Singular Value Decomposition and Principle Component Analysis

Edwin Sarver

I. Introduction

Singular Value Decomposition (SVD) and Principle Component Analysis (PCA) are mathematical algorithms that allow, primarily, for the simplification of data. This project implemented a command-line application to compress an image by using SVD. PCA was used to analyze a data-set comparing the genetic markers that may determine the type of Leukemia a patient may have.

II. Algorithms

Two main algorithms were used in this project: Singular Value Decomposition (SVD) and Principle Component Analysis (PCA).

A. Singular Value Decomposition (SVD)

Singular Value Decomposition is a mathematical method that uses Linear Algebra techniques to separate a matrix into the parts that are most mathematically important. A given matrix M can be separated into three matrices U, Σ, and V.

1) Theoretical Performance: The purpose of the SVD algorithm in this project was to produce a compressed image. The implementation did not need to be fast. The performance of the algorithm was, therefore, determined by the space-efficiency of the file that was output.

The theoretical space efficiency for the implemented program is

- 3 longs (64-bits each)
  - 1 for height of the image
  - 1 for width of the image
  - 1 for rank
- 1 char for the maximum gray-scale value
- 16-bits for each item in U, Σ, and V

The number of elements in Σ will be equal to the rank of the image. The number of elements in the U matrix will be the rank times with height of the image. Similarly, the number of elements in the V matrix will be the rank times the width of the image.

Therefore, the compression of the image will be proportional to the rank selected by the user for a given image.

The theoretical size of the image is shown in Equation 1

$$3 * 8 + 1 + 2 * rank * (height + width + 1) \quad (1)$$

The theoretical error for each image will depend on the image itself. This information, however, can be found in the singular value matrix, Σ. If the rank is k, the error in the approximation will be the element at $k+1$ in the diagonal of Σ (i.e. the element at $(k+1, k+1)$).

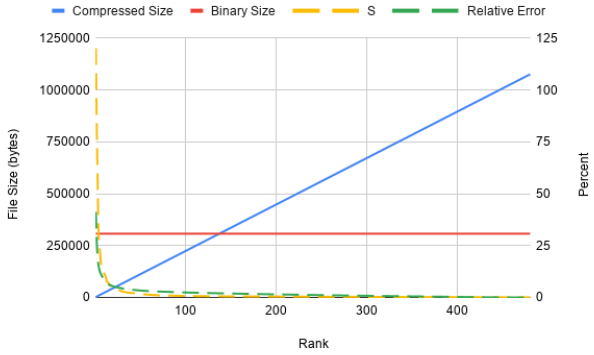

Fig. 1. File Size vs Rank

2) Experimental Performance: As expected, the space used for a given image increases proportionally as the rank increases as shown in Figure 1.

The approximation error is also shown in Figure 1 as the dotted lines. The actual error is very close to the theoretical error, but because the compressed image was stored as a half-precision float, the error does tend to be a little different.

B. Principle Component Analysis (PCA)

Principle Component Analysis simplifies the analysis of multi-dimensional data-sets by reducing the dimensionality of the data. This makes data visualization easier by finding which dimensions of the data have the greatest impact on the categorization of the data.

Reducing the dimensionality of data is important when a given problem may have many possible related factors such as in the realm of genetics. In our project we studied the factors that be related to 2 different kinds of Leukemia, Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML).

In order to process the data, it must be set up with rows being the observations and the columns being the variables. We then calculate the matrix B so that the mean of each column is subtract from it's constituent parts. The co-variance matrix can then be calculated by $S = B^T B/(n-1)$.

S can then be decomposed into eigenvectors and eigenvalues (V and D respectively) such that $S = VDV^T$. This is the same as SVD, except that the matrix $U = V$.

The D matrix shows the relative strength of the components that are represented by the columns in the V matrix.

