Matrix-Variate Discriminative Analysis, Integrative Hypothesis Testing, and Geno-Pheno A5 Analyzer

Lei Xu

Dept. of Computer Science and Engineering, Chinese Univ. of Hong Kong
Chang Jiang Chair Professor Program, School of EE&CS, Peking Univ., Beijing
lxu@cse.cuhk.edu.hk

Abstract. A general perspective is provided on both on hypothesis testing and discriminative analyses, by which matrix-variate discriminative analyses are proposed based on the matrix normal distribution, featured by a bi-linear extension of Fisher linear discriminant analysis and a further extension to binary variables. Moreover, a general formulation is proposed for integrative hypothesis testing and five typical categories are summarized. Furthermore, major techniques for variable selection are briefly elaborated. Finally, taking analyses of gene expression and exome sequencing as examples, we further propose a general procedure called *Geno-Pheno A5 Analyzer* for integrative discriminant analysis.

Keywords: Matrix-variate discriminative analysis, Bi-linear Fisher mapping, Matrix-variate logistic regression, Confusion table testing, Geno-Pheno A5 analyzer, Gene expression analysis, Exome sequencing analysis.

1 Introduction

Fisher discriminative analysis works in a multidimensional space with samples presented as vectors. However, samples are usually in matrix format for tasks such as image classification, object recognition, and various gene analyses. Considering samples in vectors actually suffers some approximation to get an easy implementation. Improvements are expected if we make discriminative analysis directly on samples in matrix format.

Working on samples from two populations or classes, discriminative analysis is featured by finding a discriminating rule that classifies each sample into one appropriate class. In a complementary aspect, two sample hypothesis test examines whether two populations are significantly different via a statistics based on samples from the populations.

This paper provides a general perspective on both the aspects, with a road not only to generalize discriminative analyses and hypothesis testing to matrix-variate samples but also to a general formulation for testing hypotheses on a set of variables organized in structures. Also, five categories are summarized according to the characteristics of testing hypotheses.

Also, major techniques are briefly elaborated for identifying which subsets of variables are responsible to testing significance and discriminative ability. Taking gene expression analysis and exome sequencing analysis as examples, we propose a Geno-Pheno A5 analyzer for integrative genotype-phonotype discriminant analysis.

J. Yang, F. Fang, and C. Sun (Eds.): IScIDE 2012, LNCS 7751, pp. 866–875, 2013.

2 KL Perspective on Hotelling Statistics and Fisher Discriminant

We start from considering two multivariate Gaussian populations:

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

with the proportional priori $\alpha^{(l)}$. The parameters are obtained by the maximum likelihood (ML) estimation with

$$\alpha^{(\ell)} = \frac{N_\ell}{N}, \; N = \sum\nolimits_\ell N_\ell, \; \mu^{(\ell)} = \frac{1}{N_\ell} \sum_{t=1}^{N_\ell} x_t^{(\ell)}, \; \Sigma^{(\ell)} = \frac{1}{N_\ell - 1} \sum_{t=1}^{N_\ell} (x_t - \mu^{(\ell)})(x_t - \mu^{(\ell)})^T.$$

Under the null hypothesis

$$H_0 : \; G(x_t \mid \mu^{(1)}, \Sigma^{(1)}) = G(x_t \mid \mu^{(0)}, \Sigma^{(0)}), \tag{2}$$

we have $\mu=\mu^{(l)}$, $\Sigma=\Sigma^{(l)}$. Putting $X^{(0)}$, $X^{(1)}$ together, the ML estimation becomes

$$\mu = \alpha^{(0)}\mu^{(0)} + \alpha^{(1)}\mu^{(1)}, \; \Sigma = \alpha^{(0)}\Sigma^{(0)} + \alpha^{(1)}\Sigma^{(1)}.$$

