Explorations – The Journal of Undergraduate Research, Scholarship and Creativity at Wright State

Volume 1
Issue 1 *Summer Undergraduate Research, Scholarship and Creative Activities Issue 2012*

Article 2

2012

Multivariate Data Analysis

Diana Copeland
Wright State University, copeland.33@wright.edu

Michael Raymer
Wright State University, michael.raymer@wright.edu

Multivariate Data Analysis

Cover Page Footnote

[1] http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html

I. INTRODUCTION

A. Multivariate Data Analysis

Multivariate analysis is the study of data that contains more than one variable per unit that is being studied [5]. Answers to very challenging questions can be obtained through analyzing multivariate data [5]. Multivariate data analysis is very important in many different fields such as Bioinformatics, Psychology, Finance, Education and many others. There have been studies conducted such as "A Multivariate Analysis of Youth Violence and Aggression: The Influence of Family, Peers, Depression, and Media Violence" [6], there are health studies such as lung cancer, breast cancer and others, and there are also environmental studies as well as studies on things such as bank loans and economy. Multivariate data is important because units differ whether the unit is a human, a lake, or an animal, there are multi variables in all of those units so it is important that we can see all the variables at once in order to make a good analysis of the data. Projections are used in multivariate analysis to represent the images and bring them to a surface (i.e computer screen). In multivariate analysis linear projection methods have been found very useful in exploration of multivariate data. Some linear projection methods include linear discriminant analysis (LDA), principal component analysis (PCA), and projection to latent structures (PLS). However, in this study we will only be concerned with PCA. PCA's main objective is to reduce data and interpretation [7]. PCA is used often for various multivariate analysis studies.

II. METHODS

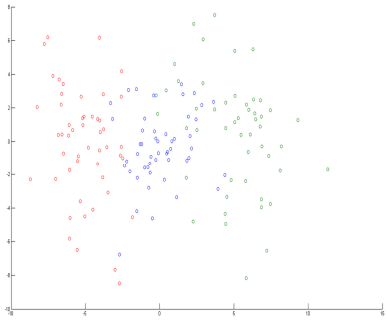
A. Program

In the programming language of Java, I developed a software tool for implementing linear projection methods. My code used several of the standard Java libraries and classes as well as two classes I created on my own. There was a significant amount of graphics programming that was to be used in order to create the projection tool. My program reads in data from a ".txt" formatted file with the first line of the file telling how many different points there are to be represented and the location of each individual point. The data is then represented as a set of 3D points which are referred to in my program as Point3D objects, each object was drawn to the screen in the shape of a circle and color coded. I used homogeneous coordinates which are actually 4D, they allow for the initial three and "1". I also had a four-by-four rotation matrix and a four-by-four translation array. The rotation array was set up differently for both rotating around the x-axis and the y-axis of the screen. It was then projected to 2D so that you can view any projection that you want.

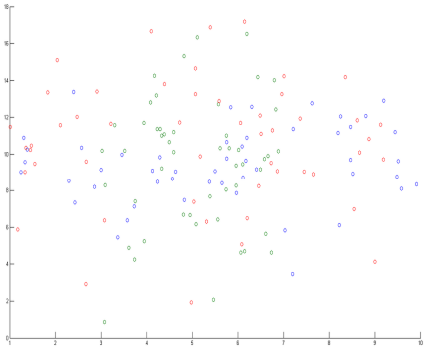
B. Manual Rotation VS. PCA

My data was provided by my research mentor Dr. Michael Raymer there were four sets of files that all differed. I then compared some of my best rotations against that of the PCA. I used MATLAB to generate the PCA results. The files that were loaded into my software program is the same files that were uploaded to MATLAB to produce its results. I first produced images of all the files using my program and rotated the data with the mouse until I found the best possible rotations and saved pictures of them. I then uploaded the files into MATLAB and took pictures of how the data started, and the results of the PCA. In both my programs and in MATLAB the points start off in the same place. When comparing I looked over the results of the PCA and my best rotations. After all cases, I would attempt to produce the very same results that were provided by MATLAB.

