# CANCER DETECTION USING COMPONENT ANALYSIS METHODS ON DNA MICROARRAY

Charles H. Lee and Michael Vodhanel

Department of Mathematics, California State University, Fullerton, USA

charleshlee@fullerton.edu and mvodhanel@fullerton.edu

**Abstract: The Principal Component Analysis (PCA) has been used widely as an effective tool for pattern recognition and feature extraction in many areas such signal enhancing for large array antennas, model reduction for simulating and controlling fluid flows, characteristics identification in criminology, etc. In this article, the mathematics for the PCA along with its application on DNA microarray data in cancer detection will be discussed. Studies based on two sets of liver and bladder cancers will be presented. Following this spirit, another feature extraction technique, called the Independent Component Analysis (ICA) will also be discussed. The ICA is known for its capability to identify multiple blind signals in speech recognition systems and medical signal processing. Its primary advantage, in contrast to the correlation-based PCA, is that not only can the ICA decorrelate the 2nd-order statistics of the signals, but it can also produce higher-order statistical dependencies, attempting to make the signals as independent as possible. The latest results on using the ICA for cancer diagnosis will also be reported. Test cases that were performed indicate that ICA modes make a much clearer distinction when comparing cancer data to cancer-free data than any of the PCA modes do.**

## Introduction

The primary objective of our studies is to investigate and develop pattern recognition algorithms for cancer classification and detection based on DNA micro-array data. Particularly, given a set of known cancer-positive DNA micro-array data, our goal was to use a pattern recognition technique to extract the underlying cancer characteristics and use the resulting representatives for cancer diagnosis and detection. In our earlier study [1, 2], the *Principal Component Analysis* (PCA) was used as a pattern recognition method for detecting liver cancer. The PCA, also known as the *Proper Orthogonal Decomposition* (POD), has been used widely as an effective tool for pattern recognition and feature extraction in many applications such signal processing, fluid dynamics, criminology, etc. In that study, the PCA enables us to efficiently mine from a sample of cancerous DNA a smallest possible set of hidden features, which can be used in diagnosis. That is, if our sample collection contains data of genomic expression of hundreds of cancerous individuals, then the resulting first few representatives will bear the underlying signatures of the gene expression, which can be used for profiling or detecting cancer. To a certain extent, the PCA detects well when the correlation between an arbitrary sample and the PCA extracted representative is positive. That is, all positive correlations with the PCA modes are perfectly and correctly identified with cancer. In addition, all non-cancerous samples carry negative correlations with PCA mined modes. However, some cancerous samples also have negative correlations. In short, our previous study concluded that positive correlations with the PCA extracted modes imply positively cancerous while another diagnosis method is needed if the correlations are negative. We believe that the PCA do not perform well with negative correlations is due to the nature of its ability to capture solely the second order statistics information. Moreover, it is possible that the cancer characteristics encrypted in the DNA microarrays may be of higher order statistics.

In this article, we investigate the feasibility of another feature extraction and pattern recognition technique to improve our previous PCA results. We particularly consider the *Independent Component Analysis* (ICA), which is known for its capability to identify multiple blind signals in speech recognition systems and medical signal processing [3]. It has also been used for discovering hidden factors in functional magnetic resonance imaging (fMRI) data [4]. Its primary advantage, in contrast to correlation-based Principal Component Analysis (PCA), is that not only can the ICA capture the second-order statistics signals but it can also include higher-order statistical dependencies. The inspiration for applying ICA to DNA-based disease detection is the similarity to the "cocktail party" problem. In the cocktail party problem, many individuals are having conversations in a room simultaneously and each conversation is considered a signal. Microphones are placed around the room so that each will record the overlapping conversations at different intensities and distances. The goal is to take the observable signals from the microphones and use them to separate and recover the original signals.

The assumption for applying ICA to DNA-based disease detection is that an individual's DNA is made up of a collection of signals that are repeated in other people's DNA each with varying intensity. The hope is that when performing ICA on a set of DNA samples that are known to have cancer, one or more of the signals

found will be cancer indicators. Once found, such cancer indicator signals can be compared with other DNA samples to determine whether or not a person has cancer. The test will then be able to give a percentage chance that an individual has cancer based on the level of intensity of cancer indicators in their DNA. The details of the algorithms and their implementation on DNA microarray data will be detailed in the next section. We will end the paper with the discussion of the results using the ICA and its improvement over the PCA.

