Multivariate Marble Analysis

San Jose State University
Branden Lopez, Nick Sobrepena, Louis Mutter

December 16, 2023
Report prepared for MATH 257
Instructor Dr. Gottlieb

TABLE OF CONTENTS

I. INTRODUCTION

The city of Ephesos, also known as Ephesus, was an ancient Greek city, the ruins of which are now in modern-day Turkey. Ephesos is an important point of interest to researchers of ancient Greece, as Ephesos is home to important landmarks such as the Temple of Artemis, one of the seven wonders of the ancient world [https://www.worldhistory.org/ephesos/]. During the city's growth over hundreds of years, many different buildings were built with marble from different quarries. Therefore, correctly identifying marble to its corresponding quarry is an essential to the history of the region's development. [https://onlinelibrary.wiley.com/doi/full/10.1111/j.1475-4754.2009.00470.x] In order to better understand the history of Ephesos and ancient Greece as a whole, Ephesos has been the subject of archeological digs and research for over one hundred years, providing an ample amount of data about these marble samples that will be used for our analysis.

There are several different methods that can be used to study the properties of the marble samples. A combination of the following are the marble properties used in this project. The simplest method is to study its most readily observable properties: grain size and color. Grain size is a common measure in geology to identify sediments and rocks, calculated by finding the average diameter of the particles that are separated by the sample through erosion [https://geologyistheway.com/sedimentary/grain-size/]. In addition, the oxygen and carbon isotopic ratios of the samples are used, which represent the ratio of two different isotopes for a single element [Isotopic Ratio | SpringerLink]. Electron paramagnetic resonance (EPR) spectroscopy produces several different measurements by measuring energy differences in atomic states. In this project, we will be studying the line width and intensity produced by EPR [What is EPR? | epr_facility (utexas.edu)].

Using this information, we seek to answer the question, "Are marbles from different groups different according to the numerical variables collected?" using Multivariate Analysis of Variance (MANOVA). The analysis will include assumption validation multivariate-normality, homogeneity of covariance matrices, multicollinearity, outliers, and include any necessary transformations.

II. DATASET

The data is taken from A.B. Yasuv et al (2011) "An Updated Multi-Method Database of Ephesos Marbles, Including White, Greco Scritto, and Bigio Varieties," [1] containing a sample of 244 marbles from 16 quarries in Ephesos with 9 variables. There are six continuous numeric variables, one discrete numeric variable, and two categorical variables. Yasuv performed this study to include Ephesos marbles, where it has been previously excluded from research.

Each of the continuous numeric variables represent different physical and chemical properties of the marble samples. The variables d180, d13C, epr_intens, epr_linwid, color, and mgs represent the Oxygen Isotopic Ratio, Carbon Isotopic Ratio, Electron Paramagnetic

Resonance intensity, Electron Paramagnetic Resonance line width, color value from 0-255 (expressed as a percent), and maximum grain size respectively. The discrete numeric variable is dolomite, representing the composition of dolomite as a rounded percent.

The two categorical variables are quarry and group. The quarry variable represents the 16 different quarries that the marble samples are taken from: Golluce, Urfali, Torbali, Ahmetli, Keftli, HC2, HC3, HC4, Kusini-Tepe, Belevi, Aya Klikiri, Farm, HC1, Belevi grey, 7 sleepers, Fault. The group variable was created in [1], grouping the previously mentioned quarry variable based on physical distance and calculated probability that a sample will belong to a group. Each group is listed from 1 to 6, representing Ephesos-1 Group (n=88), Ephesos-2 Group (n=38), Aya Klikiri Group (n=10), Farm Quarry Group (n=14), Hasancavuslar Quarry 1 (n=19), Belevi and Mt. Panayir Quarries (n=75) respectively. The group variable was derived as a result from a multivariate discriminant analysis performed in Yasuv, et al (2011).

