Lecture 3: Principal Components Analysis (PCA)

Reading: Sections 6.3.1, 10.1, 10.2, 10.4

STATS 202: Data mining and analysis

September 27, 2019

Recap: The bias variance decomposition

Training data $(x_1, y_1), \ldots, (x_n, y_n)$, a fixed test point x_0.

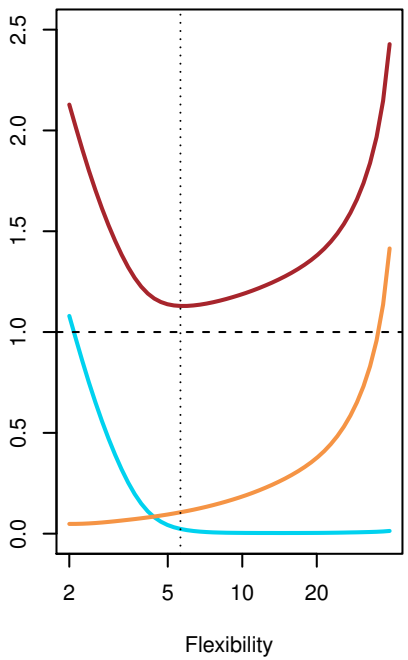
$$y_i = f(x_i) + \varepsilon_i \qquad \text{for } \varepsilon_i \text{ independent, mean } 0.$$

A regression method fit to $(x_1, y_1), \ldots, (x_n, y_n)$ produces the estimate $\hat{f}$. Then, the expected test Mean Squared Error at x_0 satisfies:

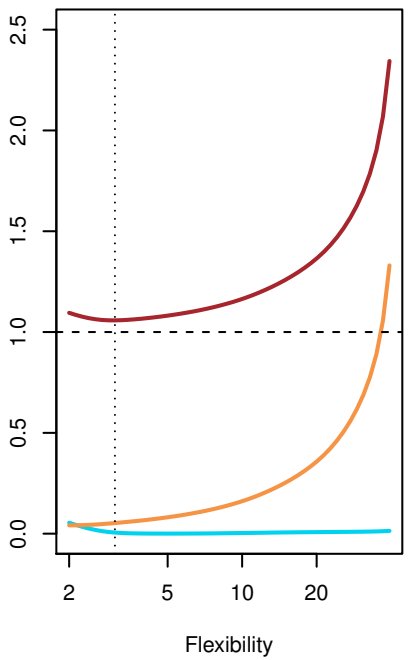
$$E(y_0 - \hat{f}(x_0))^2 = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\varepsilon).$$

Both variance and squared bias are always positive, so to minimize the MSE, you must reach a tradeoff between bias and variance.

Squiggly f, high noise



Linear f, high noise



Squiggly f, low noise

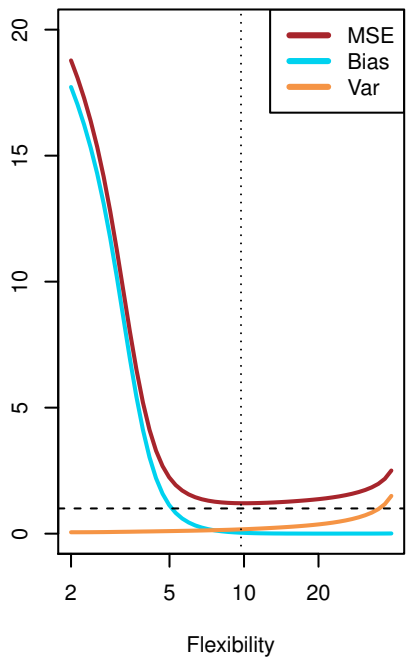


Figure 2.12

Classification problems

In a classification setting, the output takes values in a discrete set.

For example, if we are predicting the brand of a car based on a number of variables, the function f takes values in the set $\{\text{Ford, Toyota, Mercedes-Benz}, \dots\}$.

The model:

$$\cancel{Y = f(X) + \varepsilon}$$

is less relevant in this setting as Y is categorical.

Classification problems

In a classification setting, the output takes values in a discrete set.

For example, if we are predicting the brand of a car based on a number of variables, the function f takes values in the set $\{\text{Ford, Toyota, Mercedes-Benz}, \dots\}$.

