PROCEEDINGS

ISSDM2012

2012 6th International Conference on New Trends in Information Science, Service Science and Data Mining (NISS, ICMIA and NASNIT)

Santos Hotel, Taipei, Taiwan
October 23-25, 2012

IEEE Conference Record Number: 20420
IEEE PDF files Catalog Number: CFP1213H-ART
IEEE PDF files ISBN: 978-89-94364-20-9
IEEE DVD version Catalog Number: CFP1213H-DVD
IEEE DVD version ISBN: 978-89-94364-23-0
IEEE Print version Catalog Number: CFP1213H-PRT
IEEE Print version ISBN: 978-89-94364-19-3

Editors

Dr. Chien-wen Shen (National Central University, Taiwan)
Dr. Sy-Yen Kuo (IEEE Taipei Section Chair, Taiwan)
Dr. Kae Dal Kwack (IEEE Korea Council Chair, Korea)
Prof. Yen-Wei Chen(Ritsumeikan University, Japan)
Prof. Ping-Yu Hsu(National Central University, Taiwan)
Dr. Franz Ko (Dong-A University, Korea/ IBC, UK)

IEEE IEEE IEEE Korea Council AICIT SINCE 2005

Semi-Blind Bilinear Matrix System, BYY Harmony Learning, and Gene Analysis Applications

Lei Xu

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

Chang Jiang Chair Professor Program
School of Electronics Engineering and Computer Science
Peking University, Beijing, PRChina

Abstract—**A bilinear matrix system (BMS) is proposed as a general semi-blind learning framework for modeling matrix-formatted data and for extracting matrix-formatted inner factors. Different special cases of this framework lead to a family of typical learning tasks. The problem of learning such a semi-blind BMS learning is formulated as a problem of learning a particular BYY system for estimating unknown parameters and for making model selection. We develop a BYY harmony learning algorithm for learning matrix normal distribution based BMS, which relates to and also generalizes typical learning methods, such as factor analyses, 2D-PCA, and manifold learning, …, etc, featured with automatic model selection on the bi-perspective dimensions. Also, we apply this algorithm for estimating the profiles of transcriptional factor activities from gene expression data. Moreover, we briefly outline typical applications of BMS, especially a new perspective of Yang domain based hypothesis test versus Ying domain based test, exampled by schematic algorithms and genetic diagnoses applications.**

Keywords—bilinear matrix system; semi-blind learning; BYY harmon learning; 2D-PCA, manifold learning; bi-perspective factor analyses; gene regulatory; hypothesis test ; genetic diagnoses.

I. INTRODUCTION

A family of learning models have been outlined recently in [1], under a unified framework named bilinear matrix system. A $d \times N$ sample matrix X with each sample vector x_t as one column is regarded as generated from

$$X = AY + E \quad (1)$$

where a stochastic source matrix Y is called factor matrix with each column put an inner factor vector y_t per sample vector x_t, and a stochastic system matrix A that is called loading matrix with its columns forming a coordinate system. Eq.(1) is a linear system with respect to either of A and Y, and thus is called a bilinear matrix system. Moreover, E is a residual matrix that is mutually independent not only among all its elements but also with AY, i.e., $E[AYE^T] = 0$ (2)

As shown by a roadmap given in Fig.1 of [1], we are led to a family of learning tasks by imposing certain constraints on Y and A. Typical examples include

- Factor analysis (FA), binary FA (including multiple cause mixture), nonGaussian FA, and temporal extensions.
- Gaussian mixture, mean square error clustering, binary matrix factorization [2], and nonnegative matrix factorization [3], as well as a mixture of FA models and local FA (including local PCA or local subspaces).

In this paper, we further address that the linear system by Eq.(1) is merely a special case of what we should call a bilinear matrix system **(BMS)**, which covers not only Eq.(1) but also manifold learning [4-6], as well as others. This BMS provides a general formulation for semi-blind modeling of matrix-formatted data and for extracting matrix-formatted inner factors from image. Details will be introduced in Section II, where the problem of learning BMS is further formulated as learning a particular BYY system.

Section III considers matrix normal distribution and develops a BYY harmony learning algorithm, which is thus further addressed in Section IV as a unified framework that covers and improves several existing learning methods. Section V outlines typical applications, while Section VI provides a new perspective of hypothesis test. Finally, Section VI proposes four genetic analysis applications.

II. SEMI-BLIND BILINEAR MATRIX SYSTEM AND BAYESIAN YING-YANG HARMONY LEARNING

The linear system by Eq.(1) is actually a degenerated special case of the following bilinear matrix system (BMS):

$$X = AYB^T + E, \quad (3)$$

where the $d \times N$ sample matrix X is a linear function of an $m \times M$ inner factor matrix Y with both a left-multiplicative $d \times m$ matrix A and a right-multiplicative $N \times M$ matrix B. It degenerates to Eq.(1) at $M = N$ and $B=I$. Still, A is a loading matrix that transforms columns (vectors) of Y into columns (vectors) of X, while B maps rows (sequences) of Y into rows (sequences) of X. Still, E is a residual matrix that is mutually independent from AYB^T.

Alternatively, Eq.(3) can be equivalently expressed as follows:

$$X = A\ Y_B + E,\ \ Y_B = YB^T. \quad (4)$$

As discussed in Sect. of [1], even the system by Eq.(1) already has indeterminacy for which certain constraints on Y, A, and X are usually needed. The situation becomes even worse with one additional matrix B. Thus, certain structures of Y, A, and B need to be known in advances, the task is to learn the rest unknowns of this system, that is, we are dealing with semi-blind learning problems, as addressed in Sect.1 of [1].

We start to consider a particular matrix that consists of independently and identically distributed (i.i.d.) elements, or shortly i.i.d. matrix. That is, a matrix $S=\{s_{ij}\}$ is distributed by

$$q(S) = \prod_{i,j} q(s_{ij}\,|\varphi), \quad (5a)$$

with every s_{ij} coming from a same scalar distribution $q(s|\varphi)$. One example is the standard normal distribution

$$q(s|\varphi) = N(s|0,1), \quad (5b)$$

where $N(u|\mu,\Sigma)$ denotes a Gaussian density with the mean μ and the covariance Σ. Generally, $q(s|\varphi)$ may come from the exponential family (see Sect.2 of [1]).

