Mini Project 2

Jackson Passey, Gavin Hatch, Ty Hawkes

2025-10-20

Introduction

This analysis attempts to answer questions about fatty acids in olives found in two different regions. The data includes response variables palmitic, palmitoleic, stearic, oleic, linoleic, eicosanoic, linolenic, and eicosenoic acids predicted by region (Region 2 and Region 4).

This report assesses whether the olive samples from each region deviate from their historical averages and whether the two regions deviate from each other. We also explore whether linoleic and linolenic acids can be dropped from the list of fatty acids without significantly decreasing the separation of the two samples.

The data analyzed contains several missing values. Part of this report determines whether the data is missing at random and provides a new dataset with imputed values using the Expectation-Maximization (EM) algorithm and multiple imputation.

We will show that the olive samples from both regions do not deviate significantly from their historical averages, that the two regions differ significantly from each other in terms of their fatty acid profiles, and that linoleic and linolenic acids should not be dropped from the analysis due to the significant amount of separation they contribute to distinguishing between the two regions.

Data Preparation

The data for regions 2 and 4 were obtained from separate organizations, which may have used different data collection procedures. Both datasets contain missing values in various fatty acid measurements. In order to perform a comprehensive multivariate analysis, these missing values need to be addressed. It is assumed prior to this analysis that the data from both regions are distributed multivariate normal.

Under the assumption that the data is missing at random (MAR), we employed the Expectation-Maximization (EM) algorithm combined with multiple imputation to fill in the missing values. Expectation-Maximization is an iterative method that estimates the parameters of a statistical model in the presence of missing data. The algorithm alternates between estimating the missing values given the observed data (E-step) and updating the model parameters based on the complete data (M-step). This process continues until convergence. The algorithm can be summarized as follows:

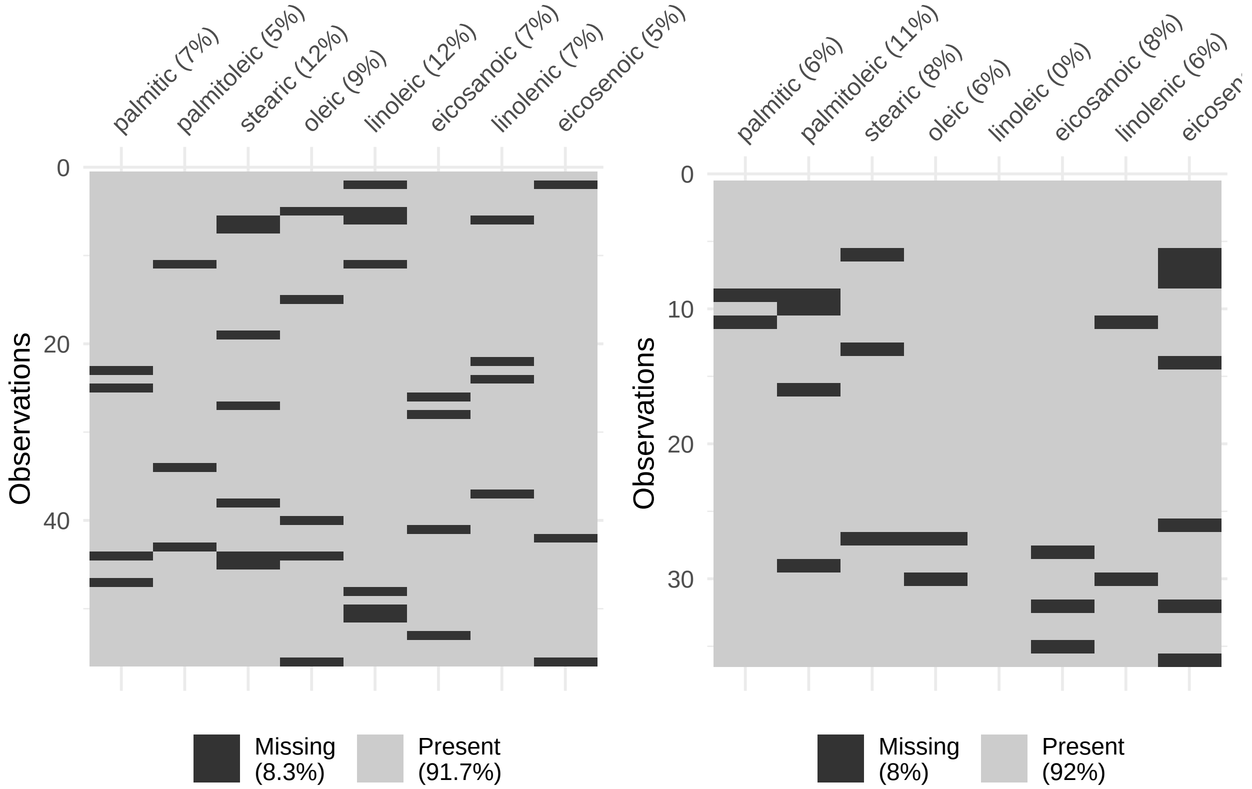
$$\text{E-step: } Q(\theta|\theta^{(t)}) = E[\log L(\theta; X_{obs}, X_{mis})|X_{obs}, \theta^{(t)}]$$

