PCA example II

US-air data

STAT 32950-24620

Spring 2023 (3/23, wk1)

PCA II - Old usair data

Data: Air Pullution in 41 US cities

- SO2: Sulphur dioxide content of air in micrograms per cubic meter
- Temp: Average annual temperature in Fahrenheit
- Manuf: Number of manufacturing enterprises employing 20 or more workers
- Pop: Population size (1970 census) in thousands
- Wind: Average annual wind speed in miles per hours
- Percip: Average annual percipitation in inches
- Days: Average number of days with persipitation per year

Source: 'A peek at some history of Chicago', from Everitt Ch3

Input data

```
usair = source("chap3usair.dat")$value; str(usair)
```

```
## 'data.frame':    41 obs. of  7 variables:
##  $ SO2     : num  10 13 12 17 56 36 29 14 10 24 ...
##  $ Neg.Temp: num  -70.3 -61 -56.7 -51.9 -49.1 -54 -57.3
##  $ Manuf   : num  213 91 453 454 412 80 434 136 207 368
##  $ Pop     : num  582 132 716 515 158 80 757 529 335 497
##  $ Wind    : num  6 8.2 8.7 9 9 9 9.3 8.8 9 9.1 ...
##  $ Precip  : num  7.05 48.52 20.66 12.95 43.37 ...
##  $ Days    : num  36 100 67 86 127 114 111 116 128 115 .
```

```
attach(usair)
```

Remarks:

- SO2 can be treated as response.
- Use (-1)*temp, so all high values mean worse environment.

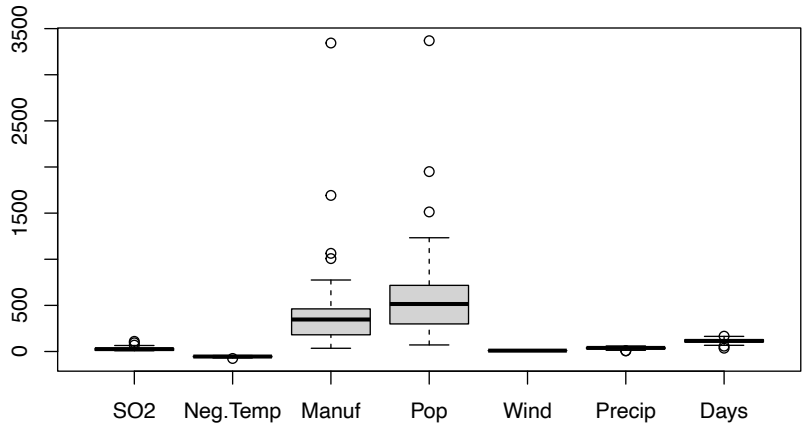

```
summary(usair[,1:3]); summary(usair[,4:7])
```

```
##       SO2           Neg.Temp         Manuf
##  Min.   :  8.0   Min.   :-75.5   Min.   :  35
##  1st Qu.: 13.0   1st Qu.:-59.3   1st Qu.: 181
##  Median : 26.0   Median :-54.6   Median : 347
##  Mean   : 30.1   Mean   :-55.8   Mean   : 463
##  3rd Qu.: 35.0   3rd Qu.:-50.6   3rd Qu.: 462
##  Max.   :110.0   Max.   :-43.5   Max.   :3344
```

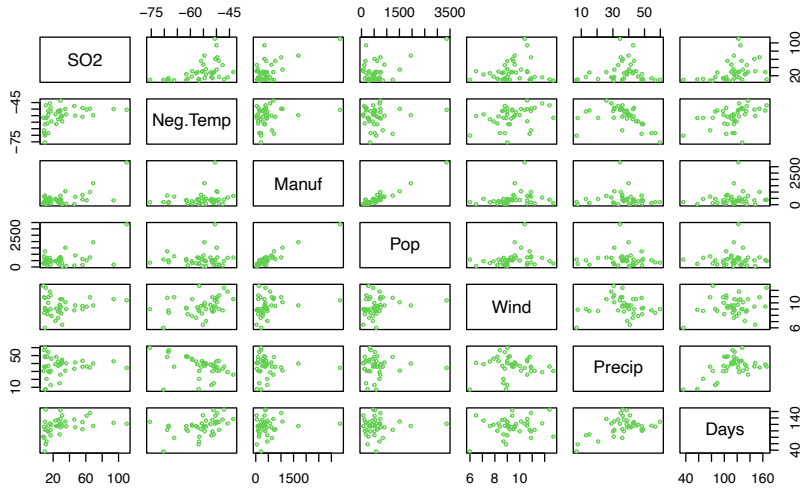
```
##       Pop            Wind          Precip           Day
##  Min.   :  71   Min.   : 6.00   Min.   : 7.05   Min.
##  1st Qu.: 299   1st Qu.: 8.70   1st Qu.:30.96   1st Qu.
##  Median : 515   Median : 9.30   Median :38.74   Median
##  Mean   : 609   Mean   : 9.44   Mean   :36.77   Mean
##  3rd Qu.: 717   3rd Qu.:10.60   3rd Qu.:43.11   3rd Qu.
##  Max.   :3369   Max.   :12.70   Max.   :59.80   Max.
```

