

On the Application of Principal Component Analysis to DNA Microarray Time Series Data

Kevin Chu

December 20, 2001

1 Introduction

Advances in DNA microarray and gene chip technologies have made it possible to simultaneously measure the gene expression level for thousands of genes. Because microarrays and gene chips produce huge amounts of data compared with traditional experimental genetic methods, new analysis techniques need to be developed to extract meaningful biological information.

Early techniques for analyzing this large amount of genetic data focused on clustering genes as a means to investigate multiple gene effects or gain insight into the functionality of poorly characterized genes [1, 2]. These techniques were primarily based on statistical clustering algorithms and often used correlations in gene expression (or some variation) across multiple cell samples as the metric for gene similarity.

More recently, there has been an effort to examine gene expression data from a global perspective [3, 4, 5, 6]. Rather than seeking a data reduction by reducing the number of genes to focus on, data reduction is achieved by identifying dominant global genetic expression states and using these to group the experimental samples. Principal component analysis (PCA) is the main tool that has been used to carry out this data reduction. Unfortunately, the criteria used to distinguish between biologically significant groupings and experimental noise do not seem to have a solid statistical foundation.

One particularly popular application of PCA has been to analyze time series microarray data. This paper reviews the current PCA-based methods for analyzing time series data, discusses theoretical and experimental problem areas, and proposes statistical tests that may potentially be able to address some of the statistical significance issues.

2 Source of experimental data

DNA microarray technology is based on the the unique matching between complementary DNA (cDNA) and messenger RNA (mRNA) strands. First, the DNA microarray is constructed by placing known cDNA strands in individual wells on a slide. Next, the cell sample to be studied is prepared by tagging the mRNA within the cell with a fluorescent marker (typically red). Concurrently, the mRNA of reference cells are tagged with a different fluorescent marker (typically green). The mRNA from the sample and the reference cells are then mixed and allowed to simultaneously hybridize with the cDNA fragments on the slide. After some experimental work up, the fluorescence levels of the sample and reference tags are measured for each well on the slide. Finally, the gene expression level for each gene is computed as

$$\log(I_{\text{sample}}/I_{\text{ref}}),$$

where I is the fluorescence level.

For analysis, the gene expression data is conventionally organized in an $n \times m$ matrix where n is the number of genes whose expression is measured and m is the number of experimental samples.

3 PCA-Based analysis of DNA microarray time series data

Before examining the application of principal component analysis to DNA microarray data, it is useful to review conventional PCA and to consider what kinds of results it produces for simple time series data.

3.1 Conventional PCA

The standard use of principal component analysis in data analysis attempts to find a low dimensional subspace of the space of centered experimental variables (*i.e.* mean removed) that accounts for a majority of the variance observed in the data. Another way to think of this is that the procedure tries to find “best fit” hyper-plane of low dimensionality for the data. It is important to note that this analysis takes place in *dependent* variable-space. The associated fit that occurs in “experiment”-space does not have the same meaning as the fit in variable-space [7]. Intuitively, the analysis is not symmetric because dependent and independent variables are not interchangeable.

However, useful information can be extracted from the fit in experiment-space. The components of the k -th singular vector in experiment-space are the relative magnitudes of measurements of the k -th principal component across the different experiments. In other words, if the *created* dependent variable associated with the k -th principal component had been measured in each experiment, the relative magnitudes in the experiments would match the relative magnitudes of the components of the k -th singular vector in experiment-space. In traditional PCA, this observation is not very interesting because the goal is to find a descriptive relationship between the dependent variables. Furthermore, since there are usually very many more data points than variables, the fit in experiment-space is probably significantly affected by any noise in the data.

