{cases} \end{aligned} \tag{39}$$

Accordingly, we extend Equation (18) into:

$$\begin{aligned} &y_t = \sum_{i=1}^{m}\sum_{j=1}^{d} w^{(i)} v_i^{(j)} x_t^{(i,j)} = Tr[\text{diag}[\mathbf{w}]\, X_t V] = \mathbf{w}^T \mathbf{x}_t^{\mathbf{v}},\ \mathbf{x}_t^{\mathbf{v}} = \left[\mathbf{v}_1^T \mathbf{x}_t^{(1)}, \cdots, \mathbf{v}_m^T \mathbf{x}_t^{(m)}\right]^T, \\ &V = [\mathbf{v}_1, \cdots, \mathbf{v}_d], \mathbf{v}_i = \left[v_i^{(1)}, \cdots, v_i^{(d)}\right]^T, \text{diag}[\mathbf{w}] = \text{diag}\left[w^{(1)}, \cdots, w^{(m)}\right], \end{aligned} \tag{40}$$

where $Tr[A]$ denotes the trace of the matrix A.

Putting it into Equation (11) and considering choice (a) in Equation (39), we get Equation (33) modified into:

$$
\begin{aligned}
&\{\mathbf{w}^*, V^*\} = \arg\max_{\mathbf{w},V} J(\mathbf{w}, V), \text{subject to}: \ \|\mathbf{v}_i\| = 1, \forall i, \\
&J(\mathbf{w}, V) = \frac{\mathbf{w}^T\left(\mathbf{c}_1^{\mathbf{v}} - \mathbf{c}_0^{\mathbf{v}}\right)\left(\mathbf{c}_1^{\mathbf{v}} - \mathbf{c}_0^{\mathbf{v}}\right)^T \mathbf{w}}{\mathbf{w}^T \Sigma_{\mathbf{v}} \mathbf{w}} = \frac{Tr^2\left[\mathrm{diag}\left[\mathbf{w}\right]\left(C_1^x - C_2^x\right) V\right]}{\sum_{\omega=0,1} \sum_{t=1}^{N_\omega} Tr^2\left[\mathrm{diag}[\mathbf{w}]\left(X_{t,\omega} - C_\omega^x\right) V\right]}, \\
&\mathbf{c}_\omega^{\mathbf{v}} = \left[\mathbf{v}_1^T \mathbf{c}_\omega^{x\,(1)}, \cdots, \mathbf{v}_d^T \mathbf{c}_\omega^{x\,(d)}\right]^T, \Sigma_{\mathbf{v}} = \sum_{\omega=0,1} \sum_{t=1}^{N_\omega} \delta \mathbf{x}_{t,\omega}^{\mathbf{v}} \delta \mathbf{x}_{t,\omega}^{\mathbf{v}\,T}, \\
&\text{where } \delta \mathbf{x}_{t,\omega}^{\mathbf{v}} = \left[\mathbf{v}_1^T\left(\mathbf{x}_{t,\omega}^{(1)} - \mathbf{c}_\omega^{x\,(1)}\right), \cdots, \mathbf{v}_m^T\left(\mathbf{x}_{t,\omega}^{(m)} - \mathbf{c}_\omega^{x\,(m)}\right)\right]^T.
\end{aligned}
\tag{41}
$$

which may be solved by iterating:

$$
\begin{aligned}
&\text{fix } \boldsymbol{w}, \ \text{get } V \ \text{by solving } \nabla_V J(\mathbf{w}, V) = 0, \\
&\qquad\qquad \text{subject to}: \ \|\mathbf{v}_i\| = 1, \forall i. \\
&\text{fix } V, \ \text{get } \mathbf{w}^* \propto \Sigma_{\mathbf{v}}^{-1}\left(\mathbf{c}_1^{\mathbf{v}} - \mathbf{c}_0^{\mathbf{v}}\right), \mathbf{w} = \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|}.
\end{aligned}
\tag{42}
$$

For simplicity, we may approximately ignore the coupling across different subscript i and get:

$$
\begin{aligned}
&\mathbf{v}_i^* \propto \Sigma^{(i)\ -1}\left(\mathbf{c}_1^{(i)} - \mathbf{c}_0^{(i)}\right), \ \mathbf{v}_i = \mathbf{v}_i^*/\|\mathbf{v}_i^*\|, \\
&\Sigma^{(i)} = \sum_{\omega=0,1} \sum_{t=1}^{N_\omega}\left(\mathbf{x}_{t,\omega}^{(i)} - \mathbf{c}_\omega^{x\,(i)}\right)\left(\mathbf{x}_{t,\omega}^{(i)} - \mathbf{c}_\omega^{x\,(i)}\right)^T.
\end{aligned}
\tag{43}
$$

This solution does not relate to $\boldsymbol{w}$, and thus, the job is done after getting $\boldsymbol{w}^*$ by Equation (34).

Also, we may update V by a gradient-based approach via $\nabla_V J(\mathbf{w}, V)$. Practically, a regularisation may be added on $J(\mathbf{w}, \mathbf{v})$ and $J(\mathbf{w}, V)$ via Gaussian priories on $\mathbf{w}, \mathbf{v}$, and V. Alternatively, we may make sparse learning via Laplace priories on $\mathbf{w}, \mathbf{v}$, and V.

Being a complementary to model-based two-sample tests that considers H_0 by Equation (22) from an overall perspective of populations, we may also perform the classification task in Table 1 to evaluate the goodness of the decomposition by Equation (10), measured by another quantity ε_C, e.g. the following rate of incorrect classification

$$
\varepsilon_C = \frac{\#X_0^{(1)} + \#X_1^{(0)}}{\#X_0 + \#X_1}. \tag{44}
$$

Classically, an optimal classification is given by:

$$
\omega = \arg\max_j [\alpha_j q(\xi|\theta_j)], \tag{45}
$$

where ξ could be either of $\boldsymbol{x}_t$ and X_t or the corresponding projections y_t and Y_t. Mapping samples into the projections helps to reduce the dimension of $\boldsymbol{x}_t$ and X_t for tackling the overfitting difficulty of task A in Table 1, especially when the size of samples is not large enough. Also, it facilitates visualisation of two populations in a low dimension (especially below 3D dimension) such that classification is made with human interaction.

Boundary-based tests

Actually, the FDA by Equation (11) finds $\boldsymbol{w}$ that defines the normal direction of the best discriminative hyperplane, as shown in Figure 3. In addition to Equation (45), the hyperplane often acts as a separating boundary as follows:

$$g(x, \mathbf{w}) = \mathbf{w}^T\mathbf{x} + w_0 = \mathbf{w}^T(\mathbf{x} - \mu) = 0,$$
$$\text{by which } \boldsymbol{x} \text{ is classified into}$$
$$\begin{cases} \text{a case sample}, & \textit{if } g(\mathbf{x}, \mathbf{w}) > 0, \\ \text{a control sample}, & \textit{if } g(\mathbf{x}, \mathbf{w}) \le 0. \end{cases} \tag{46}$$

That is, it performs task C to get the decomposition by Equation (10) on which we may directly get the measure ε_C by Equation (44).

Alternatively, testing Equation (1) may be made by the following statistics from Equation (10):

$$s = \frac{\#X_1^{(1)} + \#X_0^{(0)}}{\#X^{(1)} + \#X^{(0)}},$$
$$\text{or } s = \frac{\#X_1^{(1)} + \#X_0^{(0)}}{\#X_0^{(1)} + \#X_1^{(0)}}. \tag{47}$$

There are also two other choices in Table 2. Choice (1) is a model-based test for task B from the perspective of one-dimensional samples of $y_t = \mathbf{w}^T\mathbf{x}_t$. Focusing on a most discriminative direction, this test puts attention only on salient differences. As to be addressed later in Table 3, the test can be made together with testing H_0 by Equation (5) such that the rest of the entire sample space is taken into consideration.

Choice (2) in Table 2 provides a statistics for task B on samples without dimension reduction. The statistics s_B comes from considering that samples of $X_1^{(1)}$, $X_0^{(0)}$ should be distant from the boundary (as illustrated by two blue arrows in Figure 3) while samples of $X_0^{(1)}$, $X_1^{(0)}$ should not be far from this boundary (see two red arrows). Actually, s_B is a

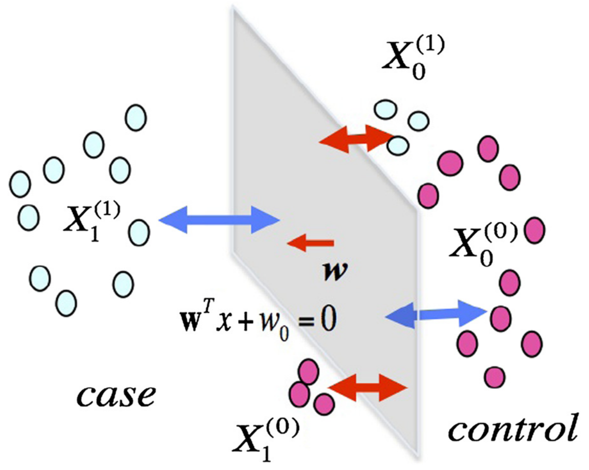

Figure 3 Linear boundary-based statistics.

Table 2 Two Boundary based tests for Task B

Type	Description
(1)	on the projected samples of $y_t = \mathbf{w}^T \mathbf{x}_t$, we use the one dimensional case of Equation (24) or the Welch's t-test to test Equation (1) merely along the normal direction of the boundary.
(2)	measuring the distances of samples from a separating boundary, we consider $s_B = \frac{\sum_{\mathbf{x}\in X_1^{(1)}\cup X_0^{(0)}} \lvert \frac{\mathbf{w}^T(\mathbf{x}-\mathbf{c}_0)}{\lVert\mathbf{w}\rVert} \rvert^q}{\sum_{\mathbf{x}\in X_0^{(1)}\cup X_1^{(0)}} \lvert \frac{\mathbf{w}^T(\mathbf{x}-\mathbf{c}_0)}{\lVert\mathbf{w}\rVert} \rvert^q + \gamma_B}, \; q \geq 0.$ with q=2 for the square distance, q=1 for the Euclidean one.

special case of the ones given by Equations (26) and (30) in (Xu 2013a). The only difference is that $\gamma_B > 0$ is added here to trade off the contribution from $X_0^{(1)} \cup X_1^{(0)}$.

Both two choices in Table 2 are based on the boundary (i.e. either Equation (10) or $y_t = \mathbf{w}^T \mathbf{x}_t$) and thus are called *boundary-based two-sample tests* or *BBT* in short. Different choices of BBT are also coupled with how $\boldsymbol{w}$ is obtained; see some examples outlined in Table 4.

Replacing Equation (11) with the matrix-variate FDA by Equation (33), we get the projection $y_t = \mathbf{w}^T X_t \mathbf{v}$ column by column along the direction $\boldsymbol{w}$ and row by row along the direction $\boldsymbol{v}$. With every appearance of $\boldsymbol{x}$ replaced by $\boldsymbol{x}_t^v = X_t \mathbf{v}$, all the above studies directly apply. Similarly, we may also consider the dual representation $y_t = \mathbf{v}^T \mathbf{x}_t^w$ with $\boldsymbol{x}_t^w = X_t^T \mathbf{w}$ to get a linear separating boundary featured by $\boldsymbol{v}$. It follows from Equations (19) and (20) that $\boldsymbol{w}$ and $\boldsymbol{v}$ jointly form a linear boundary by vec$[O]$ to separate samples of vec$[X_t]$.

Furthermore, extension can be made on the generalised bi-linear form via Equation (40) and Equation (41), with each $\boldsymbol{x}$ replaced by $\boldsymbol{x}_t^v$ given in Equation (40).

Extensions can be also made on the generalised bi-linear form by Equation (35). Samples of two populations are projected into a dimension-reduced matrix $Y_t = V^T X_t W$, and then, a matrix-variate Hotelling test can be made by Equation (28) with X_t replaced by Y_t and the subscript x replaced by y, where the matrices W, V actually take the roles of the boundary.

Table 3 Four Types of Integrative Hypothesis Tests

Types	Description
Type-1 *(model based IHT)*	For Task A, each of two populations is modelled by a parametric model, with ε_A measured by the negative log-likelihood by Equation (32) or its extension to generalisation error. For Task B, a model based test is made to compare the difference between two parametric models, with ε_B by the corresponding p-value. For Task C, we get the classification by Equation (45), with ε_C by Equation (44) or the p-value by a BBT via a statistics obtained from Equation (10).
Type-2 *(boundary based IHT)*	A separating boundary is modelled by a hyperplane with its normal **w**, based on which Task D is handled by a boundary existence test by Equation (5) with ε_D measured by the corresponding p-value. For Task C we get the classification by Equation (46) with ε_C by Equation (44) or alternatively the corresponding p-value obtained by Equation (47), and for Task B we get the p-value by one of two BBT choices in Table 2.
Type-3 *(mixing IHT)*	Mix the above two types with two populations and their separating boundary all in parametric models. A basic one uses $\varepsilon_A, \varepsilon_B$ from Type-1 and $\varepsilon_C, \varepsilon_D$ from Type-2. The other uses $\varepsilon_C, \varepsilon_D$ from Type-2 while $\varepsilon_A, \varepsilon_B$ are modified by Equation (58).
Type-4 *(Ying-Yang IHT)*	Instead of mixing, the parametric models are jointly learned for two populations of samples and their separating boundary. One example is the BYY harmony learning based formulation to be introduced after Equation (60).

