

# Chapter 10: Unsupervised Learning



IN CS, IT CAN BE HARD TO EXPLAIN  
THE DIFFERENCE BETWEEN THE EASY  
AND THE VIRTUALLY IMPOSSIBLE.

Credit: <https://xked.com/1425/>

This chapter will focus on methods intended for the setting in which we only have a set of features  $X_1, \dots, X_p$  measured on  $n$  observations. *no longer have  $Y$ !*

*We are not interested in prediction because we have no associated  $Y$ .*

Goal: *discovering interesting things about measurements  $X_1, \dots, X_p$*

- Is there an informative way to plot the data?*
- Can we discover subgroups among variables or observations?*

# 1 The Challenge of Unsupervised Learning

Supervised learning is a well-understood area.

You now have a good grasp of supervised learning.

If you were asked to predict a binary response -

Logistic regression, SVM, LDA, classification trees, RF, bagged trees, boosted trees,

and a clear understanding of how to assess your results -

validation on an independent test set

Cross-validation error.

In contrast, unsupervised learning is often much more challenging.

More subjective, no simple goal for the analysis, e.g. prediction.

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

1st part of an analysis, before fitting any models.

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

No universally accepted mechanism for performing cross-validation or validation on a test set.

Because there is no way to "check our work" with no response variable.

→ We don't know the true answer!

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

cancer research assay gene expression levels in 100 patients and look for subgroups among cancer sample to better understand the disease.

online shopping identify similar groups of shoppers show preferential items that they might be particularly interested.

Entity resolution Many noisy databases without unique identifying attributes → can we find the matches/links?

## 2 Principal Components Analysis

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

When faced w/ a large set of correlated variables, we can use principal components to summarize this set w/ a smaller number of representative variables that collectively explain most of the variability in the original data set.

Principal component directions = directions in feature space along which original data are highly variable.  
↓  
define lines and subspaces that are close as possible to the data cloud.

Principal component regression

used principal components as predictors in a regression model instead of original variables.

*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.

Unsupervised approach

involves only features  $X_1, \dots, X_p$ , no response  $Y$ .

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

Visualize observations or of variables.

as a tool for visualization / EDA.

## 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.

We could do this by examining 2D plots of the data which contain  $n$  observations based on 2 features.

ggpairs, pairs  
 $\Rightarrow \binom{p}{2} = \frac{p(p-1)}{2}$  plots. e.g. w/  $p=10 \Rightarrow 45$  plots!

- Too many to look at
- likely no plot will be informative because they contain a small fraction of information present in our data.

For visualization in high dimensions

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

Then plot observations in low-dimensional space.

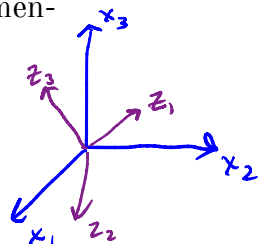
PCA provides us a tool to do just this.

It finds low dimensional representation of a data set that contains as much as possible of the variation (information).

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

PCA seeks a small number of dimensions that are as interesting as possible.

"interesting" = amount the observations vary along each dimension.



Each dimension found by PCA is a linear combination of the  $p$  factors.

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

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + \dots + \phi_{p1} X_p$$

normalized:  $\sum_{j=1}^p \phi_{j1}^2 = 1$  ← otherwise could result in arbitrarily large variances

$\phi_{11}, \dots, \phi_{p1}$  are called "loadings" of the first principal component.

$$\phi_1 = (\phi_{11}, \dots, \phi_{p1})^T = \text{"loading vector"}$$

that has the largest variance. of  $Z_1$ ,

$$\text{Var}(\phi_{11} X_1 + \dots + \phi_{p1} X_p).$$

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

① Assume each variable has been centered (i.e. columns have mean 0) — only care about maximizing variance.  
(do it)

② Look for linear combinations of the form

$$Z_{i1} = \phi_{11} x_{i1} + \phi_{21} x_{i2} + \dots + \phi_{p1} x_{ip}$$

w/ largest sample variance. subject to  $\sum_{j=1}^p \phi_{j1}^2 = 1$ .

i.e. solve the following optimization problem:

$$\text{maximize}_{\phi_{11}, \dots, \phi_{p1}} \left\{ \frac{1}{n} \sum_{i=1}^n \left( \sum_{j=1}^p \phi_{j1} x_{ij} \right)^2 \right\}$$

$$\text{subject to } \sum_{j=1}^p \phi_{j1}^2 = 1$$

can write this way because columns centered  
 $\Rightarrow \frac{1}{n} \sum_{i=1}^n x_{ij} = 0$   
 $\Rightarrow \frac{1}{n} \sum_{i=1}^n Z_{i1} = 0$   
 $\Rightarrow$  this is sample variance of  $Z_{i1}, i=1, \dots, n$ .