By choosing the top 2 or 3 principle components (the highest values in D), the corresponding weights in the V

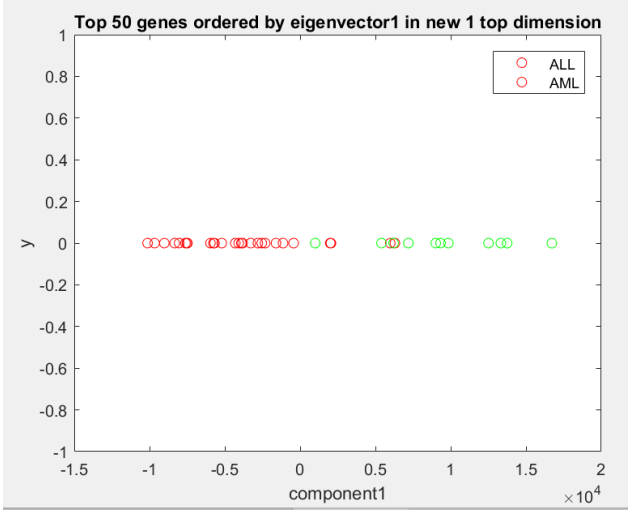

Fig. 2. Top 50 Genes, 1st Principle Component

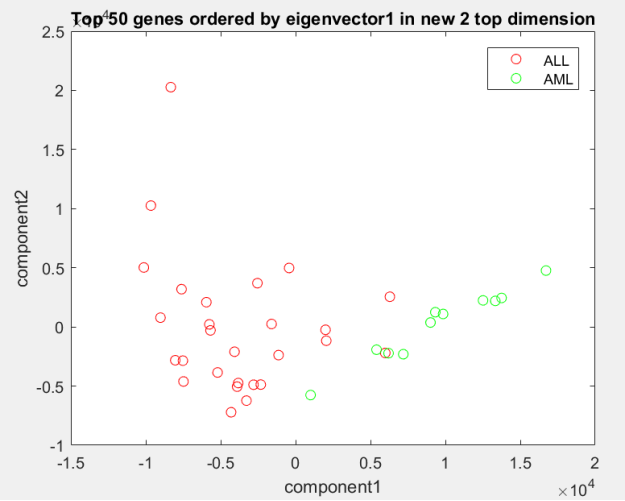

Fig. 3. Top 50 Genes, 2nd Principle Component

matrix can then be used to plot the original observations. This will cause value to clump and thus categorization of the data becomes much easier.

After applying PCA on the Leukemia data-set, the ALL and AML data is easier to categorize due to the separation between the data-points, as shown in Figures 2, 3, and 4. ALL and AML have distinct features within the 50 genes represented in the graphs of the principle components.

III. Insights

The modularity and generalizability of SVD and PCA make them valuable for many applications. Static image compression is just one of the many possible ways that SVD can be utilized. Another application is for video compression by taking the difference between frames and then compressing the resulting difference. Any application that requires the compression or analysis of highly-dimensional data can be

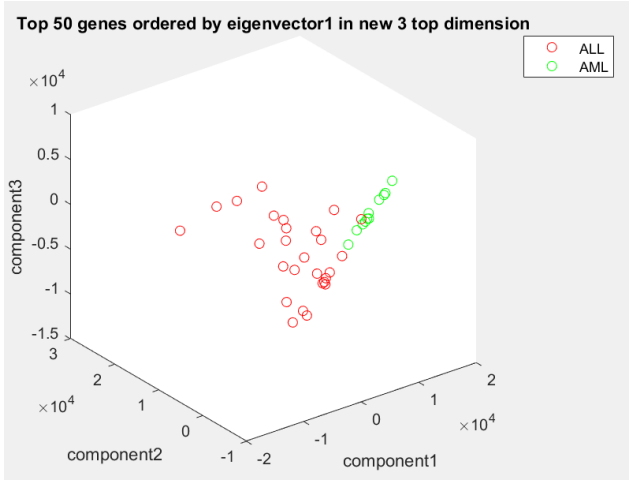

Fig. 4. Top 50 Genes, 3rd Principle Component

a possible application for SVD and, in the case of analysis, PCA.

IV. Test Cases

In this project, many forms of testing were utilized to test the image processing implementation: Unit tests, correctness tests, and performance tests.

A. Unit Tests

Unit tests were used to ensure that each part of the program functioned as intended even after changes were made to the code base. Included in these tests were tests to verify the algorithms described in section II. The PGM images used in the unit tests were very small and were only large enough to ensure that the basic functionality of the SVD algorithm was correct.

B. Correctness Tests

Correctness tests for the SVD implementation were manual because they required the analysis of images. The image used to verify the correctness of the algorithms was the provided image of the College of Arts and Sciences building on the University of Akron campus. In the course of testing, several issues were identified and corrected. Most of the issues that were identified arose from incorrectly calculating the position in the matrix that a particular element in the compressed, binary image file should occupy.