Table 4 Some choices for obtaining w

Choice	Description
(a)	get **w** via FDA by Equation (11), as addressed in the previous subsection.
(b)	estimate **w** by maximizing *L* by Equation (4), as to be addressed in the next subsection.
(c)	get **w** as the normal direction of a separating hyperplane by one of machine learning approaches, e.g., support vector machine (SVM) (Cortes and Vapnik 1995; Suykens et al. 2002).

Matrix-variate logistic regression

Testing H_0 by Equation (5) has been widely studied in the literature of logistic regression. Actually, the role of this $\boldsymbol{w}$ is the same as the one in Equation (46), i.e. a discriminative boundary that separates every sample into either $\omega = 1$ or $\omega = 0$. Thus, the choices in Table 4 can be cross-utilised for a mutual benefit, e.g. getting $\boldsymbol{w}$ via FDA by Equation (11) is relatively easy to compute and thus provides an initialization for estimating $\boldsymbol{w}$ by Equation (4), while the advantage of Equation (3) over FDA is that dummy or design variables may be taken into consideration for learning $\boldsymbol{w}$, e.g. we extend $\zeta_t = y_t + c$ in Equation (3) into:

$$\zeta_t = y_t + z_t + c, \; z_t = \mathbf{b}^T \boldsymbol{\xi}_t, \tag{48}$$

where $\boldsymbol{\xi}_t$ consists of dummy variables. Moreover, random effects may also be added, in a way similar to that of the linear mixed model by Equation (15).

Testing H_0 by Equation (5) is typically handled with the Wald test by Equation (7) or Rao's score test by Equation (8), for which the score vector and the information matrix are given as follows (Pan et al. 2014):

$$\begin{aligned} \Delta(\mathbf{w}) &= \sum_{t=1}^{N}(\omega_t - \bar{\omega})(\mathbf{x}_t - \mathbf{c}), \; \mathbf{c} = \frac{1}{N}\sum_{t=1}^{N}\mathbf{x}_t, \\ I(\mathbf{w}) &= \bar{\omega}(1-\bar{\omega})\sum_{t=1}^{N}(\mathbf{x}_t - \mathbf{c})(\mathbf{x}_t - \mathbf{c})^T, \end{aligned} \tag{49}$$

where $\bar{\omega}$ denotes the mean of ω_t.

Being different from the BBT addressed in the previous subsection, testing H_0 by Equation (5) directly aims at whether a boundary $\boldsymbol{w}$ exists. Such a test is thus named *boundary existence test*. It is widely known as a test for regression analyses. Also, we may regard it as a two-sample test that is complementary to the BBT choice (1) in Table 2. The two tests jointly cover the entire space of samples.

The boundary existence test actually tackles another essential problem of discriminative analysis, namely, task D in Table 1. Given two populations with a finite sample size, it is not difficult to draw a boundary to separate them if there is no restriction on the complexity of the boundary. However, a boundary with a high complexity will be unreliable to separate new samples that come randomly from the same populations. To be reliable, the boundary should have an appropriate complexity too. It follows from Equation (45) that an optimal separating boundary is related to the models $q(\mathbf{x}|\theta_1)$ and $q(\mathbf{x}|\theta_0)$. In other words, appropriate boundary complexity is related to an appropriate model boundary complexity. Thus, task D and task A in Table 1 are coupled.

Typically, we consider a linear boundary because of its simple complexity. In the literature of pattern recognition (Cortes and Vapnik 1995; Cover 1965) efforts on whether samples of two populations are linearly separable by a hyperplane or a maximum-margin hyperplane can be regarded as examples related to task D in Table 1.

Next, we proceed to consider matrix-variate logistic regression. Putting the case and control samples into a paired set $\{X_t, \omega_t\}, t = 1, \cdots, N$, we extend Equation (3) with the inner product $y_t = \mathbf{w}^T \mathbf{x}_t$ to be replaced by the bi-linear form by Equation (18) or its extension by Equation (40).

Given V, the above studies directly apply when $\boldsymbol{x}_t^{\mathbf{v}}$ in Equation (40) replaces $\boldsymbol{x}_t$ in Equations 3, 4, 7, and 8. The task of learning $\mathbf{w}$, V can be made via the matrix-variate FDA by Equations (34) or (42).

Alternatively, we may estimate $\mathbf{w}$, V via the maximum likelihood L by Equation (4) with the advantage of taking the effect of covariates into consideration. With $-L$ written as $J(\mathbf{w}, V)$, we get it solved by Equation (37) with $\mathbf{w}$ replaced by W, e.g. implemented by the following gradient-based updating (Hosmer et al. 2013):

$$\begin{aligned} \mathbf{w}^{\text{new}} &= \mathbf{w}^{\text{old}} - \eta_w \nabla_{\mathbf{w}} J\left(\mathbf{w}^{\text{old}}, \nu^{\text{old}}\right), \\ \nu^{\text{new}} &= \nu^{\text{old}} - \eta_V \nabla_{\mathbf{v}} J\left(\mathbf{w}^{\text{new}}, \nu^{\text{old}}\right), \end{aligned} \tag{50}$$

where $\eta_w > 0, \eta_V > 0$ are small learning step sizes.

Also, we may test the dual problem of Equation (5) as follows:

$$H_0 : \ \mathbf{v} = \mathbf{0}, \tag{51}$$

for the bi-linear form by Equation (18) simply with $\boldsymbol{v}$ replacing $\boldsymbol{w}$ in Equations 6, 7, 8, and 49. Similarly, extension may also be made to test $H_0 : \ \mathbf{v}_i = \mathbf{0}, \forall i$.

Moreover, we may also apply Equation (21) to develop a statistics as follows:

$$s_{KL} = \sum_t KL(p(\omega_t|\boldsymbol{x}_t, \theta^*)||p(\omega_t|\boldsymbol{x}_t, \hat{\theta})), \tag{52}$$

with $p(\omega_t|\boldsymbol{x}_t, \theta)$ given by Equation (3), where θ^* is estimated via maximising L by Equation (4) under H_0 by Equation (5) and $\hat{\theta}$ is estimated via maximising L by Equation (4) without H_0.

Similarly, we may get a matrix-variate Cox regression with the inner product $\mathbf{w}^T \mathbf{x}_t$ in Equation (13) replaced by the bi-linear form by Equation (18) or its extension Equation (40). Accordingly, we test the H_0 by Equation (5) and the H_0 by Equation (51), using the Wald test with Equation (7) or Rao's score by Equation (8) with $\Delta(\mathbf{w}), I(\mathbf{w})$ computed from Equation (6) but L given by the partial likelihood $L(\mathbf{w})$.

Furthermore, the univariate y_t can be extended into a vector or matrix Y_t. One typical example is a bi-linear regression of Y_t by Equation (35), that is we consider:

$$Y_t = V^T X_t W + E_t, \tag{53}$$

where E_t is independent of X_t and comes from $N(Y_t - V^T X_t W|0, \Lambda, D)$ by Equation (26), while both Λ, D are diagonal matrices.

Again, there are two choices to estimate W, V. One is the matrix-variate FDA by Equation (36). The other is maximising the following likelihood:

$$L = \sum_{t=1}^{N} N\left(Y_t - V^T X_t W|0, \Lambda, D\right). \tag{54}$$

Particularly, when $\Lambda = \lambda I, D = dI$, we are lead to the following least square error approach:

$$\min_{W,V} J(W,V),\ J(W,V) = \sum_{t=1}^{N} Tr\left[\left(Y_t - V^T X_t W\right)\left(Y_t - V^T X_t W\right)^T\right], \tag{55}$$

which may be again handled by Equation (37) with $\mathbf{w}$ replaced by W.

It can be observed that Equation (53) is an extension of Equation (17) with $\boldsymbol{F} = 0$. On the other hand, we may extend Equation (17) into a bi-linear extension as follows:

$$Y = V^T X W + Z\mathbf{F} + \mathbf{E}, \tag{56}$$

which degenerates to:

$$\mathbf{y} = V^T X\mathbf{w} + Z\mathbf{f} + \mathbf{e}, \tag{57}$$

as a bi-linear mixed model extended from Equation (15).

Integrative hypothesis test

Discriminative analysis and testing of H_0 by Equation (1) are made from either a model-based perspective (e.g. performing task A and task B in Table 1) or a boundary-based perspective (e.g. performing task C and task D in Table 1). Moreover, all the four tasks are associated with another problem called feature selection, that is, selecting a number of elements in $\boldsymbol{x}$ to form a subset $\boldsymbol{x}_f$ such that one or more of the four tasks achieves a good enough performance.

In the existing efforts, each of four tasks has been studied individually, with each having its strength and limited coverage. However, performances of these tasks are coupled, and thus, a best set of features for one task may not be necessarily the best for the others.

The complementary nature of *task B* and *task C* was preliminarily discussed in Section VI in (Xu 2012a), where a model-based test for *task B* is named as A-test (a test in the observed data domain) and a boundary-based test for *task C* is named as I-test (a test in the inner representation domain). Under the name of IHT, good performances of *task B* and *task C* are demanded jointly (Xu 2013a, 2013b). This paper further extends IHT to include *task A* and *task D*.

We start at jointly optimising the performances of *task B* and *task C*. Its necessity and feasibility are empirically justified, with help of the 2D scattering plots of ε_B by the p value for measuring the performance of *task B* and ε_C by the misclassification rate for measuring the performance of *task C*. A small ε_B indicates a big difference between $q(\mathbf{x}|\theta_0)$ and $q(\mathbf{x}|\theta_1)$ from an overall perspective, and a small ε_C indicates a well classification of samples from a separating boundary perspective. Illustrated in Figure 4 are two examples obtained from one empirical study.

As indicated by the blue vertical dashed line in Figure 4, there are many miRNAs that share a same small p value ε_B but can take different values of misclassification ε_C in a big range. Also, as indicated by the blue horizontal dashed line in Figure 4, there could be multiple miRNAs that take a same misclassification but take different p values. In other words, though the performance of one task is optimised, the performance of the other can still be poor. Thus, we need to jointly seek the good performances of both the tasks, i.e. IHT is necessary. On the other hand, it is observable from the red dots within the blue

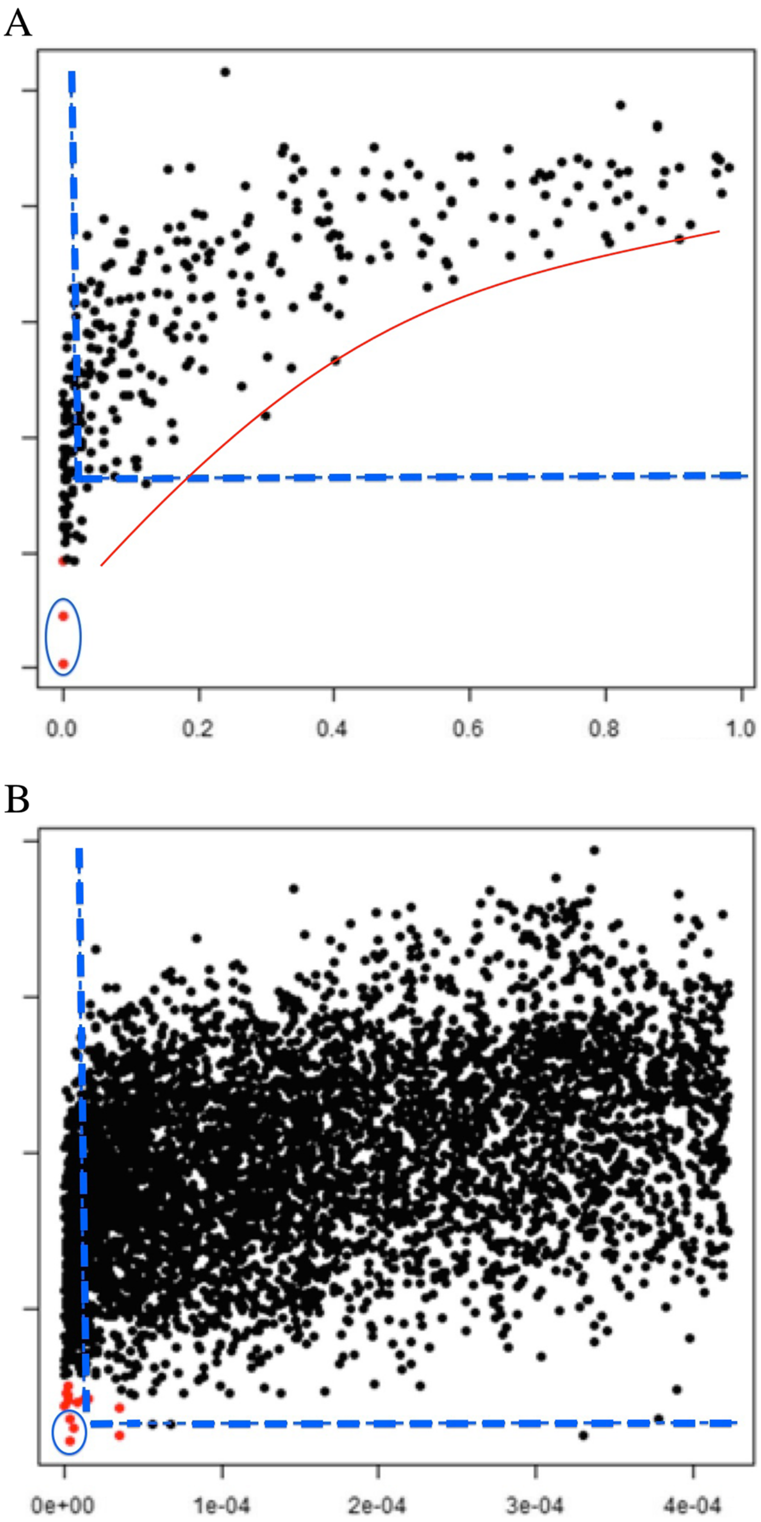

Figure 4 Necessity and feasibility of evaluating the performances of *tasks B* and *C*, on the samples of gene expressions. A scattering point denotes a performance pair with the *x*-axis for *p* value and the *y*-axis for misclassification, associated with one miRNA for **(A)** and two miRNAs for **(B)**.

circle in Figure 4 that there are indeed a few scattering points with each taking both a small p value ε_B and a small misclassification ε_C, i.e. it is also feasible to achieve the goal of IHT too.

