Homework 5

Evan Yacek ety78

This homework is due on Mar. 1, 2016 at 11:59pm. Please submit as a PDF file on Canvas.

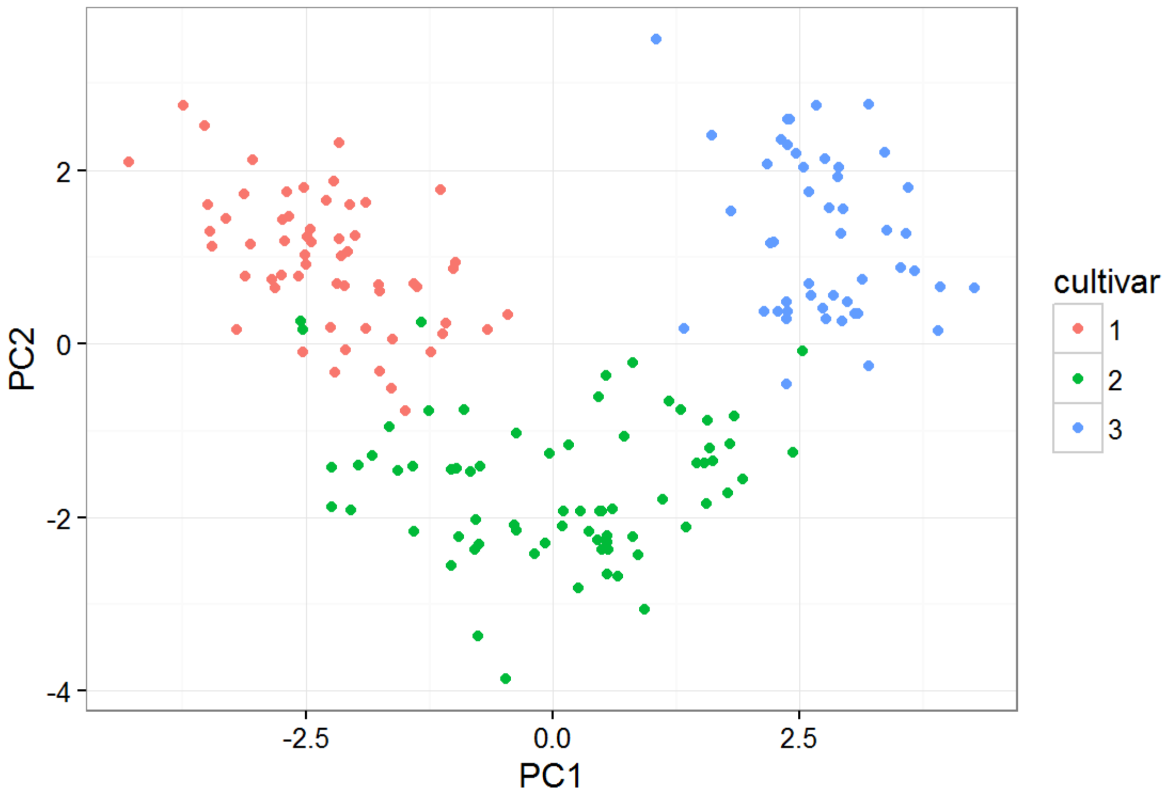
For this homework you will use the `wine` data set. The `wine` data set contains concentrations of 13 different chemical compounds (`chem1` - `chem13`) in 178 samples of wines grown in Italy. Each row is a different sample of wine, and the data set contains three different cultivars (`cultivar`) of wine.

```
wine <- read.csv("http://wilkelab.org/classes/SDS348/data_sets/wine.csv", colClasses = c("cultivar" =
"factor"))
head(wine)
```

