STA135: Multivariate Data Analysis Final Project

Student: Aditya Mittal

University of California, Davis

Department of Statistics

adimittal@ucdavis.edu

Instructor: Xiaodong Li

University of California, Davis

Department of Statistics

xdgli@ucdavis.edu

Course: STA 135 - Multivariate Data Analysis

1 Introduction

The objective of multivariate statistics encompass a wide range of applications. In this paper, I will focus on four key methodologies covered in our course STA135 - Multivariate Data Analysis lectures: one-sample inference, two-sample testing, Linear Discriminant Analysis (LDA), and Principle Component Analysis (PCA). My focus will be applying these methodologies to analyze two datasets: 'Milk Transportation Costs' and 'Abalone Measurements'. These datasets will serve as the foundation for exploring and implementing the aforementioned statistical techniques.

1.1 Milk Transportation Costs

The Milk Transportation Costs dataset comprises of transportation costs associated with gasoline and diesel trucks. Specifically, for each type of fuel, we have recorded the following three metrics: Fuel Costs, Repair Costs, and Capital Costs. The dataset includes 59 observations, with 36 and 23 observations for gasoline and diesel, respectively. Our analysis involves two types of hypothesis testing: one-sample inference and two-sample testing. Initially, I conducted one-sample inference exclusively on gasoline trucks to examine whether there is a difference in the mean Fuel, Repair, and Capital Costs are within this group. Subsequently, with the inclusion of second sample from diesel trucks, I conducted two-sample tests to explore if there is any difference in the mean population costs between the two fuel transportation categories. The following table provides a quick glance of the dataset:

Table 1: Milk Transportation Costs Dataset

Fuel Costs	**Repair Costs**	**Capital Costs**	**Type**
16.44	12.43	11.23	gasoline
7.19	2.70	3.92	gasoline
8.50	12.26	9.11	diesel

1.2 Abalone

The Abalone dataset provides several physical measurements pertaining to males, females, and infants. Before conducting any analysis, the data was pre-processed to combine "Male" and "Female" category into a singular "Adult" class. Then, I employed LDA to classify the age category of Abalone (Adult vs. Infant) based on all the given physical measurements. Note that we have 8 different physical measurements given (shown in detail using the table below). Following this, Principal Component Analysis (PCA) was utilized to reduce dimensionality of the dataset while balancing the retention of the overall proportion of variance. LDA was then rerun on the reduced dataset across each principal component to observe the impact on accuracy across smaller datasets. To assess the effectiveness of LDA and evaluate the influence of PCA, I used a test/train split to gauge the performance of this classifier on training and validation data using the misclassification rate as the accuracy measure. Ultimately, this approach aims to provide insights into the accuracy of Abalone predictions based on both the original and reduced datasets, shedding light on the effectiveness of LDA and the impact of PCA on the predictive power on Abalone age classification.

Table 2: Abalone Dataset

Age	Length	Diameter	Height	Whole Weight	Shucked Weight	Viscera Weight	Shell Weight	Rings
A	0.455	0.365	0.095	0.5140	0.2245	0.1010	0.150	15
A	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.070	7
A	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.210	9

2 Methodologies

2.1 One-Sample Inference

In general, one-sample inference is useful determine the possible range of the population mean vector $\vec{\mu}$. In the case of Milk transportation data, I am interested in using one-sample inference to determine if there is any difference in mean Fuel, Repair, and Capital costs in regards with gasoline fuel only. For this question, my null hypothesis can be defined as: H_0: $\mu_1 = \mu_2 = \mu_3$ and I will test this via Hotelling's T^2 at level of $\alpha = 0.05$. Note that, to answer these questions, I have the assumption of having the random sample follow a multivariate normal distribution:

$$X_1, X_2, X_3 \text{ i.i.d.} \sim \mathcal{N}_3(\mu, \Sigma)$$

The summary statistics can be defined as $\vec{x}$ for sample mean vector and S_1 for sample covariance matrix. I define them as the follow:

$$\vec{x} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} \quad S_1 = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{bmatrix}$$

To test our hypothesis, I can first create linear combinations as following: $Y_1 = \vec{X}_1 - \vec{X}_2$ and $Y_2 = \vec{X}_1 - \vec{X}_3$. Then, the means are $E(Y_1) = E(\vec{X}_1) - E(\vec{X}_2) = \mu_1 - \mu_2 = \mu_{Y_1}$, $E(Y_2) = E(\vec{X}_1) - E(\vec{X}_3) = \mu_1 - \mu_3 = \mu_{Y_2}$. Consequently, the new mean vector can be defined as as $\vec{Y} = CX$ with sample covariance matrix $S_y = CS_xC^T$ where the contrast matrix C is:

$$C = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}$$

Then, the linear combinations produces the new sample mean vector and covariance matrix:

$$\vec{y} = \begin{bmatrix} \mu_{Y_1} \\ \mu_{Y_2} \end{bmatrix} \quad S_y = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

With this, my new hypothesis would be testing $H_0 : \mu_{Y_1} = \mu_{Y_2} = 0$. By Hotelling's $T^2 = n(\vec{Y} - \vec{\mu}_y)^\top S_y^{-1}(\vec{Y} - \vec{\mu}_y) \sim \frac{(n-1)q}{n-q} F_{q,n-q}$, I have displayed the results via statistical testing and a constructed the mean-centered ellipse for visual intuition in the results section.

2.2 Two-Sample Testing

Furthermore, in addition to the data from gasoline trucks, I also have 'cost' measures from diesel trucks. Naturally, this leads to the question on whether there exists a difference between the population mean vector between the two types of fuels transportation methods. Furthermore, we have analogous summary statistics for both gasoline and diesel trucks – including fuel, repair, and capital costs for each.

