Available online at www.sciencedirect.com

ScienceDirect

Journal of Statistical Planning and Inference 138 (2008) 500–515

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Several biplot methods applied to gene expression data

Mira Park[a], Jae Won Lee[b,*], Jung Bok Lee[c], Seuck Heun Song[b]

[a]Department of Pre-medicine, Eulji University, 143-5 Yongdu-dong, Chung-gu, Daejon 301-832, Korea
[b]Department of Statistics, Korea University, 5-1 Anam-dong, Seongbuk-gu, Seoul 701-112, Korea
[c]Institute of Human Genomic Study, College of Medicine, Korea University, Gojan1-dong, Danwon-gu, Ansan, Gyeonggi-do, Korea

Available online 26 June 2007

Abstract

DNA microarray experiments result in enormous amount of data, which need careful interpretation. Biplot approaches show simultaneous display of genes and samples in low-dimensional graphs and thus can be used to represent the relationships between genes and samples. There are several different types of biplots, and these methods need to be evaluated because each plot provides different result.

In this paper, we review several variants of biplot methods such as principal component analysis biplot, factor analysis biplot, multidimensional scaling biplot and correspondence analysis biplot. We investigate the properties of these methods and compare their performances by analyzing various types of well-known gene expression data. We also suggest the supplementary data method as a tool for (i) classifying the previously unknown sample/gene to existing class, (ii) analyzing mixture data and (iii) presenting illustrative variables, etc. The usefulness of this approach for interpreting microarray data is demonstrated.

Keywords: Gene expression data; Biplot; Supplementary data; Principal component analysis; Factor analysis; Correspondence analysis; Multidimensional scaling

1. Introduction

DNA microarray technology has been advanced to the point that it is now possible to monitor gene expression levels on a genomic scale. Currently, two types of microarrays are in common use: 2-channel cDNA microarrays and high-density oligonucleotide microarrays chips such as Affymetrix chips. Every microarray gene experiments result in enormous amount of gene expression data, which need statistical considerations.

Traditional clustering techniques such as hierarchical clustering, k-means clustering and self-organizing map have been applied to the analysis of gene expression data (cf. Eisen et al., 1998; Tamayo et al., 1999; Golub et al., 1999, etc.). It is useful to find gene/sample clusters with similar gene expression patterns for summarizing and interpreting the microarray data. However, it would be more effective if we represent this information by drawing a low-dimensional graph. Visualization of the gene expression data helps us to find and interpret the relationships between genes/samples and to detect outliers. Principal component analysis, often performed by singular value decomposition, has been explored as a method for visualizing large-scale expression data. Raychaudhuri et al. (2000) used PCA to analyze time

[*] Corresponding author. Department of Statistics, Korea University, 5-1 Anam-dong, Seongbuk-gu, Seoul 136-701, Korea. Tel.: +82 2 3290 2237; fax: +82 2 924 9895.

E-mail address: jael@korea.ac.kr (J.W. Lee).

0378-3758/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2007.06.019

series yeast sporulation expression data. Similarly, Alter et al. (2000) and Holter et al. (2000) analyzed microarray data using SVD. On the other hand, Fellenberg et al. (2001) used correspondence analysis to visualize the relationship between genes and tissues.

In this paper, we review several variants of biplot methods as the visualization tool for exploring gene expression data. Biplot method was originally suggested by Gabriel (1971) and there have been several variants proposed by subsequent researchers (cf. Gower and Hand, 1996). These approaches can show simultaneous display of observations and variables as well as represent the relationships between observations and those between variables in low-dimensional graphs. Here we use PCA (principal component analysis) biplot, FA (factor analysis) biplot, MDS (multidimensional scaling) biplot and CA (correspondence analysis) biplot. We investigate the properties of the resulting graphs and compare the performances of different methods. Also we consider the supplementary data analysis, which is presented by Lebart et al. (1984), for exploratory analysis of microarray data. Several application methods are proposed with illustration of simulated and real data. These methods are evaluated with four well-known data: leukemia data set of Golub et al. (1999), lymphoma data set of Alizadeh et al. (2000), colon cancer data set of Alon et al. (1999) and 60 cancer cell line of Ross et al. (2000).

2. Principles of biplot methods in microarray data analysis

The gene expression data on p genes for n mRNA samples may be summarized by an $n \times p$ matrix $X = (x_{ij})$, where x_{ij} denotes the expression level of jth gene in ith mRNA sample. The expression levels might be either absolute (e.g. oligonucleotide arrays) or relative with respect to the expression levels of a suitably defined common reference sample (e.g. cDNA microarrays). Usually, the data are centered (mean zero) and/or standardized (mean zero, variance one) for each gene across the samples.

2.1. *Principal component analysis and factor analysis biplot*

The singular value decomposition of X is given by

$$X = UDV',$$

where U and V are $n \times r$ and $p \times r$ matrix, respectively, each with orthonormal columns so that $U'U = V'V = I_r$, D is a $r \times r$ diagonal matrix with elements $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_r$ in the diagonals, and r is rank of X. Let us define $D^\alpha = \mathrm{diag}(\lambda_1^\alpha, \ldots, \lambda_r^\alpha)$ and let $G = UD^\alpha$ and $H = VD^{1-\alpha}$ where $0 \leqslant \alpha \leqslant 1$. Thus X can be factorized as

$$X = UDV' = GH'$$

for a $n \times r$ matrix G and a $p \times r$ matrix H. And it can be shown that $X_{(s)} = G_{(s)}H'_{(s)}$ provides the best possible rank $s(\leqslant r)$ approximation to X, where $G_{(s)}$ and $H_{(s)}$ are the first s columns of G and H, respectively. One can obtain s-dimensional row (sample) and column (gene) plot by plotting $G_{(s)}$ and $H_{(s)}$, respectively (Gabriel, 1971).

Different values of α lead to different geometries. If we choose $\alpha = 1$, then $G_{(s)} = (\lambda_1 u_1, \ldots, \lambda_s u_s)$ and $H_{(s)} = (v_1, \ldots, v_s)$. And the Euclidean distance between two sample points in the plot represents the Euclidean distance in the complete set since $XX' \approx G_{(s)}G'_{(s)}$. Here the ith row of $G_{(s)}$, consists of the first s principal components for ith sample. We call it principal component analysis (PCA) biplot. On the other hand, if we choose $\alpha = 0$ and take the first s columns of G and H, we have the coordinates for s-dimensional plot. We call this factor analysis (FA) biplot. In FA biplot, cosine between gene points is proportional to the covariance or correlation between genes because $X'X \approx H_{(s)}H'_{(s)}$. In both plots, by superimposing sample and gene plot, we can recover original data since $X \approx G_{(s)}H'_{(s)}$.

2.2. *Correspondence analysis biplot*

CA was originally developed for 2-way contingency tables (Greenacre and Hastie, 1987). To analyze using CA, the data should be positive number. Thus it is necessary to shift the data additively to be a positive range after centering and standardization before analysis. Now let $X = (x_{ij})$ be the data matrix after shifting, x_{i+} and x_{+j} denote sum of the ith row and jth column, respectively, and x_{++} be the grand total of X. Define $F = (f_{ij})$ where $f_{ij} = x_{ij}/x_{++}$.