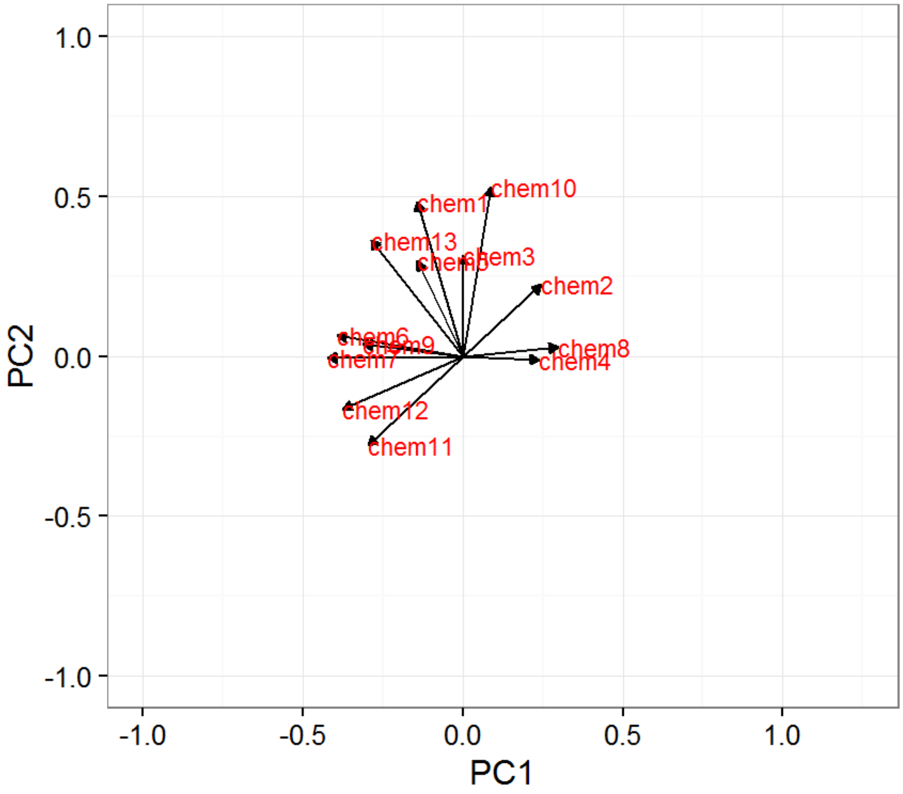
```
##   cultivar chem1 chem2 chem3 chem4 chem5 chem6 chem7 chem8 chem9 chem10
## 1        1 14.23  1.71  2.43  15.6   127  2.80  3.06  0.28  2.29   5.64
## 2        1 13.20  1.78  2.14  11.2   100  2.65  2.76  0.26  1.28   4.38
## 3        1 13.16  2.36  2.67  18.6   101  2.80  3.24  0.30  2.81   5.68
## 4        1 14.37  1.95  2.50  16.8   113  3.85  3.49  0.24  2.18   7.80
## 5        1 13.24  2.59  2.87  21.0   118  2.80  2.69  0.39  1.82   4.32
## 6        1 14.20  1.76  2.45  15.2   112  3.27  3.39  0.34  1.97   6.75
##   chem11 chem12 chem13
## 1   1.04   3.92   1065
## 2   1.05   3.40   1050
## 3   1.03   3.17   1185
## 4   0.86   3.45   1480
## 5   1.04   2.93    735
## 6   1.05   2.85   1450
```

Question 1 (3 pts): *Perform a principal components analysis (PCA). Since the chemical concentrations may span several orders of magnitude across different compounds, be sure to* ***scale*** *the data (using `scale`) before doing PCA. Create a scatterplot of PC1 vs. PC2 and color each point by cultivar. What do you observe? Visually, and without doing any calculations, do the cultivars cluster together in principal-component space?*

```
wine %>% select(-cultivar) %>%
  scale() %>%
  prcomp() ->
  pca
wine.pca <- data.frame(wine, pca$x)
ggplot(wine.pca, aes(x=PC1, y=PC2, color=cultivar)) + geom_point()
```



