

Statistics 452: Statistical Learning and Prediction

Chapter 10, part 1: Introduction to Unsupervised Learning

Brad McNeney

Supervised *versus* Unsupervised Learning

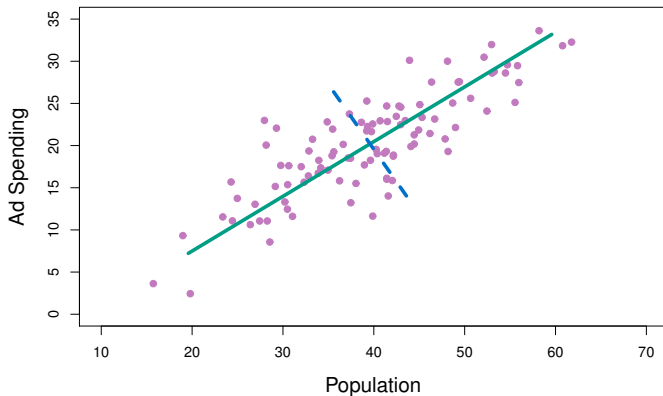
- ▶ Supervised means that there is an outcome \mathbf{y} , unsupervised means there is not.
- ▶ Supervised learning has well-defined goals like prediction.
 - ▶ Can check the fitted model by seeing how well it predicts test observations.
- ▶ Unsupervised learning is more exploratory, without an obvious goal.
 - ▶ A common theme is trying to identify simple structure underlying the feature data.
 - ▶ We will discuss dimension reduction by principal components analysis (PCA) and clustering.

Principal Components Analysis (PCA)

- ▶ Goal is low-rank approximation of the X data matrix
 - ▶ Discussed in Chapter 6 and reviewed below.
- ▶ Think of principal components (PCs) as new coordinates for the data vectors.
 - ▶ The first PC is the direction of greatest variation,
 - ▶ The second PC is the direction of second-greatest variation, orthogonal to the first,
 - ▶ And so on.

PCs for Advertising Data

- Text Figure 6.14: The green line is the first PC, the blue line the second.



PCs as Linear Combinations of X 's

- ▶ The details of how the linear combinations are derived are discussed in the text.
- ▶ In the advertising example, the first PC is

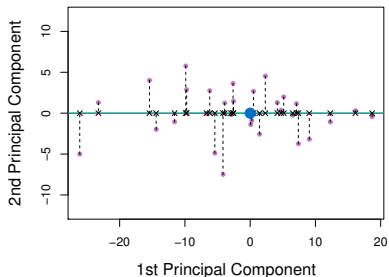
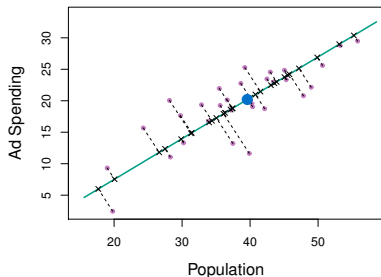
$$Z_1 = 0.838X_1 + 0.544X_2$$

where X_1 is population centred by its mean and X_2 is advertising expenditure centred by its mean.

- ▶ The coefficients of the linear combination, $\phi_{11} = 0.838$ and $\phi_{12} = 0.544$, are called the first principal component *loadings*.

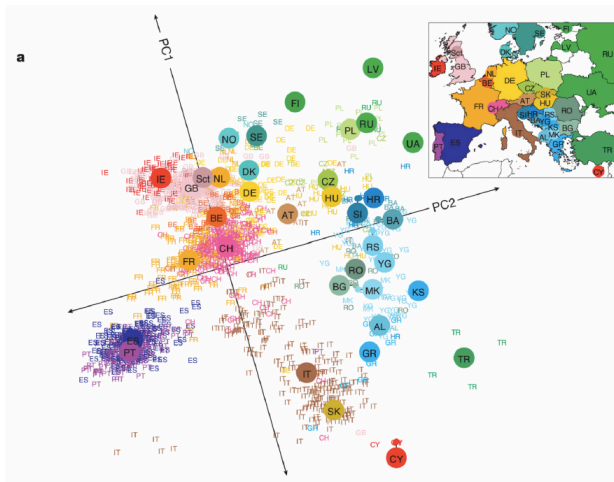
Principal Component Scores

- ▶ Projecting each point onto the PCs gives the PC scores.
 - ▶ Projecting a data vector onto a line means finding the point on the line closest to the vector.
- ▶ Text Figure 6.15: Black x's are the first PC score for each observation, distance of each purple dot from the green line is the second PC score.



