Multivariate Analysis of Olive Oil Compositions from Three Macro Areas of Italy

Introduction

The purpose of our project is to study and analyze the chemical composition of olive oil produced in different regions of Italy and test whether the chemical composition can be used to identify unique profiles of each region, both visually and mathematically.

The data set we use consists of 572 olive oil samples from three macro-regions of Italy, and each sample has eight chemical composition readings (Ref 1.). The eight chemical components are all fatty acids, including palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic and eicosenoic acids, and are measured in "parts per 10,000."

The eight fatty acids have different chain lengths in their molecules (number of carbons) and different degree of unsaturation (number of unsaturated bonds). Based on the degree of the unsaturation, we can classify the fatty acids into three types (as shown in Table 1). In this project, we studied the variation of the chemicals compositions among the regions and used the chemical measurements to identify the place of production of the olive oil samples. By exploring the acid content of olive oil, we can better understand health impacts of different oils.

Table 1: Fatty Acid Type

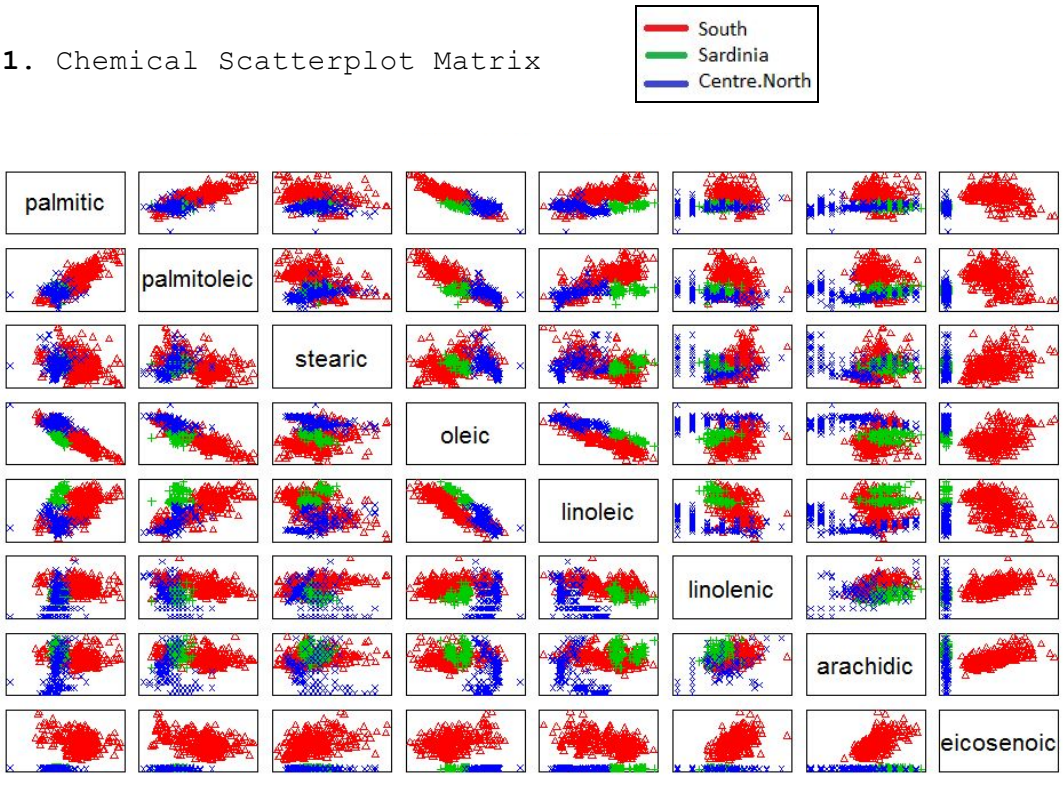
	Acid Name (# of carbon: # unsaturated bond)
Saturated acids	Palmitic (C16:0)
	Stearic (C18:0)
	Arachidic (C20:0)
Monounsaturated acids	Palmitoleic (C16:1)
	Oleic (C18:1)
	Eicosenoic (C 20:1)
Polyunsaturated acids	linoleic acid (C18:2)

	Linolenic acid (C18:3)

Methods

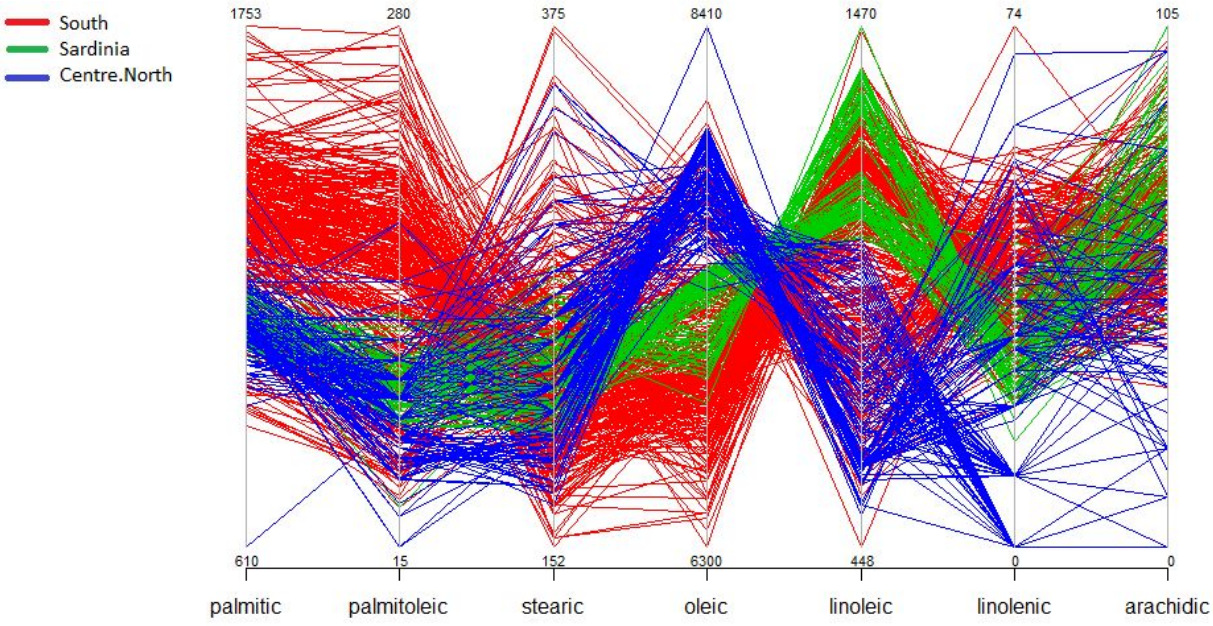
Exploratory plots were created to obtain better understanding on the data. First, we created a scatterplot matrix of all eight chemicals to look at correlations and groupings (Fig. 1). Strong correlations between palmitic, palmitoleic, oleic acids were observed. The chemical distributions are quite different among regions. In addition, we observed that eicosenoic acid, the 8th chemical, had some strange patterns. With a closer examination on the data, it is found that the measurements of eicosenoic at macro region 2 and 3 are of levels at 1, 2 and 3, which are much smaller than those for the macro region 1 (~10-40) (Please Ref Fig A1. in the appendix). The measurements are questionable. Thus the measurements on eicosenoic will be not included in all the following data analysis in this project.