3.2 Application of PCA to time series data

For time series data, the data fit generated in experiment-space takes on greater meaning because the order of the components in the k -th experiment-singular vector is significant. In this context, the components of the k -th experiment-singular vector can be interpreted as the time series that would arise by making measurements on the k -th principal component. That is, the k -th experiment-singular vector gives the time evolution of the k -th principal component. Furthermore, the time series for different principal components are orthogonal. The orthogonality can be useful in defining normal “modes” for the dynamics of the data. For the remainder of this paper, the k -th experiment-singular vector will be referred to as the k -th *principal time series* to emphasize its dynamical interpretation.

3.2.1 Sinusoidal data

Because current applications of PCA to microarray time series data often lead to a decomposition of the dynamics into sinusoidal modes, it is useful to consider what kinds of results PCA gives for data made up of only a few frequencies.

Single frequency data Suppose that there are n variables which are all varying at the same frequency, f , but at arbitrarily different amplitudes and phase shifts: $A_i(t) = \alpha_i \sin(2\pi ft + \phi_i)$. Let A_{it} represent the value of the i -th variable sampled at the t -th time point. Organizing the data into a

matrix, we have

$$A = \begin{bmatrix} A_1 \\ \vdots \\ A_n \end{bmatrix}$$

where A_i is the time series for the i -th variable. Then, PCA gives as its first two principal time series, a sine-cosine pair with frequency f and same phase shift (Figure 1). The variance in the remaining principal components is at the level of machine error, so they can all be disregarded. This result makes sense because the data vector $\mathbf{x}(t) = (A_1(t), \dots, A_n(t))$ traces out an ellipse in \mathbf{R}^n . Thus, the data points can be collapsed onto a two dimensional hyper-plane,

$$\mathbf{x}(t) = \sigma_1 U_1 V_1' + \sigma_2 U_2 V_2' \quad (1)$$

where U_i and V_i are the i -th singular vectors of the data matrix, A . Because the ellipse is traced out in time, the principal time series V_1 and V_2 must be a sine-cosine pair of with the same frequency and phase shift. Furthermore, the entire ellipse is traced out in one period the n variables, so V_1 and V_2 must have frequency f .

Multiple frequency data For data which is a linear combination of multiple sines waves, PCA produces results are not as clean as in the single frequency case. In general, the principal time series are still linear combinations of sine and cosine waves (Figure 2). However, if the amplitudes of the different frequency sine waves are sufficiently different, the principal time series come in sine-cosine pairs ordered by the magnitude of the fourier coefficient in the continuous signals (Figure 3). Intuitively, this result makes sense because when the true signal is a linear combination of multiple sine waves where different frequencies have widely differing magnitudes, the data vector $\mathbf{x}(t) = (A_1(t), \dots, A_n(t))$ roughly traces out a perturbed ellipse in \mathbf{R}^n .

Effects of data quality As with the analysis of any time series data, the frequency and duration of data sampling places severe limitations on the amount of information that can be extracted from the data. Since extracting principal time series is related to spectral analysis, it is not unreasonable to expect the limitations that arise in spectral analysis to arise in PCA. Two important data properties are sampling frequency and sampling window [8]. As in fourier analysis, the sampling frequency limits the range of frequencies that can be observed and the sampling window limits the achievable