III. METHODS

Kruskal-Wallis

While ANOVA is an excellent tool in determining the difference between means of groups, it assumes that the data follows normal distribution; when normality is violated, the non-parametric Kruskal-Wallis (KW) replaces the former. The KW test makes no assumptions about the mean or variance, making it the non-parametric equivalent to ANOVA. Only assuming groups are independent, a KW test ranks observations with respect to an ordinal or categorical variable and determines if the groups have the same mean on ranks. [Getting Started with the Kruskal-Wallis Test | University of Virginia Library Research Data Services + Sciences]

KW can be used on two or more independent groups where each group has a sample size of 5 or more. Using the ranks of the data to calculate the test-statistic, H,

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

where N is the total sample size, k is the number of groups we are comparing, is the sum of ranks for group i, and is the sample size of group i. H is compared to a cut-off value determined by a chi-squared distribution with k-1 degrees of freedom; if H is larger than the cutoff, we reject the null hypothesis.

For KW the null hypothesis is that the medians of each group are the same, meaning that all groups come from the same distribution. The alternative hypothesis is that at least one of the groups has a different median, meaning at least one comes from a different distribution than the others.

Andrews' Curves

Due to the high-dimensionality of the data, we selected to explore our data using Andrews' Curves. Discovered by D.F. Andrews in 1972 and published in the article, "Plots of High-Dimensional Data", the Andrews' Curves are a method of "plotting data of more than two dimensions" [2], where each data point in p-dimensional space, $\mathbf{x} = (x_1, x_2, \cdots, x_p)$, where each x_i for i = 1, …, n is a numeric variable, is mapped to a function

$$f_{\mathbf{x}}(t) = x_1 / \sqrt{2} + x_2 \sin t + x_3 \cos t + x_4 \sin 2t + x_5 \cos 2t + \cdots,$$

and the function is plotted on the range $-\pi < t < \pi$. Using this technique, we are not limited by the high-dimensionality of the data. Andrews describes that, "one is accustomed to examining plots of functions, and they may be infinite-dimensional. This suggests imbedding high dimensional data in a higher dimensional but easily visualized space of functions" (p. 125) is of salience in his proposed technique. Some other interesting properties of Andrews' Curves are discussed in [2], namely that the transformation described above (1) preserves means, (2) preserves distances, which are also proportional to the Euclidean distance between the corresponding points, and (3) the representation also preserves variances up to a constant. In (3), this constant differs whether *p* is even or odd. Because the distances are preserved, Andrews states, "close points will appear as close functions and distant points as distant functions." (p.132)

There are excellent remarks in Andrews (1972) regarding strategies to employ when visualizing high-dimensional data, which we used in our exploratory data analysis. Specifically, we selected to view subsetted data; Andrews suggests favoring small *n* where possible. Another strategy Andrews suggests in [2] is to use principal components. In particular, Andrews states, "low frequencies are more readily seen than high frequencies." As such, we include both the original data and the principal components of the centered data. Lastly, Andrews discusses the importance of the ordering of the variables so that the variable that is of most importance is placed first in the matrix

Principal Components

A statistical technique frequently utilized when dealing with high-dimensional data, "principal component analysis is concerned with explaining the variance-covariance structure of a set of variables through… linear combinations of these variables." (Johnson, 2007)[3] Using these linear combinations of variables, we are able to achieve a reduction in the dimensionality of the data, with a goal of retaining the interpretability of the data. By creating linear combinations of the original variables $X_1, X_2, \cdots, X_p$, with a mean and covariance "geometrically,

… these represent the selection of a new coordinate system obtained by rotating the original system with X_1, X_2, ⋯, X_p as the coordinate axes." [3] Johnson further describes the new structure as a simpler and more parsimonious description of the covariance structure.

It is of importance to note that Principal Component Analysis (PCA) does not lend itself to discrimination of the data– rather, it serves to preserve the covariance structure and facilitate interpretation of the original *p*-variables via the *k*-principal components, ideally with $k < p$. To this end, we may utilize scree plots– a plot that depicts the amount of total variance in the sample retained plotted against each individual principal component. To determine the optimal number of principal components to take, there are a myriad of heuristics to aid selection of the first few principal components. We will discuss our particular scenario later.

Johnson describes that PCA, "are more of a means to an end rather than an end in themselves… they frequently serve as intermediate steps in much larger investigations." (p.430)[3] In particular, we will illustrate the combination of PCA and Andrews' Curves to facilitate illustrating possible differences in the values among different groups. While this is not a formalized test, it does serve as a useful tool in the exploration of the data.

