Mini Project 2 Stat 666

Devin Eddington, David Kartchner, Stephen Merrill

1 Introduction

Agricultural success depends on a number of local factors, such as type of soil (i.e. clay, silt, sand, etc) and its nutrient profile. Moreover, certain crops tend to drain the soil of particular nutrients, which can lead to both lower crop yield and lower-quality crops over time. One way to examine changes in crop outcomes is to examine changes in fatty acid profile in different crops across time. In the following analysis, we consider two questions regarding two adjacent agricultural regions. First, we examine whether long-term farming causes a significant change in the fatty acids of olives grown in each area by comparing a current sample with historical averages. Second, we consider whether modern fertilization practices and cross-pollination have caused formerly heterogeneous regions to develop homogeneous crops by comparing current samples from each region with each other. We consider data taken on eight fatty acids in Kalamata olives from agricultural regions 2 and 4 of Laconia, Greece. Our datasets contain 56 and 36 observations, respectively.

2 Methods

2.1 Missing Data and Imputation

Before beginning our analysis, we observed that between the two regions, 46 of the 92 observations contained one or more fatty acid score that is missing. In order to impute the unrecorded data in an unbiased way, we must be able to confidently suppose that the data is missing at random, or that the missing values aren't systematically different from non-missing values. Since randomness of missing data cannot be directly tested, we created two-dimensional scatter plots for each of the 28 pairings of acids, noting whether observations with missing data deviated noticeably from observations without missing data or if there is some form of clustering happening among the missing data. A telling subsample of the plots for each region is pictured in Figure 1.

The clustering of red points, or observations containing missing values for fatty acids other than the fatty acids plotted on the axises, indicate that observations in region 4 were not missing at random. The spread of red points in region 2 provide no evidence that these observations are not missing at random. We acknowledge that the non-random missingness in region 4 may bias our results for this region.

After assessing the randomness of missing observations, we imputed missing values with an EM (Expectation Maximization) algorithm. The algorithm we used proceeds as follows:

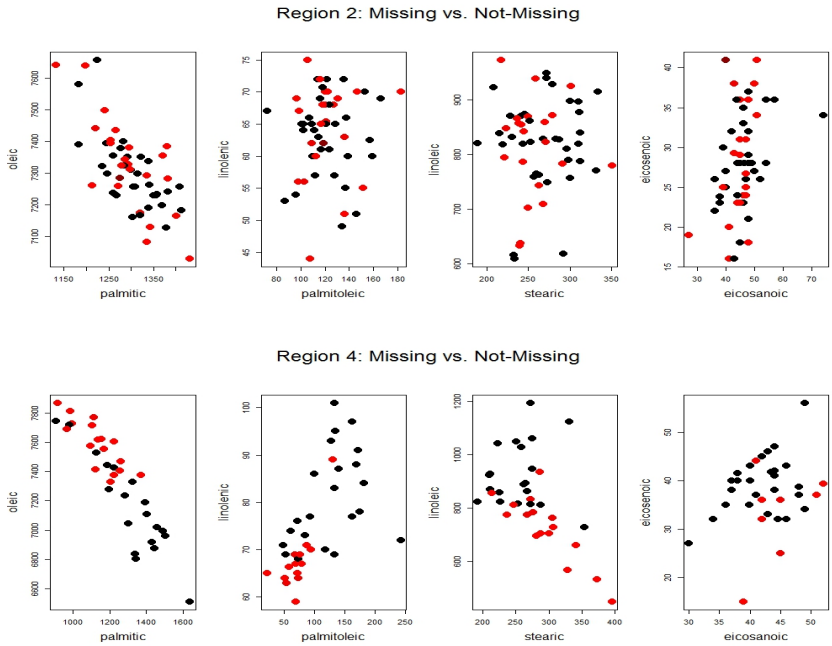

Figure 1: Scatter plots of fatty acids with and without missing data, given by black and red dots, respectively. Note that in region 2, observations with missing data do not differ noticably from those without missing data, whereas such observations differ markedly in region 4.

1. Estimate $\tilde{\boldsymbol{\mu}}_{\mathbf{0}}$ as the observed mean of each fatty acid and $\tilde{\boldsymbol{\Sigma}}$ as the sample variance after filling missing observations with column mean.

