Gene clustering by structural prior based local factor analysis model under Bayesian Ying-Yang harmony learning

Lei Shi; Shikui Tu; Lei Xu;
Dept. of Comput. Sci. & Eng., Chinese Univ. of Hong Kong, Hong Kong, China

This paper appears in: Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on
Issue Date: 18-21 Dec. 2010
On page(s): 696 - 699
Location: Hong Kong
Print ISBN: 978-1-4244-8306-8
INSPEC Accession Number: 11795056
Digital Object Identifier: 10.1109/BIBM.2010.5706655
Date of Current Version: 04 February 2011

ABSTRACT

We propose a clustering algorithm based on a structural prior based Local Factor Analysis (spLFA) model under the Bayesian Ying-Yang harmony learning, which automatically determines the hidden dimensionalities during parameter learning, reduces the number of free parameters by projecting the mean vectors onto a low dimensional manifold, imposes the sparseness by a Normal-Jeffreys prior. Experiments on the diagnostic research dataset show that BYY-spLFA outperforms the k-means clustering and single-link hierarchical clustering. The experiments on a lymphoma cancer datset further indicate the BYY-spLFA is able to uncover the number of phenotypes correctly and cluster the phenotypes more accurately. In addition, we modify BYY-spLFA to implement supervised learning and preliminarily demonstrate its effectiveness on a Leukemia data for classification.

INDEX TERMS

- **INSPEC**
  - **Controlled Indexing**
    Bayes methods , bioinformatics , diseases , genetics , learning (artificial intelligence) , medical computing , patient diagnosis , pattern classification , pattern clustering
  - **Non Controlled Indexing**
    BYY-spLFA , Bayesian Ying-Yang harmony learning , Leukemia data , Normal-Jeffreys prior , data classification , diagnostic research dataset , free parameter number reduction , gene clustering , hidden dimensionalities , low dimensional manifold , lymphoma cancer datset , parameter learning , spLFA model , structural prior based local factor analysis model , supervised learning
- **Author Keywords**
  Bayesian Ying-Yang learning , feature selection , gene clustering , sparse learning , structural prior

Gene Clustering by Structural Prior based Local Factor Analysis Model under Bayesian Ying-Yang Harmony Learning

Lei Shi, Shikui Tu, and Lei Xu*

Dept. of Computer Sci. and Eng., The Chinese University of Hong Kong, Shatin, NT, Hong Kong

Email: {shil,sktu,lxu}@cse.cuhk.edu.hk, (Lei Xu: corresponding author)*

***Abstract*—We propose a clustering algorithm based on a structural prior based Local Factor Analysis (spLFA) model under the Bayesian Ying-Yang harmony learning, which automatically determines the hidden dimensionalities during parameter learning, reduces the number of free parameters by projecting the mean vectors onto a low dimensional manifold, imposes the sparseness by a Normal-Jeffreys prior. Experiments on the diagnostic research dataset show that BYY-spLFA outperforms the k-means clustering and single-link hierarchical clustering. The experiments on a lymphoma cancer datset further indicate the BYY-spLFA is able to uncover the number of phenotypes correctly and cluster the phenotypes more accurately. In addition, we modify BYY-spLFA to implement supervised learning and preliminarily demonstrate its effectiveness on a Leukemia data for classification.**

***Keywords*-gene clustering; sparse learning; Bayesian Ying-Yang learning; structural prior; feature selection**

I. Introduction

Microarray gene expression data allow us to monitor the expression of thousands of genes under different conditions qualitatively and simultaneously [1]. Identification of groups of genes with similar expression patterns is usually treated as a clustering task in unsupervised learning. Gene clustering analysis plays an important role in discovering biologically and medically meaningful modules (e.g., revealing unknown subtypes of a disease), as well as in further investigating specific pathways or genetic mechanisms.

Traditional clustering methods may not perform well on gene data that are typically high-dimensional but have a very small sample size. Local Factor Analysis (LFA) [2], which projects the data of a cluster into a low-dimensional factor manifold, has been studied under the Bayesian Ying-Yang (BYY) harmony learning with the hidden dimensionality automatically determined during parameter learning. In this paper, we further impose a structural prior on LFA (or shortly spLFA) such that the means (centers) of all factor manifolds are located on another low-dimensional manifold. With this structural constraint, the number of free parameters can be further reduced. Moreover, we adopt the joint Normal-Jeffreys (NJ) prior studied in [3], [4] to those orthogonal vectors that span the factor manifolds and the low-dimensional mean vector manifold. Then, a learning algorithm is developed to implement the Bayesian Ying-Yang (BYY) harmony learning, and thus is named as BYY-spLFA. During learning, Not only the dimensions of all the manifolds are automatically determined during learning, but also extra parameters are shrunken to zero (or ineffective)by the NJ prior that assumes a Normal prior over a parameter and then a Jeffreys prior over the variance of the Normal prior, with advantages of no sparsity strength controlling hyper-parameters and a better sparsity realization over a Laplacian prior as shown in [4].