To adopt different scales, s_{ij} is rescaled into $u_{ij} = \lambda_i^r s_{ij} \lambda_j^c$ per column and per row, or in the following matrix form :

$$U = \Lambda_r^{0.5} S \Lambda_c^{0.5},\ \Lambda_r = diag[\lambda_1^r,\ldots,\lambda_m^r],\ \Lambda_c = diag[\lambda_1^c,\ldots,\lambda_M^c], \quad (6)$$

which consists of bi-directionally recalled and independently distributed (r.i.d.) elements, or shortly we call r.i.d. matrix.

We consider the following structural r.i.d. matrices:

$$Y = \Lambda_r^{0.5} Y_s \Lambda_c^{0.5},\ E = D_\sigma^{0.5} E_s, \quad (7)$$

where E is rescaled merely per row, i.e., columns of E are identically distributed. Both Y_s and E_s are i.i.d. matrices by Eq.(5a)&(5b). Thus, the structures of $q(Y|\theta_y)$ and $q(E|\theta_e)$ become available once $q(s|\varphi)$ is chosen. Also, we can get $q(Y_B|\theta_y)$ from $q(Y|\theta_y)$ via $Y_B = YB^T$.

It follows from Eq.(3) that we get the following structure:

$$q(X|Y,\theta_{x|y}) = q(E|\theta_e) = q(X|AYB^T,\theta_{x|y}) = q(X|AY_B,\theta_{x|y}) = q(X|Y_B,\theta_{x|y}). \quad (8)$$

To reduce the system indeterminacy, certain structures will be also added on either or both of A and B. Especially, different structures of B also lead to different learning models.

All the rest unknown parameters are estimated under a learning principle with help of an efficient implementing algorithm. One typical principle is maximizing the likelihood

$$q(X|\theta) = \int q(X|Y,\theta_{x|y}) q(Y|\theta_y) dY \ \text{ or } Y \text{ is replaced by } Y_B \quad (9)$$

for which a gradient algorithm or an EM-like algorithm is usually developed for implementation. Moreover, constraints may also be imposed via appropriate priories

$$q(\theta|\mathbf{k}) = q(A)q(B)q(\theta_{x|y})q(\theta_y) \quad (10)$$

on a part or all the parameters, based on which we may further implement Bayesian learning or variational Bayes (VB) [7,8].

Extensive experiments have shown in [9,10] that the Bayesian Ying-Yang (BYY) harmony learning outperforms not only maximum likelihood learning but also considerably outperforms Bayesian learning, VB, and minimum message length. The BYY harmony learning was proposed in 1995 [11] and developed systematically over a decade and half, which provides not only a framework that accommodates typical learning approaches from a unified perspective but also a new road that leads to improved model selection criteria and Ying-Yang learning algorithms with automatic model selection. Readers are further referred to recent overviews [1,12,13].

The BYY system considers the joint distribution of a set $\mathbf{X}_N = \{x\}_{t=1}^N$ of samples and its inner representation $\mathbf{R} = \{Y, \theta, \mathbf{k}\}$ in two types of Bayesian decomposition, where $\mathbf{k}$ consists of one or more integers that represents the complexity of the system, e.g., m, M in Eq.(3). From a modern perspective of the famous ancient Yin-Yang philosophy, one type is called Yang machine, coinciding the Yang concept with a visible domain $p(X|X_N)$ obtained from a set X_N of samples for a Yang space and a $\mathbf{X} \rightarrow \mathbf{R}$ pathway by $p(\mathbf{R}|\mathbf{X})$ as a Yang pathway. The other is called Ying machine with an invisible domain $q(\mathbf{R})$ for a Ying space and $\mathbf{R} \rightarrow \mathbf{X}$ by $q(\mathbf{R}|\mathbf{X})$ as a Ying pathway.

Specifically, we have

$$q(\mathbf{R}) = q(Y|\theta_y)q(\theta)q(\mathbf{k}) \text{ or } q(\mathbf{R}) = q(Y_B|\theta_y)q(\theta)q(\mathbf{k}), \quad (11)$$
$$q(X|\mathbf{R}) = q(X|Y,\theta_{x|y}) = q(X|Y_B,\theta_{x|y}),$$

with each of components as introduced from Eq.(5) to Eq.(8). Typically, the structure of $p(\mathbf{R}|\mathbf{X})$ is designed as some type of inverse of the Ying machine, for which details are referred to one recent overview [13], especially its Sect.3.2.

After the structures of each component in a Ying-Yang pair have been specified, all the rest unknowns are determined by maximizing the following harmony functional

$$H(p\|q) = \int p(\mathbf{R}|\mathbf{X}) p(\mathbf{X}) \ln[q(\mathbf{X}|\mathbf{R}) q(\mathbf{R})] d\mathbf{X} d\mathbf{R}, \quad (12)$$

Specifically, it consists of a parameter learning task for estimating $\theta = \{A,\ B,\ \theta_y,\ \theta_e\}$ and model selection task of selecting $\boldsymbol{k} = \{m, M\}$. Both the tasks are implemented via maximizing $H(p||q)$, *i.e.,*

$$[\theta^*, \mathbf{k}^*] = \arg\max_{\theta,\mathbf{k}} H(p\|q). \quad (13)$$

III. Normal Matrix Distribution and BYY Harmony Learning Algorithm

One typical situation of semi-blind BMS is featured with each element in Y_s and E_s coming from a standard normal distribution. Expressed in a matrix format, we have

$$q(Y|\theta_y) = N(Y|0,\Lambda_c,\Lambda_r),\ q(X|Y,\theta_{x|y}) = N(X|AYB^T, I, D_\sigma), \quad (14)$$

where the notation means the matrix normal distribution, that is, for a $d \times N$ random matrix Z we have

$$N(Z|M,\Omega,\Sigma) = \frac{\exp\{-0.5Tr[\Omega^{-1}(Z-M)^T\Sigma^{-1}(Z-M)]\}}{(2\pi)^{0.5dN}|\Sigma|^{0.5N}|\Omega|^{0.5d}},\ M = EZ, \quad (15)$$

which links to a multivariate normal distribution by

$$N(vec(Z)|vec(M), \Omega \otimes \Sigma), \quad (16)$$

where $\otimes$ denotes the Kronecker product and $vec[A]$ denotes the vectorization of a matrix A.

