BE175 Winter 2022 Final Project Report

PCA and ICA on Alzheimer's DNA Microarray Gene Expression Data

Jamie Stickelmaier, Katie Villaseñor, and Sarah O'Donovan

Introduction

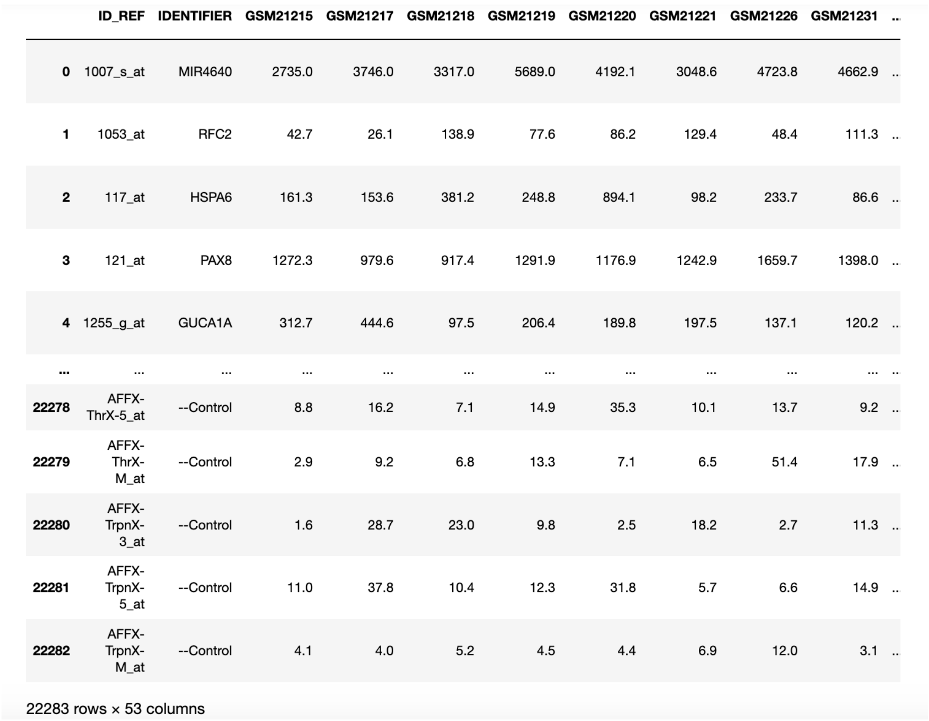
Alzheimer’s Disease (AD) is a progressive neurological disorder that primarily affects the senior population, with a 2020 report estimating that approximately 5.8 million Americans are living with the disease [1]. With numbers projected to nearly triple to 14 million by 2060, it is essential for research groups to continue to analyze the association between the onset of AD and any associated trends in gene patterns in order to grow closer to AD gene-specific treatment [1].

Machine learning methods have previously been used to analyze and classify trends of common conditions within the healthcare community, with the focus on specific genes that may be significant contributors to their occurrence. Specifically to AD, research highlighted in *Kong et al* has been done to identify gene expression profiles as linear combinations of simple expression patterns that may lead to the onset of AD [2]. A limitation, however, within the use of machine learning is the intrinsically noisy, complex, and high-dimensional natures of DNA microarrays. To combat this, researchers utilized the working methods of both principal component analysis (PCA) and independent component analysis (ICA), to analyze isolated DNA microarrays for discovery of any relevant patterns. Here, we discuss the use of PCA and ICA and how it can be applied to the discovery of genes associated with AD pathogenesis through the recreation of key figures presented in *Kong et al*.

Problem Definition

In order to initiate our approach to reproduce the data, we acquired the initial data set of 53 patients with 22283 gene expressions each as used in *Kong et al*. Our first goal of data reproduction was to interpret the data produced by microarray technology in order to assemble readable gene profiles of 13 patients. Arriving at the same problem as the initial researchers, samples with excess noise and irrelevant gene expression profiles were filtered out and eliminated from analysis entirely.

