Multivariate Analysis HW03

Luke Beebe

2024-02-11

MVN

Use the methods described in the manuscript RJ-2014-031_2.pdf from Lecture 02 files and the versicolor data set created in HW01 to assess the bivariate normality of Sepal.Length and Sepal.Width. Repeat for Petal.Length and Petal.Width.

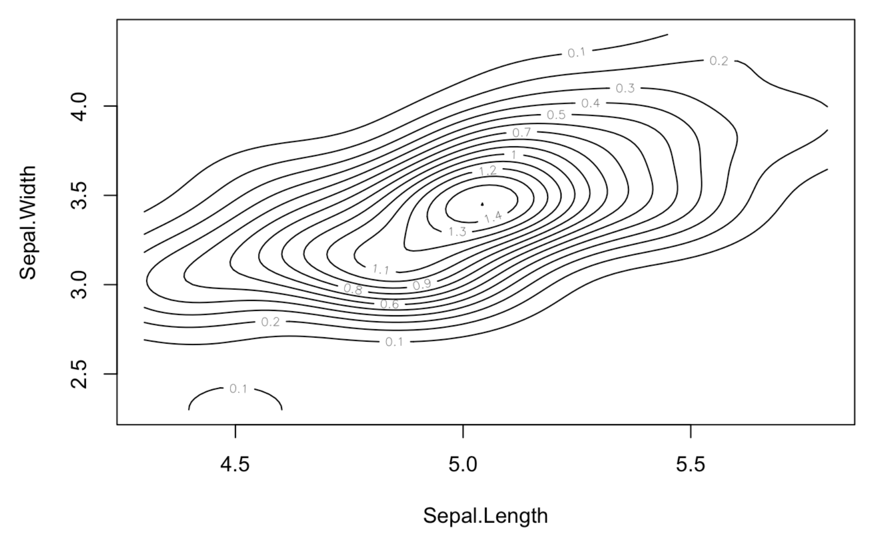
```
setosa1 <- iris[1:50, 1:2]
setosa2 <- iris[1:50, 3:4]
```

```
results <- NULL
tests <- c("mardia", "hz", "royston", "dh", "energy")
for(test in tests){
  p <- mvn(data=setosa1, mvnTest=test)$multivariateNormality
  print(paste(test, na.omit(p$`p value`)))
}
```

```
## [1] "mardia 0.943793240544736" "mardia 0.925538081956865"
## [1] "hz 0.914633595525848"
## [1] "royston 0.24457373120284"
## [1] "dh 0.0208564184084919"
## [1] "energy 0.805"
```

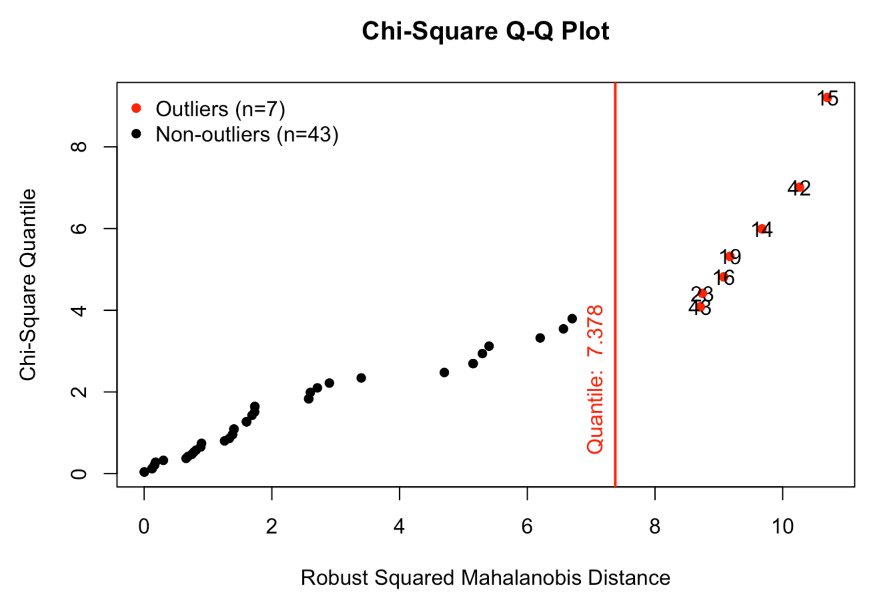
Only Doornik-Hansen rejects the null of a multivariate distribution at a significance level 0.05. Let's check the bivariate contour plot to see what the data looks like.

```
setosa1_contour <- mvn(setosa1, multivariatePlot = "contour")
```

