Independent Arrays or Independent Time Courses for Gene Expression Time Series

Sookjeong Kim, Seungjin Choi
Department of Computer Science
Pohang University of Science and Technology
San 31 Hyoja-dong, Nam-gu
Pohang 790-784, Korea
Email: {koko,seungjin}@postech.ac.kr

***Abstract*— In this paper we apply three different independent component analysis (ICA) methods, including spatial ICA (sICA), temporal ICA (tICA), and spatiotemporal ICA (stICA), to gene expression time series data and compare their performance in clustering genes and in finding biologically meaningful modes. Only spatial ICA was applied to gene expression data [3], [4]. However, in the case of yeast cell cycle-related gene expression time series data, our comparative study reveals that tICA outperforms sICA and stICA in the task of gene clustering and stICA finds linear modes that best match the cell cycle.**

I. Introduction

Microarray technology allows us to measure expression levels of thousands of genes simultaneously, producing gene expression profiles generated by gene interactions. For example, gene expression data analysis is useful in discriminating cancer tissues from healthy ones or in revealing biological functions of certain genes. Successive microarray experiments over time, produces gene expression time series data. Main issues in these experiments (over time), are to detect cellular processes underlying regulatory effects, to infer regulatory networks, and ultimately to match genes with associated biological functions.

Linear model-based methods explicitly describe expression levels of genes as linear functions of common hidden variables which are expected to be related to distinct biological causes of variations such as regulators of gene expression, cellular functions, or responses to experimental treatments. Such linear model-based methods include singular value decomposition (SVD), principal component analysis (PCA), independent component analysis (ICA), Bayes decomposition, and the plaid model. Standard clustering methods (such as k-means and hierarchical clustering) assign a gene (involving various biological functions) to one of clusters, however linear model-based methods allow the assignment of such a gene to null, single, or multiple clusters.

ICA is an exemplary linear model-based method that has been widely used in a variety of applications. Given a set of multivariate data, ICA aims at finding a linear decomposition where statistical independence is maximized over space (sICA) or over time (tICA). On one hand, tICA has been widely used in the context of blind source separation (for example, acoustic source separation, co-channel signal separation in digital communications, brain wave separation in EEG, and so on), since a set of temporally independent time courses is sought for in such applications. On the other hand, sICA was successfully applied to the field of medical image analysis (for example, fMRI and PET) where mutually independent source images and a corresponding dual set of unconstrained time courses, are of interest [5]. Spatiotemporal ICA (stICA) is a method which permits a trade-off between the mutual independence of spatial underlying variables (for example, images in fMRI) and the mutual independence of their corresponding time courses [8].

In the context of bioinformatics, Liebermeister [4] showed that expression modes and their influences, extracted by sICA, could be used to visualize the samples and genes in lower-dimensional space and a projection to expression modes could highlight particular biological functions. In addition, sICA was also used in gene clustering [3]. So far, only sICA has been considered as a tool for gene expression data analysis, because it seems to better fit in such a task. However, regarding gene expression time series data, tICA might be more suitable for gene clustering and temporal mode analysis, because it tries to maximize mutual independence over time. Numerical experimental study with several sets of yeast cell cycle-related gene expression time series data, shows that tICA outperforms sICA and stICA, which is an interesting result. Although sICA, tICA, and stICA are known methods, a main contribution of this paper, is to compare these three methods in the context of gene expression time series data analysis, showing that tICA is more suitable for gene clustering and stICA finds linear modes that best match the cell cycle.

II. Linear Models

Linear models assume that the data matrix $X = [X_{ij}] \in \mathbb{R}^{m \times N}$ (where the (i,j)-element, X_{ij} represents the expression level of the ith gene associated with the jth sample (time point), $i = 1, \ldots, m$, $j = 1, \ldots, N$.) is modelled as

$$X = SA, \tag{1}$$

where $S \in \mathbb{R}^{m \times n}$ and $A \in \mathbb{R}^{n \times N}$ are the encoding variable and linear mode matrix, or vice versa, depending on constraints over time or over space.

We briefly overview of linear model-based methods including PCA, sICA, tICA, and stICA. We follow some notations used in [8].

A. PCA

PCA is a widely-used linear dimensionality reduction technique which decomposes high-dimensional data into low-dimensional subspace components. PCA is illustrated as a linear orthogonal transformation which captures maximal variations in data.