A sample of the original unfiltered data is as follows:



	ID_REF	IDENTIFIER	GSM21215	GSM21217	GSM21218	GSM21219	GSM21220	GSM21221	GSM21226	GSM21231	...
0	1007_s_at	MIR4640	2735.0	3746.0	3317.0	5689.0	4192.1	3048.6	4723.8	4662.9	...
1	1053_at	RFC2	42.7	26.1	138.9	77.6	86.2	129.4	48.4	111.3	...
2	117_at	HSPA6	161.3	153.6	381.2	248.8	894.1	98.2	233.7	86.6	...
3	121_at	PAX8	1272.3	979.6	917.4	1291.9	1176.9	1242.9	1659.7	1398.0	...
4	1255_g_at	GUCA1A	312.7	444.6	97.5	206.4	189.8	197.5	137.1	120.2	...
...	...	...	...	...	...	...	...	...	...	...	...
22278	AFFX-ThrX-5_at	--Control	8.8	16.2	7.1	14.9	35.3	10.1	13.7	9.2	...
22279	AFFX-ThrX-M_at	--Control	2.9	9.2	6.8	13.3	7.1	6.5	51.4	17.9	...
22280	AFFX-TrpnX-3_at	--Control	1.6	28.7	23.0	9.8	2.5	18.2	2.7	11.3	...
22281	AFFX-TrpnX-5_at	--Control	11.0	37.8	10.4	12.3	31.8	5.7	6.6	14.9	...
22282	AFFX-TrpnX-M_at	--Control	4.1	4.0	5.2	4.5	4.4	6.9	12.0	3.1	...

22283 rows × 53 columns

Figure 1. Sample of the unfiltered data from 53 patients with 22283 genes each

After the filtering process, we were ultimately left with 8 control patients and 5 severe AD patients for a total of 13 patients and 3617 genes from each patient. Our next two goals of data reproduction are as follows: apply unsupervised PCA and ICA for further noise and dimensionality reduction, and use analysis to isolate genes associated with AD pathogenesis.

Our work will ultimately aim to serve as a confirmation that this method of machine learning can be applied to gene pattern analysis and identification of correlation with AD. The successful implementation of efficiently reproducing and analyzing the complex nature of DNA microarrays will also indicate whether this method can be adapted and applied to research of other disorders that are identifiable through gene patternization.

Methods:

Data Reduction

To implement our machine learning algorithms we used scikit learn's PCA and ICA decomposition functions. PCA is an important tool for data decomposition, as it allows for the break-down of a complicated data matrix denoted as X into a T matrix of scores relating the correlation between each sample, and a W matrix of loadings for the relative weight of each variable. In our project, we first calculated the scores by performing the 'fit_tranform' method included in the PCA function, specifying the use of 10 principal components as was used in the paper. It is interesting to note that although the paper chose 10 principal components, we made a scree plot of the variance explained for each component, and it seems the most efficient choice of components is around four to five. After finding the scores, we calculated the loadings matrix by finding the eigenvalues and multiplying the eigenvalues by the square root of the explained variance. We plotted the scores and loadings matrices in a new space spanned by two of the ten principal components to analyze the data in fewer dimensions. The relative scores of our Alzheimer's patients were related to their general location in the positive direction of principal component one. With this information, we then sorted by value the column of the loadings matrix corresponding to principal component one to identify the 3617 gene signatures with the highest weight.

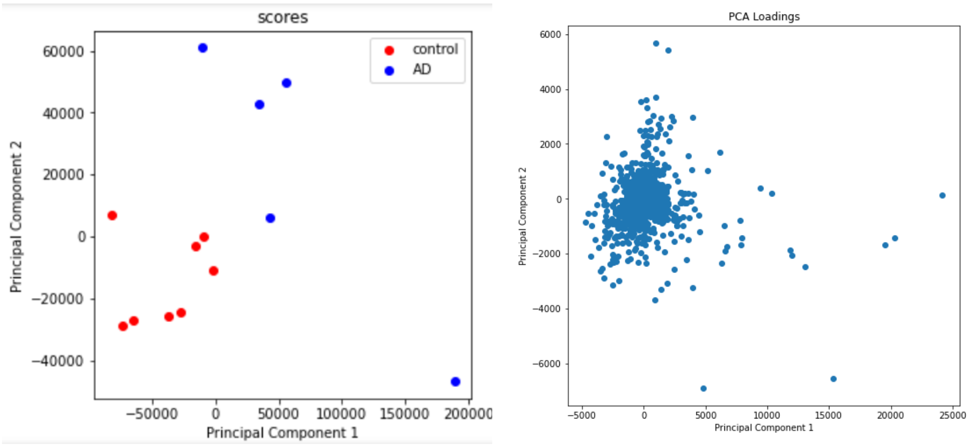


Figure 2. Plot of the scores of each patient and the loadings of each gene signature calculated via PCA

Execution of Gene Signature Analysis

With our dataset reduced to include only 3617 genes per patient, we calculated the loadings once more using both ICA and PCA. We calculated PCA in the same way as before. For ICA we decomposed the matrix into an A matrix of the samples versus 13 latent variables and an S matrix of the gene signature loadings for each latent variable. Hierarchical heat maps were plotted using a matrix containing the components and samples, determined in both ICA and PCA by transforming the loadings matrix. In accordance with the paper, each gene signature was plotted versus its corresponding loading value for all 10 components. These graphs took the form of line plots and histograms and aided us in comparing the efficiency of both algorithms.

