Chapter 10: Usupervised Learning

WHEN A USER TAKES A PHOTO, THE APP SHOULD CHECK WHETHER THEY'RE IN A NATIONAL PARK...
SURE, EASY GIS LOOKUP. GIMME A FEW HOURS.
...AND CHECK WHETHER THE PHOTO IS OF A BIRD.
I'LL NEED A RESEARCH TEAM AND FIVE YEARS.
IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.

Credit: https://xkcd.com/1425/

This chapter will focus on methods intended for the setting in which we only have a set of features $X_1, \ldots, X_p$ measured on n observations.

1 The Challenge of Unsupervised Learning

Supervised learning is a well-understood area.

In contrast, unsupervised learning is often much more challenging.

Unsupervised learning is often performed as part of an *exploratory data analysis*.

It can be hard to assess the results obtained from unsupervised learning methods.

Techniques for unsupervised learning are of growing importance in a number of fields.

2 Principal Components Analysis

We have already seen principal components as a method for dimension reduction.

Principal Components Analysis (PCA) refers to the process by which principal components are computed and the subsequent use of these components to understand the data.

Apart from producing derived variables forr use in supervised learning, PCA also serves as a tool for data visualization.

2.1 What are Principal Components?

Suppose we wish to visualize n observations with measurements on a set of p features as part of an exploratory data analysis.

Goal: We would like to find a low-dimensional representation of the data that captures as much of the information as possible.

PCA provides us a tool to do just this.

Idea: Each of the n observations lives in p dimensional space, but not all of these dimensions are equally interesting.

The *first principal component* of a set of features $X_1, \dots, X_p$ is the normalized linear combination of the features

that has the largest variance.

Given a $\boldsymbol{n} \times \boldsymbol{p}$ data set $\boldsymbol{X}$, how do we compute the first principal component?

There is a nice geometric interpretation for the first principal component.

After the first principal component Z_1 of the features has been determined, we can find the second principal component, Z_2. The second principal component is the linear combination of $X_1, \ldots, X_p$ that has maximal variance out of all linear combinations that are uncorrelated with Z_1.

Once we have computed the principal components, we can plot them against each other to produce low-dimensional views of the data.

```r
str(USArrests)
```

```
## 'data.frame':    50 obs. of  4 variables:
##  $ Murder  : num  13.2 10 8.1 8.8 9 7.9 3.3 5.9 15.4 17.4 ...
##  $ Assault : int  236 263 294 190 276 204 110 238 335 211 ...
##  $ UrbanPop: int  58 48 80 50 91 78 77 72 80 60 ...
##  $ Rape    : num  21.2 44.5 31 19.5 40.6 38.7 11.1 15.8 31.9 25.8 ...
```