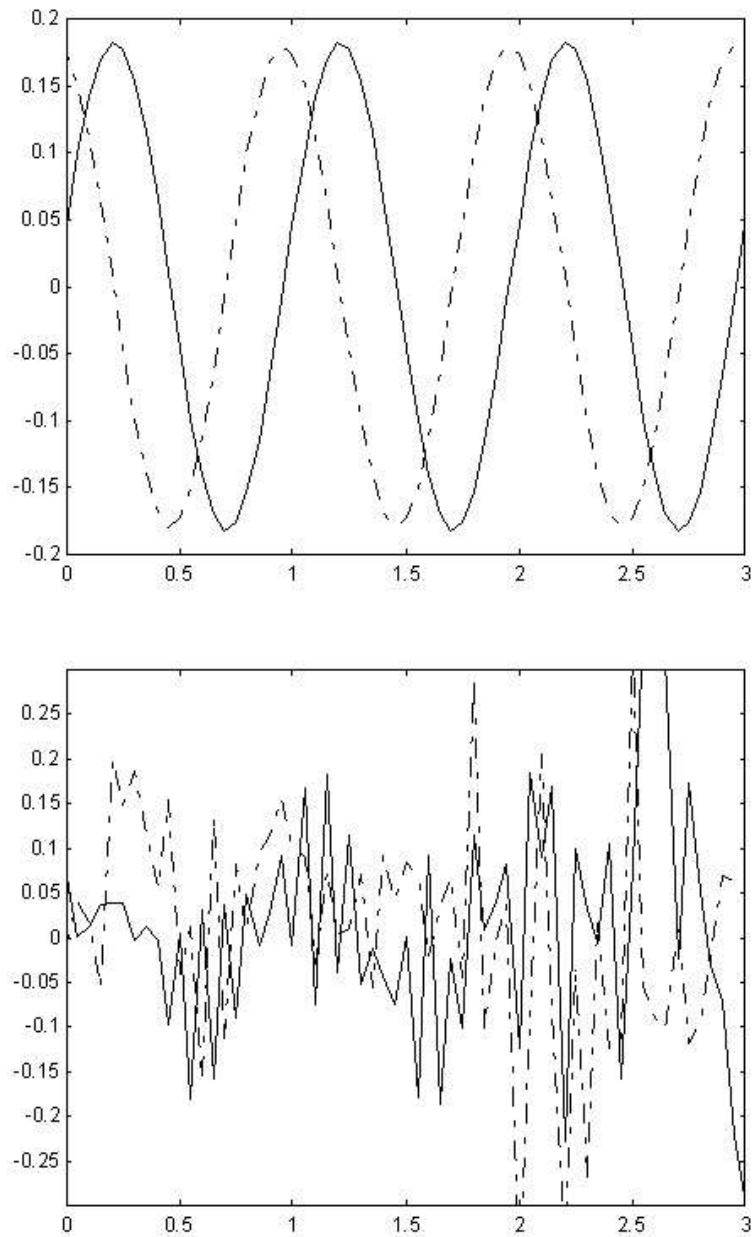


Figure 1: The two graphs show the principal time series corresponding to the largest two (top) and two other arbitrarily selected (bottom) singular values for a generated data set consisting of 100 genes which are all sinusoids varying at a frequency of $f = 1$ with random amplitudes and phase shifts. Notice that the dominant principal time series form a sine-cosine pair while the other “modes” are dominated by noise. The sampling frequency and time window are 20 Hz and 3 seconds, respectively.

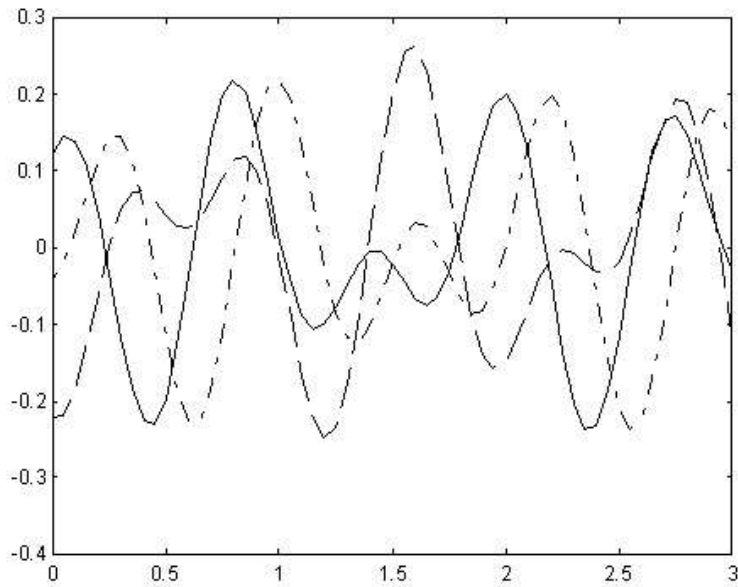


Figure 2: The graph shows the principal time series corresponding to the largest four singular values for a generated data set consisting of 100 genes which are all varying as linear combinations of the form $\sin(2\pi t + \phi_1) + \sin(\pi^2 t + \phi_2)$ where ϕ_1 and ϕ_2 are independent, gene dependent phase shifts. Notice that the dominant principal time series do not cleanly separate into sine-cosine pairs. The sampling frequency and time window are 20 Hz and 3 seconds, respectively.