2. For each observation $\boldsymbol{x_i} = \begin{bmatrix} \boldsymbol{x}_i^{(1)} \\ \boldsymbol{x}_i^{(2)} \end{bmatrix}$, define $\boldsymbol{x}_i^{(1)}$ to be the missing components of that observation and $\boldsymbol{x}_i^{(2)}$ to be the observed components. Partition $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$ similarly such that $\tilde{\boldsymbol{\mu}} = \begin{bmatrix} \tilde{\boldsymbol{\mu}}^{(1)} \\ \tilde{\boldsymbol{\mu}}^{(2)} \end{bmatrix}$ and $\tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} \tilde{\boldsymbol{\Sigma}}_{\mathbf{11}} & \tilde{\boldsymbol{\Sigma}}_{\mathbf{12}} \\ \tilde{\boldsymbol{\Sigma}}_{\mathbf{21}} & \tilde{\boldsymbol{\Sigma}}_{\mathbf{22}} \end{bmatrix}$.

3. Impute missing values using regression: $\tilde{\boldsymbol{x}}_i^{(1)} = \tilde{\boldsymbol{\mu}}_i^{(1)} + \boldsymbol{B}(\boldsymbol{x}_i^{(2)} - \tilde{\boldsymbol{\mu}}_i^{(2)})$ with the regression coefficients $\boldsymbol{B} = \tilde{\boldsymbol{\Sigma}}_{12}\tilde{\boldsymbol{\Sigma}}_{22}^{-1}$.

4. Calculate new sufficient statistics, meaning obtain new estimations of $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$ using the imputed data.

5. Repeat steps 2-4 the imputation converges (i.e. imputed values stop meaningfully changing).

The EM algorithm guarantees an increase in likelihood for each iteration which means that at worst we have obtained a locally optimal imputed values. Unfortunately, the EM algorithm we used does not fix data collection problems, meaning that non-random missingness will likely bias our imputed values for region 4. Moreover, EM predictions strictly follow the prescribed formula, which does not account for the variability that would occur in nature. We can account for some of the variance associated with our predicted values using a procedure called multiple imputation after the EM algorithm has converged.

To preform multiple imputation, let k be the number of iterations before the EM algorithm converges. Collect J new predictions using $\tilde{\boldsymbol{x}}_{[j]}^{(1)} = \tilde{\boldsymbol{\mu}}_k^{(1)} + \boldsymbol{B}_k(\boldsymbol{x}_k^{(2)} - \tilde{\boldsymbol{\mu}}_k^{(2)}) + \boldsymbol{e}_{[j]}^{(1)}$ where $\boldsymbol{e}_{[j]}^{(1)} \sim N_p(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{k}})$. From each of the J prediction matrices calculate $\bar{\boldsymbol{x}}_{[j]}$ as the column means and $\boldsymbol{S}_{[j]}$ as the covariance for the j^{th} prediction matrix. Our final estimates for $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}$ are calculated as follows:

$$\bar{\boldsymbol{x}}_{MI} = \frac{1}{J}\sum_{j=1}^{J} \bar{\boldsymbol{x}}_{[j]} \tag{1}$$

$$\boldsymbol{S}_{MI} = \frac{1}{J}\sum_{j=1}^{J} \frac{1}{n}\boldsymbol{S}_{[j]} + (1 + \frac{1}{J})\boldsymbol{B} \tag{2}$$

where $\boldsymbol{B} = \frac{1}{J-1}\sum_{j=1}^{J}(\bar{\boldsymbol{x}}_{[j]} - \bar{\boldsymbol{x}}_{MI})(\bar{\boldsymbol{x}}_{[j]} - \bar{\boldsymbol{x}}_{MI})'$

2.2 Hotelling T^2 and Discriminant Analysis

In order to compare the regions both to each other and to their respective historical means, we carried out multivariate hypothesis tests that compare the eight acids simultaneously. There are several advantages to using a multivariate test rather than a series of univariate tests, which this analysis will demonstrate. First, multivariate testing avoids the false inflation of the Type I error rate and makes use of correlation between the variables. Moreover, multivariate testing is often more powerful, meaning there is a higher probability of rejecting a false null hypothesis. We performed hypothesis tests multivariately based on a Hotelling T^2 statistic, which assumes sampling is from a $N_p(\mu, \Sigma)$. The distribution of the T^2 statistic is indexed by two parameters, the dimension ($p = 8$) and degrees of freedom ($\nu = n - 1$). The Hotelling T^2 has the general form of:

$$T^2 = n(\bar{\boldsymbol{x}} - \boldsymbol{\mu_0})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu_0}) \tag{3}$$

