STAT 501 Homework 1

Gaussian

February 7, 2018

1. (a) Read the dataset

```r
senators<-read_xls("senate_voting_data.xls")
```

(b) Plot Andrews' curves

```r
senators.names<-names(senators)[-c(1,2)]
rev.party.state.names<-lapply(X=strsplit(gsub(patterns = "[.]", replacement = "",x=senators.names), strsplit = " "),FUN = rev)

senators.party <- lapply(X = rev.party.state.names, FUN = function(x)(unlist(x)[1]))
senators.party <- unlist(senators.party)

senators.last.names <- lapply(X = rev.party.state.names, FUN = function(x)(unlist(x)[4]))
senators.last.names <- unlist(senators.last.names)


#Create new data.frame for plotting
senators_new <- as.data.frame(t(senators[,-c(1,2)]))

colnames(senators_new) <- NULL
rownames(senators_new) <- NULL

senators_new <- data.frame(senators_new, party = senators.party)

# Use the codes from Canvas
source("ggandrews.R")

# Display the Andrews' curves
ggandrews(senators_new, type = 2, clr = 543, linecol = c("blue", "purple", "red"))
```

Andrew's Curves for senators is shown in Figure 1. From Andrews' curves we can see that for each party the curve follows a similar pattern, so senators within each party have similar voting preferences. We can also see curves from three different parties are mixed together, so it would be hard to distinguish the senator's party from the voting preference.

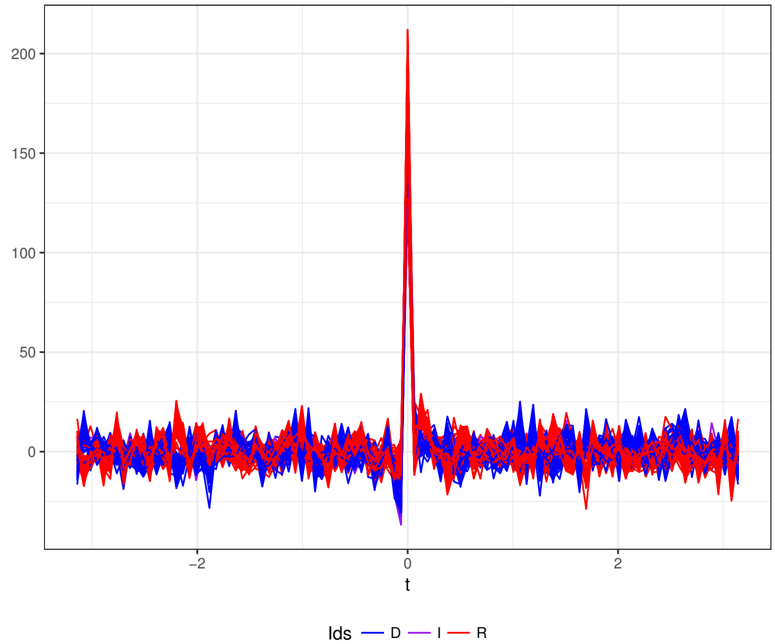


Figure 1: Andrew's Curves for senators

2. (a) Radial visualization and star coordinates

```
library(lattice)
require(dprep)
sclerosis <- read.table("sclerosis.dat", header=F)

p <- dim(sclerosis)[2]
sclerosis[, p] <- as.factor(ifelse(sclerosis[,p] == 0, "normal",
    "sclerosis"))

colnames(sclerosis) <- c("Age", "TS1", "DS1", "TS2", "DS2",
    "Disease")

# Use codes from Canvas
source("radviz2d.R")

# Display the radial visualization plot
radviz2d(dataset = sclerosis, name = "Sclerosis")

# Use the codes from Canvas
source("starcoord.R")

#Display the star coordinates
starcoord(data = sclerosis, class = TRUE)
```

Plot for radial visualization is shown in Figure 2 and plot for star coordinates is shown in Figure 3. From the radial visualization, we can see there is more variability in age within the "normal" group than that of "sclerosis" group. And the total response for stimuli S1 and

S2 are similar in both groups. The differences of response for S1 and S2 are also similar in both groups. From the star coordinates plot, we can also see that the "normal" group varies more that the "sclerosis" group in age. We can also see "normal" group varies less than the "sclerosis" in other dimensions and values for those dimensions are smaller.