Figure 1. Chemical Scatterplot Matrix



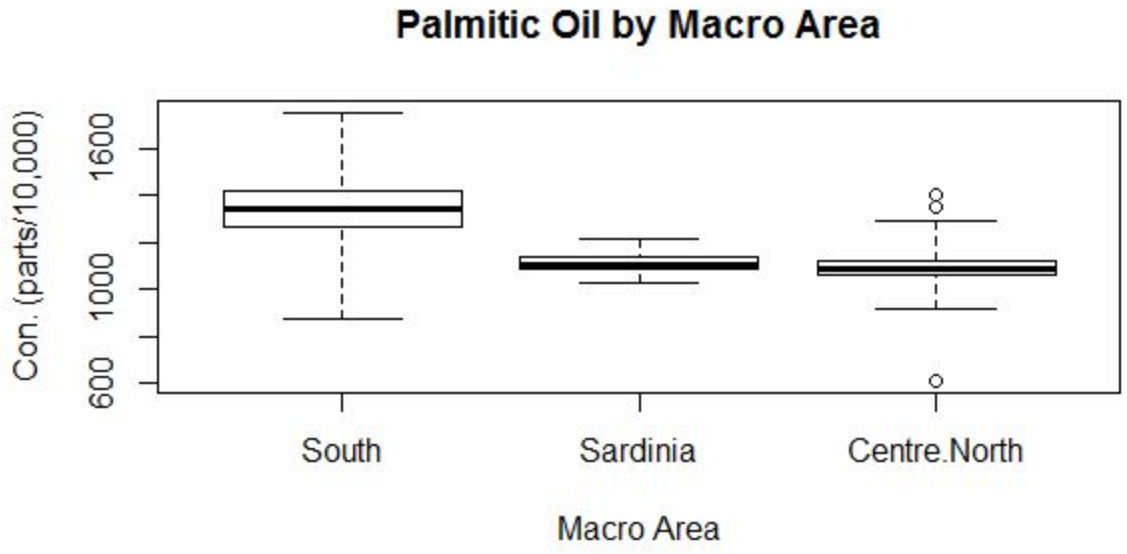
A parallel coordinate plot was plotted (Fig. 2)and was used to examine the variations of the chemicals among the macro areas. To facilitate the comparison, the data points for each chemicals were rescaled. For this plot, we observed that higher levels of oleic is associated with lower levels of palmitic and linoleic, indicating negative association between them. It was also observed that the concentrations of palmitic acid, palmitoleic and stearic acid are quite same for macro area 2 and 3. These observations will be compared with formal tests later on.

Figure 2: Parallel Coordinates Plot by Macro Area



Box plots, such as the one in Fig. 3, were used to compare chemical concentrations for each region and identify outliers. Here, we see an outlier in Centre North for Palmitic Oil (minimum value), which we removed for our future analyses. In all, we removed 3 outliers from the dataset, but further exploration could suggest more of them.

Figure 3: Boxplot of Palmitic Oil by macro area



The following sections will discuss the individual approaches used to analyze this dataset, specifically MANOVA, pairwise comparisons, principal components analysis (PCA), factor analysis (FA), and discriminant analysis (LDA,QDA).

MANOVA

To obtain a general idea of the data, the mean concentrations of the seven chemicals are listed in Table 2. Oleic is by far the most abundant chemical. Palmitic and linoleic are found at similar levels, ~ 1000.

Table 2. Mean values of the seven fatty acids of each macro area.

	palmitic	palmitoleic	stearic	oleic	linoleic	linolenic	arachidic
South	1332	155	229	7100	1034	38	63
Sardinia	1111	97	226	7269	1197	27	73
Center North	1098	84	231	7789	728	22	38

The mean vectors of the chemical compositions from three macro areas were then compared using MANOVA. The three assumptions for MANOVA analysis were examined. First, for the independency, these samples were sampled from different macro areas and are reasonable to be assumed as independent samples. Second, the normality of the chemical measurements of each region was examined using Q-Q plots. As shown in Figure A2 (q-q plots in appendix), most of the chemicals are approximately normal. In addition, the sample sizes for the three macro regions are large, 322, 97 and 150, respectively. These relatively large sample sizes also help on the normality of data. Last, the equivalence of the covariance between groups was examined. The covariance of each chemical (diagonal elements of the covariance matrix) for each macro region were plotted (Figure 4). The variance between region 2 and 3 are similar but the variance of region 1 is much larger than that of 2 and 3. The correlation matrix of the three regions are similar (Figure 5). It has been suggested that when the larger samples also have larger variance than the MANOVA test tends to be robust for type I errors (with a loss in power) (Ref 2.) Thus, the MANOVA test herein should have good control of type I errors.

Figure 4. The variances of each chemicals for the three macro regions.

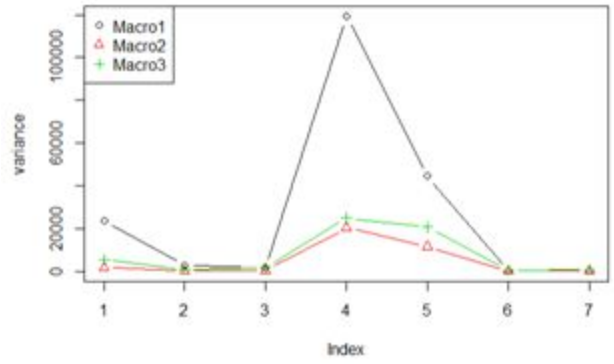
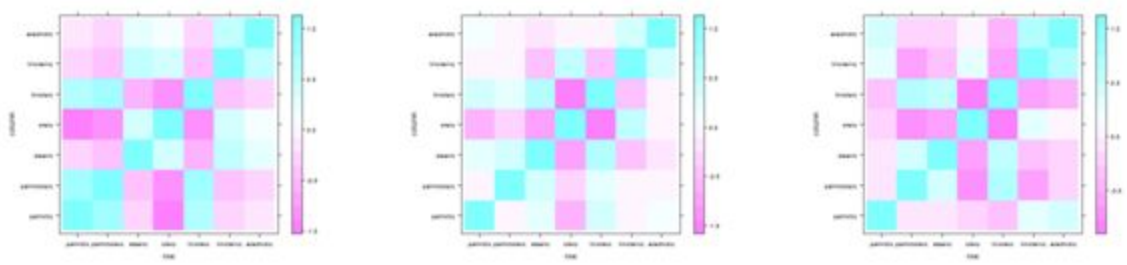


Figure 5. Correlation matrix for: macro areas 1 to 3 (from left to right).