Suppose that the singular value decomposition (SVD) of X is given by

$$X \approx UDV^\top, \tag{2}$$

where $U \in \mathbb{R}^{m\times n}$ corresponds to eigenarrays, $V \in \mathbb{R}^{n\times N}$ is associated with eigengenes, and D is a diagonal matrix containing singular values. In order to choose an appropriate value of n, we use the method, *PCA-L* which is based on the Laplace approximation [6].

In this paper, we use PCA for two reasons: (1) in order to provide a comparison with ICA methods; (2) to provide a reduced rank data set as input to ICA. Following notations in [8], we define $\widetilde{X} \approx X$ as

$$X \approx \widetilde{X} = UDV^\top = \left(UD^{1/2}\right)\left(VD^{1/2}\right)^\top = \widetilde{U}\widetilde{V}^\top. \tag{3}$$

B. Spatial ICA

Spatial ICA seeks a set of independent arrays S_S and a corresponding set of dual unconstrained time courses A_S. It embodies the assumption that each eigenarray in $\widetilde{U}$ is composed of a linear combination of n independent arrays (associated with independent component patterns), i.e., $\widetilde{U} = S_S\widetilde{A}_S$, where $S_S \in \mathbb{R}^{m\times n}$ contains a set of n independent m-dimensional arrays and $\widetilde{A}_S \in \mathbb{R}^{n\times n}$ is an encoding variable matrix (mixing matrix).

Define $Y_S = \widetilde{U}W_S$ where W_S is a permuted version of $\widetilde{A}_S^{-1}$. That is $Y_S = S_S P$ where P is a generalized permutation matrix. With this definition, the n dual time courses $A_S \in \mathbb{R}^{n\times N}$ associated with the n independent arrays, is computed by $A_S = W_S^{-1}\widetilde{V}^\top$, since $\widetilde{X} = Y_S A_S = \widetilde{U}\widetilde{V}^\top = Y_S W_S^{-1}\widetilde{V}^\top$. Each row vector of A_S corresponds to a temporal mode.

C. Temporal ICA

Temporal ICA finds a set of independent time courses and a corresponding set of dual unconstrained arrays (spatial patterns). It embodies the assumption that each eigengene in $\widetilde{V}$ consists of a linear combination of n independent sequences, i.e., $\widetilde{V} = S_T\widetilde{A}_T$, where $S_T \in \mathbb{R}^{N\times n}$ has a set of n independent temporal sequences of length N and $\widetilde{A}_T \in \mathbb{R}^{n\times n}$ is an associated mixing matrix.

Unmixing by $Y_T = \widetilde{V}W_T$ where $W_T = \widetilde{A}_T^{-1}P$, leads us to recover the n dual arrays A_T associated with the n independent time courses, by calculating $A_T = W_T^{-1}\widetilde{U}^T$, which is a consequence of $\widetilde{X}^T = Y_T A_T = \widetilde{V}\widetilde{U}^T = Y_T W_T^{-1}\widetilde{U}^T$. Fig. 1 illustrates how a set of dual arrays are calculated when tICA is applied to gene expression time series data.

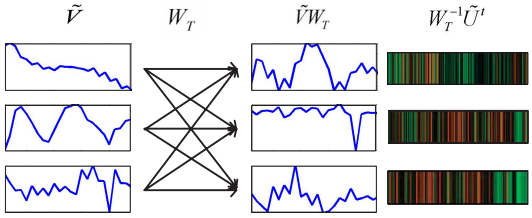

Fig. 1. An illustration of finding a set of dual arrays associated with n independent time courses, when tICA is applied to gene expression time series data. Temporal ICA first learn a mixing matrix A_T from the eigengene matrix V, in order to compute $W_T = A_T^{-1}P$. Eigengenes are linearly transformed by W_T, to produce $Y_T = VW_T$. Then n dual arrays A_T (each row of A_T corresponds to m-dimensional array) are computed by $A_T = W_T^{-1}U^\top$.

D. Spatiotemporal ICA

In linear decomposition, sICA enforces independence constraints over space, to find a set of independent arrays, whereas tICA embodies independence constraints over time, to seek a set of independent time courses. Spatiotemporal ICA finds a linear decomposition, by maximizing the degree of independence over space as well as over time, without necessarily producing independence in either space or time. In fact it allows a trade-off between the independence of arrays and the independence of time courses.