Such a 2D plot's evaluation provides a tool for better joint performances of *task B* and *task C*, by which we may interactively observe the configuration of scattering points and locate the candidate points that are nearest to the origin of the coordinate space.

Extensions can be further made to a joint evaluation of the IHT performance with *task A* and *task D* also included, such that the strengths of different tests and methods are integrated in a rather systemic way, for which we address four types of IHT in Table 3.

From the model-based perspective, the first type is an extension of the one addressed in Figure 2, with ε_C added in to get a 3D plots for a joint evaluation of ε_A, ε_B, and ε_C. Instead of Equation (45), we may get ε_C by some nonparametric classifiers, e.g. the classic kNN classifier and the kernel classifiers (Williams 2003). Moreover, we are unable to handle task D because the boundary involved here does not have an explicit expression to be tested.

From the boundary-based perspective, the second type considers samples jointly by a separating boundary and projected samples, evaluated by ε_D for the existence of boundary, ε_C for the misclassification by the boundary, and ε_B for measuring the difference of two populations either along the normal direction of the boundary or according to the sample deviations from the boundary. Again, we may use a 3D plots for a joint evaluation of ε_B, ε_C, and ε_D. However, it is difficult to handle task A merely based on the boundary.

The type of *mix-modelled IHT* combines the above two types to avoid the weak points of each type. Two typical examples are listed in Table 3. One picks $\varepsilon_A, \varepsilon_B$ from type (1) and $\varepsilon_C, \varepsilon_D$ from type (2) for a joint evaluation. The other modifies $\varepsilon_A, \varepsilon_B$ by taking the outcome by Equation (10) of the boundary in consideration, with the original estimated θ_0 and θ_1 replaced by the following maximum likelihood estimation:

$$\max_{\theta} q\left(X_0^{(0)}|\theta\right) \text{ and } \max_{\theta} q\left(X_1^{(1)}|\theta\right). \tag{58}$$

Even better, we may estimate each θ_ω by the maximum likelihood on the entire set X of samples but with the likelihood of each sample weighted by its corresponding posteriori $p(\omega|\text{sample})$ by Equation (3).

BYY-harmony-learning-based formulation

The 2D plots and 3D plots only provides a preliminary tool for IHT, we need further studies on not only appropriate combinations of multiple p values and misclassification rates but also simultaneous optimisation of multiple measures. For the latter purpose, the *mix-modelled IHT* in Table 3 is further extended via iteratively learning θ_0 and θ_1 by Equation (58) to update the models $q\left(X_0^{(0)}|\theta_0\right), q\left(X_1^{(1)}|\theta_1\right)$ and also re-estimating the boundary $\boldsymbol{w}$, e.g. by a FDA method based on the updated models.

Leaving the task D for a future study, in the sequel, we further understand the task of learning the models from a perspective of learning a Ying machine and the task of learning the boundary from a perspective of learning a Yang machine, which leads to a BYY-harmony-learning-based formulation for IHT.

We start from revisiting Equation (29) from an IHT perspective. From $\alpha_1 q(\mathbf{x}|\theta_1) = q(\mathbf{x}|\theta) - \alpha_0 q(\mathbf{x}|\theta_0)$, we consider the task B by the following measure:

$$\begin{aligned} KL_{10} &= KL(\alpha_1 q(\mathbf{x}|\theta_1)||\alpha_0 q(\mathbf{x}|\theta_0)) = \\ &= \int \alpha_1 q(\mathbf{x}|\theta_1) \ln\left[\alpha_1 q(\mathbf{x}|\theta_1)\right] d\mathbf{x} - e^c_{1,0} \\ &= L_1 - \left(e^c_{0,1} + e^c_{1,0}\right), \\ L_1 &= \int q(\mathbf{x}|\theta) \ln\left[\alpha_1 q(\mathbf{x}|\theta_1)\right] d\mathbf{x}, \\ e^c_{i,j} &= \int \alpha^{(i)} q(\mathbf{x}|\theta_i) \ln\left[\alpha^{(j)} q(\mathbf{x}|\theta_j)\right] d\mathbf{x}, \end{aligned}$$

from which we observe that a large KL_{10} comes from a large L_1 that reflects a good modelling of $\alpha_1 q(\mathbf{x}|\theta_1)$ (i.e. a good performance of task A) and a small confusion error $e^c_{0,1}+e^c_{1,0}$ that is closely related to a small misclassification (i.e. a good performance of task C). In other words, three tasks are coordinately optimised.

However, a good modelling on the control samples has not been taken in the consideration of KL_{10}, which may be further improved by considering:

$$\begin{aligned} KL_{\text{sum}} &= \frac{L_1 + L_0}{2} - \left(e^c_{0,1} + e^c_{1,0}\right), \\ L_0 &= \int q(\mathbf{x}|\theta) \ln\left[\alpha_0 q(\mathbf{x}|\theta_0)\right] d\mathbf{x}. \end{aligned} \tag{59}$$

From this KL_{sum}, we need to get $\theta_\omega, \omega = 0, 1$ by the ML learning. In other words, KL_{sum} merely takes a role of evaluating the performances of *task B* and *task C*, but do not have a port to accommodate samples for estimating $\theta_\omega, \omega = 0, 1$. Favourably, such a port is provided in the BYY harmony learning such that *task A*, *task B*, and *task C* are all jointly implemented.

Firstly, proposed in (Xu 1995) and systematically developed in the past two decades, the BYY harmony learning on typical structures leads to new model selection criteria, new techniques for implementing learning regularisation, and developing a class of algorithms that implement automatic model selection during parameter learning. Readers are referred to (Xu 2010, 2012b, 2015) for the latest introduction about the BYY harmony learning.

Briefly, a BYY system consists of a Yang machine and Ying machine corresponding to two types of decomposition, namely, Yang $p(R|X)p(X)$ and Ying $q(X|R)q(R)$, respectively. The data X is regarded as generated from its inner representation R that consists of latent variables Y and parameters θ. The harmony measure is mathematically expressed as follows:

$$H(p||q) = \int p(R|X)p(X) \ln[\, q(X|R)q(R)]\, dXdR. \tag{60}$$

Maximising this $H(p||q)$ makes this Ying Yang pair not only best matched but also have the least complexity. Such an ability can also be further observed from several perspectives (see Section 4.1 in (Xu 2010)).

Applied to $\alpha_1 q(\mathbf{x}|\theta_1)$ and $\alpha_0 q(\mathbf{x}|\theta_0)$, we have:

$$\begin{aligned} H(p||q) &= \sum_{\omega=0,1} \int p(\omega|\mathbf{x}_t)p(\mathbf{x}) \ln[\alpha_\omega q(\mathbf{x}|\theta_\omega)]\, d\mathbf{x}, \\ p(\omega|\mathbf{x}_t) &= \frac{\alpha_\omega q(\mathbf{x}|\theta_\omega)}{\sum_{\omega=0,1} \alpha_\omega q(\mathbf{x}|\theta_\omega)}, \end{aligned} \tag{61}$$

where $p(\mathbf{x})$ provides a port to accommodate samples $\{\mathbf{x}_t\}_{t=1}^N$ via an empirical $p(\mathbf{x}) = \frac{1}{N}\sum_t \delta(\mathbf{x}-\mathbf{x}_t)$ with $\delta(\mathbf{x})$ being the Dirac delta, which thus makes it possible to estimate $\theta_\omega, \omega = 0, 1$ via maximising $H(p||q)$.

It follows from $p(0|\mathbf{x}_t) + p(1|\mathbf{x}_t) = 1$ that we get:

$$\begin{aligned} H(p||q) &= L_1^H + L_0^H - \left(e_{0,1}^H + e_{1,0}^H\right), \\ L_\omega^H &= \int p(\mathbf{x}) \ln\left[\alpha_\omega q(\mathbf{x}|\theta_\omega)\right] d\mathbf{x}, \\ e_{i,j}^H &= \int p(j|\mathbf{x}_t) p(\mathbf{x}) \ln\left[\alpha_i q(\mathbf{x}|\theta_i)\right] d\mathbf{x}. \end{aligned} \tag{62}$$

Approximately considering $p(\mathbf{x}) \approx q(\mathbf{x}|\theta)$, $e_{0,1}^H + e_{1,0}^H \approx e_{0,1}^c + e_{1,0}^c$, and $L_1^H + L_0^H \approx L_1 + L_0$, we observe that $H(p||q)$ shares a nature similar to KL_{sum} in Equation (59), while a difference is that the modelling part $L_1^H + L_0^H$ is provided with a port $p(\mathbf{x})$ to accommodate samples such that task A can be performed via maximising $H(p||q)$ without a need of separately estimating θ_ω by the ML learning.

For $q(\mathbf{x}|\theta) = G(\mathbf{x}|c, \Sigma)$, we implement the maximisation of $H(p||q)$ to estimate θ_ω by directly adopting the semi-supervised BYY harmony learning for Gaussian mixture given in (Xu 2015), i.e. its *algorithm 9*, by which the performances of task A, task B, and task C are coordinated. Moreover, $H(p||q)$ can be extended into its matrix-variate counterpart. Particularly, *algorithm 9* in (Xu 2015) can be extended into the algorithm 1 given below for learning $\alpha_\omega N\left(X|C_\omega^x, \Omega_\omega^x, \Sigma_\omega^x\right)$.

Algorithm 1 *Semi-supervised learning MND mixture*

```
Require: Initialise $p_{\omega,t} = 1/k$.
  Repeat the following two steps until converged:
  Ying-Step: for $\omega = 0, 1$, we get
    $N_\omega = \sum_{t=1}^N p_{\omega,t}$, $\alpha_\omega = \frac{N_\omega}{N_0+N_1}$,
    $C_\omega^x = \frac{\sum_{t=1}^N p_{\omega,t} X_t}{N_\omega}$,
    $\Sigma_\omega^x = \frac{\sum_{t=1}^N p_{\omega,t}(X_t - C_\omega^x)(X_t - C_\omega^x)^T}{N_\omega}$,
    $\Omega_\omega^x = \frac{\sum_{t=1}^N p_{\omega,t}(X_t - C_\omega^x)^T \Sigma_\omega^{x\ -1}(X_t - C_\omega^x)}{N_\omega}$.
  Yang-Step: for $t = 1, \cdots, N$ and $\omega = 0, 1$, we get
  $p_{\omega|x_t} = \frac{[\alpha_\omega N(X|C_\omega^x, \Omega_\omega^x, \Sigma_\omega^x)]^{[\gamma\delta_{\omega,\omega_t^*}+1+\eta]/\eta}}{\sum_{\omega=0,1}[\alpha_\omega N(X|C_\omega^x, \Omega_\omega^x, \Sigma_\omega^x)]^{[\gamma\delta_{\omega,\omega_t^*}+1+\eta]/\eta}}$,
    $p_{\omega,t} = (\eta + 1 + \gamma\delta_{\omega,\omega_t^*}) p_{\omega|x_t}$,
  where $\delta_{ij}$ is the Kronecker delta.
```

Remarks:

(a) There are N samples in total, with each sample X_t associated with a label ω_t^* as follows:

$$\omega_t^* = \begin{cases} 0, & \text{for a sample from the control,} \\ 1, & \text{for a sample from the case,} \\ \#, & \text{for a sample with label missing.} \end{cases}$$

(b) The bigger the $\gamma > 0$ is, the stronger the supervision role of the label ω_t^* will be. We may let $\gamma > 0$ to start at a high value and gradually decrease towards a pre-specified value by a simulated annealing procedure.

(c) η is controlled as described in Sect.2.3 of (Xu, 2015). A simple way is letting η to start from a small value and gradually increase to a large value.

During implementation of the above algorithm, not only task A is performed but also task C can be simply handled in the Yang step by checking whether $p_{1|\mathbf{x}_t} \geq p_{0|\mathbf{x}_t}$ to classify each sample into the case or control. Also, task B can be made after learning by putting the resulted parameters into $s_{KL} = KL_{10}$ or $s_{KL} = KL_{\text{sum}}$ to get the corresponding p value.