**Materials and Methods**

All results are based on bladder cancer data set of Chen [5], which was downloaded from the Stanford Microarray Database [6]. The data set consists of samples of 125 individuals, of which the first 103 samples are known to have cancer and the remaining 22 samples are known to be cancer free. Each sample contains 6688 expression values of different genes.

For each run, we select randomly 80 out of 103 cancerous samples. The PCA is applied to the chosen 80 cancerous samples and to yield four arbitrary PCA modes. This step is necessary to bring the number of cancer-indicator modes to a more practical number. In addition, it eliminates some of the "noise" in the DNA samples. The number of modes that should be used, in practice, is not known a priory, so there is a trial-and-error aspect to this solution. Higher numbers of PCA modes have also been used and the results do no appear to be different. These PCA modes can be considered as the most dominant, resembling, and common to all cancerous samples. Moreover, these four modes can be considered as a mixture of the high-order statistics for cancer traces hidden in the expression of the genes, which can be separated or "unmixed" using the ICA.

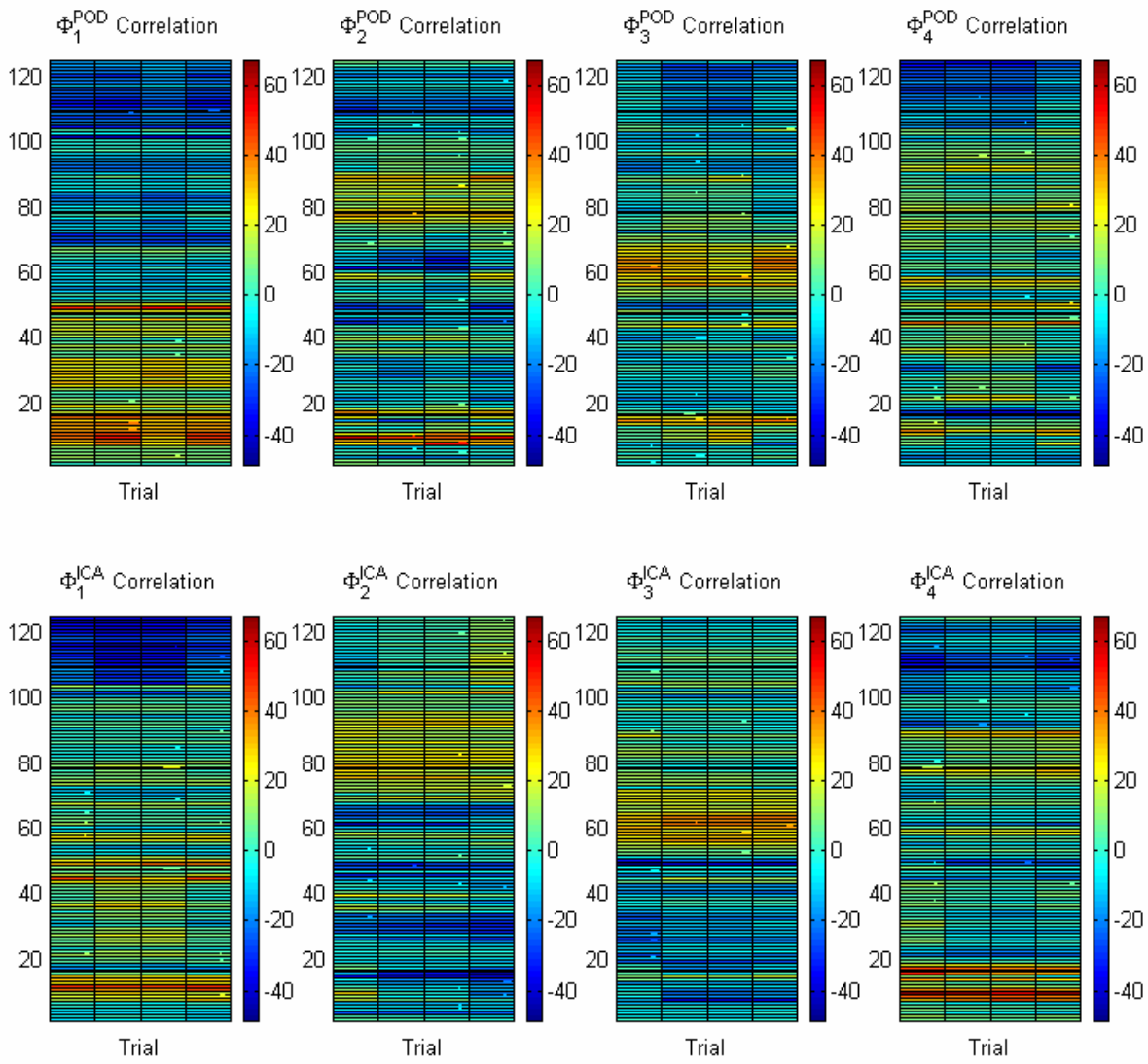


Figure 1: Correlations of five different trials between the 125 individual samples and the four PCA modes (top) and the four ICA modes (bottom)