Given $\widetilde{X} = \widetilde{U}\widetilde{V}^\top$, stICA finds the following decomposition:

$$\widetilde{X} = S_S\Lambda S_T^\top, \tag{4}$$

where $S_S \in \mathbb{R}^{m\times n}$ contains a set of n independent m-dimensional arrays, $S_T \in \mathbb{R}^{N\times n}$ has a set of n independent temporal sequences of length N, and Λ is a diagonal scaling matrix. There exist two $n\times n$ mixing matrices, W_S and W_T such that $S_S = \widetilde{U}W_S$ and $S_T = \widetilde{V}W_T$. The following relation

$$\begin{aligned}\widetilde{X} &= S_S\Lambda S_T^\top = \widetilde{U}W_S\Lambda(\widetilde{V}W_T)^\top \\ &= \widetilde{U}W_S\Lambda W_T^\top\widetilde{V}^\top = \widetilde{U}\widetilde{V}^\top,\end{aligned} \tag{5}$$

implies that $W_S\Lambda W_T^\top = I$, which leads to

$$W_T = (W_S^{-1})^\top(\Lambda^{-1})^\top. \tag{6}$$

Linear transforms, W_S and W_T, are found by jointly optimizing objective functions associated with sICA and tICA. That is, the objective function for stICA has the form

$$\mathcal{J}_{stICA} = \alpha\mathcal{J}_{sICA} + (1-\alpha)\mathcal{J}_{tICA}, \tag{7}$$

where $\mathcal{J}_{sICA}$ and $\mathcal{J}_{tICA}$ could be infomax criteria or log-likelihood functions and α defines the relative weighting for spatial independence and temporal independence. More details on stICA can be found in [8].

III. Numerical Experiments

We applied sICA, tICA, and stICA to 3 sets of yeast cell cycle-related data [1], [7] (see Table I). Procedures that we took from preprocessing till statistical significance test, are summarized below.

1) *Preprocessing:* The gene expression data matrix X was preprocessed such that each element is associated with $X_{ij} =$