Last but not least, considering semi-supervised learning, we also propose an improved procedure in Table 5 for training, testing, and validating on a small size of samples.

Integrating *p* values, inferring rejection domain, and S-space boundary-based tests

Each IHT type in Table 3 involves more than one measure, which incurs for the problem about how different measures are jointly evaluated. Though 2D or 3D plots provide a possible joint evaluation, how to appropriately scale each measure is still a challenging issue. In general, we need to integrate multiple measures into a scalar index based on which the joint performance can be evaluated, which relates closely to efforts made on combing multiple classifiers (Xu and Amari 2008; Xu et al. 1992b) and evidence combination (Barnett 2008).

For an IHT task, the final scalar index is typically the p value. When multiple measures are all in the p values, what we encounter becomes the task of p value combination, e.g. by the Fisher combination (Fisher 1948).

In Table 3, ε_B and ε_D are already given in p values. But ε_A is usually measured by a square error or negative log-likelihood, and ε_C is measured by a misclassification rate. Alternatively, ε_C may be given in a p value via the statistics in Equation (47). Let $s = -\varepsilon_A$ or generally $s = -\varepsilon$ for a monotonic measure $\varepsilon \geq 0$ that prefers values close to zero, we may get the corresponding p value with help of the permutation method.

However, p value combination has a weak point. Each p value is merely a positive number that indicates the false alarm probability, losing certain useful information already. Under the term *meta-analysis* (Evangelou and Ioannidis 2013), efforts have been made by transforming p values into multiple Z statistics such that the missing information is added in without or with help of information directly from data (Zaykin 2011).

Actually, the Hotelling T^2 statistics by Equation (24) and getting a statistics by Equation (21) may also be regarded as examples that get an integrated statistics s_f. Generally, a multivariate hypothesis test may also be regarded as an integration of multiple univariate hypothesis tests.

Table 5 Semi-supervised testing and validating

Issues	Description
Issue-1	Estimate the parameters by semi-supervised learning on the training set, from which we get the corresponding p-value p and a classifier. Using this classifier on the training set and the testing set, it follows from Equation (44) that we get ε_C^{tr} and ε_C^{te}. This is what we traditionally get.
Issue-2	Lump the training samples and testing samples together, and estimate the parameters by semi-supervised learning on the lumped set, we also get the corresponding $\tilde{p}$, $\tilde{\varepsilon}_C^{tr}$ and $\tilde{\varepsilon}_C^{te}$.
Issue-3	$\tilde{p}$ is actually more reliable than p because testing samples are used for regularising parameter estimation. This $\tilde{p}$ is also different from the traditional compounded p-value because the label information of testing samples have not been compounded.
Issue-4	Without using the label information of testing samples, $\tilde{\varepsilon}_C^{te}$ shares the concept same as ε_C^{te}, but is actually more reliable because of regularization.
Issue-5	Merging the training set and testing set to get a big training set and treating the validating set as a new testing set, which actually extends this procedure to improve the validation.

Typically, an integrated statistics $s_f = g(\mathbf{s}, \Psi) \geq 0$ comes from $\boldsymbol{s} = [s^{(1)}, \cdots, s^{(d)}]$ such that $s_f \geq 0$ monotonically increases as the situation differs far from H_0, where each $s^{(i)}$ comes from one univariate hypothesis test (e.g. $\boldsymbol{s} = \mathbf{c}_1 - \mathbf{c}_1$ in the Hotelling T^2 statistics) with a set Ψ of parameters shaping the integration (e.g. the covariance Σ in the Hotelling T^2 statistics). The set Ψ is specified without or with help of information obtained directly from input data. A critical value $\tilde{s}_f$ is computed from the original pair of the sample set X_0, X_1. Then, the false alarm probability $p(s_f > \tilde{s}_f|H_0)$ is obtained as the p value, where and hereafter $p(\cdot|H_0)$ denotes under the condition that H_0 is satisfied.

However, choices for such a $s_f = g(\mathbf{s}, \Psi)$ are very limited in the existing studies, mostly in a quadratic form such as Hotelling statistics, Rao's score by Equation (8), and the Wald test by Equation (7). This is equivalent to approximately regarding $s^{(1)}, \cdots, s^{(d)}$ from a multivariate Gaussian distribution, while other distributions are seldom studied yet.

Instead of seeking an integrated statistics s_f, we directly seek the domain $\Gamma(\tilde{\boldsymbol{s}})$ of rejecting H_0 in the space of $\boldsymbol{s}$ based on a critical vector $\tilde{\boldsymbol{s}}$ as follows:

$$\Gamma(\tilde{\boldsymbol{s}}) \text{ with } \tilde{\boldsymbol{s}}_{X_{1||0}} = I_{nf}(X_0||X_1), \tag{63}$$

where $\tilde{\boldsymbol{s}}_{X_{1||0}} = I_{nf}(X_0||X_1)$ means that $\tilde{\boldsymbol{s}}$ is inferred from the given sample set X_0, X_1 by an inferring method I_{nf}, and the subscript $X_{1||0}$ is used as the abbreviation of $X_1||X_0$, which will be used whenever its omission will not cause confusion.

Then, test is made by checking the probability that $\boldsymbol{s}$ falls in $\Gamma(\tilde{\boldsymbol{s}})$ under H_0, that is:

$$p\left(\boldsymbol{s} \in \Gamma\left(\tilde{\boldsymbol{s}}\right)|H_0\right) = p\left(\boldsymbol{s} \in \Gamma\left(\tilde{\boldsymbol{s}}_{X_{1||0}}\right)|H_0\right). \tag{64}$$

We estimate the p value by a permutation test. That is, we get a new pair of sample sets X_0^{π}, X_1^{π} from X_0, X_1 by a permutation π that shuffles each label ω of $\boldsymbol{x}_{t,\omega}$ and then we obtain:

$$p\left(\boldsymbol{s} \in \Gamma\left(\tilde{\boldsymbol{s}}\right)|H_0\right) = \frac{1}{\#\Pi}\left\{1 + \sum_{\pi \in \Pi} I\left(\boldsymbol{s}_{X^{\pi}_{1|0}} \in \Gamma\left(\tilde{\boldsymbol{s}}\right)\right)\right\},$$
$$I(u) = \begin{cases} 1, & u \text{ is true}, \\ 0, & \text{otherwise}, \end{cases} \tag{65}$$

where $\#S$ denotes the cardinality of a set S, the subscript $X^{\pi}_{1|0}$ is used as the abbreviation of $X_0^{\pi}||X_1^{\pi}$, and Π consists of a large enough set of permutations made by either enumeration or random shuffling, including that π = empty denotes the sample pair X_0, X_1.

Recalling the classic studies of getting an integrated statistics s_f, we observe that $\tilde{s}_f = g(\boldsymbol{s}, \Psi)$ actually define a closed shell or boundary that divides the space of multivariate statistics $\boldsymbol{s}$ (shortly S-space) into two parts, with its inside as the acceptance domain and its outside as the rejection domain $\Gamma(\tilde{\boldsymbol{s}})$. For example, the acceptance domain obtained by both the Hotelling statistics and Rao's score by Equation (8) is a hyper-elliptic volume. We may further extend a hyper-elliptic volume to a bounded volume in another shape. Actually, a bounded acceptance domain corresponds a probabilistic modelling by a single-mode distribution. Thus, the corresponding tests are called S-space model-based tests.

On the other hand, we have also a S-space boundary based test (BBT) as summarised in Table 6. It should not be confused with the BBTs in the space of input data (shortly D-space), as those previously addressed in Tables 2 and 3, as well as in Figure 3. Those are two-sample tests with the boundary for separating two populations in the D-space while the S-space BBTs may correspond to any tests in the D-space.

Table 6 S-space boundary based test (BBT)

Step	Description
(1)	infer $\tilde{\mathbf{s}} = I_{nf}(X_0 \| X_1)$ in the multidimensional space of statistics $\mathbf{s}$, where $\tilde{\mathbf{s}}_{X_{1\|\|0}} = I_{nf}(X_0 \| X_1)$ means that $\tilde{\mathbf{s}}$ is inferred from the given sample set X_0, X_1 by an inferring method I_{nf}, and the subscript $X_{1\|\|0}$ is used as the abbreviation of $X_1 \| X_0$, which will be used whenever its omission will not cause confusion.
(2)	use $\tilde{\mathbf{s}}$ to design an unbounded boundary that divides the space of statistics $\mathbf{s}$ into two separated and unbounded half-spaces.
(3)	let the one that does not contain the origin $\mathbf{0}$ as the rejection domain $\Gamma(\tilde{\mathbf{s}})$, with the corresponding boundary side named as the R-side. The other one is the acceptance domain.
(4)	tend to reject H_0 as $\mathbf{s}$ deviates from the R-side of boundary with a nonzero distance. The larger the distance is, the more seriously H_0 breaks.

Also, integration can be made by considering the complementarity of S-space BBTs and S-space model-based tests, via combining $\Gamma(\tilde{s})$ and the acceptance domains, obtained from not only the above complementary aspects, but also different sources, e.g. a bottom-up source from univariate tests on input data and a top-down source inversely transformed from the p values via a *meta-analysis* (Evangelou and Ioannidis 2013). Also, based on the resulted $\Gamma(\tilde{s})$, an easy computing expression $s_f = g(s, \Gamma(\tilde{s}))$ may be obtained to get an asymptotic distribution $p(s_f | \Gamma(\tilde{s}))$ for a fast estimation of the p value, see examples given after Equation (70).

S-space BBT for the multivariate zero mean

Testing H_0 by Equation (1) for the case-control studies can be formulated into testing whether a multivariate statistics $s = \left[s^{(1)}, \cdots, s^{(d)}\right]$ takes a point far away from the origin of the multidimensional space. One example is a two-sample test that examines the following null:

$$H_0 : \mathbf{s} = \mathbf{c}_1 - \mathbf{c}_1 = 0, \tag{66}$$

by the Hotelling T^2 statistics. The second example is the Wald testing statistics by Equation (7), and another example will be given in the next subsection.

In the existing studies, such a test is typically made via either the χ^2_k statistics or Hotelling's T^2 statistics. Also, Rao's score by Equation (8) is such a type of statistics. As addressed in the previous subsection, they are all featured by an integrated statistics $s_f \geq 0$ that monotonically increases as s deviates away from the origin and belong to the S-space model-based tests. Also, all these tests may be regarded as extensions of one typical univariate two-tail test (e.g. by t^2 test), that is, a univariate statistics s deviates away from the origin $s = 0$ via the value $|s|$.

The counterpart of a univariate two-tail test is a univariate one-tail test that examines how far s deviates from $(-\infty, 0]$, i.e. testing the statement $s \leq 0$. When either rejecting $s \leq 0$ or rejecting $s \geq 0$ happens, we reject $H_0 : s = 0$. Even when the statement $s \leq 0$ is not rejected, there are still chances that $H_0 : s = 0$ will be rejected.

Typical studies of univariate one-tail tests include the one-tailed t-test and one-tailed z-test. However, we are not clear what are their counterparts in multivariate tests. As addressed above, Hotelling's T^2 test can be regarded as a multivariate counterpart of a two-tailed test.

The S-space BBT given in Table 6 actually provides a road to extend univariate one-tail tests to multivariate ones. Observing univariate one-tail tests from the perspective of S-space BBT, we see that $\tilde{s} = I_{nf}(X_0||X_1)$ is actually a boundary point that results in:

$$\Gamma(\tilde{s}) = \{s : (s - \tilde{s})\text{sign}(\tilde{s}) > 0\} \tag{67}$$
$$= \begin{cases} [\tilde{s}, \infty), & if\ \tilde{s} > 0, \\ (-\infty, \tilde{s}]\,, & if\ \tilde{s} < 0. \end{cases} \text{ with Sign}[u] = \frac{u}{|u|}.$$

Given $\tilde{s}$ and thus $\Gamma(\tilde{s})$, any s obtained from the case-control samples under H_0 may cause a false alarm if s falls in $\Gamma(\tilde{s})$, which happens in a probability $p(s \in \Gamma(\tilde{s})|H_0)$, i.e. the p value by the inference $\tilde{s}$. If it is small enough, the statement $s \notin \Gamma(\tilde{s})$ will be rejected, which implies that $s = 0$ or H_0 by Equation (1) is rejected.

We further consider a statistics $\boldsymbol{s}$ in the multidimensional space from the perspective of S-space BBT given in Table 6 (2). We start by observing an orthant of the R^d space featured by $\text{sign}(\tilde{\boldsymbol{s}}) = \left[\text{sign}\left(\tilde{s}^{(1)}\right), \ldots, \text{sign}\left(\tilde{s}^{(d)}\right)\right]^T$ and consider one separating boundary, as illustrated in Figure 5A. Such a boundary is equivalent to the following decomposition:

$$\Gamma(\tilde{\boldsymbol{s}}) = \Gamma(\tilde{s}^{(1)}) \times \cdots \times \Gamma\left(\tilde{s}^{(d)}\right),$$
$$p(\boldsymbol{s} \in \Gamma(\tilde{\boldsymbol{s}})|H_0) = \prod_i p\left(\boldsymbol{s}^{(i)} \in \Gamma\left(\tilde{s}^{(i)}\right)|H_0\right), \tag{68}$$

where each $\Gamma(\tilde{s}^{(i)})$ is given by Equation (67) for computing $p\left(\mathbf{s}^{(i)} \in \Gamma\left(\tilde{s}^{(i)}\right)|H_0\right)$. This actually provides an example that extends a one-tail univariate hypothesis test to a vector-variate one.

