Topographic Independent Component Analysis of Gene Expression Time Series Data

Sookjeong Kim and Seungjin Choi

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

Abstract. Topographic independent component analysis (TICA) is an interesting extension of the conventional ICA, which aims at finding a linear decomposition into approximately independent components with the dependence between two components is approximated by their proximity in the topographic representation. In this paper we apply the topographic ICA to gene expression time series data and compare it with the conventional ICA as well as the independent subspace analysis (ISA). Empirical study with yeast cell cycle-related data and yeast sporulation data, shows that TICA is more suitable for gene clustering.

1 Introduction

Microarray technology allows us to measure expression levels of thousands of genes simultaneously, producing gene expression profiles that are 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 principal component analysis (PCA) [1], factor analysis [2] independent component analysis (ICA) [3, 4], and independent subspace analysis (ISA) [5, 6]. 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.

In the context of bioinformatics, Liebermeister [4] showed that expression modes and their influences, extracted by ICA, 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, ICA was successfully applied to gene clustering [7, 8]. ISA [9] is a generalization of ICA where invariant feature subspace is incorporated with multidimensional ICA, allowing components in the same subspace to be dependent but requiring independence between feature subspace. It was shown in [5, 6] that ISA is more useful in gene clustering and gene-gene interaction analysis, compared to ICA.

Topographic independent component analysis (TICA) is a further generalization of ISA, which aims at finding a linear decomposition into approximately independent components with the dependence between two components is approximated by their proximity in the topographic representation [10]. In other words, TICA incorporates some nonlinear dependency into a linear model, which is more suitable for gene expression time series data where there might exist some dependency between expression modes. In this paper we apply TICA to gene expression time series data and compare it with the conventional ICA as well as the independent subspace analysis (ISA). Empirical study with yeast cell cycle-related data and yeast sporulation data, shows that TICA is more suitable for gene clustering.

2 Methods: ICA, ISA, TICA

2.1 ICA

ICA is a statistical method that decomposes a multivariate data into a linear sum of non-orthogonal basis vectors with basis coefficients being statistically independent. The simplest form of ICA consider the linear generative model where the data matrix $\boldsymbol{X} = [X_{ij}]$ (where the element X_{ij} represents the expression level of gene i associated with the jth sample, $i = 1, \ldots, m$, $j = 1, \ldots, N$) is assumed to be generated by

$$\boldsymbol{X} = \boldsymbol{S}\boldsymbol{A}, \tag{1}$$

where $\boldsymbol{S} \in \mathbb{R}^{m \times n}$ is a matrix consisting of latent variables (or encoding variables) and the row vectors of $\boldsymbol{A} \in \mathbb{R}^{n \times N}$ are basis vectors corresponding to *linear modes* in [4].

Given a matrix $\boldsymbol{X} \in \mathbb{R}^{m \times N}$, its row and column vectors are denoted by $\boldsymbol{x}_i$, $i = 1, \ldots, m$ and by $\boldsymbol{x}^j$, $j = 1, \ldots, N$. Throughout this paper, we assume that the data matrix $\boldsymbol{X}$ is already whitened. In other words, the row vectors of $\boldsymbol{A}$ are confined to be orthogonal each other and to be normalized to have unit norm. Non-orthogonal factor is reflected in a whitening transform. In order to avoid an abuse of notations, we use the notation $\boldsymbol{X}$ for the whitened data matrix and $n \leq N$ represents an intrinsic dimension estimated by PCA.

ICA searches for a parameter matrix $\boldsymbol{W} \in \mathbb{R}^{n \times n}$ which maximizes the normalized log-likelihood $\mathcal{L}_{ica}$ given by

$$\mathcal{L}_{ica} = \frac{1}{m} \sum_{t=1}^{m} \sum_{i=1}^{n} \log p\left(\langle \boldsymbol{w}_i, \boldsymbol{x}_t \rangle\right) + \log \left|\det \boldsymbol{W}\right|, \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product between two arguments. The estimated parameter matrix $\boldsymbol{W}$ leads us to calculate the latent variable matrix by $\boldsymbol{S} = \boldsymbol{X}\boldsymbol{W}^T$ ($\boldsymbol{W}^T = \boldsymbol{A}^{-1}$). For an orthogonal matrix $\boldsymbol{A}$, the row vectors of $\boldsymbol{W}$ coincide with the row vectors of $\boldsymbol{A}$.