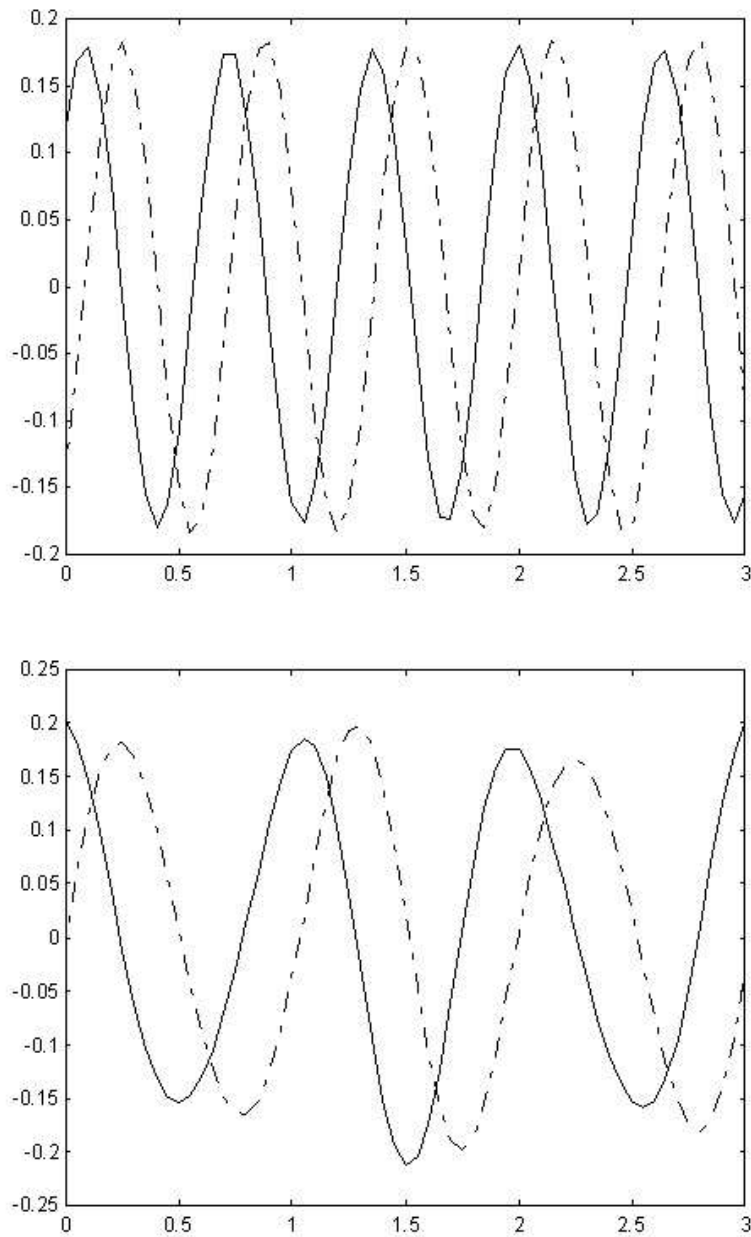


Figure 3: These two graphs shows the principal time series corresponding to the largest pair (top) and second largest pair (bottom) singular values for a generated data set consisting of 100 genes which are all varying as linear combinations of the form $\sin(2\pi t + \phi_1) + 10\sin(\pi^2 t + \phi_2)$ where ϕ_1 and ϕ_2 are independent, gene dependent phase shifts. Notice that the principal time series are fairly well separated into sine-cosine pairs. The sampling frequency and time window are 20 Hz and 3 seconds, respectively.

frequency resolution. Furthermore, a small sampling window could make it difficult to distinguish between very slow varying signals and a “monotonic” signal (such as those that were observed by Holter *et al.* [3]). For current microarray data, the sampling frequency is so low that it would be difficult to detect any dynamics that occur faster than a few times the frequency of the cell cycle. In addition, the data extends only over a few cell cycles, so it would be difficult to resolve closely spaced frequencies.