IV. RESULTS

To effectively use MANOVA, data must follow multivariate normal distribution. While many assumptions exist, we narrow in on the Multivariate Chi-squared plot. Using the residuals from a MANOVA on groups, we see that our Chi-squared plot does not follow a 45-degree line in figure N; this indicates the data is not multivariate normal. Instead, it is necessary to use a non-parametric method capable of handling non-normally distributed data, Kruskal-Wallis, as defined above. Univariate Kruskal-Wallis determines that not all group distributions are the same, and while Wilcoxon's test finds that groups 4-6 have similar distributions with respect to d180, Multivariate Kruskal-Wallis is unsuccessful. Untold before use is that Multivariate Kruskal-Wallis is not commonly implemented because the results are difficult to interpret, and i cannot handle variables of different scales [NPMV].

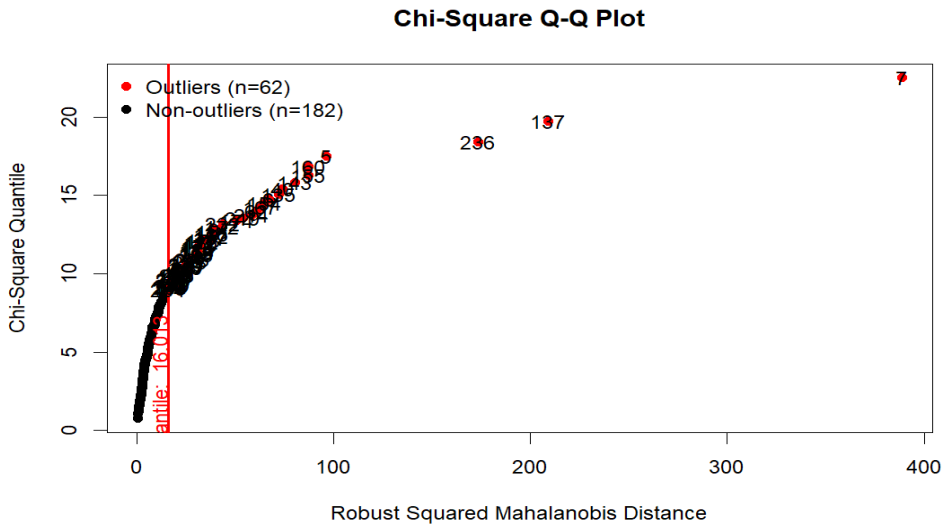

Figure 1: Multivariate Chi-sqared plot

To display the entire assortment of variables present in the data, we present the Andrews' Curves for the entire data set, using all numeric variables: d18O, d13C, dolomite, epr_intens, epr_linwid, color, and mgs in figure 1. Since we construct the Andrews' Curves with seven variables, the form of our particular transformation with the order of the variables considered by the order of importance described by Yasuv et al (2011) is given by

$$f_x(t) = \text{d18O} / \sqrt{2} + \text{d13C} \sin t + \text{dolomite} \cos t + \text{epr\_intens} \sin 2t + \text{epr\_linwid} \cos 2t + \text{color} \sin 3t + \text{mgs} \cos 3t.$$

Each curve represents a row of the original data, transformed according to the formula provided in our Methods section, and is colored according to the group it was assigned. In this case, the group variable is the result of the original analysis conducted in Yasuv et al. (2011). According to Yasuv, et al., the group is the product of a multivariate discriminant analysis that assigns the group based on the highest probability that a marble sample belongs. At a high level, we can see each group, or color, has a typical behavior but also we observe individuals whose behavior is inconsistent with the rest of the group. Specifically, we note the black, blue, and pink curves that behave more extremely and in opposite patterns from the rest of the curves of their own color, but also the vast majority of the data.