For simplicity and without losing generality, the sequel considers EX=0, i.e., the mean has be removed from samples.

Following Eq.(35) and Eq.(36) in [13], it follows from the above Eq.(15) that we consider $p(Y|X,\theta)$ by

$$p(vec(Y)|X,\theta) = N(vec(Y)|vec(EY), \Pi_\theta^{-1}), \quad (17)$$
$$EY = WX,\ \Pi_\theta = (\Lambda_r \otimes \Lambda_c)^{-1} + (A^T D_\sigma^{-1} A) \otimes (B^T B).$$

Using a set of samples directly, i.e., $p(X|X_N) = \delta(X - |X_N)$, putting Eq.(14) and Eq.(17) into Eq.(12) and following a derivation of getting Eq.(37) and Eq.(38) in [13], we have

$$H(p\|q) = L(X_N, Y^*, \theta) - 0.5[mM + vec^T(\delta_Y)\Pi_\theta vec(\delta_Y)],$$
$$L(X_N, Y, \theta) = \ln[N(X_N|AYB^T, I, D_\sigma) N(Y|0,\Lambda_c,\Lambda_r) q(\theta|\mathbf{k})], \quad (18)$$
$$\delta_Y = Y^* - WX_N,\ \ Y^* = \arg\max_Y L(X_N, Y, \theta),$$
$$vec^T(\delta_Y)\Pi_\theta vec(\delta_Y) = Tr[\Lambda_c^{-1}\delta_Y^T\Lambda_r^{-1}\delta_Y] + Tr[B^T B\delta_Y^T A^T D_\sigma^{-1} A\delta_Y]$$

Also, we consider the priori by Eq.(10) as follows:

$$q(\theta|\mathbf{k}) = q(\Lambda_r)q(\Lambda_c)q(D_\sigma)q(A, \{\eta_{ij}^A\})q(B, \{\eta_{ij}^B\}),$$
$$q(\Lambda_r) \propto |\Lambda_r|^{-0.5},\ q(\Lambda_c) \propto |\Lambda_c|^{-0.5},\ q(D_\sigma) \propto |D_\sigma|^{-0.5},$$
$$q(A, \{\eta_{ij}^A\}) \propto \prod_{ij} \{G(a_{ij}|0, \eta_{ij}^A)/\sqrt{\eta_{ij}^A}\},$$
$$q(B, \{\eta_{ij}^B\}) \propto \prod_{ij} \{G(b_{ij}|0, \eta_{ij}^B)/\sqrt{\eta_{ij}^B}\}. \quad (19)$$

The maximization of $H(p||q)$ by Eq.(18) is implemented by the following Ying-Yang alternation:

Yang step: get $Y^* = \arg\max_Y L(X_N, Y, \theta)$ by solving the root Y^* of the following Sylvester equation:

$$VY + YW = UB\Lambda_c^{old},\ \text{ with } V = (\Lambda_r^{old} A^{old\mathrm{T}} D_\sigma^{old-1} A^{old})^{-1}, \quad (20)$$

$W = B^{old\,T} B^{old} \Lambda_c^{old}$, $\mathrm{U} = (A^{old\mathrm{T}} D_\sigma^{old\,-1} A^{old})^{-1} A^{old\mathrm{T}} D_\sigma^{old\,-1} X$,

which can be solved by one of the existing techniques.

Ying step: update each part of $\theta=\{A, B, \theta_y, \theta_e\}$ by

$$\delta\theta = \theta^{new} - \theta^{old} \propto \nabla_\theta H(p\|q),$$

which consists of

$$\begin{aligned}
&\delta_Y = Y^* - W^{old} X_N,\ \mathrm{E} = X_N - A^{old} Y^* B^{old\,T},\ \delta_Y = Y^* - W^{old} X_N,\\
&\delta W = \delta_Y X_N^T - D_\sigma^{new} \Omega_W,\ \delta A = E B^{old} Y^{*T} + A^{old} \delta_Y B^{old\,T} B^{old} \delta_Y^T - D_\sigma^{new} \Omega_A,\\
&\delta B = \mathrm{E} D_\sigma^{new\text{-}1} A^{old} Y^* - B^{old} \delta_Y^T A^{old\,T} D_\sigma^{new\text{-}1} A^{old} \delta_Y - \Omega_B,\\
&\Omega_A = \{a_{ij} / \eta_{ij}^A\},\ \Omega_B = \{b_{ij} / \eta_{ij}^B\},\ \Omega_W = \{w_{ij} / \eta_{ij}^W\},\\
&\delta D_\sigma = diag[(N+1)^{-1}(EE^T + A^{old} \delta_Y B^{old\,T} B^{old} \delta_Y^T A^{old\,\mathrm{T}}) - D_\sigma^{old}],\\
&\delta \Lambda_r = diag[(M+1)^{-1}(Y^* \Lambda_c^{old\,-1} Y^{*T} + \delta_Y \Lambda_c^{old\,-1} \delta_Y^T) - \Lambda_r^{old}],\\
&\delta \Lambda_c = diag[(m+1)^{-1}(Y^{*T} \Lambda_r^{old\,-1} Y^* + \delta_Y^T \Lambda_r^{old\,-1} \delta_Y) - \Lambda_c^{old}],\\
&\delta a_{ij} = a_{ij}^2 - 2\eta_{ij}^A,\ \delta b_{ij} = b_{ij}^2 - 2\eta_{ij}^B,\ \delta w_{ij} = w_{ij}^2 - 2\eta_{ij}^W.
\end{aligned} \tag{21}$$

during which we delete

- the ℓth row of Y and the ℓth column a_ℓ of A if its corresponding $\lambda_r^{(\ell)} \|a_\ell\|^2 \to 0$, (22)
- the ℓth column of Y and the ℓth column b_ℓ of B if its corresponding $\lambda_c^{(\ell)} \|b_\ell\|^2 \to 0$, (23)

where $\lambda_r^{(\ell)}, \lambda_c^{(\ell)}$ is the ℓth diagonal element of Λ_r, Λ_c respectively.