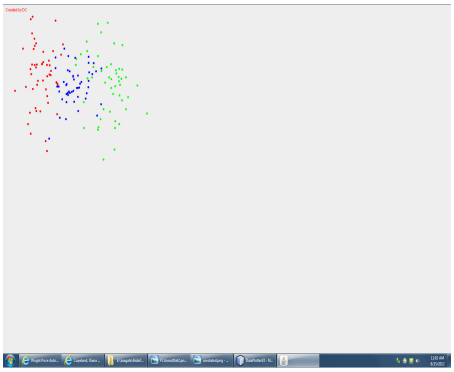
III. RESULTS



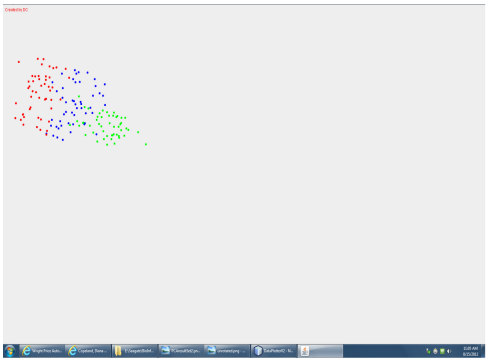
Data set 2 of 4: Unrotated



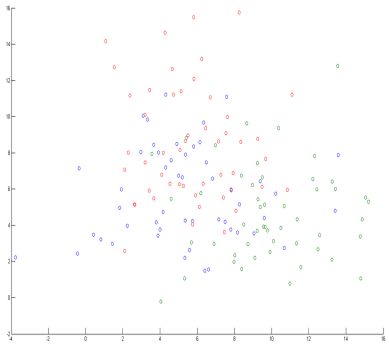
Data Set 2 of 4: PCA results for best rotation



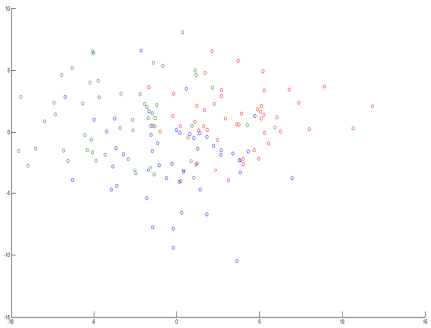
Data set 2 of 4: Manually matched PCA results



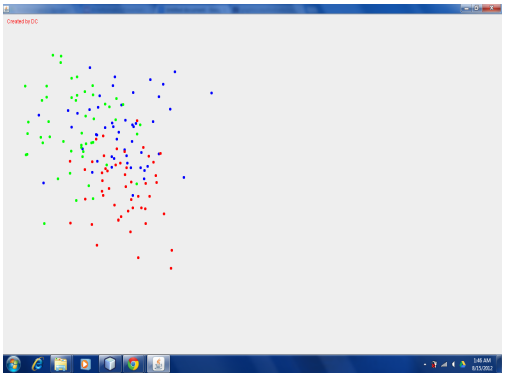
Data set 2 of 4: Best Rotation obtained manually



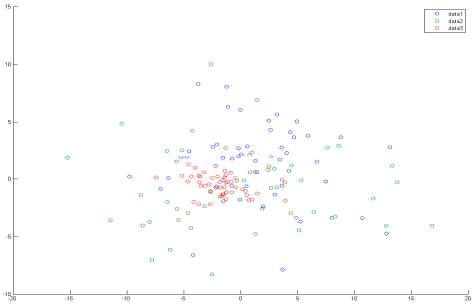
Data set 4 of 4: Unrotated



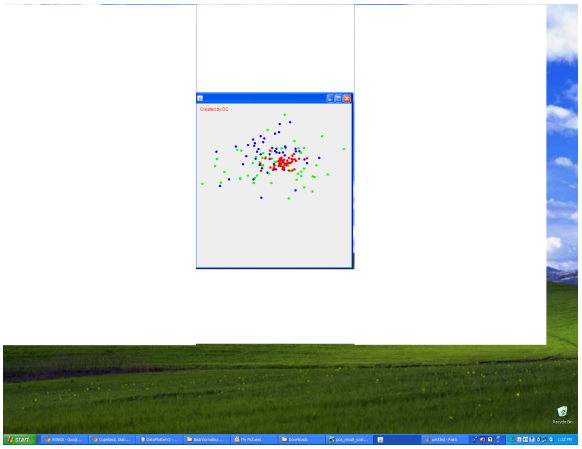
Data set 4 of 4: PCA results for best rotation

