Dimensionality Reduction & Principal Components

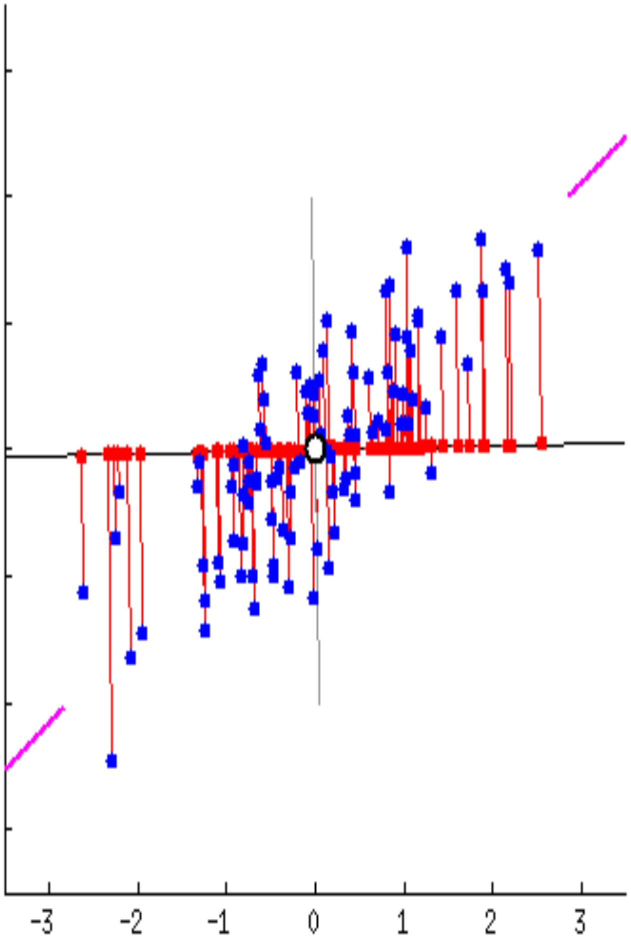
Challenge

- When we have a large number of features, there is often multi-collinearity
  - Design doesn't fill the space
  - One (or more) variable(s) highly correlated with a combination of other variables
- Some times we have more features than observations
  - Design is over-determined

Goal

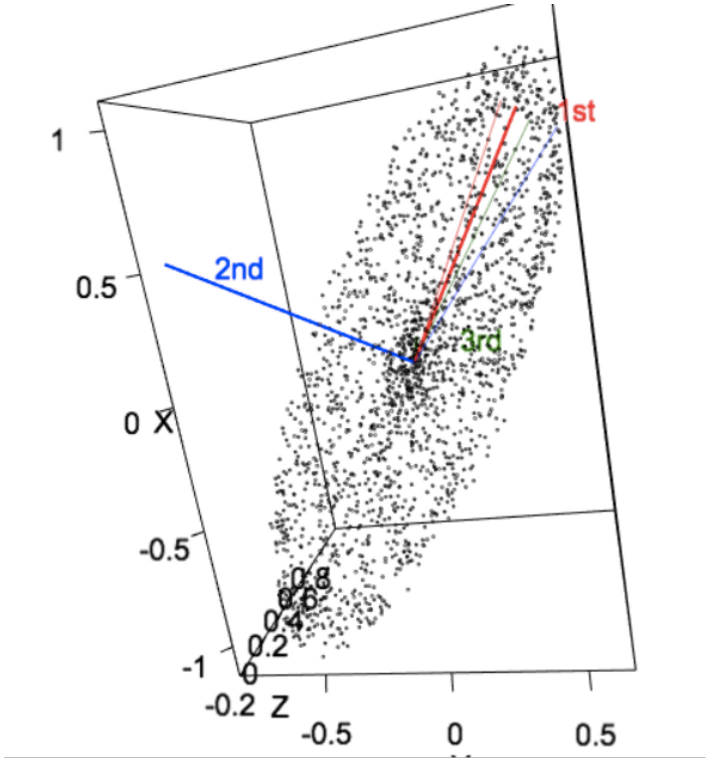
- Create a few new features that are functions (combinations) of original features and that preserve as much information as possible
- Analyze the new features
  - Cluster observations
  - Fit models with new features

