독립성분분석에 의한 유전자 발현 시계열 데이터의 공간적 패턴과 시간적 모드 분석

김숙정[0] 최승진
포항공과대학교 컴퓨터공학과
{koko[0],seungjin}@postech.ac.kr

Spatial pattern and temporal mode analysis of microarray *time-series* data by independent component analysis

Sookjeong Kim[0] Seungjin Choi
Department of Computer Science, POSTECH, Korea

Abstract

In this paper we apply several variations of independent component analysis(ICA) methods, such as spatial ICA (sICA), temporal ICA (tICA), and spatiotemporal ICA (stICA), to yeast cell cycle datasets, and compare their performance in finding components that result in gene clusters coherent with annotations and in extracting meaningful temporal modes. It turns out that the results of tICA are superior to those of PCA, sICA, and stICA in terms of gene clustering and the temporal modes extracted by stICA highlights particular cellular processes.

1. Introduction

Microarray experiments can measure the expression levels of thousands of genes simultaneously. The resulting so-called expression profiles allow, for example, investigation of differences in distinct tissue types or between healthy and diseased tissues. When microarray experiments are performed successively in time we call this experimental result a gene expression *time-series* data. The questions this experiment result tries to address are the detection of the cellular processes underlying the regulatory effects observed, inference of regulatory networks and, ultimately, assignment of function to the genes analyzed in the time courses.

Linear model based methods explicitly describe the expression levels of genes as linear functions of common hidden variables which these variables may be related to distinct biological causes of variation, like regulators of gene expression, cellular functions, or responses to experimental treatments.

In this paper, we introduce diverse ICA methods, such as sICA [1], tICA [2], and stICA [3] which are a linear model based method but which are naturally modified ICA. ICA decomposes an input data into components so that each component is statistically as independent from the others as possible.

sICA seeks a set of spatially independent patterns and a corresponding (dual) set of unconstrained temporal modes. While the spatially independent patterns extracted by sICA are approximately independent, their corresponding dual temporal modes can be correlated. In contrast, while the temporally independent modes extracted by tICA are approximately independent, their corresponding dual spatial patterns can be correlated. And stICA simultaneously maximizes statistical independence over both temporal modes and spatial patterns. Thus there exist dependence between spatially independent patterns and between temporally independent modes. The physically realistic assumption of gene expression *time-series* data is that there exist dependence between temporal modes and between spatial patterns. A reliable assumption for gene clustering exist dependence between spatial patterns. Therefore tICA is superior to other methods.

2. Methods: Linear models

We applied several linear models to yeast cell cycle gene expression *time-series* datasets, in order to compare the performance of ICA methods. We consider an $m \times N$ data matrix X whose rows correspond to genes and whose columns correspond to the time points (arrays or samples).

2.1 PCA

In order to choose the appropriate number of variables (n), we use PCA-L which is based in the Laplace approximation [4]. We choose the appropriate number of principal components, k and apply PCA to X using the SVD method. The SVD of X is said to be the factorization,

$$X \approx UDV^t, \quad (1)$$

where U is an $m \times n$ matrix of $n \leqq m$ eigenarrays, V is an $N \times n$ matrix of n eigengenes, and D is a diagonal matrix of singular values $\lambda^{1/2}$. Each singular value corresponds to the square root of an eigenvalue λ. For later use, we define $\tilde{X} \approx X$ as

$$X \approx \tilde{X} = UDV^t = (UD^{1/2})(VD^{1/2})^t = \tilde{U}\tilde{V}^t \quad (2)$$

2.2 sICA

sICA seeks a set of spatially independent component(IC) patterns and a corresponding (dual) set of unconstrained temporal modes. sICA embodies the assumption that each eigenarray in U is composed of a linear combination of n spatially IC patterns $\tilde{U} = S_S\tilde{A}_S$, where $\tilde{A}_S$ is a $n \times n$ mixing matrix and S_S is and $m \times n$ set of n statistically independent patterns $S_S = (s_{S1} | \ldots | s_{Sn})$. sICA decomposes $\tilde{U}$ into n IC patterns $y_S = \tilde{U}W_S$. The unmixing matrix W_S is a permuted version of $\tilde{A}_S^{-1}$, such that each column in y_S is a scaled version of exactly one column in S_S. sICA achieves this by maximizing the entropy $H(Y_S)$ of $Y_S = \sigma_S(y_S)$, where σ_S approximates the cumulative density function (cdf) of each of the spatial ICs. The n dual temporal modes A_S associated with the n IC patterns y_S can be recovered as follows. If $\tilde{X} = y_S A_S = \tilde{U}\tilde{V}^t$ and $\tilde{U} = y_S W^{-1}$, then

$$\tilde{X} = y_S W_S^{-1}\tilde{V}^t = y_S A_S, \quad (3)$$

where each row of $A_S = W_S^{-1}\tilde{V}^t$ contains one temporal mode.