We test BYY-spLFA on the diagnostic research dataset of small round blue-cell tumors of childhood [5] by evaluating the values of the Rand index [6] of the obtained clusters. The results show that BYY-spLFA obtains higher Rand index values than two widely used clustering methods, k-means and single-link hierarchical clustering. Moreover the clustering results on gene expression profiles of a lymphoma cancer dataset [7] indicates BYY-spLFA not only correctly detects the number of phenotypes from samples directly without human expertise, but also gives a better clustering accuracy and Rand index than k-means and single-link hierarchical clustering. Furthermore, BYY-spLFA is modified appropriately to implement supervised learning, and its effectiveness in classification is preliminarily demonstrated on a Leukemia dataset.

II. Model Specification

A structural prior based Local Factor Analysis model describes the observation distribution into a mixture of k components (clusters) with the mixing weights $\{\alpha_l\}_{l=1}^k$, $\forall \alpha_l > 0$ and $\sum_{l=1}^k \alpha_l = 1$. In each component $l = 1, \ldots, k$, the observable variable $\mathbf{x} \in R^d$ is generated via a linear mapping from an m_l-dimensional independent Gaussian factor $\mathbf{y}$ with $m_l < d$, plus mean values and a Gaussian noise which is independent of the factors[1]:

$$\begin{aligned} & q(\mathbf{x}|\mathbf{y}, l, \boldsymbol{\Theta}) = \mathcal{N}(\mathbf{x}|\mathbf{n} + \mathbf{V}\boldsymbol{\mu}_l + \mathbf{U}_l\mathbf{y}, \boldsymbol{\Psi}_l), \\ & q(\mathbf{y}|l) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \boldsymbol{\Lambda}_l), \quad q(\{\boldsymbol{\mu}_l\}) = \prod_{l=1}^{k} \mathcal{N}(\boldsymbol{\mu}_l|\mathbf{0}, \boldsymbol{\Gamma}), \\ & \text{s.t.} \ \ \mathbf{V}^T\mathbf{V} = \mathbf{I}_r, \quad \mathbf{U}_l^T\mathbf{U}_l = \mathbf{I}_{m_l} \ \ \text{for} \ \ \forall l. \end{aligned} \tag{1}$$

where each cluster l forms a factor manifold, with $\mathbf{U}_l$ being a $d \times m_l$ orthogonal loading matrix, $\boldsymbol{\Lambda}_l$ being an $m_l \times m_l$ diagonal covariance, and $\boldsymbol{\Psi}_l$ is a $d \times d$ diagonal noise covariance. The spLFA model is intuitively illustrated in Fig. 1. If $r = d$ and $\mathbf{V}$ is an identity matrix, the spLFA becomes the traditional Local Factor Analysis (LFA) [8].

[1] Here and throughout this paper, $q(\cdot)$ stands for a generative probability and $p(\cdot)$ stands for a posterior type probability.

978-1-4244-8305-1/10/$26.00 ©2010 IEEE

The number $(d-0.5r+k)(r+1)-k$ of free parameters of the centers of spLFA after projections by $\{\mathbf{n}+\mathbf{V}\boldsymbol{\mu}_l\}_{l=1}^k$ is smaller than $d \times k$ of the LFA's centers $\{\boldsymbol{\mu}'_l\}_{l=1}^k$ for a very small r. Moreover, a Gaussian prior is usually assigned on mean vectors in the literature [9], which however falls into a dilemma on determining the covariance's structure of Gaussian. On one hand, a full covariance matrix for LFA's centers $\{\boldsymbol{\mu}'_l\}_{l=1}^k$ has $d(d-1)/2$ hyper-parameters and is difficult to be determined appropriately for a large d. On the other hand, a diagonal or spherical covariance matrix may be too constrained to describe the data accurately. This problem is here overcome by assuming the centers $\{\boldsymbol{\mu}_l\}$ *a priori* distributed in an r-dimensional manifold ($r < d$), which is spanned by $\mathbf{V}$ and dispersed with a diagonal covariance $\mathbf{\Gamma}$ as given in Eq. (1), where $\mathbf{V}$ and $\mathbf{\Gamma}$ in total have $(2d-r+1) \times r/2$ free hyper-parameters to be determined.