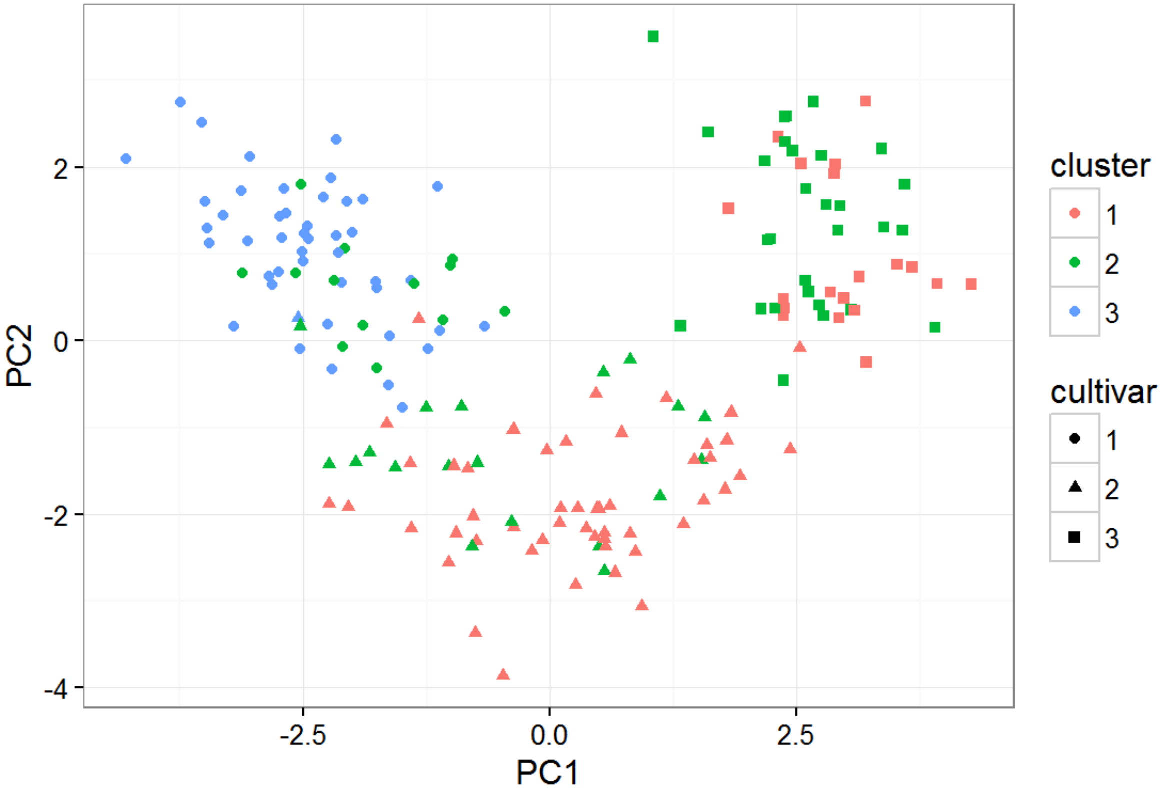
```
# capture the rotation matrix in a data frame
rotation_data <- data.frame(pca$rotation, variable=row.names(pca$rotation))
# define a pleasing arrow style
arrow_style <- arrow(length = unit(0.05, "inches"),
                     type = "closed")
# now plot, using geom_segment() for arrows and geom_text for labels
ggplot(rotation_data) +
  geom_segment(aes(xend=PC1, yend=PC2), x=0, y=0, arrow=arrow_style) +
  geom_text(aes(x=PC1, y=PC2, label=variable), hjust=0, size=3, color='red') +
  xlim(-1.,1.25) +
  ylim(-1.,1.) +
  coord_fixed() # fix aspect ratio to 1:1
```



The cultivars do not cluster together in principal component space. Each cultivar seems to be particularly distinct from one another, not having much overlap in their distribution.

Question 2 (4 pts): *Now take your matrix of **principal components coordinates** (not the raw chemical concentration data!) from Question 1 above and cluster the wines into 3 groups (`centers=3`) using k-means clustering with 10 random starts (`nstart=10`). Create a scatterplot of PC1 vs. PC2. This time, color each point by **cluster** and set the plotting symbol by **cultivar**. What do you observe?*

```
wine.pca %>% select(-cultivar,-chem1,-chem2,-chem3,-chem4,-chem5,-chem6,-chem7,-chem8,-chem9,-chem10,
-chem11,-chem12,-chem12) %>% kmeans(centers=3, nstart=10) -> km
wine_clustered <- data.frame(wine.pca, cluster=factor(km$cluster))
ggplot(wine_clustered, aes(x=PC1, y=PC2, color=cluster, shape=cultivar)) + geom_point()
```