Note that, in conducting two-sample testing, I again operate under the following assumptions: I have two independent 3-variate random samples with the same population covariance $\Sigma_1 = \Sigma_2 = \Sigma$.

$$\vec{X}_1 \sim \mathcal{N}_3(\mu_1, \Sigma) : \vec{X}_{11}, \dots, \vec{X}_{1n_1}$$
$$\vec{X}_2 \sim \mathcal{N}_3(\mu_2, \Sigma) : \vec{X}_{21}, \dots, \vec{X}_{2n_2}$$

Also, denote the summary statistics as $\vec{\bar{X}}_1$, $\vec{\bar{X}}_2$, S_1 and S_2. Note that the summary statistics associated with $\vec{\bar{X}}_1$ and S_1 are pertaining to gasoline trucks and $\vec{\bar{X}}_2$ and S_2 are pertaining to diesel trucks. Thus, I have the following hypothesis to test: $H_0 : \vec{\mu}_1 = \vec{\mu}_2$ versus $H_A : \vec{\mu}_1 \neq \vec{\mu}_2$. I can define difference in sample mean vector $(\vec{X}_1 - \vec{X}_2)$ and pooled covariance matrix S_{pooled} as the following:

$$\vec{X}_1 - \vec{X}_2 = \vec{x}_1 - \vec{x}_2 \quad S_{\text{pooled}} = \tfrac{n_1-1}{n_1+n_2-2} S_1 + \tfrac{n_2-1}{n_1+n_2-2} S_2$$

By Hotelling's $T^2 = (\vec{X}_1 - \vec{X}_2)^T \left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_{\text{pooled}}^{-1} (\vec{X}_1 - \vec{X}_2) \sim \frac{(n_1+n_2-2)p}{n_1+n_2-1-p} F_{p,n_1+n_2-1-p}$, I can conduct our hypothesis testing results are displayed in the results section.

2.3 Linear Discriminant Analysis

The objective of Linear Discriminant Analysis (LDA) is to classify a whether new observation vector $\vec{x}_0$ falls into Class 1 or Class 2. Let's consider the second Abalone dataset, which provides physical measurements for both adults and infants. I am interested in being able to predict the Abalone's age category based on it's physical measurements. Thus, I use LDA to assess whether I can accurately predict whether an abalone is an adult or an infant based on its physical attributes. In this approach, I will assign "Adult" as Class 1 and "Infant" as Class 2. As defined, **Fisher's rule** for classification involves assigning a new observation $\vec{x}_0$ to Class 1 if:

$$(\vec{x}_0 - \bar{\vec{x}}_1)^\top S_{\text{pooled}}^{-1} (\vec{x}_0 - \bar{\vec{x}}_1) \leq (\vec{x}_0 - \bar{\vec{x}}_2)^\top S_{\text{pooled}}^{-1} (\vec{x}_0 - \bar{\vec{x}}_2)$$

Equivalently, this rule can be stated as:

$$\vec{w}^\top \vec{x}_0 \geq \frac{1}{2} \vec{w}^\top (\bar{\vec{x}}_1 + \bar{\vec{x}}_2),$$

where $\vec{w} = S^{-1}\text{pooled}(\bar{\vec{x}}_1 - \bar{\vec{x}}_2)$.

Otherwise, $\vec{x}_0$ is assigned to Class 2. This approach is geometrically intuitive, as it suggests that the Mahalanobis distance of the new test point $\vec{x}_0$ will be closer to the centroid of Class 1 than it is to Class 2. Other perspectives include distributional assumptions via Bayes' Rule, and ultimately lead to the same conclusion as derived by Fisher's rule. In order to measure accuracy, I used two methods: Training Error (Apparent Error Rate) and Testing Error (using Lachenbruch's Holdout). In the textbook for our class, Lachenbruch's holdout is defined equivalently as the Leave One Out Cross Validation Procedure, where we have n - 1 training examples and the n-th data point is the test point. This process is iteratively ran across with each observation used as the test data and the average error rate is computed. In particular, I computed the error for each as following:

$$\text{Misclassification rate} = \frac{\text{Number of misclassified samples}}{\text{Total number of samples}}$$

This will help gauge the effectiveness of LDA on both training and testing (unseen) data.

2.4 Principle Component Analysis

Initially, I had employed Linear Discriminant Analysis (LDA) on the abalone dataset with all available columns, where some that these columns may contain noise and potentially impact our predictive accuracy. With this, I am interested in measuring whether we can represent this data in a smaller dimensional space, and also the the accuracy of running LDA on this reduced dataset. With this, I use Principal Component Analysis (PCA) to reduce the dimensionality of the data. By reducing the number of features, PCA aims to retain the most essential information while minimizing the effects of noise.

The first step in this process was to standardize my data so all the variables had mean zero with standard deviation of one. Let $X_1, \ldots, X_9$ be the original columns. I transformed them into $Z_1, \ldots, Z_9$ as:

$$Z_1 = \frac{X_1 - \mu_1}{\sqrt{\sigma_{11}}},\ Z_2 = \frac{X_2 - \mu_2}{\sqrt{\sigma_{22}}},\ \ldots,\ Z_9 = \frac{X_9 - \mu_9}{\sqrt{\sigma_{99}}}.$$

If I represent the covariance of the standardized data via its spectral decomposition, then k-th Principle Component can be computed as follows:

$$Y_k = v_k^T Z = v_{k1} Z_1 + \ldots + v_{kp} Z_p$$

Here v_k represent's the k-th eigenvector. This approach allows us to reduce the dataset into principle components that will retain the information present in the original data. The proportion of variance retained, i.e. the ratio of sum of the first k eigenvalues over the sum of all eigenvalues can be represented as:

$$\frac{\lambda_1 + \ldots + \lambda_k}{\lambda_1 + \ldots + \lambda_p} >= \gamma$$

Here, γ is an arbitrary threshold set by the user, which can depend heavily on the specific application and type of problem. This criterion is used to ensure that a sufficient amount of information is retained while also reducing the dimensionality of the dataset. By selecting the appropriate number of principal components, the goal is to balance between information loss with dimensionality reduction. Through this approach, I reduced the dataset into 2-dimensional columns. I also conducted LDA again to see how the reduced dataset performs in terms of training and validation accuracy. A graphical view of the geometric perspective of LDA in regards to the Mahalonibis distance is provided as well.