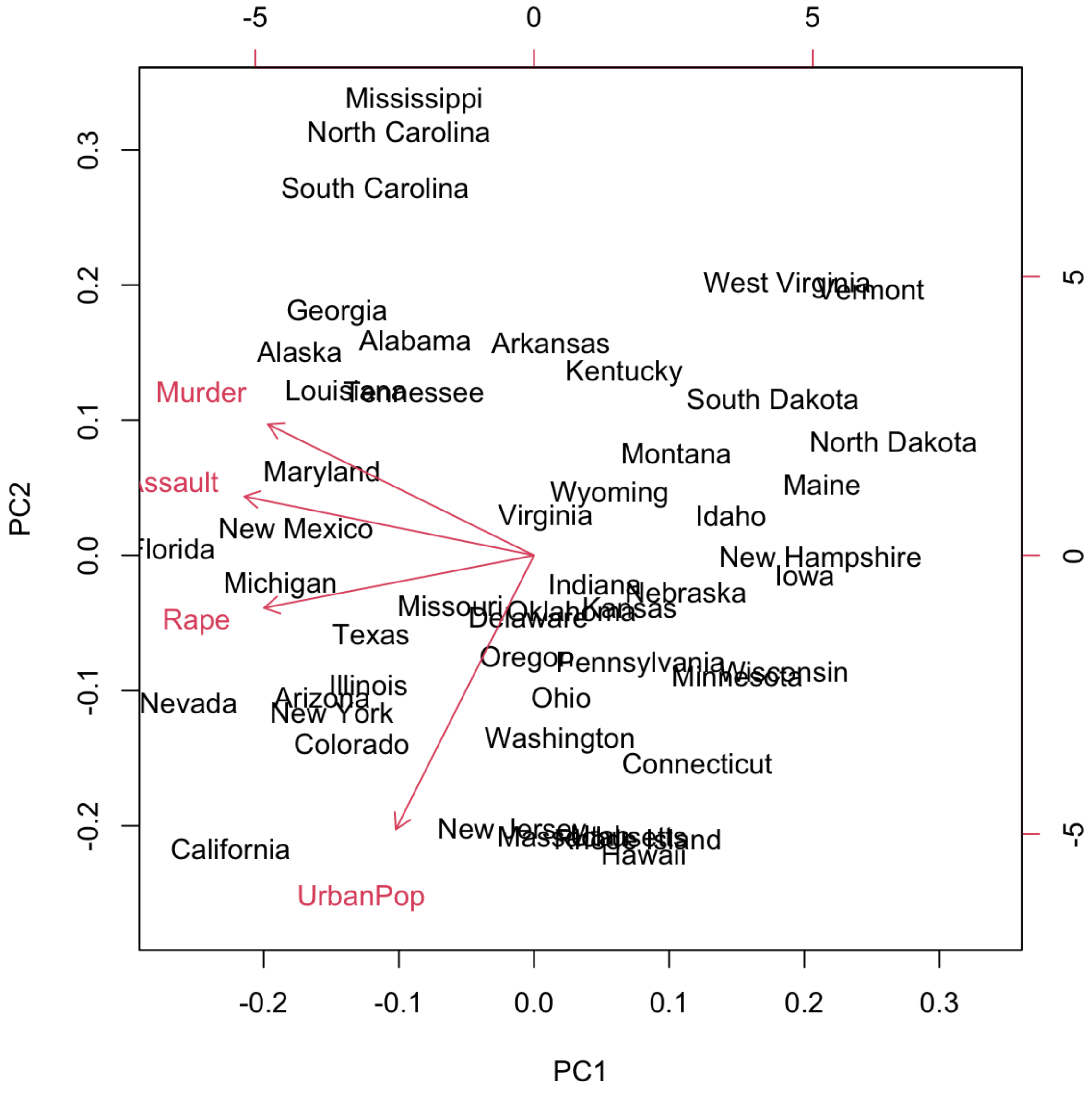
```r
pca <- prcomp(USArrests, center = TRUE, scale = TRUE) # get loadings

summary(pca) # summary
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.5749 0.9949 0.59713 0.41645
## Proportion of Variance 0.6201 0.2474 0.08914 0.04336
## Cumulative Proportion  0.6201 0.8675 0.95664 1.00000
```

```r
pca$rotation # principal components loading matrix
```

```
##                 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
```

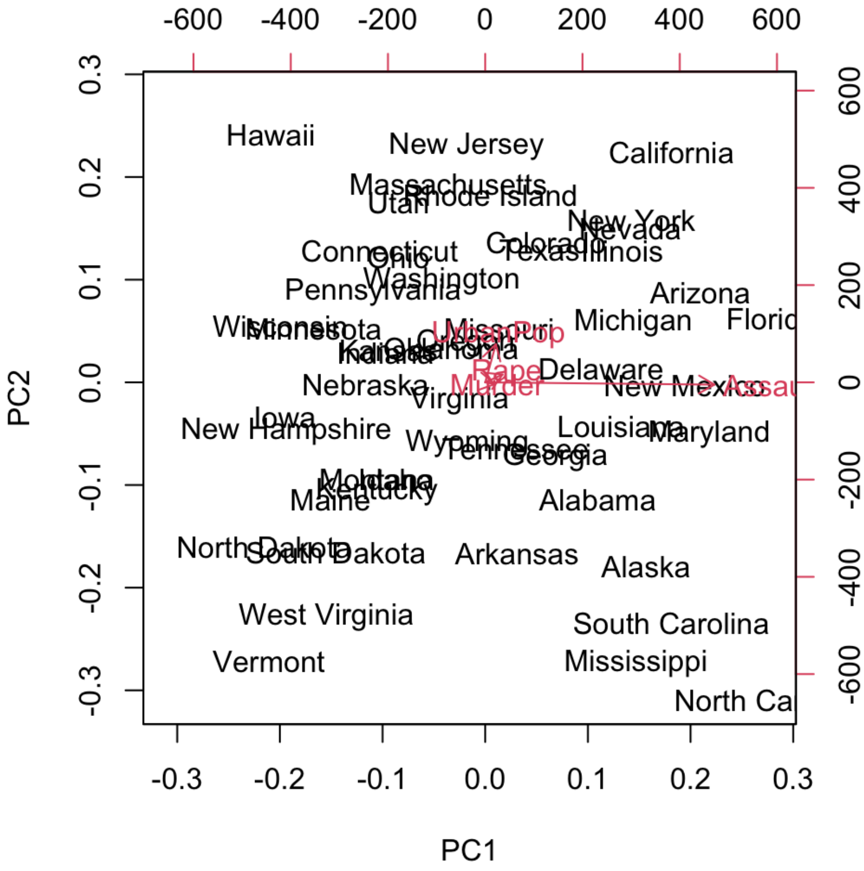
```r
## plot scores + directions
biplot(pca)
```


