

**Resampling methods for variance estimation of singular value
decomposition analyses from microarray experiments**

Debashis Ghosh

Department of Biostatistics, University of Michigan

Corresponding author:

Debashis Ghosh, Ph.D.

Department of Biostatistics

School of Public Health, University of Michigan

1420 Washington Heights, Room M4057

Ann Arbor, Michigan 48109-2029

Phone: (734) 615-9824

Fax: (734) 763-2215

Email: ghoshd@umich.edu

Abstract

Microarray experiments offer the ability to generate gene expression measurements for thousands of genes simultaneously. Work has begun recently on attempting to reconstruct genetic networks based on analyses of microarray experiments in time-course studies. An important tool in these analyses has been the singular value decomposition method. However, little work has been done on assessing the variability associated with singular value decomposition analyses. In this report, we discuss use of the bootstrap as a method of obtaining standard errors for singular value decomposition analyses. We consider use of this method both when there are replicates and when no replicates exist. The proposed methods are illustrated with an application to data from a recent study by Cho et al. (2001).

Key Words: Bootstrap; Gene expression; Synchronized time-course study.

Introduction

With the recent development of microarray technology, it has become possible to measure gene expression simultaneously on a large scale basis. Because the mRNA levels can be assessed simultaneously for thousands of genes, work has now begun on attempting to elucidate genetic networks and metabolic pathways in various model organisms. One type of experiment that is useful for understanding regulatory mechanisms is the synchronized time-course study. In these experiments, cells are suspended in a certain state and then released. At certain time points, mRNA is taken from the cells, and microarrays are run on the mRNA samples. The result is a set of measurements from chips at various points in time.

One technique proposed by several authors for analyzing microarray time-course data is the singular value decomposition (SVD) (Golub and van Loan, 1996). By using the SVD, the raw, high-dimensional microarray measurements are transformed into a set of independent variables in a lower-dimensional subspace. The uses of SVD analysis of microarray data have been manifold. One proposal involved using SVD as a means of filter and preprocessing the data (Alter et al., 2000). Another use has been to summarize microarray time-course data (Raychaudhuri et al., 2000). A third application has been to summarize the time-course data into so-called “characteristic modes” that are easier to study (Holter et al., 2000). In fact, these authors have suggested that a subset of these modes can explain much of the variation in gene expression dynamics and have begun investigations in dynamic modelling of gene expression data using these quantities (Holter et al., 2001).

However, there has been virtually little mention of variability assessment of SVD analyses. In Holter et al. (2000), a singular value decomposition of a randomly generated dataset was considered as a comparison with that of the real data. However, without any type of formal variance estimation, it is impossible to determine what are “real” patterns in the SVD analyses versus those that arise by chance. Similar issues

arise in hierarchical clustering of microarray data (Zhang and Zhao, 2000; Kerr and Churchill, 2001).

Assessing this variability requires development of a statistical framework for singular value decomposition analyses of microarray time-course data. Our focus will be on performing inference for the characteristic modes. In this article, we propose a non-parametric approach for variance estimation involving the nonparametric bootstrap (Efron and Tibshirani, 1993) and is applicable when there are replicate data available on time-course experiments. In many instances, however, such replicate data are not available. For this scenario, we describe the assumptions made in applying bootstrap methods. We will then apply the proposed methods to data from a human fibroblast study (Cho et al., 2001). While these data were generated using oligonucleotide arrays, we expect that similar considerations should hold using other types of microarrays.

Experimental Methods

Notation and Singular Value Decomposition

Before describing the singular value decomposition, we first introduce some notation. Let x_{ij} denote the gene expression measurement for the j th gene of the i th sample (collected at time t_i), $j = 1, \dots, p$, $i = 1, \dots, n$. Typically, p is on the order of several thousands, while n is on the order of 40-60. We define the vector $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$, where \mathbf{a}' is the transpose of the vector \mathbf{a} , and the $p \times n$ matrix $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_n]$. The singular value decomposition of X is defined in the following manner:

$$\mathbf{X} = \mathbf{A}\mathbf{D}\mathbf{F}, \tag{1}$$

where \mathbf{A} is an $p \times n$ matrix of loadings, \mathbf{D} is a $p \times p$ diagonal matrix of the singular values of \mathbf{X} , and \mathbf{F} is a $p \times p$ matrix. The SVD describes the structure of the matrix \mathbf{X} . For example, the number of nonzero singular values on the diagonal of \mathbf{D} is equivalent to the rank of \mathbf{X} . Furthermore, we have that $\mathbf{A}'\mathbf{A} = \mathbf{I}$ and $\mathbf{F}'\mathbf{F} = \mathbf{F}\mathbf{F}' = \mathbf{I}$, where

\mathbf{I} is an $n \times n$ identity matrix. Typically, the authors have used \mathbf{D} or \mathbf{DF} as the lower-dimensional summary of \mathbf{X} ; the rows of these $p \times p$ matrices correspond to the characteristic modes of \mathbf{X} proposed by Holter et al. (2000). The computation of the singular value decomposition is typically iterative; a good summary of algorithms for performing this task can be found in Golub and van Loan (1996).