3 Results and Analysis

This section implements the methodologies described above using the Milk Transportation data and the Abalone data.

3.1 One-Sample Inference

Using one-sample inference, I am interested in testing whether or not there is a difference between mean population Fuel, Repair, and Capital costs in regards to using with gasoline fuel trucks only. I have a sample of 36 observations in this dataset. Thus, my null hypothesis can be defined as: H_0: $\mu_1 = \mu_2 = \mu_3$ to be tested via Hotelling's T^2 at α level = 0.05. For gasoline trucks, the summary statistics for mean and covariance is defined as $\vec{x}$, S_1 as follows:

$$\vec{x} = \begin{bmatrix}12.218\\8.112\\9.590\end{bmatrix} \qquad S_1 = \begin{bmatrix}23.013 & 12.366 & 2.906\\12.366 & 17.544 & 4.773\\2.906 & 4.773 & 13.963\end{bmatrix}$$

Using the constrast matrix C as defined in the methods section, I can can create the linear combinations $Y_1 = \vec{X}_1 - \vec{X}_2$ and $Y_2 = \vec{X}_1 - \vec{X}_3$. Through this linear combinations, I computed the new mean vector and covariance matrix:

$$\vec{y} = \begin{bmatrix}4.106\\2.628\end{bmatrix} \qquad S_y = \begin{bmatrix}15.824 & 12.513\\12.513 & 31.163\end{bmatrix}$$

Note that my new hypothesis is now $H_0 : \mu_{Y_1} = \mu_{Y_2} = 0$. From the formula of Hotelling's $T^2 = n(\vec{Y} - \vec{\mu}_y)^\top S_y^{-1}(\vec{Y} - \vec{\mu}_y)$, we compare the test statistic value $T^2 = 39.003$ against the critical value using $\frac{(35)2}{34}F_{2,34}(1 - 0.05) = 6.74$. Since our computed test statistic T^2 is greater than the associated critical value, we reject the null hypothesis at α level 0.05 and conclude that there is indeed a difference between mean population Fuel, Repair and Capital Costs in the context of gasoline trucks. With this, the visualization below displays the 95% Confidence region using a mean centered ellipse.

95% Confidence Region (One Sample Inference)

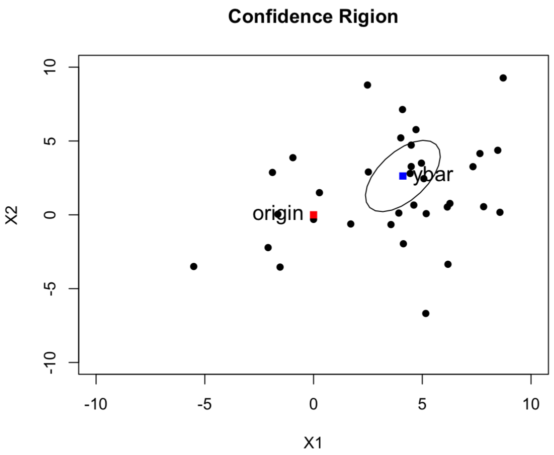


Figure 1: 95% Confidence Region Mean Centered Ellipse

Based the Mean-Centered Ellipse of our linearly transformed data, we can see that the origin falls outside 95% contour of the ellipse and is consistent with the statistical tests using Hotelling's T^2 that we should reject the NULL hypothesis. The 95% contour contains all mean difference vectors such that the corresponding Hotelling's T^2 fails, since the point δ= 0 falls outside the 95% contour, we should reject H0 at level $\alpha = 0.05$, which is consistent with the testing results as well.

I have displayed both statistical testing and graphic display for the one sample inference. From this testing procedure, we can deduce that there is indeed a difference in the mean Fuel, Repair, and Capital costs of gasoline transportation trucks. Now, I will move on to two-sample testing.

3.2 Two-Sample Testing

Furthermore, in addition to the data from gasoline trucks, we have also collected results from diesel trucks. This naturally leads to the question of whether there exists a difference in average Fuel, Repair, and Capital costs between the two types of fuels. We have analogous summary statistics for both gasoline and diesel trucks which includes fuel, repair, and capital costs. I've denoted the summary statistics as $\vec{\bar{X}}_1$, $\vec{\bar{X}}_2$, S_1 and S_2. Note that the summary statistics associated with $\vec{\bar{X}}_1$ and S_1 are pertaining to gasoline trucks and $\vec{\bar{X}}_2$ and S_2 are pertaining to diesel trucks.

$$\vec{\bar{X}}_1 = \begin{bmatrix} 12.218 \\ 8.112 \\ 9.590 \end{bmatrix} \quad S_1 = \begin{bmatrix} 23.013 & 12.366 & 2.90 \\ 12.366 & 17.544 & 4.773 \\ 2.906 & 4.773 & 13.963 \end{bmatrix} \quad \vec{\bar{X}}_2 = \begin{bmatrix} 10.105 \\ 10.762 \\ 18.167 \end{bmatrix} \quad S_2 = \begin{bmatrix} 4.362 & 0.759 & 2.362 \\ 0.759 & 25.851 & 7.685 \\ 2.362 & 7.685 & 46.654 \end{bmatrix}$$

