Project 2 Report

Ling Jin (Student ID: 011880184)

Introduction

Gene expression profiling is a powerful tool in biomedical research, particularly in the context of cancer studies. However, the massive scale of gene expression data poses considerable challenges for analysis. In the dataset provided for this project, each of the 801 subjects has measurements across 20,531 genes, with each subject classified into one of five cancer types: BRCA, COAD, KIRC, LUAD, or PRAD. The key problem to be resolved involves managing the high dimensionality of the data while extracting biologically meaningful patterns that could differentiate cancer types.

High-dimensional data often contain redundant or noisy features that mask the underlying structure. Principal Component Analysis (PCA) is a widely used dimensionality reduction method that finds orthogonal linear combinations of features (genes) that explain the largest possible variance in the data. Traditional PCA, however, produces components that involve all original variables, making interpretation difficult in biological contexts. Sparse PCA extends PCA by introducing sparsity in the loadings, leading to components influenced by only a small subset of genes, enhancing interpretability.

The proposed solution is to first clean and preprocess the gene expression data by filtering out genes with excessive zero expression, then apply PCA to explore the variance structure and assess whether a few linear combinations of genes can account for a significant proportion of the variance. To address the challenge of interpretation, Sparse PCA is employed, allowing the identification of key genes contributing to principal components. These methods together offer a pathway to manage dimensionality, visualize the structure of the data, and interpret biological insights.

Methods and Results

To address the problem of high dimensionality, data preprocessing was first performed. Genes that had zero expression for at least 300 subjects were filtered out, reducing noise and

retaining more informative genes. From the remaining genes, a random subset of 1000 genes was selected to maintain computational efficiency while ensuring sufficient diversity. This selection was made using a fixed random seed for reproducibility. The selected gene expression values were then standardized so that each gene had a mean of zero and a standard deviation of one. Standardization was critical to prevent genes with large absolute values from disproportionately influencing the principal components.

Principal Component Analysis was conducted on the standardized data. The PCA revealed that the first principal component explained approximately 12.4% of the total variance, the second explained about 10.7%, and the third explained 8.5%. Collectively, the first ten principal components captured around 53.2% of the total variance. A scree plot of the variance explained by each principal component is shown below, revealing a clear 'elbow' point around the tenth component, confirming that only the first several components are necessary to capture the dominant structure in the data.

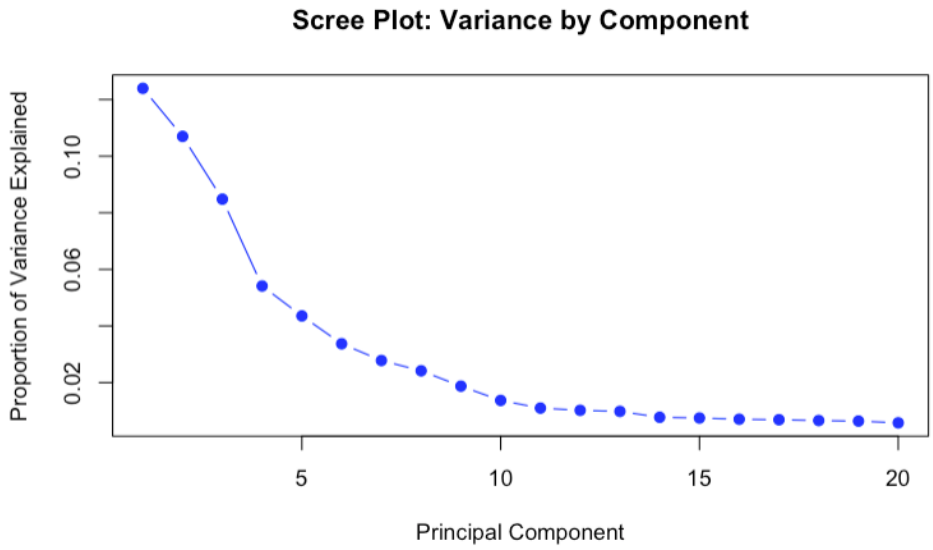

Figure 1: Scree Plot showing Variance Explained by Principal Components.

The cumulative variance explained by the first 20 principal components was plotted, as shown above. The curve levels off after about 10 to 15 components, suggesting that additional components contribute increasingly less to explaining the variance.

