EEMB 146 Lab 9 Assignment

Samantha Chen

5/31/2021

Please Grade Exercise 2 (PCA)

Exercise 2: PCA

```
# Loading in data
data(iris)
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

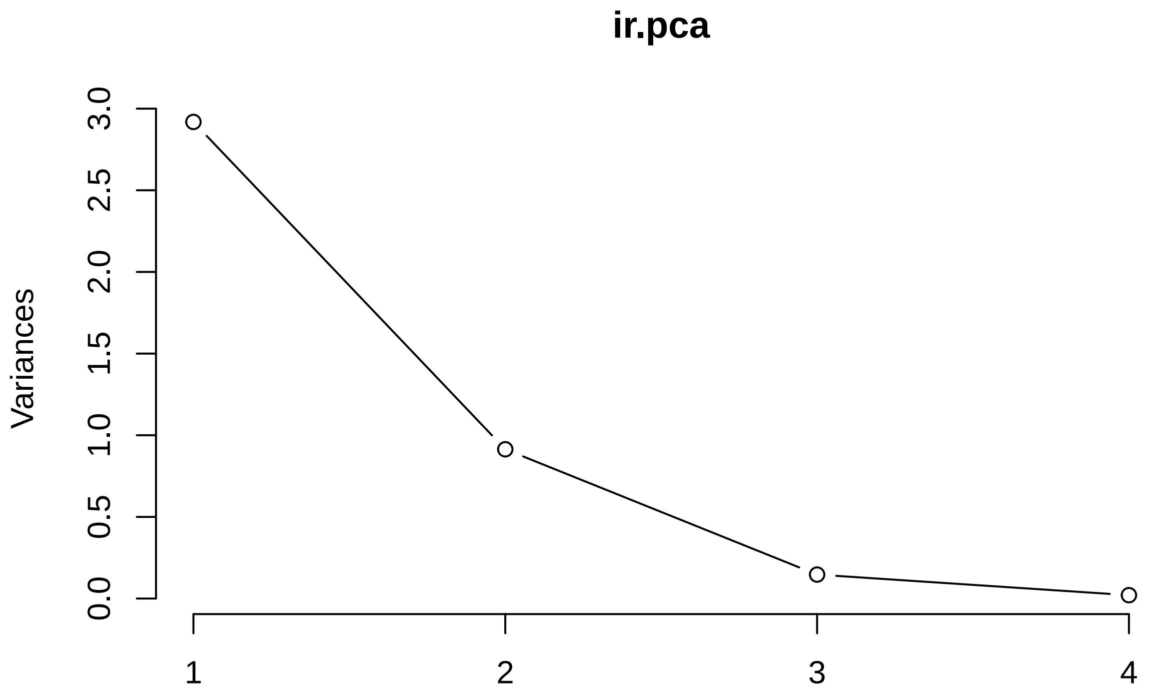
```
# cleaning data
iris.num <- iris[, -5] # removing non-numerical data

# running PCA
ir.pca <- prcomp(iris.num, center = TRUE, scale = TRUE)

print(ir.pca)
```

```
## Standard deviations (1, .., p=4):
## [1] 1.7083611 0.9560494 0.3830886 0.1439265
##
## Rotation (n x k) = (4 x 4):
##                     PC1         PC2        PC3        PC4
## Sepal.Length  0.5210659 -0.37741762  0.7195664  0.2612863
## Sepal.Width  -0.2693474 -0.92329566 -0.2443818 -0.1235096
## Petal.Length  0.5804131 -0.02449161 -0.1421264 -0.8014492
## Petal.Width   0.5648565 -0.06694199 -0.6342727  0.5235971
```