In implementation, it is not easy to get the factorization of $p(\boldsymbol{s} \in \Gamma(\tilde{\boldsymbol{s}})|H_0)$ by Equation (68). Instead, we approximately consider to remove the second-order dependence by the following decorrelation:

$$\mathbf{s}_u = \begin{cases} U^T\mathbf{s}, & \text{Choice } (a), \\ \Lambda^{-0.5}U^T\mathbf{s}, & \text{Choice } (b), \end{cases} \quad s.t.\ U^TU = I, \tag{69}$$

where Λ_u is a diagonal matrix consisting of the nonzero eigenvalues of the following covariance matrix:

$$\Sigma_\pi = \frac{\sum_{\pi\in\Pi}\left(\mathbf{s}_{X^\pi_{1|0}} - \mu^\pi\right)\left(\mathbf{s}_{X^\pi_{1|0}} - \mu^\pi\right)^T}{\#\Pi},$$
$$\mu^\pi = \frac{\sum_{\pi\in\Pi}\mathbf{s}_{X^\pi_{1|0}}}{\#\Pi}.$$

and U is a $d \times m$ matrix with its columns consisting of the eigenvectors of Σ_π such that $\Lambda_u = U^T\Sigma_\pi U$.

Another issue is that only those major components in Equation (68) are useful while some components are not only useless but also disturbing, especially when we consider a limited size of samples. To do so, one may consider that the columns of the matrix U consist of the eigenvectors of Σ_π corresponding to the m-largest diagonal elements of Λ_u. Such an implementation of Equation (69) is typically called principal component analysis (PCA). How to decide an appropriate number of components is a model selection task

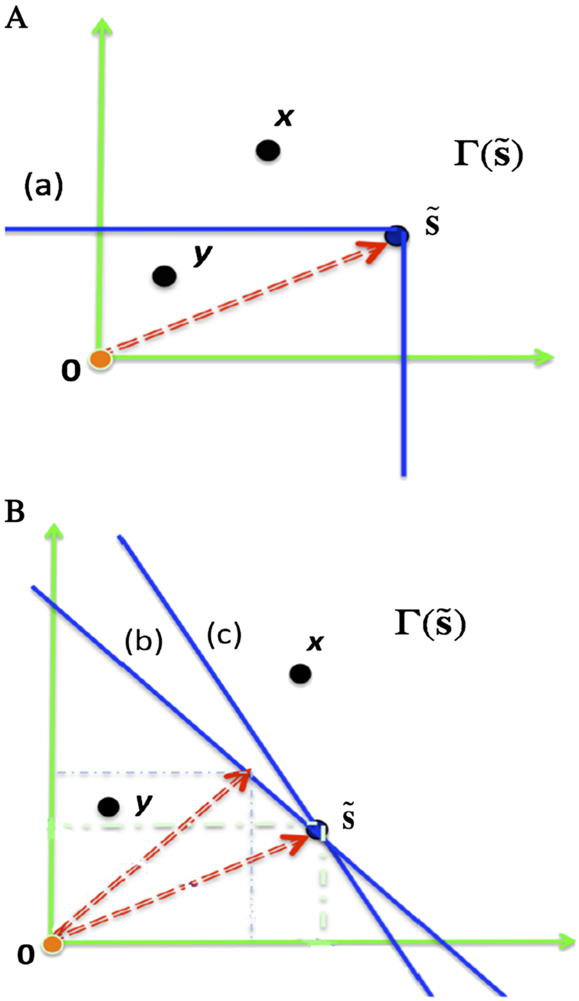

Figure 5 Rejection domain $\Gamma(\tilde{s})$, e.g. point x in the rejection domain, and y in the acceptance domain. **(A)** One separating boundary that consists of d lines with each emitting from $\tilde{s}$ to infinity in parallel to one axis of the orthant. **(B)** Choice (c) defines a hyperplane that passes $\tilde{s}$ and uses its vector direction as the normal direction. While choice (b) takes the normal direction by the primary diagonal direction of the orthant.

(Tu and Xu 2011, 2012; Xu 2011). Moreover, one novel direction for this task will be addressed later in thip paper between Equation (91) and Equation (99). Actually, Equation (69) only applies to remove the second-order dependence. One may further consider non-Gaussian factor analysis (NFA) and binary factor analysis (BFA) to remove dependencies among non-Gaussian components (Tu and Xu (2014); Xu (2003, 2009) and also Section 5 in Xu (2012b)).

Simply, we use the notation $\tilde{s} = I_{nf}(X_0||X_1)$ to denote a procedure to obtain such major components and then use this $\tilde{s}$ to get a separating boundary and its corresponding $\Gamma(\tilde{s})$. Illustrated in Figure 5 are three examples as follows:

$$\Gamma(\tilde{s}) = \begin{cases} \left\{ s : \left(s^{(i)} - \tilde{s}^{(i)}\right) \mathrm{sign}\left(\tilde{s}^{(i)}\right) > 0, \forall i \right\}, & (a), \\ \left\{ s : \left(s - \tilde{s}\right)^T \mathrm{sign}(\tilde{s}) > 0 \right\}, & (b), \\ \left\{ s : \left(s - \tilde{s}\right)^T \tilde{s} > 0 \right\}, & (c). \end{cases} \tag{70}$$

Choice (a) is illustrated in Figure 5A same as the one in Equation (68) with each $\Gamma(\tilde{s}^{(i)})$ given by Equation (67). As illustrated in Figure 5B, each of two other choices is a half space bounded by a plane and on the side away from the origin. Choice (b) is more suitable to the case after using Equation (69) in choice (b). Except for the degenerated cases that the normal direction of the hyperplane becomes in parallel to one of the coordinate axis, choice (b) and choice (c) will approximately describe a certain dependence across the components of $\boldsymbol{s}$.

After using Equation (69) to make the statistics $\boldsymbol{s}$ become an m-dimensional vector with the second-order dependence removed, we may observe that the scope of $\Gamma(\tilde{\boldsymbol{s}})$ becomes narrowed as m reduces. When $m = 1$, the scope of $\Gamma(\tilde{\boldsymbol{s}})$ is narrowed to a one-tail test along the axis of only one component.

In implementation, we obtain $p(\boldsymbol{s} \in \Gamma(\tilde{\boldsymbol{s}})|H_0)$ by Equation (64) via the permutation by Equation (65). Also, choice (b) and choice (c) may be understood from getting an integrated statistics as follows:

$$\begin{aligned} \boldsymbol{s_w} &= \boldsymbol{w}^T\boldsymbol{s}, \\ \boldsymbol{w} &= \text{sign}(\tilde{\boldsymbol{s}}) \; or \; \boldsymbol{w} = \tilde{\boldsymbol{s}}. \end{aligned} \tag{71}$$

Approximately, $\boldsymbol{s_w}$ comes from a normal distribution with the mean $\mu_{\mathbf{w}}$ and the variance $\boldsymbol{s_w}$, based on which we can make a one univariate test.

SPD test and SPD discriminative analysis

Proposed in (Xu 2013a), the SPD method firstly examines the delta $\delta(x, y)$ by pairing every case sample $x \in X_1$ and every control sample $y \in X_0$ and then summarises such deltas as follows:

$$D(X_1||X_0) = \frac{1}{\#X_0\#X_1}\sum_{x\in X_1}\sum_{y\in X_0}\delta(x,y). \tag{72}$$

Generally, $\delta(x, y)$ could be either symmetric or antisymmetric. One simple symmetric example is:

$$\begin{aligned} \delta(x,y) &= \frac{(x-y)^2}{\alpha_1\sigma_1^2+\alpha_0\sigma_0^2}, \\ D(X_{1||0}) &= 1+\frac{(c_1-c_0)^2}{\alpha_1\sigma_1^2+\alpha_0\sigma_0^2}-\frac{r_{xy}}{\alpha_1\sigma_1^2+\alpha_0\sigma_0^2}, \end{aligned} \tag{73}$$

where $c_\omega, \sigma_\omega^2, \alpha_\omega$ is the sample mean, variance, and proportion of the samples in X_ω, respectively, and r_{xy} is the mutual correlation between x and y.

The above example can be extended to the case that both $\boldsymbol{x}, \mathbf{y}$ are vectors with:

$$\delta(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T[\alpha_0\Sigma_0+\alpha_1\Sigma_1]^{-1}(\mathbf{x}-\mathbf{y}).$$

Also, we may consider an antisymmetric delta:

$$\delta(x,y) = \rho(x-y), \; d\rho(u)/du > 0, \tag{74}$$

where $\rho(u)$ is a monotonic function. One simplest example is $\rho(u) = u$ as follows:

$$\delta(x,y) = x-y, \tag{75}$$

which is equivalent to testing the difference of two sample means. To find the collective inclining structure, we classify $\delta(x,y)$ into three groups by $x > y, x = y, x < y$ and get the following decomposition:

$$\begin{aligned} D(X_{1||0}) &= D_{+}(X_{1||0}) - D_{-}(X_{1||0}), \\ D_{+}(X_{1||0}) &= \sum_{x>y}(x-y), \\ D_{-}(X_{1||0}) &= \sum_{y<x}(y-x) \end{aligned} \tag{76}$$

with $D\left(X_{1||0}\right) < 0$ indicating that there is a collective inclining dominance (i.e. the representations of cases are bigger than the ones of controls), $D(X_{1||0}) < 0$ indicating a reversed dominance, and $D\left(X_{1||0}\right) = 0$ indicating no dominance.

Recalling Equation (66), it follows from $\tilde{s} = D(X_{1||0}) = c_1 - c_0$ that $D(X_{1||0})$ is approximated from a normal distribution. Thus, the above collective inclining dominance can be tested by the one-tailed t-test and one-tailed z-test addressed in the previous subsections. We may get the mean $\mu\left(X^{\pi}_{1||0}\right)$ and the variance $\sigma^2\left(X^{\pi}_{1||0}\right)$ from $\left\{D(X^{\pi}_{1||0}, \pi \in \Pi\right\}$ and then approximately compute the p value by a univariate one-tail z-test.

When $\boldsymbol{x}, \mathbf{y}$ are vectors, we consider:

$$\mathbf{s} = \left[D^{(1)}(X_{1||0}), \cdots, D^{(d)}(X_{1||0})\right]^{T}, \tag{77}$$

with each $D^{(i)}(X_{1||0})$ by Equation (76). The task is detecting whether there is a collective inclining dominance, i.e. whether $\boldsymbol{s}$ deviates far away from the origin such that H_0 by Equation (1) breaks. The task can be handled by the S-space BBT in Table 6 as a multivariate extension of a one-tail univariate hypothesis test, following the method introduced from Equations (68) to (71) given previously.

Also, we may consider this multivariate SPD study from a perspective similar to the FDA by Equation (11). When $\boldsymbol{x}, \mathbf{y}$ are the d-dimensional vectors, we extend Equation (74) into:

$$\delta(\mathbf{x},\mathbf{y}) = \rho(\mathbf{x}-\mathbf{y})^{T}\mathbf{w}, \tag{78}$$

where $\rho(\mathbf{u}) = \left[\rho(u^{(1)}, \cdots, \rho(u^{(d)}\right]^{T}$ and $\rho(u)$ is the same as the one in Equation (74). That is, the difference $\boldsymbol{x}-\mathbf{y}$ is projected onto a most reasonable direction $\mathbf{w}$. In the simplest case $\rho(u) = u$, we get $\delta(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^{T}\mathbf{w}$ given in Equation (72) and thus leads to $\mathbf{s}_{\mathbf{w}} = \mathbf{w}^{T}\mathbf{s}$ in Equation (71) as follows:

$$\mathbf{s}_{\mathbf{w}} = \frac{1}{\#X_0\#X_1}\sum_{x\in X_1}\sum_{y\in X_0}(\mathbf{x}-\mathbf{y})^{T}\mathbf{w} = \mathbf{w}^{T}\mathbf{s}. \tag{79}$$

Without losing generality, we consider that the components of $\boldsymbol{s}$ are mutually independent, e.g. obtaining a second-order independence by Equation (69). Then, we seek how to choose an appropriate $\boldsymbol{w}$.

Under H_0, we expect that $\mathbf{s}^{\pi}_{\mathbf{w}} = D_{\mathbf{w}}\left(X^{\pi}_{1||0}\right), \pi \in \Pi$ varies around its mean that is typically zero according to Equation (75), that is, we expect that the following standard deviation of $\boldsymbol{s}^{\pi}_{\mathbf{w}}$ is minimised:

$$\begin{aligned} \sigma_{\pi}(\mathbf{w}) &= \sqrt{\mathbf{w}^{T}\Sigma_{\pi}\mathbf{w}}, \\ \Sigma_{\pi} &= E\left[\left(\mathbf{s}^{\pi}_{\mathbf{w}} - E\mathbf{s}^{\pi}_{\mathbf{w}}\right)\left(\mathbf{s}^{\pi}_{\mathbf{w}} - E\mathbf{s}^{\pi}_{\mathbf{w}}\right)^{T}\right]. \end{aligned} \tag{80}$$

Also, we expect that $\boldsymbol{s}_{\mathbf{w}}$ best preserves discriminative information underlying X_1, X_0, for which we maximise $|\mathbf{s_w}|$. We apply a bootstrapping method to enhance the reliability by maximising:

$$\rho_\gamma(\mathbf{w}) = \sum_{\omega \in \Omega} |\mathbf{w}^T \mathbf{s}^\phi|^\gamma, \ \gamma > 0, \tag{81}$$

which may tend to ∞ if it is unbounded. To avoid it, some bound will be imposed on $\boldsymbol{w}$.