Thus, I have the following hypothesis to test: $H_0 : \vec{\mu}_1 = \vec{\mu}_2$ versus $H_A : \vec{\mu}_1 \neq \vec{\mu}_2$. I can define $(\vec{X}_1 - \vec{X}_2)$ and S_{pooled}:

$$\vec{X}_1 - \vec{X}_2 = \begin{bmatrix} 2.11 \\ -2.64 \\ -8.57 \end{bmatrix} \quad S_{\text{pooled}} = \begin{bmatrix} 15.814 & 7.886 & 2.696 \\ 7.886 & 20.750 & 5.897 \\ 2.696 & 5.897 & 26.580 \end{bmatrix}$$

$T^2 = (\vec{X}_1 - \vec{X}_2)^T \left(\frac{1}{n_1} + \frac{1}{n_2}\right) S^{-1}_{\text{pooled}} (\vec{X}_1 - \vec{X}_2)$

$\frac{(n_1+n_2-2)p}{n_1+n_2-1-p} F_{p,n_1+n_2-1-p}$

From the definition of Hotelling's $T^2 = (\vec{X}_1 - \vec{X}_2)^T \left(\frac{1}{n_1} + \frac{1}{n_2}\right) S^{-1}_{\text{pooled}} (\vec{X}_1 - \vec{X}_2)$, I compared the value of the test statistic $T^2 = 50.91$ against the critical value $\frac{(n_1+n_2-2)p}{n_1+n_2-1-p} F_{p,n_1+n_2-1-p}(1-\alpha) = 8.620$. Since the computed test statistic is greater than the associated critical value at α level 0.05, I rejected the null hypothesis. This indicates strong evidence to support the conclusion that there does exist a significant difference between the mean fuel, repair, and capital costs of gasoline versus diesel trucks. In a practical sense, the two-sample testing provides evidence that there is indeed a disparity in the average costs between gasoline and diesel trucks and potentially underscores the importance of considering fuel type when assessing overall costs in the context of trucking operations.

3.3 Linear Discriminant Analysis

Using LDA, I want to predict whether or not Abalone can be classified as an adult or infant based on its physical measurements. In this case, the data includes 8 columns of different measurements. As defined above by the Fisher's Rule, I can classify Abalone as an "Adult" if:

$$\vec{w}^\top \vec{x}_0 \geq 5.698254$$

This threshold serves as a decision boundary value separating the two classes, with observations falling above it being classified as adults, and those below it as infants. To test the effectiveness of Linear Discriminant Analysis classifier, I tested the accuracy of our predictions on both training error and unseen testing error using Lachenbruch's Holdout (as defined in the methods section above). Since I have a binary classification problem, I use misclassification rate as the measure of accuracy. The results from applying this method on all 8 predictor columns provides promising results. The table below shows the error rates for both:

Table 3: Misclassification Rate using LDA on Abalone Dataset

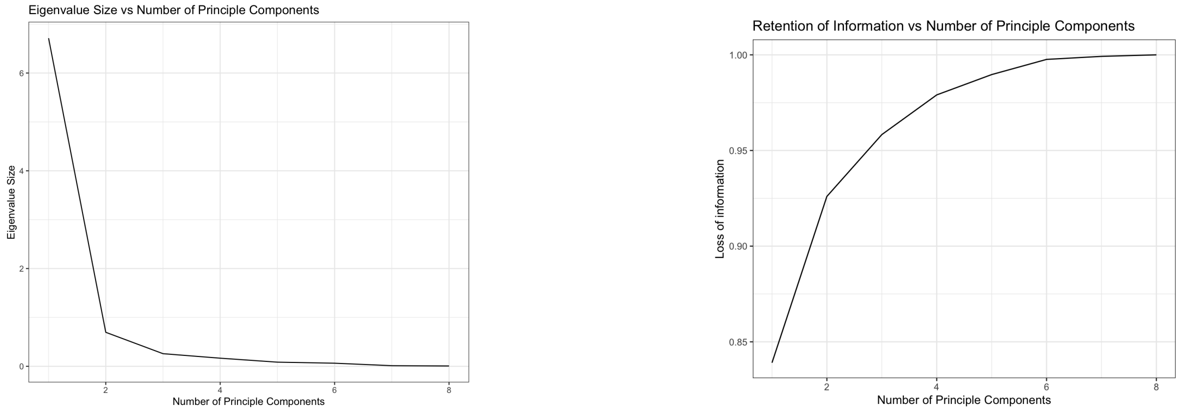
Training Error	0.2027771
Testing Error	0.2032559

Evidently, the results from LDA is quite promising on both training and testing data. The error rate on both training and testing data appears to be approximately equal at 20%, suggesting that the LDA classifier is quite accurate in predicting whether or not Abalone is an adult or infant based on it's physical measurements. The empirical analysis on the Abalone dataset demonstrates the effectiveness of the LDA classifier in generating accurate predictions. Note that, for this particular dataset, the testing error was very similar to the training error; however that may not generally be the case. Additional experiments using other datasets such as forest.fires.csv may highlight the discrepancy of the LDA classifier in performing well on both training and testing data. In this case, the model's ability to discern between adult and infant abalones based on their physical measurements holds promise for its practical application in real-world scenarios involving abalone population studies and management.

3.4 Principle Component Analysis

For conducting Principle Component Analysis (PCA), I first standardized the data to center the data with mean 0 and standard deviation 1. This decision was made as the 'Rings' column displayed a significantly higher mean and variance compared to the other 7 predictors. This disparity would likely have influenced the outcomes of PCA. The standardization ensures that all variables have comparable scales and variances to prevent any one variable from overtly influencing the analysis. The results from PCA are displayed below. The plot on the left highlights the eigenvalue size versus the number of principle components. The plot on the right highlights the proportion of variance retained of the original data as with regards to the number of principle components.