2.2 ISA

In contrast to ICA, multidimensional ICA [11] assumes that latent variables $s_i = \langle \boldsymbol{w}_i, \boldsymbol{x} \rangle$ are divided into J number of κ-tuples (where κ represents the dimension of subspace) and find a linear decomposition such that J κ-tuples are independent with allowing the components in the same tuple to be dependent. For the sake of simplicity, we assume identical dimension, κ for every feature subspace. ISA [9] incorporates the invariant feature subspace into the multidimensional ICA. To this end, the pooled energy $E_j(\boldsymbol{x})$ for the jth feature subspace $\mathcal{F}_j$ is defined by

$$E_j(\boldsymbol{x}) = \sum_{i \in \mathcal{F}_j} \langle \boldsymbol{w}_i, \boldsymbol{x} \rangle^2 . \tag{3}$$

With these definitions, the normalized log-likelihood $\mathcal{L}_{isa}$ of the data given the ISA model, is given by

$$\mathcal{L}_{isa} = \frac{1}{m} \sum_{t=1}^{m} \sum_{j=1}^{J} \log p \left(\sum_{i \in \mathcal{F}_j} \langle \boldsymbol{w}_i, \boldsymbol{x}_t \rangle^2 \right) + \log |\det \boldsymbol{W}| , \tag{4}$$

where $p\left(\sum_{i \in \mathcal{F}_j} s_i^2\right) = p_j(s_i, i \in \mathcal{F}_j)$ represents the probability density inside the jth κ-tuple of s_i.

The parameter matrix $\boldsymbol{W}$ which maximizes the log-likelihood (4), finds a linear decomposition such that pooled energies $E_j(\boldsymbol{x})$ are independent but the components $s_i \in \mathcal{F}_j$ are allowed to be dependent. Learning $\boldsymbol{W}$ can be carried out by a gradient-ascent method. More details on ISA can be found in [9].

2.3 TICA

TICA is a further generalization of ISA, which aims at finding a linear decomposition into approximately independent components with the dependence between two components is approximated by their proximity in the topographic representation [10].

The following normalized log-likelihood $\mathcal{L}_{tica}$ was considered for TICA,

$$\mathcal{L}_{tica} = \frac{1}{m} \sum_{t=1}^{m} \sum_{j=1}^{n} \Psi \left(\sum_{i=1}^{n} h(i,j) \langle \boldsymbol{w}_i, \boldsymbol{x}_t \rangle^2 \right) + \log |\det \boldsymbol{W}| , \tag{5}$$

where $\Psi(\cdot)$ is a function of local energies that plays a similar role to the log-density in the conventional ICA and $h(i,j)$ is a neighborhood function. See [10] for more details.

3 Experiments and Results

3.1 Datasets

Our experiments were conducted with publicly available yeast cell cycle datasets [12, 13] and yeast sporulation dataset [14] (see Table 1).

Table 1. Four datasets used in our experiments, are summarized. First three datasets are yeast cell cycle-related data that was also used in [12, 13] and the last dataset is yeast sporulation data [14]. Experiments in yeast cell cycle-related data, are named by the method used to synchronize yeast cells. The number of open read frames (ORFs) is the number of time-series that have no missing values in them, time interval is the interval between measurements, # time points is the number of measurements, and # of eigenvectors indicated the number of eigenvectors chosen by PCA-L [15].

no	experiment	# of ORFs	time interval	# time points	# of eigenvectors
1	alpha	4579	7 min	18	6
2	cdc15	5490	10-20 min	24	8
3	cdc28	3167	10 min	17	6
4	sporulation	6118	0.5-3 hr	7	4

3.2 Procedures

Procedures that we took from a 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 $X_{ij} = \log R_{ij} - \log G_{ij}$ where R_{ij} and G_{ij} represent red and green light intensity, respectively. In practice, gene expression data usually contain missing values. We removed genes whose profiles have missing values more than 10%. Then we applied the *KNNimput* method [16], in order to fill in missing values. The data matrix was doubly centered such that each row and each column have zero mean.