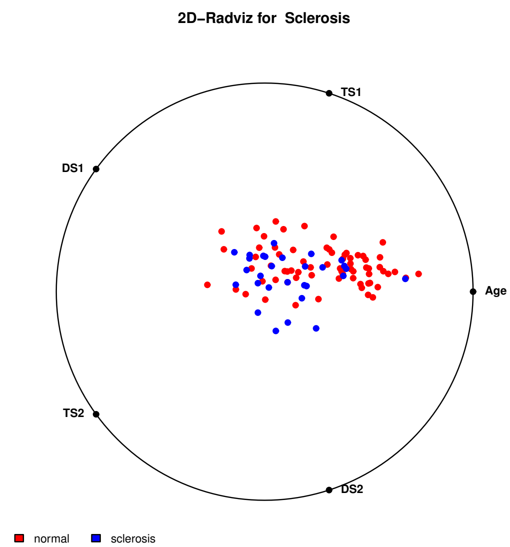


Figure 2: Radial visualization for sclerosis data

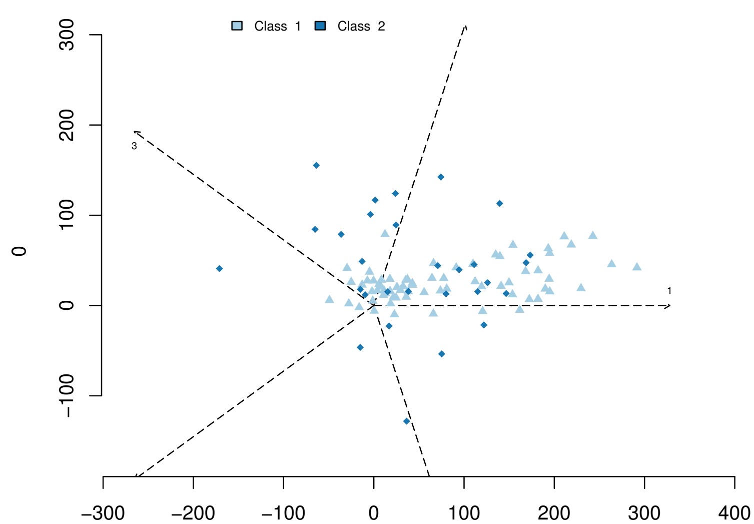


Figure 3: Star coordinates for sclerosis data

(b) Calculate the means for each group

```
normal_mean<-colMeans(sclerosis[sclerosis[,p]=="normal", -p])
normal_mean

sclerosis_mean<-colMeans(sclerosis[sclerosis[,p]=="sclerosis",-p])
sclerosis_mean
```

Means for group “normal”:

```
       Age        TS1        DS1        TS2        DS2
 37.985507 147.289855   1.562319 195.602899   1.620290
```

Means for group “sclerosis”:

```
      Age       TS1       DS1       TS2       DS2
 42.06897 178.26897  12.27586 236.93103  13.08276
```

(c) Display the Chernoff faces

```
sclerosis_mean_data <- as.matrix(rbind(normal_mean,
    sclerosis_mean))
library(TeachingDemos)
faces(sclerosis_mean_data, labels = c("normal", "sclerosis"))
```

The Chernoff faces for the two groups are shown in Figure 4.