Principle Component Analysis on Abalone Data



(a) Eigenvalue Size vs Number of Principle Components

(b) Proportion of Variance Retained vs Number of Principle Components

Figure 2: Results of PCA on Abalone Data

From the plot on the left, the eigenvalue analysis indicates a significant drop-off in size after the first two principal components, with the eigenvalue associated with the first component at 6.7 and subsequent values below 0.5. This is supported by the plot on the right, which illustrates that the majority of information is preserved within the first two components. According to the PCA results, over 92% of the proportion of variance is retained by these initial two components. In the method's section, I denoted γ as an arbitrary threshold, which users can set as an accepted standard (varying based on context) for the overall proportion of variance retained. Suppose our threshold for this case is set to be greater than or equal to 90%. In that case, the first two principal components meet this criterion, effectively transforming the dataset into a 2-dimensional space. With this, a main benefit of PCA is to transform high-dimensional data into a 2-dimensional representation, helping with visualizations. The graph below facilitates the visualization of the Abalone dataset in a 2-dimensional space, showcasing the distinction between Adult and Infant Abalones based on their physical measurements.

Visualization of Abalone data using PCA

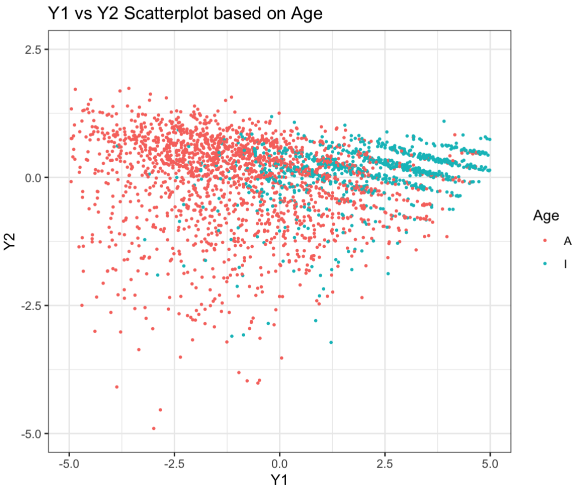


Figure 3: 2-dimensional visualization of Adult vs Infant Abalones

The initial dataset comprised of 8 different physical measurements. Using PCA, I reduced this dataset into a 2-dimensional space and visualized it using a scatterplot using the first two principle components. In this case, the first two principle components retain majority of the variance present in the original data. With this, we can see that there is fair difference between the clusters of adult abalones and infant abalones based on their physical measurements. This approach will help us gain geometric intuition of LDA in the section below.

It is important to note that the individual interpretations and contributions of the principle components gets quite complex due to the number of original columns, and is outside the scope of this project. In any case, the primary objective for using PCA on this paper was to gauge the effectiveness of LDA using the reduced data. For instance, I am interested in measuring this classifier's performance on the 2-dimensional data from above as compared to the original dataset. Thus, at this stage, I will proceed with running the LDA model again on the reduced dataset in order to the new accuracy rates on training and testing data depending on the number of principle components.

3.5 Linear Discriminant Analysis (on Reduced Dataset)

First, let's begin with a geometric interpretation on classification using LDA using on the 2-dimensional scatterplot above. Since we are in 2-dimensional space, we can show a visualization of how a new test point is classified based on the Mahalanobis distance. In particular, the plot below offers an intuitive view on LDA classification, providing valuable insights into how new data points are categorized based on their proximity to class centroids.

LDA on Reduced Abalone Dataset

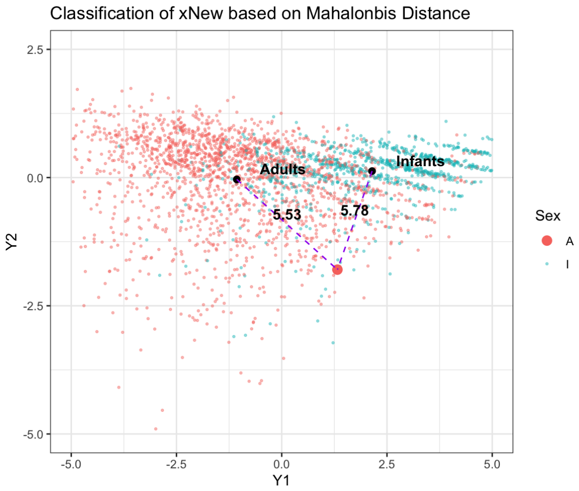


Figure 4: Mahalanobis Distance Geometric Intuition

Considering a new test data point, X_{new}, representing an "Adult" Abalone with coordinates (1.32, -1.79) in the two-dimensional space defined by the principal components. I calculated the Mahalanobis distance from X_{new} to the class centroids (means) of adults and infants. Upon analysis, I found that the distance from X_{new} to the centroid of Adult Abalones is shorter than its distance to the centroid of Infant Abalones. Consequently, I classified X_{new} as an Adult. This decision stems from the geometric intuition that X_{new} lies closer to the cluster of Adult Abalones in the reduced-dimensional space, suggesting a higher likelihood of belonging to that class.

As mentioned previously, I am interested in comparing the accuracy of the LDA classifier on the reduced dataset to understand the relationship between the proportion of variance retained and the the predictive power of this classifier. Thus, the table below displays the results for training and testing data in terms of accuracy and proportion retained in context with principle components.