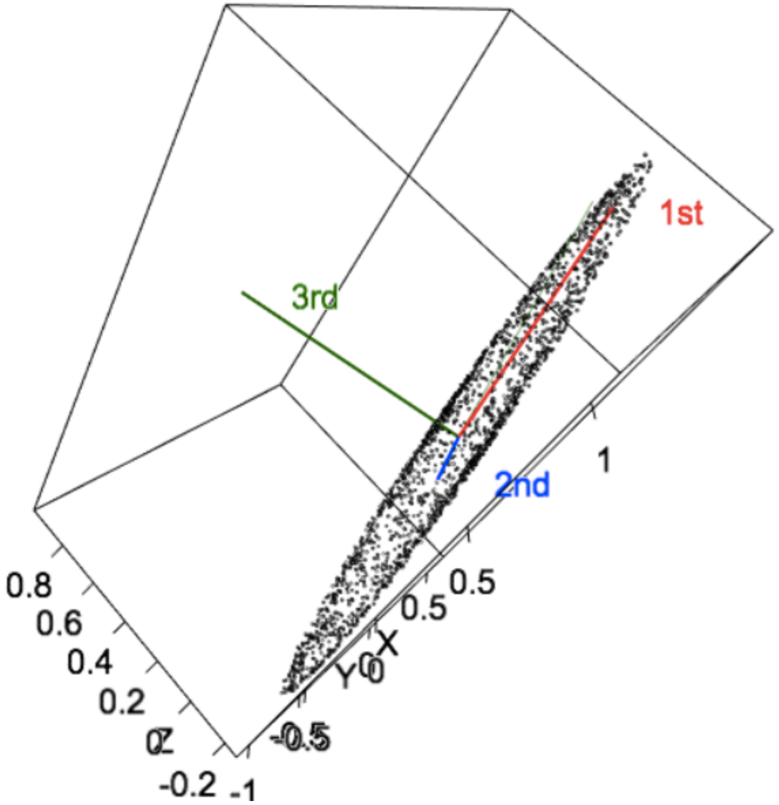
2D Simple Example



Which dimension has the greatest variability?

3D Simple Example





Theory

Principal Component Idea

- Data X = [x_1, x_2, …, x_p]
  - x_j is a feature with n values (one for each record)
- Standardize data Z = [z_1, z_2, …, z_p]
  - z_j has mean 0 and variance 1
- Transform to new basis ZA = W = [w_1, …, w_p]
  - w_1 accounts for max collective variation in Z
  - w_2 orthogonal to w_1 and accounts for max variation, etc.
- Reduce to the first few w_j

Set Up

- Consider the first vector

$$\mathrm{w}_1 = \mathrm{a}_{11}\mathrm{z}_1 + \mathrm{a}_{12}\mathrm{z}_2 + \cdots + \mathrm{a}_{1\mathrm{p}}\mathrm{z}_\mathrm{p}$$

$$= Za_1$$

- Maximize w$_1$ variance,

$$s_1^2 = w_1^t w_1 / (n-1)$$

$$= a_1^t Z^t Z a_1 / (n-1)$$

$$= a_1^t R_{xx} a_1$$

Maximization

- Maximize w$_1$ variance, $s_1^2 = a_1^t R_{xx} a_1$
- Subject to the constraint $a_1^t a_1 = 1$

$$\max_{a_1 \lambda_1} a_1^t R_{xx} a_1 - \lambda_1 (a_1^t a_1 - 1)$$

- Differentiate $2R_{xx} a_1 - 2\lambda_1 a_1$
- Set to 0

$$R_{xx} a_1 = \lambda_1 a_1$$

Eigenvector of R_{xx} and corresponding eigenvalue

Repeat

Properties of the new basis

- w_1 is the first principal component
- The w_j are orthogonal
- w_j has variance λ_j
- The eigenvalues are decreasing
  $$\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p > 0$$
- The sum of the eigenvalues is p, which equals the sum of the variances of z_j

How to use this information

- Collinearity occurs when some of the basis vectors are very short (dimensions are nearly collapsed)
- The relative size of variances (the eigenvalues) serves as indicator for collinearity
- The loadings (the eigenvectors) can reveal important relationships among the features
- The new basis can reveal clusters in the data