Scales of variables - Important

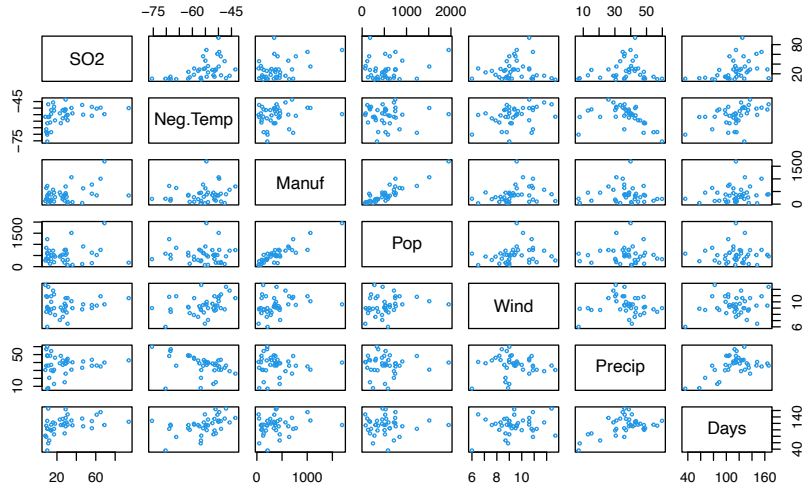
```
boxplot(usair)
```



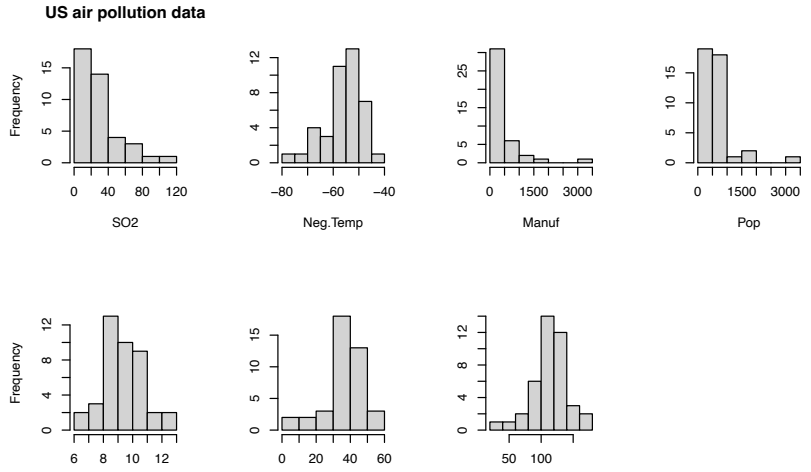

```
pairs(usair,cex=.5,col=3) #outlier: obs 11 (Chicago)
```




```
pairs(usair[-11,],cex=.5,col=4) # w/o outlier obs 11
```




```
par(mfrow=c(2,4)) # Histograms of usair vars
hist(SO2,main="US air pollution data")
hist(Neg.Temp, ylab="",main="")
hist(Manuf, ylab="",main=""); hist(Pop, ylab="",main="")
hist(Wind, main="")
hist(Precip, ylab="",main=""); hist(Days, ylab="",main="")
```



```
round(cor(usair),2)
```

```
##             SO2 Neg.Temp Manuf   Pop  Wind Precip Days
## SO2        1.00     0.43  0.64  0.49  0.09   0.05 0.37
## Neg.Temp   0.43     1.00  0.19  0.06  0.35  -0.39 0.43
## Manuf      0.64     0.19  1.00  0.96  0.24  -0.03 0.13
## Pop        0.49     0.06  0.96  1.00  0.21  -0.03 0.04
## Wind       0.09     0.35  0.24  0.21  1.00  -0.01 0.16
## Precip     0.05    -0.39 -0.03 -0.03 -0.01   1.00 0.50
## Days       0.37     0.43  0.13  0.04  0.16   0.50 1.00
```

```
round(cov(usair),2)
```

```
##              SO2 Neg.Temp    Manuf      Pop   Wind  Prec
## SO2       550.95    73.56   8527.7   6712.0   3.18    15
## Neg.Temp   73.56    52.24    774.0    262.4   3.61   -32
## Manuf    8527.72   773.97 317502.9 311718.8 191.55  -215
## Pop      6711.99   262.35 311718.8 335371.9 175.93  -178
## Wind        3.18     3.61    191.6    175.9   2.04    -0
## Precip     15.00   -32.86   -215.0   -178.1  -0.22   138
## Days      229.93    82.43   1969.0    646.0   6.21   154
```