Instead, we will adopt the notation:

$$
\begin{aligned}
P(X, Y) &: \text{joint distribution of } (X, Y), \\
P(Y \mid X) &: \text{conditional distribution of } Y \text{ given } X, \\
\hat{y}_i &: \text{prediction for } x_i.
\end{aligned}
$$

Loss function for classification

There are many ways to measure the error of a classification prediction. One of the most common is the 0-1 loss:

$$E(\mathbf{1}(y_0 \neq \hat{y}_0))$$

As with squared error, we can compute average test predition error (called **test error rate** under 0-1 loss) using previously unseen test data $\{(x_i', y_i'); i = 1, \ldots, m\}$:

$$\frac{1}{m} \sum_{i=1}^{m} \mathbf{1}(y_i' \neq \hat{y}_i')$$

Similarly, we can compute the (usually optimistic) **training error rate**

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(y_i \neq \hat{y}_i)$$

Bayes classifier

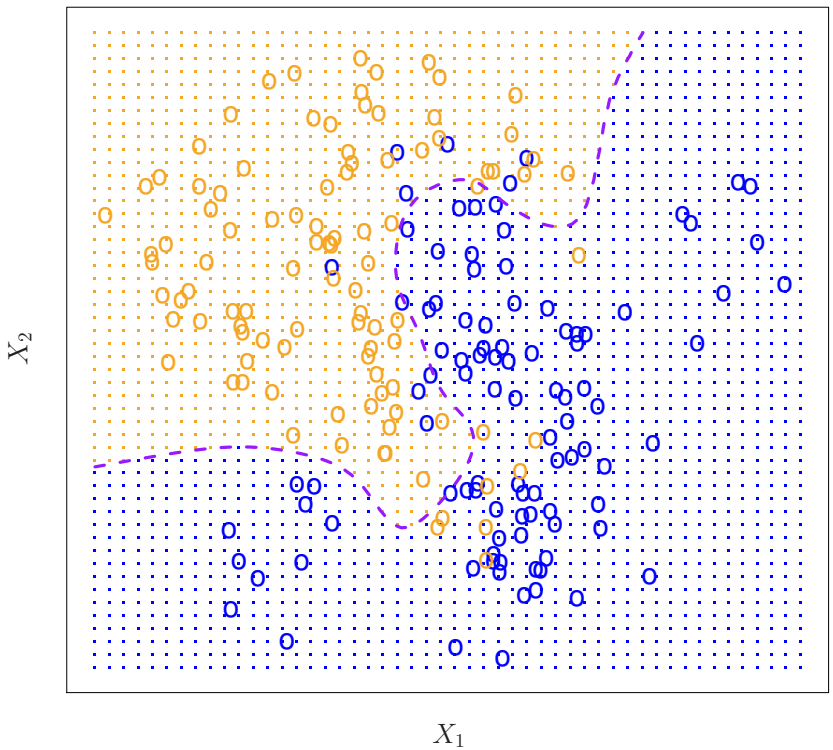


Figure 2.13

If we knew the conditional probability $P(Y \mid X)$, we could compute the optimal classifier under the 0-1 loss.

The Bayes classifier predicts

$$\hat{y}_i = \text{argmax}_j \;\; P(Y = j \mid X = x_i)$$

In practice we don't know $P(Y \mid X)$, but many classification methods operate by estimating these conditional probabilities.

K-nearest neighbors

To assign a color to an input x, we look at its $K = 3$ nearest neighbors and predict the color of the majority of the neighbors.