Results

Comparing PCA and ICA

In recreating the results of the paper, we had difficulty demonstrating that ICA was a stronger method for identifying significant AD genes than PCA. When we plotted clustermaps of the latent variables and the principal components after the data was analyzed with ICA and PCA respectively, neither method displayed hierarchical clustering that could separate control samples and AD samples from each other. We contribute this inconsistency and deviation from the paper's findings due to our elimination of three samples from their data. The paper eliminated one control patient and two AD patients due to noise, but offered no other specifics on how they analyzed this noise. We eliminated the same number of patients by running initial clustermaps and removing subjects that seemed to consistently group out of place, but there was no way for us to check if these were the same subjects that the paper eliminated in their methods.

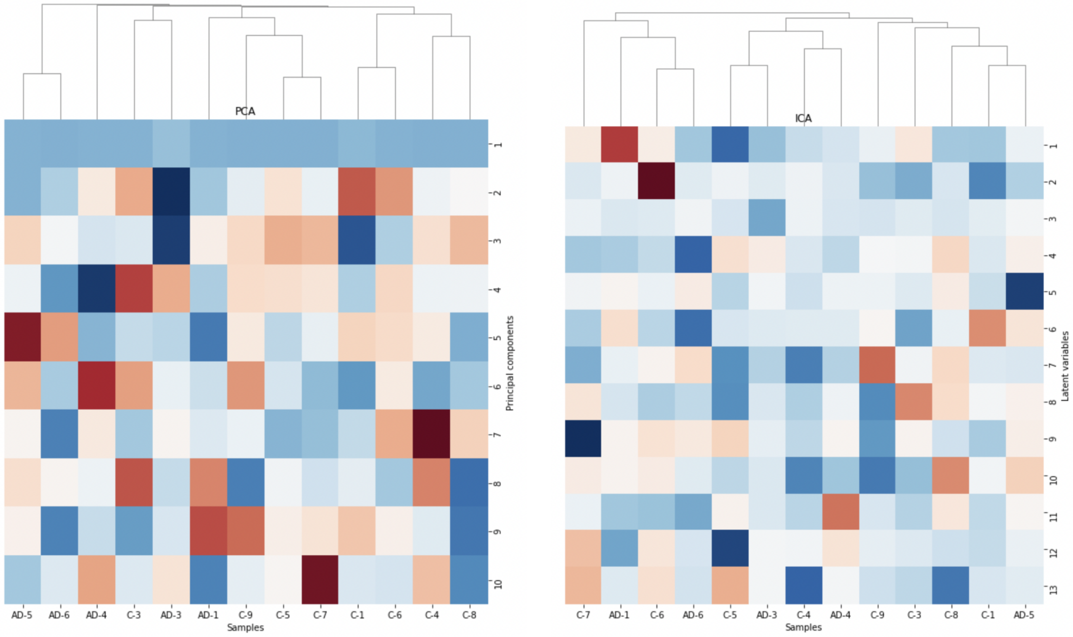


Figure 3. Hierarchical clustering of the PCA outputs (left) and ICA outputs (right). Neither clustermap demonstrates clear grouping between control and AD samples.

Additionally, the plots for our gene signatures did not demonstrate significant difference between PCA and ICA analysis. It was expected that PCA would affect a large number of genes from the microarray data and that ICA would have sparser effects and generate high signal intensities for fewer genes; however, both PCA and ICA showed a similar level of sparseness in the intensity of their gene signatures. The histograms of these gene signatures were also supposed to be sparser for ICA, and yet the PCA histogram seems to be more super-gaussian than the ICA histogram, which is a reversal of the findings of the paper. This may be due to the way we selected which genes from the dataset to analyze, where we selected 3617 genes based on our scores and loadings plot, because this information was also not described specifically by the researchers.