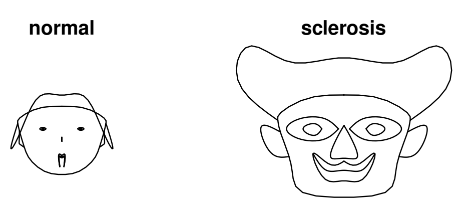


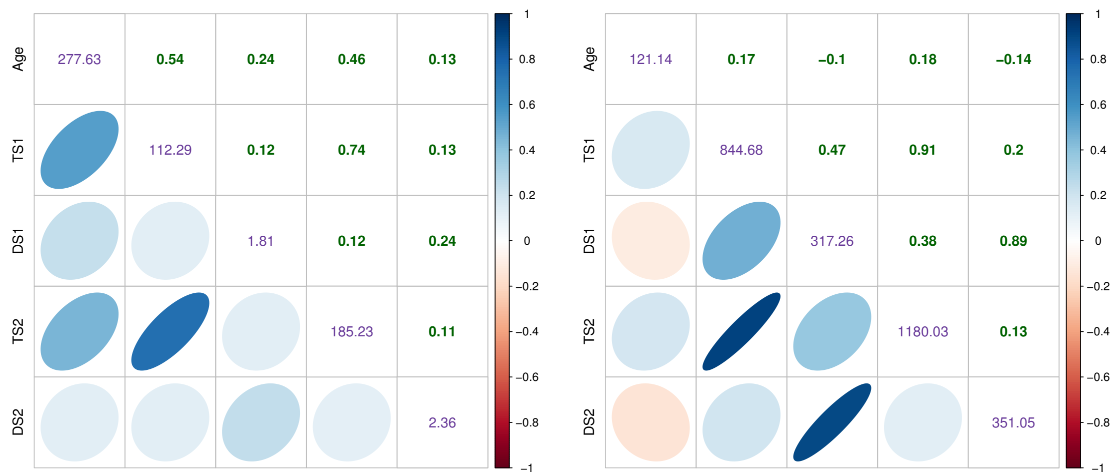
Figure 4: Chernoff faces for two groups of sclerosis data

(d) Display the correlation matrix for each group.

```r
source("plotcorr.R")
# normal
plot.corr(xx = sclerosis[sclerosis[,p]=="normal", -p])

# sclerosis
plot.corr(xx = sclerosis[sclerosis[,p]=="sclerosis", -p])
```

Correlation plots are shown in Figure 5. We can see for both groups, total response with



(a) normal

(b) sclerosis

Figure 5: Correlation plots for two groups

stimuli S1 and total response with stimuli S2 are highly correlated. For the "sclerosis" group we can also see difference in response with stimuli S1 and stimuli S2 are also highly correlated. Total response and age for "normal" group shows a stronger association than "sclerosis" group. We can also see the variance of age for "normal" group is larger than that of "sclerosis" group, and the variances for "normal" group are smaller than those of "sclerosis" group in other dimensions.

3. (a) Formulate the correlation plot

```r
#Read the dataset
Tornado<-read.table(file
    ="https://www.nssl.noaa.gov//users/brooks//public_html//
feda//datasets//tornf1p.txt", col.names = c("year", "jan", "feb",
    "mar","apr", "may","jun","jul", "aug","sep", "oct", "nov",
    "dec"))
#Create correlation plot
source("plotcorr.R")
plot.corr(Tornado[2:13])
```

The correlation plot is shown in Figure 6. Generally it is observed that months next to each other or near to each other like Jan-Feb, Jan-March has positibe correlation while month far away from each other/ in oppite season o year like Jan-June, Jan-July has negative correlation although it is not true for all the cases. From the correlation plot it is also observed that the variance is high for months during spring/early summer Apr-June compared to other months.