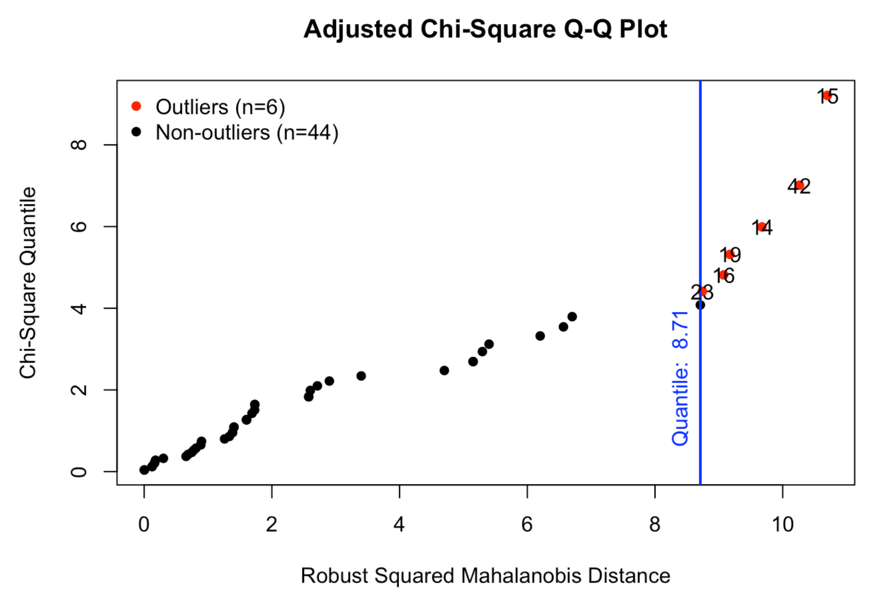


The plot shows an outlier in the bottom left corner of the plot. Lets see if there's statistical significance to the outliers.

```
setosa1_outlier <- mvn(data=setosa1, multivariateOutlierMethod="quan")
```




```
setosa1_outlier_adj <- mvn(data=setosa1, multivariateOutlierMethod="adj")
```



Both the quantile and adjusted quantile plots show multivariate outliers. Because of this we cannot assume bivariate normality.

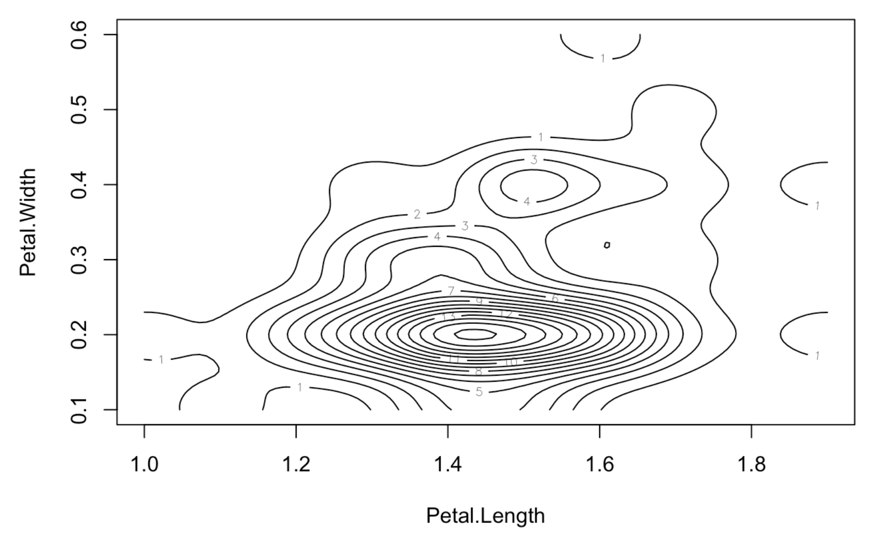
Let's check the second setosa set.

```
results <- NULL
tests <- c("mardia", "hz", "royston", "dh", "energy")
for(test in tests){
  p <- mvn(data=setosa2, mvnTest=test)$multivariateNormality
  print(paste(test, na.omit(p$`p value`)))
}
```

```
## [1] "mardia 0.0134389894271658" "mardia 0.135415238392804"
## [1] "hz 0.00139475678690615"
## [1] "royston 4.18884528186296e-07"
## [1] "dh 1.93142414990043e-10"
## [1] "energy 0"
```

Of the tests, only one fails to reject the null: Mardia Kurtosis. Let's look at a contour plot to see what this looks like.

```
setosa1_contour <- mvn(setosa2, multivariatePlot = "contour")
```



Using the tests as inference, we can see how this contour plot is not normal. There is a dense space towards the bottom, but a lot of erratic variance elsewhere. We cannot assume the normality of this set.