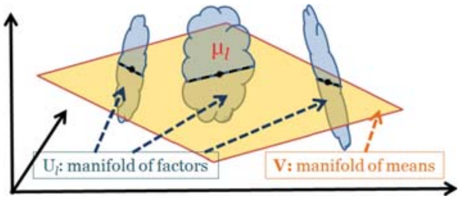

Figure 1. Illustration of structural prior based Local Factor Analysis. In each component l, the factor manifold is drawn in blue, and its origin $\boldsymbol{\mu}_l$ further distributes in the yellow colored manifold.

The matrices $\mathbf{V}$ and $\{\mathbf{U}_l\}$ still have many free parameters when d is large, we further impose sparsity on them with feature selection made automatically during learning, in help of the following Normal-Jeffreys prior [3], [4]:

$$\begin{aligned}
&q(\mathbf{V},\boldsymbol{\gamma}) = \prod_{i,j}[q(v^{(ij)}|\gamma^{(ij)})q(\gamma^{(ij)})],\\
&q(v^{(ij)}|\gamma^{(ij)}) = \mathcal{N}(v^{(ij)}|0,\gamma^{(ij)}), \quad q(\gamma^{(ij)}) \propto 1/\gamma^{(ij)},\\
&q(\mathbf{U}_l,\boldsymbol{\tau}_l) = \prod_{i,j}[q(u_l^{(ij)}|\tau_l^{(ij)})q(\tau_l^{(ij)})],\\
&q(u_l^{(ij)}|\tau_l^{(ij)}) = \mathcal{N}(u_l^{(ij)}|0,\tau_l^{(ij)}), \quad q(\tau_l^{(ij)}) \propto 1/\tau_l^{(ij)}, \qquad (2)
\end{aligned}$$

where a Normal prior is considered with its variance further in a Jeffreys prior. Instead of having explicit hyper-parameters to control (e.g., in a Laplacian prior), the sparseness strength of $\mathbf{V}$ and $\{\mathbf{U}_l\}$ is adjusted by $\boldsymbol{\gamma}$ and $\{\boldsymbol{\tau}_l\}$ [4]. Particularly, an element $v^{(ij)}$ in $\mathbf{V}$ approaches zero as the corresponding $\gamma^{(ij)}$ approaches zero. Also, it is similar for each $\mathbf{U}_l$.

III. BYY Harmony Learning on spLFA

Given a finite size of samples, the learning tasks on the spLFA model consisting of making sparse learning on the parameters and determining r for the mean vector manifold and $\{m_l\}$ for each factor manifold. For these purposes, we develop a learning algorithm under the BYY harmony learning framework, enhanced with the Eq. (2) for imposing sparsity.

For a set of i.i.d. observations $\mathbf{X}_N = \{\mathbf{x}_t\}_{t=1}^N$ generated from the hidden representation $\{\mathbf{Y},\mathbf{Z}\}$, we consider the BYY system as shown in Box I and Box II, where the parameters are divided into two parts $\mathbf{\Theta} = \mathbf{\Theta}_a \cup \mathbf{\Theta}_b$, with $\mathbf{\Theta}_a = \{\mathbf{n},\{\mathbf{\Lambda}_l,\mathbf{\Psi}_l\}\}$ and $\mathbf{\Theta}_b = \{\mathbf{V},\{\boldsymbol{\mu}_l,\mathbf{U}_l\}\}$, and $\mathbf{\Xi} = \{\mathbf{\Gamma},\boldsymbol{\gamma},\{\boldsymbol{\tau}_l\}\}$. Also, we have a set $\mathbf{\Theta} = \mathbf{\Theta}_a \cup \mathbf{\Theta}_b$ of hyper-parameters. Moreover, $\mathbf{Z} = \{z_{it}\}$ is a set of binary variables for $i = 1,\ldots,k$ and $t = 1,\ldots,N$, with each $z_{lt} \in \{0,1\}$, $\sum_{l=1}^k z_{lt} = 1$ and $z_{lt} = 1$ iff sample $\mathbf{x}_t$ belongs to class l.

All the unknowns in the BYY system are determined by the BYY harmony learning, which is implemented via maximizing the harmony measure given in Eq. (3), where $\mathbf{\Theta}_b^* = arg\max_{\mathbf{\Theta}_b} H(p||q,\mathbf{\Theta},\mathbf{\Xi})$ and is approximately replaced by the available estimation of $\mathbf{\Theta}_b$ in the last round of iteration. Moreover, $\mathbf{\Pi}_{\Theta_b}$ is the Hessian matrix of the harmony measure with respect to the parameters $\mathbf{\Theta}_b$. Since the exact calculation of $\mathbf{\Pi}_{\Theta_b}$ is too difficult, we approximately consider the block-diagonal structure shown in Box III. Therein, the $\mathbf{W}_l^{(j*)}$ refers to the j-th row vector of $\mathbf{W}_l$. The algorithm details of BYY-spLFA are summarized in Box IV.