High-Dimensional Example: Genes Reflect Geography

- First 2 PCs from 197,146 genetic markers on 1,387 European individuals (Novembre *et al.* 2008)



US Arrests Data

- ▶ Dataset that comes with R.
- ▶ From the help file: “This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.”

```
data(USArrests) # help(USArrests)
head(USArrests)
```

##		Murder	Assault	UrbanPop	Rape
##	Alabama	13.2	236	58	21.2
##	Alaska	10.0	263	48	44.5
##	Arizona	8.1	294	80	31.0
##	Arkansas	8.8	190	50	19.5
##	California	9.0	276	91	40.6
##	Colorado	7.9	204	78	38.7


```
# defaults in prcomp() are to center, but not scale
pcout <- prcomp(USArrests, scale=TRUE)
pcout$rotation # loadings
```

##		PC1	PC2	PC3	PC4
##	Murder	-0.5358995	0.4181809	-0.3412327	0.64922780
##	Assault	-0.5831836	0.1879856	-0.2681484	-0.74340748
##	UrbanPop	-0.2781909	-0.8728062	-0.3780158	0.13387773
##	Rape	-0.5434321	-0.1673186	0.8177779	0.08902432

```
head(pcout$x) # scores
```

##		PC1	PC2	PC3	PC4
##	Alabama	-0.9756604	1.1220012	-0.43980366	0.154696581
##	Alaska	-1.9305379	1.0624269	2.01950027	-0.434175454
##	Arizona	-1.7454429	-0.7384595	0.05423025	-0.826264240
##	Arkansas	0.1399989	1.1085423	0.11342217	-0.180973554
##	California	-2.4986128	-1.5274267	0.59254100	-0.338559240
##	Colorado	-1.4993407	-0.9776297	1.08400162	0.001450164

Scree Plot

- ▶ A scree plot shows the variance (or proportion of total variance) in the direction of each PC.
- ▶ If the variance drops and then levels out, the “elbow” where it levels out is a reasonable choice for a reduced number of PCs that captures most of the variation in the **X**.

```
screeplot(pcout) # or just plot(pcout)
```



- ▶ No obvious elbow.

Iris Data

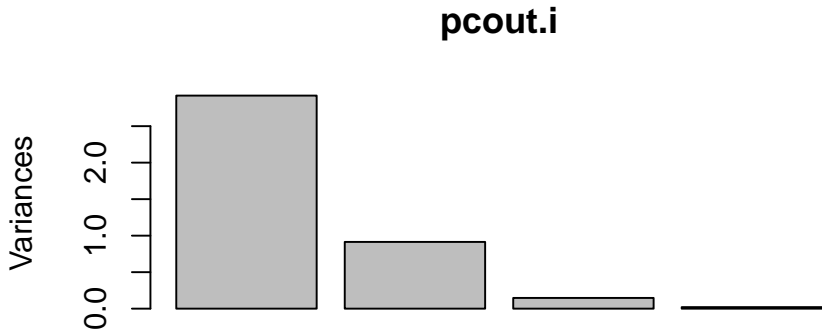
```
data(iris) # help(iris)
head(iris)
```

```
##   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
```

```
pcout.i <- prcomp(iris[,-5],scale=TRUE)
```

- ▶ For the iris data, two PCs appear to explain most of the variation.

```
screepplot(pcout.i)
```



Interpretation of Loadings

- ▶ The first PC is a contrast between sepal width and the other variables.
- ▶ The second PC is a weighted average of sepal length and width.