```
#plot PCA as line plot
plot(ir.pca, type = "line")
```

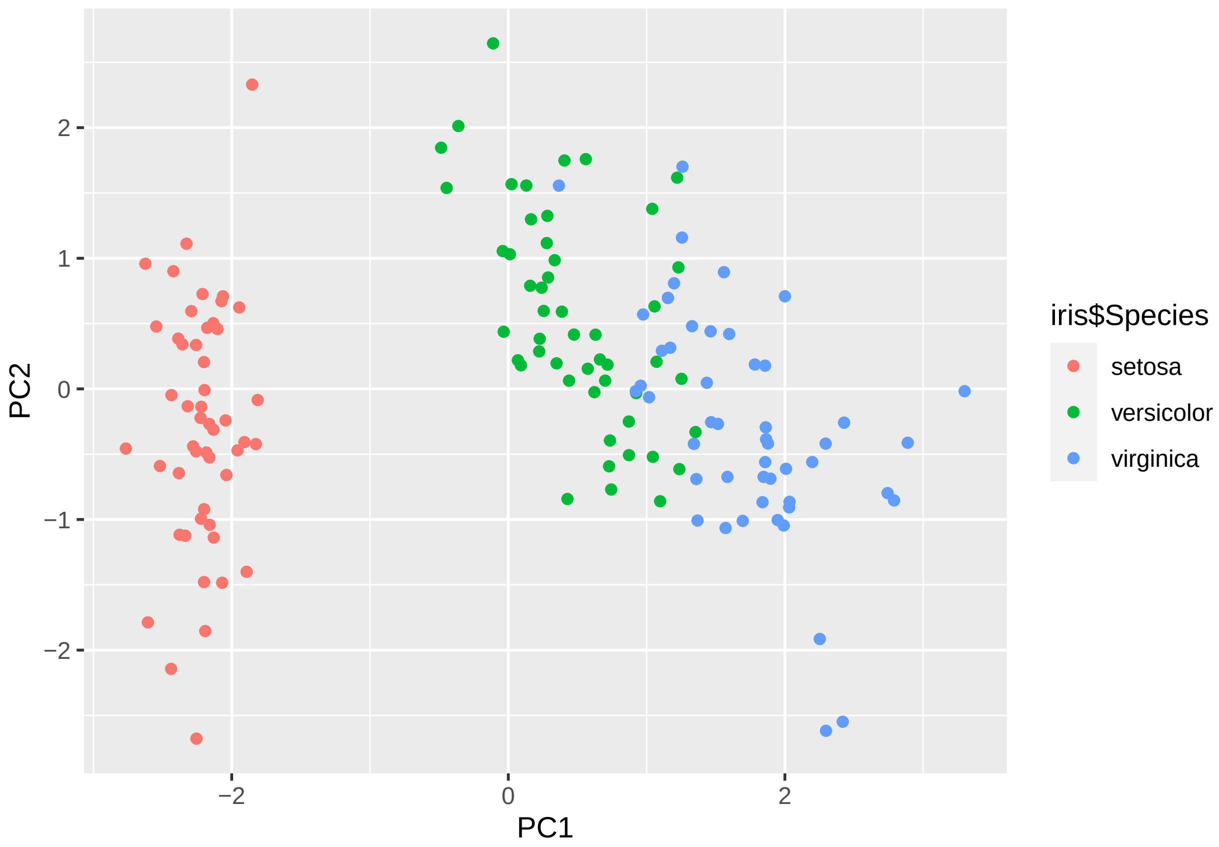


```
summary(ir.pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.7084 0.9560 0.38309 0.14393
## Proportion of Variance 0.7296 0.2285 0.03669 0.00518
## Cumulative Proportion  0.7296 0.9581 0.99482 1.00000
```

```
# Extract PC values for each observation
out <- as.data.frame(ir.pca$x)

#plot PC values
ggplot(out, aes(x = PC1, y = PC2, color = iris$Species)) +
  geom_point()
```



Excercise 2 Questions

1. How many PCs will a dataset with 10 variables return? A dataset with 100?

In a dataset the total number of principal components is determined by the original number of variables. In a dataset with 10 variables there will be 10 PCs and in a dataset with 100 variables there will be 100 PCs.

2. For the iris PCA, which variable is the most important for each PC?

The most important variable for PC1 is Petal Length (loading value = 0.5804131), for PC2 is Sepal Width (loading value = -0.92329566), for PC3 is Sepal Length (loading value = 0.7195664), and for PC4 is Petal Length (loading value = -0.8014492)

3. How many PCs would you use to describe this dataset? Why?

I would use the first two PCs to describe this dataset because according to the importance of components table, PC1 explains around 73% of the total variances and PC2 explains around 23% of the total variances. PC3 and PC4 barely explain 1% of the total variances so it won't be useful to the analysis.

Exercise 5: Survival Analysis