Box I: components in Ying machine

$$\boxed{\begin{aligned}
&q(\mathbf{X}|\mathbf{Y},\mathbf{Z},\mathbf{\Theta}) = \textstyle\prod_{t=1}^N\prod_{l=1}^k q(\mathbf{x}_t|\mathbf{y},l,\mathbf{\Theta})^{z_{lt}},\\
&q(\mathbf{Y}|\mathbf{Z},\mathbf{\Theta}) = \textstyle\prod_{t=1}^N\prod_{l=1}^k q(\mathbf{y}|l)^{z_{lt}},\\
&q(\mathbf{Z}|\mathbf{\Theta}) = \textstyle\prod_{t=1}^N\prod_{l=1}^k \alpha_l^{z_{l,t}},\\
&q(\mathbf{\Theta}_b|\mathbf{\Xi}) = q(\mathbf{V}|\boldsymbol{\gamma})\textstyle\prod_{l=1}^k q(\mathbf{U}_l|\boldsymbol{\tau}_l),\\
&q(\mathbf{\Xi}) = q(\boldsymbol{\gamma})\textstyle\prod_{l=1}^k q(\boldsymbol{\tau}_l),\\
&\text{with details referred to Eqs. (1) and (2).}
\end{aligned}}$$

Box II: components in Yang machine

$$\boxed{\begin{aligned}
&p(\mathbf{Z}|\mathbf{X}_N) = \textstyle\prod_{t=1}^N\prod_{l=1}^k p(l|\mathbf{x}_t,\mathbf{\Theta})^{z_{lt}},\\
&p(\mathbf{Y}|\mathbf{X}_N,\mathbf{Z},\mathbf{\Theta}) = \textstyle\prod_{t=1}^N\prod_{l=1}^k p(\mathbf{y}|\mathbf{x}_t,l,\mathbf{\Theta})^{z_{lt}},\\
&\text{with } p(l|\mathbf{x}_t,\mathbf{\Theta}) = \alpha_l q(\mathbf{x}_t|l,\mathbf{\Theta})/[\textstyle\sum_{l=1}^k \alpha_l q(\mathbf{x}_t|l,\mathbf{\Theta})].,\\
&\qquad p(\mathbf{y}|\mathbf{x}_t,l,\mathbf{\Theta}) = \mathcal{N}(\mathbf{y}|\mathbf{W}_l(\mathbf{x}_t-\mathbf{n}-\mathbf{V}\boldsymbol{\mu}_l),\widetilde{\mathbf{\Lambda}}_l),\\
&\qquad \mathbf{W}_l = \mathbf{\Lambda}_l\mathbf{U}_l^T(\mathbf{U}_l\mathbf{\Lambda}_l\mathbf{U}_l^T+\mathbf{\Psi}_l)^{-1},\\
&\qquad \widetilde{\mathbf{\Lambda}}_l = (\mathbf{U}_l^T\mathbf{\Psi}_l^{-1}\mathbf{U}_l+\mathbf{\Lambda}_l^{-1})^{-1}.
\end{aligned}}$$

Box III: Hessian of parameters $\mathbf{\Pi}_{\Theta_b}$

$$\boxed{\begin{aligned}
\mathbf{\Pi}_{\Theta_b} &= \text{Block-Diag}[\mathbf{\Pi}_V,\mathbf{\Pi}_{U_1},\ldots,\mathbf{\Pi}_{U_k},\mathbf{\Pi}_{\mu_1},\ldots,\mathbf{\Pi}_{\mu_k}],\\
\mathbf{\Pi}_V &= \text{Block-Diag}[\mathbf{\Pi}_V^{(1)},\ldots,\mathbf{\Pi}_V^{(r)}],\\
\mathbf{\Pi}_{U_l} &= \text{Block-Diag}[\mathbf{\Pi}_{U_l}^{(1)},\ldots,\mathbf{\Pi}_{U_l}^{(m_l)}],\\
\mathbf{\Pi}_V^{(j)} &= \textstyle\sum_{t=1}^N\sum_{l=1}^k p(l|\mathbf{x}_t)(\mu_l^{(j)})^2[(\mathbf{I}-\mathbf{U}_l\mathbf{W}_l)^T\mathbf{\Psi}_l^{-1}(\mathbf{I}-\mathbf{U}_l\mathbf{W}_l)\\
&\quad +\mathbf{W}_l^T\mathbf{\Lambda}_l^{-1}\mathbf{W}_l] + \text{diag}[1/\gamma^{(1j)},\ldots,1/\gamma^{(rj)}],\\
\mathbf{\Pi}_{U_l}^{(j)} &= \textstyle\sum_{t=1}^N p(l|\mathbf{x}_t)\mathbf{\Psi}_l^{-1}[\mathbf{W}_l^{(j*)}(\mathbf{x}_t-\mathbf{n}-\mathbf{V}\boldsymbol{\mu}_l)]^2\\
&\quad +\text{diag}[1/\tau_l^{(1j)},\ldots,1/\tau_l^{(m_lj)}],\\
\mathbf{\Pi}_{\mu_l} &= \textstyle\sum_{t=1}^N p(l|\mathbf{x}_t)\mathbf{V}^T[(\mathbf{I}-\mathbf{U}_l\mathbf{W}_l)\mathbf{\Psi}_l^{-1}(\mathbf{I}-\mathbf{U}_l\mathbf{W}_l)\\
&\quad +\mathbf{W}_l^T\mathbf{\Lambda}_l^{-1}\mathbf{W}_l]\mathbf{V}+\mathbf{\Gamma}^{-1},
\end{aligned}}$$