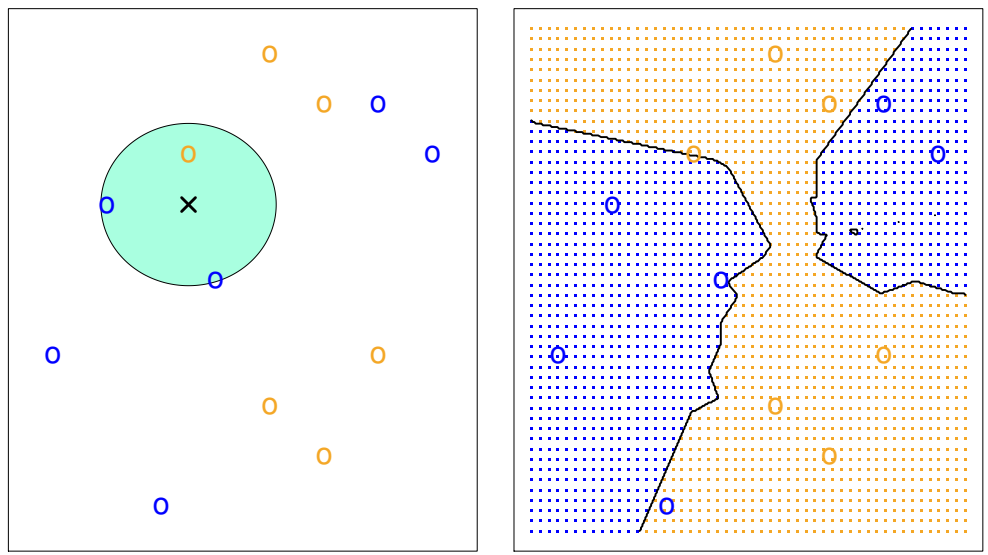


Figure 2.14

K-nearest neighbors gives a similar decision boundary

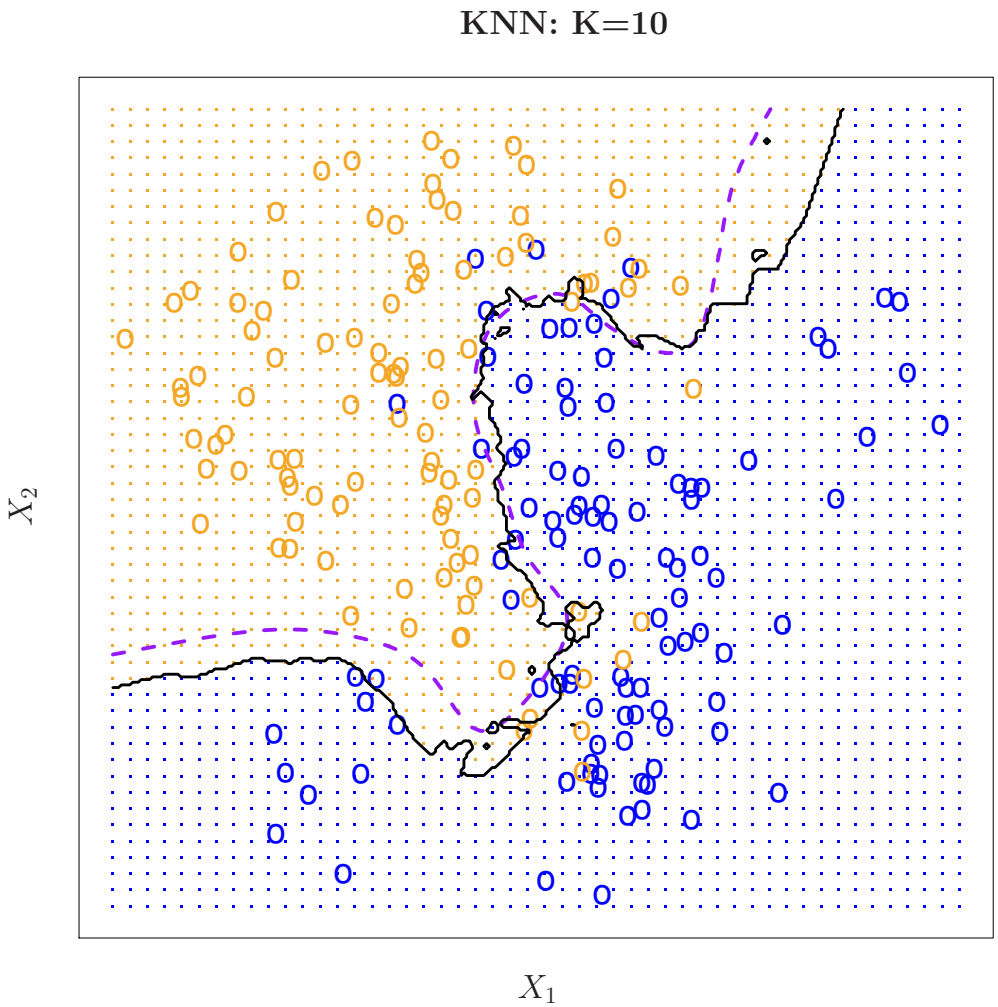