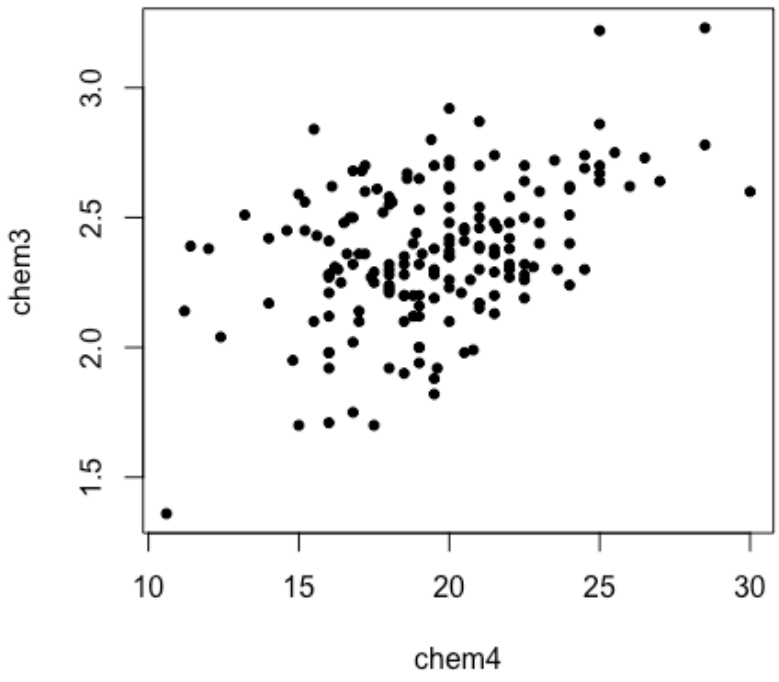
Italian Wines

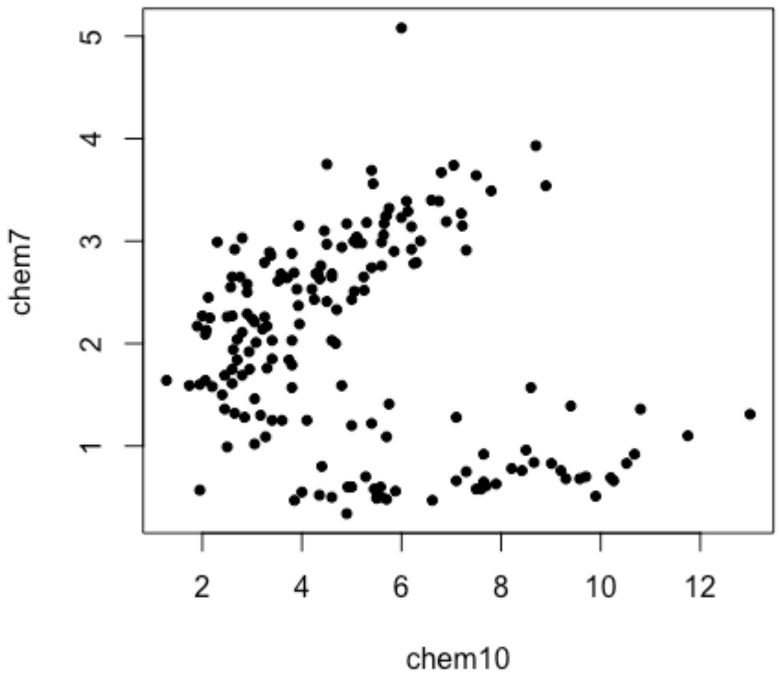
Italian Wines

- Chemical analysis of wines grown in the same region in Italy
- Derived from 3 different cultivars.
- Features - quantities of 13 chemicals found in each wine sample.
- 178 records, one for each sample.