2) *Data whitening:* Given the gene expression data matrix $X \in \mathbb{R}^{m \times N}$ where m is the number of genes and N is the number of arrays (time points), we chose the dimension n using the *PCA-L* method [15]. Data whitening was carried out through PCA with n principal eigenvectors.

3) *Decomposition by ICA, ISA, and TICA:* We applied ICA, ISA, and TICA algorithms, to whitened data matrix, in order to estimate the parameter matrix $W \in \mathbb{R}^{n \times n}$.

4) *Gene clustering:* For each column vector $\boldsymbol{s}^i$, genes with strong positive and negative values are grouped, which leads to two clusters related to induced and repressed genes. We considered standard deviation σ for each column vector

as a threshold. Genes with expression levels higher than $c \times \sigma$ and with expression levels lower than $-c \times \sigma$, are grouped as two significant clusters. In our experiments, we chose $c = 1.5$.

5) *Statistical significance test:* To determine statistical significance of functional category enrichment for each cluster, we used the Gene Ontology (GO) annotation database [17] where genes were assigned to an associated set of functional categories. We calculated p-values for statistical significance test, using the hypergeometric distribution that is used to obtain the chance probability of observing the number of genes from a particular GO category within each cluster. The p-value is the probability to find at least k genes from a functional category within a cluster of size c:

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{f}{i}\binom{g-f}{c-i}}{\binom{g}{c}}, \tag{6}$$

where f is the total number of genes within a functional category and g is the total number of genes within the genome [18].

3.3 Results

It was shown in [7] that ICA-based gene clustering method outperformed several existing methods such as PCA, k-means, and hierarchical clustering. However, the conventional ICA model do not take into account the temporal dependence of gene expression time series data. The inherent time dependencies in the data suggest that clustering techniques which reflect those dependencies yield improved performance. ISA and TICA-based gene clustering methods consider somewhat dependencies of gene expression patterns. Clustering results with 4 different data sets (described in Table 1), confirm that TICA and ISA indeed yield better clustering, compared to ICA (see Fig. 1).

For TICA, a square neighborhood function of size 3×3, was used. The intrinsic dimension n determined by the PCA-L for each data set is summarized in Table 1. For ISA, the number of feature subspace, was chosen as $J = 2$ for Dataset 1, 3, 4 and $J = 4$ for Dataset 2. Thus, the dimension of the feature subspace is $\kappa = 3$ for Dataset 1, 3, 4 and $\kappa = 2$ for Dataset 2.

For each data set, we determined n latent variables by ICA, ISA, and TICA and investigate the biological coherence of $2n$ clusters consisting of genes with significantly high and low expression levels within independent components. For each cluster, we calculated p-values and considered only p-values less than 10^{-5}. Scatter plots of the negative logarithm of p-value, are shown in Fig. 1.

The TICA decomposition of the gene expression data matrix of rank n, leads to n temporal modes (corresponding to n basis vectors). Each temporal mode defines two gene clusters that show a strong positive or negative response. These clusters contain subgroups related to particular biological functions, mostly consistent with the temporal modes. Fig. 2 depicts 3 temporal modes during sporu-