The resulting four ICA modes are obtained when the ICA is applied. Each mode, PCA or ICA, is then correlated with the individual 125 samples. Figure 1 shows the resulting correlations for five different trials, each time a different random set of 80 cancerous samples is used. For each graph in Figure 1, the x-axis represents the trial number and y-axis represents individual samples. As shown in Figure 1, the dark blue represents low covariance while the red represents high covariance and the other colors fall between. When a mode is a good cancer indicator, it is expected to show a large difference in covariance between the first 103 samples (cancerous) and the last 22 samples (cancer-free).

**Results**

The PCA mode correlations in the top row of Figure 1 do not show very much consistent separation. However, the ICA modes do show more distinctive separation. Particularly, the first mode shows a great deal of separation with all of the cancer-free samples having low covariance. The other ICA modes also seem to contain more useful information. It seems that having very low covariance with the second mode or very high covariance with the third mode can both indicate cancer. It can be observed that the patterns in Figure 1 suggest that ICA has found three different cancer indicators. These may indicate different DNA characteristics such as cancer stage, other illnesses, etc. In any case, the evidence suggests that this ICA method can be used in successful cancer detection. This point can be re-emphasized further by observing the binary plot of Figure 2, where all correlations above –18 are set to positive (1) and all values under –18 are set to negative (0). The 101st sample never indicates having cancer in

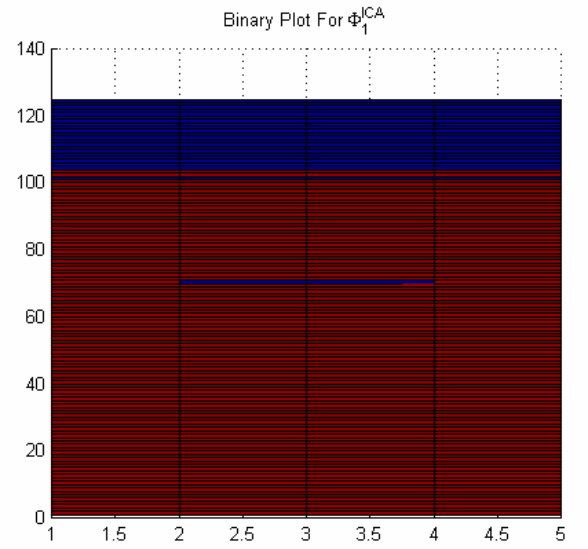


Figure 2: Binary plot of correlations of the first ICA mode with the cut-off value of -18

