Proceedings of the 6th Nordic Signal Processing Symposium - NORSIG 2004
June 9 - 11, 2004
Espoo, Finland

Disease Detection Technique Using the Principal Orthogonal Decomposition on DNA Microarray Data

David Peterson and Charles H. Lee
Department of Mathematics
California State University, Fullerton

Abstract—In this paper we demonstrate the feasibility in disease detection using a pattern recognition technique on DNA data expressed in microarrays. Specifically, we employ the Principal Orthogonal Decomposition (POD) method (also known as the Karhunen-Loève method) to extract the characteristics of a disease from a collection of DNA samples of individuals who suffer the disease. The resulting primary component captures the dominant features of the original samples. Such representatives are then correlated with an arbitrary sample to determine whether the sample carries the disease. Though the approach can be applied to other diseases, in our study we showed that the POD method could be applied to the DNA microarrays to positively detect liver and bladder cancers.

TABLE OF CONTENTS

1. INTRODUCTION

DNA microarrays store the expressions of thousands of individual genes on a single surface that is about the size of a microscope slide. Such image allows one to see genes that are induced or repressed in an experiment. As a result, signatures of a disease may be encrypted in DNA microarrays, and once found, can be used for diagnoses. Our goal in this study is to apply a pattern recognition technique, called the *principal orthogonal decomposition*, to extract the characteristics of a disease from an ensemble of samples known to carry the disease and to use the extracted feature for disease detection.

Our significant achievement is to demonstrate that the POD technique can be applied to DNA microarray data collected from cancerous tissue samples to detect liver and bladder cancers. Namely, we obtain the two sets of DNA microarray data from the liver cancer [1] and the gastric cancer [2] studies. Both sets are stored in the Stanford Microarray Database, genome-www5.stanford.edu. The detail can be found in the case-studies section. It is noteworthy to mention that although our study focuses on liver and bladder cancers, the method is not necessarily restricted to these types of diseases.

2. MATHEMATICAL FORMULATION FOR POD

2.1 Principal Orthogonal Decomposition

This section provides a brief summary of the Principal Orthogonal Decomposition (POD) method. The POD method has received much attention in recent years as a tool to reduce the complexity and dimensions of dynamical models in engineering and science [3]-[5]. In principle, one begins with an ensemble of data, called *snapshots*, collected from an experiment or a numerical procedure characterizing a physical system. The POD technique is then used to produce a few principal elements that can be used to reconstruct the entire snapshot collection.

In the POD method we are given a series of *images* or *snapshots*, $\{V_i(\vec{x})\}_{i=1}^{n_s}$. Our goal is to construct a set of basis functions $\{\Phi_i(\vec{x})\}_{i=1}^{n_s}$, whose first few elements can capture all the dominant features of the entire snapshots collection. In other words, the primary component Φ_1 captures *most* of the essential features of the original ensemble, while subsequent basis elements capture more of the smaller and finer variability between the snapshots. As a result, we wish to choose the primary component Φ_1 such that the quantity

$$\sum_{i=1}^{n_s} |\langle V_i, \Phi_1 \rangle|^2 \tag{1}$$

is as large as possible with $\langle \cdot, \cdot \rangle$ denotes the inner product. It is natural to assume Φ_1 to be of the linear combination of the snapshots,

$$\Phi_1(\vec{x}) = \sum_{j=1}^{n_s} w_j V_j(\vec{x}), \tag{2}$$

where $\vec{w} = [\, w_1 \quad w_2 \quad \cdots \quad w_{n_s} \,]^T$ is the weighting vector assigned to the snapshots. Thus maximizing the quantity in (1) is equivalent to maximizing the following

$$\| \theta \vec{w} \|^2 = \vec{w}^T \theta^2 \vec{w}, \tag{3}$$

where θ is the covariance matrix of the snapshots with its (i, j) component, $\theta_{i,j}$, defined by

$$\theta_{i,j} = \langle V_i, V_j \rangle, \quad i = 1, \cdots, n_s, j = 1, \cdots, n_s. \tag{4}$$