3.14

Consider the data matrix X in Exercise 3.1. We have n=3 observations on p=2 variables X1 and X2. Form the linear combinations

```
X <- matrix(c(9,1,5,3,1,2), nrow=3, ncol=2, byrow=T)
b <- c(2,3)
c <- c(-1, 2)
```

a. Evaluate the sample means, variances, and covariance of b'X and c'X from first principles. That is, calculate the observed values of b'X and c'X, and then use the ample mean, variance, and covariance formulas.

```
bX <- b %*% t(X)
cX <- c %*% t(X)
mean(bX)
```

```
## [1] 16
```

```
var(t(bX))[1,]
```

```
## [1] 49
```

```
mean(cX)
```

```
## [1] -1
```

```
var(t(cX))[1,]
```

```
## [1] 28
```

```
cov(t(bX), t(cX))[1,]
```

```
## [1] -28
```

mean(bX) = 16 var(t(bX)) = 49 mean(cX) = -1 var(t(cX)) = 28 cov(t(bX), t(cX)) = -28

b. Calculate the sample means, variances, and covariance of b'X and c'X using (3-36). Compare the results in (a) and (b).

```
X_ <- apply(X, 2, mean)
S <- cov(X)
(b %*% X_)[1,]
```

```
## [1] 16
```

```
(b %*% S %*% b)[1,]
```

```
## [1] 49
```

```
(c %*% X_)[1,]
```

```
## [1] -1
```

```
(c %*% S %*% c)[1,]
```

```
## [1] 28
```

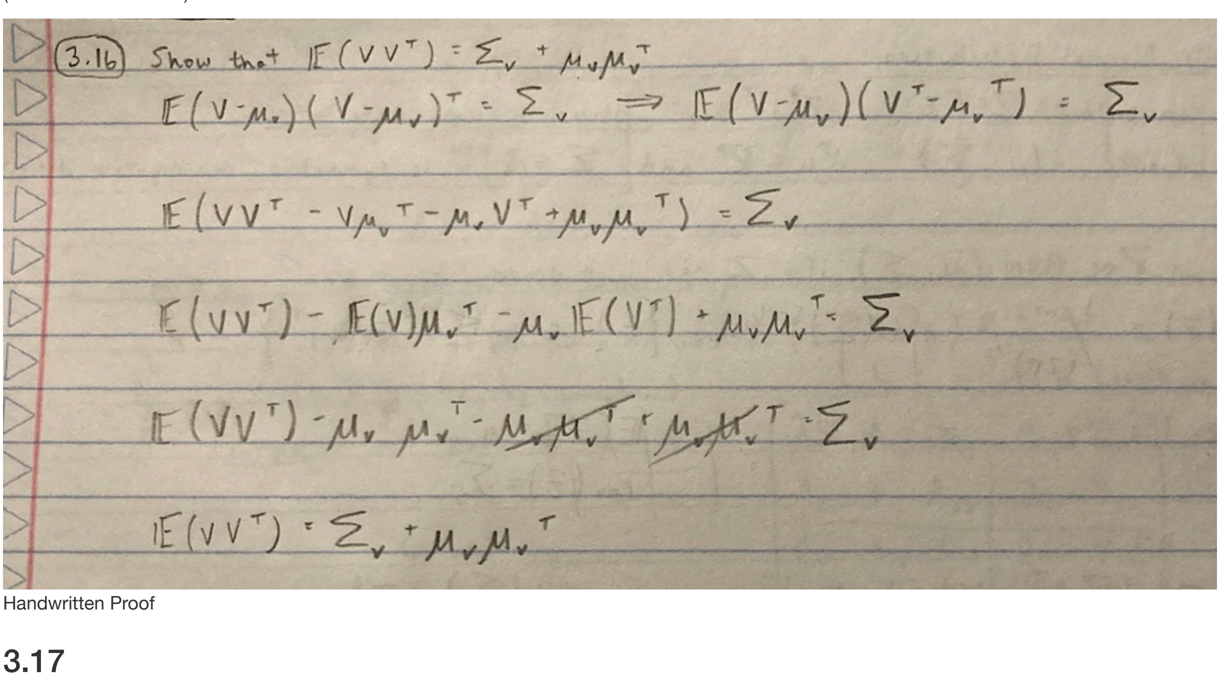
```
(b %*% S %*% c)[1,]
```

```
## [1] -28
```

The results are the same between (a) and (b).