Figure 2.15

Higher K results in a smoother decision boundary

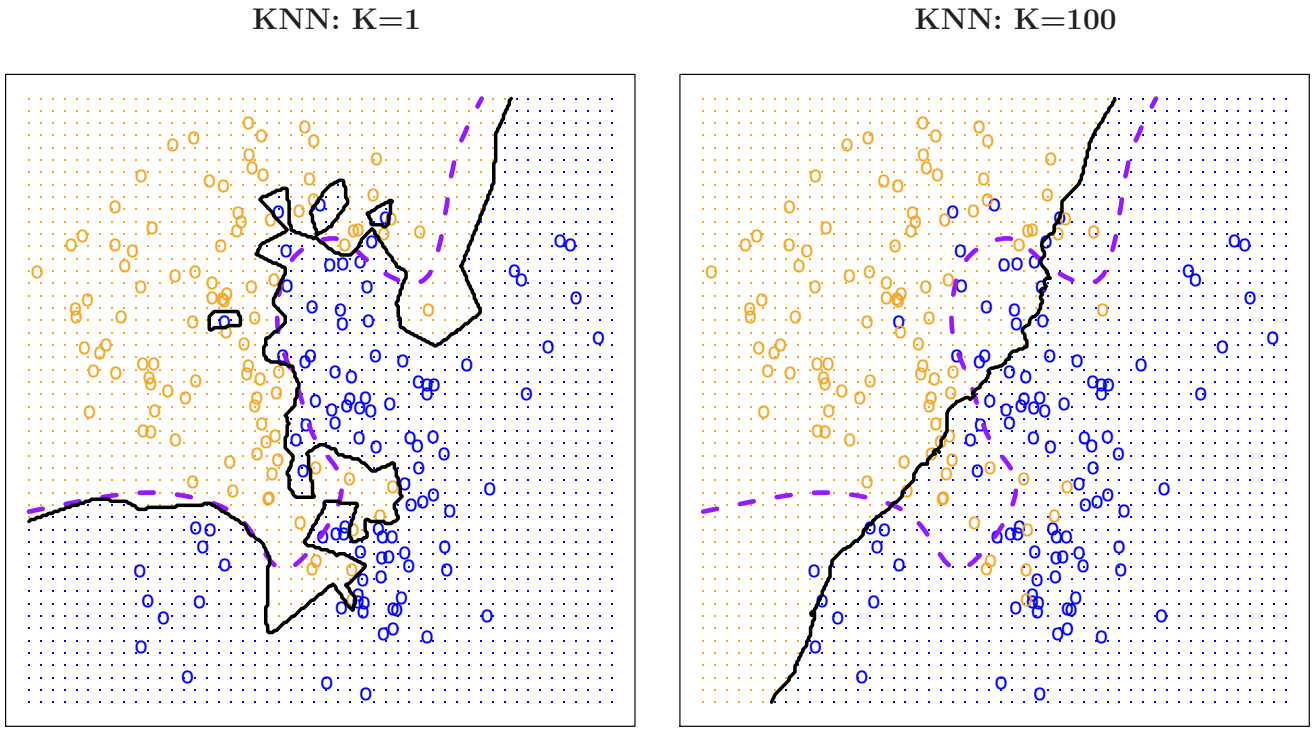
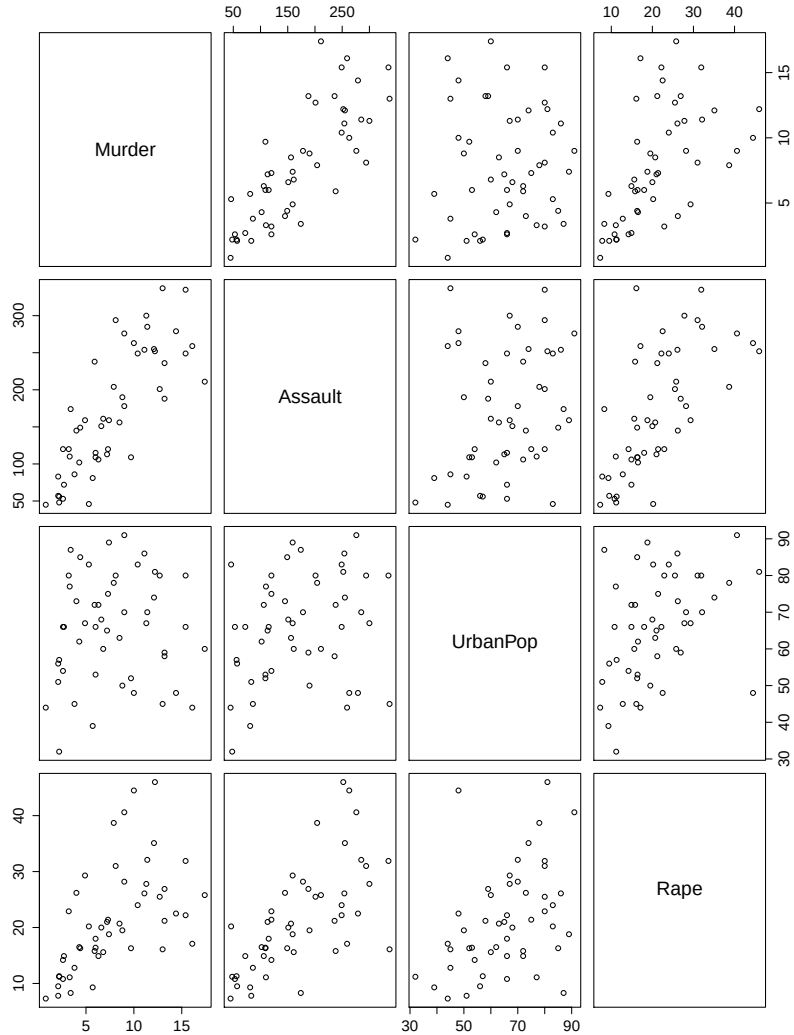