Being different from the maximum likelihood, an important nature of maximizing this harmony measure in Eq. (3) leads to not only a best matching between the Ying-Yang pair, but also a compact model with a least complexity. Such an model selection ability can be understood from several perspectives [2], [8]. Specifically, maximizing the harmony measure in Eq. (3) will provide an intrinsic force to push α_l to approach zero if cluster l is extra and thus discarded. Also, $\mathbf{\Lambda}_l^{(j)}$ of a cluster l is pushed to approach zero if the corresponding j-th hidden dimension in cluster l is extra

$$
\begin{aligned}
H(p||q,\mathbf{\Xi}) &= H(p||q,\mathbf{\Theta},\mathbf{\Xi}) - 0.5\mathrm{tr}[(\mathrm{vec}(\mathbf{\Theta}_b) - \mathrm{vec}(\mathbf{\Theta}_b^*))^T \mathbf{\Pi}_{\Theta_b}(\mathrm{vec}(\mathbf{\Theta}_b) - \mathrm{vec}(\mathbf{\Theta}_b^*)] - 0.5k(\mathbf{\Theta}_b),\\
H(p||q,\mathbf{\Theta},\mathbf{\Xi}) &= \int \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}_N,\mathbf{\Theta})p(\mathbf{Y}|\mathbf{X}_N,\mathbf{Z},\mathbf{\Theta}) \ln[q(\mathbf{X}_N|\mathbf{Y},\mathbf{Z},\mathbf{\Theta})q(\mathbf{Y}_N|\mathbf{Z})q(\mathbf{Z}|\mathbf{\Theta})q(\mathbf{\Theta}_b|\mathbf{\Xi})q(\mathbf{\Xi})]d\mathbf{Y}.
\end{aligned}
\tag{3}
$$