C. Performance Tests

Performance test were also manual. An image of Saturn was downloaded from [2]. The image was translated into a binary representation, then the original pgm was split into a header and data file with a python script. The split files were then compressed using the SVD compression algorithm across all possible ranks. The resulting binary files were then compared against the binary files that were generated directly from the original PGM files for size. The compressed files were then uncompressed to pgm files and were inspected

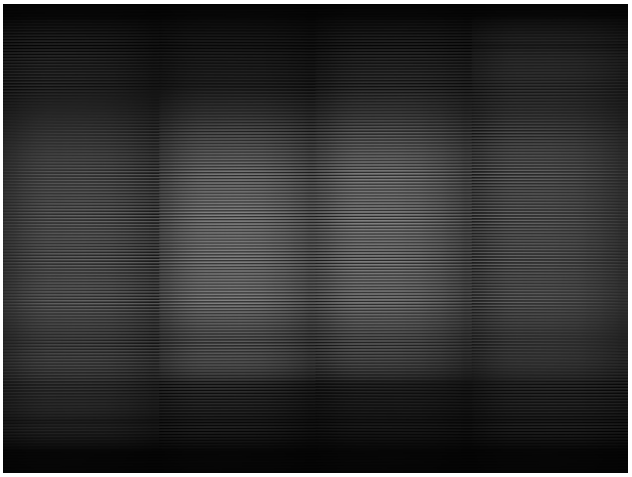

Fig. 5. SVD approximation at Rank 1

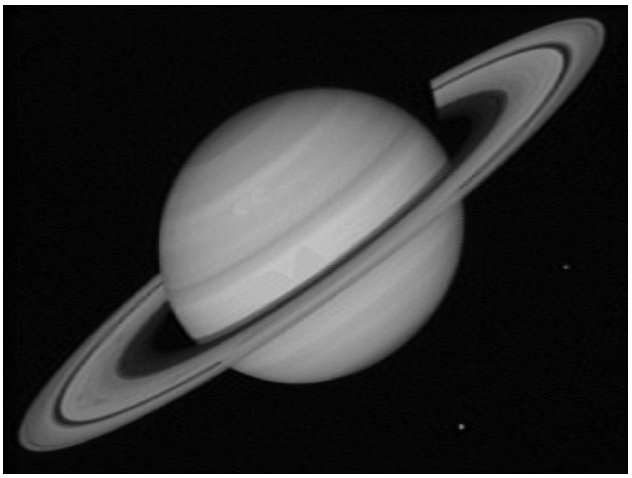

Fig. 6. SVD approximation at Rank 91

for visual fidelity and the error from the original image was calculated. The relative error for each rank on each image could therefore be calculated. The images at various stages of approximation are shown in Figures 5, 6, and 7.

The relative error can also be calculated from the singular values in the Σ matrix. The program was therefore modified to print the singular values to a separate file for further analysis.

V. Conclusion

Singular Value Decomposition is a highly generalizable, powerful algorithm that creates a simplified mathematical representation of a data-set that describes which components are most important to the overall representation of the data-set. This allows for image compression by removing less important components of the image, but also allows for data analysis by the same mechanism when used in Principle Component Analysis.

References

[1] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, Introduction to Algorithms, 3rd ed., The MIT Press, 2009, pp. 594–602, 709–731

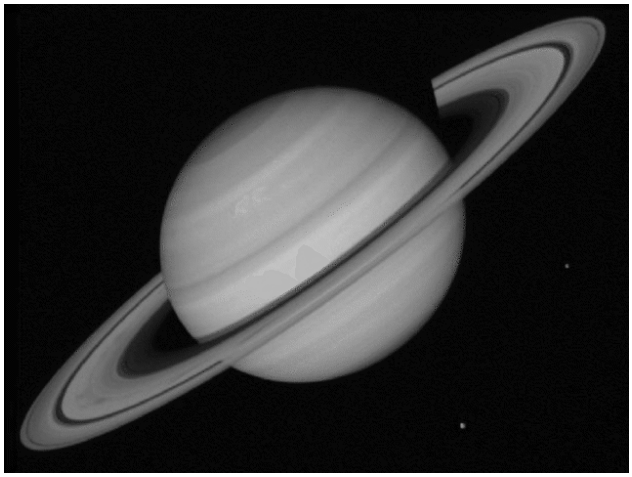

Fig. 7. SVD approximation at Rank 480

[2] ”PGMA Files,” fsu.edu, Jun 10, 2011. [Online]. Available: https://people.sc.fsu.edu/ jburkardt/data/pgma/pgma.html. [Accessed: Nov. 11, 2019].