Homework 1 solution

Due: 2/4/2019 before 11pm. Submit in Canvas (file upload). Rmd file and the html output file (submit both files) are strongly recommended, but not required.

1. (10 pts)

Let $X = (X_1, \dots, X_n)'$ be the $n \times p$ data matrix, where $X_i = (X_{i1}, \dots, X_{ip})'$ is the ith onservation. Let $\bar{X} = n^{-1}\sum_{i=1}^n X_i$ be the sample mean. Let $s_{j_1j_2} = n^{-1}\sum_{i=1}^n (X_{ij_1} - \bar{X}_{j_1})(X_{ij_2} - \bar{X}_{j_2})$ be the sample covariance between the j_1th and j_2th variables. Let $S = (s_{j_1j_2})$ be the sample covariance matrix. Show that

$$S = \frac{1}{n}X'X - \bar{X}'\bar{X}.$$

Proof: Let $A = \frac{1}{n}X'X - \bar{X}'\bar{X}$. The (j_1, j_2)th element of A is

$$a_{j_1,j_2} = \frac{1}{n}\sum_{i=1}^n X'_{j_1i}X_{ij_2} - \bar{X}_{j1}\bar{X}_{j2} = \frac{1}{n}\sum_{i=1}^n X_{ij_1}X_{ij_2} - \bar{X}_{j1}\bar{X}_{j2}.$$

On the other hand,

$$s_{j_1j_2} = \frac{1}{n}\sum_{i=1}^n (X_{ij_1} - \bar{X}_{j_1})(X_{ij_2} - \bar{X}_{j_2}) = \frac{1}{n}\sum_{i=1}^n X_{ij_1}X_{ij_2} - \frac{1}{n}\sum_{i=1}^n \bar{X}_{j_1}\bar{X}_{j_2} = \frac{1}{n}\sum_{i=1}^n X_{ij_1}X_{ij_2} - \bar{X}_{j_1}\bar{X}_{j_2}.$$

Therefore, $a_{j_1,j_2} = s_{j_1j_2}$, and $S = \frac{1}{n}X'X - \bar{X}'\bar{X}$.

2. (10 pts)

Find ALL the eigenvalues and their eigenvectors for the following matrices

- $\Sigma = \sigma\mathbb{1}\mathbb{1}'$ where $\mathbb{1} = (1, 1, \dots, 1)'$ is the p dimensional vector of 1.

  Solution: Since $\text{rank}(\Sigma) = 1$, we have $\ker(\Sigma) = p - 1$. Thus, $\lambda = 0$ is a root of the characteristic polynomial with multiplicity at least $p - 1$. On the other hand, $tr(\Sigma) = p\sigma \neq 0$. Thus, Σ has an eigenvalue $p\sigma$. So Σ has eigenvalue 0 with multiplicity $p - 1$ and eigenvalue $p\sigma$ with multiplicity 1. The eigenvectors corresponding to 0 solves the equation $\{\boldsymbol{x} = (x_1, x_2, \dots, x_p)' | x_1 + x_2 + \cdots + x_p = 0\}$. One possible solution system would be

  $$\left\{ \begin{bmatrix} 1 \\ -1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ -1 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \dots, \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ -1 \end{bmatrix} \right\}.$$

  The eigenvector corresponding to $p\sigma$ can be chosen as $\mathbb{1}$.

- $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_p)$ is a diagonal matrix.

  Solution: We have $\Sigma = \text{diag}\{\sigma_1, \dots, \sigma_p\} = \sigma_1\boldsymbol{e}_1\boldsymbol{e}_1' + \sigma_2\boldsymbol{e}_2\boldsymbol{e}_2' + \cdots + \sigma_p\boldsymbol{e}_p\boldsymbol{e}_p'$ where $\boldsymbol{e}_i$ is the p dimensional vector with the ith coordinate being 1 and others being 0 for $i = 1, 2, \dots, p$. Since $\{\boldsymbol{e}_i\}$ are orthogonal, the eigenvalues of Σ are $\{\sigma_1, \sigma_2, \dots, \sigma_p\}$ and the corresponding eigenvectors are $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_p\}$.

3. (10 pts)

Show that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.