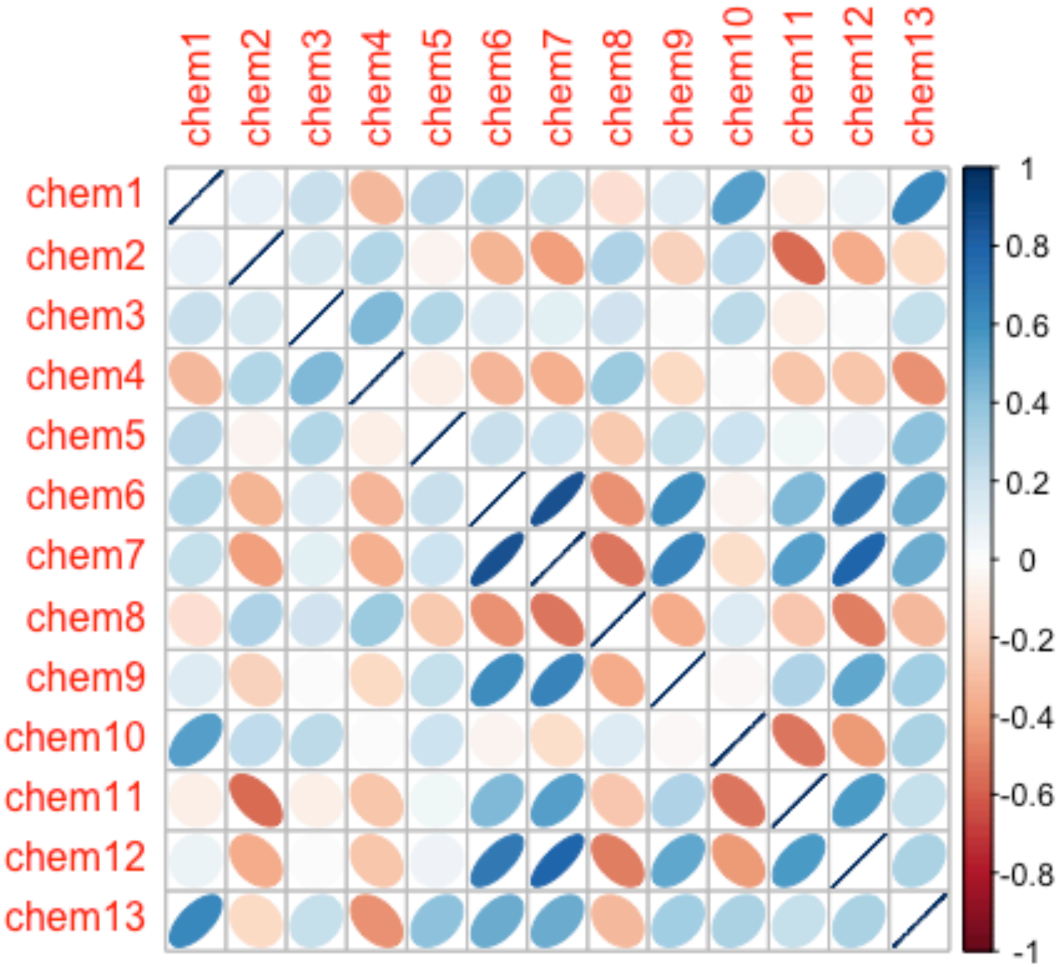
Is the chemical composition useful in clustering the wine samples?

Seeing clusters in 13 dimensions…





Visualizing Pairwise Correlations



Interpretation of Loadings

```
             PC1          PC2         PC3         PC4
chem1  -0.144329395  0.483651548 -0.20738262  0.01785630
chem2   0.245187580  0.224930935  0.08901289 -0.53689028
chem3   0.002051061  0.316068814  0.62622390  0.21417556
chem4   0.239320405 -0.010590502  0.61208035 -0.06085941
chem5  -0.141992042  0.299634003  0.13075693  0.35179658
chem6  -0.394660845  0.065039512  0.14617896 -0.19806835
chem7  -0.422934297 -0.003359812  0.15068190 -0.15229479
chem8   0.298533103  0.028779488  0.17036816  0.20330102
chem9  -0.313429488  0.039301722  0.14945431 -0.39905653
chem10  0.088616705  0.529995672 -0.13730621 -0.06592568
chem11 -0.296714564 -0.279235148  0.08522192  0.42777141
chem12 -0.376167411 -0.164496193  0.16600459 -0.18412074
chem13 -0.286752227  0.364902832 -0.12674592  0.23207086
```