For $\gamma = 1$, we usually consider:

$$\max_{\mathbf{w}} \rho_{\gamma=1}(\mathbf{w}), \ s.t. \ w^{(i)} \in \left[a^{(i)}, b^{(i)}\right], \forall i. \tag{82}$$

by which the solution of $\boldsymbol{w} = [\, w^{(1)}, \ldots, w^{(d)}]^T$ is reached at one vertex, i.e. $w^{(i)}$ takes either $a^{(i)}$ or $b^{(i)}$. Particularly, when Ω consists of only one pair X_1, X_0, the above maximisation leads to choice (b) in Equation (70) if we let $-a^{(i)} = b^{(i)} = 1$ and to choice (c) if we let $-a^{(i)} = b^{(i)} = |D^{(i)}(X_{1||0})|$.

For $\gamma = 2$, we consider:

$$\max_{\mathbf{w}, \ s.t. \ \|\mathbf{w}\|^2 = 1,} \rho_{\gamma=2}(\mathbf{w}) = \mathbf{w}^T \Sigma^\phi \mathbf{w}, \tag{83}$$

with its solution given by the eigenvector that corresponds to the largest eigenvalue of $\Sigma^\phi = \sum_{\omega \in \Omega} \mathbf{s}^\phi \mathbf{s}^{\phi \ T}$.

Integrating Equations (80) and (81), we consider to maximise $\rho_\gamma(\mathbf{w})$ with $\sigma_\pi^\gamma(\mathbf{w})$ minimised simultaneously or subject to a constraint $\sigma_\pi^\gamma(\mathbf{w}) \leq$ constant.

Alternatively, we may consider:

$$\max_{\mathbf{w}} J(\mathbf{w}), \ J(\mathbf{w}) = \frac{\rho_\gamma(\mathbf{w})}{\sigma_\pi^\gamma(\mathbf{w})}, \tag{84}$$

which shares a spirit similar to the FDA by Equation (11). At the typical case $\gamma = 2$, it becomes

$$\max_{\mathbf{w}} J(\mathbf{w}), \ J(\mathbf{w}) = \frac{\mathbf{w}^T \Sigma^\phi \mathbf{w}}{\mathbf{w}^T \Lambda \mathbf{w}}, \tag{85}$$

with its solution given by the eigenvector that corresponds to the largest eigenvalue of $\Sigma_\pi^{-0.5} \Sigma^\phi \Sigma_\pi^{-0.5}$.

Furthermore, we proceed to consider that each $D^{(i)}(X_{1||0})$ in Equation (79) is not a simple difference by Equation (76) but the following 1×2 row vector:

$$D^{(i)}(X_{1||0}) = \left[D_+^{(i)}(X_{1||0}), -D_-^{(i)}(X_{1||0})\right]. \tag{86}$$

Also, we may extend $\boldsymbol{x} - \mathbf{y}$ with each element $x^{(i)} - y^{(i)}$ becoming a row vector $\left[x^{(i)}, -y^{(i)}\right]$. Accordingly, we get:

$$\begin{aligned} \mathbf{x} - \mathbf{y} &= \Delta_{x-y} \mathbf{v}, \\ \delta(x, y) &= \mathbf{w}^T \Delta_{x-y} \mathbf{v}, \end{aligned} \tag{87}$$

where $\mathbf{v} = \left[v^{(1)}, v^{(2)}\right]^T$ and Δ_{x-y} is a $d \times 2$ matrix with the i-th row being $[\, x^{(i)}, -y^{(i)}]$. It follows from Equation (72) that the above Equation (87) leads $D^{(i)}(X_{1||0})$ to:

$$\begin{aligned} D^{(i)}(X_{1||0}) &= \left[c_1^{(i)}, -c_0^{(i)}\right], \\ \mathbf{s} &= D_M(X_{1||0}) \mathbf{v}, \end{aligned} \tag{88}$$

where $D_M(X_{1||0})$ is a $d \times 2$ matrix with $D^{(i)}(X_{1||0})$ as its i-th column. Accordingly, the inner product by Equation (79) becomes:

$$\mathbf{s}_{\mathbf{w}} = \mathbf{w}^T D_M(X_{1||0})\mathbf{v}. \tag{89}$$

Given $\boldsymbol{v}$ as fixed, the study from Equations (79) and (84) applies directly for us to get $\mathbf{w}$.

Given $\boldsymbol{w}$ as fixed, $\mathbf{w}^T D_M\left(X^{\pi}_{1||0}\right) = D^T_c\left(X_{1||0}\right)$ becomes a two-dimensional row vector and, it follows from Equation (89) that we have $\boldsymbol{s}_{\mathbf{w}} = \mathbf{v}^T D^T_c\left(X_{1||0}\right)$ in the same form as Equation (79). With $\boldsymbol{v}$ in the place of $\boldsymbol{w}$ and $D_c\left(X_{1||0}\right)$ in the place of $\boldsymbol{s}$, similarly, the study from Equations (79) and (84) applies directly for us to get $\boldsymbol{v}$. Generally, we iteratively update $\mathbf{v}$ with a fixed $\boldsymbol{w}$ and update $\mathbf{w}$ with a fixed $\boldsymbol{v}$, for a number of circles getting converged. Still, whether such an alternative iterating procedure can converge is an open issue that demands further investigation.

The *p* values and testing complexity control

Recalling Equation (64) and Table 6, based on a given sample pair $X_{1||0} = X_0\|X_1$, we get a statistics vector $\tilde{\boldsymbol{s}}_{X_{1||0}} = I_{nf}(X_0||X_1)$ and a rejection domain $\Gamma = \Gamma\left(\tilde{\boldsymbol{s}}_{X_{1||0}}\right)$ by the inferring method I_{nf}. Then, we compute the following false alarm probability:

$$p_{X_{1||0}} = p(\mathbf{s} \in \Gamma |\, I_{nf}, X_{1||0}, H_0) \tag{90}$$

as the p value. This concept is the same as the one used in the conventional literature where $X_{1||0}$ and I_{nf} are usually implied but not spelled out.

Being different from those studies considering a univariate statistics, the p value by a multidimensional statistics vector $\boldsymbol{s}$ highly depends on the dimension m of this vector or the complexity of the testing space. Given a limited sample size, the p value by Equation (90) will reduce as the value of m increases, causing a phenomenon similar to the overfitting problem in the studies of machine learning and statistical modelling. In other words, we encounter a 'dimension curse' in hypothesis testing too. Therefore, we need to appropriately control the complexity of testing space, i.e. selecting one appropriate m.

Given a criterion $J(m)$, the problem of selecting a best subset is a typical problem of feature selection. Generally, it involves an exhaustive evaluation of all the combinations of m features (i.e. m components of $\boldsymbol{s}$) and all the possible values of m, which is a NP hard problem. Usually, the branch and bound policy (Narendra and Fukunaga 1977; Somol et al. 2004) and the best first strategy are used to save computing cost (Xu et al. 1988). In this paper, we only consider one simple selection strategy that evaluates the components of $\boldsymbol{s}$ incrementally one by one.

To facilitate it, we perform Equation (69) to make the components of $\boldsymbol{s}$ become decorrelated and start to pick one component that corresponds to the smallest value of a given criterion $J(m)$. Then, we successively add in one component such that $J(m)$ gets a bigger drop further and so on and so forth until no further reduction is caused. Finally, the selected components form the resulted feature set with a size m^*.

For this purpose, using the p value by Equation (90) as $J(m)$ does not work well because of its tendency of reducing as m increases, resulting in one m^* that is usually much bigger than the appropriate one. Instead, we consider another false alarm probability as follows:

$$p(\mathbf{s} \in \Gamma |\, I_{nf}, H_0) = \int p\left(X^{\pi}_{1||0}\right) p\left(\mathbf{s} \in \Gamma |\, I_{nf}, X^{\pi}_{1||0}, H_0\right) dX^{\pi}_{1||0}, \tag{91}$$

which is obtained on all the possible sets of $X^{\pi}_{1||0}$ that come under H_0 instead of merely on a given pair $X_{1||0}$.

Though this probability is useless to judge whether $X_{1||0}$ contains enough information to reject H_0, it reflects how the complexity of testing space affects a background portion of the false alarm probability. Actually, it reflects an inverse of the effective volume of the support that the statistics $\mathbf{s}$ locates. As m increases, the volume increases exponentially, and thus, $p(\mathbf{s} \in \Gamma | I_{nf}, H_0)$ will reduce negative-exponentially. Such an exponentially decreasing tendency is also contained in $p(\mathbf{s} \in \Gamma | I_{nf}, X_{1||0}, H_0)$ for the same reason, which affects the accuracy of the estimated p value.

To reduce this background disturbance, we consider Equations (90) and (91) jointly by the following *a posteriori* version of the p value:

$$pp_{X_{1||0}} = p(\neg H_0 | I_{nf}, X_{1||0}, H_0) = \frac{\int_{p_{X^{\pi}_{1|0}} \le p_{X_{1|0}}} p\left(X^{\pi}_{1||0}\right) p\left(\mathbf{s} \in \Gamma | I_{nf}, X^{\pi}_{1||0}, H_0\right) dX^{\pi}_{1||0}}{p(\mathbf{s} \in \Gamma | I_{nf}, H_0)},$$

where and hereafter $\neg H_0$ denotes rejecting H_0. The denominator aims at cancelling out the disturbing portion in the numerator, such that $pp_{X_{1||0}}$ provides not only a better estimation of false alarm probability of rejecting H_0 but also a better criterion $J(m)$ for selecting a best subset of the components of $\mathbf{s}$ and thus inferring one appropriate m^*.

Instead of directly handling the above integral, we get a large set Π of sample pairs X^{π}_1, X^{π}_0, with each pair X^{π}_1, X^{π}_0 resulted from a permutation of X_0 and X_1. Using every pair X^{π}_1, X^{π}_0 to infer $I_{nf}\left(X^{\pi}_0 || X^{\pi}_1\right) = \tilde{\mathbf{s}}_{X^{\pi}_{1|0}}$, we get a set of p values as follows:

$$P_{\Pi} = \left\{p_{X^{\pi}_{1|0}}, \pi \in \Pi\right\}, \text{ with } p_{X^{\pi}_{1|0}} = p\left(\mathbf{s} \in \Gamma | I_{nf}, X^{\pi}_{1||0}, H_0\right) \tag{92}$$

based on which we compute:

$$\begin{aligned} pp_{X_{1||0}} &= \frac{\sum_{\pi \in \Pi_{\Gamma}} p_{X^{\pi}_{1|0}}}{\sum_{\pi \in \Pi} p_{X^{\pi}_{1|0}}} = pp^{o}_{X_{1||0}} rp_{X_{1||0}}, \\ pp^{o}_{X_{1||0}} &= \frac{n_{\Gamma}}{n_{\Pi}}, \; rp_{X_{1||0}} = \frac{\mu_{\Gamma}}{\mu}, \\ n_{\Gamma} &= \#\Pi_{\Gamma}, \; n_{\Pi} = \#\Pi, \\ \Pi_{\Gamma} &= \left\{\pi : p_{X^{\pi}_{1|0}} \le p_{X_{1|0}}, \forall \pi \in \Pi\right\}, \\ \mu_{\Gamma} &= \frac{\sum_{\pi \in \Pi_{\Gamma}} p_{X^{\pi}_{1|0}}}{n_{\Gamma}}, \; \mu = \frac{\sum_{\pi \in \Pi} p_{X^{\pi}_{1|0}}}{n_{\Pi}}. \end{aligned} \tag{93}$$

We observe that the pp value has two factors. One is $pp^{o}_{X_{1||0}}$ that describes the proportion of the pairs of X^{π}_1, X^{π}_0 with the corresponding $p_{X^{\pi}_{1|0}} \le p_{X_{1|0}}$, that is, on each of these pairs we should also reject H_0 if we reject H_0 on $X_{1||0}$. In other words, $pp^{o}_{X_{1||0}}$ reflects the information of relative difference contained in P_{Π}. The other factor $rp_{X_{1||0}}$ is the ratio of the average false alarm probability per pair over the disturbing background per pair, reflecting the strength of discriminative information contained in P_{Π}.

In implementation, we may use $rp_{X_{1||0}}$ to make an initial screening. When $rp_{X_{1||0}} > 1$, inference is nonsense and no further computing should be made. Generally, $rp_{X_{1||0}}$ will be much smaller than 1, and thus, $pp_{X_{1||0}}$ will be much smaller, while $pp^{o}_{X_{1||0}}$ provides a worst case upper bound of $pp_{X_{1||0}}$.