Figure 2.16

Principal Components Analysis

- This is the most popular unsupervised procedure ever.
- Invented by Karl Pearson (1901).
- Developed by Harold Hotelling (1933). $\leftarrow$ Stanford pride!
- **What does it do?** It provides a way to visualize high dimensional data, summarizing the most important information.

Scatterplot matrix



Biplot (from PCA)

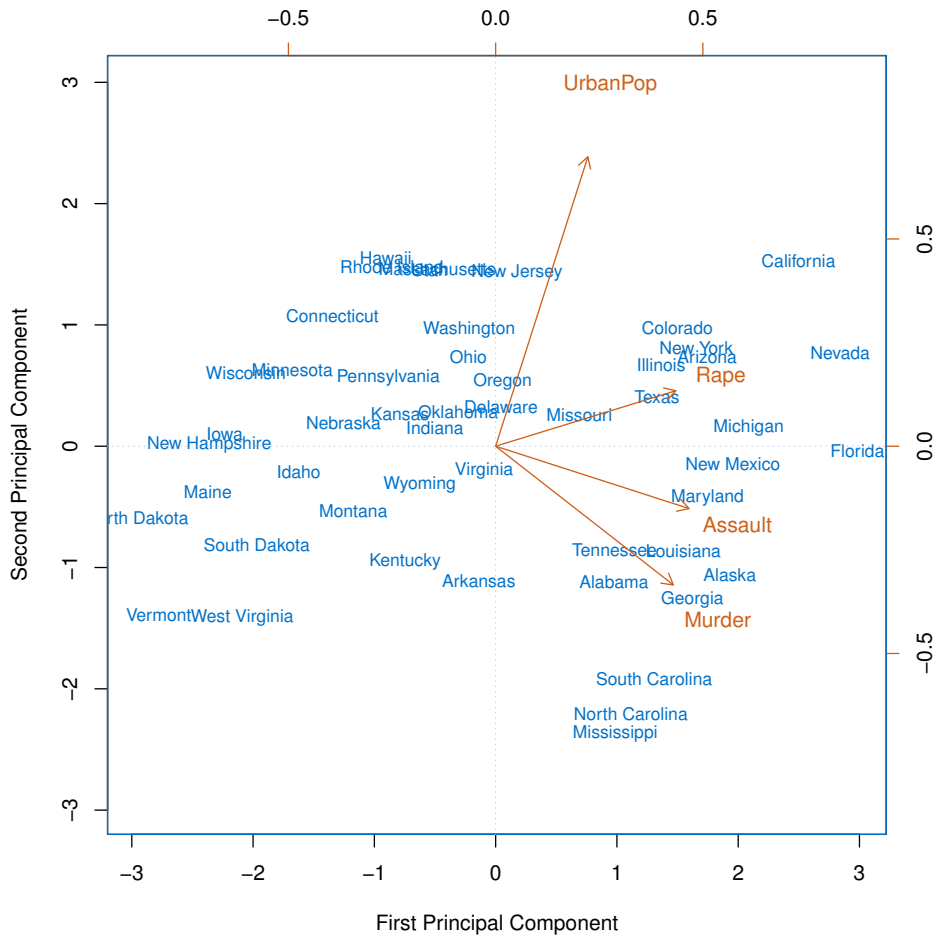
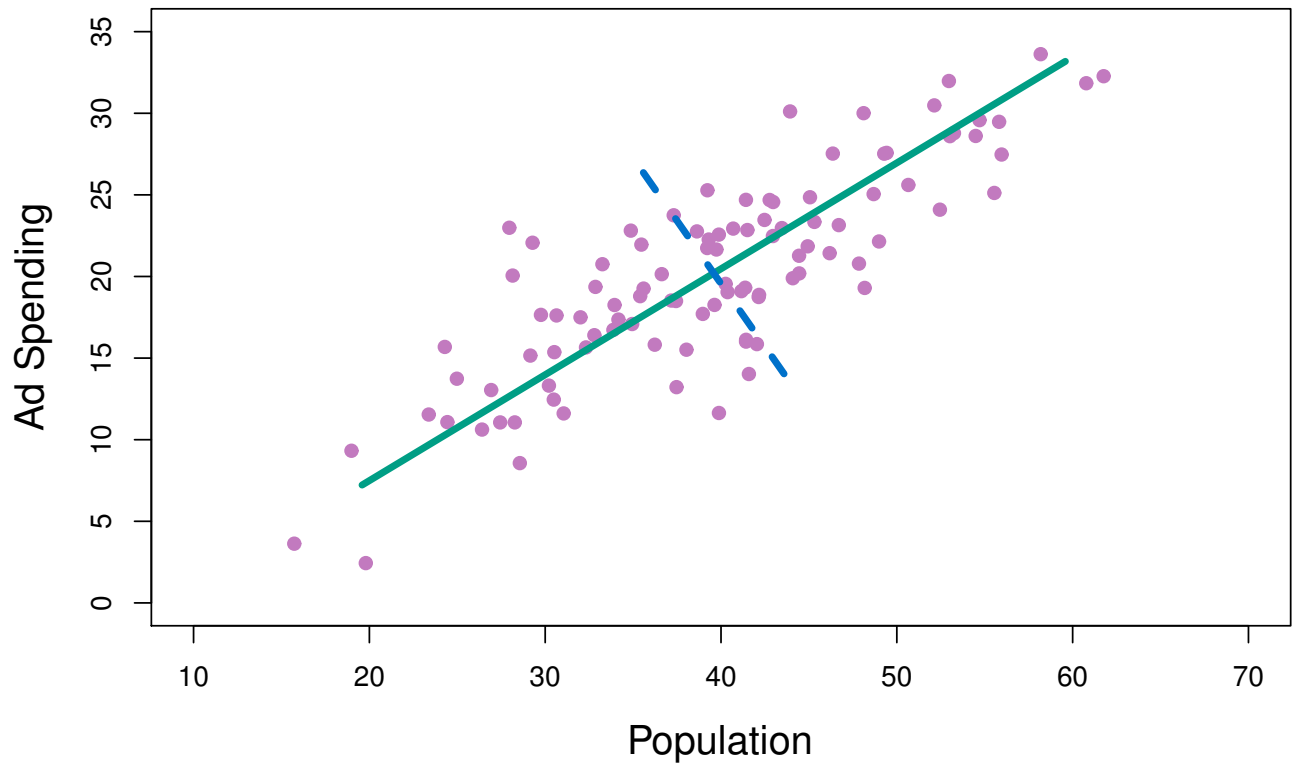


Figure 10.1

What is the first principal component?

It is the direction of the line that is closest to the datapoints, in terms of squared Euclidean distance.



That is, the PC direction minimizes the average squared length of the dotted lines.