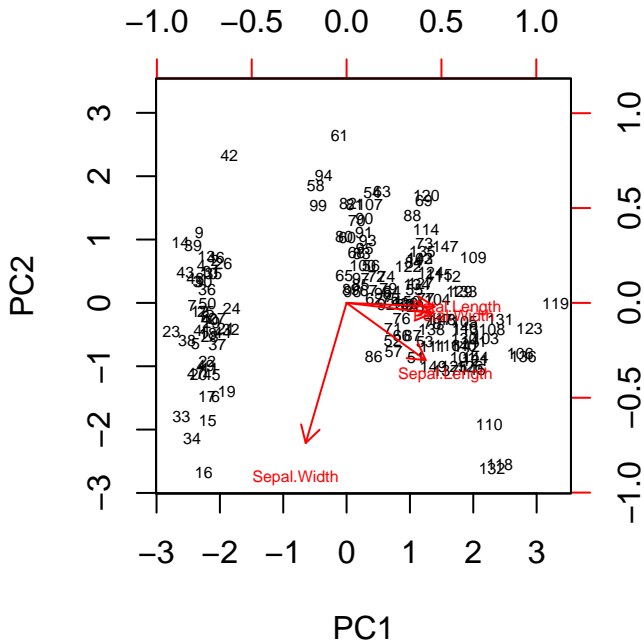
```
pcout.i$rotation
```

##		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

Biplot of First Two PCs

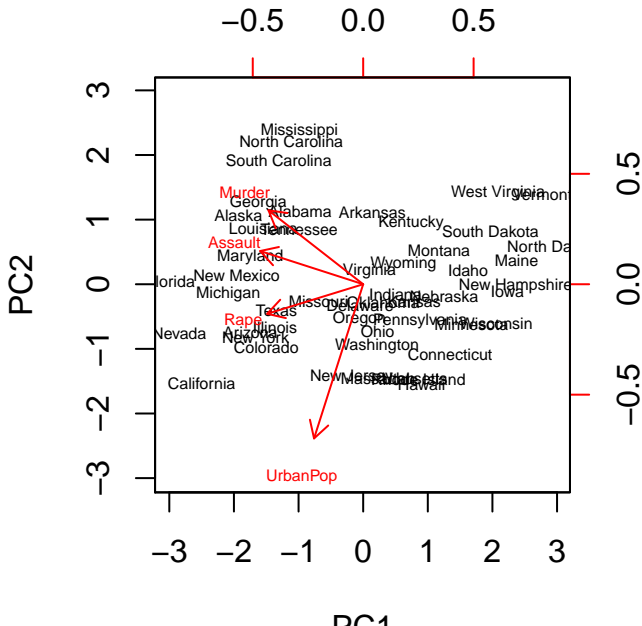
- ▶ We can visualize the first two PCs on a scatterplot.
- ▶ A biplot shows
 - (i) the PC scores for observational units (see left and bottom axes), and
 - (ii) the loadings of the features that define the first two PCs (see top and right axes)

```
biplot(pcout.i,cex=.5,scale=0) #scale=0 avoids scaling of points on plot
```



Biplot of IIS Arrests Data

```
biplot(pcout, cex=.5, scale=0) #scale=0 avoids scaling of points on plot
```



- ▶ Note different appearance from text: PCs are only unique up to a sign change

Multiple Correspondence Analysis (MCA)

- ▶ An exploratory analysis methodology for multivariate datasets with categorical variables.
- ▶ In basic form, it is PCA on dummy variables that represent the categorical variables.
- ▶ Illustrate with the health utilities index (HUI) variables from the Canadian Community Health Survey - Healthy Aging

HUI Data

- The “NOT STATED” response to questions is missing data.

```
library(tidyverse)
hui <- read.csv("HUI.csv.gz", na.strings = "NOT\ STATED")
hui[hui=="NA"] <- NA
hui <- na.omit(hui)
names(hui)
```

```
## [1] "GEO_PRV" "GEOGMA2" "DHHGAGE" "DHH_SEX" "HUIDCOG" "HUIGDEX"
## [7] "HUIDEMO" "HUIGHER" "HUIDHSI" "HUIGMOB" "HUIGSPE" "HUIGVIS"
## [13] "WTS_M"
```

```
dim(hui)
```

```
## [1] 30106    13
```

Summaries

```
summary(hui) # WTS_M are sampling weights
```

```
##      GEO_PRV      GEOGMA2      DHHGAGE      DHH_SEX
##  ONT :6377   CMA      :16980  55 TO 59 YEARS:4727  FEMALE:17118
##  QUE :5154   NON - CMA:13126  60 TO 64 YEARS:4471  MALE :12988
##  BC  :3747
##  AB  :2672
##  NS  :2217
##  NB  :2151
##  (Other):7788      (Other)      :7352
##
##      HUIDCOG      HUIDDEX      HUIDEMO
##  COG. ATT. LEVE 1:21131  LIM. HANDS/F : 379  EMOT. ATT. LEV.1:22542
##  COG. ATT. LEVE 2: 743   NA      : 0      EMOT. ATT. LEV.2: 6161
##  COG. ATT. LEVE 3: 5713  USE OF HANDS/F.:29727  EMOT. ATT. LEV.3: 1086
##  COG. ATT. LEVE 4: 1847      EMOT. ATT. LEV.4: 254
##  COG. ATT. LEVE 5: 593      EMOT. ATT. LEV.5: 63
##  COG. ATT. LEVE 6: 79      NA      : 0
##  NA      : 0
##
##      HUIGHER      HUIDHSI      HUIGMOB
##  NA      : 0      0.973 : 8787  NA      : 0
##  NO PROBLEMS :26463  0.905 : 3763  NEED MECH. SUPP: 2332
##  PROB./CORR. : 2450  1      : 2966  NO AID REQUIRED: 464
##  PROB./NOT CORR.: 1193  0.931 : 1058  NO PROBLEMS :26478
##      0.842 : 752  REQUIRES HELP : 832
##      0.919 : 719
##      (Other):12061
##
##      HUIGSPE      HUIGVIS      WTS_M
##  NA      : 0      NA      : 0      Min. : 10.00
##  NO PROBLEMS :29879  NO PROBLEMS : 6386  1st Qu.: 91.74
##  PARTIAL/NOT UND.: 227  VISUAL P. UNCOR.: 975  Median : 231.26
##      VISUAL PROB. COR:22745  Mean : 445.34
##      3rd Qu.: 518.12
##      Max. :23740.26
##
```