Proof: Let $A = (a_{ij})_{p\times q}$ and $B = (b_{ij})_{q\times p}$. (Note that A and B don't require to be square matrices.) Without loss generality, assume $p \le q$.

$$\mathrm{tr}(AB) = \sum_{i=1}^{p}\sum_{j=1}^{q} a_{ij}b_{ji} = \sum_{j=1}^{q}\sum_{i=1}^{p} b_{ji}a_{ij} = \mathrm{tr}(BA).$$

4. (10 pts)

Given two variables in the data matrix X (say the j_1th and j_2th variables). Show that their sample correlation will not change by standardization.

Proof: Let $x_{(j_1)}, x_{(j_2)}$ be the j_1th and j_2th variables and $\tilde{x}_{(j_1)}, \tilde{x}_{(j_2)}$ be the standardization of $x_{(j_1)}$ and $x_{(j_2)}$. Then, $\tilde{x}_{(j_1)} = \frac{1}{\sqrt{s_{j_1j_1}}}(x_{(j_1)} - \bar{x}_{j_1}\mathbb{1})$ and $\tilde{x}_{(j_2)} = \frac{1}{\sqrt{s_{j_2j_2}}}(x_{(j_2)} - \bar{x}_{j_2}\mathbb{1})$. The sample correlation between $\tilde{x}_{(j_1)}$ and $\tilde{x}_{(j_2)}$ is

$$\begin{aligned}\tilde{r}_{j_1j_2} &= \frac{1}{n}\sum_{i=1}^{n}\tilde{x}_{ij_1}\tilde{x}_{ij_2} = \frac{1}{n}\sum_{i=1}^{n}\frac{x_{ij_1}-\bar{x}_{j_1}}{\sqrt{s_{j_1j_1}}}\frac{x_{ij_2}-\bar{x}_{j_2}}{\sqrt{s_{j_2j_2}}} \\ &= \frac{1}{\sqrt{s_{j_1j_1}}\sqrt{s_{j_2j_2}}}\frac{1}{n}\sum_{i=1}^{n}(x_{ij_1}-\bar{x}_{j_1})(x_{ij_2}-\bar{x}_{j_2}) \\ &= \frac{s_{j_1j_2}}{\sqrt{s_{j_1j_1}}\sqrt{s_{j_2j_2}}} = r_{j_1j_2}.\end{aligned}$$

5. (10 pts)

Given a data matrix X as in Question 1. Assume the means of the p variables are zero. Let $S = \frac{1}{n}X'X$ be the sample covariance matrix. Let $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_p$ be the ordered eigenvalues of S. Let $e_1, \ldots, e_p$ be their corresponding orthogonal eigenvectors with unit length. In multivariate analysis, we usually want to use the first few eigenvalues and eigenvectors to represent the original data, as a tool of dimension reduction.

- On one aspect, let
$$S_m = \lambda_1 e_1 e_1' + \ldots + \lambda_m e_m e_m'$$
be an approximate of S for $m < p$. Calculate $\mathrm{tr}\{(S - S_m)^2\}$ and $\mathrm{tr}\{(S - S_m)^2\}/\mathrm{tr}(S^2)$, where $\mathrm{tr}(S^2)$ can be regarded as the total variation of the data.

  Solution: Note that $\mathrm{tr}\{(S - S_m)^2\} = \mathrm{tr}(S^2) + \mathrm{tr}(S_m^2) - 2\mathrm{tr}(SS_m)$, and $\mathrm{tr}(S^2) = \sum_{i=1}^{p}\lambda_i^2$, $\mathrm{tr}(S^2) = \sum_{i=1}^{m}\lambda_i^2$. We also have $\mathrm{tr}(SS_m) = \mathrm{tr}(S_m^2)$ Therefore,
$$\frac{\mathrm{tr}\{(S - S_m)^2\}}{\mathrm{tr}(S^2)} = \frac{\sum_{i=m+1}^{p}\lambda_i^2}{\sum_{i=1}^{p}\lambda_i^2}.$$

- On another aspect, $\{Xe_1, \ldots, Xe_m\}$ are the transformed data by the eigenvectors. Calculate the sample covariance S_t of $\{Xe_1, \ldots, Xe_m\}$ (regard the sample mean as 0). What is $\mathrm{tr}(S_t^2)$ comparing to $\mathrm{tr}(S^2)$?