We further use the following Kullback–Leibler (KL) divergence

$$\mathrm{KL}(\mathrm{p} \,\|\, \mathrm{q}) = \int p(x) \ln[p(x)/q(x)] dx \tag{3}$$

to measure the deviation contributed by the difference of $\mu^{(l)}$, resulting in

$$s_{KL} = KL(G(x \mid \mu^{(1)}, \Sigma) \,\|\, G(x \mid \mu, \Sigma)) = \mathrm{Tr}[(\mu^{(0)} - \mu^{(1)})(\mu^{(0)} - \mu^{(1)})^{\mathrm{T}} \Sigma^{-1}], \tag{4}$$

from which we observe the following Hotelling two-sample T-squared statistics [1]

$$T^2 = \alpha_0 \alpha_1 s_{KL} \tag{5}$$

Moreover, considering the following linear projection

$$y_t = w^T x_t, \;\; \mu_y^{(\ell)} = w^T \mu^{(\ell)}, \; \Sigma_y = w^T \Sigma w, \tag{6}$$

we use Eq.(4) to measure two resulted scalar Gaussian populations and get

$$s_{KL}(w) = w^T (\mu^{(0)} - \mu^{(1)})(\mu^{(0)} - \mu^{(1)})^T w (w^T \Sigma w)^{-1}, \tag{7}$$

which could be further maximized to obtain an optimal w^*. That is, we reach the popular Fisher linear discriminant analysis (LDA).

3 Matrix-Variate Discriminative Analysis

Following Sect. 3 in [2], we consider a $d \times m$ matrix variate X from the following matrix normal distribution [3] :

$$N(X \mid M, \Omega, \Sigma) = \frac{\exp\{-0.5Tr[\Omega^{-1}(X - M)^T \Sigma^{-1}(X - M)]\}}{(2\pi)^{0.5md} \, |\Sigma|^{0.5d} \, |\Omega|^{0.5m}}, \; M = EX, \tag{8}$$

where an $m \times m$ matrix Ω describes the cross-column dependence of X and a $d \times d$ matrix Σ describes the cross-row dependence of X. This matrix distribution links to

a multivariate Gaussian distribution $G(vec(X)\,|\,vec(M),\Sigma\otimes\Omega)$, where $\otimes$ denotes the Kronecker product and $vec[A]$ is the vectorization of a matrix A.

Given samples of the following populations:

$$\mathbf{X}^{(\ell)}=[X_1^{(\ell)},\ldots,X_{N_\ell}^{(\ell)}],\ \ X_t^{(\ell)}\ from\ N(X\,|\,M^{(\ell)},\Omega^{(\ell)},\Sigma^{(\ell)}),\ \ \ell=0,1, \tag{9}$$

the parameters are estimated as follows

$$\alpha^{(\ell)}=\frac{N_\ell}{\sum_\ell N_\ell},\ \ M^{(\ell)}=\frac{1}{N_\ell}\sum_{t=1}^{N_\ell}X_t^{(\ell)},\Sigma^{(\ell)}=\frac{1}{N_\ell-1}\sum_{t=1}^{N_\ell}(X_t^{(\ell)}-M^{(\ell)})(X_t^{(\ell)}-M^{(\ell)})^T, \tag{10}$$

where $\Omega^{(l)}$ depends on $\Sigma^{(l)}$ and one estimate is given by

$$\Omega^{(\ell)}=\frac{1}{N_\ell-1}\sum_{t=1}^{N_\ell}(X_t-M^{(\ell)})^T\Sigma^{(\ell)-1}(X_t-M^{(\ell)}). \tag{11}$$

With the normal distribution, we get the following observations. First, we can get the following matrix variate Bayes discriminative rule:

$$\ell^*=\arg\max_\ell g^{(\ell)}(X_t),\ \ \ g^{(\ell)}(X_t)=\ln N(X^{(\ell)}\,|\,M^{(\ell)},\Omega^{(\ell)},\Sigma^{(\ell)})+\ln\alpha^{(\ell)}. \tag{12}$$

Second, it follows from Eq.(4) that we similarly obtain

$$s_{KL}=KL(N(X\,|\,M^{(1)},\Omega,\Sigma)\,\|\,N(X\,|\,M,\Omega,\Sigma))=\mathrm{Tr}[\Omega^{-1}(M^{(1)}-M^{(0)})^T\Sigma^{-1}(M^{(1)}-M^{(0)})], \tag{13}$$

with the following estimates:

$$M=\frac{1}{N}\sum_{\ell=0,1}\sum_{t=1}^{N_\ell}X_t^{(\ell)},\ \Sigma=\frac{1}{N-1}\sum_{\ell=0,1}\sum_{t=1}^{N_\ell}(X_t^{(\ell)}-M)(X_t^{(\ell)}-M)^T,$$

$$\Omega=\frac{1}{N-1}\sum_{\ell=0,1}\sum_{t=1}^{N_\ell}(X_t^{(\ell)}-M^{)})^T\Sigma^{-1}(X_t^{(\ell)}-M).$$