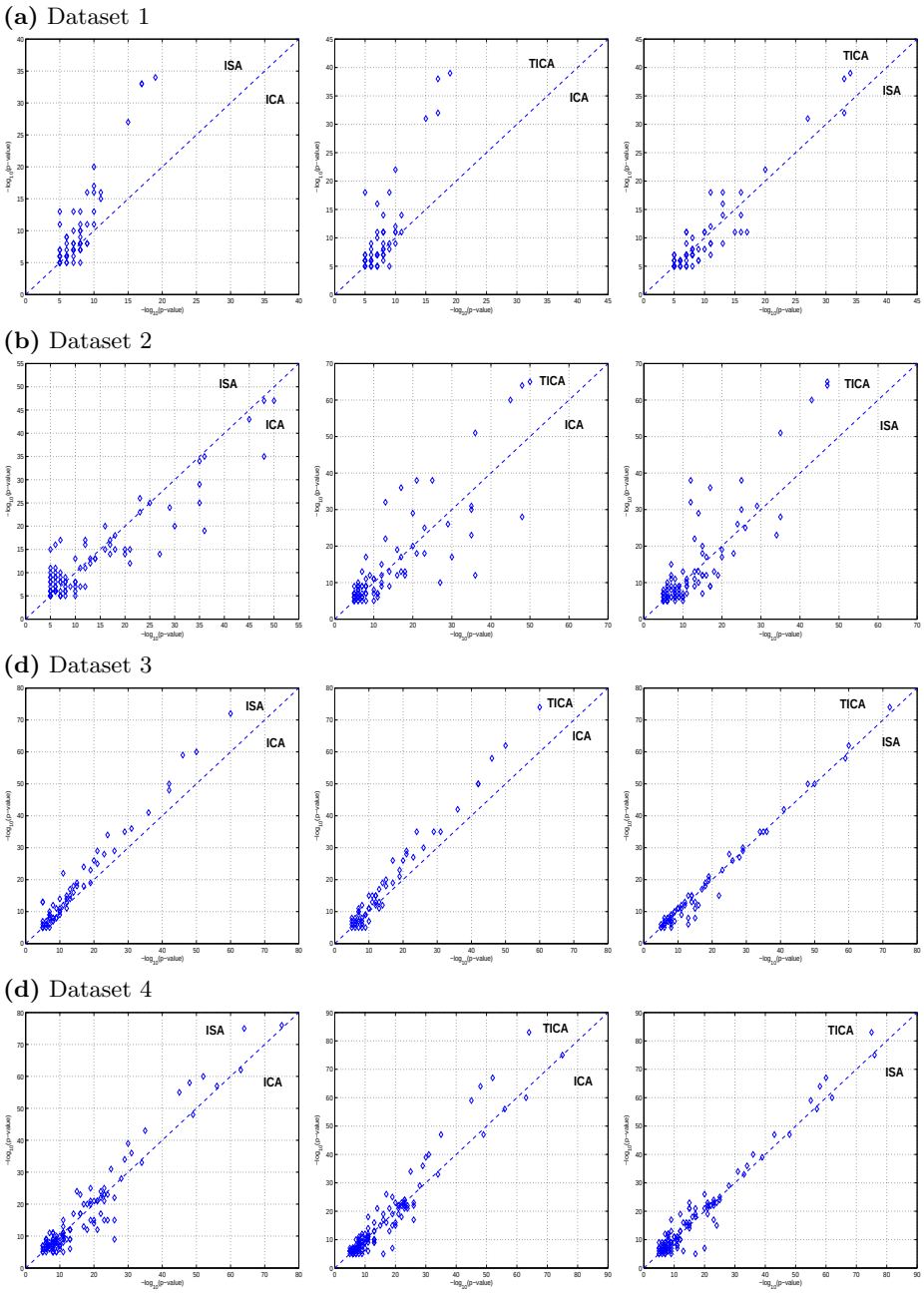

Fig. 1. Performance comparison of three different ICA methods (ICA, ISA, and TICA) with yeast cell cycle-related data and yeast sporulation data (see Table 1). Through Dataset 1-4, TICA has more points above the (anti-diagonal) line representing equal performance, compared to ICA and ISA, which indicates the enrichment of the TICA-based clustering.

Table 2. Temporal modes extracted by TICA from Dataset 4 (sporulation). The modes were characterized according to functionally related clusters.

Mode	Induced functions	Repressed functions
1	sporulation, spore wall assembly,	alcohol metabolism, carbohydrate metabolism, oxidoreductase activity
2	ribosome biogenesis and assembly, rRNA processing, rRNA metabolism, cytosolic ribosome, ribosome, structural constituent of ribosome	organic acid metabolism, carboxylic acid metabolism, amine metabolism
3	sulfur metabolism, cytosolic ribosome (sensu Eukarya), ribosome	cell cycle, cell proliferation, nuclear division, chromosome

lation. The temporal modes mainly reflect the sporulation behavior (see also Table 2).