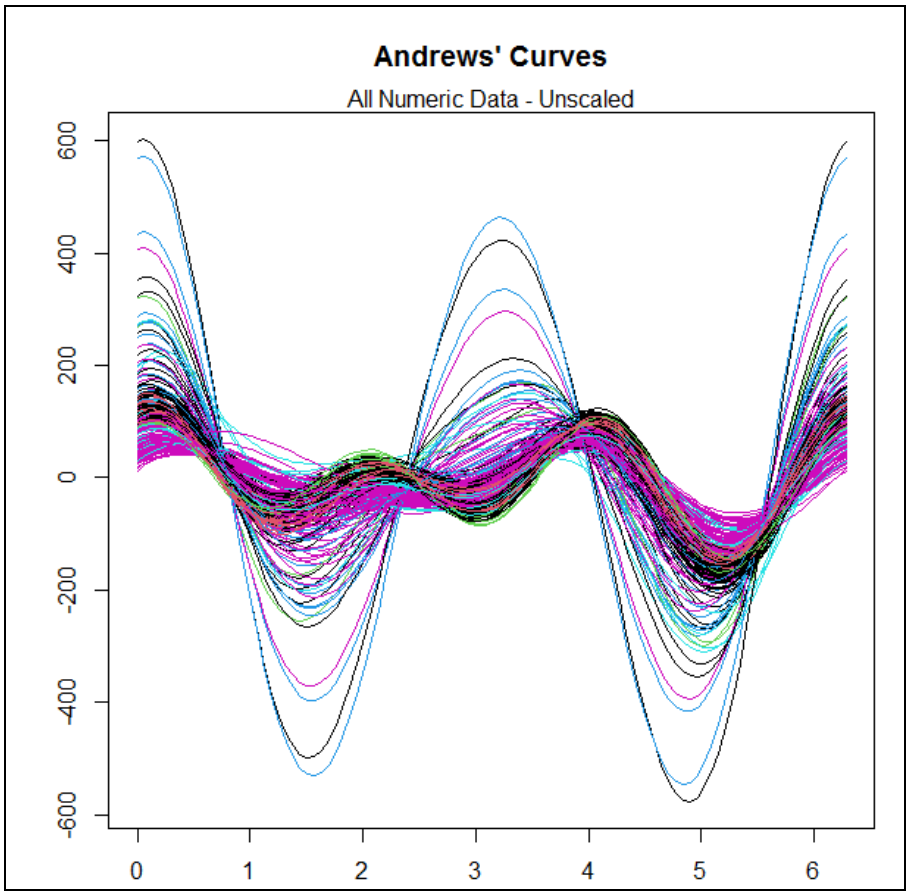

Figure 1. Andrews' Curves for all numeric variables in the original data.

One strategy outlined by Andrews in [2] discusses utilizing smaller n in each plot. To get a better understanding of how the members of each group performed within a particular group, figure 2 breaks out the Andrews' Curves such that each plot contains exactly one group.

We see that among each group in figure 2, there exists at least one curve per panel that does not adhere to the common group's performance. This makes sense because overall, the correct discrimination rate described in Yasuv et all (2011) was reported to be 71% (p.230). As such, we would expect to see a considerable amount of variability, even within a group.

Group 1 (white marbles: Ephesos-1 group, $n_1 = 88$) appears to have approximately nine observations that deviate from the behavior of the rest of the group, and whose samples originate from eight different quarries. Group 2 (white marbles: Ephesos-2 group, $n_2 = 38$) has consistent performance on a high-level, but upon closer inspection, we can see there is some spread within the group between $t = 1$ and $t = 3$. The samples from group 2 originate from two quarries. Group 3 (white marbles: Aya Klikiri, $n_3 = 10$) has two members that behave quite differently from the rest of the group; these all originate from the same quarry. Group 4 (white marbles: Farm quarry group, $n_4 = 14$) appear diverse, marked by few overlapping curves– even considering all of these samples are from a single quarry. Group 5 (Grecco Scritto marbles:

Hasancavuslar quarry-1, $n_5 = 19$) have both a considerable amount of spread for marbles sampled from a singular quarry and members of the group that behave atypically, where approximately 5 of the observations are consistently achieving a different value for a given *t*. Group 6 (Bigio Antico marbles: Belevi and Mt Panayir, $n_6 = 75$) displays a recognizable group pattern near the horizontal axis, but we also observe approximately 15 curves that deviate from the typical pattern, which is at its most variable at $t = \pi/2$ and $3\pi/2$.