The data was fit using MANOVA, and test statistic is 0.082, with p-value< 2.2 e-16 (Table 4). As discussed earlier, the larger sample size (region 1) have larger variation, the type I error is under good control. Thus, the fitting results show strong evidence that the mean vectors of the chemical compositions of the three regions are different. Pairwise comparisons on each of the chemicals from different regions were examined to explore the variations of the chemicals among regions. The test statistics (T) are listed in Table 3 (based on the sample dimension, the critical t-value was calculated as 3.052). For macro regions 1 vs. 2 and 1 vs. 3, all fatty acids (except stearic) show significant differences, with |T| greater than the critical value. For macro regions 2 vs. 3, oleic, linoleic, linolenic and arachidic show significant differences, while palmitic palmitoleic and stearic are of the same level, which is consistent with the previous exploratory analysis.

Table 3. Values of the t test statistic of for the pair-wise comparisons of chemicals from different regions. Critical value: $t_{0.05/(2*21),\ df\ =\ 562} = 3.052$.

	palmitic	palmitoleic	stearic	oleic	linoleic	linolenic	arachidic
1 vs. 2	15.63	12.17	0.58	-5.24	-7.82	8.74	-4.98
2 vs. 3	0.81	2.56	-0.99	-14.33	19.97	3.77	15.30
1 vs. 3	19.39	17.63	-0.62	-25.02	17.17	15.21	14.34

Table 4. Results of MANOVA fitting.

	Df	Wilks	approx F	num Df	den Df	Pr(>F)
regionf	2	0.081994	199.38	14	1120	< 2.2e-16 ***

Principle Component Analysis (PCA)

Principle component analysis (PCA) was conducted to examine the potential data reduction of the seven chemical measurements. We did a close examination on the scatter plot in Figure 1. The scatter plot

show the chemicals have some different degrees and directions of correlations. In one group (Group 1), palmitic, palmitoleic,linoleic linolenic arachidic, showed positive correlation with each other. This group of acids shows negative correlation with stearic and oleic acids(group 2). Since all these compounds add to 10,000, it is likely, the first group of compounds change with the same direction, and opposite of the second two acids.

PCA analysis was performed on the seven chemical measurements. The proportions of variance for each of the seven principal components and the cumulative proportion of variance were listed in Table 5 and plotted in Figure 6. The first four principal components (PC) already explained ~ 95% of the variation in the data. The four principle components are listed in Table 6. As can be seen, PC1 is actually the contrast of the two groups of fatty acids we observed in the scatter plot. The last two acids, linolenic and arachidic, have most contribution of PC2. However, we cannot explain further on the PCs due to the complexity of the data.

Table 5. Proportion of variance for each principal component.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	1.8551	1.2415	0.9745	0.8293	0.5080	0.3452	0.0517
Proportion of Variance	0.4916	0.2202	0.1356	0.0983	0.0369	0.0170	0.0004
Cumulative Proportion	0.4916	0.7118	0.8475	0.9457	0.9826	0.9996	1.0000

Figure 6: The scree plot for the cumulative proportion of variance.