Box IV: BYY-spLFA algorithm details

(1) for $\forall l = 1,\ldots,k$ and $t = 1,\ldots,N$, calculate:
$p_{lt} = p(l|\mathbf{x}_t)[1 + \ln p(l|\mathbf{x}_t) - \sum_{j=1}^{k} p(j|\mathbf{x}_t)\ln p(j|\mathbf{x}_t)]$,
$\mathbf{e}_{lt} = \mathbf{x}_t - \mathbf{n} - \mathbf{V}\boldsymbol{\mu}_l$, $\mathbf{W}_l = \mathbf{\Lambda}_l\mathbf{U}_l^T(\mathbf{U}_l\mathbf{\Lambda}_l\mathbf{U}_l^T + \mathbf{\Psi}_l)^{-1}$,
(2) update (hyper-)parameters:
$\alpha_l^{new} = (1-\eta)\alpha_l + \sum_{t=1}^{N} p_{lt}/N$
$\mathbf{\Psi}_l^{new} = (1-\eta)\mathbf{\Psi}_l + \mathrm{diag}[\sum_{t=1}^{N} p_{lt}(\mathbf{I}_d - \mathbf{U}_l\mathbf{W}_l)\mathbf{e}_{lt}\mathbf{e}_{lt}^T$
$\quad\cdot(\mathbf{I}_d - \mathbf{U}_l\mathbf{W}_l)^T]/\sum_{t=1}^{N} p_{lt}$,
$\mathbf{\Lambda}_l^{new} = (1-\eta)\mathbf{\Lambda}_l + \mathrm{diag}[\sum_{t=1}^{N} p_{lt}\mathbf{W}_l\mathbf{e}_{lt}\mathbf{e}_{lt}^T\mathbf{W}_l^T]/\sum_{t=1}^{N} p_{lt}$,
$\mathbf{U}_l^{new} = \mathbf{U}_l + \eta(\mathbf{G}_{\mathbf{U}_l} - \mathbf{U}_l\mathbf{G}_{\mathbf{U}_l}^T\mathbf{U}_l)$, with $\mathbf{G}_{\mathbf{U}_l} = \sum_{t=1}^{N}\mathbf{\Psi}_l^{-1}$
$\quad\cdot(\mathbf{I}_d - \mathbf{U}_l\mathbf{W}_l)\mathbf{e}_{lt}\mathbf{e}_{lt}^T\mathbf{W}_l^T - \mathbf{U}_l./\boldsymbol{\tau}_l + [\Delta_{U_l}^{(1)},\ldots,\Delta_{U_l}^{(m_l)}]$,
$\quad$ and $\Delta_{U_l}^{(j)} = \mathbf{\Pi}_{U_l}^{(j)}(\hat{\mathbf{U}}_l^{(*j)} - \mathbf{U}_l^{(*j)})$,
$\boldsymbol{\mu}_l^{new} = \boldsymbol{\mu}_l + \eta\sum_{t=1}^{N} p_{lt}\mathbf{V}^T[(\mathbf{I}_d - \mathbf{U}_l\mathbf{W}_l)^T\mathbf{\Psi}_l^{-1}(\mathbf{I}_d - \mathbf{U}_l\mathbf{W}_l)$
$\quad+\mathbf{W}_l^T\mathbf{\Lambda}_l^{-1}\mathbf{W}_l]\mathbf{e}_{lt} - \mathbf{\Gamma}^{-1}\boldsymbol{\mu}_l + \mathbf{\Pi}_{\mu}^{(l)}(\hat{\boldsymbol{\mu}}_l - \boldsymbol{\mu}_l)$,
$\mathbf{V}^{new} = \mathbf{V} + \eta(\mathbf{G}_{\mathbf{V}} - \mathbf{V}\mathbf{G}_{\mathbf{V}}^T\mathbf{V})$, with $\mathbf{G}_{\mathbf{V}} = \sum_{l=1}^{k}\sum_{t=1}^{N} p_{lt}$
$\quad\cdot[(\mathbf{I}_d - \mathbf{U}_l\mathbf{W}_l)^T\mathbf{\Psi}_l^{-1}(\mathbf{I}_d - \mathbf{U}_l\mathbf{W}_l) + \mathbf{W}_l^T\mathbf{\Lambda}_l^{-1}\mathbf{W}_l]\mathbf{e}_{lt}\boldsymbol{\mu}_l^T$
$\quad-\mathbf{V}./\boldsymbol{\rho} + [\Delta_V^{(r)},\ldots,\Delta_V^{(r)}]$, and $\Delta_V^{(j)} = \mathbf{\Pi}_V^{(j)}(\hat{\mathbf{V}}^{(*j)} - \mathbf{V}^{(*j)})$,
$\mathbf{\Gamma}^{new} = (1-\eta)\mathbf{\Gamma} + \eta\mathrm{diag}\sum_{l=1}^{k}[\boldsymbol{\mu}_l\boldsymbol{\mu}_l^T + (\boldsymbol{\mu}_l - \hat{\boldsymbol{\mu}}_l)(\boldsymbol{\mu}_l - \hat{\boldsymbol{\mu}}_l)^T]/k$,
$\boldsymbol{\tau}_l^{(*j)\,new} = (1-\eta)\boldsymbol{\tau}_l^{(*j)} + \eta\{(\mathbf{U}_l^{(*j)})^2 + \mathrm{diag}[(\mathbf{U}_l^{(*j)} - \hat{\mathbf{U}}_l^{(*j)})$
$\quad\cdot(\mathbf{U}_l^{(*j)} - \hat{\mathbf{U}}_l^{(*j)})^T]\}$,
$\boldsymbol{\rho}^{(*j)\,new} = (1-\eta)\boldsymbol{\rho}^{(*j)} + \eta\{(\mathbf{V}^{(*j)})^2 + \mathrm{diag}[(\mathbf{V}^{(*j)} - \hat{\mathbf{V}}^{(*j)})$
$\quad\cdot(\mathbf{V}^{(*j)} - \hat{\mathbf{V}}^{(*j)})^T]\}$,
where η is a learning rate that takes a small positive value.

and thus discarded. As long as the component number and $\{m_l\}$ are initialized at large enough values, model selection will be conducted automatically during parameter learning. Similarly, r is also determined automatically. Readers are referred to [8] for a recent systematic overview of the BYY harmony learning.