Note that with distinct snapshots, the covariance matrix θ is symmetric positive definite and thus the weighting vector that maximizes (3) will also maximize

$$J(\vec{w}) = \vec{w}^T \theta \vec{w}, \quad (5)$$

In this process, the weighting vector for the primary component is exactly the dominant eigenvector of θ corresponding to the largest eigenvalue. Let us denote the eigenpairs of θ by $\left\{\left(\lambda_j, \vec{u}^{(j)}\right)\right\}_{j=1}^{n_s}$ and sort the eigenvalues in decreasing order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{n_s} \geq 0$. It follows that

$$\Phi_j(\vec{x}) = \sum_{i=1}^{n_s} u_i^{(j)} V_i(\vec{x}), \quad (6)$$

and

$$\sum_{i=1}^{n_s} \left|\langle V_i, \Phi_1 \rangle\right|^2 \geq \sum_{i=1}^{n_s} \left|\langle V_i, \Phi_2 \rangle\right|^2 \geq \cdots \geq \sum_{i=1}^{n_s} \left|\langle V_i, \Phi_{n_s} \rangle\right|^2 . \quad (7)$$

2.2 Projection and Pattern Recognition

In our study, we seek solely Φ_1 and once the primary component of a set of images has been determined it can be used for comparison with other images. Projection forms a simple way of implementing the comparison. If V is an arbitrary image to be tested, then

$$P_\Phi(V) = \frac{\langle V, \Phi_1 \rangle}{\langle \Phi_1, \Phi_1 \rangle}, \quad (8)$$

measures of the correlation between V with Φ_1. The larger the magnitude of $P_\Phi(V)$ the greater the correlation there is between the image V and the original set of images.

3. CASE STUDIES

As an application of the method, we examined DNA microarray data from references [1] and [2]. The data were obtained from the Stanford Microarray Database at *genome-www5.stanford.edu*. This analysis used the log(base 2) of the R/G normalized ratio (mean). Data for each of these references contain normal tissue samples in addition to the samples from tumorous tissue. Genes were only included in the analysis if good data were present in over 80% of the samples. For samples, which were missing data for a particular gene, the missing value was imputed with the average of the values for that gene from the other samples. After imputing the missing data, the average value for each gene was removed.

The principal component was determined using a random selection of the tumorous tissue samples. Projections onto this principal component were performed for all the tumorous samples, as well as all of the normal tissue samples. We then compare the projections for the normal and tumorous samples. If the principal component derived from the tumorous samples is significantly different from that for a normal tissue sample, the normal tissue projections should differ considerably from the tumorous tissue projections.

3.1 Chen Liver Cancer Data

Reference [1] contained data from 76 normal tissue samples and 105 primary liver tumor samples. The POD analysis was performed 100 times, each time using a different set of 85 of the tumor samples selected at random. Thus, in this case we are finding the principal component of the tumorous samples.

The projections for all of the samples onto the principal components for each case are summarized in Figure 1. In this figure, the horizontal axis is the case number and the vertical axis represents the projections for the samples onto the principal component. Samples 1 through 105 were the tumorous samples whereas samples 106 through 181 were the normal tissue samples. Figure 1 shows that a very large percentage of the normal tissue samples (samples 106 through 181) have negative projections onto the principal component. The tumorous samples (samples 1 through 105) show more variability, but about 75% of them show positive projections.

We generated statistics for the projections of the tumorous samples to find the sample mean and standard deviation for each of the 100 cases. We then averaged these values to determine an average mean and standard deviation value for tumorous samples. The projections for the tumorous samples tend to be normally distributed. To show this, we 'normalized' the projections by subtracting the mean and dividing the result by the standard deviation. The projections were then sorted into ascending order, and the percentile values were plotted against those from a standard normal distribution. The results are shown in the top plot of Figure 2. If the projections are normally distributed, the percentile values should fall close to the middle line shown on the Figure. The top and bottom line on the Figure show the mean plus and minus three sigma values. The percentile values for the tumorous sample projections line up fairly well with those from the standard normal. Thus, it is a reasonable to assume that these projections are normally distributed.

Similar statistics were generated for the projections from the normal tissue samples, with the percentile values plotted against those from a standard normal distribution in the bottom plot of Figure 2. From this figure, it also seems reasonable to consider the projections from the normal tissue samples as being normally distributed.