3.3 Analysis of DNA microarray time series data

Many recent analyses of DNA microarray time series data based on a PCA approach seem to be trying to extract principal time series information from the data. Part of the reason for this may be nature of microarray data – there are far fewer data points than variables. This situation is the reverse of the data which is commonly analyzed using PCA. Because there are so few data points in microarray data, analysis in the gene expression-space is inherently noisy. The subspace determined by *all* of the data points is already of very low dimension compared to the dimensionality of the entire variable-space. As a result, it seems that researchers may have looked past the asymmetry between the variable and experiment spaces in applying PCA.

Currently, there is no consensus in how PCA should be applied to the data or how the results should be interpreted. Two main points of contention are:

1. the type of “normalization” that should be done on the data,
2. the criterion for discarding singular vectors, and

Data normalization The goal of data normalization seems to be the removal of the steady state gene expression from the data set. The competing methods are:

1. for each gene, compute the average expression of the gene across experiments and subtract this value from all the expression values for that gene [3], and
2. after computing the singular value decomposition (SVD) for the data set, discard the principal component associated with the largest singular value [4, 5].

Some sources also suggest that the average gene expression for each experiment be removed so that all columns have zero mean [3]. However, such a procedure seems inconsistent with removing the mean for each gene (*i.e.* rows having zero mean) and does not have a clear “biological” meaning. From a traditional, PCA perspective, the procedure of removing the mean on a gene by gene basis is the one that makes most sense. While the principal component with the largest singular gives an approximation of the mean gene expression for each gene, there does not seem to any gain to using the SVD to compute the mean when direct computation is more accurate. Furthermore, procedure (2) has the “unphysical” side effect of forcing all variations around the mean genetic state to be orthogonal to the mean.

Criterion for discarding principal components Once PCA has been applied to the gene expression data, principal components and principal time series associated with “small” singular values are discarded. However, there is no consistent way that this is done; the cut-off value used to determine significance is quite arbitrary [5]. Raychaudhuri *et al.* chose to use a $(70/n)\%$ rule, discarding any component that accounted for less than $(70/n)\%$ of the total variance [5]. Holter *et al.* take an alternative approach of rejecting all principal components with singular values that are not “significantly” larger than the singular values of a appropriately chosen random matrix [3]. In all of these approaches, it is rare that any principal components other than the top two are retained. However, in some cases it is ambiguous whether more principal components should be kept [3, 4].

While the current approaches may seem to lack adequate rigor, they seem to be in the right spirit. The remainder of this paper will discuss some statistical tests that may be able help determine which results are significant.

4 Statistical significance tests

The singular values of the data matrix, A , are critical in PCA analysis of microarray data, so it is natural to consider statistics based on the Wishart distribution. However, because microarray data is such that there are more variables than experiments, organization of the data matrix so that it is “tall and thin” requires that rows represent genes and columns represent time points. So, normalizing the data so that each gene has zero mean expression does not produce columns that have zero mean. Thus, any statistical tests must be based on the noncentral Wishart distribution. For convenience of

notation in the following discussion, we adopt Muirhead's notation $\text{etr}(A)$ to denote $\exp(\text{tr}(A))$ [9].