Solution: Note that $\{Xe_1, \dots, Xe_m\} = X(e_1, \dots, e_m)$. This leads to

$$\begin{aligned}
S_t &= \frac{1}{n}\left(X(e_1,\dots,e_m)\right)' X(e_1,\dots,e_m) \\
&= \frac{1}{n}\begin{pmatrix} e_1' \\ e_2' \\ \vdots \\ e_m' \end{pmatrix} X'X(e_1,e_2,\dots,e_m) = \begin{pmatrix} e_1' \\ e_2' \\ \vdots \\ e_m' \end{pmatrix} S(e_1,e_2,\dots,e_m) \\
&= \begin{pmatrix} e_1' \\ e_2' \\ \vdots \\ e_m' \end{pmatrix} (\lambda_1 e_1 e_1' + \cdots + \lambda_p e_p e_p')(e_1,\dots,e_m) = \begin{pmatrix} \lambda_1 e_1' \\ \lambda_2 e_2' \\ \vdots \\ \lambda_m e_m' \end{pmatrix} (e_1,e_2,\dots,e_m) \\
&= \mathrm{diag}(\lambda_1,\dots,\lambda_m).
\end{aligned}$$

Therefore, $\mathrm{tr}(S_t^2) = \sum_{i=1}^m \lambda_i^2 = \mathrm{tr}(S_m^2)$.

- What can you conclude on the dimension reduction by eigenvectors from the above two points?

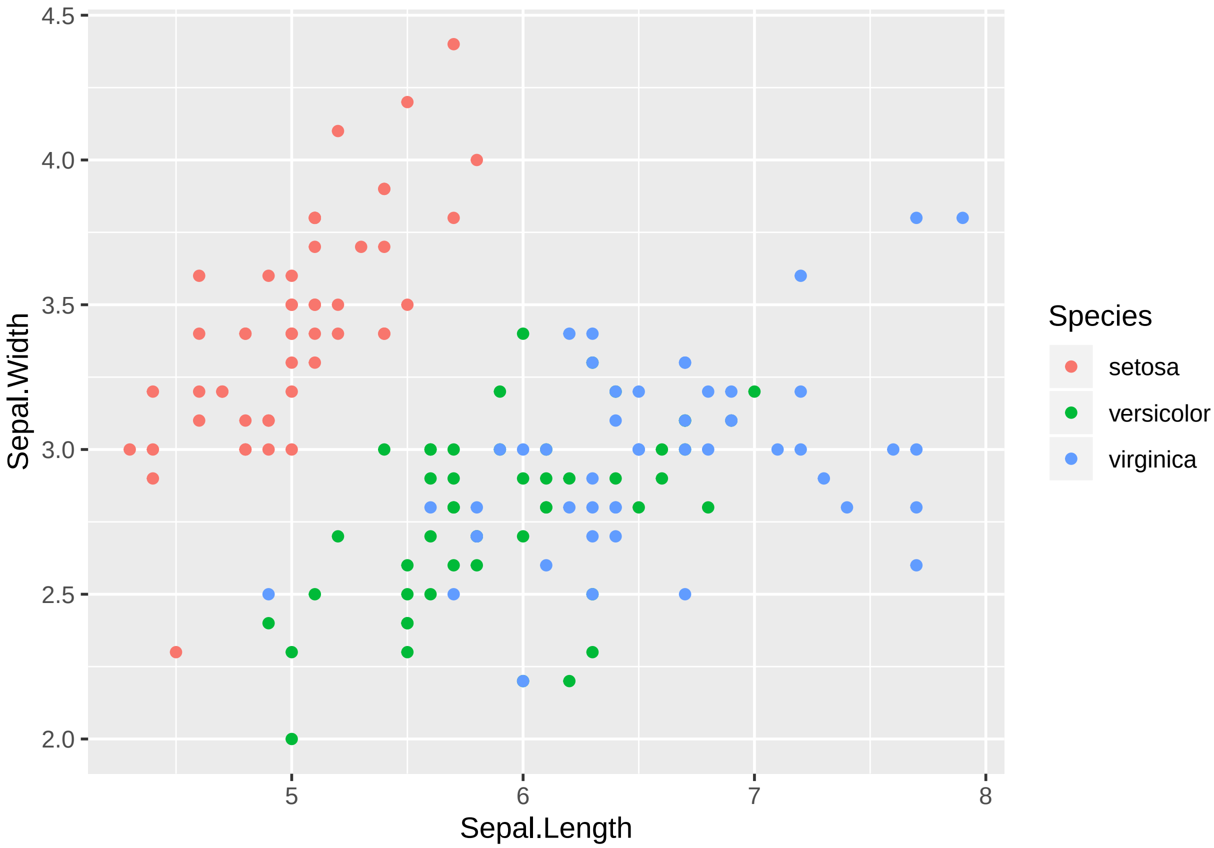
  Solution: If there is a few leading eigenvalues dominating the summation $\sum_{i=1}^p \lambda_i^2$, dimension reduction by eigenvectors can preserve most of the total variation of the original data while decrease the dimensions of the variables.