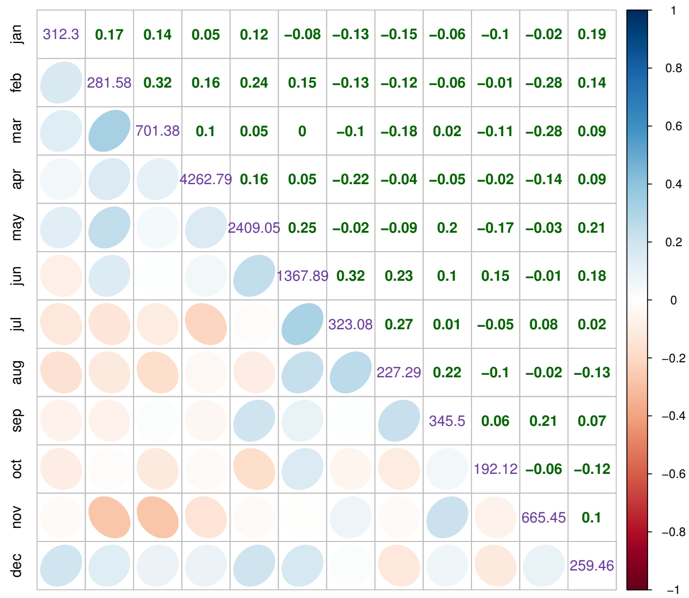


Figure 6: Correlation plot for Tornado data

(b) i. Parallel-coordinates plot for each group

```r
#Create 3 group and a column of group in the dataframe
Tornado$Period<-cut(x =
    Tornado$year,breaks=c(1954,1974,1994,2014),labels =
    c("I","II","III"),include.lowest = F)

#Create parallel plot with colour by the group
source("parcoordplot.R")
parcoordplot(xx =Tornado[-1,2:13],cl =
    as.factor(Tornado$Period[-1]),FUN=mean,alpha = 0.2)
```

The parallel-coordinates plot is shown in Figure 7. From the parallel plot along with

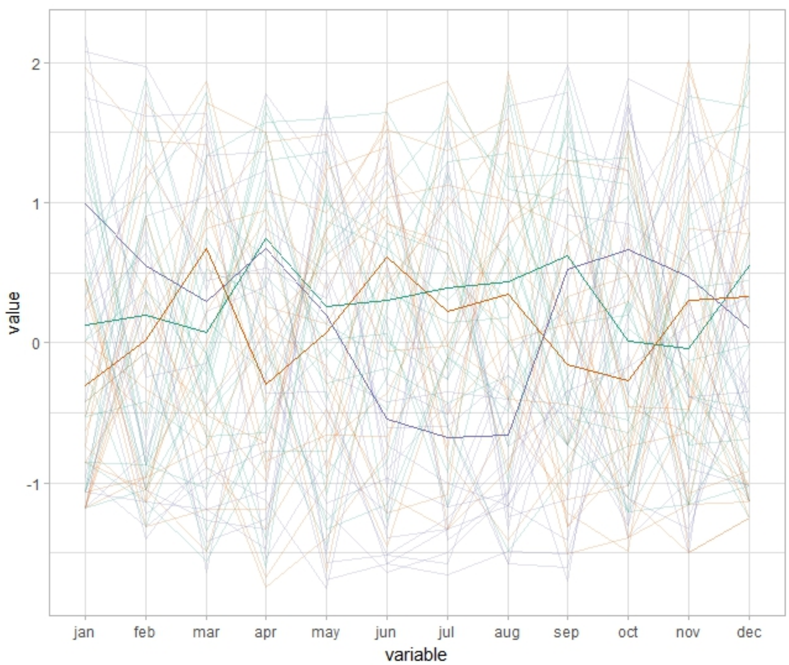
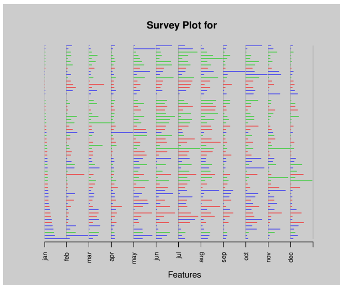


Figure 7: Parallel-coordinates plot for Tornado data

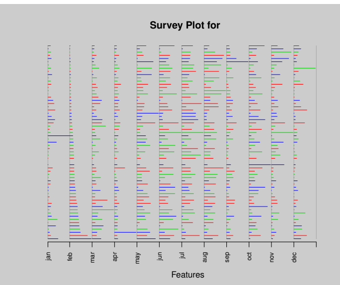
superimposed mean it can be seen that during different period the frequency of tornado varied largely across the months. For example during third period the frequency was very low from June to August while it was high during 1st and 2nd period. Any specific pattern however is hard to observe due to the variation.

ii. Create survey plots ordered by each of 12 months

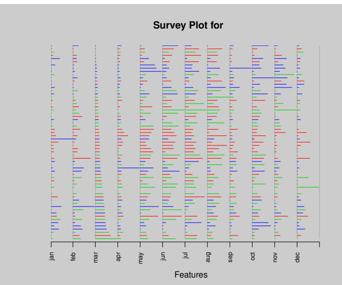
```
source("surveyplot.R")
for (i in 1:12){
  surveyplot(cbind(Tornado[,2:13],
      as.numeric(as.factor(Tornado[,14]))), order = i)}
```