3.16

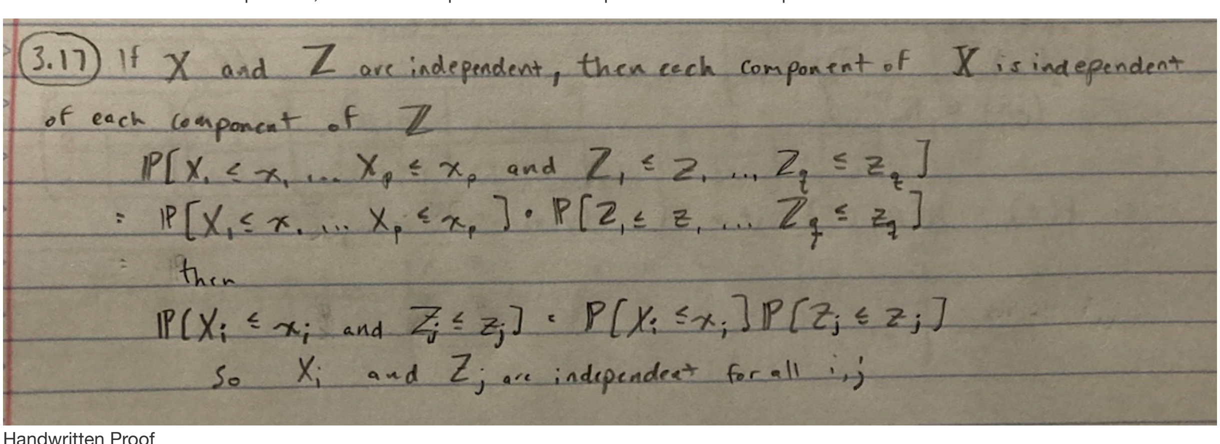
Let *V* be a vector random variable with mean vector E(*V*)=u and covariance matrix = E(*V*-u) * t(*V*) Show that E(*V* * t(*V*)) = CovMatrix + u * t(u) (Second moment of *V*)



Handwritten Proof

3.17

Show that if *X* and *Z* are independent, then each component of *X* is independent of each component of *Z*.



Handwritten Proof

3.18

Energy consumption in 2001, by state, from the major sources is recorded in quadrillions of BTUs. The resulting mean and covariance matrix are:

```
x_ <- c(0.766, 0.508, 0.438, 0.161)
s <- matrix(c(0.856, 0.635, 0.173, 0.096,
              0.635, 0.568, 0.128, 0.067,
              0.173, 0.127, 0.171, 0.039,
              0.096, 0.067, 0.039, 0.043),
            nrow=4, ncol=4, byrow=T)
```

a. Using the summary statistics, determine the sample mean and variance of a state's total energy consumption for these major sources

```
print("mean(total energy consumption)")
```

```
## [1] "mean(total energy consumption)"
```

```
sum(x_)
```

```
## [1] 1.873
```

```
print("var(total energy consumption)")
```

```
## [1] "var(total energy consumption)"
```

```
sum(s)
```

```
## [1] 3.913
```

b. Determine the sample mean and variance of the excess of petroleum consumption over natural gas consumption. Also find the sample covariance of this variable with the total variable in part (a).

```
print("petroleum - natural gas")
```

```
## [1] "petroleum - natural gas"
```

```
x_[1]-x_[2]
```

```
## [1] 0.258
```

```
print("var(petroleum - natural gas)")
```

```
## [1] "var(petroleum - natural gas)"
```

```
b <- s[1,1]+s[2,2]-2*s[1,2]
print(b)
```

```
## [1] 0.154
```

```
print("cov(petroleum - natural gas, total energy consumption)")
```

```
## [1] "cov(petroleum - natural gas, total energy consumption)"
```

```
matrix(c(b, (b*sum(s))^2, (sum(s)*b)^2, sum(s)),
       ncol=2)
```

```
##           [,1]      [,2]
## [1,] 0.1540000 0.3631292
## [2,] 0.3631292 3.9130000
```