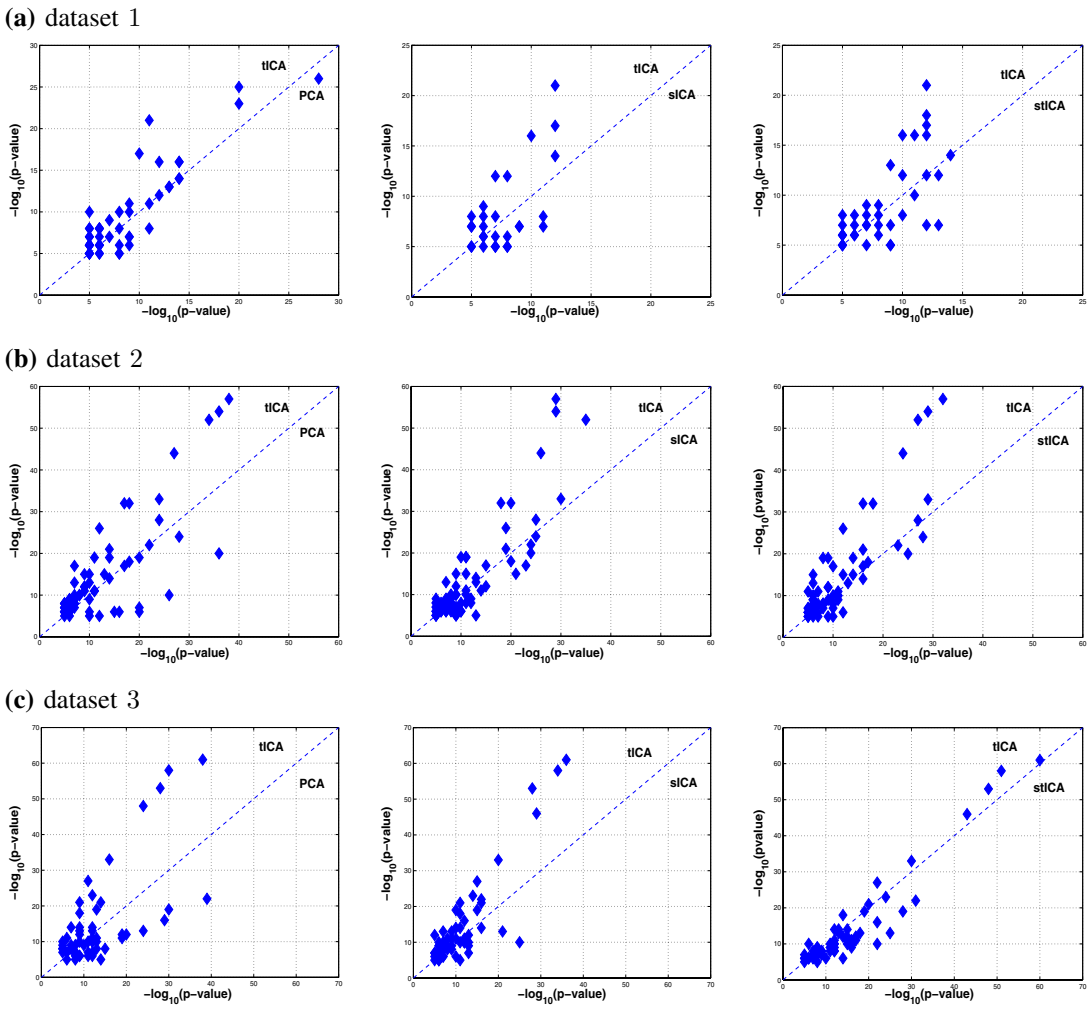

Fig. 2. Performance comparison of three different ICA methods (such as sICA, tICA and stICA) and PCA on **(a)** dataset 1 (alpha) **(b)** dataset 2 (cdc15) and **(c)** dataset 3 (elutriation). Each point corresponds to $-\log_{10}(p-value)$ of a GO annotation (biological function).

TABLE I
DATASETS AND THEIR PROPERTIES.

no	experiment	# of ORFs	# time points	# of eigenvectors
1	alpha	4579	18	6
2	cdc15	5490	24	7
3	elutriation	5981	14	4

$\log R_{ij} - \log G_{ij}$ where R_{ij} and G_{ij} represent red and green light intensity, respectively. We removed genes whose profiles have missing values more than 10%. Then we applied the *KNNimput* method [10], in order to fill in missing values. The data matrix was centered such that each row and each column have zero mean.

2) *Dimensionality Selection:* We chose the dimension n using the *PCA-L* method [6] that is based on the Laplace approximation. Then SVD was applied to find the decomposition in (3).

3) *Decomposition by ICA:* A conjugate gradient method was used to find independent components for sICA, tICA, and stICA. The initial condition for the unmixing matrix were randomly chosen, but were identical for three ICA methods. The hypothesized density for ICA algorithms were chosen as a super-Gaussian distribution.

4) *Gene Clustering:* Column vectors of S_S in sICA are independent arrays. For each column vector, genes with strong positive and negative values are grouped, which leads to two clusters related to induced and repressed genes. For tICA, the same method is applied to the row vectors of A_T containing unconstrained dual arrays.

5) *Statistical Significance Test:* For each cluster, we measured the enrichment with genes of known functional annotations. Using the Gene Ontology (GO) annotation databases [2], we calculated the p-value for each cluster with every annotated genes. The hypergeometric distribution was used to obtain the chance probability of observing the number of genes from a particular GO functional category within each cluster [9].

IV. Results

Our analysis was carried out in accordance with procedures mentioned above on the publicly available yeast cell cycle-related data. The expression levels of genes were preprocessed to be log-ratios $X_{ij} = \log R_{ij} - \log G_{ij}$ where R_{ij} and G_{ij} denote red and green intensities, respectively. Then we applied ICA algorithms (including tICA, sICA, stICA) and PCA to the gene expression matrix X. All results were based on the analysis of a reduced rank data $\widetilde{X}$ of rank n. We computed n independent components and n principal components by ICA and PCA, respectively. Each eigenarray or independent array (or dual array) went through grouping into two clusters, each of which contains 10% genes with significantly high or low influences. Statistical significance for each cluster was evaluated by computing p-values which tell us how well the genes in a cluster match a certain functional category. Only p-values less than 10^{-5} were considered. Scatter plots of the negative logarithm of the best p-value for each cluster are shown in Fig. 2. Among three ICA methods, tICA was the best in all cases. In addition, ICA methods produced significantly lower p-value than PCA did.