Third, we get a bi-linear projection by a $d_s\times d$ matrix W and a $m\times m_s$ matrix U

$$Y_t=W^TX_tU,\ \ M_y^{(\ell)}=W^TM^{(\ell)}U. \tag{14}$$

It follows from Eq.(10) and Eq.(11) that we have

$$\Sigma_y=\frac{1}{N-1}W^T\sum_{\ell=0,1}\sum_{t=1}^{N_\ell}(X_t^{(\ell)}-M^{(\ell)})UU^T(X_t^{(\ell)}-M^{(\ell)})^TW,$$

$$\Omega_y=\frac{1}{N-1}U^T\sum_{\ell=0,1}\sum_{t=1}^{N_\ell}(X_t^{(\ell)}-M^{(\ell)})^TW\Sigma_y^{-1}W^T(X_t^{(\ell)}-M^{(\ell)})\,U. \tag{15}$$

Then, we use s_{KL} in Eq.(13) to measure the two matrix normal populations

$$s_{KL}(W,U)=\mathrm{Tr}[\Omega_y^{-1}U^T(M^{(1)}-M^{(0)})^TW\Sigma_y^{-1}W^T(M^{(1)}-M^{(0)})U], \tag{16}$$

which is further maximized to obtain an optimal W^*, U^*. That is, we have

$$\{W^*,U^*\}=\arg\max_{W,U}s_{KL}(W,U), \tag{17}$$

which is a bi-linear extension of Fisher linear discriminant analysis.

Particularly, when d_s=m_s =1, we get the following Fisher criterion

$$s_{KL}(w,u)=\frac{(\mu_y^{(1)}-\mu_y^{(0)})^2}{\sigma_y^2},\quad y_t=w^T X_t u,\quad \mu_y^{(\ell)}=w^T M^{(\ell)}u,$$
$$\sigma_y^2=\Omega_y\Sigma_y=\frac{1}{N-1}\sum_{\ell=0,1}\sum_{t=1}^{N_\ell}[w^T(X_t^{(\ell)}-M^{(\ell)})u]^2. \tag{18}$$

Moreover, $y_t= w^T X_t u$ may also be turned into a regression to a label $I_t=1$ as X_t comes from $X_t^{(1)}$ or label $I_t=0$ as X_t comes from $X_t^{(0)}$, i.e.,

$$p(I_t=1\,|\,y_t)=1/[1+\exp(-\beta y_t)].$$

We may estimate w*, u* by maximizing the following likelihood

$$L(w,u)=-\sum_{t=1}^{N_1}\ln(1+\exp[-\beta w^T X_t^{(1)}u])-\sum_{t=1}^{N_0}\ln(1+\exp[\beta w^T X_t^{(0)}u]). \tag{19}$$

Then, we further develop an algorithm for maximizing the measure either by Eq.(16) or Eq.(19). Simply, we may implement an alternative gradient descent updating to maximize *J(W,U)*, that is, we alternatively update

$$W^{new}=W^{old}+\eta\nabla_W J(W,U),\quad U^{new}=U^{old}+\eta\nabla_U J(W,U). \tag{20}$$

Moreover, a term of L1 norm or lasso penalty or Laplace priori may be added to *J(W,U)* to regularize the coefficients of W, U such that extra parameters will be eliminated, e.g., adding to *w* by Eq.(7) leads to sparse LDA [4].

Last but not least, we may further consider two even general cases.

One is the following general matrix normal populations

$$s_{KL}=KL(N(X|M^{(1)},\Omega^{(1)},\Sigma^{(1)})||N(X|M,\Omega,\Sigma)), \tag{21}$$
$$\Sigma=\sum_{j=1}^{k}\lambda_j\phi_j\phi_j^T+\sigma^2 I,\ k\le m,$$

where $\lambda_j,\phi_j, j=1,\cdots,k$ are the first k largest eigenvalues of Σ and the corresponding eigenvectors. Typically, we consider $k<d$ for a small size of samples, which also provides modification to s_{KL} by Eq. (6) and s_{KL} by Eq.(13) at the special case $\Omega=I$.