6. (10 pts)

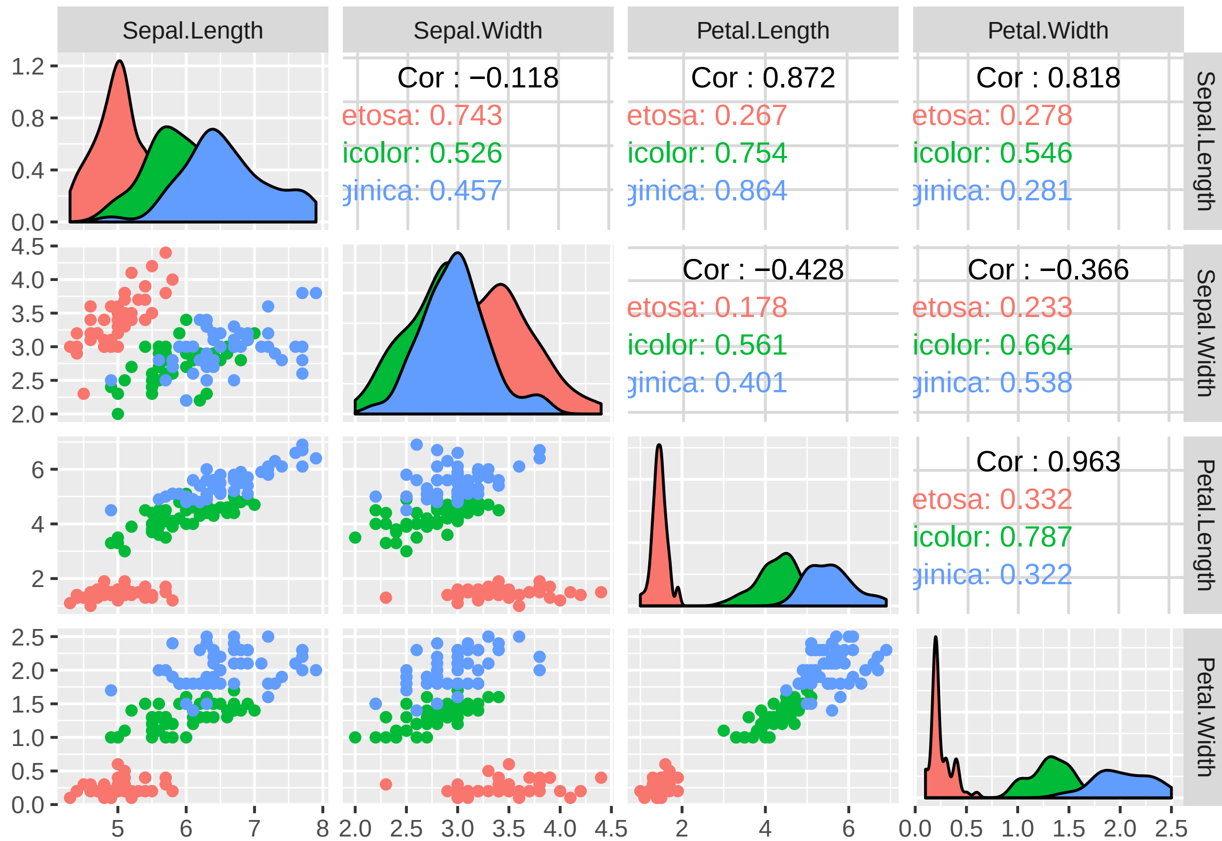
Use the **ggplot2** package to visualize the data `iris` in R.