As addressed in Sect.4.2 of [13], the least complexity nature of the BYY harmony learning will enforce Eq.(22) when the ℓth row of Y is extra and enforce Eq.(23) when the ℓth column of Y is extra. That is, automatic model selection happens during the implementation of the above Ying-Yang alternation .

For the structure by Eq.(4), we modify Eq.(14) into

$$\begin{aligned}
&q(Y_B \mid \theta_y) = N(Y_B \mid 0,\ \Omega_B, \Lambda_r),\ \Omega_B = B \Lambda_c^{old} B,\\
&q(X \mid Y, \theta_{x|y}) = N(X \mid AY_B, I, D_\sigma).
\end{aligned} \tag{24}$$

For a shared use of equations (18)-(23), we subsequently drop the subscript of $\mathrm{Y_B}$ whenever it will not cause confusion.

Using Eq.(24), Eq.(17), Eq.(18), and Eq.(20), we accordingly have the following modifications:

$$\begin{aligned}
&L(X_N, Y, \theta) = \ln[N(X_N \mid AY, D_\sigma, I) N(Y \mid 0, \Omega_B, \Lambda_r) q(\theta \mid \mathbf{k})],\\
&\Pi_\theta = \Lambda_r^{-1} \otimes \Omega_B^{-1} + (A^T D_\sigma^{-1} A) \otimes I.\\
&vec^T(\delta_Y) \Pi_\theta vec(\delta_Y) = Tr[\Omega_B^{-1} \delta_Y^T \Lambda_r^{-1} \delta_Y] + Tr[\delta_Y^T A^T D_\sigma^{-1} A \delta_Y].
\end{aligned} \tag{25}$$

Though it follows from Eq.(9) that two formulations are equivalent in term of the maximum likelihood learning, the formulation by Eq.(14) is different from the formulation by Eq.(24) in term of the BYY harmony learning, with a further comparison to be addressed elsewhere.

IV. A UNIFIED LEARNING FRAMEWORK FOR CROSS-ROW DEPENDENCE AND CROSS-COLUMN DEPENDENCE

Eq.(3) provides a unified framework for modeling a $d \times N$ sample matrix X with cross-row dependence described by A and cross-column dependence described by B. Either or both of $A \neq I$ and $B \neq I$ features different types of learning models.

One most widely studied family is featured by $A \neq I$ and $B = I$, e.g., Eq.(1) in Sect. I and summarized on a roadmap in Fig.1 of [1]. It covers not only factor analysis (FA) and various nonGaussian extensions but also mean square error clustering and Gaussian mixture when each column of Y is a class label vector (i.e., only one element is 1 while all the others are zeros), with details referred to Sect.3.2 of [1].

By Eq.(3) with $A \neq I$ and $B \neq I$, all the models of the above family can be extended to cover certain type of cross-column dependence by one structure $B \neq I$. First, when $\mathrm{D}_\sigma = \sigma^2 I$, $A^T A = I$, and $B^T B = I$, from Eq.(1) & Eq.(2) we observe

$$\begin{aligned}
&E[XX^T] = E[AYY^T A^T] + E[EE^T] = A\Lambda_r A^T + \sigma^2 I,\\
&E[X^T X] = E[BY^T Y B^T] + E[E^T E] = B\Lambda_c B^T + I,\\
&E[A^T XX^T A] = \Lambda_r + \sigma^2 I,\ E[B^T X^T XB] = \Lambda_c + I.
\end{aligned} \tag{26}$$

Maximizing $E[A^T XX^T A]$ leads us to the traditional PCA for getting A to make $A^T X$, while maximizing $E[B^T X^T XB]$ leads to the 2D-PCA for getting B to make XB [14], as well as to the 2-directional 2D-PCA by using such obtained A, B to make $A^T XB$ [15]. The semi-blind BMS learning extends these studies with not only A, B estimated jointly by either the maximum likelihood learning or the BYY harmony learning, but also the mapping $X \to Y^*$ made by Eq.(20) subject to $A^T A = I$ and $B^T B = I$.

Second, beyond either one or ones of $\mathrm{D}_\sigma = \sigma^2 I$, $A^T A = I$, and $B^T B = I$, the semi-blind BMS learning can be regarded as an extension of factor analysis by Eq.(1) into a bi-directional linear model by Eq.(3) for getting the matrix-formatted inner factors Y that describes the *motifs* of X with a greatly reduced redundancy, which may be named as the bi-perspective FA. Also, we get a bi-perspective nonGaussian FA when the factors by Eq.(5a) are nonGaussians. Particularly, when σ^2 is very small, it can be regarded as a bi-perspective extension of independent factor analysis (ICA).

Third, the semi-blind BMS learning by Eq.(25) also relates to and further extends manifold learning [4-6]. With $p(Y|X, \theta)$ by Eq.(17) simply degenerated into $\delta(Y\text{-}WX)$ and thus $\delta_Y = 0$ in Eq.(25), we let $\Lambda_r = I$ and denote $L = (B\Lambda_c B^T)^{-1}$, from which maximizing $H(p||q)$ becomes equivalent to maximizing

$$L(X_N, Y, \theta) = \ln[N(X_N \mid AY, D_\sigma, I) q(\theta \mid \mathbf{k})] + \ln N(Y \mid 0, I, L^{-1}), \tag{27}$$

while it follows the notation in Eq.(15) that the maximization of the second term with respect to W is further equivalent to

$$\min_W \mathrm{Tr}[\mathrm{Y^T L Y}]_{Y=WX}, \tag{28}$$

which leads to the Laplacian eigenmaps for manifold learning [5] if L is given by the graph Laplacian. Moreover, the first term in Eq.(27) tends to improve the Laplacian eigenmaps by letting its reconstruction AY to be close to data and the estimated parameters θ to be a regularized via a priori $q(\theta|\mathbf{k})$.