$Z_{11}, \dots, Z_{n1}$  are called "scores" of 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  $X_1, \dots, X_p$  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.

```
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 ...
```

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

```
summary(pca) # summary
```

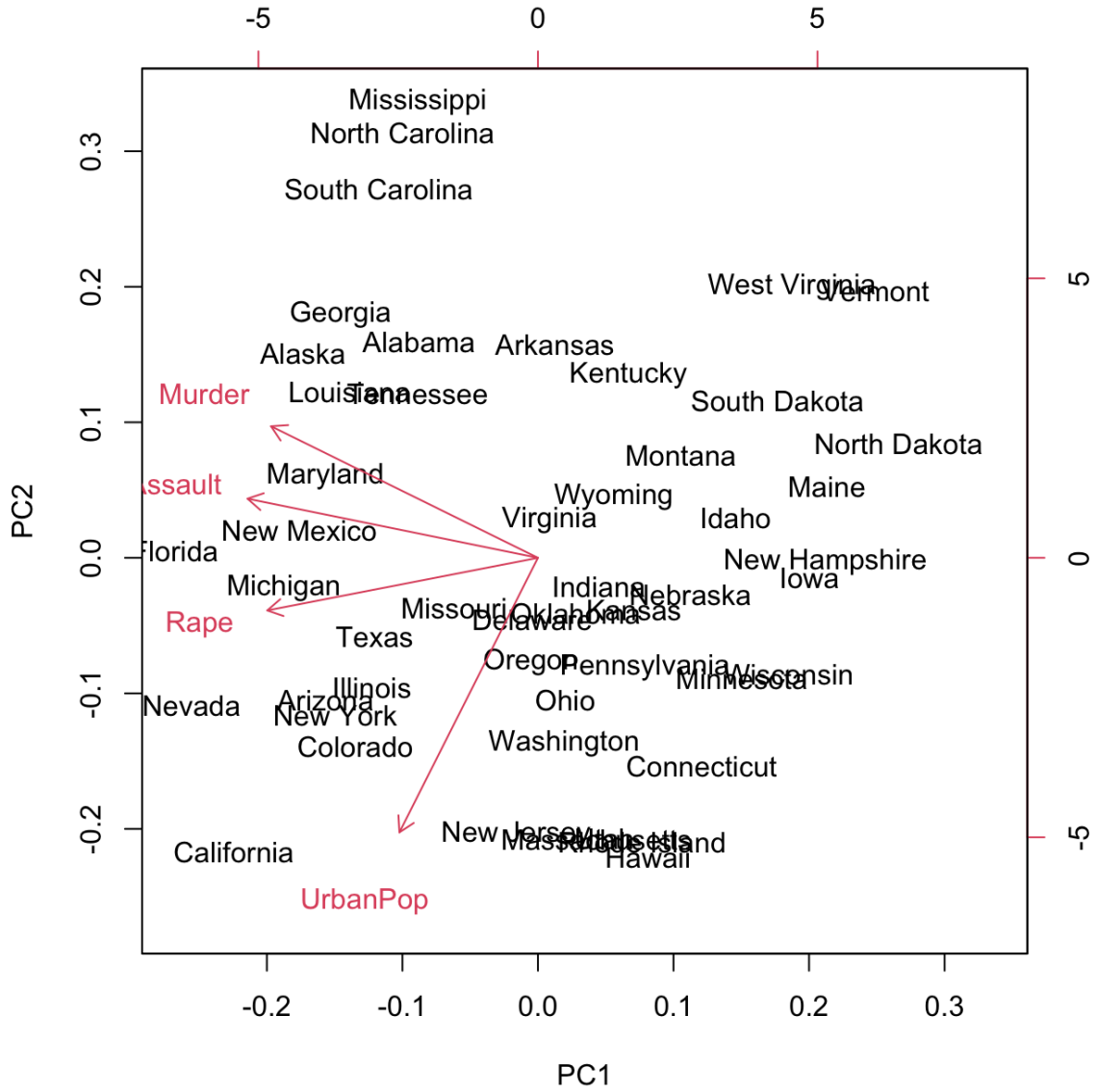
```
## Importance of components:
##              PC1      PC2      PC3      PC4
## Standard deviation  1.5749 0.9949 0.59713 0.41645
## Proportion of Variance 0.6201 0.2474 0.08914 0.04336
## Cumulative Proportion 0.6201 0.8675 0.95664 1.00000
```

```
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
```

```
## plot scores + directions
```

```
biplot(pca)
```