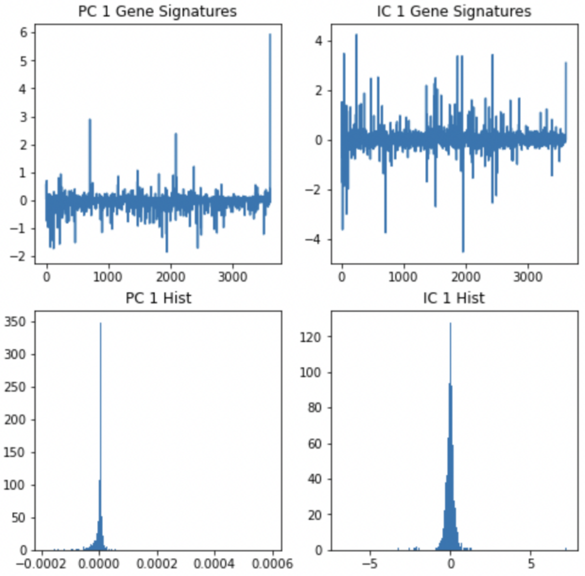


Figure 4. The first principal component (top left) and first independent component (top right) from PCA and ICA analysis of the gene signatures respectively. In these top graphs, the x-axis is genes and the y-axis is signal intensity. On the bottom depicts the gene signature histograms of the first principal component (bottom left) and the first independent component (bottom right).

Identifying Significant Genes

Despite our difficulties replicating the comparative analysis of PCA and ICA, we still analyzed which genes were deemed significant by ICA analysis. Upon looking specifically at independent components 4 and 5, we were able to set the same signal intensity threshold as the paper did, a threshold of 2, which would distinguish significant genes from those that were unaffected by ICA. With their ICA analysis, the paper distinguished 94 genes that were either up-regulated or down-regulated in severe AD samples. Our ICA analysis singled out an average of 99 affected genes after our data was scaled and the threshold was applied. Though the genes we looked at were not the same as those of the paper due to the elimination of parts of our dataset from the beginning, we were still able to capture the essence of ICA highlighting gene pathways that are significant in AD onset.

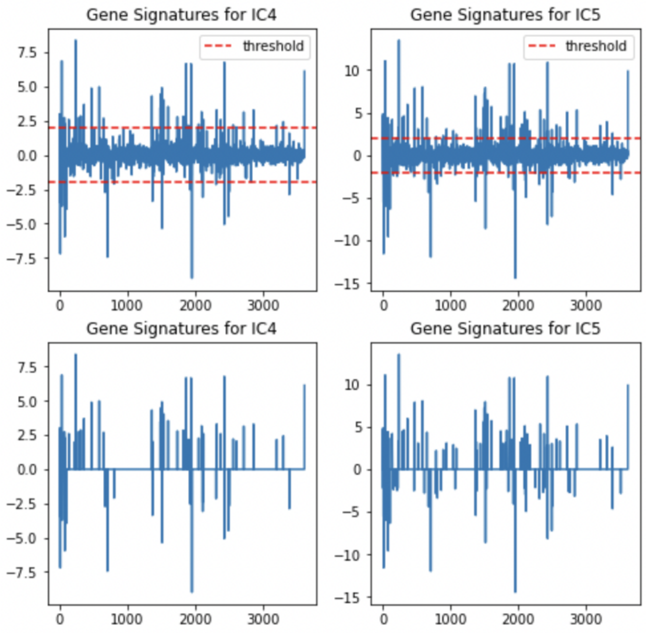


Figure 5. Gene signatures for independent component 4 (top left) and independent component 5 (top right), with genes on the x-axis and signal intensity on the y-axis. The threshold was set to 2, and signal intensities below this threshold were set to zero in the bottom graphs for independent component 4 (bottom left) and independent component 5 (bottom right).

Citations

[1] Matthews, K.A., Xu, W., Gaglioti, A.H., Holt, J.Ba., Croft, J.B., Mack, D. and McGuire, L.C. (2019), Racial and ethnic estimates of Alzheimer's disease and related dementias in the United States (2015–2060) in adults aged ≥65 years. Alzheimer's & Dementia, 15: 17-24. https://doi.org/10.1016/j.jalz.2018.06.3063

[2] Kong, W., Mou, X., Liu, Q. et al. Independent component analysis of Alzheimer's DNA microarray gene expression data. Mol Neurodegeneration 4, 5 (2009).

https://doi.org/10.1186/1750-1326-4-5

[3] Blalock EM, Geddes JW, Chen KC, Porter NM et al. Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. Proc Natl Acad Sci U S A 2004 Feb 17;101(7):2173-8. PMID: 14769913

[4] Ringnér, M. What is principal component analysis?. Nat Biotechnol 26, 303–304 (2008). https://doi.org/10.1038/nbt0308-303

[5] Xuyang Lu, Qijia Jiang, Boying Meng, Comparison of Three Different Matrix Factorization Techniques for Unsupervised Machine Learning. OpenStax CNX. Dec 18, 2013 http://cnx.org/content/col11602/1.1/