Chp. 12: Unsupervised Learning

Unsupervised Learning

- In supervised learning, we typically have a set of p features $X_1, X_2, ..., X_p$ measured on n observations and a response Y also measured on those same n observations
- Unsupervised learning can be more challenging b/c we don't have any Y labels to test our model against a 'correct' answer.

Principal Components Analysis

- Find a low dimensional representation of the data that captures as much as possible of the variation.
- PCA finds a small number of dimensions (principal components) that are each a linear combination of the p features.

Principal Components Analysis

The first principal component of a set of features $X_1, X_2, ..., X_p$ is the normalized linear combination of the features that has the largest variance

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

Principal Components Analysis

The first **principal component** of a set of features $X_1, X_2, ..., X_p$ is the normalized linear combination of the features that has the largest variance

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

Principal Components Analysis

The first **principal component** of a set of features $X_1, X_2, ..., X_p$ is the normalized linear combination of the features that has the largest variance

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

First **principal component score** for the *i*th observation:

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + ... + \phi_{p1}x_{ip}$$

Principal Components Analysis

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

$\phi_{11}, ..., \phi_{p1}$ are the *loadings* of the first principal component (PC) and can be written as the loading vector
$\phi_1 = (\phi_{11} \quad \phi_{21} \quad ... \quad \phi_{1p})^T$

Principal Components Analysis

The first principal component of a set of features $X_1, X_2, ..., X_p$ is the **normalized** linear combination of the features that has the largest variance

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

By normalized, we mean that $\sum_{j=1}^{p} \phi_{j1}^2 = 1$. This is done to ensure the ϕ's do not become arbitrarily large.

PCA - Computing the First PC

Goal: look for the linear combination of the sample feature values that has the largest sample variance, subject to normalization.

PCA - Computing the First PC

Goal: look for the linear combination of the sample feature values that has the largest sample variance, subject to normalization.

$$\begin{aligned} &\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximize}} \quad \left\{ \frac{1}{n} \sum_{i=1}^{n} (z_{i1} - \bar{z}_{i1})^2 \right\} \\ &\text{subject to} \quad \sum_{j=1}^{p} \phi_{j1}^2 = 1 \end{aligned}$$

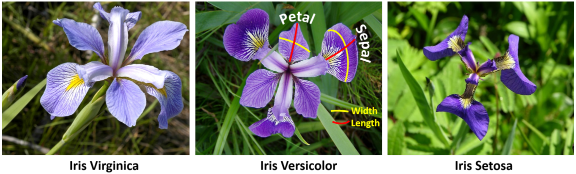
PCA - Computing the Next PC

After the first principal component, Z_1 has been determined, we find the second PC, Z_2 as the linear combination of $X_1, ..., X_p$ that has maximal variance out of all linear combinations that are uncorrelated with Z_1.

Imposing this constraint is equivalent to constraining the loading vector ϕ_2 to be orthogonal (perpendicular) to to ϕ_1.

PCA - Interpretation with Biplots

Biplots display both the principal compononent scores and the principle component loadings.