This statistic has a mathematical connection to the F-distribution, which we use to calculate p-values:

$$\frac{\nu - p + 1}{\nu p} T_{p,\nu}^2 = F_{p,\nu-p+1} \tag{4}$$

Similarly, to assess difference between regions, we calculated:

$$T^2 = \frac{n_2 n_4}{n_2 + n_4}\Big((\bar{\boldsymbol{x}}_{\boldsymbol{2}} - \bar{\boldsymbol{x}}_{\boldsymbol{4}}) - (\boldsymbol{\mu_2} - \boldsymbol{\mu_4})\Big)' \boldsymbol{S_{pl}}^{-1}\Big((\bar{\boldsymbol{x}}_{\boldsymbol{2}} - \bar{\boldsymbol{x}}_{\boldsymbol{4}}) - (\boldsymbol{\mu_2} - \boldsymbol{\mu_4})\Big) \tag{5}$$

where

$$\boldsymbol{S_{pl}} = (n_2 - 1)\boldsymbol{S_2} + (n_4 - 1)\boldsymbol{S_4} \tag{6}$$

for the covariance matrices $\boldsymbol{S_2}$ and $\boldsymbol{S_4}$ of regions 2 and 4, respectively.

To analyze the factors driving the difference between regions, we calculated the discriminant function $\boldsymbol{a}$, which indicates the direction that maximizes the separation between the means of both regions as follows:

$$\boldsymbol{a} = \boldsymbol{S_{pl}}^{-1}(\bar{\boldsymbol{x_2}} - \bar{\boldsymbol{x_4}}) \tag{7}$$

where $\boldsymbol{S_{pl}}$ is the pooled covariance matrix for the two regions. We then find the relative contributions of each variable x_j by standardizing the discriminant function as follows:

$$\boldsymbol{a^*} = diag(\boldsymbol{S_{pl}})\boldsymbol{a} \tag{8}$$

2.3 Test for Redundant Information

While some fatty acids are easy to measure, others are not. Accordingly, we tested our data to see if linoleic and linolenic acids (both of which are costly to measure) add additional information regarding the separation of the groups above and beyond that given by the other six acids. To do so, we partition $\bar{\boldsymbol{x_2}} = \begin{pmatrix} \bar{\boldsymbol{y_2}} \\ \bar{\boldsymbol{z_2}} \end{pmatrix}$ and $\bar{\boldsymbol{x_4}} = \begin{pmatrix} \bar{\boldsymbol{y_4}} \\ \bar{\boldsymbol{z_4}} \end{pmatrix}$, where $\bar{\boldsymbol{y_2}}$ and $\bar{\boldsymbol{y_4}}$ are the $p = 6$ measurements of the easy-to-measure acids and $\bar{\boldsymbol{z_2}}$ and $\bar{\boldsymbol{z_4}}$ are the $q = 2$ measurements of the difficult-to-measure acids from each respective region. To test whether $\bar{\boldsymbol{z_2}}$ and $\bar{\boldsymbol{z_4}}$ significantly increase the separation between means, we calculated the following test statistic:

$$F_{add} = \left(\frac{\nu - p - q + 1}{q}\right)\frac{T^2_{p+q} - T^2_p}{\nu + T^2_p} \sim F_{q,\nu-p-q+1} \tag{9}$$

where

$$T^2_{p+q} = \frac{n_1 n_2}{n_1 + n_2}(\bar{\boldsymbol{x_1}} - \bar{\boldsymbol{x_2}})'\boldsymbol{S_{pl}}^{-1}(\bar{\boldsymbol{x_1}} - \bar{\boldsymbol{x_2}}) \tag{10}$$

$$T^2_p = \frac{n_1 n_2}{n_1 + n_2}(\bar{\boldsymbol{y_1}} - \bar{\boldsymbol{y_2}})'\boldsymbol{S_{yy}}^{-1}(\bar{\boldsymbol{y_1}} - \bar{\boldsymbol{y_2}}) \tag{11}$$

with $\boldsymbol{S_{pl}}$ is as defined in Eq. (6) and $\boldsymbol{S_{yy}}$ is the subsection of this matrix corresponding to $\bar{\boldsymbol{y_i}}$. Rejecting the null hypothesis in this case means that linoleic and linolenic acids contribute significantly to the separation of the means of regions 2 and 4.