4.1 The noncentral Wishart distribution

Definition: Let Z be a $n \times m$ matrix distributed as $N(M, I_n \otimes \Sigma)$ with M a $n \times m$ and Σ a $m \times m$ positive definite, symmetric matrix. Then the $m \times m$ matrix, $A = Z'Z$, has *noncentral Wishart distribution* with n degrees of freedom, covariance matrix Σ , and matrix of noncentrality parameters $\Omega = \Sigma^{-1}M'M$. The distribution of A is denoted by $W_m(n, \Sigma, \Omega)$. If $n \geq m$, then the density function of A is

$$\frac{1}{2^{mn/2} \Gamma_m(\frac{1}{2}n) (\det \Sigma)^{n/2}} \text{etr}\left(-\frac{1}{2}\Sigma^{-1}A\right) (\det A)^{(n-m-1)/2} \text{etr}\left(-\frac{1}{2}\Omega\right) {}_0F_1\left(\frac{1}{2}n; \frac{1}{4}\Omega\Sigma^{-1}A\right). \quad (2)$$

In the case where $M = \mathbf{1}\mu'$ where $\mu \in R^m$ is the expected value of each row, Ω takes on the special form $\Omega = n\Sigma^{-1}\mu\mu'$.

The joint density for eigenvalues of A may be obtained by using the following

Theorem: Let a $m \times m$ positive definite matrix have density $f(A)$. Then the joint density of the eigenvalues, l_1, l_2, \dots, l_m of A is

$$\frac{\pi^{m^2/2}}{\Gamma_m(\frac{1}{2}m)} \prod_{i < j} (l_i - l_j) \int_{O(m)} f(HLH') (dH) \quad (3)$$

where $l_1 > l_2 > \dots > l_m$ and $L = \text{diag}(l_1, l_2, \dots, l_m)$ [9].

Thus, for a matrix A from a noncentral Wishart distribution, we find that

$$\frac{\pi^{m^2/2}}{2^{mn/2} \Gamma_m(\frac{1}{2}m) \Gamma_m(\frac{1}{2}n) (\det \Sigma)^{n/2}} \prod_{i=1}^m l_i^{(n-m-1)/2} \prod_{i < j} (l_i - l_j) \cdot \text{etr}\left(-\frac{1}{2}\Omega\right) \int_{O(m)} \text{etr}\left(-\frac{1}{2}\Sigma^{-1}HLH'\right) {}_0F_1\left(\frac{1}{2}n; \frac{1}{4}\Omega\Sigma^{-1}HLH'\right) (dH). \quad (4)$$

When analyzing a sample covariance matrix, it is nS that has distribution $W_m(n, \Sigma, \Omega)$. If l_1, l_2, \dots, l_m are the eigenvalues of S then the joint density of the eigenvalues, is given by equation (4), with the change of variables $l_i \rightarrow nl_i$:

$$\left(\frac{n}{2}\right)^{nm/2} \frac{\pi^{m^2/2}}{\Gamma_m(\frac{1}{2}m) \Gamma_m(\frac{1}{2}n) (\det \Sigma)^{n/2}} \prod_{i=1}^m l_i^{(n-m-1)/2} \prod_{i < j} (l_i - l_j) \cdot \text{etr}\left(-\frac{1}{2}\Omega\right) \int_{O(m)} \text{etr}\left(-\frac{1}{2}n\Sigma^{-1}HLH'\right) {}_0F_1\left(\frac{1}{2}n; \frac{1}{4}n\Omega\Sigma^{-1}HLH'\right) (dH). \quad (5)$$

As for the central Wishart distribution, it should be possible to obtain an asymptotic approximation to this density in the limit of large n using the multivariate generalization of Laplace's method described in Muirhead [9]. Unfortunately, the hypergeometric function ${}_0F_1$ does not have as simple a form as ${}_0F_0$. Furthermore, finding an asymptotic approximation for ${}_0F_1(\frac{1}{2}n; \frac{1}{4}n\Omega\Sigma^{-1}A)$ based on its integral representation is complicated by the fact that the orthogonal group that the integral is taken over has dimension equal to the limiting variable, n .