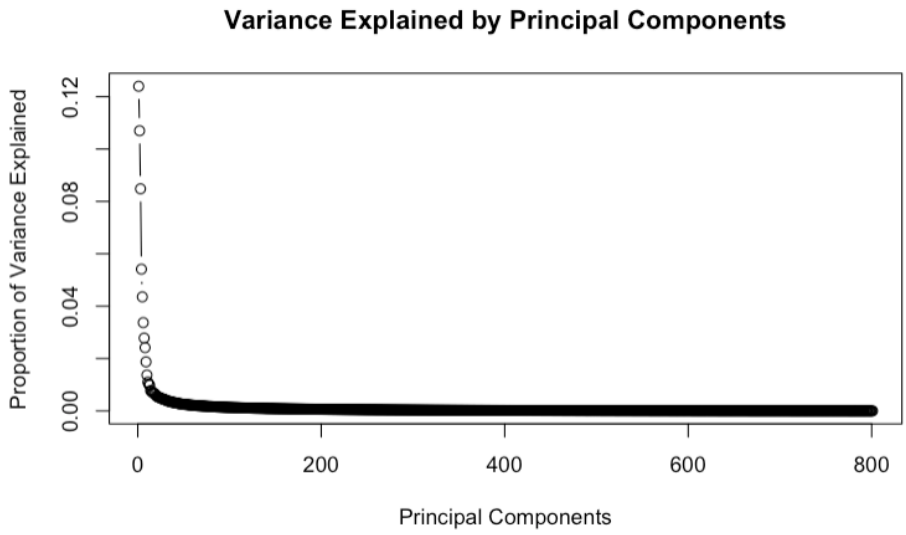

Figure 2: Variance Explained by Principal Components.

The variance explained by all principal components was plotted. As shown above, the majority of variance is captured by the first few components, while later components contribute very little.

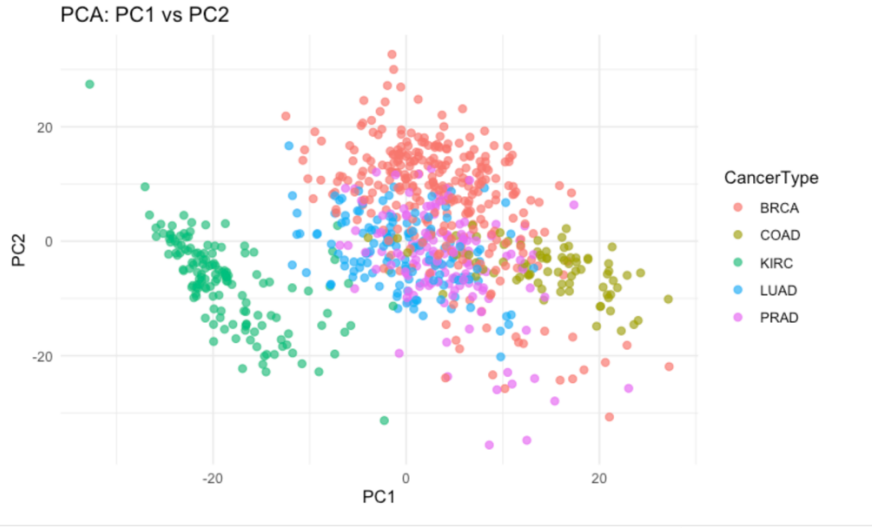

Figure 3: PCA Scatterplot of PC1 vs PC2 Colored by Cancer Type.

The principal components were also visualized by projecting the data onto the first two PCs. When colored by cancer type, distinct clusters become apparent. In particular, KIRC

samples form a very tight and distinct cluster, whereas BRCA and LUAD samples are more overlapping.

Additionally, a scatterplot of PC1 versus PC3 was generated, showing even better separation for certain cancer types, particularly PRAD, which clustered more distinctly along PC3 than PC2.