The other considers that all the elements in X are binary valued, with help of the following Gibbs measure:

$$p(X|\beta,\Phi,\Psi)\propto e^{-\beta E(X|\Phi,\Psi)},\ E(X|\Phi,\Psi)=\sum_{i,j,k,\ell}\phi_{ij}\psi_{k\ell}x_{ik}x_{j\ell}, \tag{22}$$

which is used to replace $N(X|\cdot,\cdot,\cdot)$ in Eqs.(12) (21).

4 Integrative Hypothesis Testing and Variable Selection

4.1 One General Formulation and Typical Categories

Originally, studies on hypothesis testing consider samples from populations of one random variable, e.g., a SNP takes one of genotypes in the popular PLINK study [5].

For practical problems, a hypothesis is typically made on multi-variables and studies have also proceeded to multivariate hypothesis test [6,7], which recently gets ever-interested in gene analyses [8,9]. However, these studies work on vector samples while we actually encounter hypotheses on a set of variables organized in structures beyond vector. Thus, we need to perform testing by integrating information associated with these structured variables. Informally, we refer efforts towards this direction to a term called integrative hypothesis testing (IHT).

Similar to Eqs.(4), (13) & (21), we propose a general IHT formulation with the help of the KL divergence by Eq.(3). Given samples of the following populations

$$\mathbf{X}^{(\ell)} = [X_1^{(\ell)}, \ldots, X_{N_\ell}^{(\ell)}], \;\; X_t^{(\ell)} \; from \; p(X \mid \Theta^{(\ell)}), \quad \ell = 0,1,$$

with variables of $X_{\mathbf{t}}^{(l)}$ in certain specific structure, we estimate $\Theta^{(l)}$ from $\mathbf{X}^{(l)}$ by a learning principle, e.g., by the following maximum likelihood estimation

$$\Theta^{(1)} = \arg\max_{\Theta} \ln p(\mathbf{X}^{(1)} \mid \Theta),$$

$$\Theta^{(0)} = \arg\max_{\Theta} \begin{cases} \ln p(\mathbf{X}^{(0)} \mid \Theta), & \text{Choice (a)}, \\ [\ln p(\mathbf{X}^{(1)} \mid \Theta) + \ln p(\mathbf{X}^{(0)} \mid \Theta)], & \text{Choice (b)}. \end{cases}$$

Then, we test the following null hypothesis

$$H_0: \; p(X \mid \Theta^{(1)}) \text{ is not different from } p(X \mid \Theta^{(0)}), \tag{23}$$

by the following statistics

$$s_{KL} = KL(p(X \mid \Theta^{(1)}) \,\|\, p(X \mid \Theta^{(0)})). \tag{24}$$

Also, there could be many other statistics for testing H_0 by Eq.(23). Even further, there could be various formulations for this testing, which are roughly classified into the following five categories.

(1) Decomposing H_0 by Eq.(23) with $X=\{\xi_1,...,\xi_d\}$ into the sub-hypotheses

$$H_0^{(j)}: \; p(\xi_j \mid \theta_j^{(1)}) \text{ is not different from } p(\xi_j \mid \theta_j^{(0)}), \; j = 1, \ldots, d, \tag{25}$$

such that H_0 by Eq.(23) is equivalent to a composition of the sub-hypotheses. Testing H_0 by Eq.(23) is made by combining the tests of sub-hypotheses, e.g., combining the resulted p-values by the *Fisher's combined probability test* [10], the Kost's method [11], and others [12,13].

(2) Targeting at computing s_{KL} by effectively exploring the structure of X.

When X consists of independent variables ξ_1, ..., ξ_d, we have

$$s_{KL} = \sum_j s_j, \;\; s_j = KL(p(\xi_j \mid \theta_j^{(1)}) \,\|\, p(\xi_j \mid \theta_j^{(0)})), \tag{26}$$

that is, we make a test by a statistics that is simply a sum of individual ones, E.g., if each $s_j \sim \chi(k)$, we simply consider $s_{KL} \sim \chi(kd)$.

When ξ_1, ..., ξ_d are joint variables in vector or matrix formats, not only the Hotelling T^2 statistics [1] can be computed from s_{KL} by Eq.(5) but also H_0 by Eq.(23) can be tested by an extension of the Hotelling statistics by Eq.(13).