IV. Experiments on Gene Expression Data

A. Clustering on SRBCT Data

We apply BYY-spLFA to the diagnostic research dataset of small round blue-cell tumors (SRBCTs) of childhood [5]. As described in Table I, the SRBCT dataset consists of 83 samples from four categories, and the gene expression levels of 2308 genes were measured using cDNA microarray.

Table I
Description of the SRBCT dataset.

# genes	2308
# samples	83
4 diagnostic classes	Ewing's sarcoma (EWS): 29 samples Burkitt's lymphoma (BL): 11 samples neuroblastoma (NB): 18 samples rhabdomyosarcoma (RMS): 25 samples

Since we have already known the true partition of the dataset, the clustering performance can be evaluated by comparing the learned partition and the true partition in terms of external criteria, one representative of which is the Rand index (RI) [6]. Particularly, supposing the true partition is $\mathbf{A}$ and the learned partition is $\mathbf{B}$, a score s_1 is defined as the number of pairs of samples that are in the same cluster in $\mathbf{A}$ and in the same cluster in $\mathbf{B}$, and another score s_2 is defined as the number of pairs of samples that are in different clusters in $\mathbf{A}$ and in different clusters in $\mathbf{B}$. Considering totally C_N^2 possible pairs of samples, the Rand index is defined as:

$$
RI = (s_1 + s_2)/C_N^2, \tag{4}
$$

which intuitively describes the degree of agreement between two partitions $\mathbf{A}$ and $\mathbf{B}$. The Rand index has a value between 0 and 1, and a higher score indicates better agreement of the learned clusters with the ground truth. Particularly, $RI = 0$ if two partitions do not agree on any pair of samples, and $RI = 1$ if the two partitions are exactly the same.

Table II
Best Rand index scores on SRBCT dataset after 10 independent runs. The parentheses include cluster numbers used for k-means and HC, and the error-bar for BYY-spLFA. BYY-spLFA selected 4.4 clusters in average.

method	k-means	HC	BYY-spLFA (avg±std)
Rand index	0.52 (3) 0.61 (4) 0.57 (5)	0.34 (3) 0.52 (4) 0.53 (5)	0.92 (0.85±0.04)

Since the k-means and the (single-link) hierarchical clustering (HC) algorithm can not determine the cluster number k, we implement k-means for $k = \{3, 4, 5\}$ respectively, and cut the dendrogram of HC to generate $k = \{3, 4, 5\}$ clusters. After 10 independent runs, the best Rand index values of k-means, HC and BYY-spLFA are compared in Table II. As shown, BYY-spLFA outperforms k-means and HC, relatively by about 60% and 80% in terms of the best Rand index. Moreover, BYY-spLFA selects 4.4 clusters in average of the 10 runs.

B. Clustering on Lymphoma Data

Here BYY-spLFA is applied on gene expression profiles of lymphoma cancer data from [7], which includes 96 samples with expression levels of 4026 genes. There are 9 phenotypes (classes) in total, and we pick the 4 largest phenotypes withs 76 samples in total: P_1) 46 samples of

Table III
Contingency tables on lymphoma data.

BYY-spLFA
Rand index: 0.85
accuracy: 94.7%

	P_1	P_2	P_3	P_4
C_1	42	0	0	0
C_2	0	11	0	0
C_3	1	0	10	0
C_4	3	0	0	9

k-means
Rand index: 0.62
accuracy: 82.9%

	P_1	P_2	P_3	P_4
C_1	38	1	0	0
C_2	2	9	0	0
C_3	4	0	8	1
C_4	2	1	2	8

HC
Rand index: 0.52
accuracy: 79.0%

	P_1	P_2	P_3	P_4
C_1	36	1	1	0
C_2	3	8	0	0
C_3	2	2	8	1
C_4	5	0	1	8

diffuse large B-cell lymphoma, P_2) 11 samples of follicular lymphoma, P_3) 10 samples of Activated Blood cell B-cell, P_4) 9 samples of chronic lymphocytic leukemia. Our purpose is to uncover these phenotypes from samples directly without human expertise.

BYY-spLFA is initialized as 10 clusters and is able to detect 4 clusters. The k-means and HC algorithms are implemented for a fixed cluster number $k = 4$. Their clustering results are shown in the form of contingency tables in Table III. Therein, $C_1 \sim C_4$ refer to the four learned clusters, and a value v_{ij} in column P_i and row C_j is the number of samples in common between P_i and C_j. We can observe that all phenotypes are correctly recovered except for P_1 with 4 samples wrongly clustered into C_3 and C_4. Table III shows that BYY-spLFA not only successfully detects the phenotype number 4, but also clusters the samples more accurately than k-mean and HC, with the highest RI score.