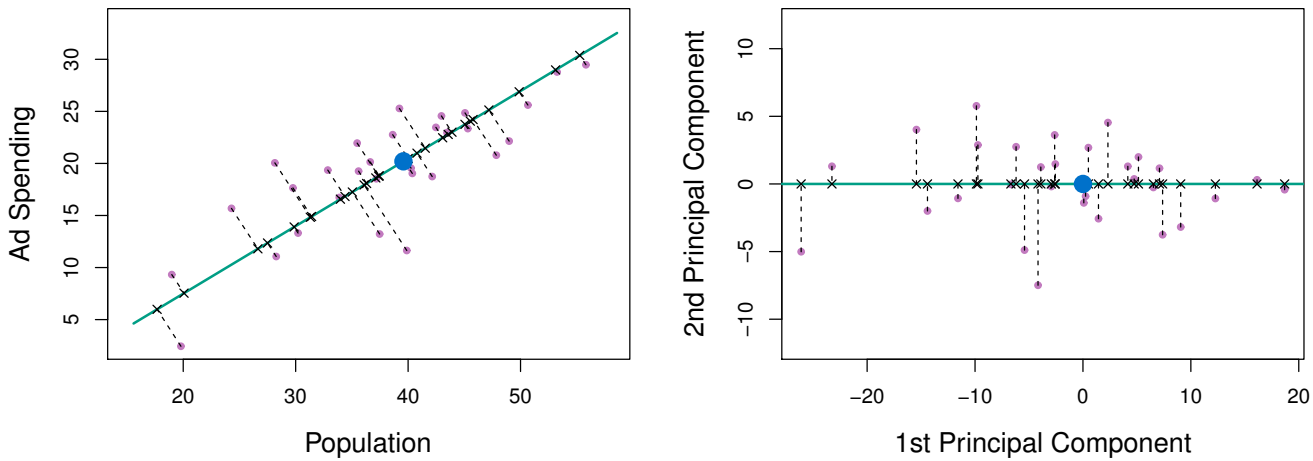
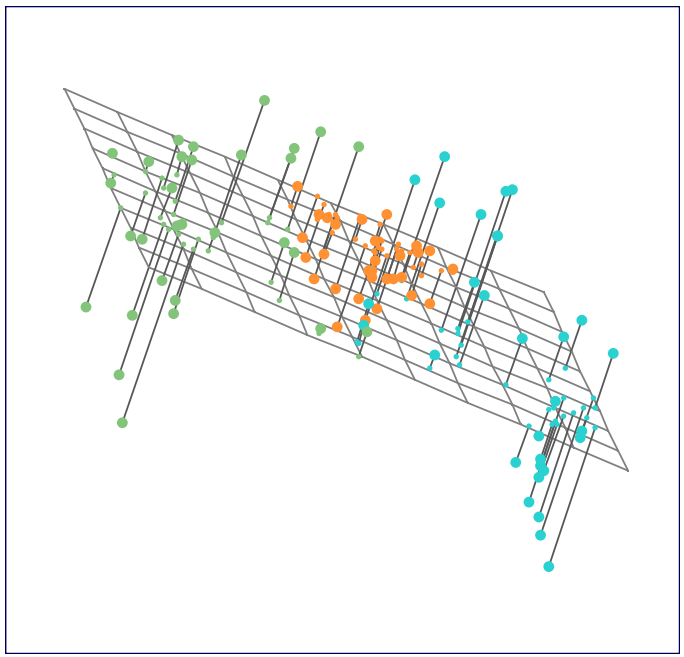


Figure 6.15

What does this look like with 3 variables?

The first two principal components span a plane which is closest to the data.



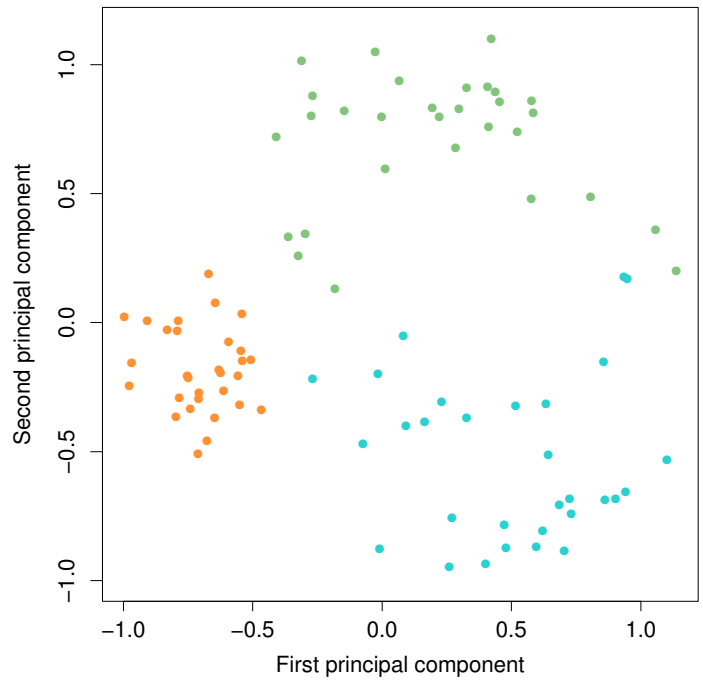


Figure 10.2

A second interpretation

The projection onto the first principal component is the one with the **highest variance**.