For other dependence structures, e.g., $p(\xi_1, \xi_2, \xi_3) = p(\xi_1)p(\xi_2|\xi_1)p(\xi_3|\xi_2)$, we have $s_{KL}=s_1+s_{2|1}+s_{3|2}$ with $s_1=KL(p(\xi_1)\|q(\xi_1))$ and $s_{i|j}=KL(p(\xi_i|\xi_j)\|q(\xi_i|\xi_j))$.

Moreover, we may consider a statistics other than s_{KL}, e.g., Dempster's statistics [6], and others [7].

(3) Mapping H_0 by Eq.(23) into one or a set of hypotheses on certain inner representations of $X=\{\xi_1,...,\xi_d\}$. A rejection of these mapped hypotheses means that H_0 by Eq.(23) should be rejected since the truth of H_0 by Eq.(23) implies the truth of these mapped hypotheses. Being different from those tests directly made on the visible data domain of X (called Yang domain and thus named as Yang-test or shortly A-test), these mapped hypotheses are tested in an inner representation domain (called Ying domain), which forms another type of hypothesis testing that is named as Ying test or shortly I-test. In Sect. 6 of [2], two I-test approaches have been proposed which are elaborated below:

(a) FDA based one variable test Considering the projection y_t from either Eq.(6) or Eq.(18), we test the following null hypothesis

$$H_0: G(y_t \mid \mu_y^{(1)}, \sigma_y^{(1)2}) \text{ is not same as } G(y_t \mid \mu_y^{(0)}, \sigma_y^{(0)2}), \tag{27}$$

which can be performed by the Welch's t test. For Eq.(6), a best direction w^* is obtained by *Fisher discriminant analysis* (FDA) by Eq.(7).

(b) Confusion table test (CTT) Instead of making a linear projection, we use a classifier to map the samples of $X^{(0)}, X^{(1)}$ into binary labels C_0 and C_1, resulting in a confusion table shown in Fig.1(a) on which we further test

$$H_0: \text{ two rows have no difference.} \tag{28}$$

Again, its rejection means that H_0 by Eq.(23) should be rejected. This classifier is trained for a minimum classification error towards $N_{10}=0$ and $N_{01}=0$, i.e., two rows become as different as possible. Then, the hypothesis H_0 by Eq.(28) can be tested by the following Pearson chi-square statistics

$$t = \sum_{i=0,1} [\frac{(N_{i1} - pN_i)^2}{pN_i} + \frac{(N_{i0} - (1-p)N_i)^2}{(1-p)N_i}] \sim \chi^2(1), \; p = \frac{N_{01} + N_{11}}{N_0 + N_1}. \tag{29}$$

(4) Integrating tests on a number of the null hypotheses by Eq.(28) on the confusion tables obtained from a number of classifiers that either come from Eq.(6) or Eq.(18) for different values of w,u or come from different implementations (e.g., different learning methods, different initializations, or by A-3 given in the next section). The integration is made by one of the methods in the above Category (1).

(5) Making integrative hypothesis testing by combining a number of classifiers with help of either of majority voting, Bayes voting, product rule, Dempster-Shafer rule (see Tab.2 in [14]), resulting in a final confusion table for testing the null hypotheses by Eq.(28). Moreover, we may learn a tree classifier to handle a tree-structured variables $\xi_1, ..., \xi_d$, resulting in a final confusion table for testing a general null hypotheses by Eq.(23).

4.2 Feature Selection: Bottom-Up Search versus Top-Down Search

Once H_0 is rejected on the joint variables of X, we further identify which one or subset of variables are responsible to this rejection, which is called feature selection and implemented by one of the following typical techniques.

Bottom-up search is featured by searching subset of variables with the size of subsets increasing from 1 to a desired one. Typical efforts may be summarized into the following two classes:

(1) *Examining all the combinations* by the J –value (i.e., the value of J). First, each variable of X is examined. Second, each pair of variables are examined. At each m, each of C_d^m subsets of X is examined, until m reaches to a desired one.

(2) *Incremental stepwise search* First, the best variable ξ_1^* (i.e., a biggest J value) is selected from X, resulting in X_{d-1}. Second, each pairing of ξ_1^* with each of $d-1$ variables of X_{d-1} is examined by the J-value, with the best $\xi_1^*\xi_2^*$ selected and X_{d-1} reduces to X_{d-2}, …, so forth, until reaching a desired size.