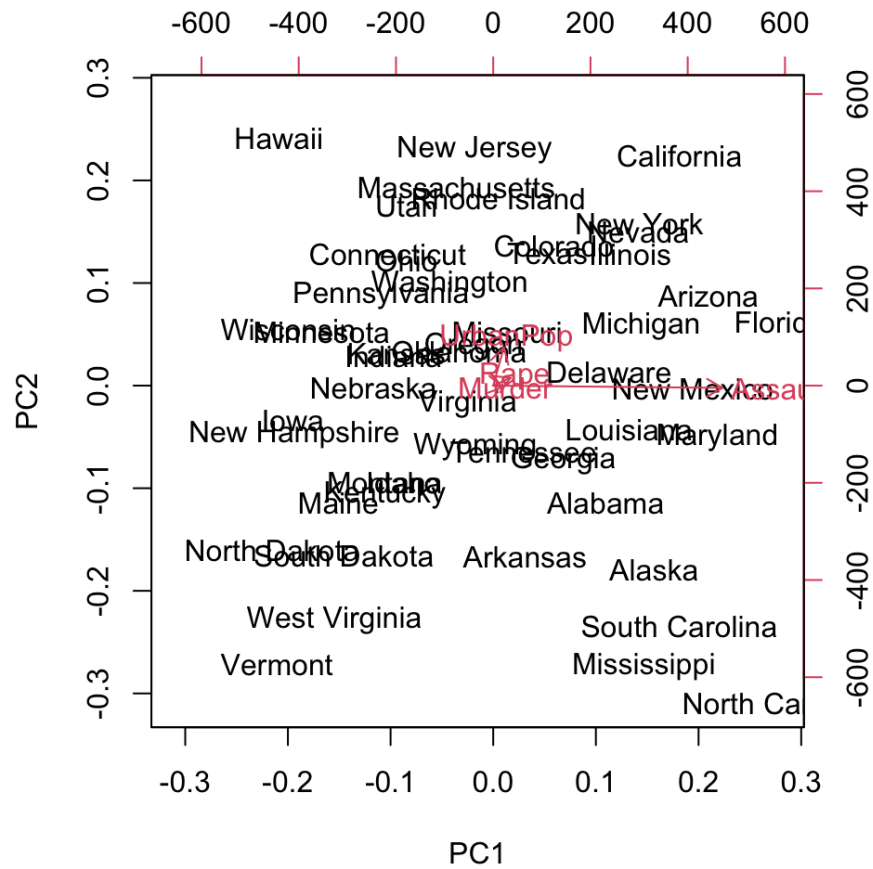




## 2.2 Scaling Variables

We've already talked about how when PCA is performed, the variables 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 we can summarize 50 observations in 4 dimensions using just the first two principal component score vectors and the first two principal component vectors.

**Question:**

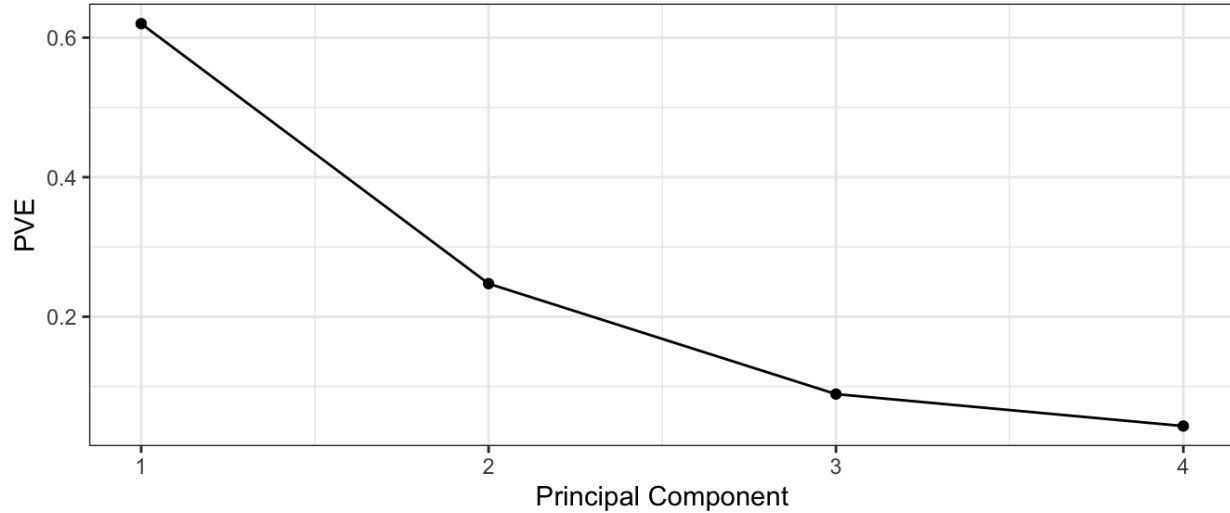
More generally, we are interested in knowing the *proportion of variance explained (PVE)* by each principal component.

## 2.5 How Many Principal Components to Use

In general, a  $n \times p$  matrix  $\mathbf{X}$  has  $\min(n - 1, p)$  distinct 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  $M \ll p$  principal components.

This can lead to *less noisy* results.