Fast and Tractable Disease Detection Technique Using Principal Component Analysis

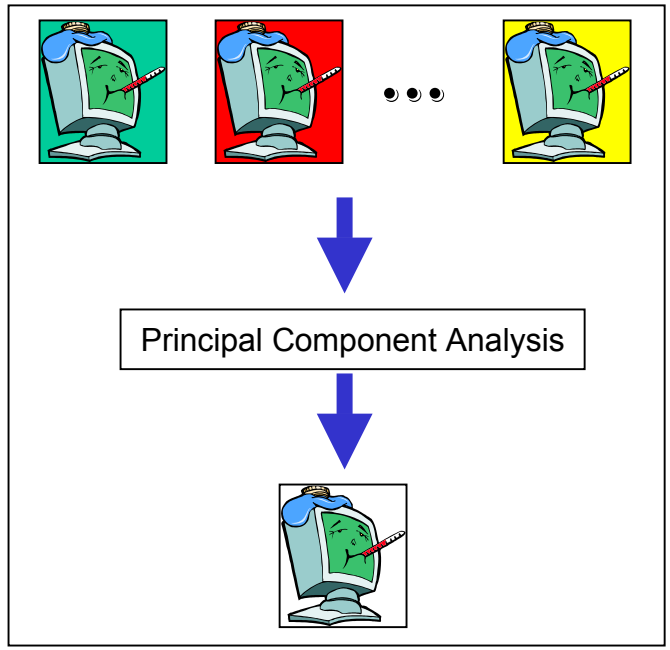
David Peterson & Michael Vodhanel
Advisor: Dr. Charles H. Lee

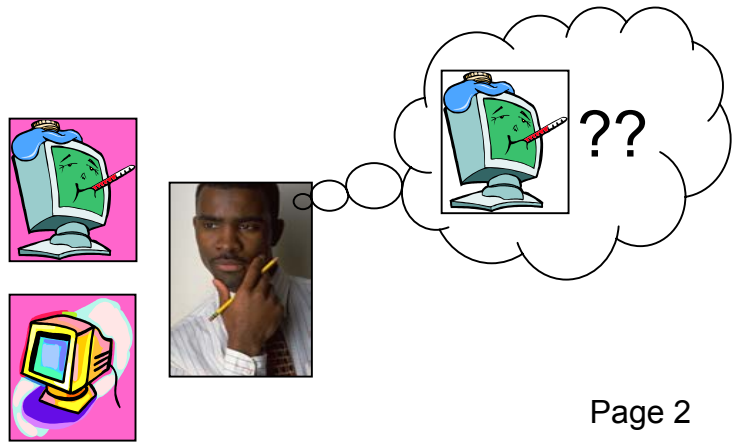
Problem Description

- Given a set of DNA microarray data from diseased samples

- Apply Principal Component Analysis (PCA) techniques to extract the primary component of the diseased samples (captures the diseased features)

- Perform simple disease detection tests by finding the projection of arbitrary samples onto the principal component

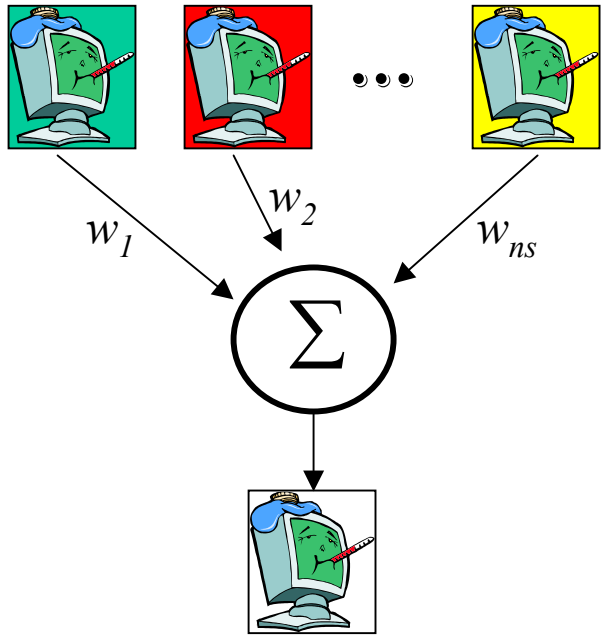




Mathematics of the PCA

- We begin with a series of "snapshots"

$$\{V_i(\vec{x})\}_{i=1}^{n_s}$$



- Assume that the principal component is a linear combination of the snapshots, with weighting factors w_i.