Andrews (1972) discussed the use of principal components in his proposed technique, recommending its use (p.135). We created principal components using the centered numeric data, and using the cumulative proportion of sample variance retained, we selected the first two principal components. The first principal component retains approximately 86.91% of the total sample variance, and the second principal component combined with the first retains approximately 95.28% of the total sample variance. We acknowledge that there exist other methods of selecting the *k*-principal components, and believe that our decision to retain a cumulative 95% of the original sample variance is both representative and parsimonious, as we have $k < p$.

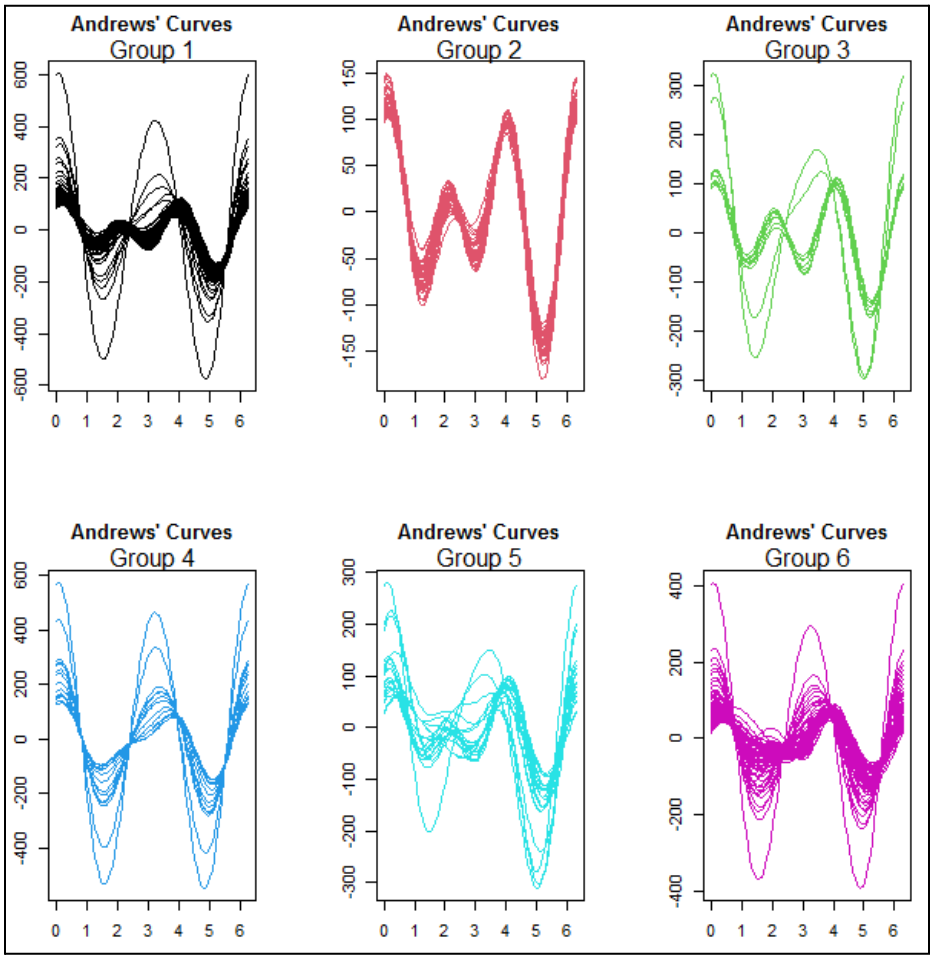

Figure 2. Andrews' Curves for original data, broken out by group.

The following interpretations are based on Table 1 below. PC1 is a contrast between a low isotopic oxygen ratio (d18O) with a high isotopic carbon ratio, a large percentage of dolomite content, a high electron paramagnetic resonance intensity, a high electron paramagnetic resonance linewidth, a higher color percentage (i.e., closer to white), and a higher minimum grain size. PC1 has higher values for marbles with lower values of oxygen isotopic ratio and higher values of carbon isotopic ratio, dolomite content, electron paramagnetic resonance intensity, electron paramagnetic resonance linewidth, color, and minimum grain size. However, PC1 appears to be dominated by dolomite content.