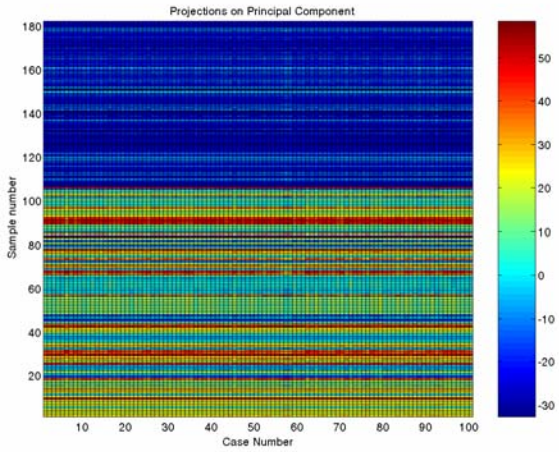

Figure 1 – Projections onto Principal Component – Chen Liver Cancer Data

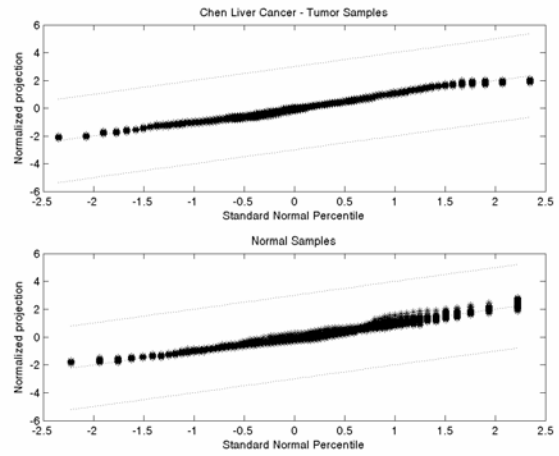

Figure 2 – Percentile Limits vs. Standard Normal Distribution – Chen Liver Cancer Data

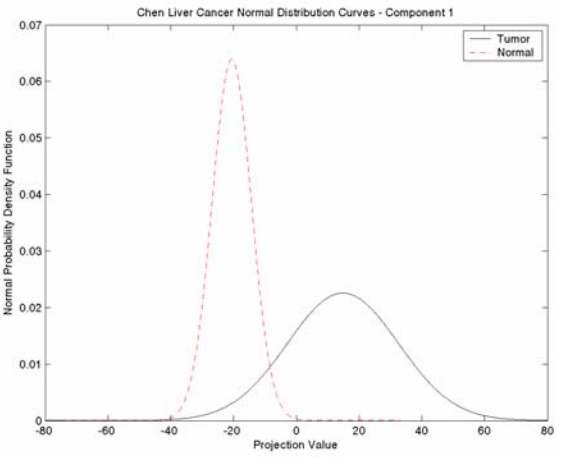

Figure 3 – Normal Density Functions – Chen Liver Cancer Data

The normal probability density functions for the projections are shown in Figure 3. This figure shows that if the projection of a sample is positive, the sample is almost certainly tumorous. There is about a 25% probability of a tumorous sample having a negative projection.

3.2 Chen Bladder Cancer Data

A similar analysis was performed using the Chen bladder cancer data from Reference [2]. The data used in this analysis consisted of 103 cancerous tissue samples and 21 normal tissue samples. Similar to the analysis in section 3.1, the principal component for the tumorous samples was performed 100 times, each time using 83 randomly selected samples for the principal component analysis. The resulting projections onto the principal components are shown in Figure 4.

Examination of the figure shows that there is much more variability for the projections. About 40% of the projections from tumorous samples are negative.

We once again generated statistics for the projections of the mean and tumorous samples, and plotted percentile values against those for a standard normal distribution (Figure 5). Once again, we can see that the normal distribution assumption is not unreasonable. The normal probability distribution functions are plotted in Figure 6. Once again, the Figure shows that if a sample has a positive projection, it is almost certainly tumorous. If the projection is negative, however, there is about a 40% probability that the sample is tumorous.