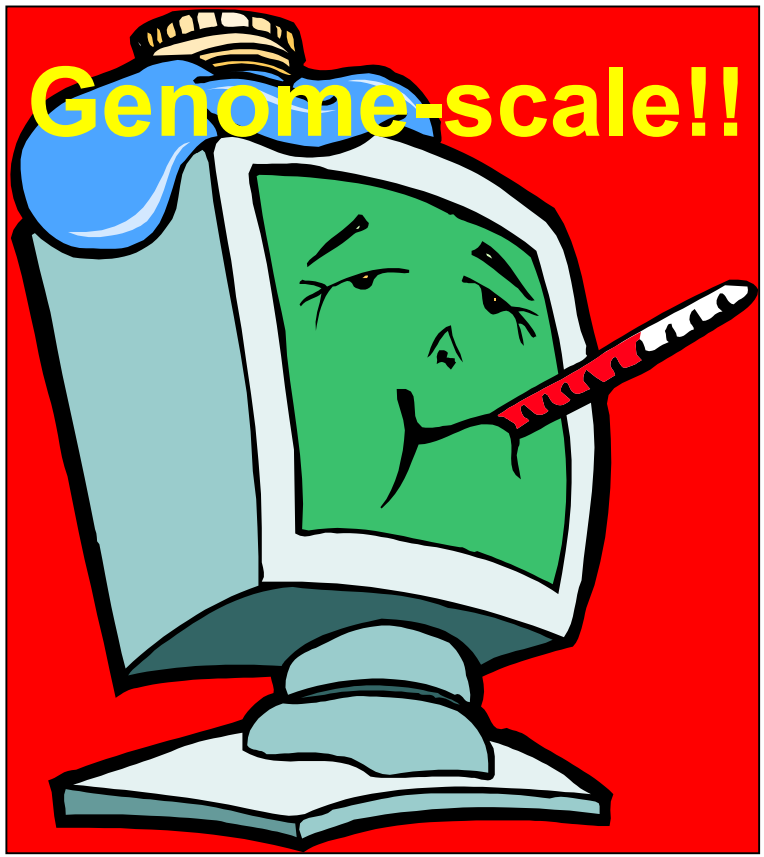
$$\Phi_1(\vec{x}) = \sum_{i=1}^{n_s} w_i V_i(\vec{x})$$

- The weighting factors can be shown to be the components of the primary eigenvector of the covariance matrix θ, where the *(i,j)* component of θ is defined as:

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

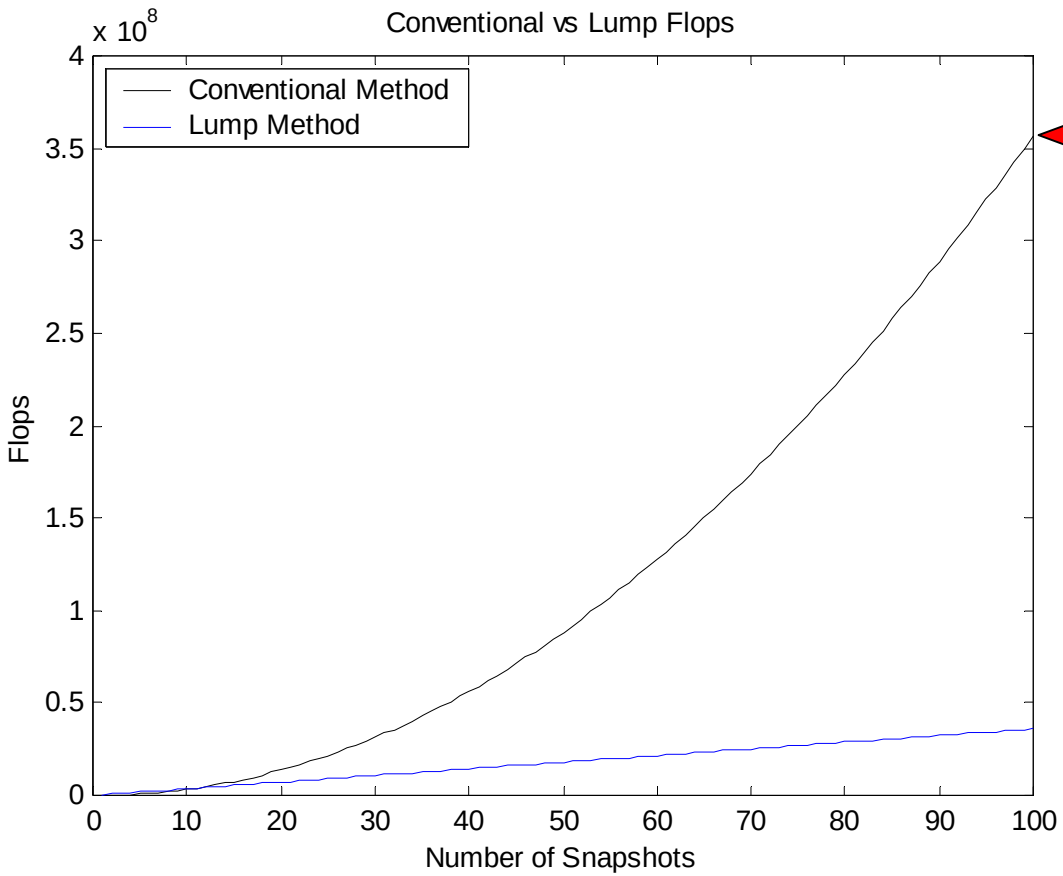
Computational Problem for PCA

The conventional POD method is $O(n \times Ns^3)$, where
n is the snapshot size and
Ns is the number of snapshots.



The Lump Method (many samples)

we developed the matrix-free lump method, which is O(n × Ns).



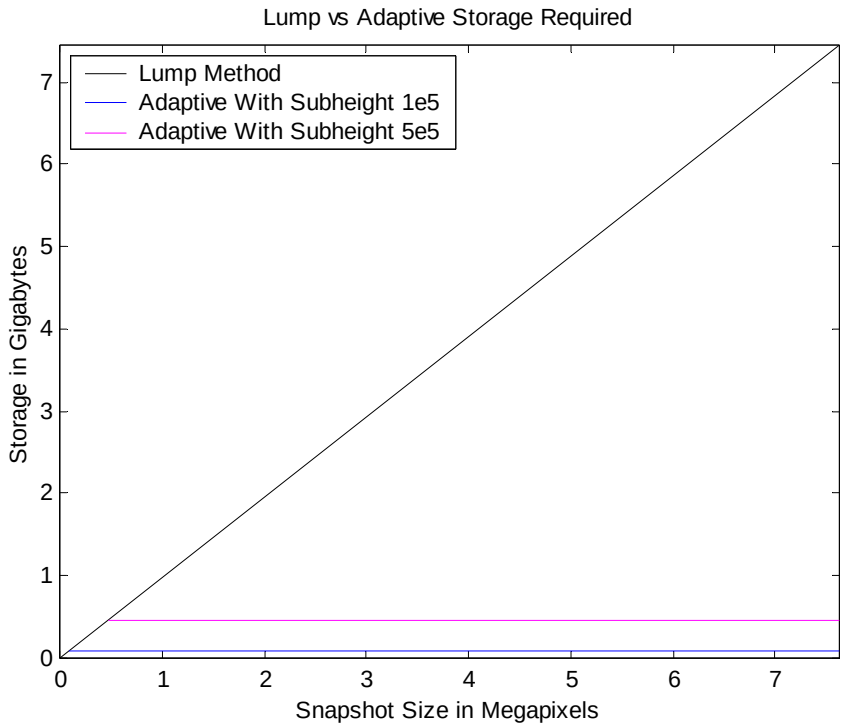
CONVENTIONAL METHOD

SIGNIFICANT computational saving.

OUR LUMP METHOD

The Adaptive Method (large-size samples)

The adaptive algorithm is also a matrix-free PCA method that is O(n × Ns). However, the adaptive method allows for the snapshots to be read and analyzed in small portions. Because the entire snapshots do not need to be stored, this saves memory