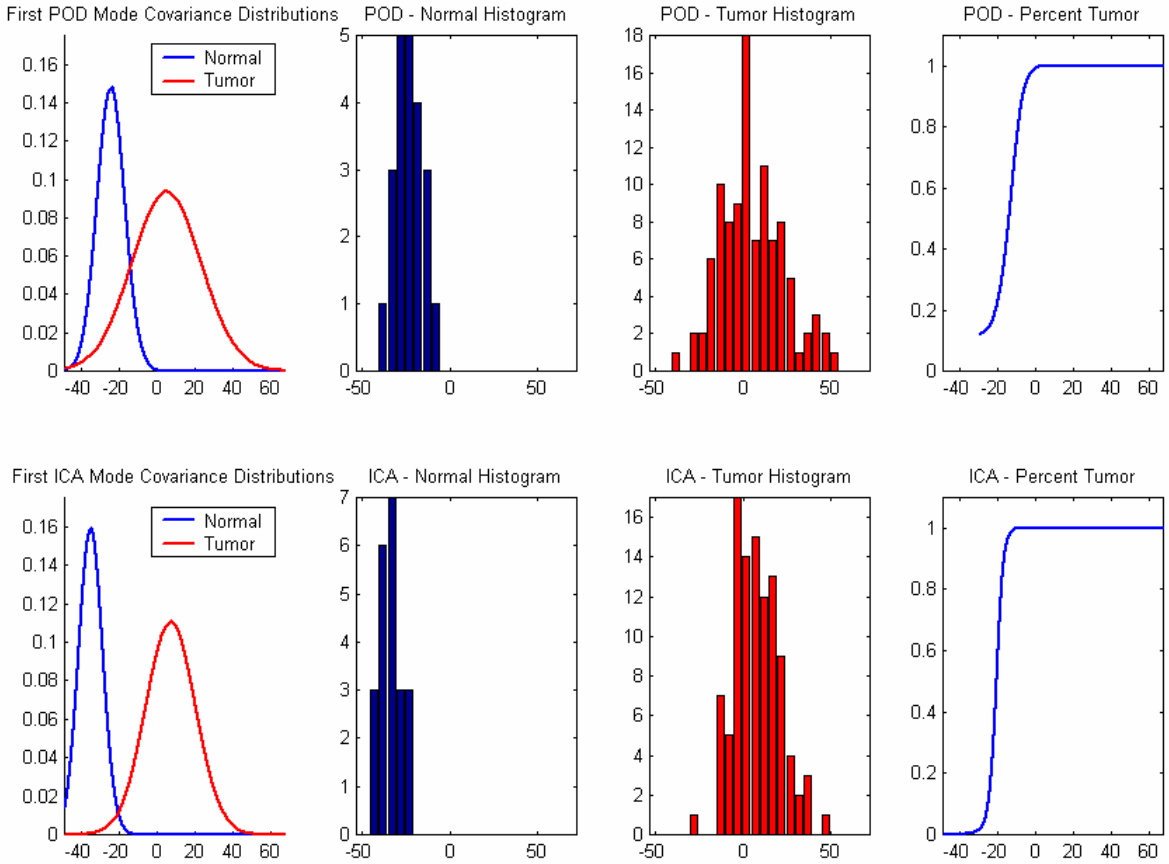


Figure 3: Statistical analyses for bladder cancer detection using the PCA (top) and the ICA (bottom)

any simulation even though it supposedly belongs to the set that contains cancer. Excluding this sample, this binary test correctly identifies cancer-free DNA 100% of the time and identifies DNA that has cancer over 99% of the time. The best iterations have 100% success in both directions.

Analysis of the correlation results was further performed to illustrate the effectiveness of both the PCA and ICA analyses. Normal plots, histograms, and percentage plots were created based only on the dominant PCA mode and first independent ICA component and are shown in Figure 3. The normal distributions clearly show that there is little overlap between covariance of the samples with cancer and the samples without cancer when using the ICA test. The histograms also show that the ICA test separates the two sets better than the PCA test. Furthermore, the percentage plot shows a very steep climb for the ICA test, which is desirable because it shows that the test is very decisive and is likely to return results near 0% or 100%. Results that are not near 0% or 100% are not desirable because they do not give a clear answer whether cancer is present or not.

**Conclusions**

Important findings were made for both the PCA and ICA as DNA-based disease detection methods. For the PCA, it was found that using the sum of the modes instead of just the dominant mode makes a superior test. By itself, the dominant mode is insufficient and does not separate the cancer and cancer-free data sets well enough. As for the ICA, it was found that the first mode could be used to successfully to detect which samples contain cancer and which do not by itself. Furthermore, it appears that the other modes might be useful as secondary tests. Thus, the investigation suggests that independent component analysis could be successfully applied to DNA-based disease detection in general.

**References**


[1] DAVID PETERSON AND CHARRLES H. LEE, "Disease Detection Technique Using the Principal Orthogonal Decomposition on DNA Microarray Data", Proceedings of the 6th Nordic Signal Processing Symposium, NORSIG 2004, Espoo, Finland, pp. 33-36, 2004.

[2] CHARLES H. LEE AND DAVID PETERSON, "A DNA-based Pattern Recognition Technique in Cancer Detection", Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, San Francisco, 2004.

[3] AAPO HYVARINEN, JUHA KARHUNEN, AND ERKKI OJA, (2001) 'Independent Component Analysis: Algorithms and Applications', John Wiley & Sons.

[4] V.CALHOUN, T.ADALI, G.PEARLSON, AND J.PEKAR, "A Method for Making Group Inferences From Functional MRI Data Using Independent Component Analysis Human Brain Map", vol. 14, pp. 140-151, 2001.

[5] 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.

[6] STANFORD MICROARRAY DATABASE, http://genome-www5.stanford.edu/