After convergence, multiple imputation is performed by generating several complete datasets, each with different imputed values drawn from the estimated distribution of the missing data. This approach accounts for the uncertainty associated with the imputation process. This is done by adding a random error term to each imputed value based on the estimated covariance structure of the data.

It is important to note that because of the missing data, the degrees of freedom for the analysis will be affected. In this analysis, we use Rubin's rules to adjust the degrees of freedom accordingly. We computed the adjusted degrees of freedom using the formula: $\nu = (M-1)(1+\frac{1}{r})^2$ where $r = (1+\frac{1}{M})tr(B\bar{U}^{-1})$

Assessment of Missing at Random (MAR) Assumption

In order to use the EM algorithm and multiple imputation correctly, we must first assess whether the missing data mechanism is consistent with the Missing at Random (MAR) assumption. MAR assumes that the probability of missingness is related to the observed data but not to the unobserved data. To test this assumption, we examined the patterns of missingness in the datasets for region 2 and region 4 visually and statistically with fisher's exact p-values.



An initial look at the missing data patterns for both regions indicates that the missingness does not appear to be systematically related to any particular variable. The visualizations show that missing values are scattered throughout the datasets without any obvious patterns.

To assess the MAR assumption more formally, we conducted pairwise Fisher's exact tests between the missingness indicators of each pair of variables. Fishers exact tests use contingency tables to determine if there are nonrandom associations between two categorical variables. In this case, we created 2x2 tables for each pair of variables, where one variable indicates whether a value is missing (1) or observed (0) for the dependent variable, and the other variable indicates the same for the predictor variable. The null hypothesis for each test is that the missingness of the dependent variable is independent of the missingness of the predictor variable.

The resulting p-values for each unique pair of variables (shown in the appendix) were all above common significance thresholds (e.g., 0.05), suggesting that there is no strong evidence to reject the MAR assumption. Based on these findings, we proceed under the assumption that the data is missing at random (MAR), allowing us to utilize the EM algorithm and multiple imputation for handling missing values in our multivariate analyses.

Analysis for Region 2

For Region 2, the sample mean vector was compared to the historical average representing the eight fatty acids (palmitic, palmitoleic, stearic, oleic, linoleic, eicosanoic, linolenic, and eicosenoic). Using the EM algorithm followed by multiple imputation to address missing data, a multivariate hypothesis test was performed. For each imputed dataset, the sample mean vector and within-imputation covariance matrix were computed. These results were then combined using Rubin's rules to form the following:

$$\bar{\boldsymbol{\theta}} = \frac{1}{M}\sum_{m=1}^{M}\boldsymbol{\theta}_m, \quad \mathbf{U}_{\text{bar}} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{U}_m, \quad \mathbf{B} = \frac{1}{M-1}\sum_{m=1}^{M}(\boldsymbol{\theta}_m - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta}_m - \bar{\boldsymbol{\theta}})^\top.$$

The total covariance matrix, incorporating both within- and between-imputation variability, was then

$$\mathbf{T} = \mathbf{U}_{\text{bar}} + \left(1 + \frac{1}{M}\right)\mathbf{B}.$$

Using $\bar{\boldsymbol{\theta}}$ and $\mathbf{T}$, a pooled Hotelling's T^2 statistic was computed as

$$T^2 = (\bar{\boldsymbol{\theta}} - \boldsymbol{\mu}_0)^\top \mathbf{T}^{-1}(\bar{\boldsymbol{\theta}} - \boldsymbol{\mu}_0),$$

which follows an approximate F-distribution under the null hypothesis $\mathcal{H}_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$. For Region~2, this test produced

$$T^2 = 8.5188722, \quad F = 0.9293315, \quad p = 0.4981799.$$

This p-value does not provide enough evidence that the Region~2 sample mean vector significantly deviates from its historical average. The pooled mean estimates after combining across imputations were

$$\bar{\boldsymbol{\theta}} = (1300.6265065,\ 121.3068211,\ 264.3366436,\ 7309.8880432,\ 818.2672166,\ 45.651223,\ 63.5366868,\ 28.3552148)$$

Examination of the sample means relative to the hypothesized values suggests that most components are close to their historical levels.