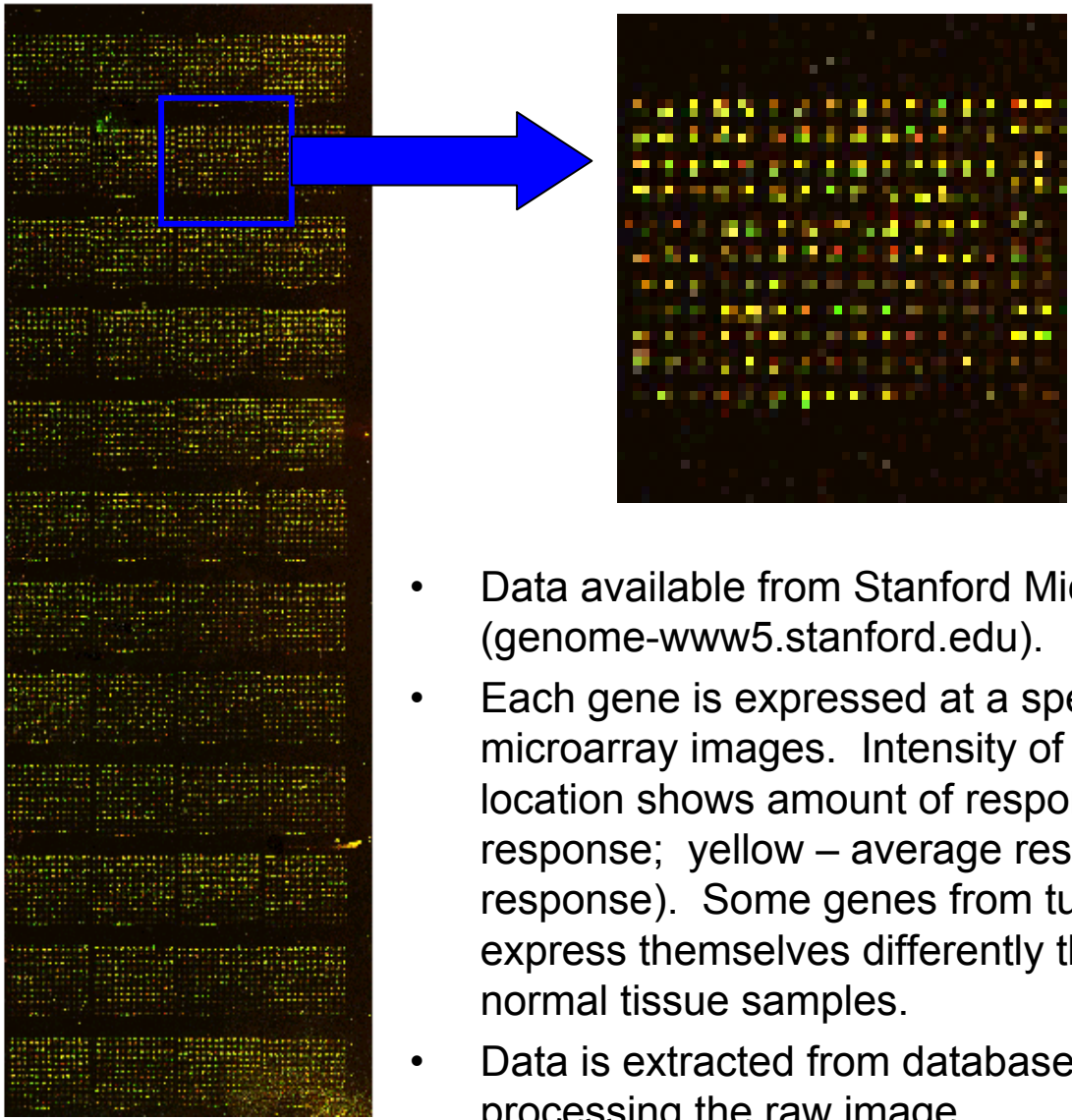


CONVENTIONAL METHOD

MUCH LESS REQUIRED MEMORY

OUR ADAPTIVE METHOD

DNA Microarray Data Images



- Data available from Stanford Microarray Database (genome-www5.stanford.edu).
- Each gene is expressed at a specific grid location in the microarray images. Intensity of the image at the grid location shows amount of response (green – minimal response; yellow – average response; red – maximal response). Some genes from tumorous samples will express themselves differently than the same genes from normal tissue samples.
- Data is extracted from database in tabular form, rather than processing the raw image.