2.3 tICA

tICA seeks a set of temporally independent component (IC) modes and a corresponding (dual) set of unconstrained spatial patterns, tICA embodies the assumption that each eigengene in $\tilde{V}$ is a linear combination of n temporally independent modes $\tilde{V} = S_T\tilde{A}_T$, where $\tilde{A}_T$ is a $n \times n$ mixing matrix S_T is an $N \times n$ set of n statistically independent temporal modes $S_T = (s_{T1} | \ldots | s_{Tn})$. tICA decompose $\tilde{V}$ into n IC modes $y_T = \tilde{V}W_T$. W_T is a permuted version of A_T^{-1}, such that each column vector in y_T is a scaled version of exactly one column in S_T. This is achieved by maximizing the entropy $H(Y_T)$ of $Y_T = \sigma_T(y_T)$, where σ_T approximated the cdf of the temporal source signals.

The n dual spatial patterns A_T associated with the n IC modes y_T can be recovered as follows. If $\tilde{X} = \tilde{V}\tilde{U}^t$ and $\tilde{V} = y_T W_T^{-1}$, then

$$\tilde{X}^t = y_T W_T^{-1}\tilde{U}^t = y_T A_T, \quad (4)$$

where each row of $A_T = W_T^{-1}\tilde{U}^t$ contains one spatial pattern.

2.4 stICA

stICA maximizes the degree of independence over spatial pattern and temporal mode, without necessarily producing independence in either spatial pattern or temporal mode. That is, stICA permits a tradeoff between the mutual independence of spatial patterns and the mutual independence of their corresponding temporal modes. stICA embodies the assumption that

$$\tilde{X} = S_S \Lambda S_T^t, \quad (5)$$

where S_S is an $m \times n$ matrix of n spatially independent patterns, S_T is an $N \times n$ matrix of n temporally independent modes, and Λ is required to ensure that S_S and S_T have amplitudes appropriate to their respective cdfs σ_S and σ_T.

If $\tilde{X} = \tilde{U}\tilde{V}^t$, then two $n \times n$ unmixing matrices W_S and W_T exist such that $S_S = \tilde{U}W_S$ and $S_T = \tilde{V}W_T$

$$\tilde{X} = S_S \Lambda S_T^t = \tilde{U}W_S\Lambda(\tilde{V}W_T)^t = \tilde{U}W_S\Lambda W_T^t\tilde{V}^t = \tilde{U}\tilde{V}^t, \quad (6)$$

implying that $W_S \Lambda W_T^t = I$, from which we can derive

$$W_T = (W_S^{-1})^t(\Lambda^{-1})^t. \quad (7)$$

The matrices W_S and W_T can be found by simultaneously maximizing a function h_{ST} of the spatial and temporal entropy of extracted signals. The function h_{ST} to be maximized is defined

$$h_{ST}(W_S, \Lambda) = \alpha H(Y_S) + (1-\alpha)H(Y_T), \quad (8)$$

where α defines the relative weighting afforded to spatial and temporal entropy. We used $\alpha = 0.5$ for results presented here.

3. Results

3.1 Datasets

experiment	ORFs	time interval	time points
alhpa	4579	7 min	18
cdc15	5490	10-20 min	24
elutriation	5981	30 min	14

Table1. Datasets and their properties

3.2 Procedures

1) Preprocessing : Logarithm transformation
2) Choice of dimensionality : PCA-L
3) Decomposition by PCA and ICA algorithms
4) Gene clustering
5) Determination of statistical significance : Using Gene Ontology (GO) [5] gene annotation database

3.3 Results

Neither independent spatial patterns nor independent temporal modes are strictly independent in gene

expression *time-series* data. The inherent time dependencies in the data suggest that clustering techniques which reflect those dependencies yield improved performance. In case of gene clustering, tICA and stICA-based gene clustering methods consider somewhat dependency of independent spatial patterns. To test our hypothesis, we compared tICA and stICA with sICA on datasets. And PCA method was compared with tICA. The all ICA algorithms were tested on each preprocessed dataset under the procedures described above.