The ICA decomposition of the gene expression data matrix of rank n, leads to n temporal modes (or dual time courses). These modes were already shown to be related to cell cycle behavior when those modes were calculated by sICA [4]. Here we calculated (dual) temporal modes by three ICA methods. An interesting point that we found in our numerical experiments, was that temporal modes calculated by stICA exhibited the cell cycle behavior more clearly, compared to other two ICA methods. Fig. 3 shows the temporal behavior of mode 2, 3, 4 when stICA calculated 6 temporal modes. The result is also summarized in Table II).

TABLE II

THE THREE MOST SIGNIFICANT TEMPORAL MODES ON DATASET 1 (ALPHA). THE TEMPORAL MODES WERE CHARACTERIZED ACCORDING TO FUNCTIONALLY RELATED CATEGORIES.

mode	Induced functions	Repressed functions
2	sexual reproduction, cell wall, bud	protein amino acid glycosylation, nucleosome
3	cell cycle, cell proliferation, DNA replication, response to stress, chromosome, replication fork,	ribosome biogenesis, rRNA processing, nucleolus, RNA helicase activity
4	cell proliferation, cell cycle, DNA repair, chromosome, cell wall	

V. Discussion

We have applied three different ICA methods, including sICA, tICA, and stICA, to a problem of gene expression time series analysis. We compared three ICA methods in the context of gene expression time series data analysis, showing that tICA is more suitable for gene clustering in all datasets and stICA finds linear modes that best match the cell cycle. tICA and stICA would be expected to reflect the specific characteristics of gene expression time series data. For gene clustering, ICA methods performed well than PCA did. Consequently, the linear temporal modes and independent arrays (spatial patterns) will help to highlight particular biological functions in gene expression time series data.

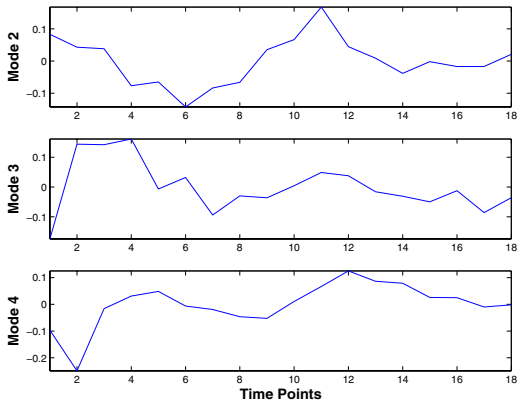

Fig. 3. Temporal modes that best match the cell cycle on dataset 1 (alpha).

Acknowledgment

We thank J.V. Stone *et al.* for sharing their stICA MATLAB code with us. This work was supported by Systems Bio-Dynamics Research Center and POSTECH Basic Research Fund.

References

[1] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lcokhart, and R. W. Davis, "A genome-wide transcriptional analysis of the mitotic cell cycle," *Mol. Cell*, vol. 2, pp. 65–73, 1998.

[2] Gene Ontology Consortium, "Creating the gene ontology resource: Design and implementation," *Genome Research*, vol. 11, pp. 1425–1433, 2001.

[3] S. Lee and S. Batzoglou, "ICA-based clustering of genes from microarray expression data," in *Advances in Neural Information Processing Systems*, vol. 16. MIT Press, 2004.

[4] W. Liebermeister, "Linear modes of gene expression determined by independent component analysis," *Bioinformatics*, vol. 18, no. 1, pp. 51–60, 2002.

[5] M. J. McKeown, T. Jung, S. Makeig, G. Brown, S. S. Kindermann, T. W. Lee, and T. J. Sejnowski, "Spatially independent activity patterns in functional magnetic resonance imaging data during the stroop color-naming task," *Proc. Natl. Acad. Sci. USA*, vol. 95, pp. 803–810, 1998.

[6] T. P. Minka, "Automatic choice of dimensionality for PCA," in *Advances in Neural Information Processing Systems*, vol. 13. MIT Press, 2001.

[7] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization," *Molecular Biology of the Cell*, vol. 9, pp. 3273–3297, Dec. 1998.

[8] J. V. Stone, J. Porrill, N. R. Porter, and I. W. Wilkinson, "Spatiotemporal independent component analysis of event-related fmri data using skewed probability density functions," *NeuroImage*, vol. 15, no. 2, pp. 407–421, 2002.

[9] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of generic network architecture," *Nature Genetics*, vol. 22, pp. 281–285, 1999.

[10] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.