- ▶ Cognitive function (our focus) with levels:
 1. Able to remember most things, think clearly and solve day to day problems
 2. Able to remember most things, but have a little difficulty when trying to think and solve day to day problems
 3. Somewhat forgetful, but able to think clearly and solve day to day problems
 4. Somewhat forgetful, and have a little difficulty when trying to think or solve day to day problems
 5. Very forgetful, and have great difficulty when trying to think or solve day to day problems
 6. Unable to remember anything at all, and unable to think or solve day to day problems

```
library(dplyr)
levels(hui$HUIDCOG) <- c(as.character(1:6), "NA")
hui %>% group_by(HUIDCOG) %>% summarize(n = sum(WTS_M))
```

```
## # A tibble: 6 x 2
##   HUIDCOG      n
##   <fct>      <dbl>
## 1 1      9965049.
## 2 2      291964.
## 3 3     2293952.
## 4 4      660207.
## 5 5      173239.
## 6 6       23129.
```

Pairwise summaries

► Relationship between HUIDCOG and others

```
stab <- hui %>% group_by(DHHGAGE,HUIDCOG) %>%  
  summarize(n = sum(WTS_M)) %>% spread(HUIDCOG,n)  
stab[,2:7] <- round(stab[,2:7]/rowSums(stab[,2:7]),3)  
stab
```

```
## # A tibble: 9 x 7  
## # Groups:   DHHGAGE [9]  
##   DHHGAGE      `1`      `2`      `3`      `4`      `5`      `6`  
##   <fct>      <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 45 TO 49 YEARS 0.762 0.022 0.165 0.042 0.008 0.001  
## 2 50 TO 54 YEARS 0.775 0.025 0.145 0.043 0.012 0.001  
## 3 55 TO 59 YEARS 0.772 0.017 0.161 0.037 0.013 0  
## 4 60 TO 64 YEARS 0.754 0.016 0.18 0.04 0.009 0.002  
## 5 65 TO 69 YEARS 0.774 0.023 0.161 0.035 0.006 0.001  
## 6 70 TO 74 YEARS 0.715 0.016 0.204 0.052 0.01 0.003  
## 7 75 TO 79 YEARS 0.666 0.029 0.212 0.072 0.019 0.002  
## 8 80 TO 84 YEARS 0.634 0.024 0.202 0.112 0.025 0.003  
## 9 85 AND OLDER 0.536 0.043 0.21 0.132 0.06 0.017
```

Pairwise summaries, cont.

► Relationship between HUIDCOG and HUIDEX.

```
stab <- hui %>% group_by(HUIDEX,HUIDCOG) %>%  
  summarize(n = sum(WTS_M)) %>% spread(HUIDCOG,n)  
  
stab[,2:7] <- round(stab[,2:7]/rowSums(stab[,2:7]),3)  
stab
```

```
## # A tibble: 2 x 7  
## # Groups:   HUIDEX [3]  
##   HUIDEX      `1`    `2`    `3`    `4`    `5`    `6`  
##   <fct>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 LIM. HANDS/F 0.464 0.03 0.204 0.146 0.132 0.024  
## 2 USE OF HANDS/F. 0.746 0.022 0.171 0.048 0.012 0.002
```

► And so on ...

MCA with dummy variables

- ▶ We could expand the categorical variables as dummy variables and do a biplot of the result, but it is not straightforward to incorporate the weights

MCA with FactoMineR

- ▶ The R package FactoMineR includes many useful functions for multivariate data analysis, including MCA.
 - ▶ Their `MCA()` function includes a weighting argument.

```
library(FactoMineR)
h <- sample_n(hui,size=1000)
hHUI <- select(h,HUIDCOG,HUIGDEX,HUIDEMO,HUIGHER,HUIGMOB,
               HUIGSPE,HUIGVIS)
res.mca <- MCA(hHUI,row.w = h$WTS.M)
```

MCA factor map