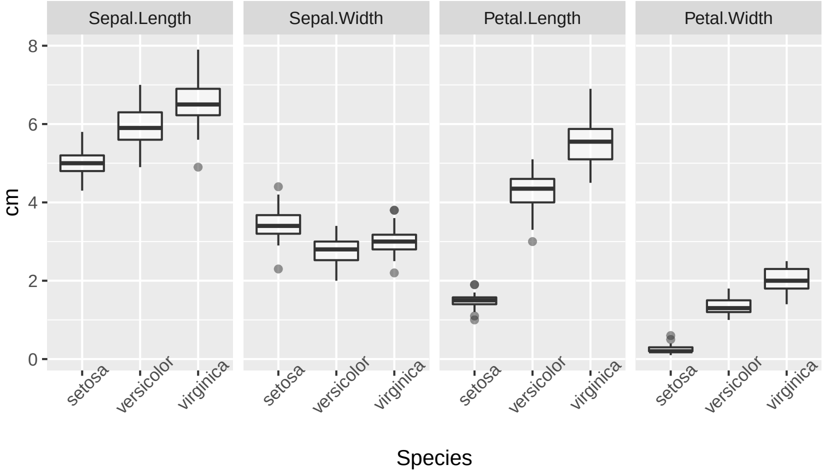
Iris Virginica

Iris Versicolor

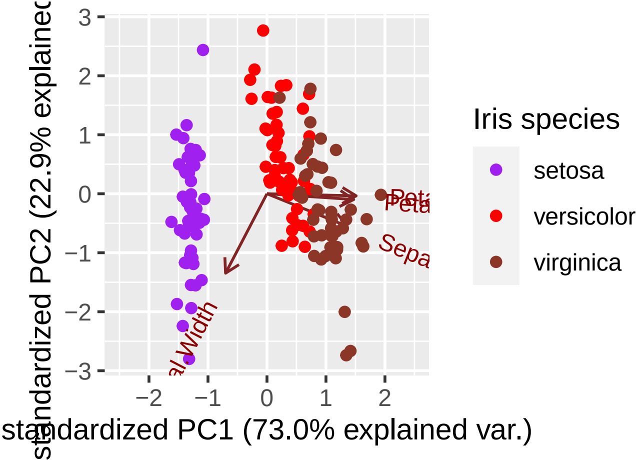
Iris Setosa

Figure 1: iris

PCA - Interpretation with Biplots



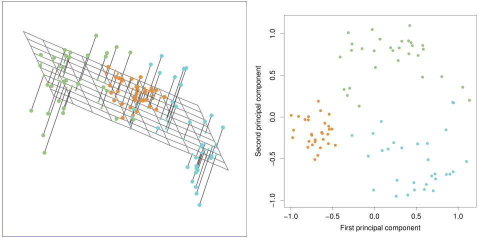
PCA - Interpretation with Biplots



```
##                        PC1         PC2
## Sepal.Length  0.5210659 -0.37741762
## Sepal.Width  -0.2693474 -0.92329566
## Petal.Length  0.5804131 -0.02449161
## Petal.Width   0.5648565 -0.06694199
```

PCA - Alt Interpretation

Principal components provide low-dimensional linear surfaces that are closest to the observations in terms of the average squared Euclidean distance.



ISL Fig. 10.2: PC1 and PC2 give the coordinates of the projection of 90 observations onto a plane that best fits the data with the variance in the plane maximized.

Preprocessing

Most importantly, before PCA is performed the variables should be **centered** to have mean zero so that the resulting components are only looking at the variance within the dataset, and not capturing the overall mean of the dataset as an important variable (dimension)

Preprocessing

Unless features are measured in the same units, we usually scale each variable to have standard deviation one before PCA. Here, 'Assault' measured as number per 100,000 people ($\sigma^2 = 6945$).

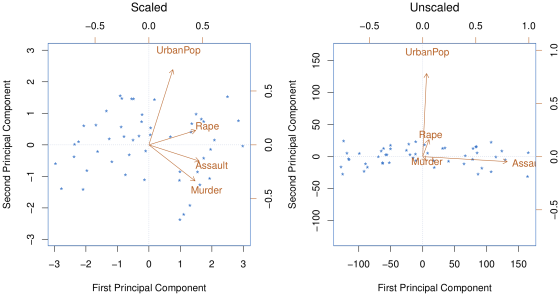


Figure 2: fig10.3a

Proportion of Variance Explained

(After centering & scaling) the total variance present in a dataset is:

$$\sum_{j=1}^{p} \mathrm{Var}(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2$$

and the variance explained by the mth PC is:

$$\frac{1}{n} \sum_{i=1}^{n} z_{im}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2$$

Proportion of Variance Explained

The proportion of variance explained is then just:

$$\frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{jm}x_{ij}\right)^2}{\sum_{j=1}^{p}\sum_{i=1}^{n}x_{ij}^2}$$

Proportion of Variance Explained

How many PCs to keep?

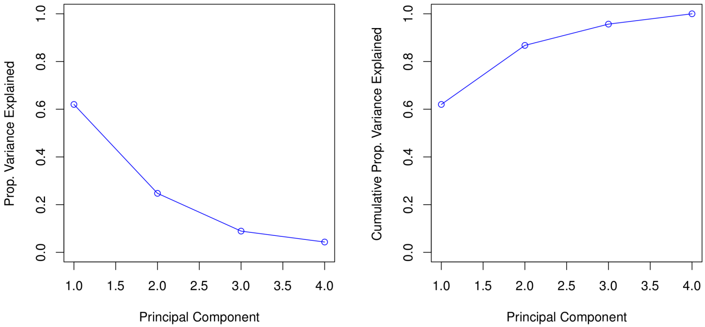


Figure 3: fig10.4

Summary

PCA is a low dimensional representation of the data that captures as much as possible of the variation.

One of the most commomly used methods for unsupervised learning.

Can be important to scale & center variables.