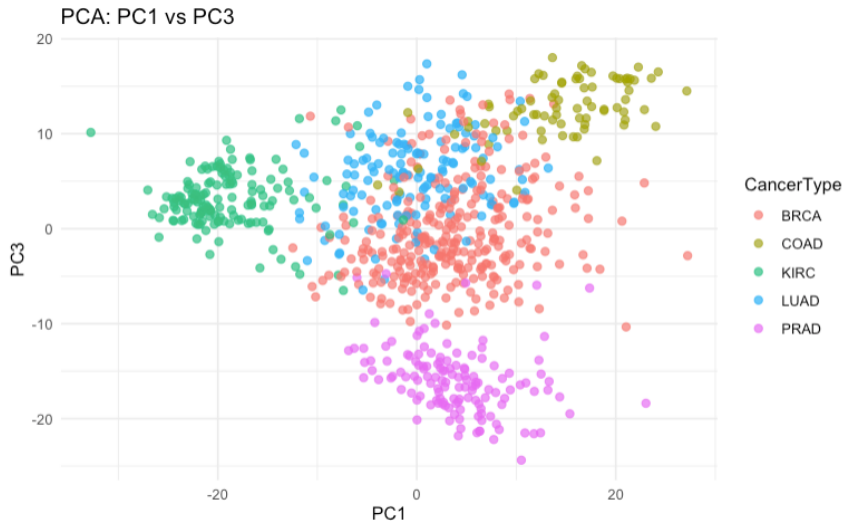

Figure 4: PCA Scatterplot of PC1 vs PC3 Colored by Cancer Type.

To enhance biological interpretability, Sparse PCA was applied. Although Sparse PCA explained slightly less variance compared to standard PCA, the tradeoff was worthwhile due to the improved interpretability.

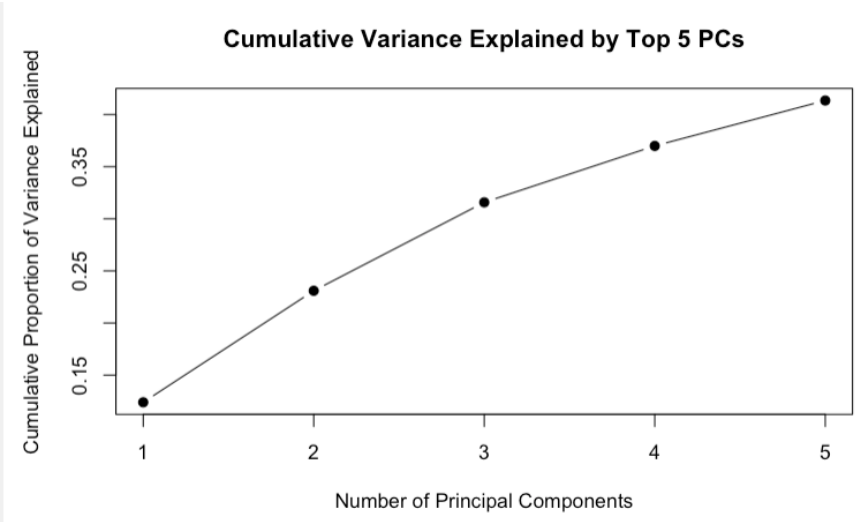

Figure 5: Cumulative Variance Explained by Top 5 Sparse Principal Components.

A scatterplot for Sparse PCA, displaying the first two sparse components, shows that the clustering pattern was similar to that of standard PCA, with KIRC maintaining clear separation.

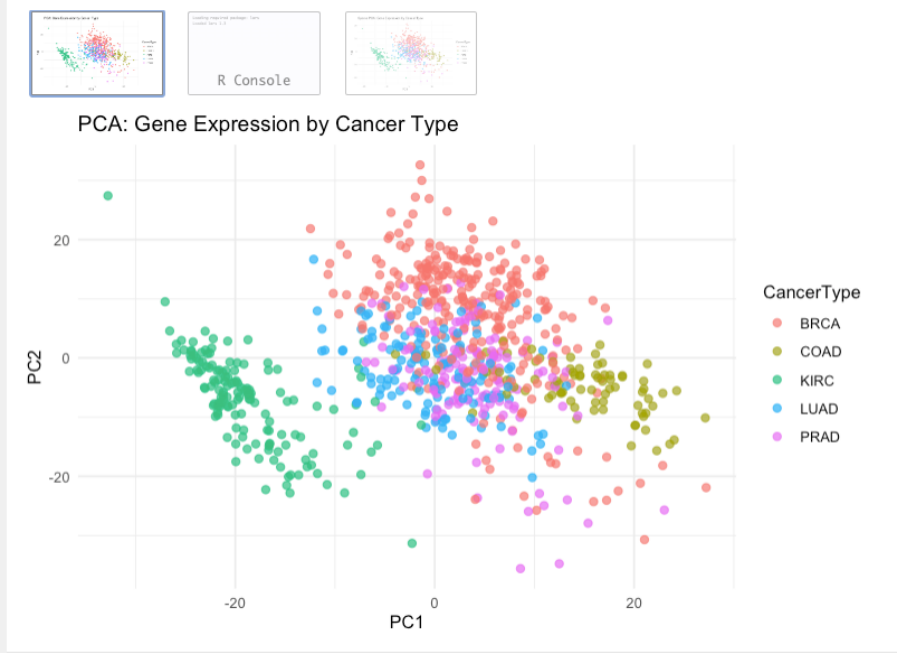

Figure 6: Sparse PCA Scatterplot of Sparse PC1 vs Sparse PC2 Colored by Cancer Type.