- Make a scatter plot for the variables `Sepal.Length` and `Sepal.Width` colored by `Species`. What can you see?
- Make a scatter matrix for every pairs of the variables, colored by `Species`. What can you see?
- Calculate the sample mean for each species.
- Make a star plot for the sample means of each species to illustrate their potential differences. Comment.

```
library(tidyverse)
ggplot(data = iris, aes(x = Sepal.Length, y = Sepal.Width, colour = Species)) +
    geom_point()
```



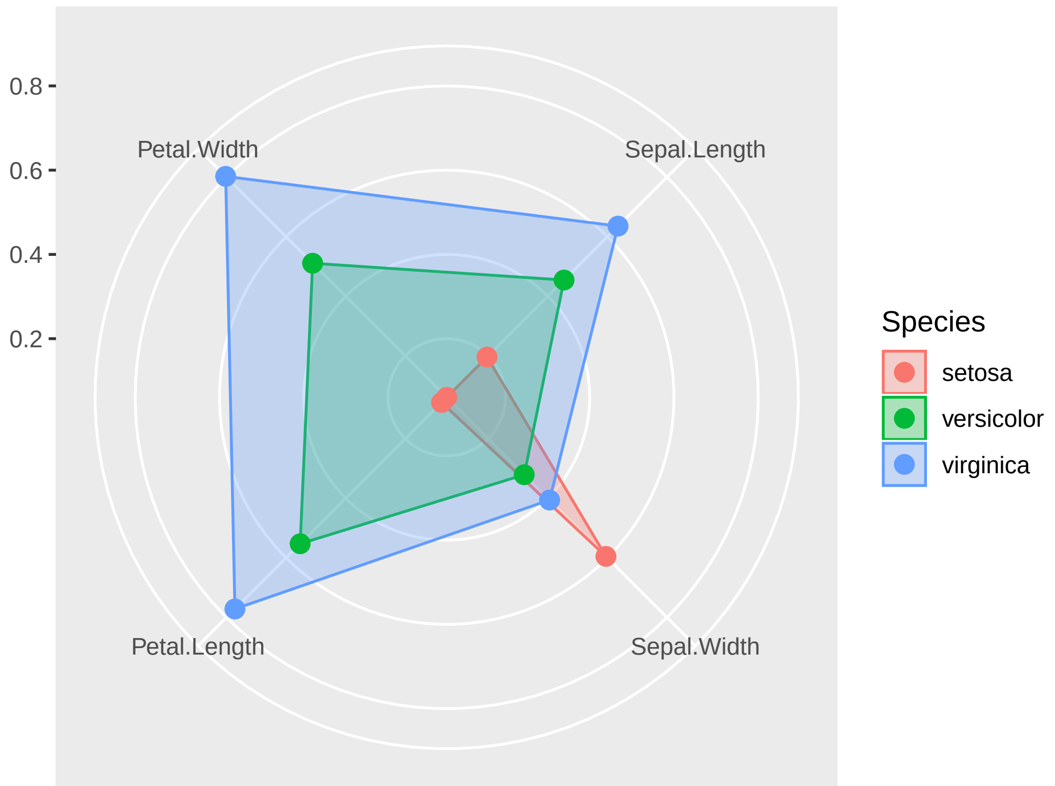

```
GGally::ggpairs(iris, columns = 1:4, aes(colour = Species))
```




```
iris %>% group_by(Species) %>% summarise_all(mean)
```

```
## # A tibble: 3 x 5
##   Species    Sepal.Length Sepal.Width Petal.Length Petal.Width
##   <fct>             <dbl>       <dbl>        <dbl>       <dbl>
## 1 setosa             5.01        3.43         1.46       0.246
## 2 versicolor         5.94        2.77         4.26       1.33
## 3 virginica          6.59        2.97         5.55       2.03
```

```
ggiraphExtra::ggRadar(iris, aes(colour = Species), legend.position = "right")
```




```r
# Another star plot function.
# devtools::install_github("ricardo-bion/ggradar", dependencies = TRUE)
# iris_mean %>%
#     mutate_at(vars(-Species), scales::rescale) %>%
#     ggradar::ggradar(axis.label.size = 3, group.line.width = .5,
#                      group.point.size = 2, legend.text.size = 12)
```