Assessing Outliers and Multivariate Normality in Fatty Acids

Stephen Merrill & Maddie Phan

Introduction

A critical and initial step in data analysis is evaluating whether the data meets the assumptions associated with corresponding methods. Violation of the assumptions may affect and potentially invalidate the results. Assumptions often include: distribution, linearity and, independence. The goal of this analysis is to assess whether 8 fatty acids collected on olives in a particular agricultural region follow a multivariate normal distribution, as well as identify potential outliers. Evaluating distributional assumptions will help us to move forward with potential analyses with confidence. Additionally, outliers that are a result of miskeyed information can be identified easier after distributional assumptions are made.

Data & Methods

In order to assess the claim of multivariate normality, we first assessed univariate normality. Even though univariate normality does not necessarily mean that all the variables are jointly normal, it is the first step. Univariate normality is assessed using histograms and QQ-plots. Figure 1 contains histograms of fatty acids. Note that Oelic and Linoleic are slightly left skewed whereas other acids such as Eicosenoic are much dramatically skewed right skewed. The skewness among all the acids indicate that each acid may not be distributed normally.

Additionally, we assessed bivariate normality using pairs plots to explore relationships between pairs of acids. The relationship between Oelic and Linoleic is curved which creates doubts about bivariate normality as well. These initial concerns about univariate and bivariate normality suggest that we will need to utilize a transformation to support multivariate normality. This transformation took the form recommended by Box and Cox (1964) and given below in (1).

$$x_i^{(\lambda)} = \begin{cases} \dfrac{x_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \\ \ln(x_i) & \text{if } \lambda = 0, \end{cases} \tag{1}$$

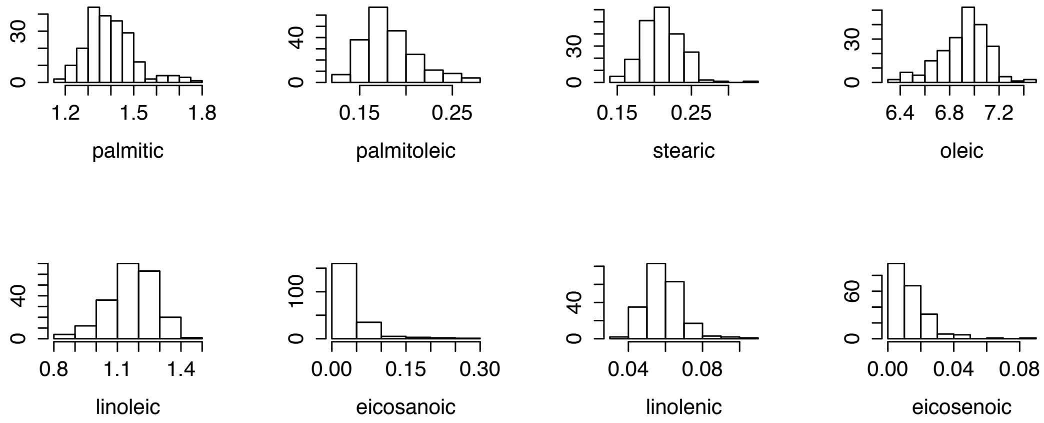

Figure 1: Histograms of Fatty Acids

We first applied the transformation univariately, and then used the resulting λ values in an iterative algorithm that maximized (2) in order to obtain the λ values for use in a multivariate transformation.

$$l(\boldsymbol{\lambda}) = -\frac{n}{2} ln|\boldsymbol{S_{\lambda}}| + \sum_{j=1}^{p} \left[(\lambda_j - 1) \sum_{i-1}^{n} ln(x_{ij}) \right] \tag{2}$$

Once those λ values were obtained, we identified potential outliers by calculating a standardized distance D_i^2 given in (3). This is a distance measurement from the transformed data to each column mean. Then, we found which observations have the largest distances. However, this method only identified the observation with a potential outlier, not the specific acid. In order to further identify the outlier, we found the sampling distribution of each acid and then found whether each acid observation was farther than 3 standard deviations from the mean. This was appropriate as we were looking at the data on the transformed scale and therefore under an assumption of normality.

$$D_i^2 = (\boldsymbol{x_i} - \bar{\boldsymbol{x}})' \boldsymbol{S}^{-1} (\boldsymbol{x_i} - \bar{\boldsymbol{x}}) \tag{3}$$

Results

Initially, we applied a Box-Cox transformation to each variable independent of the other 7 acids. Table 1 shows these λ values, which are a starting point for the multivariate box-cox transformation. A positive λ is correcting for left skewness and negative λ values correct right skewness. For example, Figure 1 shows that Palmitic is right skewed and the box-cox transformation returned a lambda of -2.17. We also used QQ-plots to assess normality after the univariate transformations. They are not included, but all plots look mostly linear with some concerns about univariate normality in the tails. Figure 2 shows a diagnostic χ^2 quantile plot of the data transformed using the univariate λ values. The χ^2 quantile plot is used to assess the multivariate normality and also has concerning tail behavior.