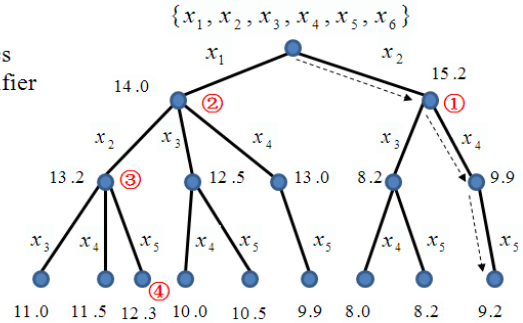
Choices of Selection Criterion J

- the value of s_{KL} or one other statistics
- the misclassification error of a classifier
- the p-value

(c)

	H_1 reject	H_0 not reject
N_1 samples of case $X^{(1)}$	N_{11} samples	N_{10} samples
N_0 samples of control $X^{(0)}$	N_{01} samples	N_{00} samples

(a)



(b)

Fig. 1. (a) A confusion table obtained by classification. (b) Illustration of the depth-first search (dashed line) versus the best-first search for feature selection. (c) Selection criterion J.

Top-down search is featured by searching subsets of variables with the size of subsets decreasing from d to a desired one. Typical efforts may be summarized into the following four classes:

(a) *Testing individual variable* Testing the significance of the coefficient associated with each variable, e.g., by the likelihood ratio test or the Wald statistics test [15].

(b) *Backward elimination* The simple one is the depth-first search. It gradually eliminates a variable that is the least important one, as d reduces to a desired one. At each m, we examine each of m subsets X_{m-1} resulted from eliminating x from X_m and keep the best subset by the J-value, as shown by the dashed line in Fig.1(b). However, this search is easy to fall in a local optimum. Instead, as shown by the circled numbers in Fig.1(b), a better alternative is the best-first search proposed in [16], which will find the optimum without enumerating all the possibilities when the J-value is monotonic to the cardinality of variable sets.

(c) *Sparse learning* that eliminates the coefficients associated with extra variables during learning, as previously discussed after Eq.(19).

Mixed search is featured by alternatively using bottom-up and top-down search in a specific procedure, e.g., forward-backward stepwise regression [17]. Moreover, we may make sparse learning with different initializations that lead different subsets of variables, and then use the union of these subsets to start a top-down search.

5 Integrative Geno-Pheno A5 Analyzer

The proposed methods in Sect.3 and Sect.4 are applicable to tasks such as image recognition, object detection, fault and disease diagnosis, with improvements that come from considering samples in matrix format. Particularly, the methods are more appealing to those real problems for which samples are actually in matrix format but used approximately in vectors merely for an easy implementation.

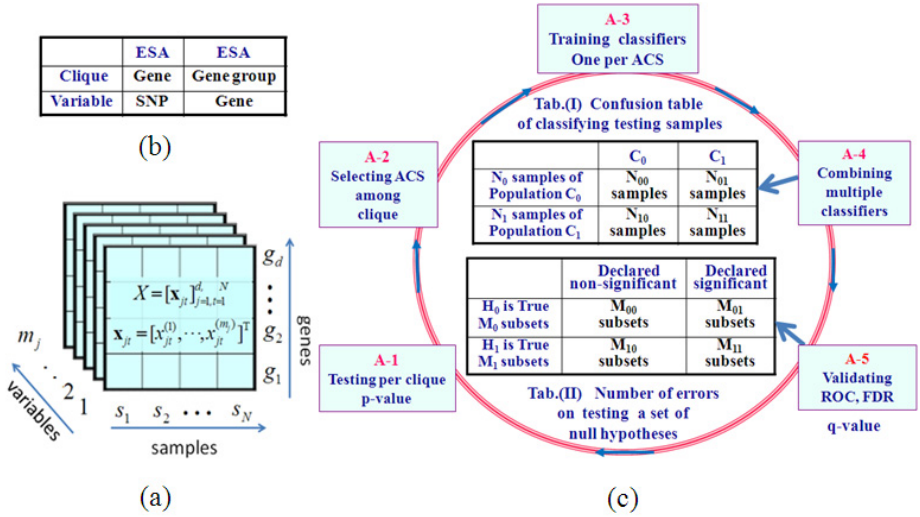


Fig. 2. (a) general data type widely encountered in gene analyses. (b) two levels of variable search. (c) integrative *Geno-Pheno A5 Analyzer*.