Beyond, Eq.(25) also provides the following new features :

(1) As $\Lambda_r \neq 0$ is determined during the learning of Eq.(20) and Eq.(21) with help of Eq(22), the least complexity nature of the BYY harmony learning will determine the dimension of manifold during learning.

(2) With $p(Y|X, \theta) \neq \delta(Y\text{-}WX)$, learning is also regularized by the second term of $H(p||q)$ in Eq.(18) to make learning progress more balanced among different unknown parts.

Not all types of cross-column dependence are covered by $Y_B = YB^T$, *e.g.,* a temporal dependence $y_t = By_{t-1} + \varepsilon_t$. As shown by Eq.(95) in [1], this dependence can be rewritten into

$vec(Y_B)=Bvec(Y)$, but it becomes $Y=B^TYI_{-1}+E$, where I_{-1} is a matrix with its first lower diagonal elements being 1 while all the other elements being zeros. However, we may regard $Y \approx YI_{-1}$ and thus $Y=(I-B^T)^{-1}E$, when the column number of Y is large. In other words, the semi-blind BMS learning provides an alternative to make temporal factor analysis approximately.

V. TYPICAL TASKS OF APPLICATIONS

Modeling-analyses: 2D-factors, key-features, & time-structure By Eq.(3) and Eq.(4), data matrix X is modeled in three parts Y, A, and B. First, we get the matrix-formatted inner factors Y to describe the *motifs* of X for subsequent analyses. Second, A describes a cross-row dependence of X, from which we identify not only those major rows as key features but also how these features are generated from which parts of 2*D-factors*. Moreover, B describes a cross-column dependence or temporal structure of X. Third, based on a semi-blind BMS by Eq.(3) and Eq.(4) as a whole, we may also perform estimation of missing parts and prediction of new samples.

Pattern classification and object recognition A pattern classification task refers to a classification of a sample X into one of classes C_j, $j=1,...,k$, where a sample X could be an image or a sub-image. With redundancy reduction by $X \rightarrow Y$, a classifier is obtained from samples of Y by one of the existing methods. Also, we may individually learn the semi-blind BMS from training samples of each class C_j, that is, rewriting the distributions in Eq.(14) with the class label added to each parameter set as a superscript, we estimate the set $\theta^{(j)}$ of

$$q(Y|\theta_y^{(j)}),\ q(X|Y,\theta_{x|y}^{(j)}),\ \text{or}\ q(X|\theta^{(j)})=\int q(X|Y,\theta_{x|y}^{(j)})q(Y|\theta_y^{(j)})dY. \quad (29)$$

and classify X into the j^*th class by

$$j^*=\begin{cases} Max_j\ \ln[\alpha^{(j)}q(X|\theta^{(j)})], & \text{choice (a)},\\ Max_j\ \ln[\alpha^{(j)}q(Y|\theta_y^{(j)})q(X|Y,\theta_{x|y}^{(j)})], & \text{choice (b)}, \end{cases} \quad (30)$$

where $\alpha^{(j)}$ is the proportional priori of the class C_j.

Fault detection and disease diagnosis The problems consider samples from two classes or called two populations. Samples of one class C_0 come from a normal population, while samples of the other class C_1 deviate from the normal, usually referred as *fault* in engineering fields or as *disease* in medical diagnosis. Based on both samples of C_0 and samples of C_1, the task is to test whether the population C_1 is significantly different from the population C_0, and which part of information from $X^{(l)}$ is responsible for this difference.

To be more specific, we consider

$$X^{(\ell)}=[x_1^{(\ell)},...,x_{N_\ell}^{(\ell)}],\ \ell=0,1,$$
$$x_t^{(\ell)}=[x_{1,t}^{(\ell)},...,x_{d,t}^{(\ell)}]^T \text{ from } q(x_t|\theta^{(\ell)}),\ x_t=[x_{1,t},...,x_{d,t}]^T, \quad (31)$$

the tasks typically consist of

(1) Estimate $\hat{\theta}^{(0)},\hat{\theta}^{(1)}$ from $X^{(0)},X^{(1)}$, (32a)

Usually by a maximum likelihood learning. Alternatively, we use the BYY learning algorithm in Sect.III.

(2) Test the following null hypothesis

$$H_0:\ q(x_t|\hat{\theta}^{(1)}) \text{ is not different from } q(x_t|\hat{\theta}^{(0)}), \quad (32b)$$

in a sense that the probability of rejecting this hypothesis is less than a value p (usually called p-value) when both $X^{(0)}$ and $X^{(1)}$ come from the same distribution $q(x_t|\theta^{(0)})$.

(3) Find out which part of information is responsible for the rejection, e.g., which rows of $X^{(l)}$. (32c)

Typically, $q(x_t|\theta^{(0)})$ and $q(x_t|\theta^{(1)})$ share a same function form $q(x_t|\theta)$ with the difference coming from parameters, and thus the hypothesis by Eq.(32b) could be replaced by

$$H_0:\hat{\varphi}^{(1)}=\hat{\varphi}^{(0)},\quad \varphi\subseteq\theta. \quad (33)$$

A typical example of the subset φ is the mean $m=\int x_t q(x_t|\theta)dx_t$.

Usually, the columns of $X^{(l)}$ are assumed to be independent, which is helpful for implementing Eq.(32a). When the rows of $X^{(l)}$ are also independent, we have

$$q(x_t|\hat{\theta}^{(\ell)})=\prod_j q(x_{j,t}|\hat{\theta}_j^{(\ell)}),\ \hat{\theta}^{(\ell)}=\{\hat{\theta}_j^{(\ell)}\}. \quad (34)$$

The test of the hypothesis by Eq.(32b) is thus decomposed into testing everyone of the following hypotheses:

$$H_0^{(j)}:\ q(x_{j,t}|\hat{\theta}_j^{(1)}) \text{ is not different from } q(x_{j,t}|\hat{\theta}_j^{(0)}),$$
$$or\ \ H_0^{(j)}:\ \hat{\varphi}_j^{(1)}=\hat{\varphi}_j^{(0)},\quad \varphi_j\subseteq\theta_j. \quad (35)$$

VI. YAND DOMAIN TEST VERSUS YIND DOMAIN TEST

A mathematical formulation is needed to implement the hypothesis tests, for which we need a statistics s to measure the discriminative power such that $s=0$ when two populations are same while $s\geq 0$ increases as it deviates from H_0. As shown in Fig.1(a), the p-value indicates the shadow area and becomes significantly small as s goes beyond a threshold.