PCA using scaled data

```
usair.pc = princomp(usair[,-1],cor=T)
summary(usair.pc,loading=T, digit =3)
```

```
## Importance of components:
##                           Comp.1 Comp.2 Comp.3 Comp.4 Comp.
## Standard deviation         1.482  1.225 1.1810 0.8719 0.338
## Proportion of Variance     0.366  0.250 0.2324 0.1267 0.019
## Cumulative Proportion      0.366  0.616 0.8485 0.9752 0.994
##
## Loadings:
##           Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## Neg.Temp   0.330  0.128  0.672  0.306  0.558  0.136
## Manuf      0.612 -0.168 -0.273  0.137  0.102 -0.703
## Pop        0.578 -0.222 -0.350                0.695
## Wind       0.354  0.131  0.297 -0.869 -0.113
## Precip            0.623 -0.505 -0.171  0.568
## Days       0.238  0.708         0.311 -0.580
```


Variation explained by PCs

The first three PCs account for 85% of variations (!?!)

- Principal component 1:

  "Quality of life or goodness environment"

- Principal component 2:

  "Wetness of weather"

  (amount and duration of rainfall)

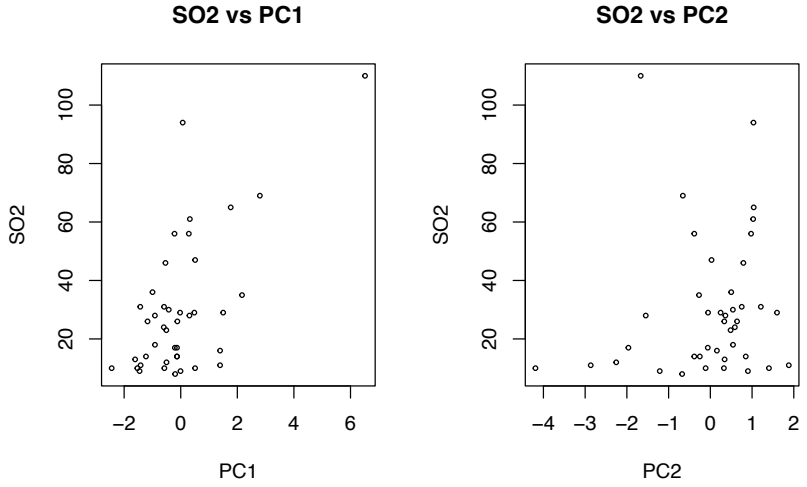
- Principal component 3:

  "Climate type": hot-and-wet vs cold-and-dry

  (contrast between rainfall and negative temperature)

SO_2 vs PCs

```
par(mfrow=c(1,2))
plot(usair.pc$scores[,1],SO2,xlab="PC1",cex=0.5); title("SO
plot(usair.pc$scores[,2],SO2,xlab="PC2",cex=0.5); title("SO
```

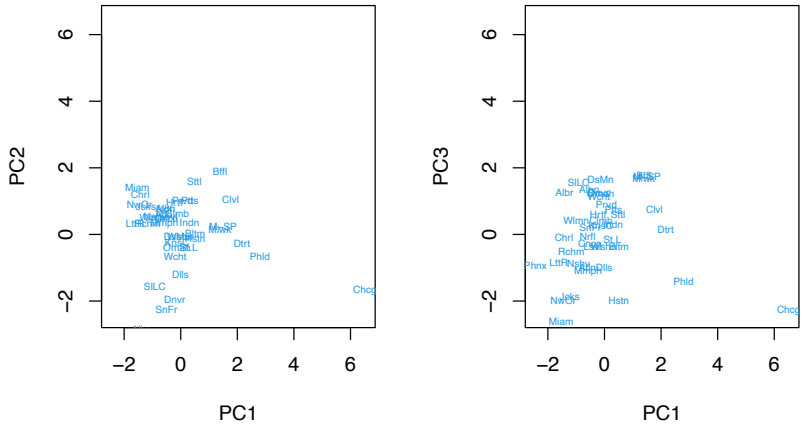


PC coordinates of obs labeled by city names (code)