We should observe $pp_{X_{1||0}}$, $pp^{o}_{X_{1||0}}$, and $rp_{X_{1||0}}$ at not only one same value of m but also an appropriate m^*. In addition to using $pp_{X_{1||0}}$ by Equation (93) as $J(m)$ for making an incremental selection, we may also consider $pp^{o}_{X_{1||0}}$ or $rp_{X_{1||0}}$ as $J(m)$, resulting in m^*_o or m^*_{rp}. Also, it follows from some mathematical derivation that we have $m^* \geq m^*_{rp} \geq m^*_o$ with m^*_o being a most conservative lower bound. We will be more confident when all these values are identical or not different too much. Moreover, further insights can be obtained from the following considerations.

On one side, we desire that the exponentially decreasing tendency contained in $p(\mathbf{s} \in \Gamma|\, I_{nf}, X_{1||0}, H_0)$ is removed via the normalisation by $p(\mathbf{s} \in \Gamma|\, I_{nf}, H_0)$ such that $pp_{X_{1||0}}$ in Equation (93) will no longer have such a decreasing tendency. With $p_{X^{\pi}_{1|0}} = p(\mathbf{s} \in \Gamma|\, I_{nf}, X^{\pi}_{1||0}, H_0)$ in Equation (92) replaced by $pp_{X_{1||0}}$, $pp^{o}_{X_{1||0}}$, and $rp_{X_{1||0}}$, we may turn P_{Π} into its counterparts P_{pp}, P_{pp^o}, and P_{rp}. We compute not only the varying curve for each of $pp_{X_{1||0}}$, $pp^{o}_{X_{1||0}}$, and $rp_{X_{1||0}}$ as m increases, but also the varying curve of the mean of the elements in each of P_{pp}, P_{pp^o}, and P_{rp} as m increases. Then, we compare each curve with its corresponding mean curve and desire that the mean curve is as flat as possible or at least flat around m^*.

On the other side, desiring a flat mean curve is not a sole principle. W also desire that the discriminative information should be kept in each of $pp_{X_{1||0}}$, $pp^{o}_{X_{1||0}}$, and $rp_{X_{1||0}}$ as much as possible. Observing the factorization $pp_{X_{1||0}} = pp^{o}_{X_{1||0}} rp_{X_{1||0}}$ in Equation (93), the strength of discriminative information is contained in $rp_{X_{1||0}}$ with an exponentially decreasing tendency that is supposed to be mutually cancelled out by the denominator and the numerator but perhaps not completely, while the discriminative information of relative difference is contained in $pp^{o}_{X_{1||0}}$ and kept unchanged as long as every inequality between $p_{X^{\pi}_{1|0}}$ and $p_{X_{1|0}}$ remains unchanged.

Bi-test, twin *p* values, and P-space BBT

Putting the above two sides together, we observe that a S-space multivariate test is actually a bi-test that tests H_0 together with the following hypothesis:

$$I_0 \ : \ \text{the inference is not reliable.} \tag{94}$$

We examine a decision that both H_0 and I_0 are rejected, featured with two p values.

As addressed after Equation (91), the multivariate statistics $\boldsymbol{s}$ inferred by I_{nf} suffers a systematic bias that will make I_{nf} unreliable. This unreliability varies with the dimension m that takes an important role in I_{nf}. Though corrected by the denominator in Equation (93), there are still some residuals that will not be completely cancelled out, the effect of which still varies with m and reduces the reliability of I_{nf}. The test I_0 is formulated for this reliability via controlling an appropriate m^* and a level of false alarm probability of rejecting I_0.

One should notice the difference between testing H_0 and testing I_0. Testing H_0 examines only the input, while testing I_0 examines both the input and the performance of testing H_0. The inference I_{nf} gets $X_{1||0}$ as the input and the outcomes $p_{X_{1||0}}$,$pp_{X_{1||0}}$, $pp^{o}_{X_{1||0}}$, and $rp_{X_{1||0}}$. Using $o_{X_{1||0}}$ to denote anyone of these indices, regarding I_{nf} as reliable on $X_{1||0}$ actually implies that it should also be regarded as reliable on any pair X^{π}_1, X^{π}_0 with the corresponding $o_{X^{\pi}_{1||0}}$ being smaller than $o_{X_{1||0}}$. Thus, the false alarm probability of rejecting I_0 is computed by $p\left(o_{X^{\pi}_{1||0}} \leq o_{X_{1||0}} | \neg H_0, H_0\right)$.

Interestingly, some mathematical derivation shows that letting $o_{X_{1||0}}$ to be anyone of $p_{X_{1||0}}, pp_{X_{1||0}}, pp^o_{X_{1||0}}$, and $rp_{X_{1||0}}$ will always result in the same false alarm probability as follows:

$$p(\neg I_0|\neg H_0, H_0) = p\left(p_{X^{\pi}_{1|0}} \le p_{X_{1|0}}|H_0\right) = pp^o_{X_{1||0}}, \tag{95}$$

where and hereafter $\neg I_0$ denotes rejecting I_0. Reflecting the discriminative information of relative difference, this p value of rejecting I_0 will be not affected as long as the exponentially decreasing tendency will not change every inequality between $p_{X^{\pi}_{1|0}}$ and $p_{X_{1|0}}$.

As summarised in Table 7, a multivariate test is actually a bi-test that tests not only the classic null but also a null about the 'dimension curse'. The rejection of H_0 is controlled by a given level α. If $pp_{X_{1||0}} \ge \alpha$, H_0 will not be rejected, and thus, there is no need to test I_0. Accordingly, Equation (93) for the p value of rejecting I_0 is also modified in Table 7. The bi-test is implemented with or without using stochastic simulation. Table 7 (2) outlines those previously addressed points for implementation via stochastic simulation, while Table 7 (3) outlines an alternative implementation that does not need stochastic simulation.

This alternative comes from considering Γ in the choice (a) of Equation (70) by which we have:

$$p\left(\mathbf{s} \in \Gamma|\, I_{nf}, X^{\pi}_{1||0}, H_0\right) = \prod_{i \le m^*} p_i \prod_{i > m^*} \delta_m,$$

Table 7 Multivariate Bi-test and Implementations

Type	Description
	Test bi-hypotheses and twin p-values
test H_0	whether the case-control populations are different, by an inference I_{nf} in the space of multivariate statistics $\mathbf{s}$ based on samples from the two populations. H_0 is rejected if $pp_{X_{1||0}} \le \alpha$, where the false alarm probability $pp_{X_{1||0}} = pp^o_{X_{1||0}} rp_{X_{1||0}}$ is given by Equation (93) and α is a prespecified level.
test I_0	whether the dimension m of $\mathbf{s}$ is appropriate such that I_{nf} is reliable, with the p-value given by $p(\neg I_0|\neg H_0, H_0) = p(pp_{X^{\pi}_{1|0}} \le \alpha | pp_{X_{1||0}} < \alpha, H_0),$ which is not smaller than $pp^o_{X_{1||0}}$ that reflects the relative discriminative information among $pp_{X_{1||0}}$ while ignoring $rp_{X_{1||0}}$ that reflects the strength of discriminative information.
	Bi-text Implementations
Stochastic way	(a) Make the components of $\mathbf{s}$ decorrelated by Equation (69). (b) Get $p(\mathbf{s} \in \Gamma|\, I_{nf}, X^{\pi}_{1||0}, H_0) = p(\mathbf{s} \in \Gamma(\tilde{\mathbf{s}})|H_0)$ by Equation (68) with $\Gamma(\tilde{\mathbf{s}})$ taking one of three choices in Equation (70), and then getting P_{Π} by Equation (92). (c) Get $pp_{X_{1||0}}, pp^o_{X_{1||0}}, rp_{X_{1||0}}$ by Equation (93) and then getting $p(\neg I_0|\neg H_0, H_0)$ as above. (d) Using $pp^o_{X_{1||0}}$ or $p(\neg I_0|\neg H_0, H_0)$ as $J(m)$ to infer an appropriate m^*_o and select the m^*_o best components of $\mathbf{s}$.
Nonstochastic way	(a) Make the components of $\mathbf{s}$ decorrelated by Equation (69). (b) Get $\{p_i\}$ with each p-value p_i obatined by an univariate test. (c) Get $pp^o_{X_{1||0}}$ by Equation (99) and $rp_{X_{1||0}}$ by Equation (97) with $p_{X_{1|0}} = \prod_i p_i$, as well as getting $p(\neg I_0|\neg H_0, H_0)$ as above. (d) The same as the above (2)(d).

$$p_i = p\left(\mathbf{s}^{(i)} \in \Gamma\left(\tilde{s}^{(i)}\right) |H_0\right), \tag{96}$$

where the extra components of s will contribute a constant factor $\prod_{i>m^*} \delta_i$ that will be cancelled out via the denominator and the numerator in Equation (93).

In such a case, we may get $rp_{X_{1||0}} = \mu_\Gamma/\mu$ without stochastic simulation. First, we have $\mu = \prod_i \mu^{(i)}$. Each $\tilde{s}^{(i)}$ under H_0 is a random variable with a zero mean, and its corresponding false alarm probability p_i is uniformly distributed over $[0, 0.5]$. Thus, we get $\mu^{(i)} = 1/4$. Second, we also get $\mu_\Gamma \le p_{X_{1|0}}$ by letting $p(\mathbf{s} \in \Gamma|\, I_{nf}, X^\pi_{1||0}, H_0) \le p_{X_{1|0}}$ for each $\pi \in \Pi_\Gamma$ to be approximated by its upper bound $p_{X_{1|0}}$. Putting the two together, we have:

$$rp_{X_{1||0}} = \frac{\mu_\Gamma}{\mu} \le \frac{p_{X_{1|0}}}{\mu},\ \mu = \begin{cases} \frac{1}{4^m}, \text{ one tail,} \\ \frac{1}{2^m}, \text{ two tails.} \end{cases} \tag{97}$$

Next, $pp^o_{X_{1||0}}$ is also considered without stochastic simulation. From Equation (95), we have $p\left(\neg I_0|\neg H_0, H_0\right) = pp^o_{X_{1||0}} = p\left(\prod_i p_i^\pi \le \prod_i p_i|H_0\right) = p\left(\prod_i \left(p_i^\pi\right)^2 \le \prod_i p_i^2|H_0\right)$, which leads us to the well-known Fisher combination (Fisher 1948) that makes a test on the false alarm probabilities $\{p_i\}$ by the following combination:

$$\begin{aligned} p_F &= p\left(\prod_i \left(p_i^\pi\right)^2 < \prod_i p_i^2 \mid H_0\right) \\ &= p\left(\chi^2_{2m} > -2\sum_i \ln p_i\right),\ \chi^2_{2m} = -2\sum_i \ln p_i{}^\pi. \end{aligned} \tag{98}$$

This link provides new insights from two perspectives. On one perspective, we may adopt the Fisher combination approach to estimate $pp^o_{X_{1||0}}$ as follows:

$$pp^o_{X_{1||0}} = p\left(\chi^2_{2m} > -2\sum_i \ln p_i\right). \tag{99}$$

Together with Equation (97), we get $pp_{X_{1||0}} = pp^o_{X_{1||0}} rp_{X_{1||0}}$ for testing both H_0 and I_0 without stochastic simulation via permutation.

On the other perspective, we observe that the traditional p value p_F of the Fisher combination is actually the false alarm probability by Equation (95), only reflecting the discriminative information of relative difference between $\prod_i p_i^\pi$ and $\prod_i p_i$ but ignoring the strength of discriminative information contained in $\prod_i p_i$. In other words, the Fisher combination just provides a half story for combining $\{p_i\}$, and we can use the formulation $pp_{X_{1||0}} = pp^o_{X_{1||0}} rp_{X_{1||0}}$ to complete the whole story, using $pp^o_{X_{1||0}}$ by Equation (99) and $rp_{X_{1||0}}$ by Equation (97) with $p_{X_{1|0}} = \prod_i p_i$.

The last but not least, one should notice that the p value of testing H_0 measures the chances in the S-space (i.e. the space of multivariate statistics), and the p value of testing I_0 measures an event in the P-space (i.e. the space of false alarm probabilities). In other words, testing H_0 involves a S-space BBT while testing I_0 involves a P-space BBT.

Discussions

Gene expression analyses

Gene expression analyses take important roles in bioinformatics and computational genetics. Expression profiles are featured by data matrix with its row indicating expressions of different samples $t = 1, \cdots, N$ while its column consisting of expressions $i = 1, \cdots, m$ from different genes, miRNAs, and lncRNAs.

In recent years, developments of data acquisition techniques lead us to consider expression profiles in a cubic or even a high-dimension array. As illustrated in Figure 1, one additional dimension $j = 1, \cdots, d$ is added for examining expressions under different conditions (Ji et al. 2009; Persson et al. 2011) and across different time points (Bar-Joseph et al. 2012). For examples, current cancer studies consider each basic unit (i.e. a gene, a miRNA, a lncRNA) in paired expressions of normal and tumour tissue from the same individual, that is, each individual is featured at least by a $2 \times d$ matrix X_t. Generally, each example X_t is a $m \times d$ matrix. In Table 7, we suggest a list of topics for such matrix-variate-based applications.