Table 4: LDA Accuracy Rates on Reduced Dataset

Number of Principle Components	**Proportion of Variance Retained**	**Training Error**	**Testing Error**
1	0.839	0.2095241	0.2093478
2	0.926	0.2080441	0.2082835
3	0.958	0.2087623	0.2087623
4	0.979	0.2080441	0.2080441
5	0.989	0.2051712	0.2074567
6	0.997	0.2046924	0.2056799
7	0.999	0.2027771	0.2032559
8	1.000	0.2027771	0.2032559

The results displayed in the table highlight the relationship between the proportion of variance retained and the predictive power of the LDA classifier on the reduced dataset. As observed, as the number of principle components increases, there is a trend of decreasing training and testing errors. However, this reduction in error is rather marginal. Notably, using all 8 components (representing the entire dataset) yields the same error rates as running LDA on the original data. Evidently, the performance of this classifier is slightly better when more variance is retained in the reduced data. I must note that the difference in reduction of error for this dataset with the reduced dataset is marginal, empirical results using other data (such as forest fires with more columns) may yield more pronounced outcomes and provide a more nuanced understanding. For instance, with datasets featuring more complex relationships or higher dimensionality, the benefits of PCA in reducing overfitting and improving generalization could be more pronounced. For future studies, further empirical analysis across various datasets will likely provide additional insights into the relationship between the proportion of retained variance, dimensionality reduction, and predictive accuracy.

4 Discussion

Prior to the conclusion, it is important to address some potential areas for improvements in this project. For our hypothesis testing, I assumed that my sample of the Milk transportation data followed a multivariate normal distribution. To enhance the robustness of our test results, it would be critical to assess the normality of the data and consider conducting transformations as an initial step for a more refined analysis. With regards to LDA, it will be interesting to compare the performance of other machine learning models (Logistic Regression, SVM, KNN, etc.) in predicting Abalone Age category. For PCA, to aid interpretations of the individual principle components, finding the determination of each variate associated with the first two principle components would have helped as a part of more detailed analysis.

In summary, this paper offers a thorough analysis of two datasets using several multivariate statistical methodologies, including one-sample inference, two-sample testing, Linear Discriminant Analysis (LDA), and Principal Component Analysis (PCA). For the Milk Transportation Costs dataset, one-sample inference was initially conducted to examine differences in mean Fuel, Repair, and Capital costs within gasoline trucks. The analysis revealed that we rejected the null hypothesis to conclude that there are indeed significant differences in these costs. Subsequently, two-sample testing was performed to compare the mean costs between gasoline and diesel trucks, which also indicated a significant disparity. Moving on to the Abalone dataset, LDA was employed to predict the age category (Adult vs. Infant) based on physical measurements. The results showed high accuracy rates on both training and testing data, suggesting the effectiveness of LDA in classifying Abalones based on their physical attributes. Furthermore, PCA was utilized to reduce the dimensionality of the Abalone dataset while retaining majority of the proportion of variance. The LDA classifier was then trained again on the reduced dataset, resulting in comparable accuracy rates to the original dataset. This showed the effectiveness of dimensionality reduction through PCA without significant loss of predictive power. Ultimately, this paper demonstrates the practical application of multivariate statistical techniques in analyzing real-world datasets, providing valuable insights into transportation costs and Abalone classification based on physical characteristics.