Analysis for Region 4

For Region~4, the sample mean vector was compared to the historical average (see appendix) using the same EM–multiple-imputation procedure described above. The resulting combined multivariate test yielded

$$T^2 = 16.5587032, \quad F = 1.6558703, \quad p = 0.1223715.$$

The test again does not provide enough evidence that the Region~4 fatty-acid profile differs significantly from its historical mean vector. The pooled mean estimates

$$\bar{\boldsymbol{\theta}} = (1229.5758095,\ 104.9693478,\ 273.1801133,\ 7358.7336155,\ 834.7222222,\ 42.3996755,\ 75.7251884,\ 37.6732712)$$

are remarkably close to the hypothesized values. It's clear that there is not much difference between the current and historical fatty-acid profiles for Region~4.

Comparison between Regions 2 and 4

To compare whether Regions 2 and 4 are the same, we used Hotelling's 2-Sample T test. We calculated a T^2 value of 189.7534332 which converts to an F-Statistic of 21.8743541. Using the F distribution with numerator degrees of freedom of 8 and a denominator degrees of freedom of 52.0714314(calculated using Ruben's rules for degrees of freedom that weights observed rows and imputed rows) results in a P-Value of $3.8746784 \times 10^{-14}$‘.

This is statistically significant and provides overwhelming evidence that the overall fatty-acid profiles of regions 2 and 4 differ. The agronomist's belief that region 2 and region 4 olives have evolved to have essentially the same profile in terms of the eight fatty acids is not supported by the data. The observed differences are unlikely to have come from random sampling variation alone.

It is important to note that the two samples were drawn by different organizations with potentially different data collection procedures, chemical analysis tools, and data censoring mechanisms. This combined with missing data that was imputed using MVI exposes limitations of this analysis. We do not know if significant variation comes from the different data collection procedures.

Table 1: Discriminant Function Coefficients (Transposed)

Variable	palmitic	palmitoleic	stearic	oleic	linoleic	eicosanoic	linolenic	eicosenoic
Coefficient	-0.0766	-0.1225	-0.1852	-0.1073	-0.1147	0.5183	-0.7425	-0.3166

Looking at the discriminant function, we can see that the largest coefficients are for Linolenic, Eicosanoic, and Eicosenoic acids (in that order). This suggests that these fatty acids contribute the most to the difference between regions 2 and 4. Further investigation into these specific fatty acids may provide more insights into the differences in olive profiles between the two regions.

Assessment of Linoleic and Linolenic Acids

Using a conditional Wilk's Λ test, we assessed whether linoleic acid and linolenic acid were important in contributing to the significant difference observed between regions 2 and 4. We set up the following null hypothesis:

$$H_0 : \boldsymbol{B_2} = \mathbf{0}$$

where $\boldsymbol{B_2}$ = mean vector of interest

Where the full model includes all 8 fatty acids and the reduced model excludes linoleic and linolenic acids.

With a Wilk's Λ of 0.5948317 converted to an F statistic, we calculated a P-value of $< .05$ and reject the null hypothesis. To support this rejection, we found that the discriminant functions (seen in appendix) separating regions also indicate some form of separation for linoleic and linolenic acids. In this case, we see that the standardized discriminant scores for each of them have high positive influence indicating that observations with higher values of linoleic and linolenic acids more likely reside in Region 2.

We conclude that linoleic and linoeic acids are important in contributing to the separation of the two regions beyond the information available from the other 6 acids.

Conclusions and Recommendations

The missing data appeared consistent with the MAR assumption, supporting the use of the EM algorithm followed by multiple imputation to generate valid inferences. For Region~2, the pooled Hotelling's T^2 provided weak evidence that the sample mean vector differed from its historical average.

Similarly, Region~4 showed weak evidence with its T^2. We conclude that both regions are in line with their historical averages.

