# Predicting Alzheimer's Disease with Microarray Gene Expression Data

UMBC REU Site: Interdisciplinary Program in High Performance Computing
Trevor V. Adriaanse[1], Meshach Hopkins[2], Rebecca Rachan[3], Subodh R. Selukar[4],
Graduate assistant: Elias Al-Najjar[5] Faculty mentor: Kofi P. Adragni[5], Client: Nusrat Jahan[6]
[1]Bucknell U., [2]CSEE, UMBC, [3]North Central College, [4]UNC-Chapel Hill, [5]Math & Stat, UMBC, [6]James Madison U.

## Predicting Alzheimer's Disease

Alzheimer's Disease (AD) is a fatal neurological disorder chiefly present in the elderly. AD has no cure and evidence points to a genetic link. With microarray data from [1], our goal is to find a relationship between gene expression and presence of AD. The three-stage process is: screening, sparse sufficient dimension reduction (SDR), and hierarchical clustering. We parallelize the existing R code to enhance execution speed and conduct a performance analysis.

## Methodology

We begin with $X = (X_1, \ldots, X_p)^T$ gene expressions and $Y$ the response variable, where $Y = 0$ if AD is not present and $Y = 1$ if AD is present. We use an inverse regression approach $X|Y$ because of the data sampling scheme and we seek the most significant genes that can best predict the presence or absence of AD.

The high number of gene expressions (32312) and small sample size (79) motivates a SDR. A reduction $R : \mathbb{R}^p \to \mathbb{R}^d, d < p$ is sufficient if $Y \perp\!\!\!\perp X|R(X)$. Our method of dimension reduction is the Principal Fitted Component Model:

$$X_y = \mu + \Gamma\beta y + \Delta^{1/2}\varepsilon$$

The parameter $\Gamma$ gives the relationship between AD and gene, while $\Delta$ illustrates the interdependence of genes and $\Gamma^T\Delta^{-1}X$ is a SDR of $X$.

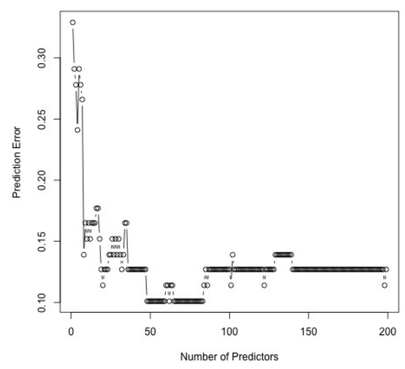
Our method is to first screen data with $t$-tests, then perform sparse estimation and cross-validation, and finally to cluster genes exhibiting mutual dependence together.

## Implementation

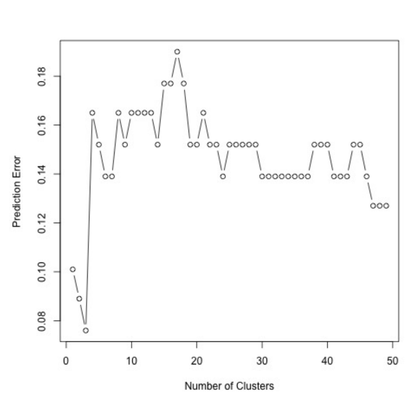
We adapted the works of [2] and [3] to implement our methodology in R. We parallelized our main function on the maya 2013 cluster using snow [4] and ran simulations as a performance analysis. Finally, we applied it to the AD data set.

## Results

Prediction Error for the 200 Most Significant Genes




Prediction Error Using the 49 Genes that Best Predict AD

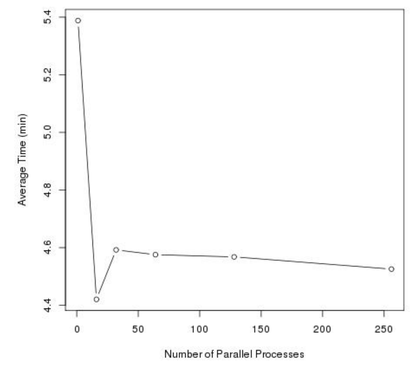



Prediction error is minimized for 49 predictors and 3 clusters.

| | | | | | |
|---|---|---|---|---|---|
| Group 1 | 8028380 | 8062844 | 8096663 | 7937275 | 7985757 |
| | 8030448 | 8075637 | 8135172 | 8178561 | 8050060 |
| | 7950284 | 8098576 | 8170891 | 7960689 | 7963235 |
| | 8062880 | 8081620 | 8176230 | 7903507 | 7982564 |
| | 7974895 | 8051773 | | | |
| Group 2 | 8039378 | 7905817 | 8028791 | 8002041 | 8037079 |
| | 8015835 | 7992447 | 8036252 | 8180371 | 7902435 |
| | 8041225 | 8123739 | 8014794 | 7899841 | 7931479 |
| | 8062796 | 8091452 | 7908867 | 7979663 | 8026155 |
| | 7981566 | 7997352 | | | |
| Group 3 | 7894596 | 7893808 | 7894185 | 8024436 | 8121130 |

## Parallelization

Performance Study of Our Main R Function




Parallelization improved performance speed most substantially using one node with sixteen processes per node.

## Conclusions

Findings:

- Three clusters of genes to predict AD
- Parallel implementation is most effective on one node with sixteen proceses per node

In the future we will extend our research by comparing our model with a logistic regression model to determine the efficacy of our methodology. Further, we will identify the biological relevance of our findings.

## References & Acknowledgments

[1] Hokama M, Oka S, Leon J, Ninomiya T et al. *Cereb Cortex* 2014.
[2] Adragni et al. *Computational Statistics*, *to appear* 2015.
[3] Adragni and Xi *Statistics*, 2015.
[4] Tierney et al. 2013.

Full technical report: HPCF–2015–23 `hpcf.umbc.edu` > Publications

REU Site: `hpcreu.umbc.edu`

NSF, NSA, DOD, UMBC, HPCF, CIRC