```
# loading in data
data("lung")
head(lung) #sex --> binary variable
```

```
##   inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
## 1    3  306      2  74   1       1       90       100     1175      NA
## 2    3  455      2  68   1       0       90        90     1225      15
## 3    3 1010      1  56   1       0       90        90       NA      15
## 4    5  210      2  57   1       1       90        60     1150      11
## 5    1  883      2  60   1       0      100        90       NA       0
## 6   12 1022      1  74   1       1       50        80      513       0
```

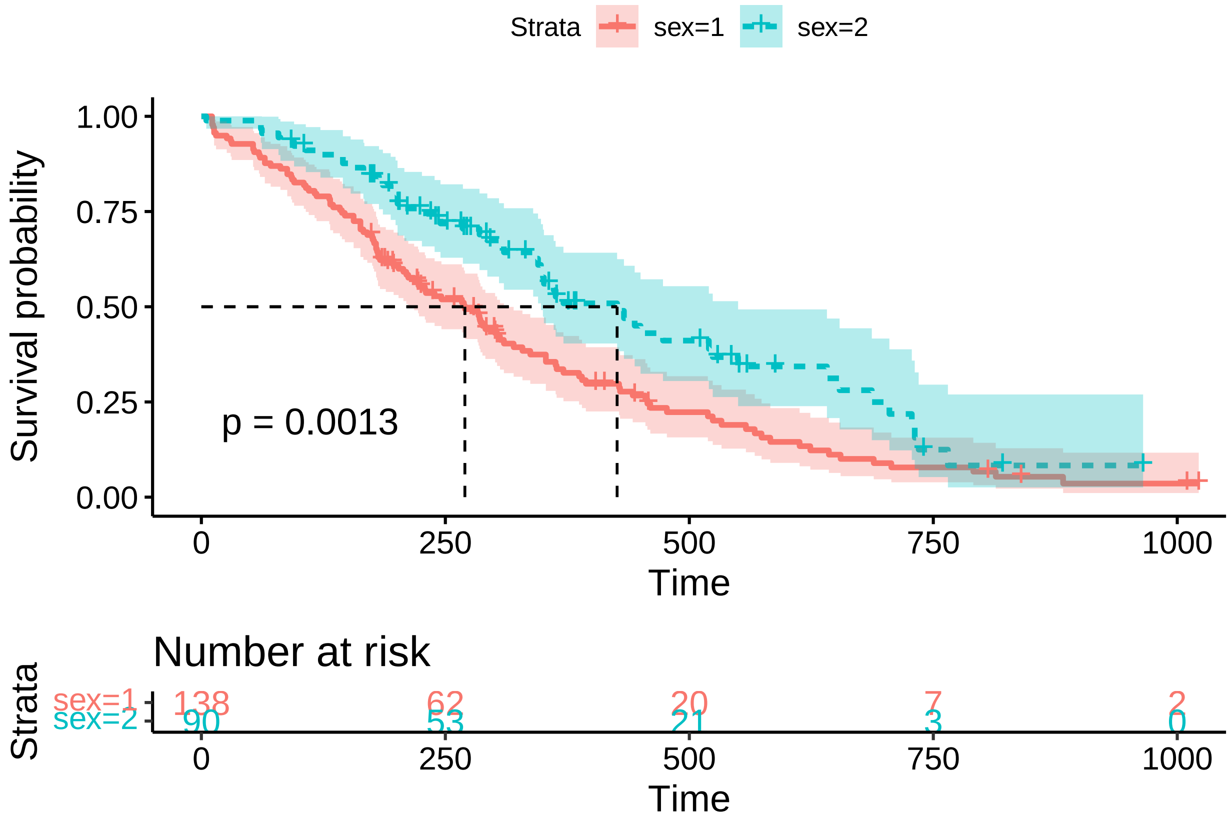
```
# creating model
srv <- survfit(Surv(time, status) ~ sex, data = lung)
print(srv) # median look significantly different from each other
```

```
## Call: survfit(formula = Surv(time, status) ~ sex, data = lung)
##
##         n events median 0.95LCL 0.95UCL
## sex=1 138    112    270     212     310
## sex=2  90     53    426     348     550
```

```
# survival analysis
res.sum <- surv_summary(srv)
head(res.sum)
```

```
##   time n.risk n.event n.censor      surv    std.err     upper     lower strata
## 1   11    138       3        0 0.9782609 0.01268978 1.0000000 0.9542301  sex=1
## 2   12    135       1        0 0.9710145 0.01470747 0.9994124 0.9434235  sex=1
## 3   13    134       2        0 0.9565217 0.01814885 0.9911586 0.9230952  sex=1
## 4   15    132       1        0 0.9492754 0.01967768 0.9866017 0.9133612  sex=1
## 5   26    131       1        0 0.9420290 0.02111708 0.9818365 0.9038355  sex=1
## 6   30    130       1        0 0.9347826 0.02248469 0.9768989 0.8944820  sex=1
##   sex
## 1   1
## 2   1
## 3   1
## 4   1
## 5   1
## 6   1
```

```
# plotting survival curve
ggsurvplot(srv, pval = T, conf.int = T,
           risk.table = T,
           risk.table.col = "strata",
           linetype = "strata",
           surv.median.line = "hv")
```



Exercise 5 Questions

1. Do you think sex is a good predictor of time-to-death? Which sex (1 or 2) has the shorter median survival time?

I think sex is a good predictor of time-to-death because the datapoints falls within the confidence levels (shaded areas) for both sexes. Sex 1 has the shorter median survival time, with a median time of 270 days.

2. Why isn't there a survival probability prediction for every single day?

The KM curve is only looking at the expected duration of time until occurence of an event of interest (in this case death by lung cancer). Sometimes some data is censored (removed) in which the event may not be observed for a person within the study time period. Either way, we are interested in the time until death occurs.

3. What does a verticle drop in the KM curve represent?

The vertical drop indicates an event occurring, therefore a drop in survival probability. In the case of our KM curve it means a participant in the study has died from lung cancer.