-5
0
5
Mississippi
North Carolina
South Carolina
West Virginia
Vermont
Georgia
Alabama
Arkansas
Alaska
Kentucky
Murder
Louisiana
Tennessee
South Dakota
North Dakota
Montana
Maryland
Maine
ssault
Wyoming
Virginia
Idaho
New Mexico
lorida
New Hampshire
Iowa
Michigan
Indiana
Nebraska
Rape
Missouri
Delaware
Oklahoma
Kansas
Texas
Oregon
Pennsylvania
Wisconsin
Minnesota
Illinois
Arizona
New York
Nevada
Ohio
Colorado
Washington
Connecticut
New Jersey
Massachusetts
Utah
Rhode Island
Hawaii
California
UrbanPop
0.3
0.2
0.1
0.0
-0.1
-0.2
5
0
-5
-0.2
-0.1
0.0
0.1
0.2
0.3
PC1
PC2

2.2 Scaling Variables

We've already talked about how when PCA is performed, the varriables should be centered to have mean zero.

This is in contrast to other methods we've seen before.



2.3 Uniqueness

Each principal component loading vector is unique, up to a sign flip.

Similarly, the score vectors are unique up to a sign flip.

2.4 Proportion of Variance Explained

We have seen using the `USArrests` data that e can summarize 50 observations in 4 dimensions using just the first two principal component score vectors and the first two principal component vectors.

Question:

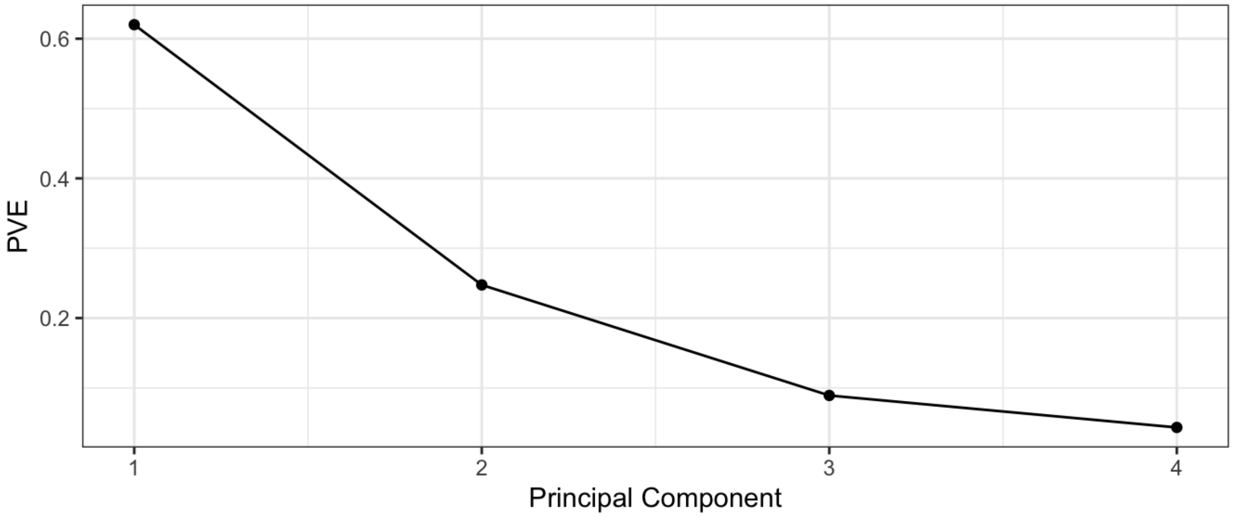
More generally, we are interested in knowing the *proportion of vriance explained (PVE)* by each principal component.

2.5 How Many Principal Components to Use

In general, a $ntimesp$ matrix $\boldsymbol{X}$ has $\min(n-1,p)$ distinctt principal components.

Rather, we would like to just use the first few principal components in order to visualize or interpret the data.

We typically decide on the number of principal components required by examining a *scree plot*.



2.6 Other Uses for Principal Components

We've seen previously that we can perform regression using the principal component score vectors as features for dimension reduction.

Many statistical techniques can be easily adapted to use the $n \times M$ matrix whose columns are the first $M << p$ principal components.

This can lead to *less noisy* results.