Comparison of Dimension Reduction Methods for Microarray Data with Survival Time

Stacey Ackerman[1], Cyrus Aghili[2], Israel Cabello[3], and John Ratana[4]

[1]Brown University, [2]Columbia University, [3]Texas State University, [4]Binghamton University

Abstract

This project compares the effectiveness of two dimension reduction strategies when applied to microarray data. Microarrays can measure which genes are expressed when a patient is undergoing cancer treatment. The statistical problem is that the number of observation (patientss' tissue samples) is much smaller than the number of variables (genes), which introduces multicollinearity and prevents the application of typical multivariate techniques. We compare two methods: principal components analysis (PCA) and partial least squares (PLS). The PCA model attempts to maximize the variance of the original data represented in the new set and assumes that this will predict survival. In contrast, the PLS model accounts for survival by selecting for genes hightly correlated with the response variable. Thus, we hypothesized that (1) PLS would more efficiently model survival time, and (2) the genes highly weighted by PLS would be more relevant to more survival time. We used the Cox Proportional Hazard Model to find the components that are significant predeictors of survival time in each reduced dataset. Modeling survival using only the first PLS component yields a significant p-value (less than 0.0005) for the likelihood ratio test, while multiple PCA components are required for significance (p-value < 0.01). This result confirmed part (1) of our hypothesis. Regarding part (2), we found that the twenty genes most heavily weighted by each method overlapped by more than 50% across datasets, and we found similar weighting structures in most cases. This indicates a strong similarity in the selection of genes that predict survival. Overall, our results indicate that PCA is able to select some of the same significant genes identified by PLS, but PLS provides a much more concise representation of the data.