Chen Liver Cancer Case Study

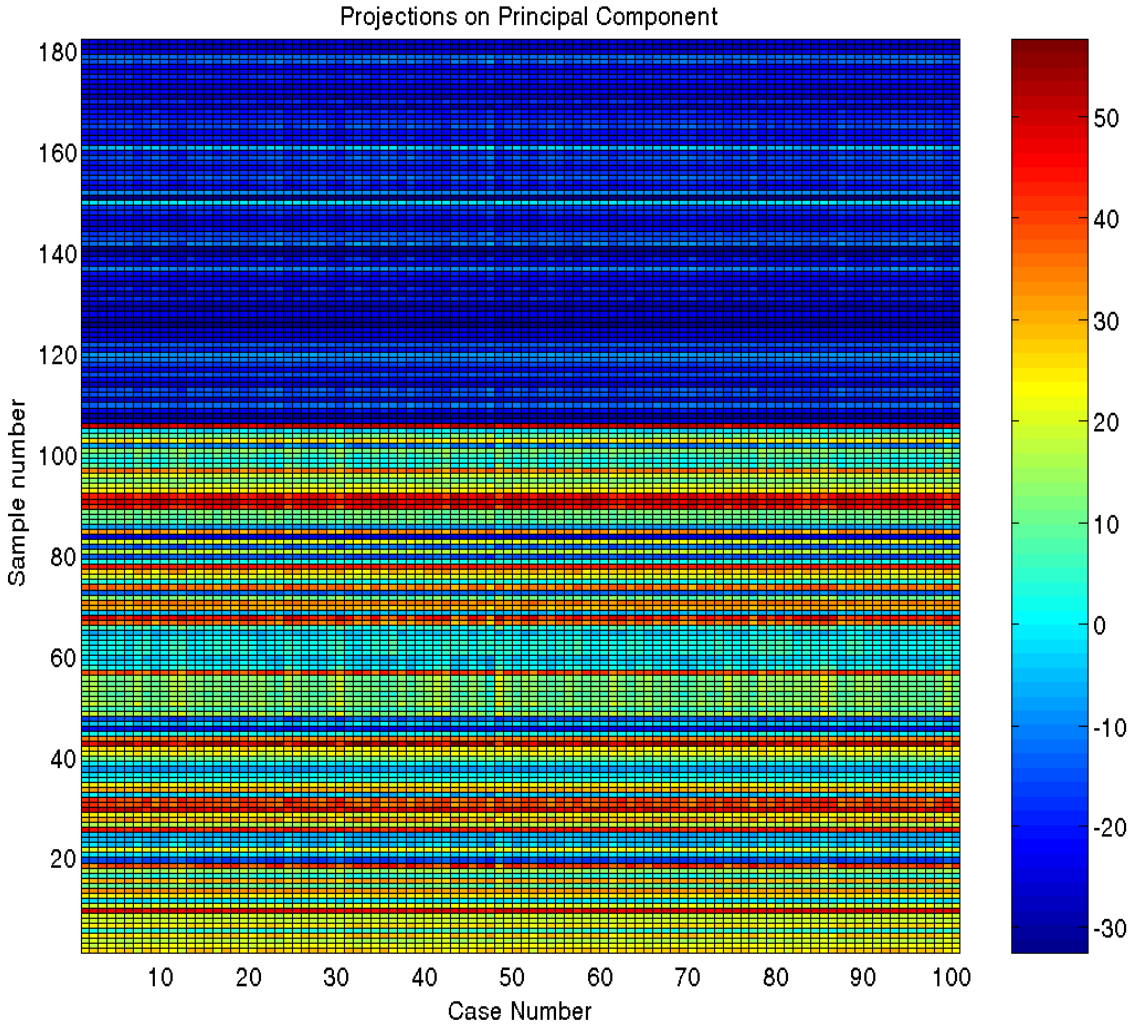
- Data for analysis was obtained from Chen, Xin, et. al, “Gene Expression Patterns in Human Liver Cancers”, Molecular Biology of the Cell, Vol. 13, 1929-1939, June 2002

- Reference provided DNA data for:
  - **76 normal** tissue samples
  - **105 primary liver tumor** samples.

- Data for **5520 genes** were extracted
  - In order for a gene to be included in this analysis, data for that gene had to be present in at least 80% of the samples
  - If a sample is missing data for a particular gene, the value was imputed by using the mean of the values from the remaining samples.