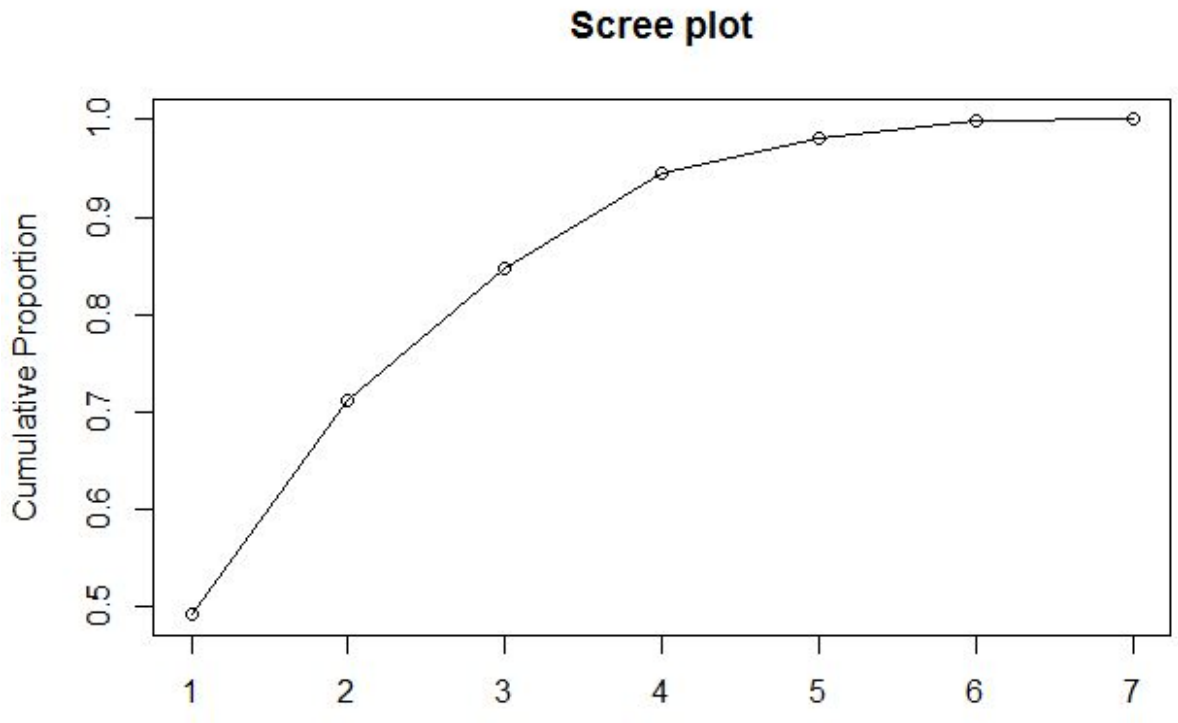
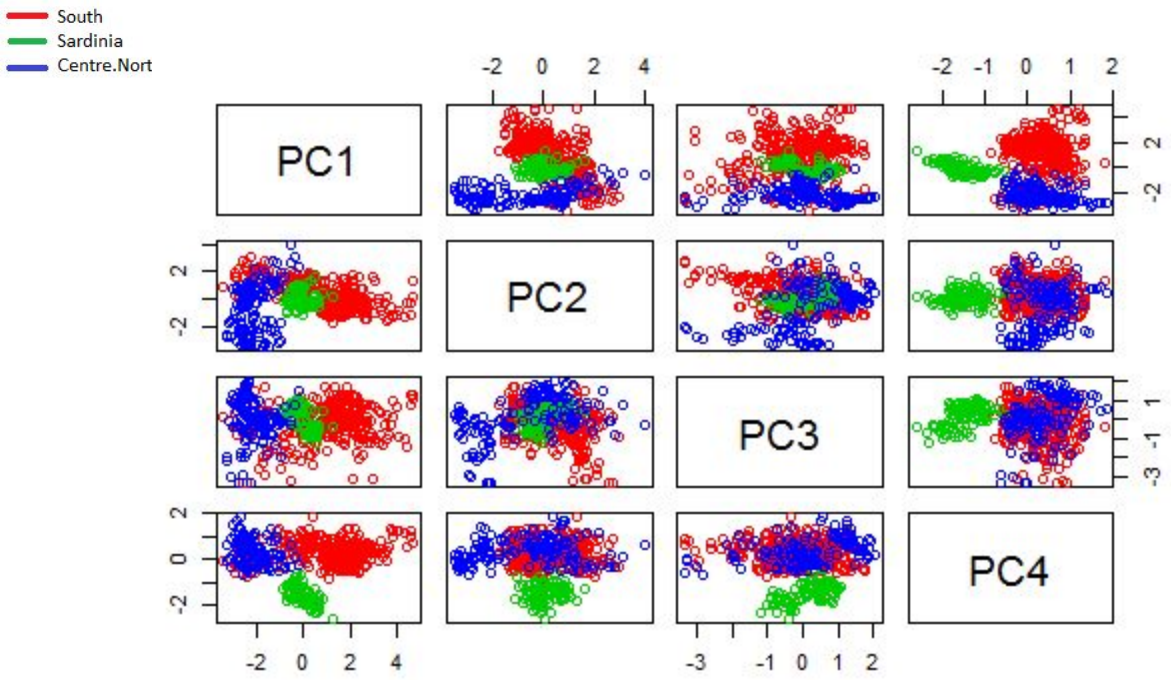


Table 6. First 4 Principal Components.

	PC1	PC2	PC3	PC4
palmitic	0.47	0.00	-0.07	0.48
palmitoleic	0.48	-0.20	-0.06	0.28
stearic	-0.14	0.14	-0.97	-0.04
oleic	-0.53	0.07	0.16	0.12
linoleic	0.42	-0.22	-0.04	-0.63
linolenic	0.17	0.70	0.06	0.28
arachidic	0.21	0.63	0.10	-0.45

We plot the first 3 principal components against each other in Figure 7. We can see that PC1 and PC2 help to distinguish between Centre North (blue) and the other regions. PC 4 is very good at identifying observations from Sardinia (green), and the combination of PC 1 with PC4 seems to be very good at separating the macro areas. The other P combinations do not have as clear of an effect, but they are still useful.

Figure 7: Comparison of first 4 principal components



Factor Analysis (FA)

Factor analysis was conducted to examine whether some latent factors can be identified to describe the data. According to previous PCA analysis, four factors were chosen for the analysis.

Table 7 shows the loading coefficients for the four factors after Varimax rotation. Factor 1 shows that chemicals palmitic, palmitoleic and oleic are strongly correlated. Oleic is highly negatively correlated with the other two chemicals in this group, which is consistent with observations in the scatter plots and PCA analysis. Factor 2 can represent the last two acids, which are also of the least abundant components among the seven chemicals. Factor 3 can represent chemical stearic since it has a loading coefficient -0.993, a lot higher than any other chemicals. It is worthing noting that from pairwise comparison, it was observed that stearic is the only chemical that of the same level for all three group. Factor 4 can represent chemical linoleic since it has a loading coefficient 0.888, a lot higher than any other chemicals. From Table 8, we can conclude that the four factors together explain 94.5% of the variation in our data.

Table 7. Loading coefficients for the four factors after Varimax rotation.

```
Loadings:
            F1     F2     F3     F4
palmitic    -0.943  0.190
palmitoleic -0.905         0.118  0.298
stearic                   -0.993
oleic        0.785 -0.183        -0.586
linoleic    -0.413                0.882
linolenic   -0.233  0.883        -0.251
arachidic           0.903         0.309
```

Table 8. Proportion of variance for each factor and the cumulative proportion of variance.