Nonparametric bootstrap with replicate data

We first consider the scenario of replicated time-course experiments. In this instance, there will be multiple \mathbf{X} matrices. One can use the bootstrap (Efron and Tibshirani, 1993) to sample the individual columns of \mathbf{X} from the available replicate experiments. Alternatively, one could sample the rows of \mathbf{X} from the experiments. Say we do this B times. For each of the B bootstrapped gene expression matrices, we then apply the SVD (1). This yields a set of B bootstrapped DF matrices. We can then plot the characteristic modes for the observed data, along with the corresponding modes for the bootstrapped datasets. We can use the bootstrapped datasets to construct pointwise confidence intervals for the modes at each of the time points. Since we are considering many time points simultaneously, we take a conservative approach with respect to cutoff values. We normally generate $B = 10,000$ bootstrap samples and take the confidence limits to be based on the 0.01th percentile and the 99.9th percentile of the distribution of characteristic modes at each time point. It is important to note that these procedures make no assumption about the dependence of gene expression measurements within and across the p columns of \mathbf{X} .

Nonparametric bootstrap with no replicate data

In many situations, however, we only have data from one time course experiment and no replicate experiments. Suppose we apply the bootstrap procedures described in the previous paragraph. If we generate bootstrapped datasets by resampling from the rows

or of \mathbf{X} , then we assume that the correlation between any two genes are the same. If we resample from the columns of \mathbf{X} , then there is an assumption that for all genes, the correlation between the gene expression measurements at any two time points is the same. These are important assumptions that need to be carefully examined. By not replicating the time-course experiment, we lose the flexibility to assume arbitrary dependence among genes and among experiments across time points. In addition, any effects due to time are confounded with the between-chip variation. Such aspects needed to be taken into account when analyzing the data from a single time-course experiment.

Implementation

As was mentioned before, any of the standard software packages that fit the singular value decomposition can be used to implement the methods described in the Algorithms section. All of these methods presented in the Results section using the R language, a public domain statistical software package that can be downloaded at the following website: <http://cran.r-project.org/> . The commands used to analyze the data can be found at the author's webpage, at the following URL:

<http://www.sph.umich.edu/~ghoshd/COMPBIO/SVD/index.html>.

Results

Human fibroblast data

We apply the resampling ideas described above to a series of experiments conducted by Cho et al. (2001). In these experiments, primary fibroblasts were prepared from human foreskin and then arrested in the late G_1 stage using a thymidine-block protocol (Rao and Johnson, 1970). The cells were then released and collected every two hours for 24 hours. Using high-density oligonucleotide arrays, mRNA was measured at 12

time points. The authors of the study carried out the entire experiment in duplicate; we will use their notation and refer to the two experiments as N2 and N3. While Cho et al. (2001) were interested in determining cell-cycle regulated transcripts, they averaged the data from the N2 and N3 experiments. we use their data simply to illustrate the techniques described in the algorithms. The data can be found at the following website: <http://www.salk.edu/docs/labs/chipdata>.

While there are measurements available on 7129 genes, we excluded genes that had fewer than two positive expression measurements for either N2 or N3. This was done because the transcript levels for these genes were so low that it would be difficult to distinguish signal from background expression. This left a total of 5914 genes. The next step was to apply a transformation of $\log(x+200)$ using base 2; this was done to stabilize the distribution of the gene expression measurements. Then the genes on each chip were centered and scaled to have mean 0.0 and variance 1. This was done to adjust for between-chip variation in hybridization. The second transformation corrected for difference in between-gene variation by centering and scaling the measurements such that for gene, the mean is 0.0 and variance is 1.0.

We first utilize the replicate data from both N2 and N3 and apply the bootstrap. We generated 10,000 bootstrapped datasets for each study. We did two separate scenarios; in the first, the columns of the matrix were resampled while for the second, the rows of the matrix were resampled. The first four characteristic modes for the observed data (based on N2), along with the bootstrapped confidence intervals, are given in Figures 1 for the first scenario and in Figure 2 for the second scenario. The modes themselves appear to suggest periodic variation. However, much of the observed patterns in the modes can be explained by chance under both the first scenario. Similar types of results hold using the second resampling scenario.