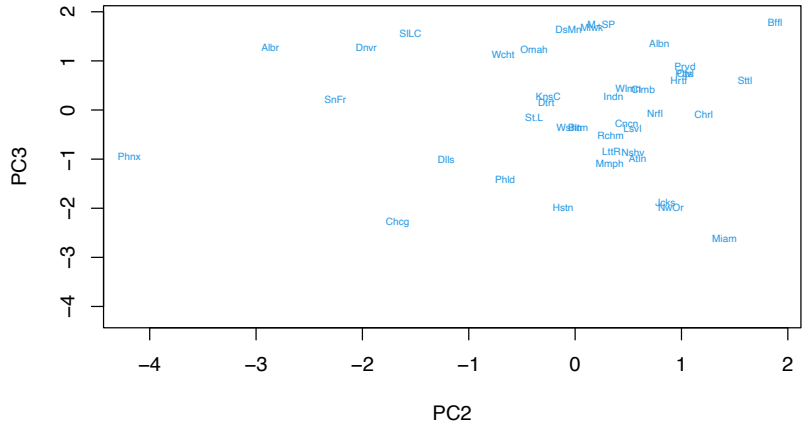
```
par(pty="m") # "s" for square, "m" for max
plot(usair.pc$scores[,1],usair.pc$scores[,2],
     ylim=range(usair.pc$scores[,1]),
     xlab="PC1",ylab="PC2",type="n",lwd=2)
text(usair.pc$scores[,1],usair.pc$scores[,2],
     labels=abbreviate(row.names(usair.dat)),
     cex=0.5,lwd=2,col=4)
```


Plots of PC coordinates of obs (cities)

Any interesting patterns?



Pattern revealed by PC plots of obs



Scaling size; comparison with non-scale PC and loading

```
round(usair.pc$scale,3) # the scaling applied to each var
```

```
## Neg.Temp    Manuf      Pop     Wind   Precip     Days
##    7.139  556.560  572.007    1.411   11.627   26.181
```

Compare with non-standardized data

```
summary(princomp(usair[,-1]))$sdev
```

```
##  Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6
## 789.128 119.620  25.756  10.769   3.512   1.247
```

```
round(princomp(usair[,-1])$loading[,1],5)
```

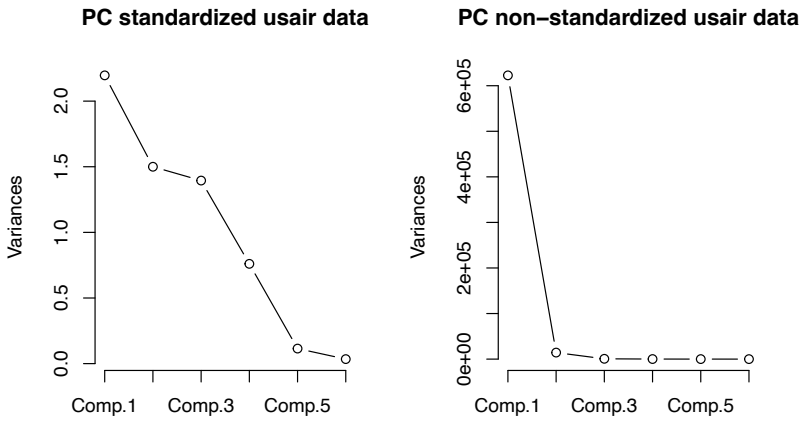
```
## Neg.Temp    Manuf      Pop     Wind   Precip     Days
##  0.00114  0.69690  0.71716  0.00041 -0.00043  0.00288
```

To scale or not to scale

For this example, Whether the variables are scaled to variance = 1 makes huge differences in PCA, as illustrated in the scree plots, which show the variance captured by each PC.

```
# Scree plots (original vs scaled data; cov vs corr)
par(mfrow=c(1,2))
screeplot(princomp(usair[,-1],cor=T),type="l",
          main="PC standardized usair data")
screeplot(princomp(usair[,-1]),type="l",
          main="PC non-standardized usair data")
```


Very different results - Which one is misleading?



Plot observations on PCs (code)

```
plot(usair.pc$scores[,1],usair.pc$scores[,2],
  ylim=range(usair.pc$scores[,2]),
  xlab = "PC1",ylab="PC2",type="n")
symbols(usair.pc$scores[,1],usair.pc$scores[,2],
        circles=sqrt(usair$SO2),
  inches=0.2,add=TRUE,bg="pink",fg="black")
text(usair.pc$scores[,1],usair.pc$scores[,2],
     labels=abbreviate(row.names(usair.dat)),
     cex=0.5,lwd=2,col=4)
title("City sized by SO2")
```


Plot of observations (cities) on (PC1, PC2)