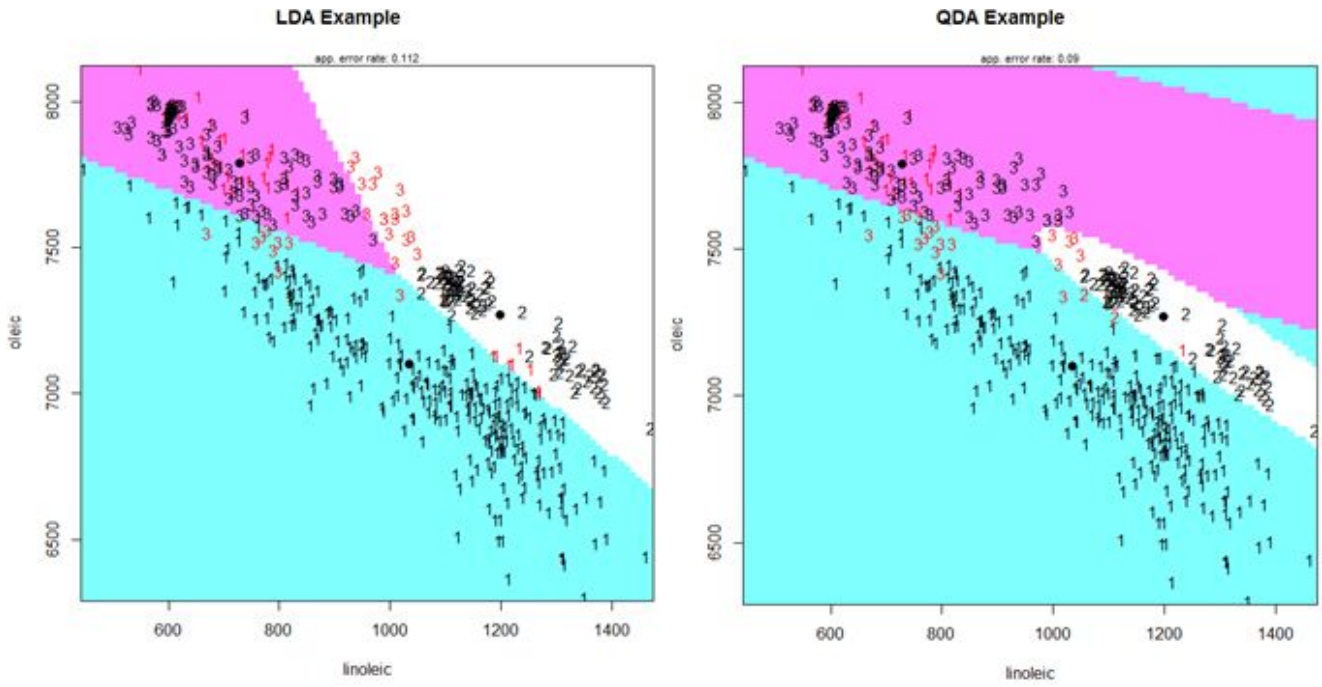
```
                  F1    F2    F3    F4
SS loadings    2.558 1.666 1.015 1.381
Proportion Var 0.365 0.238 0.145 0.197
Cumulative Var 0.365 0.603 0.748 0.946
```

Discriminant Analysis

We applied Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) to see which combinations of chemicals could help us predict the macro area that an observation came from. LDA and QDA both assume normality of the data, and LDA assumes equal covariance between groups (macro areas), which was addressed earlier and found to be acceptable.

The following plots show a sample of how the LDA and QDA classifiers categorized macro area (with 1 = South, 2 = Sardinia, 3 = Centre North)

Figure 8. LDA and QDA for oleic and linoleic acid. Note the improvement by using QDA.



The confusion matrices for LDA and QDA show that both methods work very well on classifying this dataset, but the QDA shows marginal improvements.

Table 8: LDA Confusion Matrix (APER = 0.03866432)

	Predicted		
Actual	South	Sardinia	Centre.North
South	311	0	11
Sardinia	3	94	0
Centre.North	8	0	142

Table 9: QDA Confusion Matrix (APER = 0.01581722)

	Predicted		
Actual	South	Sardinia	Centre.North
South	315	0	7
Sardinia	1	96	0
Centre.North	1	0	149

The LDA results also included 2 linear discriminants, allowing for some general interpretation of which chemicals contributed to the classification.

Table 10: Coefficients of linear discriminants:

	LD1	LD2
palmitic	0.022393057	-0.0009854534
palmitoleic	0.008389556	0.0139767576
stearic	0.019025852	-0.0044704194
oleic	0.022305847	-0.0038759415
linoleic	0.019303739	-0.0106917949
linolenic	-0.040871299	0.0205372197
arachidic	0.018437768	-0.0461095673

The first set of coefficients suggest a contrast in direction between linolenic acid and the others, a lower weight for palmitoleic oil, and equal positive weighting of the others. Although the QDA gave more accurate predictions, it is harder to interpret which combinations of chemicals are improving the APER since no coefficients are provided.

Conclusion

A data set on the concentrations of seven fatty acids in olive oil from three macro areas of Italy was analyzed. MANOVA was used to compare vector means from the three macro areas (South, Sardinia, and Centre North), and a significant difference among the areas was observed. Detailed pairwise comparisons on each fatty acids from different macro areas were conducted. PCA and FA analyses were performed to help understand the variations of the data. In both analyses, the contrast between two groups of fatty acids, one group including palmitic, palmitoleic, linoleic, linolenic, arachidic, and other group including stearic and oleic acids or oleic acid only (for FA), was observed. LDA and QDA were exploited to classify the olive oil samples to the three macro areas based on the fatty acids measurements. Both methods obtained APER < 0.05, indicating proper classifications. Further work on this data may include studying the subregions of oil origin within each macro area, to more precisely determine where olive oil comes from. This may help in industries such as marketing, i.e. “The coastal Sardinia olive oil has less saturated fats than others.”

References:

1. Data obtained from CRAN pdfCluster package https://cran.r-project.org/web/packages/pdfCluster/
2. http://www.real-statistics.com/multivariate-statistics/multivariate-analysis-of-variance-manova /manova-assumptions/

Appendix

Figure A1. Plot of eicosenoic responses vs. obs from three macro regions.