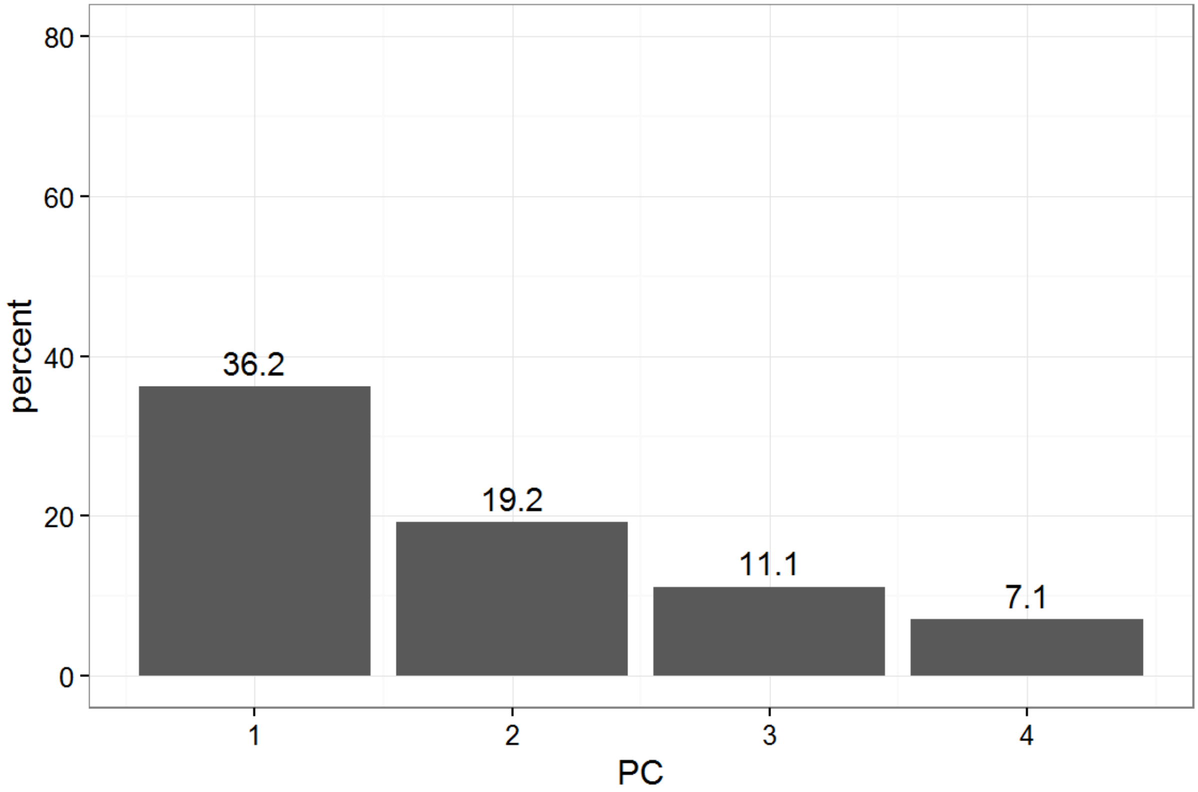


The majority of Cultivar 2 was clustered in group 2. Meanhwile, Cultivar 3 was clustered in groups 3 and 2. Finally Cultivar 1 primarirly clustered in group 1 and 3.

Question 3 (3 pts): *Create a bar plot that shows the percent variance explained by each principal component. State how much variance is explained by each of the principal components 1 through 4.*

```
percent <- 100*pca$sdev^2/sum(pca$sdev^2)


perc_data <- data.frame(percent=percent, PC=1:length(percent))
newPerc <- head(perc_data, 4)
ggplot(newPerc, aes(x=PC, y=percent)) +
  geom_bar(stat="identity") +
  geom_text(aes(label=round(percent, 1)), size=4, vjust=-.5) +
  ylim(0, 80) +
  scale_x_continuous(breaks=1:9) # make sure each PC gets an axis tick
```



The first component explains 36.2 percent of the variance, the second 19.2 percent, the third 11.1 percent, and finally the fourth 7.1 percent.