5 Code Appendix

Listing 1: R Code Appendix

```
### ------------------------- STA 135: Final Project Code ---------------------------
### -------------------------------------------------------------------------------
### Milk Transportation Costs Data
### Methods:
### One Sample Test (on gasoline)
### Two Sample Test (gasoline vs diesel sample)

# data preparation
milk_data <- read.table("~/Desktop/final_project/T6-10.dat", header = FALSE, sep = ",")
colnames(milk_data) <- c('Fuel', 'Repair', 'Capital', 'Type') # column types
milk_data[] <- lapply(milk_data, trimws) # trim white spaces
milk_data[,-4] <- lapply(milk_data[,-4], as.numeric) # convert to numeric

# split for gasoline and diesel samples
gasoline_costs <- milk_data[1:36,-4]
diesel_costs <- milk_data[37:59,-4]

###. ------------------------------ One sample test ---------------------------------
# get sample statistics
x_bar <- colMeans(gasoline_costs)
Sx <- cov(gasoline_costs); Sx; dim(Sx)
n = 36
p = 3
alpha = 0.05

# linear transformation constract matrix C
C <- matrix(c(1,-1,0,
               1,0,-1), nrow = 2 , ncol = 3, byrow = TRUE); C
q = 2 # new dimension for critical value

# new summary statistics
Sy <-  C %*% Sx %*% t(C); Sy
y_bar <- C %*% x_bar; y_bar

# test statistic
t = n * t(y_bar) %*% solve(Sy) %*% y_bar

# critical value
test_crit = (n-1)*q/(n-q) * qf(1-alpha, q, n-q)

t > test_crit # decision rule: true, reject H0

# plot mean centered-ellipse
library(ellipse)
X <- cbind(gasoline_costs[,1] - gasoline_costs[,2],
           gasoline_costs[,1] - gasoline_costs[,3])
sd <- sqrt(c(Sy[1,1], Sy[2,2]))
corr<- Sy[1,2]/sqrt(Sy[1,1]*Sy[2,2])
xbar = (1/n)*t(X)%*%rep(1,n)
```

```
plot(X, pch = 16, xlim=c(-10,10), ylim=c(-10,10),
     main="Confidence Rigion", xlab="X1", ylab="X2")
lines(ellipse(corr, scale = sd, centre = xbar,
                t = sqrt(test_crit/n), npoints = 25))
points(y_bar[1],y_bar[2], pch=15, col="blue")
text(y_bar[1], y_bar[2], "ybar", cex=1.3, pos=4)
points(0, 0, pch=15, col="red")
text(0, 0, "origin", cex=1.3, pos=2)


### ------------------------------ Two Sample test ------------------------------
# summary statistics for both samples
x_bar_1 <- colMeans(gasoline_costs)
S1 <- cov(gasoline_costs); S1; dim(S1)
x_bar_2 <- colMeans(diesel_costs)
S2 <- cov(diesel_costs); S2; dim(S2)
n <- c(36,23)
p = 3

# get combined result S_pooled and D
d = x_bar_1 - x_bar_2
Sp<-((n[1]-1)*S1+(n[2]-1)*S2)/(n[1] + n[2] -2)

# critical value
alpha = 0.05
test_crit = (n[1] + n[2] -2)*p/(n[1] + n[2] -p-1)*
  qf(1-alpha,p,n[1] + n[2] -p-1); test_crit

# Hotelling's T-square
Sp_norm <- (1/n[1] + 1/n[2])*Sp
T_square <- t(d)%*%solve(Sp_norm)%*%d; T_square

T_square > test_crit # decision rule: reject H0

### -----------------------------------------------------------------------------
### Abalone Data
### Methods:
### LDA for classification
### PCA for dimensionality reduc.

# data-prep
abalone_df<- read.table("~/Desktop/final_project/abalone.data", header = FALSE, sep = ",
colnames(abalone_df) <- c('Sex', 'Length', 'Diameter', 'Height', 'Whole_Weight',
                            'Shucked_Weight', 'Viscera_weight','Shell_weight',
                            'Rings')

### ----------------------------------- LDA -----------------------------------
## predict gender: adults vs. infants
abalone_df$Sex <- ifelse(abalone_df$Sex %in% c("M", "F"), "A", abalone_df$Sex)

### summary values for each group
adult_data <- abalone_df[abalone_df$Sex == "A",-1]
```

```
infant_data <- abalone_df[abalone_df$Sex == "I",-1]
n <- c(nrow(adult_data), nrow(infant_data))

g1_bar <- colMeans(adult_data)
S1 <- cov(adult_data)
g2_bar <- colMeans(infant_data)
S2 <- cov(infant_data)
S_pooled <- ((n[1]-1)*S1+(n[2]-1)*S2)/(n[1] + n[2] -2)

# LDA decision boundary
w <- solve(S_pooled) %*% (g1_bar - g2_bar)
boundary = 0.5 * (t(w) %*% (g1_bar + g2_bar))

### Apparent error rate (training error)
preds <- c()
for (i in 1:nrow(abalone_df)) {
  row <- abalone_df[i,-1]
  val <- t(w) %*% t(as.matrix(row))
  if (val >= boundary[1,1]) {
    preds <- c(preds, "A")
  } else {
    preds <- c(preds,"I")
  }
}
sum(preds != abalone_df[,1])/nrow(abalone_df) # misclassified samples: 0.2027771

### Test Error: leave one out C.V using Lachenbruch's Holdout
preds <- c()
for (i in 1:nrow(abalone_df)) {
  print(i)
  trn_data <- abalone_df[-i,]
  tst_data <- abalone_df[i,-1]

  adult_data <- trn_data[trn_data$Sex == "A",-1]
  infant_data <- trn_data[trn_data$Sex == "I",-1]
  n <- c(nrow(adult_data), nrow(infant_data))

  g1_bar <- colMeans(adult_data)
  S1 <- cov(adult_data)
  g2_bar <- colMeans(infant_data)
  S2 <- cov(infant_data)
  S_pooled <- ((n[1]-1)*S1+(n[2]-1)*S2)/(n[1] + n[2] -2)

  w <- solve(S_pooled) %*% (g1_bar - g2_bar)
  boundary = 0.5 * (t(w) %*% (g1_bar + g2_bar))

  val <- t(w) %*% t(as.matrix(tst_data))
  if (val >= boundary[1,1]) {
    preds <- c(preds, "A")
  } else {
    preds <- c(preds,"I")
  }
```

```
}; preds

sum(preds != abalone_df[,1])/nrow(abalone_df) # misclassified samples: 0.2032559

###-------------------------------------- PCA ----------------------------------------
# function does PCA as by number of components
performPCA <- function(data, num_components = 1, scaleData = TRUE) {
  if (scaleData) {

    ### using correlation matrix R
    data <- scale(data, center = TRUE, scale = TRUE)
    covX <- cov(data)
  } else {

    ### using covariance matrix S
    covX <- cov(data)
  }