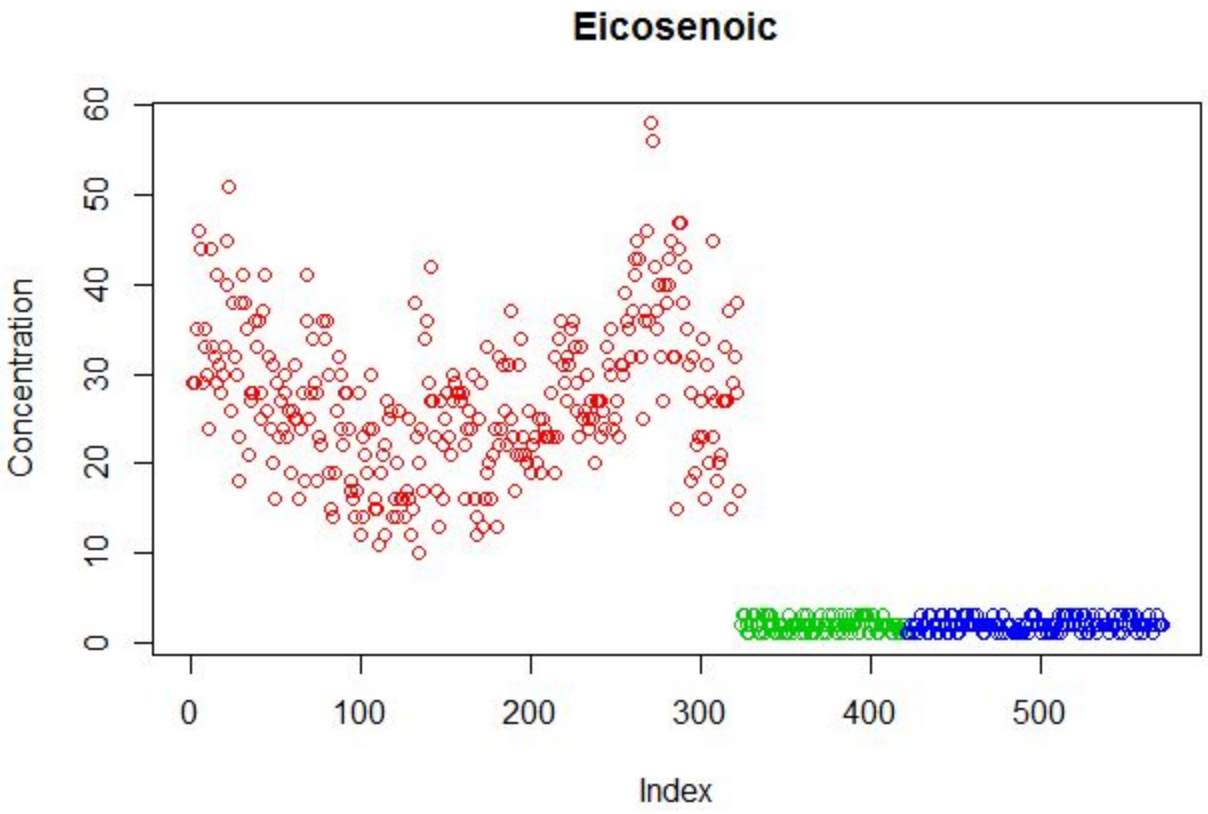
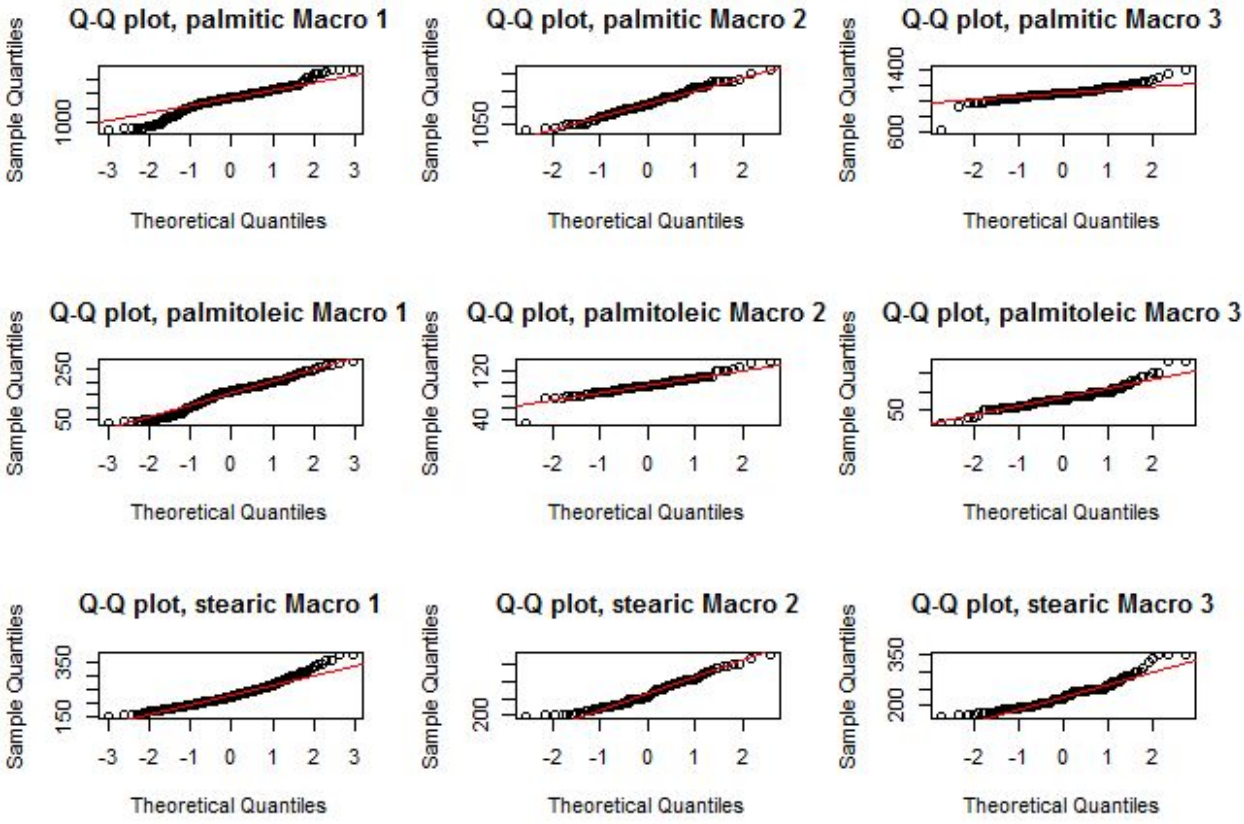
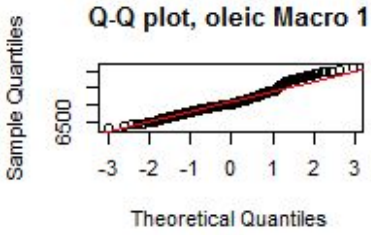
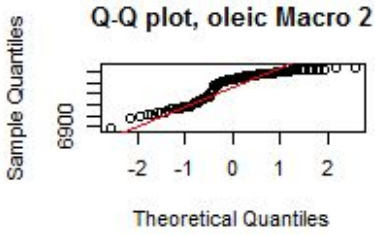
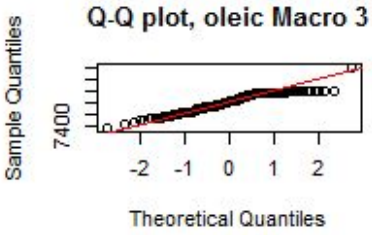


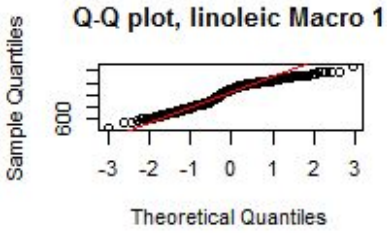
Figure A2. Q-Q plots.