4.2 Sphericity test

For conventional PCA, the *Sphericity test* is commonly applied to determine if only the first k eigenvalues of the true covariance matrix for the variables are different. If this is found to be the case and if the estimate of the value of the last $m - k$ eigenvalues is small compared to the k -th eigenvalue, the last $m - k$ principal components are discarded.

The test proceeds in stages. At stage k , the hypothesis, H_k , that the last $m - k$ are equal is tested. The value of k is incremented until H_k is not rejected or $k = m$. If H_k passes at some stage with $k < m$, the conclusion is that the first k eigenvalues differ but the last $m - k$ are equal. If testing terminates with $k = m$, the conclusion is that all m eigenvalues of the true covariance matrix are different, so no data reduction is possible.

Suppose that A is a $N \times m$ data matrix with rows representing data samples and columns representing variables:

$$A = \begin{bmatrix} A_1 \\ \vdots \\ A_N \end{bmatrix}.$$

Let

$$S = \frac{1}{n} \sum_{i=1}^N (A_i - \bar{A})'(A_i - \bar{A}) \quad (n = N - 1)$$

be the sample covariance matrix for A , and let l_1, l_2, \dots, l_m be the eigenvalues of S . Then the likelihood ratio statistic for testing the k -th null hypothesis,

$$H_k : \lambda_{k+1} = \dots = \lambda_m \quad (= \lambda, \text{ unknown})$$

is

$$\begin{aligned}
V_k &= \frac{\prod_{i=k+1}^m l_i}{\left(\frac{1}{m-k} \sum_{i=k+1}^m l_i\right)^{m-k}} \\
&= \left[\frac{\left(\prod_{i=k+1}^m l_i\right)^{\frac{1}{m-k}}}{\frac{1}{m-k} \sum_{i=k+1}^m l_i} \right]^{m-k}.
\end{aligned} \tag{6}$$

Note that this is the ratio of the geometric mean and arithmetic mean of the eigenvalues. Alternatively, this statistic could have been written in terms of the singular values of A :

$$V_k = \left[\frac{\left(\prod_{i=k+1}^m \sigma_i^2\right)^{\frac{1}{m-k}}}{\frac{1}{m-k} \sum_{i=k+1}^m \sigma_i^2} \right]^{m-k} \tag{7}$$

where $\sigma_1, \dots, \sigma_m$ are the singular values of A . It can be shown that as $n \rightarrow \infty$, the asymptotic distribution of $-n \log V_k$ is $\chi_{(m-k+2)(m-k-1)/2}^2$ [9].

For DNA microarray data analysis, the sphericity test must be modified because the columns of the data matrix no longer have zero mean and removing column means would destroy the property that rows have zero mean. Thus, we must construct a sphericity test based on the noncentral Wishart distribution.

In the sphericity test, the hypothesis

$$H_k : \lambda_{k+1} = \dots = \lambda_m \quad (= \lambda, \text{ unknown})$$

is being tested against the alternative that there is no relation between $\lambda_{k+1}, \dots, \lambda_m$. Thus, the likelihood ratio statistic is

$$\Lambda = \frac{\sup L(M, \hat{\Sigma})}{\sup L(M, \Sigma)} \tag{8}$$

where $\hat{\Sigma}$ is allowed to vary over all positive definite, symmetric matrices that have $\lambda_{k+1} = \dots = \lambda_m = \lambda$ (λ unspecified) and Σ is allowed to vary over all positive definite, symmetric matrices. Since the likelihood function is just the probability density function under a given set of assumptions on the parameters of the distribution, the likelihood function, $L(M, \Sigma)$ for a matrix with a noncentral Wishart distribution is given by the expression in equation (2).