PC2 is a contrast between isotopic oxygen ratio, isotopic carbon ratio, electron paramagnetic resonance intensity, color, and minimum grain size with dolomite content and electron paramagnetic resonance linewidth. PC2 scores are larger for higher values of oxygen isotopic ratio, carbon isotopic ratio, EPR intensity, color, and minimum grain size and lower values of dolomite content and EPR linewidth. PC2 appears to be primarily driven by color and EPR intensity.

Variable	**PC1**	**PC2**
d18O	-0.0110848	0.0394056
d13C	0.0034791	0.0325409
dolomite	0.9987883	-0.0291103
epr_intens	0.0180563	0.2798981
epr_linwid	0.0362384	-0.0010757
color	0.0254506	0.9580138
mgs	0.0000869	0.0201560

Table 1. First two principal components of the centered numeric data.

As mentioned before, Andrews (1972) recommends the use of principal components in his proposed technique. Using the two principal components described in Table 1, we repeated the plotting of Andrews curves to compare the effect of performing PCA on the data. At a high-level, we observed that using fewer principal components presented much simpler curves in the plot. This is another reason to support using fewer principal components when we retain a large degree of original sample variance. The function for plotting the PC Andrews' Curves is given by

$$f_{PC1,\,PC2}(t) = PC1 / \sqrt{2} + PC2 \sin t.$$

Contrasting the above function with $f_x(t)$ from the Andrews' Curves corresponding to the original data, we can see that $f_{PC1,\,PC2}(t)$ is easily interpretable by comparison. We have the first term responsible for the vertical height of the curve, dependent on the score for PC1. However, PC1 does not control any curvature in each particular observation, or curve. The curvature that we observe in each curve is wholly dependent on PC2. And, since we can intuit and interpret which variables had the most influence contributing to each principal component, we can tease apart why, in particular, some curves differ within each group. This is an additional benefit of selecting to work with just two principal components that retain a large proportion of the sample variance.

In the group 1 pane of Figure 3, we note that nine observations have varying degrees of separation from the rest of the group. The primary method in which these curves differ from the majority of the group is due to PC1. In contrast to the previous plot, where we had to account for all seven variables in decreasing importance, all seven play a quantifiable role in determining the vertical curve height. It makes sense that these curves likely differ from the rest of the group due to higher values of dolomite composition, EPR linewidth and/or intensity, and color, or low values of oxygen isotopic ratio on average. In group 1, all of the curves correspond to a smaller value of PC2.

In group 2 pane of Figure 3, we observe the same general group behavior as we previously did, but now have the added benefit of having intra-group separation in two members. There is almost an even distribution among which samples yielded a positive value of PC1 and a negative value for PC1. The two group members whose curves are near-flat are due to the coefficient PC2 modifying the cyclical component, sin *t,* only slightly. Group 2 also still behaves most similarly at the end points for *t*; the distances between curves for $t = 0$ is within a much smaller range than for the other groups. Aside from the two samples yielding a near-flat curve, group 2 appears to have a negative PC2 generally. This likely corresponds to the larger contribution of color to PC2, indicating that the marbles in this group were mostly white.

In the group 3 pane of Figure 3, we observe a large difference in two of the observations, stemming from much higher values of PC1. All curves have approximately a similar curvature, indicating that they attained near similar negative scores for PC2.

In the group 4 pane of Figure 3, the curves in the plot contain a curvature opposite to the curvatures observed of most group behaviors for groups one through three, with the exception of two individual curves that, like group 2, are near-flat. This observation aligns with positive values for PC2. According to the vertical axis, group 4 also has completely non-negative PC1 values.

In the pane of Figure 3 containing group 5, we can see quite a variety of different individuals with respect to both PC1 and PC2. Any similarity we saw in the original data's plot

for group 5 has been separated out considerably. There are examples of individuals whose principal component scores create a curve that can upwards, downwards, and also remain flat. This means that for PC2, group 5 has the widest variety of values. The vertical axis shows that group 5 can have both negative and positive values for PC1. It is visually the most proportionally diverse group of the 6 groups that have been broken out.

For the group 6 pane, we observe approximately four marble samples whose PC2 makes a negligible contribution, as they are near-flat. For the general group's performance, they largely have a negative value for PC1 based on vertical position on the ends of the pane's plot, and a positive value for PC2, based on the curve shape. We have a fair amount of spread observed at *t* = 0, and one sample whose PC1 value places it squarely above the rest.