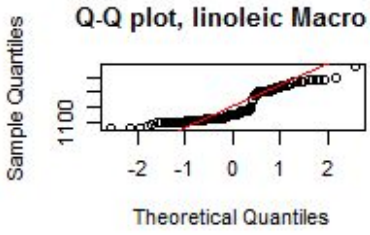

Q-Q plot, palmitic Macro 1
Sample Quantiles
1000
-3 -2 -1 0 1 2 3
Theoretical Quantiles
Q-Q plot, palmitic Macro 2
Sample Quantiles
1050
-2 -1 0 1 2
Theoretical Quantiles
Q-Q plot, palmitic Macro 3
Sample Quantiles
600 1400
-2 -1 0 1 2
Theoretical Quantiles
Q-Q plot, palmitoleic Macro 1
Sample Quantiles
50 250
-3 -2 -1 0 1 2 3
Theoretical Quantiles
Q-Q plot, palmitoleic Macro 2
Sample Quantiles
40 120
-2 -1 0 1 2
Theoretical Quantiles
Q-Q plot, palmitoleic Macro 3
Sample Quantiles
50
-2 -1 0 1 2
Theoretical Quantiles
Q-Q plot, stearic Macro 1
Sample Quantiles
150 350
-3 -2 -1 0 1 2 3
Theoretical Quantiles
Q-Q plot, stearic Macro 2
Sample Quantiles
200
-2 -1 0 1 2
Theoretical Quantiles
Q-Q plot, stearic Macro 3
Sample Quantiles
200 350
-2 -1 0 1 2
Theoretical Quantiles

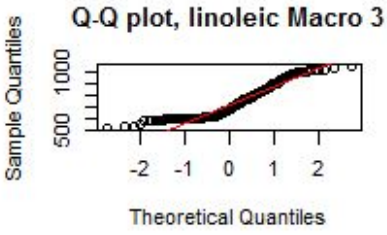

Q-Q plot, oleic Macro 1
Sample Quantiles
6500
-3 -2 -1 0 1 2 3
Theoretical Quantiles

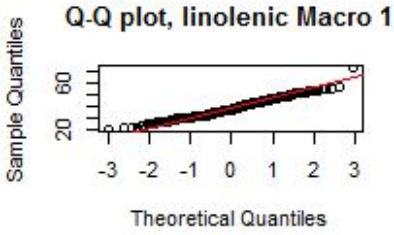

Q-Q plot, oleic Macro 2
Sample Quantiles
6900
-2 -1 0 1 2
Theoretical Quantiles

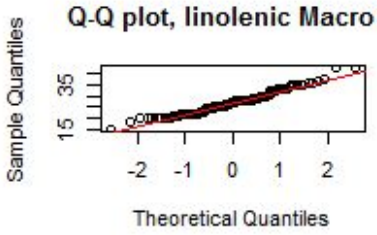

Q-Q plot, oleic Macro 3
Sample Quantiles
7400
-2 -1 0 1 2
Theoretical Quantiles

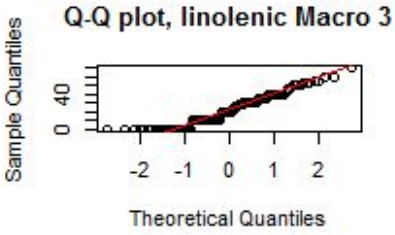

Q-Q plot, linoleic Macro 1
Sample Quantiles
600
-3 -2 -1 0 1 2 3
Theoretical Quantiles

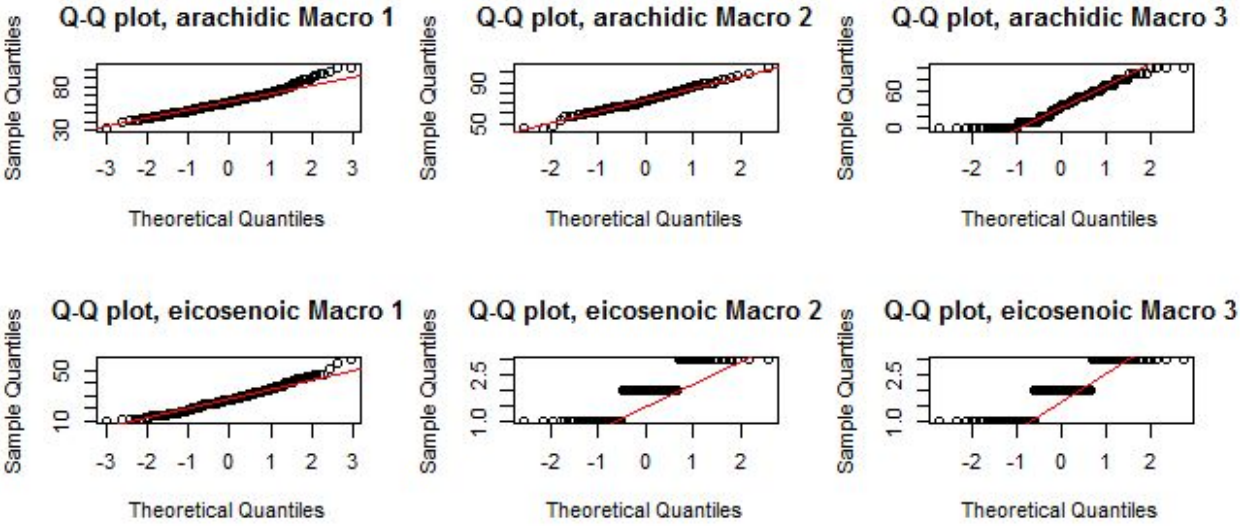

Q-Q plot, linoleic Macro 2
Sample Quantiles
1100
-2 -1 0 1 2
Theoretical Quantiles


Q-Q plot, linoleic Macro 3
Sample Quantiles
500 1000
-2 -1 0 1 2
Theoretical Quantiles


Q-Q plot, linolenic Macro 1
Sample Quantiles
20 60
-3 -2 -1 0 1 2 3
Theoretical Quantiles


Q-Q plot, linolenic Macro 2
Sample Quantiles
15 35
-2 -1 0 1 2
Theoretical Quantiles


Q-Q plot, linolenic Macro 3
Sample Quantiles
0 40
-2 -1 0 1 2
Theoretical Quantiles