Projections Onto The Principal Component

Chen Liver Cancer Study

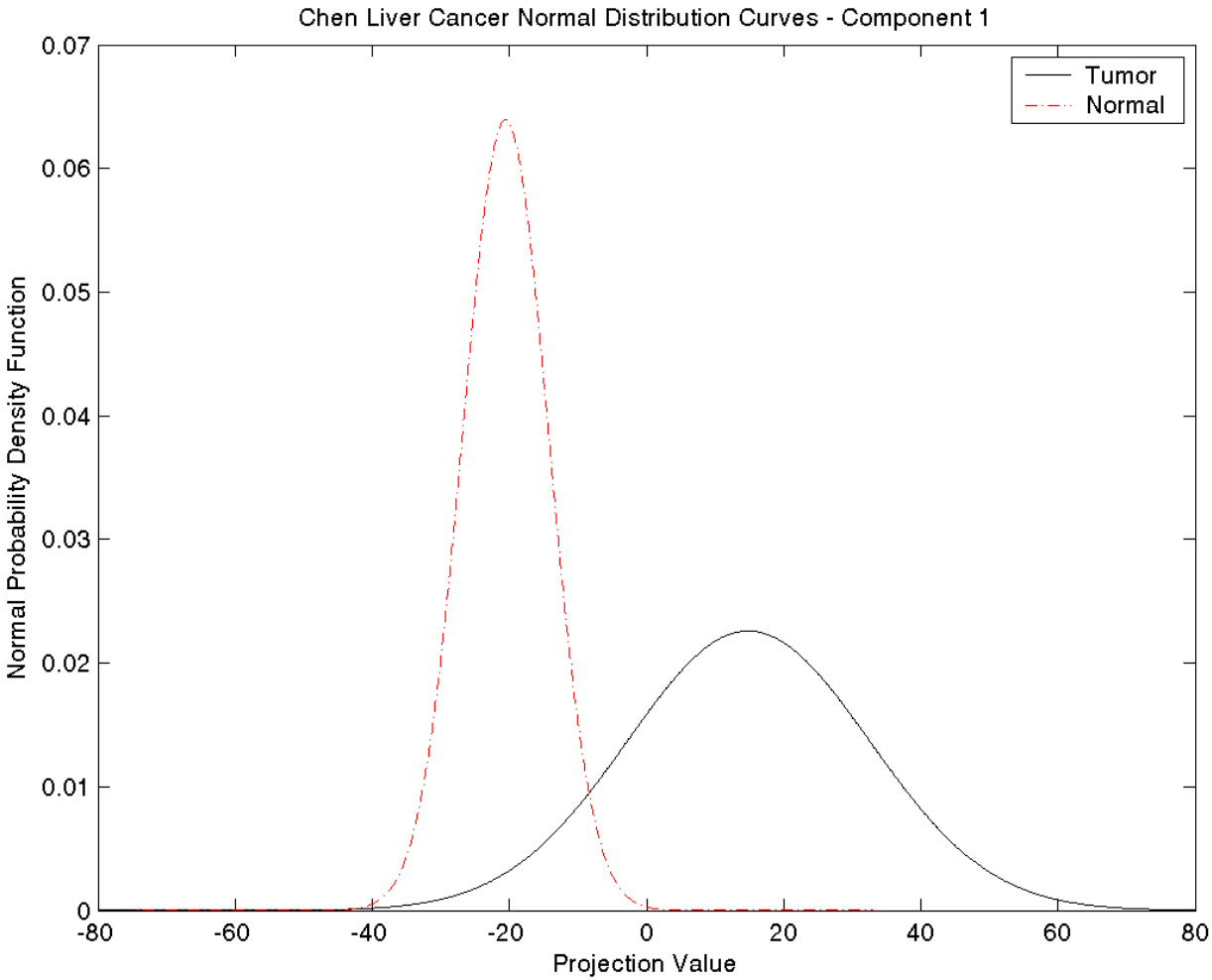


- **Tumorous tissue samples (1 - 105)**
- **Normal tissue samples (106 - 181)**
- **Principal component analysis performed 100 times**
  - **The principal component was extracted using 85 tumorous tissue samples (selected at random)**
- **Figure shows the projections of the samples onto the principal component**