3 Results

Using the procedures described in the previous section, we tested the differences between each region and its historical average but found no significant differences, although region 4 narrowly missed rejecting the null with a p-value slightly above our α level of 0.05. These results are summarized in Table 1. Next, we tested the both whether acid profiles of the olives differed significantly between regions and whether linoleic and linolenic acids contributed to this difference, summarized in Table 2. The multivariate test strongly rejected the null hypothesis of no difference, but the univariate tests gave mixed results, with Palmitoleic, Stearic, Oleic and Linoleic Acids each failing to reject the null of no difference. We further tested whether linoleic and linolenic acids contribute to the difference between regions and found strong

Location	Null Hypothesis	Test Statistic	p-value
Region 2	$H_0 : \mu_2 = \mu_0$	$F_{8,48}$: 1.107	0.375
Region 4	$H_0 : \mu_4 = \mu_0$	$F_{8,28}$: 2.179	0.061

Table 1: Tests comparing each region to its historical average. Note that neither region differs significantly from its historical average for $\alpha = 0.05$.

Test	Null Hypothesis	Test Statistic	p-value
Overall (Multivariate)	$H_0 : \mu_2 = \mu_4$	$F_{8,83}$: 21.356	3.034e-17
Palmitic	$H_0 : \mu_{21} = \mu_{41}$	t_{90}: 2.711	0.004
Palmitoleic	$H_0 : \mu_{22} = \mu_{42}$	t_{90}: 2.213	0.015
Stearic	$H_0 : \mu_{23} = \mu_{43}$	t_{90}: -1.059	0.146
Oleic	$H_0 : \mu_{24} = \mu_{44}$	t_{90}: -0.956	0.171
Linoleic	$H_0 : \mu_{25} = \mu_{45}$	t_{90}: -0.671	0.252
Eicosanoic	$H_0 : \mu_{26} = \mu_{46}$	t_{90}: 2.600	0.005
Linolenic	$H_0 : \mu_{27} = \mu_{47}$	t_{90}: -6.574	1.560e-9
Eicosenoic	$H_0 : \mu_{28} = \mu_{48}$	t_{90}: -6.583	1.499e-9
Redundant Information	$H_0 : x_5, x_7$ Redundant	$F_{2,81}$: 25.301	2.908e-9

Table 2: Tests regarding whether the acid profiles of olives differ between regions. Observe that the multivariate test is more significant than any of the univariate tests. For each of the univariate tests, bote that the rejection region for each of the individual acids is subject to a Bonferroni correction of $\frac{\alpha}{p}$, so with $\alpha = 0.05$ we reject the null if the p-values are below 0.00625. The redundant information test considers whether linoleic and linolenic significantly contribute to the separation between regions.

evidence against the null that their contribution to the inter-region separation is negligable.

The univariate tests suffer from diminished power to reject the null and are unable to account for correlation between the variables. This is apparent when considered in the context of the discriminant analysis summarized in Table 3. According to these results, Oleic and Linoleic Acids have the two strongest contributions to the multivariate difference between the regions, yet neither of these acids was found to be univariately different. Clearly univariate testing does not present the entire profile of this comparison between regions.

Palmitic	Palmitoleic	Stearic	Oleic	Linoleic	Eicosanoic	Linolenic	Eicosenoic
-3.191	-1.293	-3.278	-9.698	-5.203	1.892	-3.523	-1.032

Table 3: Values of the scaled discriminant function $\boldsymbol{a}^*$ corresponding to each acid. It appears that Oleic and Linoleic acids contribute most to the separation of the two regions

4 Conclusion

Our analysis successfully answered the relevant research questions of historical and regional differences in the fatty acid profiles of Kalamanta olives. We used multivariate hypothesis tests, which resulted in findings of no evidence of a historical difference and strong evidence of a regional difference. However, a sequence of univariate tests obtained no evidence of a difference between specific acids. This dichotomy is a result of the shortcomings of univariate testing on multivariate data.

Before any analysis was carried out the missing values in the data had to be imputed, which was done successfully through an EM algorithm with multiple imputation. We were also interested in determining if the missingness of the data was completely at random or if there was some underlying structure that was affecting where the data was missing. By examining where the missing data occurred in relation to the complete data, we concluded that the missingness in Region 4 was not at random. Although this had the potential to bias the results, we proceeded with that caveat acknowledged.

