Available online at www.sciencedirect.com
ScienceDirect
Neurocomputing 71 (2008) 2377–2387
NEUROCOMPUTING
www.elsevier.com/locate/neucom

Independent arrays or independent time courses for gene expression time series data analysis

Sookjeong Kim, Jong Kyoung Kim, Seungjin Choi*

Department of Computer Science, Pohang University of Science and Technology, San 31 Hyoja-dong, Nam-gu, Pohang 790-784, Korea

Available online 25 February 2008

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. Up to now, only spatial ICA was applied to gene expression data analysis. However, in the case of yeast cell cycle-related gene expression time series data, our comparative study shows that tICA turns out to be more useful than sICA and stICA in the task of gene clustering and that stICA finds linear modes that best match cell cycles, among these three ICA methods. The underlying generative assumption on independence over temporal modes corresponding to biological process gives the better performance of tICA and stICA compared to sICA.

Keywords: DNA microarray; Gene expression data; Independent component analysis; Principal component analysis

1. Introduction

Microarray technology allows us to measure expression levels of thousands of genes simultaneously, making it possible to explore genome-wide biological problems. For example, gene expression data analysis is useful in discriminating cancer tissues from healthy ones or in revealing biological functions of unknown genes. Gene expression patterns measured in microarray experiments over time produce gene expression time series data. Genome-wide gene expression profiles over time can be analyzed to detect underlying cellular processes, to infer transcriptional regulatory networks, and ultimately to relate genes with their associated biological functions [31].

Various computational methods have been applied to gene expression time series data. These include: (1) hierarchical clustering [9,11], *k*-means clustering [31], and a model-based clustering [25]; (2) singular value decomposition (SVD) [1,14,27] or principal component analysis (PCA) [26]; (3) Bayesian networks [10]; (4) Bayesian decomposition [22]; (5) differential equation modelling [4]; (6) hidden Markov models [28]; (7) independent component analysis (ICA) [15,21] and independent subspace analysis (ISA) [17,16].

Among the aforementioned methods, linear model-based methods including SVD, PCA, and ICA are a promising way to model generative biological processes of gene expression such as transcription factor binding and response to environmental change [21]. Such linear model-based methods were recently used in the task of clustering genes from expression data [15,20]. Although standard clustering methods such as *k*-means and hierarchical clustering assign a gene involving various biological functions to one of the clusters, 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 (spatial ICA; sICA) or over time (temporal ICA; 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,

*Corresponding author. Tel.: +82 54 279 2259; fax: +82 54 279 2299.
E-mail address: seungjin@postech.ac.kr (S. Choi).
URL: http://www.postech.ac.kr/~seungjin (S. Choi).

0925-2312/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2007.05.015

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 [23]. 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 [30].

In the context of gene expression data analysis, Liebermeister [21] 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 [15,20]. 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 studies with yeast cell cycle-related gene expression time series data, show that tICA outperforms sICA and stICA, which is an interesting result. Preliminary results were presented in [18]. Although sICA, tICA, and stICA are known methods, the 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 that stICA finds linear modes that best match cell cycles.

The rest of this paper is organized as follows. Next section illustrates PCA, sICA, tICA, and stICA in the context of gene expression data analysis. In fact, this was mainly motivated from the work in [30], where stICA was first proposed and extensively studied for fMRI data analysis. In addition, we also briefly discuss ICA for latent variables with non-symmetric probability distributions (skew-ICA), where its importance was first stressed out in [7] and further elaborated in [30]. Section 3 describes experimental results and comparison, applying these ICA methods to yeast cell cycle-related data sets. Finally, conclusions are drawn in Section 4.

2. Methods: linear models

Gene expression time series data can be organized as a matrix X where rows represent genes and columns are associated with time points. 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 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 linear model-based methods including PCA, sICA, tICA, and stICA, most of which were extensively studied for fMRI [30].

2.1. PCA

PCA is a widely used linear dimensionality reduction technique which decomposes high-dimensional data into low-dimensional subspace components. On one hand, PCA is illustrated as a linear orthogonal transformation which captures maximal variations in data. On the other hand, it finds a linear orthogonal mapping which minimizes the reconstruction error. These two approaches turn out to produce an identical result where the linear transformation is constructed by spectral decomposition of the data covariance matrix or SVD of the data matrix itself.

Suppose that the SVD of X is given by

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

where $U \in \mathbb{R}^{m\times n}$ corresponds to eigenarrays, $V \in \mathbb{R}^{N\times n}$ ($n \leqslant N$) 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 [24].

Raychaudhuri et al. [26] applied PCA to the publicly released yeast sporulation data set [8]. Alter et al. [1] introduced the use of PCA in yeast cell cycle analysis [29].

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 [30], we define $\widetilde{X} \approx X$ as

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

2.2. Spatial ICA

sICA 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). Note that "dual time courses A_S", which is another mixing matrix for the data matrix $\widetilde{X}$ is introduced to distinguish it from $\widetilde{A}_S$.

Define $Y_S = \widetilde{U}W_S$ where W_S is a permuted version of $\widetilde{A}_S^{-1}$. That is $Y_S = S_SP$ 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 $\widetilde{X} = Y_SA_S = \widetilde{U}\widetilde{V}^{\top} = Y_SW_S^{-1}\widetilde{V}^{\top}$. Each row vector of A_S corresponds to a temporal mode.