Eigenvector with
greatest variance

Interpretation of Loadings

```
               PC1          PC2         PC3         PC4
chem1  -0.144329395  0.483651548 -0.20738262  0.01785630
chem2   0.245187580  0.224930935  0.08901289 -0.53689028
chem3   0.002051061  0.316068814  0.62622390  0.21417556
chem4   0.239320405 -0.010590502  0.61208035 -0.06085941
chem5  -0.141992042  0.299634003  0.13075693  0.35179658
chem6  -0.394660845  0.065039512  0.14617896 -0.19806835
chem7  -0.422934297 -0.003359812  0.15068190 -0.15229479
chem8   0.298533103  0.028779488  0.17036816  0.20330102
chem9  -0.313429488  0.039301722  0.14945431 -0.39905653
chem10  0.088616705  0.529995672 -0.13730621 -0.06592568
chem11 -0.296714564 -0.279235148  0.08522192  0.42777141
chem12 -0.376167411 -0.164496193  0.16600459 -0.18412074
chem13 -0.286752227  0.364902832 -0.12674592  0.23207086
```

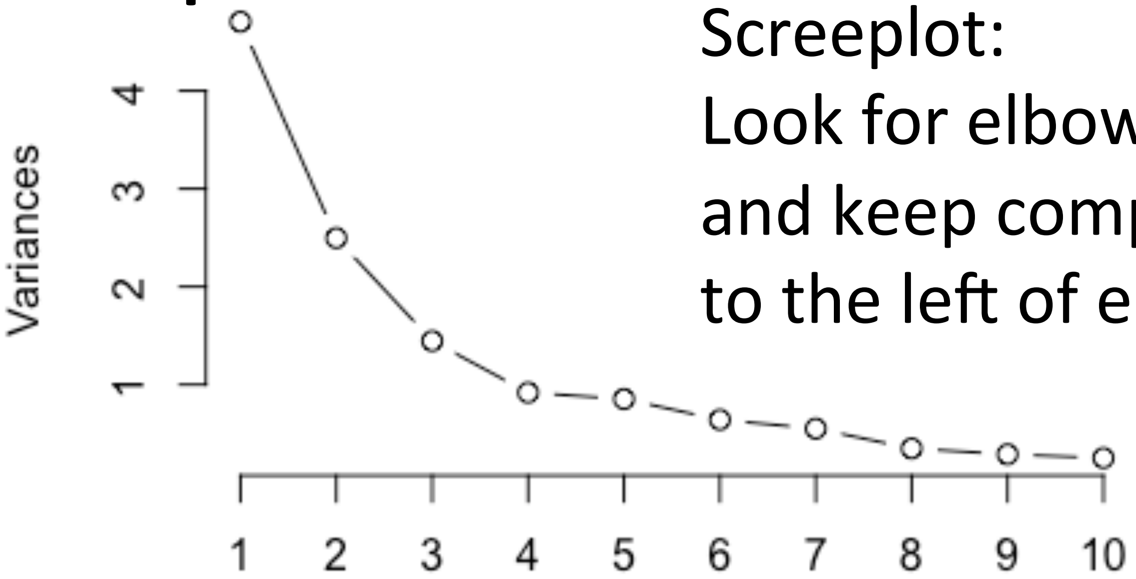
Contrast chem2&4&8
against chem6&7&9&12

Interpretation of Loadings

```
               PC1          PC2         PC3         PC4
chem1  -0.144329395  0.483651548 -0.20738262  0.01785630
chem2   0.245187580  0.224930935  0.08901289 -0.53689028
chem3   0.002051061  0.316068814  0.62622390  0.21417556
chem4   0.239320405 -0.010590502  0.61208035 -0.06085941
chem5  -0.141992042  0.299634003  0.13075693  0.35179658
chem6  -0.394660845  0.065039512  0.14617896 -0.19806835
chem7  -0.422934297 -0.003359812  0.15068190 -0.15229479
chem8   0.298533103  0.028779488  0.17036816  0.20330102
chem9  -0.313429488  0.039301722  0.14945431 -0.39905653
chem10  0.088616705  0.529995672 -0.13730621 -0.06592568
chem11 -0.296714564 -0.279235148  0.08522192  0.42777141
chem12 -0.376167411 -0.164496193  0.16600459 -0.18412074
chem13 -0.286752227  0.364902832 -0.12674592  0.23207086
```