5 Appendix

5.1 Code

```
olive2 <- read.table("https://tofu.byu.edu/stat666/datasets/oliver2a",header=T)
olive4 <- read.table("https://tofu.byu.edu/stat666/datasets/oliver4a",header=T)

em.impute <- function(x,tolerance) {
  miss <- is.na(x)
  mu <- matrix(apply(x,2,mean,na.rm=T),nrow=nrow(x),ncol=ncol(x),byrow=T)
  x[miss] <- 0
  x <- x + miss * mu
  sigma <- cov(x)
  tol <- 1
  iter <- 0
  while(tol > tolerance) {
    old.x <- x
    mu <- matrix(apply(x,2,mean,na.rm=T),nrow=nrow(x),ncol=ncol(x),byrow=T)
    sigma <- cov(x)
    for(i in 1:nrow(x)) {
      sig22 <- sigma[!miss[i,],!miss[i,]]
      sig12 <- sigma[miss[i,],!miss[i,]]
      B <- sig12%*%solve(sig22)
      xstar <- mu[1,miss[i,]] + B%*%as.numeric(x[i,!miss[i,]]-mu[1,!miss[i,]])
      x[i,miss[i,]] <- xstar
    }
    iter <- iter + 1
    tol <- max((x-old.x)/old.x)
  }
  x
}

olive2.em <- em.impute(olive2,0.0000001)
olive4.em <- em.impute(olive4,0.0000001)

#add error
library(expm)
mult.impute <- function(x,miss) {
  sig <- cov(x)
  for(i in 1:nrow(x)) {
    if(!all(!miss[i,])) {
      sigstarhat <- sig[miss[i,],miss[i,]]-sig[miss[i,],!miss[i,]]%*%
        solve(sig[!miss[i,],!miss[i,]])%*%sig[!miss[i,],miss[i,]]
      if(sum(miss[i,])==1)
        e <- 1/sqrt(sigstarhat)*rnorm(1)
      else
        e <- solve(sqrtm(sigstarhat))%*%rnorm(sum(miss[i,]))
      nummiss <- 1
      for(j in 1:ncol(x)) {
        if(miss[i,j]) {
          x[i,j] <- x[i,j]+e[nummiss]
          nummiss <- nummiss + 1
        }
      }
    }
  }
  x
```

```
}

num.imputes <- 500
p <- ncol(olive2)
betas2 <- matrix(0,nrow=num.imputes,ncol=p)
betas4 <- matrix(0,nrow=num.imputes,ncol=p)
covs2 <- matrix(0,nrow=p,ncol=p)
covs4 <- matrix(0,nrow=p,ncol=p)

for(i in 1:num.imputes) {
  o2 <- mult.impute(olive2.em,is.na(olive2))
  o4 <- mult.impute(olive4.em,is.na(olive4))
  betas2[i,] <- colMeans(o2)
  betas4[i,] <- colMeans(o4)
  covs2 <- covs2 + cov(o2)
  covs4 <- covs4 + cov(o4)
}

sum2 <- matrix(0,nrow=p,ncol=p)
sum4 <- matrix(0,nrow=p,ncol=p)
for(i in 1:num.imputes) {
  sum2 <- sum2 + (betas2[i,]-colMeans(betas2))%*%t(betas2[i,]-colMeans(betas2))
  sum4 <- sum4 + (betas4[i,]-colMeans(betas4))%*%t(betas4[i,]-colMeans(betas4))
}

B2 <- (1/num.imputes-1)*sum2
B4 <- (1/num.imputes-1)*sum4

var.xbar2 <- (covs2/num.imputes)+(1+1/num.imputes)*B2
var.xbar4 <- (covs4/num.imputes)+(1+1/num.imputes)*B4

imputed <- TRUE

# T2 for diff in means
n2 <- nrow(olive2.em)
n4 <- nrow(olive4.em)
if(!imputed) {
  Spl <- 1/(n2+n4-2)*((n2-1)*cov(olive2.em)+(n4-1)*cov(olive4.em))
  ybar2 <- colMeans(olive2.em)
  ybar4 <- colMeans(olive4.em)

} else {
  Spl <- 1/(n2+n4-2)*((n2-1)*var.xbar2+(n4-1)*var.xbar4)
  ybar2 <- colMeans(betas2)
  ybar4 <- colMeans(betas4)

}
T2.regions <- (ybar2-ybar4)%*%solve((1/n2+1/n4)*Spl)%*%t(t((ybar2-ybar4)))
Fstat <- T2.regions*(n2+n4-p-1)/((n2+n4-2)*p)
pval <- pf(Fstat,p,n2+n4-p-1,lower.tail=F)

#individual differences
results <- matrix(0,nrow=8,ncol=2)
colnames(results) <- c("t-score","p-value")
rownames(results) <- colnames(olive2)
t1 <- (ybar2[1]-ybar4[1])/sqrt(((n2+n4)/(n2*n4))*Spl[1,1])