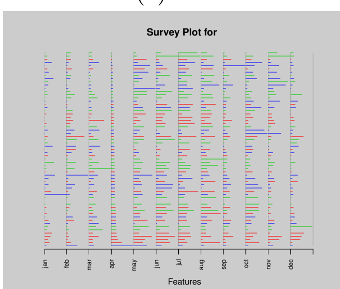
(a) Jan



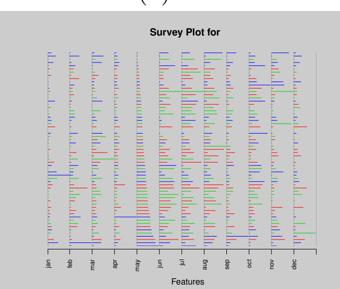
(b) Feb



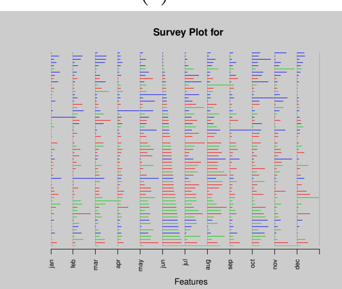
(c) Mar



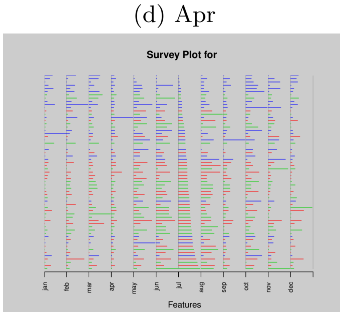
(d) Apr



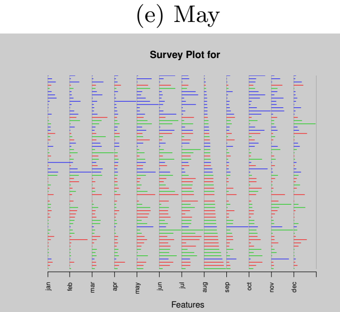
(e) May



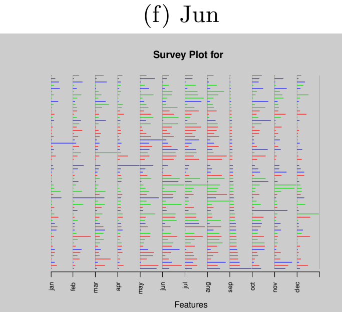
(f) Jun



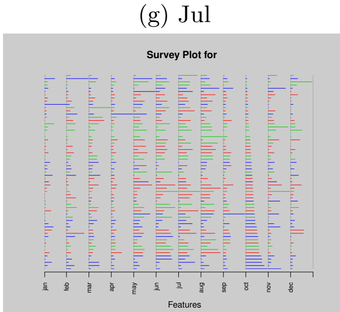
(g) Jul



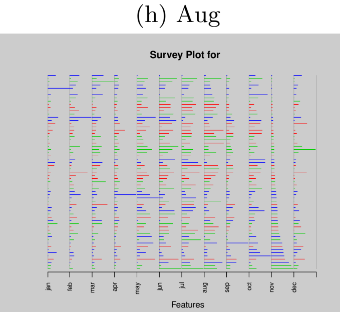
(h) Aug



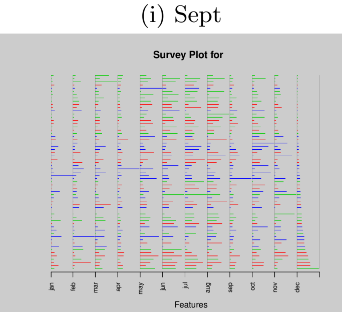
(i) Sept



(j) Oct



(k) Nov



(l) Dec

Figure 8: survey plots with different orders

The survey plots with 12 different orders are shown in Figure 8. From the survey plot distinct patterns were not visible and ordering by none of the month provided a clear separation between the classes

iii. Create the stars and Chernoff faces plot for the mean of three group

```
library(TeachingDemos)
# Compute the mean for three groups
Means<-aggregate(x =
    Tornado[-1,2:13],by=list(Tornado$Period[-1]),FUN = mean)

#Create chernoff face for the means
faces(xy = Means[,2:13])
#Create stars for the group means
stars(x = Means[,2:13],labels =
    as.character(Means$Group.1),scale = T,full = T,radius = T)
```

Plot of Chernoff faces is shown in Figure 9 and stars in shown in Figure 10. From both Chernoff face and stars, differences are observed between the group means.