Among ICA algorithms, tICA was the best in all dataset. Among both ICA algorithms and PCA, ICA algorithms performed well. (**Figure 1**).

The each temporally mode is associated with the each spatially pattern. 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. Cell cycle behavior is manifested by the temporal modes (**Figure 2**).

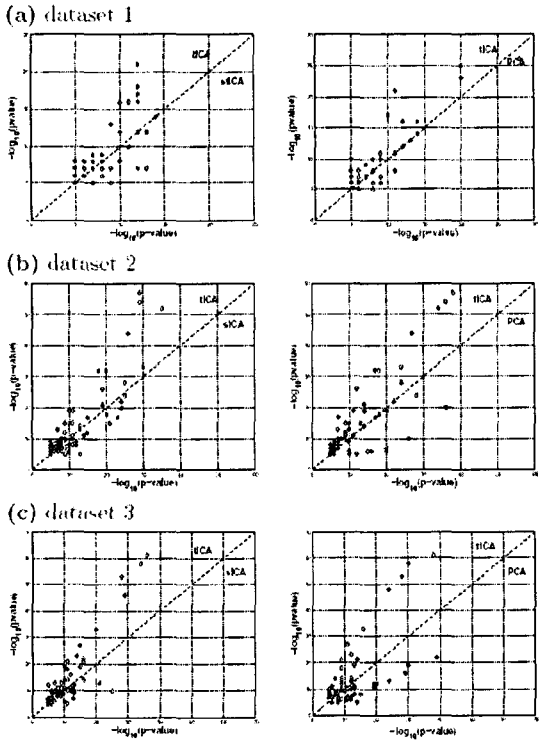

Figure 1. Scatter plots of $-log_{10}(p-value)$ of GO annotations for ICA algorithms and PCA. Each point corresponds to one functional category.

4. Conclusion and Discussion

Gene expression *time-series* experiments are an increasingly popular method for studying a wide range of biological processes. However, for analyzing these experiments, algorithms specifically designed for *time-series* data are required. We have introduced a new approach to the identification of information from gene expression *time-series* data. We tested ICA algorithms (sICA, tICA, and stICA) having *different constraints* to several gene expression *time-series* datasets. For the problem of gene clustering, tICA was the best in all datasets. Accordingly tICA would be expected to reflect the characteristics of gene expression *time-series* data.

5. Acknowledgements

This work was supported by POSTECH Research Fund and BK 21 in POSTECH.

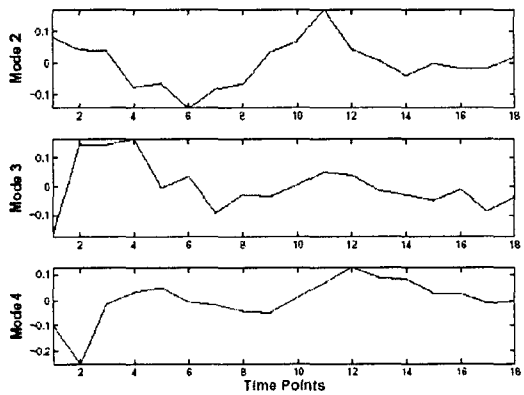

Figure 2. Temporal modes of clusters from dataset 1

6. Reference

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

[2] S. Makeig and T. Jung and A. Bell and D. Ghahremani and T. Sejnowski, Blind separation of auditory event-related brain responses into independent components, Proc. Natl Acad. Sci. USA, 94, 10979-10984, 1997

[3] J. V. Stone and J. Porrill and N. R. Porter and I. D. Wilkinson, Spatiotemporal independent component analysis of event-related fMRI data using skewed probability density functions, NeuroImage, 15, 407-421, 2002

[4] T. P. Minka, Automatic choice of dimensionality for PCA, Technical Report 514 , MIT Media Laboratory, Perceptual Computing Section, 2000

[5] Gene Ontology Consortium, Creating the gene ontology resource: design and implementation, Genome Research, 11, 1425-1433, 2001