For H_0 by Eq.(32b), one such statistics is given as follows:

$$s=KL(q(x_t|\hat{\theta}^{(1)})\|q(x_t|\hat{\theta}^{(0)})),\ \text{KL}(p\|q)=\int p(x)\ln\frac{p(x)}{q(x)}dx. \quad (36a)$$

When the rows of $X^{(l)}$ are independent, we further have the following additive decomposition:

$$s=\sum_j s^{(j)},\ s^{(j)}=KL(q(x_{j,t}|\hat{\theta}_j^{(1)})\|q(x_{j,t}|\hat{\theta}_j^{(0)})). \quad (36b)$$

For two Gaussian populations, we have

$$s^{(j)}=\frac{s_f^{(j)}-\ln\rho_j-1}{2},\ s_f^{(j)}=\frac{\sigma_j^{(1)2}+(\mu_j^{(0)}-\mu_j^{(1)})^2}{\sigma_j^{(0)2}},\ \rho_j=\frac{\sigma_j^{(1)2}}{\sigma_j^{(0)2}}, \quad (36c)$$
$$\text{for } q(x_{j,t}|\hat{\theta}_j^{(\ell)})=G(x_{j,t}|\mu_j^{(\ell)},\sigma_j^{(\ell)2}),\ \ell=0,1.$$

However, we lost the additive decomposition when the rows of $X^{(l)}$ are not independent. For which the counterpart of Eq.(36c) becomes

$$\text{For } q(x_t|\hat{\theta}^{(\ell)})=G(x_t|\mu^{(\ell)},\Sigma^{(\ell)}),\ l=0,1,$$
$$s=0.5(s_f-\ln\rho-1),\ \rho=|\Sigma^{(1)}|/|\Sigma^{(0)}|, \quad (36d)$$
$$s_f=\text{Tr}[\{\Sigma^{(1)}+(\mu^{(0)}-\mu_j^{(1)})(\mu_j^{(0)}-\mu^{(1)})^T\}\Sigma^{(0)-1}],$$

where the covariance $\Sigma^{(l)}$ may have too many free parameters to estimate, which will cause an over-fitting problem when d is not small and N is not large.

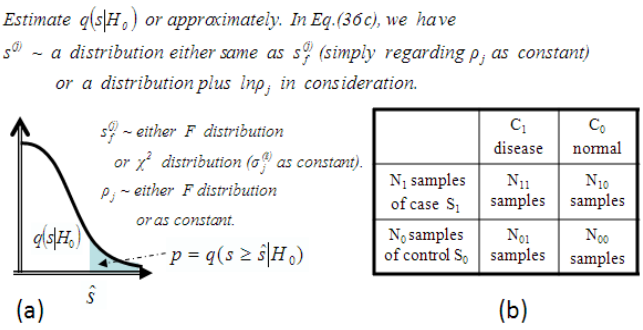
One effort towards such a problem is *Fisher discriminant analysis* (FDA). Being different from all the above hypotheses that consider samples of two populations in the domain of the observations $X^{(l)}$, FDA maps the columns of $X^{(l)}$ by $y_t=w_t^Tx_t$ into the one dimensional scalars along the axis of the vector w_t such that the following Fisher discriminant is maximized:

$$s^y=(\mu_y^{(0)}-\mu_y^{(1)})^2/(\sigma_y^{(0)2}+\sigma_y^{(1)2}) \text{ for } G(y_t|\mu_y^{(\ell)},\sigma_y^{(\ell)2}),\ell=0,1. \quad (37a)$$

Then, we use s^y to replace s in Fig.1(a) to get the p-value for testing the following hypothesis :

$$H_0: G(y_t \mid \mu_y^{(1)}, \sigma_y^{(1)2}) \text{ is not same as } G(y_t \mid \mu_y^{(0)}, \sigma_y^{(0)2}), \quad (37b)$$

which can be tested much easier than the one by Eq.(32b).



	C_1 disease	C_0 normal
N_1 samples of case S_1	N_{11} samples	N_{10} samples
N_0 samples of control S_0	N_{01} samples	N_{00} samples

Fig.1 Statistical test and confusion table

In Sect.6.2 of [13], the hypothesis H_0 by Eq.(32b) is also mapped into one that is much simple to test, via pattern classification of a two class problem. A confusion table in Fig.1(b) is obtained by a classifier. Then, testing H_0 by Eq.(32b) is mapped into testing the following one :

$$H_0: N_{11} = N_{10} \text{ and } N_{01} = N_{00}. \text{ jointly} \quad (37c)$$

This confusion table testing (CTT) approach shares with FDA by a common sprit. That is, mapping samples into much simplified inner representations with the mapped samples of two populations kept to be best discriminative, such that the hypothesis H_0 by Eq.(32b) is also mapped into an equivalent one that is much simple to test. Instead of a linear mapping $y_t=w_t^T x_t$ by FDA to maximize s^y, a classifier makes a nonlinear mapping of the samples of $X^{(0)}$ and $X^{(1)}$ into binary labels C_0 and C_1, with discriminative power best kept via training a classifier with the minimum classification error, i.e., towards $N_{10}=0$ and $N_{01}=0$. Instead of testing H_0 by Eq.(37b) via s^y, the hypothesis H_0 by Eq.(37c) is tested by a statistics that verifies whether a confusion table is significantly deviated.