C. Classification by Supervised Learning on Leukemia Data

When cluster labels are given for training, we encounter a supervised learning task. Under BYY harmony learning [8], supervised learning can be considered by designing $p(\mathbf{Z}|X_N) = \delta(\mathbf{Z} - \bar{\mathbf{Z}})$ while keeping the other components of Yang-machine unchanged in Box II, where $\bar{\mathbf{Z}}$ is a set of binary variables representing the given labels of the data X_N. Correspondingly, we calculate $p_{\ell t} = 1$ iff $\mathbf{x}_t$ belongs to ℓ-th cluster (or class) instead in Step(1) of Box IV. We test the supervised implementation of BYY-spLFA on a Leukemia dataset [10] which includes 2 leukemia subtypes: ALL (acute lymphocytic leukemia) and AML (acute myeloid leukemia). The leave-one-out testing is repeatedly implemented, i.e., in each trial one sample is randomly picked for testing and the rest are used for training. The classification accuracies by BYY-spLFA are presented in comparison with the Sparse Linear Discriminant Analysis (SLDA) [11] in Table IV, which preliminarily indicates that the supervised version of BYY-spLFA is effective in classification and feature selection.

Table IV
THE LEAVE-ONE-OUT AVERAGE CLASSIFICATION ACCURACIES ON THE LEUKEMIA DATASET WHICH MEASURES 7129 GENES ON 72 SAMPLES.

method	classification acc. (%)	avg. No. selected genes
SLDA	91.67	26
BYY-spLFA	98.61	18

V. CONCLUSION

For the gene clustering task, we have proposed a sparse clustering algorithm BYY-spLFA by imposing a structural prior on LFA and implementing sparsity during the BYY harmony learning. This BYY-spLFA is tested on the diagnostic research dataset of small round blue-cell tumors of childhood according to the Rand index, showing that BYY-spLFA is better than two widely used clustering methods, k-means and single-link hierarchical clustering. Also, we apply BYY-spLFA on a lymphoma cancer dataset. BYY-spLFA not only correctly detects the number of phenotypes from samples directly without human expertise, but also gives a better clustering accuracy and random index than k-means and hierarchical clustering. In addition, we modify BYY-spLFA to implement supervised learning, and preliminarily test its effectiveness on classifying a Leukemia dataset.

ACKNOWLEDGMENT

The work described in this paper was fully supported by a grant of the General Research Fund (GRF) from the Research Grant Council of the Hong Kong SAR (Project No:CUHK4180/10E), and by National Key Basic Research & Development Program 973 under Grant No. 2009CB825404.

REFERENCES

[1] P. O. Brown and D. Botstein, “Exploring the new world of the genome with DNA microarrays,” *Nature Genetics*, vol. 21, pp. 33–37, 1999.

[2] L. Xu, “Bayesian Ying Yang Learning,” in *Scholarpedia 2(3):1809, http://scholarpedia.org/article/Bayesian_Ying_Yang_Learning*, 2007.

[3] M. A. T. Figueiredo, “Adaptive sparseness using Jeffreys prior,” in *NIPS*, 2001, pp. 679–704.

[4] Y. Guan and J. Dy, “Sparse probabilistic principal component analysis,” in *AISTAT*, vol. 5, 2009, pp. 185–192.

[5] J. Khan *et al.*, “Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks,” *Nature medicine*, vol. 7, no. 6, pp. 673–679, June 2001.

[6] W. M. Rand, “Objective criteria for the evaluation of clustering methods,” *Journal of the American Statistical Association (American Statistical Association)*, vol. 66, no. 336, pp. 846–850, 1971.

[7] A. A. Alizadeh *et al.*, “Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling,” *Nature*, vol. 403, pp. 503–511, 2000.

[8] L. Xu, “Bayesian Ying-Yang System, Best Harmony Learning, and Five Action Circling,” *A special issue on Emerging Themes on Information Theory and Bayesian Approach, Journal of Frontiers of Electrical and Electronic Engineering in China*, vol. 5, no. 3, pp. 281–328, 2010.

[9] K. P. Burnham and D. Anderson, *Model Selection and Multi-Model Inference*. Springer, July 2002.

[10] T. R. Golub *et al.*, “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring,” *Science*, vol. 286, pp. 531–537, 1999.

[11] J. D. Tebbens and P. Schlesinger, “Improving implementation of linear discriminant analysis for the high dimension/small sample size problem,” *Computational Statistics and Data Analysis*, vol. 52, no. 1, pp. 423–437, 2007.