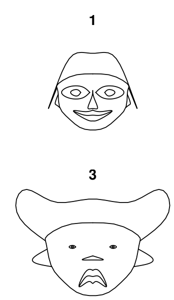


Figure 9: Chernoff faces for Tornado data

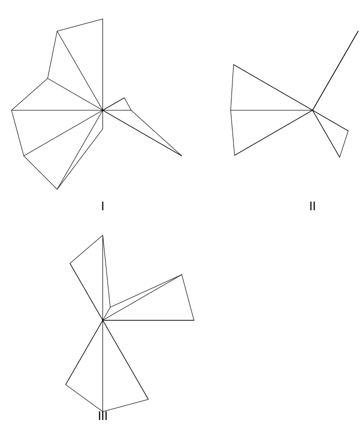


Figure 10: Stars for Tornado data

4. Let $\boldsymbol{a} = \boldsymbol{X} - \boldsymbol{X}^T\mathbf{1}/p, \boldsymbol{b} = \boldsymbol{Y} - \boldsymbol{Y}^T\mathbf{1}/p$. Then

$$\|\boldsymbol{X}^\circ - \boldsymbol{Y}^\circ\|^2 = \left\| \frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} - \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|} \right\|^2 = \left\langle \frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} - \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|}, \frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} - \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|} \right\rangle,$$

where $\langle \cdot, \cdot \rangle$ is inner product.

Since $\mathrm{Var}(\boldsymbol{X}) = \sum_{i=1}^p (x_i - \bar{x})^2 = \|\boldsymbol{X} - \boldsymbol{X}^T\mathbf{1}/p\|^2 = \|\boldsymbol{a}\|^2$, $\mathrm{Var}(\boldsymbol{Y}) = \sum_{i=1}^p (y_i - \bar{y})^2 = \|\boldsymbol{Y} - \boldsymbol{Y}^T\mathbf{1}/p\|^2 = \|\boldsymbol{b}\|^2$ and $\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^p (x_i - \bar{x})(y_i - \bar{y}) = \left\langle \boldsymbol{X} - \boldsymbol{X}^T\mathbf{1}/p, \boldsymbol{Y} - \boldsymbol{Y}^T\mathbf{1}/p \right\rangle = \langle \boldsymbol{a}, \boldsymbol{b} \rangle$.
Hence

$$\mathrm{Corr}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y})}{\sqrt{\mathrm{Var}(\boldsymbol{X})}\sqrt{\mathrm{Var}\boldsymbol{Y}}} = \frac{\langle \boldsymbol{a}, \boldsymbol{b} \rangle}{\|\boldsymbol{a}\|\|\boldsymbol{b}\|}$$

Thus

$$\begin{aligned}
\|\boldsymbol{X}^\circ - \boldsymbol{Y}^\circ\|^2 &= \left\langle \frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} - \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|}, \frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} - \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|} \right\rangle \\
&= \left\langle \frac{\boldsymbol{a}}{\|\boldsymbol{a}\|}, \frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} \right\rangle + \left\langle \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|}, \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|} \right\rangle - 2 \left\langle \frac{\boldsymbol{a}}{\|\boldsymbol{a}\|}, \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|} \right\rangle \\
&= \frac{\|\boldsymbol{a}\|^2}{\|\boldsymbol{a}\|^2} + \frac{\|\boldsymbol{b}\|^2}{\|\boldsymbol{b}\|^2} - \frac{2\langle \boldsymbol{a}, \boldsymbol{b} \rangle}{\|\boldsymbol{a}\|\|\boldsymbol{b}\|} \\
&= 2 - 2\mathrm{Corr}(\boldsymbol{X}, \boldsymbol{Y})
\end{aligned}$$