In this paper, we further propose a better alternative of H_0 by Eq.(37c). Noticing that $X^{(1)}$ *or* $q(x_t|\theta^{(1)})$ is mapped into the first row p_1 =[N_{11}/ N_1,N_{01}/ N_1] of the table in Fig.1(b) and that $X^{(0)}$ or $q(x_t|\theta^{(0)})$ is mapped into the second row p_0 =[N_{01}/ N_0,N_{00}/ N_0]. the hypothesis H_0 by Eq.(37c) is improved into

$$H_0: p_1 \text{ is not different from } p_0. \quad (38)$$

for which we may simply make Pearson chi-squared test or the Kolmogorov–Smirnov test.

FDA and CTT are complementary. For Gaussian samples, FDA is favorable. For other populations, FDA could be used approximately. We may also develop new methods from Eq.(36b). Anyway, CTT provides an easy implementation to tackle the problems since there are already many existing methods for training classifiers.

Even generally, both FDA and CTT are two examples of another type of hypothesis tests, which are complementary with the conventional type such as the hypotheses by Eq.(32b), Eq.(33), and Eq.(35). From the perspective of the BYY system introduced in Sect.II, we can summarize them into two types:

***Yang domain test* (A-test)**: made on observations or Yang domain with help of Ying transform, e.g., $X^{(0)}$, $X^{(1)}$ are examples of $\mathbf{X}$, and $q(x_t|\theta^{(0)})$, $q(x_t|\theta^{(1)})$ are examples of q($\mathbf{X}$|$\mathbf{R}$).

***Ying domain test* (I-test)**: made on inner representations or Ying domain with help of Yang transform, e.g., $\mathbf{R}$ includes y_t, C_0 , C_1, and p($\mathbf{R}$|$\mathbf{X}$) covers $y_t=w_t^T x_t$ and classifiers..

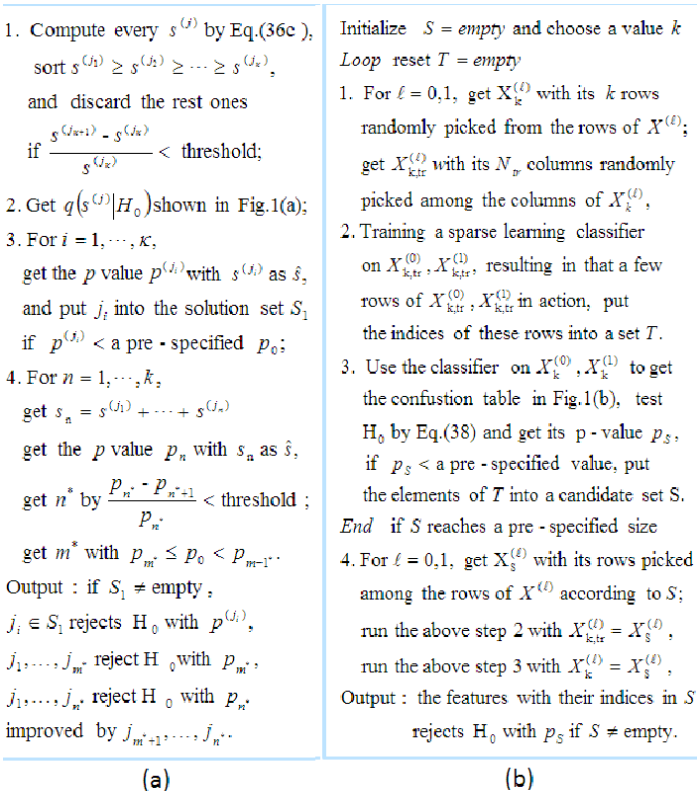


Fig.2 Two schematic algorithms
(a) A-test is used when the independence by Eq.(34) holds,
(b) I-test is used on data matrices with cross row dependences.

In implementation, A-test is featured by getting the p-value based on the estimated model parameter θ from a training set X_{TR} of samples by Eq.(32a). With a testing set X_{TE} of samples available, θ needs to be updated to adapt X_{TE} for getting a updated p-value. Such a p-value could be too optimistic when θ is estimated from a small size of X_{TR}.

For an I-test, we estimate the parameter θ of a transform T=T(θ) from observations to Ying domain and then get the p-value based on the obtained T. There are two choices for getting the p-value on X_{TR}, X_{TE}. With T_{TR} obtained from X_{TR}, we use it to map X_{TR} into Ying domain to get p_{TR} and map X_{TE} to get p_{TE}. Similar to A-test, we may also update T to adapt X_{TE} and then map X_{TE} to get another p-value. Though p_{TE} could be much less significant than p_{TR}, it could be also an indicator for evaluating the test. E.g., CTT gets, from X_{TR}, X_{TE} respectively, different versions of the table in Fig.1(b). For FDA without adapting X_{TE}, $y_t=w_t^T x_t$ obtained merely from X_{TR} may also be used on X_{TE} to get p_{TE}, though the choice is seldom considered.

The schematic algorithm in Fig.2(a) relies on the

independence by Eq.(34) to consider $s_i^{(l)}$ in a descending order for an easy implementation of A-test. If eq.(34) does not hold, we have not this easy implementation. Conceptually, a best subset of the columns of $X^{(l)}$ could be searched via evaluating s by Eq.(36d), which can be made by a forward-backward selection similar to statistical stepwise regression. Even so, such a procedure can not guarantee a full enumeration of all combinatorial choices for a best subset.

Alternatively, we may make I-test for easy implementation by the schematic algorithm in Fig.2(b), via learning sparse classifier (e.g., sparse Fisher discriminative analysis).

Also, we may examine A-test and I-test jointly. Instead of using an existing approach for either making Eq.(32a) or training a classifier, the above Ying-Yang perspective further suggests us to make the BYY harmony learning to get a BYY system from which we obtain not only $q(\mathbf{X}|\mathbf{R})q(\mathbf{R})$ for $q(x_t|\theta^{(l)})$ via Eq.(29) but also $p(\mathbf{R}|\mathbf{X})$ for $y_t=w_t^T x_t$ and classifiers.