Fatty Acids	Palmitic	Palmitoleic	Stearic	Oleic	Linoleic	Eicosanoic	Linolenic	Eicosenoic
Optimal λ	-2.17	-1.34	-0.29	6.53	2.63	-0.32	-0.04	-0.11

Table 1: Optimal λ by Acid Using Univariate Box-Cox Procedure

Table 2 contains the λ for the multivariate box-cox transformations. These values seem like reasonable adjustments from the univariate λ values, although a couple larger alterations that include sign changes, most notably Linolenic Acid, may be met with skepticism. In order to assess the performance of the resulting multivariate transformation of the data, we include Figure 3, the Beta quantile plot proposed by Gnandadesikan and Kettenring (1972).

Fatty Acids	Palmitic	Palmitoleic	Stearic	Oleic	Linoleic	Eicosanoic	Linolenic	Eicosenoic
Optimal λ	-0.327	-0.214	0.154	3.822	-0.178	-0.329	0.245	-0.081

Table 2: Optimal λ by Acid Using Multivariate Box-Cox Procedure

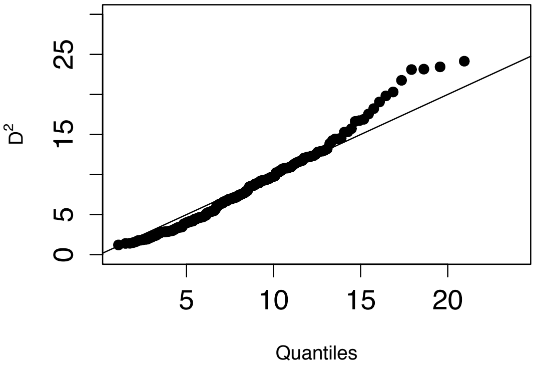

Figure 2: χ^2 QQ plot for Univariate Box-Cox

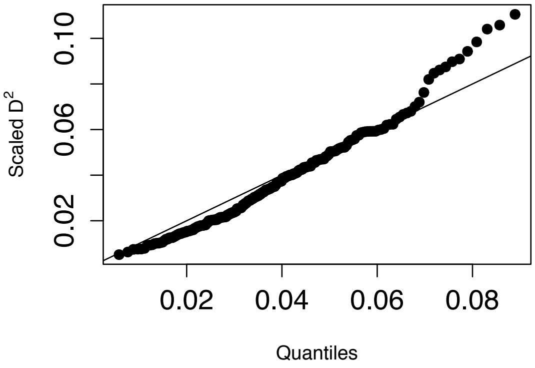

Figure 3: Beta QQ plot for Multivariate Box-Cox

These diagnostic plots are presented with notable outliers removed, since an inital check showed evidence of outliers creating apparent departures from normality. We calculated standardized distances for each observation as detailed previously and found one observation that had a significant D^2 value and several others that had elevated values. Examination of the diagnostic plots showed three points that were causing notable problems with normality, each of which had D^2 values that were previously noted. In order to then determine which individual acid measurements in these three observations might be outliers, we checked each to see if they were outside a range of three standard deviations from the acid mean. This approach is motivated by the assumption that the data has been transformed to be normal. Three points violated this check, and are therefore our guesses at a miskeyed observation: (1) Observation 40, Palmitic Acid; (2) Observation 40, Linoleic Acid; (3) Observation 176, Stearic Acid.