Contrast chem1 and chem10 against chem11

Rules of Thumb for Selecting Components

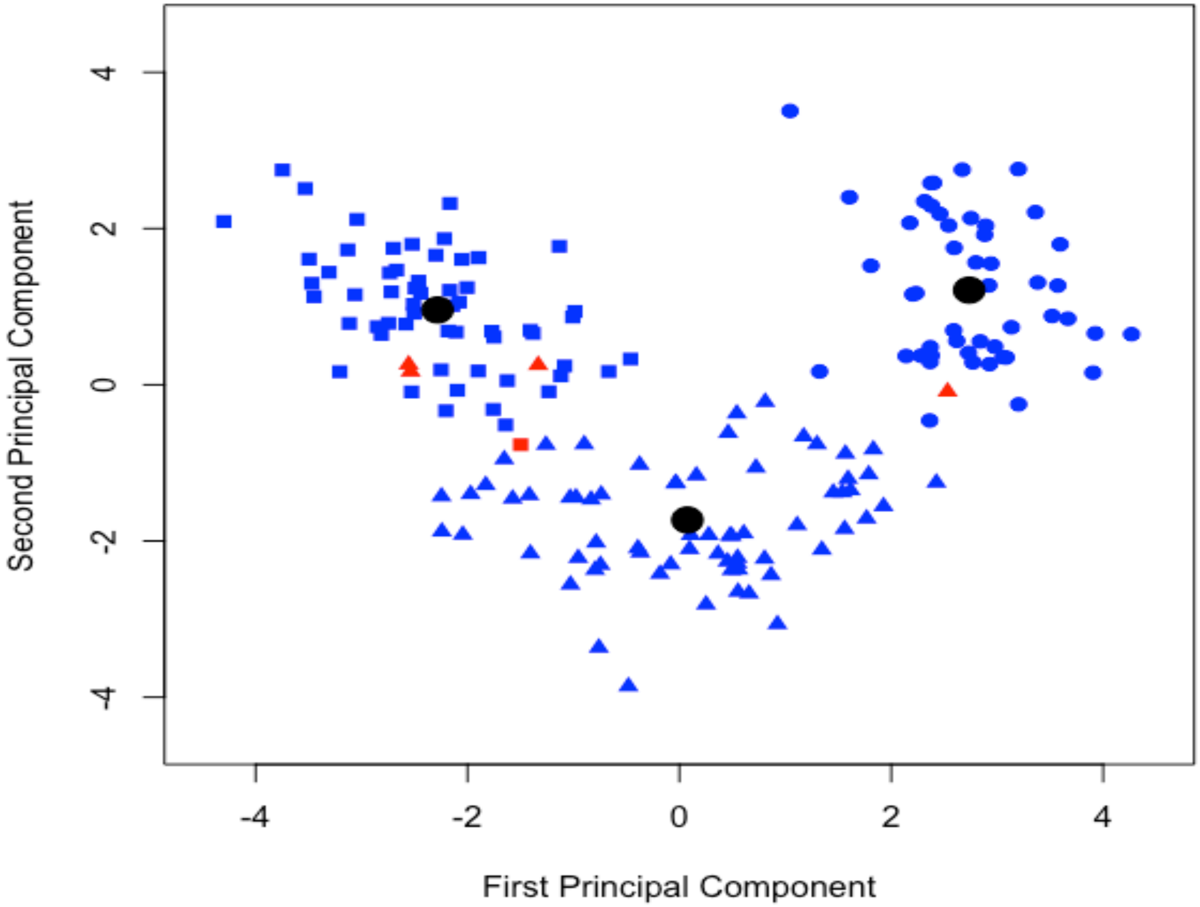


Screeplot:
Look for elbow in plot and keep components to the left of elbow

Choose components such that at least 85% of variance captured

Ratio of largest to smallest variance > 10 indicates collinearity

3-means cluster on PC 1 & 2



4 wines incorrectly clustered

Applications

- Portfolio management in finance
- High throughput genetics data
- Image analysis – pattern recognition and compression
- Intrusion detection in network traffic