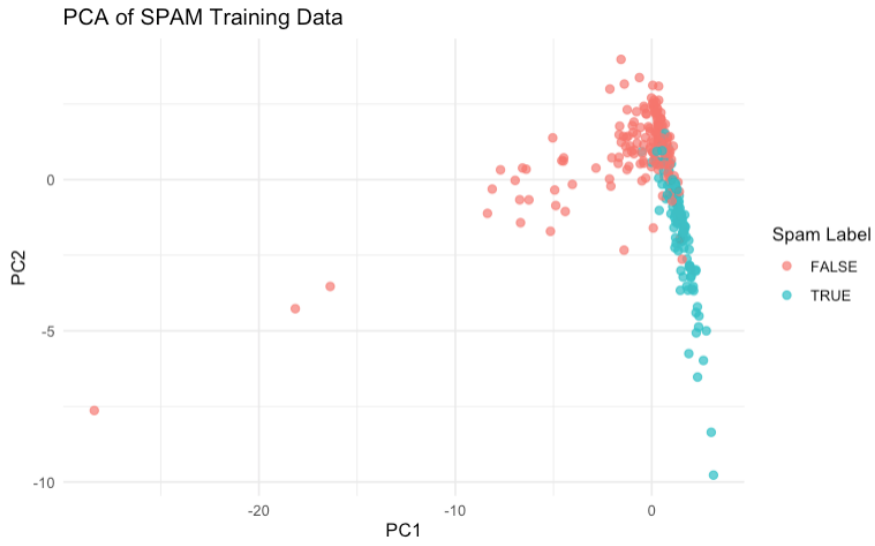

Figure 7: PCA Scatterplot of SPAM Training Data Colored by Spam Label.

Additionally, a PCA analysis was conducted on the SPAM training dataset. The first two principal components were plotted and colored based on whether the email was spam (TRUE) or not (FALSE). As shown in Figure 7, the two classes exhibit some separation, suggesting that PCA captures meaningful structure in the spam classification problem.

Discussion

While the PCA and Sparse PCA analyses were successful in revealing meaningful patterns within the gene expression data, several aspects warrant further discussion and potential improvement. One limitation of the approach is that PCA and Sparse PCA are inherently linear methods and may fail to capture complex, non-linear relationships that exist within biological systems. Techniques such as t-distributed Stochastic Neighbor Embedding (t-SNE) or Uniform Manifold Approximation and Projection (UMAP) could be applied in future work to explore non-linear structures and possibly achieve better separation between cancer types.

Another limitation lies in the random selection of 1000 genes, which, while necessary for computational feasibility, might omit important genes by chance. Alternative strategies, such as selecting genes based on variance or differential expression, could potentially yield even better results. Furthermore, the class imbalance among cancer types could affect the interpretation of the PCA plots, suggesting a need for class-balancing techniques or more careful stratified analysis.

An important extension would be biological validation of the genes identified by Sparse PCA. Functional annotation and pathway enrichment analyses could confirm whether these genes are indeed relevant to cancer development or progression. Additionally, using the reduced dimensions from PCA or Sparse PCA as features in classification models, such as support vector machines (SVM), could evaluate how well the extracted structure supports predictive tasks.

Conclusion

In this project, dimensionality reduction techniques were effectively applied to high-dimensional gene expression data to uncover meaningful biological patterns. PCA revealed that a small number of principal components captured a substantial portion of the total variance, facilitating data visualization and interpretation. Sparse PCA further enhanced biological

interpretability by identifying key genes contributing to variance while maintaining much of the structural information. PCA applied to the SPAM training dataset demonstrated that the technique generalizes well to different types of data, successfully separating spam from non-spam emails in reduced dimensions. While the analyses provided valuable insights, future work could involve exploring non-linear dimensionality reduction methods, improving feature selection strategies, and validating biological relevance through functional enrichment analyses. Overall, the combination of careful preprocessing and principled dimensionality reduction provided an effective framework for managing and interpreting large-scale high-dimensional datasets.

Appendix

The full, relevant computer codes that have been used to conduct this analysis are attached separately in the file titled "stat437Proj2.Rmd".