Many problems in gene analyses handle data as shown in Fig.2(a), e.g. eight types of data shown in Fig.5 of [18]. Here, we focus on two typical examples. One is *gene/RNA expression analysis (GEA).* One simplest case consists of an array with each column representing a sample and each row corresponding gene. Moreover, each $\mathbf{x}_{jt}$ corresponds to the expression of one gene g_j. Generally, we may also consider the expressions of one gene under an additional control, represented by several variables for describing different times, conditions, …, etc, where the number m_j may vary for different genes. In such cases, $\boldsymbol{X}$ is a data cubic with each sample being a $m_j \times d$ matrix.

The other is *exome sequencing analysis (ESA), e.g.,* see Sect.6.3 in [19]. For each gene, $\mathbf{x}_{jt}$ is a vector that consists of several SNPs with each in a genotype. Specifically, for the *l*th SNP of the gene *j* we take one of the following two choices for coding genotypes:

(a) $\mathbf{x}_{jt}^{(l)}$=1 for AA, $\mathbf{x}_{jt}^{(l)}$=2 for Aa, $\mathbf{x}_{jt}^{(l)}$=3 for aa, and $\mathbf{x}_{jt}^{(l)}$=0 if missing genotype;

(b) $\mathbf{x}_{jt}^{(l)}$=[1,0,0] for AA, $\mathbf{x}_{jt}^{(l)}$=[0,1,0] for Aa, $\mathbf{x}_{jt}^{(l)}$=[0,0,1] for aa, and $x_{jt}^{(l)}$=[1/3,1/3, 1/3] when its genotype is missing, i.e., each $\mathbf{x}_{jt}$ is an $m_j \times 3$ matrix.

Among the problems with the data type shown in Fig.2(a), one widely studied family consists of association analyses on the relations of diseases or external phenomenon to genes or inner causes, e.g., whether a disease relates to some genes or SNPs within some

genes. Such a study is featured by its scope of focusing. E.g., the scope of GEA considers all genes in ***X,*** which usually incurs for a huge computing cost; while the scope of GWAS covers the SNPs of all the genes and incurs an even huge computing cost such that enumerating the joint-effects of multiple SNPs is computationally not feasible.

To tackle the issue, we propose to divide the scope of focusing into two levels, as illustrated in Fig.2(b). A study starts from considering samples per clique, e.g., a clique for ESA is a gene that contains its SNPs while a clique for GEA is a group of genes in a same pathway. Each clique may also be formed according to some domain knowledge, e.g., molecular biology knowledge for ESA and GEA. Without this knowledge, we may technically divide data in Fig.2(a) along the vertical direction, e.g., clustering analysis on gene expression data to group genes into cliques or even randomly picking genes to form a clique. This level investigates cliques one by one, and then the other level further examines subsets of each clique.

We propose a general procedure that integrates hypothesis test and discriminative analysis for genotype-phonotype association (also applicable to those tasks mentioned at the beginning of this section), shortly called *Geno-Pheno A5 Analyzer*. It is featured by five actions as shown in Fig.2(c), which is actually a special exemplar of the A5 problem solving paradigm. For details, readers are referred to Sect.4 in [20], Sect.3.1 and Appendix B in [21].

The *Geno-Pheno A5 Analyzer* performs the following five actions on each clique:

A-1 *(*acquisition): test a null hypothesis H_0 on samples of this clique to check whether there is a significant difference between two populations C_0, C_1 (e.g, by one method of Category (3) introduced in Sect.4.1). Return to A-1 if the test is not significant, and put the clique into a set named REST.

A-2 *(*assumption): evaluate subsets of the clique for ones responsible to the difference between C_0, C_1. Each subset that passes a significance level is identified as one assumed candidate subset (ACS), resulting in a family of M_0 such ACSs.

A-3 *(*amalgamation): build up a classifier per ACS by training samples, resulting in a set of classifiers for the set of ACSs.

A-4 *(*apex-seeking) : use each ACS classifier to assign testing samples into one of C_0, C_1 and combine the results by different ACS classifiers to get a final assignment, together with the confusion table Tab.(1) in Fig.2(b), e.g, by one method of Category (5) introduced in Sect.4.1

A-5 *(*affirmation) : validating two complementary hypotheses as follows:

(1) *Vertical direction* Each of M_0 ACSs is tested by testing samples, resulting in the first row of Tab.(2) in Fig.2(b). Also, M_1 subsets from the set REST are tested as H_1 hypotheses, resulting in the second row of Tab.(2) in Fig.2(b). Then, a validation is made by the concept of false discovery rate (FDR), e.g., by the Benjamini–Hochberg procedure based on the resulted p-values [22, 23]. Also, we may return to A-4 for combining those ACS classifiers that passed this validation.