results[1,1] <- t1
pval1 <- pt(t1,n2+n4-2,lower.tail = F)
```

```
results[1,2] <- pval1
t2 <- (ybar2[2]-ybar4[2])/sqrt(((n2+n4)/(n2*n4))*Spl[2,2])
results[2,1] <- t2
pval2 <- pt(t2,n2+n4-2,lower.tail = F)
results[2,2] <- pval2
t3 <- (ybar2[3]-ybar4[3])/sqrt(((n2+n4)/(n2*n4))*Spl[3,3])
results[3,1] <- t3
pval3 <- pt(t3,n2+n4-2,lower.tail = T)
results[3,2] <- pval3
t4 <- (ybar2[4]-ybar4[4])/sqrt(((n2+n4)/(n2*n4))*Spl[4,4])
results[4,1] <- t4
pval4 <- pt(t4,n2+n4-2,lower.tail = T)
results[4,2] <- pval4
t5 <- (ybar2[5]-ybar4[5])/sqrt(((n2+n4)/(n2*n4))*Spl[5,5])
results[5,1] <- t5
pval5 <- pt(t5,n2+n4-2,lower.tail = T)
results[5,2] <- pval5
t6 <- (ybar2[6]-ybar4[6])/sqrt(((n2+n4)/(n2*n4))*Spl[6,6])
results[6,1] <- t6
pval6 <- pt(t6,n2+n4-2,lower.tail = F)
results[6,2] <- pval6
t7 <- (ybar2[7]-ybar4[7])/sqrt(((n2+n4)/(n2*n4))*Spl[7,7])
results[7,1] <- t7
pval7 <- pt(t7,n2+n4-2,lower.tail = T)
results[7,2] <- pval7
t8 <- (ybar2[8]-ybar4[8])/sqrt(((n2+n4)/(n2*n4))*Spl[8,8])
results[8,1] <- t8
pval8 <- pt(t8,n2+n4-2,lower.tail = T)
results[8,2] <- pval8

#discriminant function for diff in means
a <- solve(Spl)%*%(t(t((ybar2-ybar4))))
D.1o2 <- diag(sqrt(diag(Spl)))
astar <- D.1o2%*%a

# T2 for comp to historical data
hist2 <- c(1300,120,265,7310,820,45,65,28)
hist4 <- c(1230,105,275,7360,830,41,75,38)
if(!imputed) {
  S2 <- cov(olive2.em)
  S4 <- cov(olive4.em)
} else {
  S2 <- var.xbar2
  S4 <- var.xbar4
}

T2.2 <- n2*(ybar2-hist2)%*%solve(S2)%*%t(t((ybar2-hist2)))
nu2 <- n2-1
Fstat2 <- T2.2*(nu2-p+1)/(nu2*p)
pval2 <- pf(Fstat2,p,nu2-p+1,lower.tail=F)

T2.4 <- n4*(ybar4-hist4)%*%solve(S4)%*%t(t((ybar4-hist4)))
nu4 <- n4-1
Fstat4 <- T2.4*(nu4-p+1)/(nu4*p)
pval4 <- pf(Fstat4,p,nu4-p+1,lower.tail=F)

#test for additional info
q <- 2
```

```
SplReduced <- Spl[-c(5,7),-c(5,7)]
T2Reduced <- n2*n4*t(ybar2[-c(5,7)]-ybar4[-c(5,7)]) %*% solve(SplReduced) %*%
        (ybar2[-c(5,7)]-ybar4[-c(5,7)]) / (n2 + n4)
T2add <- ((n2+n4-2)-p)*(T2.regions-T2Reduced)/((n2+n4-2) + T2Reduced)
FstatAdd <- ((n2+n4-2)-p-1+1)*(T2.regions-T2Reduced)/(q*(n2+n4-2+T2Reduced))
pvalReduced <- pf(FstatAdd, q, (n2+n4-2)-p-q+1, lower.tail = F)

# Graphics to show missing at random
index <- combn(8,2)
index <- index[,c(3,12,15,27)]
miss2 <- is.na(oliver2a)
par(mfrow=c(1,4), oma = c(0, 0, 2, 0))
for(i in 1:ncol(index)){
  k <- colnames(oliver2a)[index[,i][1]]
  l <- colnames(oliver2a)[index[,i][2]]
  color <- apply(miss2[,-index[,i]],1,sum)+1
  plot(New.oliver2a[,index[,i]],
       col=c("black","red1","red2","red4")[color],
       pch=19, xlab=paste(k), ylab=paste(l), cex=2, cex.lab=1.5)
}
mtext("Region 2: Missing vs. Not-Missing", outer = T, cex=1.5)

miss4 <- is.na(oliver4a)
par(mfrow=c(1,4), oma = c(0, 0, 2, 0))
for(i in 1:ncol(index)){
  k <- colnames(oliver2a)[index[,i][1]]
  l <- colnames(oliver2a)[index[,i][2]]
  color <- apply(miss4[,-index[,i]],1,sum)+1
  plot(New.oliver4a[,index[,i]],
       col=c("black","red1","red2","red4")[color],
       pch=19, xlab=paste(k), ylab=paste(l), cex=2, cex.lab=1.5)
}
mtext("Region 4: Missing vs. Not-Missing", outer = T, cex=1.5)
```