VII. Gene Analysis Applications

Gene regulatory and network component analysis Regulatory signal reconstruction is formulated under the name network component analysis (NIC) [16-18] that considers the system by Eq.(1), with X being gene expression profiles and Y being the unknown transcriptional factor activities (TFA) while $A=\{a_{ij}\}$ describing the transcription regulatory network with a given network topology, in that $a_{ij}=0$ if there is no connection from the *i-th* transcriptional factor to the *j-th* gene. The rest of unknowns in A and also the TFA activities Y are estimated via minimizing the norm $\|E\|_2$.

This NIC has shortcomings. First, the problem is usually indeterminable except some special cases. Second, a given network topology is generally a rough approximation. Third, the transcriptional factors chosen as the rows of Y come from finding those proteins binding to specific DNA sequences that control transcriptions from DNA to gene expressions. But not necessarily all the binding proteins are really in action, while NIC treats them indifferently. Fourth, we already know some controlling topology about how the binding proteins act in a complex, while NIC fails to use this information.

The semi-blind BMS learning by Eqs.(3)&(4) provides a framework towards these problems. The system indeterminacy is reduced by adding a distribution structure on Y. Also, rigidly shutting off a connection by $a_{ij}=0$ is replaced by a sparse learning via $q(A,\{\eta^A_{ij}\})$ in Eq.(19) with its corresponding η^A_{ij} initialized by a small value. Moreover, we may detect whether a binding factor does not in action via Eq.(22). Furthermore, transcriptional topology is taken in consideration via either letting $L=(B\Lambda_c B^T)^{-1}$ in Eq.(25) by graph Laplacian or treating B similarly to A by sparse learning.

Genetic disease diagnosis Given $X^{(0)}$ coming from normal individuals and $X^{(1)}$ coming from individuals with a type of disease, and each row of $X^{(l)}$ labeling a gene and columns indicating different individuals, we encounter a typical fault detection task as introduced in Sect. VI for finding which genes associate the disease. When the expression of each gene is regarded as independent from ones of other genes, we can handle the problem by the schematic algorithm in Fig.2(a).

Genome-wide association studies (*GWAs*) With each row labeling a SNP instead of a gene, we may also use the schematic algorithm given in Fig.2(a) to find multiple SNPs that cause a type of disease.

Exome sequencing analysis With each column vector of $X^{(l)}$ representing each exon sequence, we may test whether the exon associates with a type of disease, implemented by either the schematic algorithm in Fig.2(b) or even the one in Fig.2(a) under the assumption of the row independence by Eq.(34).

Acknowledgements This work was supported by RGC Direct grant project 2050502, and the National Basic Research Program of China (973 Program) (No. 2009CB825404).

References

[1] Xu L. "Codimensional matrix pairing perspective of BYY harmony learning: Hierarchy of bilinear systems, joint decomposition of data-covariance, and applications of network biology," A special issue on Machine Learning and Intelligence Science: IScIDE2010 (A). Frontiers of Electrical and Electronic Engineering in China, 2011, 6(1): 86–119.

[2] Tu S K, Chen R S, Xu L. A binary matrix factorization algorithm for protein complex prediction. Proteome Science, 2011, 9(Suppl 1): S18.

[3] Lee, D.D. & Seung, H.S., Learning the parts of objects by non-negative matrix factorization, Nature, 1999, 401: 788–791.

[4] Roweis, ST & Saul, LK., Nonlinear dimensionality reduction by locally linear embedding, Science, 2000, 290 (5500): 2323-2326.

[5] Belkin,M & Niyogi,P, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Computation, 2003, 15(6): 1373–1396.

[6] He, X. & Lin, B., Tangent Space Learning and Generalization, Front. Electr. Electron. Eng. China 2011, 6(1): 27–42.

[7] Williams, P.M., Bayesian regularization and pruning using a Laplace prior, Neural Computation, 1995, 7(1): 117–143.

[8] Tibshirani R. Regression shrinkage and selection via the lasso. J. Royal Statistical Society, Series B: Methodological, 1996, 58(1): 267–288

[9] Shi L, Tu S K, Xu L. Learning Gaussian mixture with automatic model selection: A comparative study on three Bayesian related approaches, Front. Electr. Electron. Eng. China 2011, 6(2): 215–244.

[10] Tu S.K., & Xu, L, Parameterizations make different model selections : empirical findings from factor analysis, Front. Electr. Electron. Eng. China 2011, 6(2): 256–274.

[11] Xu L. Bayesian-Kullback coupled YING-YANG machines: Unified learning and new results on vector quantization. In: Proc. ICONIP95, 1995, 977–988 (A further version in NIPS8. In: Touretzky D S, et al. eds. Cambridge: MIT Press, 444–450).

[12] Xu L. Bayesian Ying Yang learning. Scholarpedia, 2007, 2(3): 1809 http://scholarpedia.org/article/Bayesian Ying Yang learning

[13] Xu L. On essential topics of BYY harmony learning: Current status, challenging issues, and gene analysis applications, Front. Electr. Electron. Eng. China, 2012,7(1): 147–196.

[14] Yang J, Zhang D, Frangi, AF, & Yang JY, Two-dimensional PCA: a new approach to appearance-based face representation and recognition, *IEEE Tr. PAMI*, 2004, 26(1):131-137.

[15] Zhang D. & Zhou Z.-H., $(2D)^2$PCA: 2-directional 2-dimensional PCA for efficient face representation and recognition, *Neurocomputing*, 2005, 69(1-3): 224-231.

[16] Liao, JC, Boscolo, R, & Sabatti, C, & Roychowdhury ,VP., Network component analysis: reconstruction of regulatory signals in biological systems. Proc. Natl. Acad. Sci. USA, 2003, 100(26): 15522–15527.

[17] Lemmens, K., Dhollander, T., & De Bie T. et al., Inferring transcriptional modules from ChIP-chip, motif and microarray data., Genome Biology 2006; 7(5): R37.

[18] Youn, A., Reiss, D.J., & Stuetzle, W. , Learning transcriptional networks from the integration of ChIP–chip and expression data in a non-parametric model, Bioinformatics, 2010, 26(15): 1879–1886.