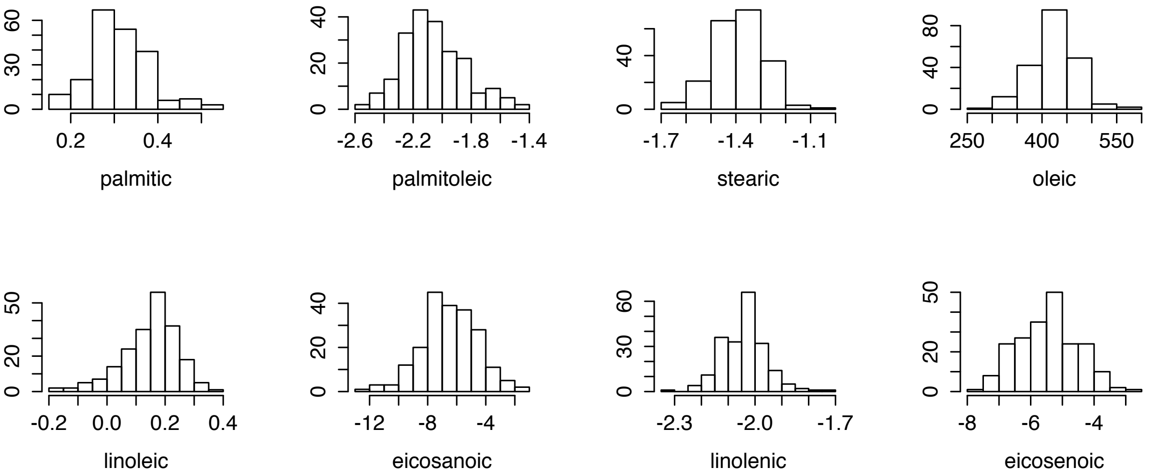

Figure 4: Histograms of After Multivariate Transformations

Conclusions

This olive acid data was clearly shown to be skewed and therefore required a transformation in order to be multivariate normally distributed. We successfully employed both univariate and multivariate Box-Cox transformations on the data and showed that the resulting data was multivariate normally distributed. After transforming the data, it became obvious that outliers were skewing the distribution. We calculated standardized distances and were able to identify three possible locations where an observation might have been miskeyed.

Although our analysis was successful, we have some reservations about the multivariate transformation. The adjustment from the univariate λ values in the Box-Cox transformation to the multivariate values was larger than anticipated, and included some questionable sign changes. The multivariately transformed values produce univariate histograms that are more skewed than the histograms produced by the univariate transformation. The quantile plots from both transformations look appropriate, but the slight skew in the data after the multivariate transformation is a clear shortcoming in our analysis. Nevertheless, we believe the transformation to appropriately approximate multivariate normality and allow for the identification of outliers.

Appendix

Code

```
dat <- read.table('https://tofu.byu.edu/stat666/datasets/oliver3b.txt', header = TRUE)
##################################################
#How to check for Normality
##################################################
#lambda
lam <- function(x, lambda){
  lambs <- matrix(0, ncol=length(lambda), nrow=length(x))
  for(i in 1:length(lambda)){
    if(lambda[i] != 0){
      lambs[,i] <- (x^(lambda[i])-1)/(lambda[i])
    }else{
      lambs[,i] <-log(x)
    }
  }
  return(lambs)
}
#Box-Cox likelihood
boxy <- function(x, lambda){
  lambdas <- lam(x, lambda)
  like <- numeric(length(lambda))
  for(i in 1:ncol(lambdas)){
    s2lam <- (1/length(x))*sum((lambdas[,i]-mean(lambdas[,i]))^2)
    like[i] <- -length(x)/2*log(s2lam) + (lambda[i]-1)*sum(log(x))
  }
  return(like)
}
##############################################
#MULTIVARIATE BOXCOX
##############################################
multivbc <- function(evals) {
  n <- 206
  evals2 <- numeric(8)
  MLE <- matrix(0, ncol = 3, nrow = 8)
  for(i in 1:length(evals)){
    eigs <- zapsmall(c(evals[i], evals[i] + .001, evals[i] -0.001))
    dats <- matrix(0, ncol =8, nrow = 206)
    for(j in 1:8){
      dats[,j]<- lam(dat[,j], evals[j])
    }
    #test which of the 3 lambda is best
    for(k in 1:3){
      evals[i] <- eigs[k]
      dats[,i] <- lam(dat[,i],eigs[k])
      #estimate S with cov
      #Plug into MLE
      MLE[i,k] <- -n/2 * log(det(cov(dats))) + sum((evals-1)*apply(dat,2, function(x) sum(lo
    }
    evals2[i] <- eigs[which.max(MLE[i,])]
  }
  evals2
}
##################################################
#Get univariate lambdas
seqs <- seq(-10, 10, by = 0.01)
evals <- numeric(8)
for(i in 1:8){
  temp <- boxy(dat[,i], seqs)
  evals[i] <- seqs[which.max(temp)]
}
first <- evals
#Iterate until all 8 rows don't change
comp <- rep(FALSE,8)
iter <- 0
```

```
while(!all(comp)) {
  current <- multivbc(first)
  first <- multivbc(current)
  comp <- first==current
  iter <- iter+1
}
#######################################################
#How to identify outliers
#######################################################
dat.trans <- dat
for(i in 1:ncol(dat)) {
  dat.trans[,i] <- (dat[,i]^first[i]-1)/first[i]
}
D2 <- numeric(nrow(dat))
Sinv <- solve(cov(dat.trans))
for(i in 1:nrow(dat)) {
  D2[i] <- t(t(dat.trans[i,]-colMeans(dat.trans)))%*%Sinv
    %*%t(dat.trans[i,]-colMeans(dat.trans))
}
#Beta qq plot
n <- nrow(dat)
p <- ncol(dat)
a <- p/2
b <- (n-p-1)/2
alpha <- (p-2)/(2*p)
beta <- (n-p-3)/(2*(n-p-1))
v <- qbeta(((1:n) - alpha)/(n-alpha-beta+1),a,b)
u <- (n*D2)/((n-1)^2)
par(mfrow=c(1,1))
plot(v,sort(u),ylab=expression(paste("Scaled ", D^2)),xlab="Quantiles")
abline(0,1)
```