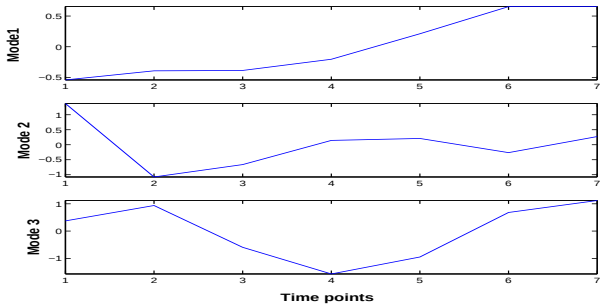

Fig. 2. Temporal modes of clusters computed by TICA are shown for Dataset 4 (sporulation).

4 Conclusions

In this paper we have applied the method of topographic ICA to gene expression time series data, in order to evaluate its performance in the task of gene clustering. Empirical comparison to the conventional ICA and the independent subspace analysis, have shown that the topographic ICA is more suitable in grouping genes into clusters containing genes associated with similar functions.

Acknowledgments: This work was supported by National Core Research Center for Systems Bio-Dynamics and POSTECH Basic Research Fund.

References

1. Raychaudhuri, S., Stuart, J.M., Altman, R.B.: Principal components analysis to summarize microarray experiments: Application to sporulation time series. In: Proc. Pacific Symp. Biocomputing. (2000) 452–463
2. Girolami, M., Breitling, R.: Biologically valid linear factor models of gene expression. Bioinformatics **20** (2004) 3021–3033
3. Hori, G., Inoue, M., Nishimura, S., Nakahara, H.: Blind gene classification based on ICA of microarray data. In: Proc. ICA, San Diego, California (2001)
4. Liebermeister, W.: Linear modes of gene expression determined by independent component analysis. Bioinformatics **18** (2002) 51–60
5. Kim, H., Choi, S., Bang, S.Y.: Membership scoring via independent feature subspace analysis for grouping co-expressed genes. In: Proc. Int'l Joint Conf. Neural Networks, Portland, Oregon (2003)
6. Kim, H., Choi, S.: Independent subspaces of gene expression data. In: Proc. IASTED Int'l Conf. Artificial Intelligence and Applications, Innsbruck, Austria (2005)
7. Lee, S., Batzoglou, S.: ICA-based clustering of genes from microarray expression data. In: Advances in Neural Information Processing Systems. Volume 16., MIT Press (2004)
8. Kim, S., Choi, S.: Independent arrays or independent time courese for gene expression data. In: Proc. IEEE Int'l Symp. Circuits and Systems, Kobe, Japan (2005)
9. Hyvärinen, A., Hoyer, P.O.: Emergence of phase and shift invarianct features by decomposition of natural images into independent feature subspaces. Neural Computation **12** (2000) 1705–1720
10. Hyvärinen, A., Hoyer, P., Inki, M.: Topographic independent component analysis. Neural Computation **13** (2001) 1525–1558
11. Cardoso, J.F.: Multidimensional independent component analysis. In: Proc. ICASSP, Seattle, WA (1998)
12. Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lcokhart, D.J., Davis, R.W.: A genome-wide transcriptional analysis of the mitotic cell cycle. Mol. Cell **2** (1998) 65–73
13. Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B.: Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. Molecular Biology of the Cell **9** (1998) 3273–3297
14. Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O., Herskowitz, I.: The transcriptional program of sporulation in budding yeast. Science **282** (1998) 699–705
15. Minka, T.P.: Automatic choice of dimensionality for PCA. In: Advances in Neural Information Processing Systems. Volume 13., MIT Press (2001)
16. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B.: Missing value estimation methods for DNA microarrays. Bioinformatics **17** (2001) 520–525
17. Gene Ontology Consortium: Creating the gene ontology resource: Design and implementation. Genome Research **11** (2001) 1425–1433
18. Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., Church, G.M.: Systematic determination of genetic network architecture. Nature Genetics **22** (1999) 281–285