(2) *Horizontal direction.* We validate the confusion table Tab.(1) by the rate of misclassification and the curve of Receiver Operating characteristics (ROC).

We may perform the above *Geno-Pheno A5 Analyzer* at the level of cliques one by one and then combine classifiers obtained from different cliques for a global test on whether two populations are significantly different.

Acknowledgments. This work was supported by the National Basic Research Program of China (973 Program) (No. 2009CB825404).

References

1. Hotelling, H.: The generalization of Student's ratio. Annals of Mathematical Statistics 2(3), 360–378 (1931)
2. Xu, L.: Semi-Blind Bilinear Matrix System, BYY Harmony Learning, and Gene Analysis Applications. In: Proc. of 6th International Conf. on New Trends in Information Science, Service Science and Data Mining, Taipei, October 23-25, pp. 661–666 (2012)
3. Dawid, A.P.: Some matrix-variate distribution theory: Notational considerations and a Bayesian application. Biometrika 68(1), 265–274 (1981)
4. Clemmensen, L., Hastie, T., Witten, D., Ersbøll, B.: Sparse discriminant analysis. Technometrics 53, 406–413 (2011)
5. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D., Maller, J., Sklar, P., de Bakker, P., Daly, M.J., Sham, P.C.: PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics 81 (2007), http://pngu.mgh.harvard.edu/purcell/plink/
6. Dempster, A.P.: A high dimensional two sample significance test. Ann. Math. Statist. 29, 995–1010 (1958)
7. Baringhaus, L., Franz, C.: On a new multivariate two-sample test. J. Multivariate Anal. 88, 190–206 (2004)
8. Glezko, G.V., Emmert-Streib, F.: Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets. Bioinformatics 25, 2348–2354 (2009)
9. Hummel, M., Meister, R., Mansmann, U.: GlobalANCOVA: exploration and assessment of gene group effects. Bioinformatics 24, 78–85 (2008)
10. Fisher, R.A.: The Statistical Utilization of Multiple Measurements. Annals of Eugenics 8, 376–386 (1938)
11. Kost, J., McDermott, M.: Combining dependent P-values. Statistics & Probability Letters 60(2), 183–190 (2002)
12. Whitlock, M.C.: Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. J. Evol. Biol. 5(18), 1368–1373 (2005)
13. Chen, Z.: Is the weighted z-test the best method for combining probabilities from independent tests? J. Evol. Biol. 24(4), 926–930 (2011)
14. Xu, L., Amari, S.: Combining classifiers and learning mixture-of experts. In: Ramón, J., et al. (eds.) Encyclopedia of Artificial Intelligence, pp. 318–326. IGI Global Pub. (2008)
15. Hilbe, J.M.: Logistic Regression Models. Chapman & Hall/CRC Press (2009)
16. Xu, L., Yan, P., Chang, T.: Best first strategy for feature selection. In: Proc. of 9th Intl Conf on Pattern Recognition, Rome, November 14-17, vol. 2, pp. 706–708 (1988)
17. Stepwise regression - Wikipedia, http://en.wikipedia.org/wiki/Stepwise_regression
18. Xu, L.: Codimensional matrix pairing perspective of BYY harmony learning: hierarchy of bilinear systems, joint decomposition. Front. Electr. Electron. Eng. China 6(1), 86–119 (2011)
19. Xu, L.: On essential topics of BYY harmony learning: Current status, challenging issues, and gene analysis applications. Front. Electr. Electron. Eng. 7(1), 147–196 (2012)
20. Xu, L.: A unified perspective and new results on RHT computing, mixture based learning, and multi-learner based problem solving. Pattern Recognition 40(8), 2129–2153 (2007)
21. Xu, L.: Bayesian Ying-Yang system, best harmony learning, and five action circling. Front. Electr. Electron. Eng. China 5(3), 281–328 (2010)
22. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. of Royal Statistical Society B 57(1), 289–300 (1995)
23. False discovery rate - Wikipedia, http://en.wikipedia.org/wiki/False_discovery_rate