The multivariate analysis comparing regions 2 and 4 revealed strong evidence that the fatty acid profiles differ between the two regions. Specifically, Hotelling's two-sample T^2 test indicated a statistically significant difference leading us to reject the null hypothesis of equal mean vectors. This suggests that olives from regions 2 and 4 possess distinct chemical compositions in terms of their fatty acid content.

To further understand which variables contributed most to this separation, a discriminant function was constructed. The resulting discriminant coefficients indicated that linolenic, eicosanoic, and eicosenoic acids were the strongest contributors to the separation of regions. In addition, our assessment of linoleic and linolenic acids confirmed their importance in contributing to the observed differences.

Limitations of this analysis include the fact that we do not know any bias that may come from collection techniques between two different organizations. Also, missing data adds unknowns to the analysis.

Future Work: Further investigation into specific fatty acids like linolenic, eicosanoic, and eicosenoic acids may provide more insights into the differences in olive profiles between the two regions.

Additionally, looking at other regions (e.g. 1 and 3) could provide additional information on regional differences. Finally, it would be helpful to look into the source of the missingness in the data.

Appendix

```r
pairwise_missing_fisher_p_values(oliver2a)
```

```
##             palmitic palmitoleic stearic  oleic linoleic eicosanoic linolenic
## palmitic          NA      1.0000  0.4231 0.3196   1.0000          1    1.0000
## palmitoleic   1.0000          NA  1.0000 1.0000   0.3354          1    1.0000
## stearic       0.4231      1.0000      NA 0.5008   1.0000          1    0.4231
## oleic         0.3196      1.0000  0.5008     NA   0.5008          1    1.0000
## linoleic      1.0000      0.3354  1.0000 0.5008       NA          1    0.4231
## eicosanoic    1.0000      1.0000  1.0000 1.0000   1.0000         NA    1.0000
## linolenic     1.0000      1.0000  0.4231 1.0000   0.4231          1        NA
## eicosenoic    1.0000      1.0000  1.0000 0.2487   0.3354          1    1.0000
##             eicosenoic
## palmitic        1.0000
## palmitoleic     1.0000
## stearic         1.0000
## oleic           0.2487
## linoleic        0.3354
## eicosanoic      1.0000
## linolenic       1.0000
## eicosenoic          NA
```

```r
pairwise_missing_fisher_p_values(oliver4a)
```

```
##             palmitic palmitoleic stearic  oleic eicosanoic linolenic eicosenoic
## palmitic          NA      0.2127  1.0000 1.0000     1.0000    0.1095     1.0000
## palmitoleic   0.2127          NA  1.0000 1.0000     1.0000    1.0000     0.5658
## stearic       1.0000      1.0000      NA 0.1619     1.0000    1.0000     0.4882
## oleic         1.0000      1.0000  0.1619     NA     1.0000    0.1095     1.0000
## eicosanoic    1.0000      1.0000  1.0000 1.0000         NA    1.0000     0.4882
## linolenic     0.1095      1.0000  1.0000 0.1095     1.0000        NA     1.0000
## eicosenoic    1.0000      0.5658  0.4882 1.0000     0.4882    1.0000         NA
```

```r
hotellings_with_imputed_datasets(oliver2a_imputed, mu2a_not)
```

```
## $mean
##    palmitic palmitoleic     stearic       oleic    linoleic  eicosanoic
##  1300.62651   121.30682   264.33664  7309.88804   818.26722    45.65122
##   linolenic  eicosenoic
##    63.53669    28.35521
##
## $total_variance
##                 palmitic palmitoleic     stearic        oleic    linoleic
## palmitic      74.8005379  3.11728705   1.2222254 -109.2972020   28.026665
## palmitoleic    3.1172871  7.59888951  -4.5248404   -0.9579184   -5.576883
## stearic        1.2222254 -4.52484040  23.4000436  -20.2102177    0.904422
## oleic       -109.2972020 -0.95791842 -20.2102177  306.5253762 -164.503016
## linoleic      28.0266649 -5.57688289   0.9044220 -164.5030159  137.396227
## eicosanoic    -0.5976989 -0.49355745   0.1927225   -2.9732275    2.686066
## linolenic      1.4292205  0.07734285  -0.9665461   -4.7879597    2.580195
## eicosenoic    -0.5995984 -0.20395873   0.1647232   -2.5896504    2.338393
##             eicosanoic   linolenic eicosenoic
## palmitic    -0.5976989  1.42922052 -0.5995984
## palmitoleic -0.4935575  0.07734285 -0.2039587
## stearic      0.1927225 -0.96654611  0.1647232
## oleic       -2.9732275 -4.78795969 -2.5896504
## linoleic     2.6860657  2.58019478  2.3383933
## eicosanoic   0.8291380  0.27792504  0.2119555
## linolenic    0.2779250  0.84316680  0.2701748
## eicosenoic   0.2119555  0.27017482  0.8021323
##
## $T2
## [1] 8.518872
##
## $F_stat
## [1] 0.9293315
##
## $p_value
## [1] 0.4981799
##
## $nu
## [1] 68.92341
```

```
hotellings_with_imputed_datasets(oliver4a_imputed, mu4a_not)
```

```
## $mean
##    palmitic palmitoleic     stearic       oleic    linoleic  eicosanoic
##  1229.57581   104.96935   273.18011  7358.73362   834.72222    42.39968
##   linolenic  eicosenoic
##    75.72519    37.67327
##
## $total_variance
##                palmitic palmitoleic       stearic       oleic     linoleic
## palmitic      885.95070  211.105636  -83.53692763 -1514.32152   425.740696
## palmitoleic   211.10564   67.198947  -29.16556923  -389.41197   115.032802
## stearic       -83.53693  -29.165569   62.77511971   159.46298  -106.455132
## oleic       -1514.32152 -389.411975  159.46297905  3170.71907 -1209.802538
## linoleic      425.74070  115.032802 -106.45513177 -1209.80254   694.083510
## eicosanoic    -11.47009   -3.973257    4.99368395    17.91368    -7.760370
## linolenic      36.13729    9.374599   -6.36828208   -72.39302    26.366467
## eicosenoic    -12.55200   -2.343926   -0.04328994    10.04935     1.650356
##              eicosanoic   linolenic   eicosenoic
## palmitic    -11.4700927  36.1372928 -12.55200269
## palmitoleic  -3.9732567   9.3745989  -2.34392621
## stearic       4.9936839  -6.3682821  -0.04328994
## oleic        17.9136812 -72.3930225  10.04935396
## linoleic     -7.7603704  26.3664669   1.65035564
## eicosanoic    0.6599370  -0.4075205   0.26996266
## linolenic    -0.4075205   3.4583579   0.37992001
## eicosenoic    0.2699627   0.3799200   1.55532218
##
## $T2
## [1] 16.5587
##
## $F_stat
## [1] 1.65587
##
## $p_value
## [1] 0.1223715
##
## $nu
## [1] 79.81564
```