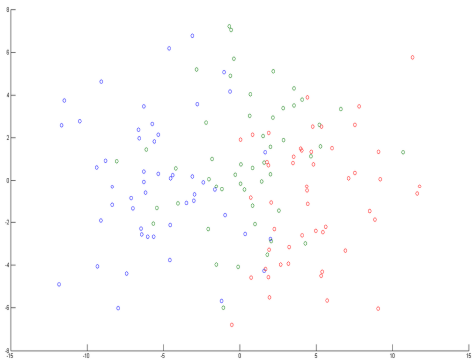


Data set 4 of 4 : Manually matched PCA results



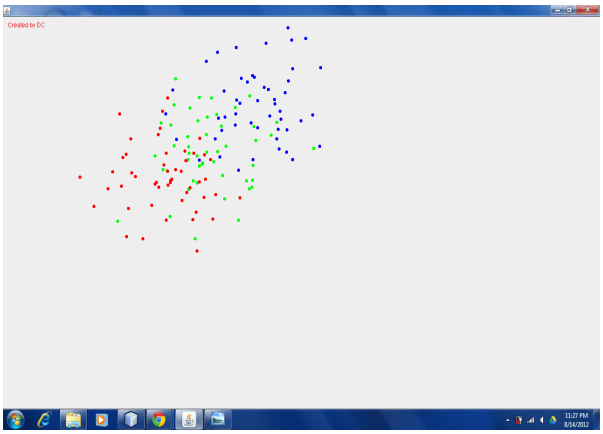
Data set 1 of 4: PCA results for best rotation





Data set 1 of 4: Best rotation obtained manually

Data set 3 of 4: PCA results for best rotation



Data set 3 of 4: Best rotation obtained manually

IV. Conclusion

Overall PCA is a good tool to use for multivariate data analysis, however, my software program was able to match the results PCA provided for the different sets of data and it could also achieve better results in some three of the four sets of data used. Manual rotation could aid experts a lot when analyzing multivariate data, because you don't just get the best rotation, you could find several. In two of the four sets of data used, I did find more than one best rotation, and that is not including the software programs matching of the PCA results. I found that greater results are achieved by using the manual rotation tool because it allows you to see the data as it is being separated and it allows for options as to where to separate the data when looking for the best rotation.

V. References

[1] http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html

[2] Zhang, D., Jing, X., & Yang, J. (2006). Linear Discriminant Analysis. In D. Zhang, X. Jing, Jing, & J. Yang (Eds.), Biometric Image Discrimination Technologies: Computational Intelligence and its Applications Series (pp. 41-64). Hershey, PA: Idea Group Publishing.

[3] Jolliffe, I. T. (1986). Principal component analysis. New York: Springer.

[4] Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, 58(2), 109–130.

[5] A. Afifi, V. A. Clark, and S. May, *Computer-Aided Multivariate Analysis* 4th ed. Boca Raton, FL.: Chapman & Hall/CRC, 2004. Print.

[6] C. J. Ferguson, C. S. Miguel, and R. D. Hartley, "*A Multivariate Analysis of Youth Violence and Aggression: The Influence of Family, Peers, Depression, and Media Violence," Journal of Pediatrics.,* vol. 155, issue 6, pp. 904-908.e3, 2009.

[7] F. Miguez, "*Introduction to R for Multivariate Data Analysis*" Available : https://netfiles.uiuc.edu/miguez/www/Teaching/MultivariateRGGobi.pdf