5.2 EM Algorithm Results

5.2.1 Region 2

	palmitic	palmitoleic	stearic	oleic	linoleic	eicosanoic	linolenic	eicosenoic
1	1315.00	139.00	230.00	7299.00	832.00	42.00	60.00	32.00
2	1321.00	136.00	217.00	7174.00	973.31	43.00	63.00	29.27
3	1359.00	115.00	246.00	7234.00	874.00	45.00	63.00	18.00
4	1378.00	111.00	272.00	7127.00	940.00	46.00	64.00	23.00
5	1295.00	109.00	245.00	7326.61	842.61	43.00	62.00	38.00
6	1275.00	121.00	270.05	7285.00	859.71	40.00	65.42	41.00
7	1336.00	120.00	334.35	7083.00	915.00	50.00	70.00	38.00
8	1309.00	122.00	241.00	7257.00	870.00	46.00	72.00	35.00
9	1340.00	114.00	189.00	7337.00	820.00	48.00	72.00	21.00
10	1299.00	116.00	253.00	7309.00	823.00	40.00	69.00	27.00
11	1221.00	122.12	221.00	7441.00	794.40	54.00	70.00	28.00
12	1245.00	72.00	283.00	7395.00	829.00	44.00	67.00	28.00
13	1285.00	129.00	244.00	7323.00	819.00	57.00	65.00	36.00
14	1248.00	107.00	313.00	7299.00	840.00	46.00	66.00	33.00
15	1356.00	106.00	236.00	7229.22	866.00	48.00	75.00	36.00
16	1260.00	102.00	228.00	7354.00	870.00	49.00	64.00	28.00
17	1261.00	121.00	312.00	7238.00	877.00	47.00	65.00	25.00
18	1304.00	124.00	279.00	7160.00	928.00	48.00	61.00	37.00
19	1344.00	117.00	310.92	7129.00	897.00	51.00	65.00	41.00
20	1323.00	96.00	300.00	7351.00	757.00	47.00	54.00	26.00
21	1292.00	117.00	215.00	7351.00	839.00	48.00	61.00	32.00
22	1254.00	118.00	244.00	7394.00	786.00	46.00	70.72	24.00
23	1320.93	131.00	259.00	7167.00	939.00	41.00	69.00	20.00
24	1213.00	109.00	301.00	7261.00	925.00	47.00	64.98	31.00
25	1341.68	98.00	351.00	7262.00	780.00	41.00	56.00	16.00
26	1266.00	97.00	263.00	7435.00	743.00	47.94	69.00	29.00
27	1298.00	99.00	312.90	7311.00	787.00	45.00	67.00	23.00
28	1272.00	116.00	279.00	7258.00	872.00	50.02	72.00	27.00
29	1278.00	87.00	332.00	7379.00	771.00	44.00	53.00	24.00
30	1184.00	112.00	311.00	7391.00	819.00	48.00	57.00	28.00
31	1382.00	110.00	268.00	7241.00	828.00	39.00	60.00	30.00
32	1183.00	146.00	292.00	7580.00	618.00	38.00	51.00	23.00
33	1261.00	153.00	219.00	7355.00	818.00	52.00	70.00	26.00
34	1198.00	137.13	239.00	7639.00	633.00	27.00	55.00	19.00
35	1225.00	134.00	232.00	7658.00	616.00	36.00	49.00	26.00
36	1339.00	166.00	208.00	7190.00	923.00	40.00	69.00	25.00
37	1132.00	157.00	240.00	7641.00	638.00	45.00	62.51	31.00
38	1381.00	183.00	233.06	7385.00	609.00	47.00	70.00	25.00
39	1409.00	128.00	257.00	7257.00	759.00	43.00	57.00	16.00
40	1306.00	127.00	250.00	7257.59	869.00	47.00	68.00	24.00
41	1372.00	120.00	250.00	7355.00	702.00	46.39	68.00	28.00
42	1336.00	113.00	242.00	7293.00	855.00	38.00	60.00	23.79
43	1401.00	135.90	238.00	7164.00	857.00	45.00	72.00	36.00
44	1279.15	119.00	271.01	7323.88	823.00	40.00	62.00	41.00
45	1432.00	152.00	272.20	7029.00	949.00	39.00	55.00	25.00
46	1412.00	124.00	298.00	7182.00	790.00	45.00	68.00	28.00
47	1369.80	147.00	291.00	7197.00	783.00	51.00	70.00	34.00
48	1383.00	118.00	273.00	7282.00	749.06	45.00	68.00	29.00
49	1283.00	102.00	263.00	7400.00	763.00	54.00	65.00	28.00
50	1296.00	136.00	260.00	7380.00	764.47	48.00	51.00	18.00
51	1287.00	108.00	287.00	7343.00	827.47	44.00	44.00	23.00
52	1351.00	159.00	296.00	7229.00	810.00	36.00	60.00	22.00
53	1241.00	97.00	268.00	7499.00	709.00	43.90	69.00	36.00
54	1267.00	101.00	300.00	7230.00	898.00	74.00	65.00	34.00
55	1235.00	138.00	252.00	7322.00	861.00	54.00	66.00	36.00
56	1255.00	103.00	223.00	7404.65	848.00	47.00	56.00	26.72