  # do spectral decomposition
  eigenDecomp <- eigen(covX)
  sorted_indices <- order(eigenDecomp$values, decreasing = TRUE)
  eigenvalues <- eigenDecomp$values[sorted_indices]
  eigenvectors <- eigenDecomp$vectors[, sorted_indices]

  # compute principle components
  pComponents = as.matrix(data) %*% as.matrix(eigenvectors)
  pca_scores <- pComponents[,1:num_components]
  return(list(scores = pca_scores,
              allEigen = eigenvalues,
              eigenvalues = eigenvalues[1:num_components],
              eigenvectors = eigenvectors[,1:num_components]))
}

# loss of information with regards to principle components
loss <- c()
eign <- c()
for (i in 8:1) {
  pca_result <- performPCA(abalone_df[,-1],i, scaleData = T)
  pca_scores <- pca_result$scores
  pca_eigen <- pca_result$eigenvalues
  pca_all <- pca_result$allEigen
  l = sum(pca_eigen)/sum(pca_all)
  loss = c(loss,l)
  eign <- pca_all
}; loss; eign

# plot the loss
dd <- data.frame(x = 8:1, y = loss)
library(ggplot2)
ggplot(dd, aes(x = x, y = y)) +
  geom_line() +
  labs(x = "Number␣of␣Principle␣Components",
         y = "Proportion of Variance Retained") +
  ggtitle("Retention of Information vs Number of Principle Components") +
  theme_bw()

# plot the eigenvalue vs number of principle components
dd <- data.frame(x = 1:8, y = eign)
ggplot(dd, aes(x = x, y = y)) +
  geom_line() +
  labs(x = "Number of Principle Components",
         y = "Eigenvalue Size") +
  ggtitle("Eigenvalue Size vs Number of Principle Components") +
  theme_bw()

# Do LDA based on n principle components - same as top
comp <- 5
pca_result_2 <- performPCA(abalone_df[,-1],comp)
pca_scores <- pca_result_2$scores
pca_eigen <- pca_result_2$eigenvalues
pca_all <- pca_result_2$allEigen
l = sum(pca_eigen)/sum(pca_all) # loss 0.92

newAba_df <- as.data.frame(cbind(abalone_df[,1], pca_scores))
colnames(newAba_df) <- c("Age", "Y1", "Y2","Y3")
newAba_df[,-1] <- lapply(newAba_df[,-1], as.numeric)

### summary values for each group
adult_data <- newAba_df[newAba_df$Age == "A",-1]
infant_data <- newAba_df[newAba_df$Age == "I",-1]
n <- c(nrow(adult_data), nrow(infant_data))

g1_bar <- colMeans(adult_data)
S1 <- cov(adult_data)
g2_bar <- colMeans(infant_data)
S2 <- cov(infant_data)
S_pooled <- ((n[1]-1)*S1+(n[2]-1)*S2)/(n[1] + n[2] -2)

# LDA decision boundary
w <- solve(S_pooled) %*% (g1_bar - g2_bar)
boundary = 0.5 * (t(w) %*% (g1_bar + g2_bar))

### Apparent error rate (training error)
preds <- c()
for (i in 1:nrow(newAba_df)) {
  row <- newAba_df[i,-1]
  val <- t(w) %*% t(as.matrix(row))
  if (val >= boundary[1,1]) {
    preds <- c(preds, "A")
  } else {
    preds <- c(preds,"I")
  }
}
sum(preds != newAba_df[,1])/nrow(newAba_df) # misclassified samples: 0.2080441
```

```
### Test Error: leave one out C.V using Lachenbruch's Holdout
preds <- c()
for (i in 1:nrow(newAba_df)) {
  print(i)
  trn_data <- newAba_df[-i,]
  tst_data <- newAba_df[i,-1]

  adult_data <- trn_data[trn_data$Age == ”A”,-1]
  infant_data <- trn_data[trn_data$Age == ”I”,-1]
  n <- c(nrow(adult_data), nrow(infant_data))

  g1_bar <- colMeans(adult_data)
  S1 <- cov(adult_data)
  g2_bar <- colMeans(infant_data)
  S2 <- cov(infant_data)
  S_pooled <- ((n[1]-1)*S1+(n[2]-1)*S2)/(n[1] + n[2] -2)

  w <- solve(S_pooled) %*% (g1_bar - g2_bar)
  boundary = 0.5 * (t(w) %*% (g1_bar + g2_bar))

  val <- t(w) %*% t(as.matrix(tst_data))
  if (val >= boundary[1,1]) {
    preds <- c(preds, ”A”)
  } else {
    preds <- c(preds,”I”)
  }
}; preds

sum(preds != newAba_df[,1])/nrow(newAba_df) # misclassified samples: 0.2082835

### plots
# plot principle components
library(ggplot2)
ggplot(newAba_df, aes(x = Y1, y = Y2, color = Age)) +
  geom_point(size = 0.5) +
  labs(x = ”Y1”, y = ”Y2”, color = ”Age”) +
  ggtitle(”Y1␣vs␣Y2␣Scatterplot␣based␣on␣Age”) +
  theme_bw() + ylim(-5,2.5) + xlim(-5,5)

## visualize geometric interpretation of LDA with Y1 and Y2
mean_values <- aggregate(cbind(Y1, Y2) ~ Age, data = newAba_df, FUN = mean)
x_o <- newAba_df[1,]
a1 = as.matrix((x_o[,-1] - g1_bar)) %*%
  solve(S_pooled) %*% t(as.matrix((x_o[,-1] - g1_bar)))
a2 = as.matrix((x_o[,-1] - g2_bar)) %*%
  solve(S_pooled) %*% t(as.matrix((x_o[,-1] - g2_bar)))
if (a1 < a2) { # for user intuition
  print(”A”)
} else {
  print(”I”)
}
```

```
mean_values$distance <- c(round(a1,2), round(a2,2))

ggplot(newAba_df, aes(x = Y1, y = Y2, color = Age)) +
  geom_point(size = 0.5, alpha = 0.4) +
  geom_point(data = mean_values, aes(x = Y1, y = Y2),
             color = "black", size = 2) +
  geom_text(data = mean_values[1, ], aes(label = paste("Adults")),
            vjust = -0.5, hjust = -0.5, color = "black", fontface = "bold") +
  geom_text(data = mean_values[2, ], aes(label = paste("Infants")),
            vjust = -0.5, hjust = -0.5, color = "black", fontface = "bold") +
  geom_point(data = data.frame(x_o), aes(x = Y1, y = Y2, color = Age), size = 3) +
  geom_segment(data = mean_values,
               aes(x = Y1, y = Y2, xend = x_o$Y1, yend = x_o$Y2),
               color = "purple", linetype = "dashed") +
  geom_text(data = mean_values,
            aes(x = (Y1 + x_o$Y1) / 2, y = (Y2 + x_o$Y2) / 2,
                                    label = paste(distance)),
            color = "black", hjust = 0.5, vjust = -0.5, fontface = "bold") +
  labs(x = "Y1", y = "Y2", color = "Age␣Group") +
  ggtitle("Classification␣of␣xNew␣based␣on␣Mahalonbis␣Distance") +
  theme_bw() + ylim(-5,2.5) + xlim(-5, 5)
```