- **Projections for normal tissue samples are almost always negative.**

Normal Distribution Curves

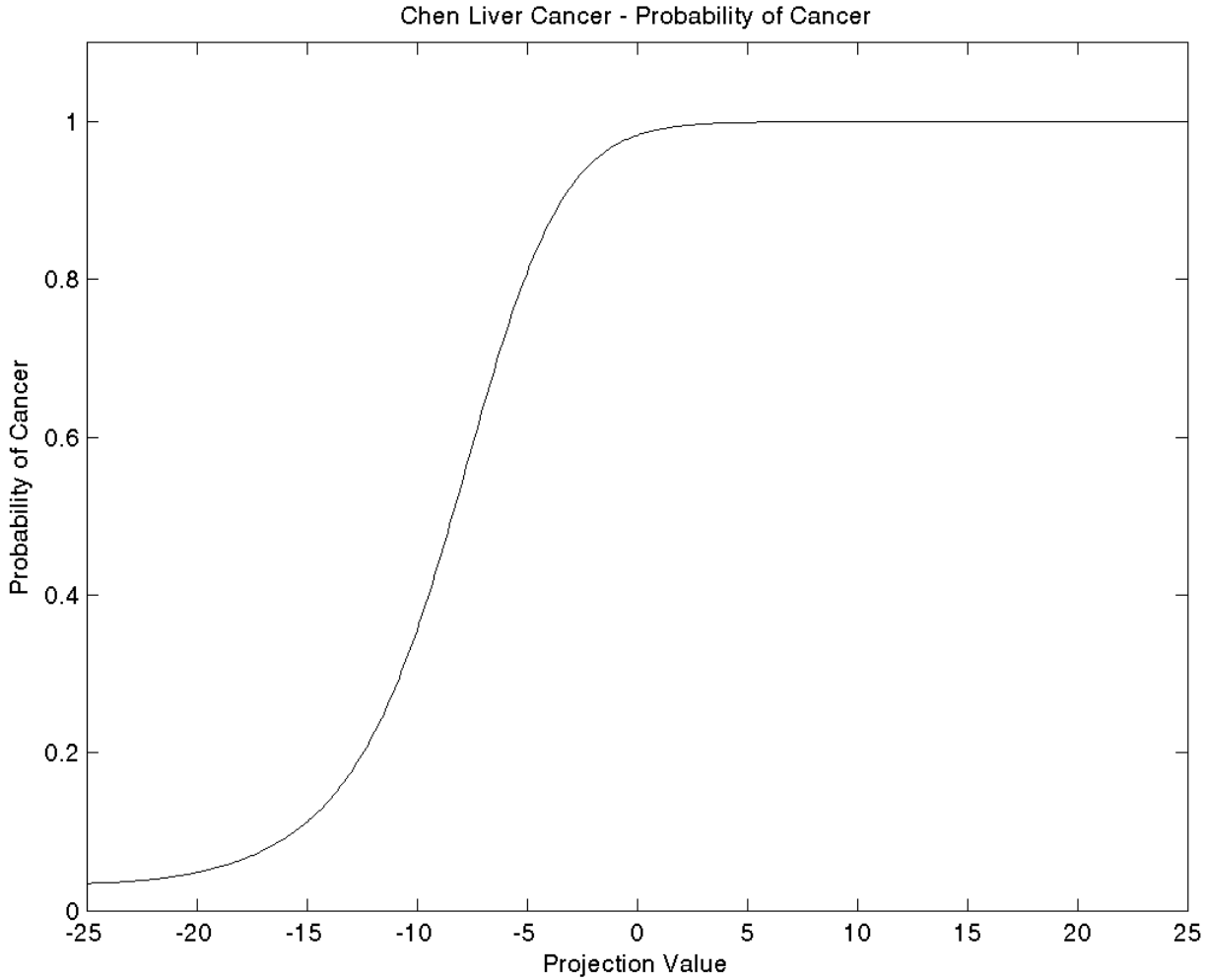
Chen Liver Cancer Study



- **Statistics generated for normal and tumorous tissue samples.**
- **Data tended to be normally distributed.**
- **Normal distribution curves are shown here.**

- **If projection is positive, sample is almost certainly tumorous.**
- **If projection is negative, there is approximately a 25% chance that the sample is tumorous.**

Probability of Cancer vs. Projection



Conclusions

- Applying Principal Component Analysis techniques to DNA microarray data can be used as a basis for simple disease detection applications.
- The method was demonstrated using data from Chen's liver cancer study, although the method is general and could be applied to any type of disease.
- The case study presented in this analysis showed that the method could be prone to false negatives; i.e., in the Chen liver cancer study 25% of the tumorous samples had negative projections.
- Increased reliability of the method might be achieved by including more components, other than just the principal component, in the analysis

Acknowledgement: This research was funded by the California State University Fullerton Faculty Research Program and by the State Special Fund for Research, Scholarly, and Creative Activities.
Dr. Charles H. Lee and David Peterson would like to thank Prof. Janey Youngblom (CSU Stanislaus) for introducing us to DNA Microarrays through the DNA Microarray Workshop sponsored by the California State University Program for Education and Research in Biotechnology (CSUPERB).