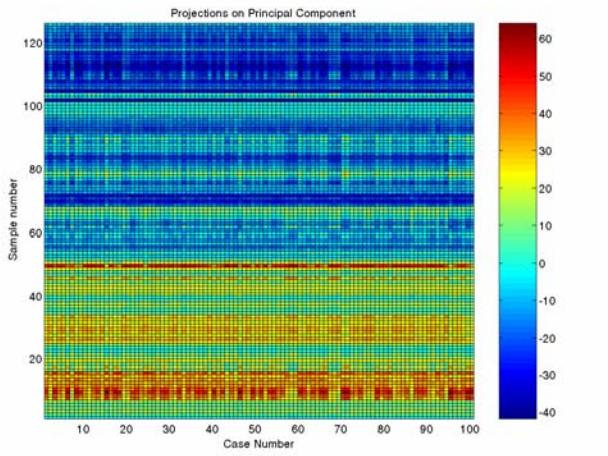

Figure 4 – Projections onto Principal Component – Chen Bladder Cancer Data

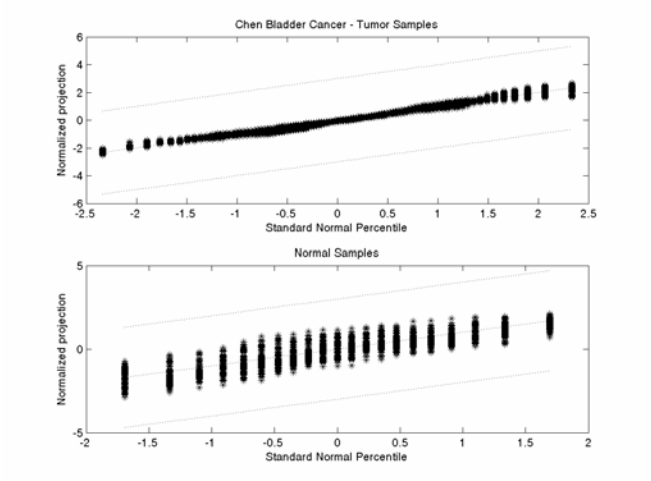

Figure 5 - Percentile Limits vs. Standard Normal Distribution – Chen Bladder Cancer Data

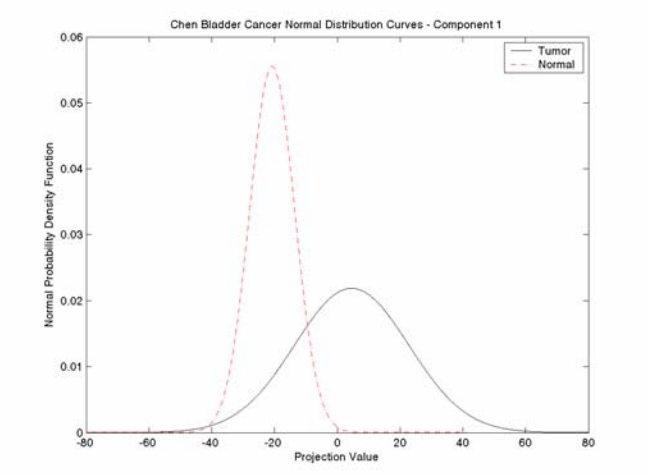

Figure 6 – Normal Density Functions – Chen Bladder Cancer Data

4. Summary and Conclusions

The above study showed an example of how the Proper Orthogonal Decomposition method can be used for a simple pattern recognition application. The principal component of a set of images is found. The magnitude of the projection of an arbitrary image onto the principal component is a measure of the correlation of an arbitrary image with the original set of data.

As a practical application, the process was used to form the principal components for a set of DNA microarray data for tumorous samples. Then projections were made for normal tissue samples, as well as other tumorous samples, against the principal components.

The analysis was performed using data from two different studies. In both cases, positive projections indicate tumorous samples. However, the method is prone to false negatives; in the liver cancer study 25% of the tumorous samples had negative projections, while 40% of the tumorous samples in the bladder cancer study had negative projections.

References

[1] Chen, X., et. al., "*Variation in Gene Expression Patterns in Human Liver Cancers*", Mol Biol Cell. 2002 Jun; 13(6): 1929-39.

[2] Chen, X., et. al., "*Variation in Gene Expression Patterns in Human Gastric Cancers*", Mol Biol Cell. 2003 Aug; 14(8): 3208-15. Epub 2003 Apr 17.

[3] H.V. Ly and H.T. Tran, "*Modeling and Control of Physical Processes using Proper Orthogonal Decomposition*" Computers and Mathematics with Applications, vol. 33 (2001) pp. 223-236.

[4] H.V. Ly and H.T. Tran, *"Proper Orthogonal Decomposition for Flow Calculations and Optimal Control in a Horizontal CVD Reactor,"* Quarterly of Applied Mathematics, Vol. 60 No. 4 (2002) pp. 631-656

[5] C.H. Lee and H.T. Tran, "*Reduced-Order Feedback Control for Thin Film Flows*," Journal of Computational and Applied Mathematics (2004), to appear.