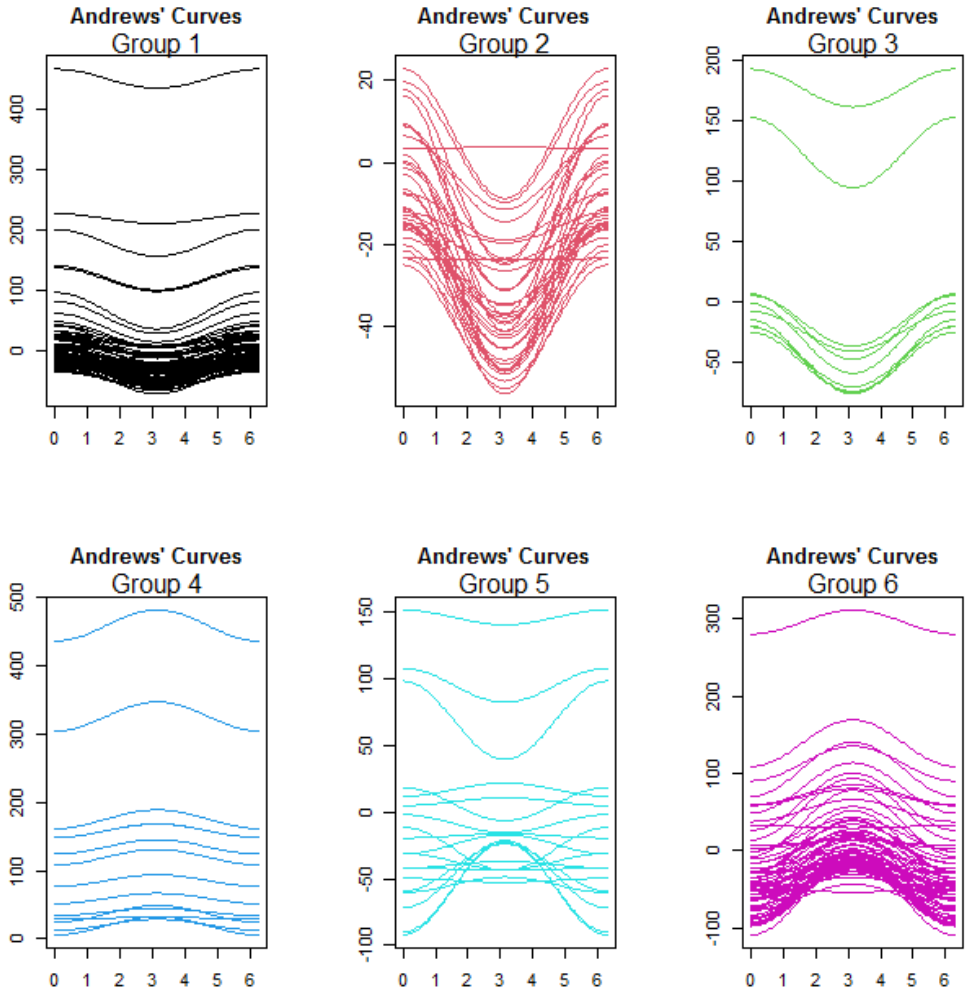

Figure 3. Andrews' Curves post-principal components analysis, plotted by group.

V. DISCUSSION

- Some groups still had a large sample size, making their curves difficult to view
- Data thus far have been proven to be non-normal, and I have not been able to normalize via box cox
- Groups were assigned via a discriminant analysis in [1], perhaps we would get more consistent results had we broken out by quarry instead of group

REFERENCES

[1]. A. B. YAVUZ, M. BRUNO, and D. ATTANASIO, “An updated, multi-method database of Ephesos Marbles, including White, Greco Scritto and BIGIO varieties*,” Archaeometry, vol. 53, no. 2, pp. 215–240, 2011.

[2]. D.F. ANDREWS, “Plots of High-Dimensional Data,” Biometrics, vol. 28, no. 1, pp. 125–136, 1972.

[3]. JOHNSON, RICHARD and WICHERN, DEAN, *Applied Multivariate Statistical Analysis*, 6th ed, Pearson, pp. 430-480, 2007.