5.2.2 Region 4

	palmitic	palmitoleic	stearic	oleic	linoleic	eicosanoic	linolenic	eicosenoic
1	1222.00	133.00	227.00	7425.00	824.00	36.00	69.00	35.00
2	1639.00	172.00	331.00	6510.00	1124.00	46.00	91.00	32.00
3	1345.00	133.00	272.00	6801.00	1194.00	48.00	83.00	37.00
4	1339.00	170.00	275.00	6838.00	1060.00	46.00	88.00	43.00
5	1194.00	135.00	263.00	7277.00	889.00	44.00	95.00	41.00
6	1112.00	68.00	395.40	7770.00	448.00	52.00	69.00	39.35
7	1222.00	70.00	329.00	7605.00	566.00	48.00	67.00	38.67
8	1136.00	72.00	341.00	7616.00	661.00	49.00	65.00	34.09
9	985.91	21.96	277.00	7815.00	784.00	45.00	65.00	25.00
10	1105.00	74.52	373.00	7714.00	532.00	51.00	68.00	37.00
11	1096.12	79.00	305.00	7576.00	763.00	45.00	67.00	36.00
12	1284.00	93.00	265.00	7235.00	893.00	43.00	77.00	46.00
13	1120.00	69.00	275.05	7416.00	946.00	42.00	59.00	36.00
14	916.00	52.00	281.00	7870.00	694.00	42.00	64.00	44.91
15	905.00	49.00	288.00	7747.00	812.00	49.00	71.00	56.00
16	1206.00	60.79	287.00	7329.00	935.00	44.00	74.00	42.00
17	1457.00	182.00	267.00	7020.00	863.00	41.00	84.00	37.00
18	1327.00	140.00	193.00	7328.00	823.00	36.00	87.00	35.00
19	1303.00	100.00	251.00	7045.00	1049.00	40.00	86.00	40.00
20	1444.00	175.00	259.00	6876.00	1027.00	34.00	78.00	32.00
21	1505.00	243.00	226.00	6962.00	858.00	30.00	72.00	27.00
22	1429.00	162.00	223.00	6917.00	1041.00	37.00	77.00	40.00
23	1491.00	162.00	211.00	6994.00	928.00	37.00	97.00	38.00
24	1393.00	128.00	211.00	7189.00	870.00	38.00	93.00	40.00
25	1404.00	134.00	210.00	7110.00	923.00	40.00	101.00	43.00
26	1222.00	130.00	214.00	7374.00	856.00	38.00	89.00	41.46
27	1153.00	74.00	288.25	7623.94	705.00	42.00	64.00	32.00
28	1169.00	76.00	307.00	7553.00	728.00	44.66	69.00	32.00
29	1369.00	118.64	237.00	7375.00	775.00	39.00	70.00	15.00
30	993.00	58.00	267.00	7730.00	773.00	41.00	66.35	44.00
31	980.00	53.00	254.00	7719.00	815.00	44.00	69.00	47.00
32	967.00	55.00	273.00	7692.00	833.00	43.49	63.00	41.78
33	1128.00	73.00	354.00	7527.00	728.00	44.00	76.00	38.00
34	1188.00	85.00	273.00	7445.00	814.00	44.00	73.00	42.00
35	1257.00	95.00	247.00	7405.00	812.00	39.88	70.00	35.00
36	1262.00	88.00	301.00	7471.00	704.00	43.00	71.00	33.03