Q-Q plot, arachidic Macro 1
Sample Quantiles
Theoretical Quantiles
Q-Q plot, arachidic Macro 2
Sample Quantiles
Theoretical Quantiles
Q-Q plot, arachidic Macro 3
Sample Quantiles
Theoretical Quantiles
Q-Q plot, eicosenoic Macro 1
Sample Quantiles
Theoretical Quantiles
Q-Q plot, eicosenoic Macro 2
Sample Quantiles
Theoretical Quantiles
Q-Q plot, eicosenoic Macro 3
Sample Quantiles
Theoretical Quantiles

```
########################################
# ST 590 Multivariate Project
########################################

# retrieve data file
library(pdfCluster)
data("oliveoil")
head(oliveoil)

#################################################
## Methods
#################################################

## set color by macro area
col <- as.numeric(factor(oliveoil$macro.area))+1

## Figure 1: Pairs plot
pairs(oliveoil[,3:9], col=col)

## Figure 2: Parallel Coordinates plot
library(MASS)
parcoord(oliveoil[,3:9],col=col, var.label=TRUE,
         main="Parallel Coordinate Plot by Macro Area")

## Figure 3: Boxplot for Palmitic oil
boxplot(palmitic ~ macro.area, range=3,data=oliveoil,
        main="Palmitic Oil by Macro Area", xlab="Macro Area",
ylab="Con. (parts/10,000)")

# remove outliers
datanew <- oliveoil[-c(79,390,522),]
mregion <- as.numeric(factor(datanew$macro.area))

#################################################
## MANOVA
#################################################

## Table 2: Chemical means by macro area
mu1<- colMeans(na.omit(datanew[mregion==1,3:9]))
mu2<- colMeans(na.omit(datanew[mregion==2,3:9]))
mu3<- colMeans(na.omit(datanew[mregion==3,3:9]))

# means of each macro region
```

```
means<-t(round(cbind(mu1,mu2,mu3)))
rownames(means)<- c("South", "Sardinia","Center North")
means


## Figure 4: Variance plot by macro area
plot(diag(var(na.omit(datanew[mregion==1,3:9]))), ylab="variance",
pch=1,type="b")
points(diag(var(na.omit(datanew[mregion==2,3:9]))),col=2,pch = 2,type
= "b")
points(diag(var(na.omit(datanew[mregion==3,3:9]))),col=3,pch=3,type =
"b")

legend(x="topleft", legend = c("Macro1", "Macro2","Macro3"), pch=1:3,
col=1:3)

## Figure 5: Correlation matrices by macro area
Gammahat1<- round(cor(na.omit(datanew[mregion==1,3:9])),2)
Gammahat2<- round(cor(na.omit(datanew[mregion==2,3:9])),2)
Gammahat3<- round(cor(na.omit(datanew[mregion==3,3:9])),2)

library(lattice)
levelplot(Gammahat1)
levelplot(Gammahat2)
levelplot(Gammahat3)

## Manova test
X <- as.matrix(datanew[, 3:9])
regionf = as.factor( datanew[, 1])
fit <- manova(X ~ regionf)
summary(fit, test="Wilks")

W<- summary(fit,test="Wilks")$SS$Residuals
g <-3  # number of regions to be compared
p<- 7  # number of variables
n<-dim(datanew)[1] # number of total obs
critical_value_t = qt(0.05/(p*g*(g-1)), (n-p), lower.tail = F)

# sample size
n1<- sum(mregion==1)
n2<- sum(mregion==2)
n3<- sum(mregion==3)
```

```
# t-test for the mean difference between macro regions
t_12<- (mu1-mu2)/sqrt(diag(W)/(n-g)*(1/n1+1/n2))
t_23<- (mu2-mu3)/sqrt(diag(W)/(n-g)*(1/n2+1/n3))
t_13<- (mu1-mu3)/sqrt(diag(W)/(n-g)*(1/n1+1/n3))

ttest<- round(cbind(t_12,t_23,t_13),2)
colnames(ttest)<- c("1 vs. 2", "2 vs. 3", "1 vs. 3")
## Table 3: t-test of pairwise macro areas
t(ttest)

#####################################################
## Principal Components
#####################################################

## subset data
chemicals <- as.matrix(datanew[,3:9])
apply(chemicals,2,summary)

## Run PCA function
pcaall <- prcomp(chemicals, center = TRUE, scale. = TRUE)
## Table 4: Proportion of variance
summary(pcaall)

## Figure 6: Scree plot
plot(1:7, summary(pcaall)$importance[3,],type="o",main="Scree
Plot",ylab="Cumulative Proportion")

## Table 5: First 4 PC's
PC = pcaall$rotation
round(PC[,1:4], 2)

## Figure 7: PCA Pairs Plot
pairs(pcaall$x[,1:4],col=col)

#####################################################
## Factor Analysis
#####################################################

#correlation matrix
R<-cor(chemicals)
#eigen value &eigen vectors
e<-eigen(R)$vectors
lambda<-eigen(R)$values
```

```
#factors
F1<-sqrt(lambda[1])*e[,1]
F2<-sqrt(lambda[2])*e[,2]
F3<-sqrt(lambda[3])*e[,3]
F4<-sqrt(lambda[4])*e[,4]


L.pc<-cbind(F1, F2,F3,F4)
rownames(L.pc)<-colnames(chemicals)

## Tables 6 and 7: VARIMAX rotation scores and proportion of variance
varimax(L.pc, normalize=T)


#####################################################
## Discriminant Analysis
#####################################################
library(klaR)

area <- as.factor(datanew$macro.area)

## LDA for 3 macro areas
lda.fit = lda(chemicals, grouping = area)  ## perform LDA
lda.pred = predict(lda.fit, chemicals)$class  ## predicting the
classes for each data point

## Apparent error rate for LDA
APER.lda = mean(lda.pred != area)

## QDA for 3 macro areas
qda.fit = qda(chemicals, grouping = area)
qda.pred = predict(qda.fit, chemicals)$class  ## predicting the
classes for each data point

APER.qda = mean(qda.pred != area) ## Apprent error rate for LDA


## Figure 8: LDA and QDA examples
partimat(chemicals[,c(4,5)],grouping=as.factor(mregion),method="lda",
main="LDA Example")
```

```
partimat(chemicals[,c(4,5)],grouping=as.factor(mregion),method="qda",
main="QDA Example")

## Table 8: LDA confusion matrix
table(area, lda.pred)

## Table 9: QDA confusion matrix
table(area, qda.pred)

## Table 10: LDA linear discriminant vectors
lda.fit$scaling

#####################################################
## Appendix
#####################################################

## Figure A1: Scatterplot of Eicosenoic oil
plot(oliveoil[,10], col=col, ylab= "Concentration",
main="Eicosenoic")

## Figure A2: Q-Q plots of each chemical in each macro area
par(mfrow=c(3,3))
for (i in 1:8) {
  for (j in 1:3){
    qqnorm(datanew[mregion==j,i+2],
           main = paste("Q-Q plot,", colnames(data)[i+2],"Macro",j ))
    qqline(datanew[mregion==j,i+2],col=2)
  }
}
```