Typically, the number d of rows (i.e. gene, miRNA, and lnclRNA) is huge, while the sample size n is small. It is difficult and also unreliable to consider the entire $m \times d$ matrix as a sample X_t. Instead, we pick k- tuple out of m rows to form a $m \times k$ matrix as a sample X_t. Without losing generality, we focus on that each sample X_t is a $2 \times k$ matrix from paired expressions of normal tissue and tumour tissue.

In the existing studies, there are two types of efforts for dealing with such format of samples. The first one reduces each sample $X_t = \left[x_t^{(i,j)}\right], i = 1, 2; j = 1, \cdots, k$ into a $1 \times k$ matrix $x_t = \left[x_t^{(1)}, \cdots, x_t^{(k)}\right]$ for multivariate hypothesis test. A typical reduction is given by:

$$x_t^{(j)} = \ln x_t^{(1,j)} - \ln x_t^{(2,j)}. \tag{100}$$

The second type of efforts is a paired difference test, e.g. a paired t-test when $k = 1$ and paired Hotelling's square test when $k \geq 2$. In Table 8, comparative empirical IHT studies are suggested on the samples of X_t in a $2 \times k$ matrix versus in a $1 \times k$ vector.

Exome sequencing analyses

The case-control study is also a major problem in a genome-wide association study (GWAS) or exome-sequencing analysis (DePristo et al. 2011; Purcell et al. 2007). Typically, a digit score (i.e. $0, 1, 2$) is assigned to a Single Nucleotide Polymorphism (SNP) allele per site and per individual. In such a representation, each sample is univariate when each site is considered one by one. One variate two-sample test takes a fundamental role for detecting a single SNP in the GWAS, e.g. the PLINK provides one widely used tool box (Purcell et al. 2007).

Moreover, each sample can be a vector when multiple sites are considered jointly. Recently, there have been ever-increasing efforts on finding multiple SNVs jointly (DePristo et al. 2011; Derkach et al. 2013; Evangelou and Ioannidis 2013; Lin et al. 2014; Liu et al. 2014; Pan et al. 2014). Also, we may test whether there is a collective inclining dominance of the representations of case samples over the ones of control samples, or vice versa, with help of the method proposed from Equations (79) and (84), as well as the extension introduced around Equations (87) and (89).

Alternatively, we may also consider a SNP allele per site and per individual with $\delta(x, y)$ in Equation (75) replaced by one 3×3 matrix $\Delta(x, y) = \left[\delta_{x-y}^{(i,j)}\right]$ with:

$$\delta_{x-y}^{(i,j)} = \begin{cases} \text{sign}(x - y), & i = x, y = j, \\ 0, & \text{otherwise.} \end{cases} \tag{101}$$

Table 8 ***Several IHT Applications***

IHT types	Applications
Model based and Mix-modelled	(a) Starting at the case that X_t is degenerated into an 1×2 matrix, we conduct the Hotelling test by Equation (2) and its extension KL_{sum} in Equation (31), in comparison with both univariate t-test and a paired t-test. (b) For the general case with $k \geq 2$, we conduct a matrix-variate test by Equation (28), as well as by the matrix-variate counterparts of $KL_{1,0}$, KL_{sum}, and KL_{sum*}, in comparison with not only the Hotelling's T-square test on the k dimensional vector x_t obtained from Equation (100) but also the paired Hotelling's T-square test on $2 \times k$ matrix-variate samples of X_t. (c) Considering each sample X_t in a $2 \times k$ matrix, we investigate the bi-linear discriminant analysis by Equations 18, 33, and 34, in comparison with the classic FDA by Equation (11) on the k dimensional vector x_t obtained from Equation (100). (d) Investigate the generalised bi-linear discriminant analysis by Equations 40, 41, and 34. For simplicity, we get $\mathbf{v}_i, i = 1, \cdots, d$ by Equation (43) and then solve $\mathbf{w}$ by Equation (34). When k becomes too big, we further regularise the learning of $\mathbf{v}_i$ by minimising $J_y = \frac{\alpha_0 \sigma_0^{y\,2} + \alpha_1 \sigma_1^{y\,2}}{(c_0^y - c_1^y)^2} + \sum_{i=1}^m \gamma_i \sum_{j=1}^d \vert u_i^{(j)} \vert^q$, with $q = 2$ for Tikhonov, $q = 1$ for sparse learning.
Boundary based and Mix-modelled	(a) Consider a logistic regression by Equation (3) with $\mathbf{w}$ in one of the ways given in Table 4, we test Equation (5) by the Rao's score Equation (8), and get ε_C by Equation (44), and ε_B by the p-value with one of choices in Table 2. (b) Extend all the above studies on Equation (3) with $y_t = \mathbf{w}^T \mathbf{x}_t$ replaced by the bi-linear form Equation (18). (c) Make a survival analysis via the Cox regression by Equation (13) in comparison with its bi-linear extension by Equations (18) or (40). Again, IHT is made by ε_D, ε_C, and ε_B in a way similar to the above.
BYY harmony	(a) Use either *Algorithm 9* in Ref. (Xu, 2015) to get $\boldsymbol{\alpha}^{(i)}, \mathbf{c}^{(i)}, \Sigma^{(i)}, i = 0, 1$ or *Algorithm 1* to get $\boldsymbol{\alpha}^{(i)}, C^{(i)}, \Sigma^{(i)}, \Omega^{(i)}, i = 0, 1$ for model based IHT. (b) Perform the procedure given in Table 5 for training, testing and validating in a small size of samples.

It follows from Equation (72) that we get $D(X_{1\vert\vert 0})$ to be also a 3×3 matrix as a collective measure, which may be further examined to test whether two populations differ significantly. We may visualise the matrix by plotting them in two 2D histograms and observe their configurations.

Conclusions

Statistical analyses for case-control studies have been addressed rather comprehensively. First, a Kullback-Leibler divergence-based formulation is suggested to develop testing statistics and discriminative criterion for the case-control studies. Based on this formulation, typical existing methods are revisited, and their matrix-variate counterparts are developed. Second, a bi-linear matrix form is proposed to obtain the matrix-variate counterparts from existing multivariate statistical analyses, such as discriminative analysis, logistic regression, Cox model, and linear mixed model. Third, the necessity and feasibility of integrative hypothesis tests (IHT) are addressed from the complementarity of BMTs and BBTs in the D-space, together with empirical illustration. Moreover, four basic components of IHT are elaborated, and four IHT types are summarised according to how the components are integrated. Then, in the space of multiple statistics (shortly S-space), the S-space BBT is proposed to perform BBT based on an unbounded boundary, with the help of information-preserved decoupling. Moreover, a S-space BBT-based extension of univariate one-tail z-test is developed to test the null of multivariate zero mean and then applied to a multivariate SPD test for detecting a collective inclining dominance for the case-control studies. Also, a SPD discriminative analysis is proposed with this multivariate SPD test improved and extended to matrix-variate ones. Furthermore, a multivariate bi-test is proposed to test not only the classic null but also a null about inference reliability due to the complexity of testing space, including a new insight on and a further development of the Fisher combination. Finally, possible applications have been suggested for expression-profile-based biomarker finding and exome-sequencing-based joint SNV detection.

Competing interests
The author declares that he has no competing interests.

Acknowledgements
This work was supported by a CUHK Direct grant project 4055025 and by the Zhi-Yuan chair professorship by Shanghai Jiao Tong University.

Received: 26 September 2014 Accepted: 10 February 2015
Published online: 07 May 2015

References

Bar-Joseph Z, Gitter A, Simon I (2012) Studying and modelling dynamic biological processes using time-series gene expression data. Nat Rev Genet 13(8):552–564

Barnett JA (2008) Computational methods for a mathematical theory of evidence. In: Yager L, Liu L (eds). Classic Works of the Dempster-Shafer Theory of Belief Functions. Studies in Fuzziness and Soft Computing. Springer, Berlin Heidelberg. pp 197–216

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297

Cox DR, Oakes D (1984) Analysis of survival data. CRC Press, Chapman & Hall, Boca Raton, Florida

Cover TM (1965) Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. Electronic Computers, IEEE Transactions on 14(3):326–334

Demidenko E (2013) Mixed models: theory and applications with R. Probability and Statistics. John Wiley & Sons, Hoboken, New Jersey

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43(5):491–498

Derkach A, Lawless JF, Sun L (2013) Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. Genet Epidemiol 37(1):110–121

Dutilleul P (1999) The mle algorithm for the matrix normal distribution. J Stat Comput Simul 64(2):105–123

Engle RF (1984) Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. Handb Econometrics 2:775–826

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

Fisher RA (1948) Questions and answers# 14. Am Stat 2(5):30–31

Gibson G (2012) Rare and common variants: twenty arguments. Nat Rev Genet 13(2):135–145

Hosmer Jr DW, Lemeshow S, Sturdivant RX (2013) Applied logistic regression. John Wiley & Sons, Hoboken, New Jersey

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

Ji J, Shi J, Budhu A, Yu Z, Forgues M, Roessler S, Ambs S, Chen Y, Meltzer PS, Croce CM, Qin L-X, Man K, Lo C-M, Lee J, Ng IOL, Fan J, Tang Z-Y, Sun H-C, Wang XW (2009) Microrna expression, survival, and response to interferon in liver cancer. New Engl J Med 361(15):1437–1447

Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER (2013) The next-generation sequencing revolution and its impact on genomics. Cell 155(1):27–38

Lin W-Y, Lou X-Y, Gao G, Liu N (2014) Rare variant association testing by adaptive combination of p-values. PloS one 9(1):85728

Liu DJ, Peloso GM, Zhan X, Holmen OL, Zawistowski M, Feng S, Nikpay M, Auer PL, Goel A, Zhang H, Peters U, Farrall M, Orho-Melander M, Kooperberg C, McPherson R, Watkins H, Willer CJ, Hveem K, Melander O, Kathiresan S, Abecasis GR (2014) Meta-analysis of gene-level tests for rare variant association. Nat Genet 46(2):200–204

Narendra PM, Fukunaga K (1977) A branch and bound algorithm for feature subset selection. Comput IEEE Trans 100(9):917–922

Pan W, Kim J, Zhang Y, Shen X, Wei P (2014) A powerful and adaptive association test for rare variants. Genetics 197(4):1081-1095

Persson H, Kvist A, Rego N, Staaf J, Vallon-Christersson J, Luts L, Loman N, Jonsson G, Naya H, Hoglund M, Borg A, Rovira C (2011) Identification of new microRNAs in paired normal and tumor breast tissue suggests a dual role for the erbb2/her2 gene. Cancer Res 71(1):78–86

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, Sham PC (2007) Plink: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559–575

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464

Simon RM, Korn EL, McShane LM, Radmacher MD, Wright GW, Zhao Y (2003) Design and analysis of DNA microarray investigations. Springer-Verlag, New York

Somol P, Pudil P, Kittler J (2004) Fast branch & bound algorithms for optimal feature selection. Pattern Anal Mach Intell IEEE Trans 26(7):900–912

Stone M (1974) Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society. Series B (Methodological) 36(2):111–147

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

Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J (2002) Least squares support vector machines. World Scientific Publishing, Singapore

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

Tu S, Xu L (2012) A theoretical investigation of several model selection criteria for dimensionality reduction. Pattern Recognit Lett 33(9):1117–1126

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

Williams CKI (2003) Learning kernel classifiers. J Am Stat Assoc 98(462):489–490

Xu L, Yan P, Chang T (1988) Best first strategy for feature selection. In: 9th International Conference on Pattern Recognition. IEEE Computer Society Press, Piscataway, New Jerse. pp 706–708

Xu L (1995) Bayesian-Kullback coupled ying-yang machines: unified learnings and new results on vector quantization. In: Proc. Int. Conf. Neural Information Process (ICONIP '95). Publishing House of Electronics Industry, Beijing. pp 977–988

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: Encyclopedia of Artificial Intelligence. IGI Global IGI Global Snippet, Hershey, Pennsylvania. pp 892–901

Xu L (2010) Bayesian ying-yang system, best harmony learning, and five action circling. Front Electrical Electronic Eng China 5(3):281–328

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 (2012a) Semi-blind bilinear matrix system, BYY harmony learning, and gene analysis applications. In: Proceedings of The 6th International Conference on New Trends in Information Science, Service Science and Data Mining: 23-25 October 2012, IEEE. pp 661–666

Xu, L (2012b) On essential topics of BYY harmony learning: current status, challenging issues, and gene analysis applications. Front Electrical Electronic Eng 7(1):147–196

Xu L (2013a) Integrative hypothesis test and A5 formulation: sample pairing delta, case control study, and boundary based statistics. In: Intelligence Science and Big Data Engineering. LNCS. Springer, Berlin Heidelberg. pp 887–902

Xu L (2013b) Matrix-Variate discriminative analysis, integrative hypothesis testing, and geno-pheno A5 analyzer. In: Intelligent Science and Intelligent Data Engineering. LNCS. Springer, Berlin Heidelberg Vol. 8261. pp 866–875

Xu L (2015) Further advances on Bayesian ying yang harmony learning. Applied Informatics 2(5)

Xu L, Amari SI (2008) Combining classifiers and learning mixture-of-experts. In: J Ramon e.a. (ed). Encyclopedia of Artificial Intelligence. IGI Global, Hershey: PA. pp 318–326

Xu L, Krzyzak A, Suen CY (1992b) Several methods for combining multiple classifiers and their applications in handwritten character recognition. IEEE Trans Syst Man Cybernet 22:418–435

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