Region 2 Historical Means

$$\boldsymbol{\mu}_{02} = (1300,\ 120,\ 265,\ 7310,\ 820,\ 45,\ 65,\ 28),$$

Region 4 Historical Means

$$\boldsymbol{\mu}_{04} = (1230,\ 105,\ 275,\ 7360,\ 830,\ 41,\ 75,\ 38),$$

Comparison between Regions 2 and 4

```r
cat("Hotelling's T^2 Statistic:", as.numeric(t_squared), "\n") # 189.7534
```

```
## Hotelling's T^2 Statistic: 189.7534
```

```r
cat("F-statistic:", as.numeric(f_statistic), "\n") # 21.87435
```

```
## F-statistic: 21.87435
```

```r
cat("Numerator degrees of freedom:", df1, "\n")
```

```
## Numerator degrees of freedom: 2
```

```r
cat("Denominator degrees of freedom:", df2, "\n")
```

```
## Denominator degrees of freedom: 83
```

```r
cat("P-value:", as.numeric(p_value), "\n")
```

```
## P-value: 3.874678e-14
```

Assessment of Linoleic and Linolenic Acids

```r
paste("F-stat for region effect:", Fs)
```

```
## [1] "F-stat for region effect: 28.2676278919721"
```

```r
paste("df1:", df1, "df2:", df2)
```

```
## [1] "df1: 2 df2: 83"
```

```r
paste("P-value:", 1-pf(Fs, df1, df2))
```

```
## [1] "P-value: 4.33867164417734e-10"
```

Table 2: Discriminant Function Coefficients for Conditional Wilk's Lambda Test

Variable	Coefficient
palmitic	5.3399
palmitoleic	2.4534
stearic	4.3190
oleic	14.4057

linoleic	7.8321
eicosanoic	-1.7658
linolenic	3.7817
eicosenoic	1.2415