[Figure 1 about here]

[Figure 2 about here]

We now use only the data from N2 so that we have no replicate data available. In Figure 3, we show the results for the first four characteristic modes and the associated confidence intervals when the columns of the gene expression matrix are used for re-sampling. Note that there is an assumption in this procedure that for each gene, the correlation between measurements at any two time points is the same. In addition, we are unable to incorporate the variability from replicate time-course experiments. Comparing these graphs with those in Figures 1 and 2 suggests that there is substantial variability between replicate time-course experiments. However, we still find that much of the observed trends in the characteristic modes are consistent with results that are attributable to chance.

[Figure 3 about here]

Discussion

The goal of this paper is to bring out the importance of assessment of variability of singular value decomposition analyses in microarray experiments. In most instances, microarray data are not sufficient for determination of patterns relative to those that can be generated by chance. Since data from high-throughput technologies are extremely multivariate, there will be a large number of trends that will arise due to randomness. It is important not to treat these findings as confirmatory without statistical and/or external experimental validation.

The methods proposed here would be useful in the situation where replicate time-course data exist. While replicating microarray time-course experiments can be expensive, such an analysis could still be performed if one were performing a meta-analysis of time-course experiments using a publicly available microarray database. However, in many situations, it is not feasible to perform replicate time-course experiments. It would be desirable to develop model-based techniques for singular value decomposition analyses of microarray time-course data in the absence of replicate experimental data.

In addition, such an approach can allow for performing inference on the number of characteristic modes. We are currently looking into this and plan to communicate our results in a separate report.

Another issue has to do with the spacing of the time-course experiments. The information on the time points at which the measurements are taken is not utilized in the singular value decomposition analysis. However, in practice, measurements will be taken at potentially irregularly spaced time points. Aach and Church (2001) have proposed using dynamic programming-type algorithms for aligning gene expression time series data. An alternative approach would be to treat the sequence of gene expression measurements over time as a function and to consider functional singular value decomposition techniques for analyzing gene expression dynamics. This is another interesting area for research that we are currently exploring.

References

- Aach, J. and Church, G. M. (2001). Aligning gene expression time series with time warping algorithms. *Bioinformatics* **17**: 495–508.
- Alter, O., Brown P. O. and Botstein D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences* **97**, 10101–10106.
- Cho, R. J, Huang, M., Campbell, M. J., Dong, H., Steinmetz, L., Sapinoso, L., Hampton, G., Elledge, S. J., Davis, R. W., and Lockhart, D. J. (2001). Transcriptional regulation and function during the human cell cycle. *Nature Genetics* **27**, 48–54.
- Efron, B. and Tibshirani, E. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.
- Golub, G. H. and van Loan, C. F. (1996). *Matrix Computations*. Baltimore: John Hopkins University Press.
- Holter, N. S., Maritan, A., Cieplak, M., Fedoroff, N. V., and Banavar, J. R. (2001). Dynamic modeling of gene expression data. *Proceedings of the National Academy of Sciences* **98**, 1693–1698.
- Holter, N. S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J. and Fedoroff, N. V. (2000). Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proceedings of the National Academy of Sciences* **97**, 8409–8414.
- Kerr, M. K. and Churchill, G. A. (2001). Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Sciences* **98**, 8961–8965.
- Rao, P. N. and Johnson, R. T. (1970). Mammalian cell fusion: studies on the regulation of DNA synthesis and mitosis. *Nature* **225**: 159–164.

Zhang, K. and Zhao, H. (2000). Assessing reliability of gene clusters from gene expression data. *Functional and Integrative Genomics* **1**, 156–173.

Figure 1. Plot of first four characteristic modes (solid line) for N2 data from Cho et al. (2001), along with 99% pointwise confidence intervals. Intervals obtained by applying bootstrap to columns of data matrices from N2 and N3.

Figure 2. Plot of first four characteristic modes (solid line) for N2 data from Cho et al. (2001), along with 99% pointwise confidence intervals. Intervals obtained by applying bootstrap to rows of data matrices from N2 and N3.

Figure 3. Plot of first four characteristic modes (solid line) for N2 data from Cho et al. (2001), along with 99% pointwise confidence intervals. Intervals obtained by applying bootstrap to columns of data matrices from N2.