Theoretically, it should be possible to compute the probability distribution for the likelihood ratio statistic, Λ . With this distribution (or an

asymptotic approximation), it should be straightforward to compute a rejection cut-off, Λ_0 , for Λ with significance level

$$\alpha = \Pr(\Lambda \leq \Lambda_0 \mid H_k). \tag{9}$$

4.3 Crude singular value test

To determine whether the k largest singular values are significant, the k -th singular value could be used as a crude statistic. For a data matrix A that is distributed $W_m(n, \Sigma, \Omega)$, it should be possible to compute an asymptotic distribution for the largest singular value, σ_1 . To test the hypothesis that the k largest singular values are significant, the k -th largest singular value could be compared to a critical value, c , which is determined by the criterion:

$$\alpha = \Pr(\sigma_1 > c \mid W_m(n, \Sigma, \Omega)) \tag{10}$$

where α is the “significance level” of the test. If $\sigma_k > c$, then $\sigma_1, \dots, \sigma_k$ are taken to be significant because the probability that $\sigma_k > c$ is less than the probability $\sigma_1 > c$.

4.4 Open issues

In both the sphericity test and crude singular value test, it is likely that there will be some unspecified parameters in the probability distribution of the statistics. In particular, the mean matrix, M , and the covariance matrix, Σ , need to be specified. For the case where $M = \mathbf{1}\mu$, μ will need to be specified instead. For microarray data, a naive estimate for these parameters might be the column means and covariance matrix. However, further investigation would be needed to determine the appropriateness of these estimates.

5 Conclusions and future directions

Principal component analysis holds promise as a method for analyzing DNA microarray data. For producing principal time series, it appears to be useful in extracting orthogonal temporal “modes” of genetic variation. However, further testing on higher quality data sets will need to be done before a solid conclusion can be reached. In addition, further research (possibly just journal research) of the noncentral Wishart distribution is necessary in order to develop the statistical tools necessary to analyze the significance of PCA results.

While current data only allows genetic dynamics at the frequency of the cell cycle to be studied, it is likely that much of the interesting dynamics occurs at higher frequency and much lower amplitudes. From the resilience of living organisms, it makes sense that there should be only a few very dominant modes in genetic dynamics. However, nontrivial cellular response to environmental stimuli are probably hidden in small perturbations (of both higher and lower frequency) to the stable genetic dynamics of the cell cycle.

For both the development of data analysis techniques and deeper understanding of gene dynamics, it will be necessary to obtain higher quality data than is currently available. In particular, data taken at higher frequency and over more cell cycles will be needed to improve the frequency resolution and range that can be studied. However, experimental limitations may make collecting this data difficult.

References

- [1] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein. Cluster Analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.*, 95:14863-14868, 1998.
- [2] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286:531-537, 1999.
- [3] Neal S. Holter, Madhusmita Mitra, Amos Maritan, Marek Cieplak, Jayanth R. Banavar, and Nina V. Fedoroff. Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *Proc. Natl. Acad. Sci.*, 97:8409-8414, 2000.
- [4] Orly Alter, Patrick O. Brown, and David Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci.*, 97:10101-10106, 2000.
- [5] Soumya Raychaudhuri, Joshua M. Stuart, and Russ B. Altman. Principal components analysis to summarize microarray experiments: Application to sporulation time series.
- [6] T. Gregory Dewey and David J. Galas. Dynamic models of gene expression and classification. *Funct. Integr. Genomics*, 1:269-278, 2001.
- [7] Ludovic Lebart, Alain Morineau, and Kenneth M. Warwick. *Multivariate Descriptive Statistical Analysis*, John Wiley & Sons, New York, NY, 1984.
- [8] Alan V. Oppenheim, Ronald W. Schaffer, and John R. Buck. *Discrete-Time Signal Processing*, Prentice Hall, Upper Saddle River, NJ, 1999, p. 693-706.
- [9] Robb J. Muirhead. *Aspects of Multivariate Statistical Theory*, John Wiley & Sons, New York, NY, 1982.
- [10] Robert B. Hogg and Allen T. Craig *Introduction to Mathematical Statistics*, Prentice Hall, Upper Saddle River, NJ, 1995.