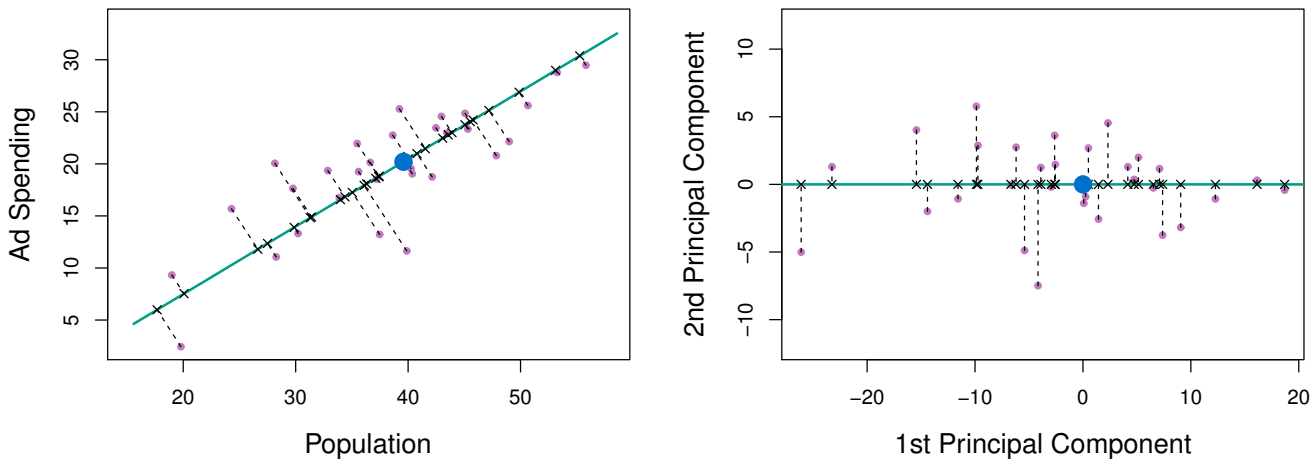


Figure 6.15

Intuition: High variance directions are often interesting directions.

How do we say this in math?

Let $\mathbf{X}$ be a data matrix with n samples, and p variables. From each variable, we subtract the mean of the column; i.e. we **center** the variables.

To find the first principal component $\phi_1 = (\phi_{11}, \ldots, \phi_{p1})$, we solve the following optimization

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

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

Projection of the ith sample onto ϕ_1. Also known as **the score** z_{i1}

How do we say this in math?

Let $\mathbf{X}$ be a data matrix with n samples, and p variables. From each variable, we subtract the mean of the column; i.e. we **center** the variables.

To find the first principal component $\phi_1 = (\phi_{11}, \ldots, \phi_{p1})$, we solve the following optimization

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

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

Sample variance of the scores Z_{i1}
(i.e. sample variance of projections of datapoints onto ϕ_1.)

How do we say this in math?

To find the second principal component $\phi_2 = (\phi_{12}, \ldots, \phi_{p2})$, we solve the following optimization

$$\max_{\phi_{12},\ldots,\phi_{p2}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j2} x_{ij} \right)^2 \right\}$$

$$\text{subject to } \sum_{j=1}^{p} \phi_{j2}^2 = 1 \quad \text{and} \quad \sum_{j=1}^{p} \phi_{j1}\phi_{j2} = 0.$$

First and second principal components must be orthogonal.

Equivalent to saying that the scores $(z_{11}, \ldots, z_{n1})$ and $(z_{12}, \ldots, z_{n2})$ are uncorrelated.

Solving the optimization

This optimization is fundamental in linear algebra. It is satisfied by either:

- The singular value decomposition (SVD) of $\mathbf{X}$:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{\Phi}^T$$

  where the ith column of $\mathbf{\Phi}$ is the ith principal component ϕ_i, and the ith column of $\mathbf{U}\mathbf{\Sigma}$ is the ith vector of scores $(z_{1i}, \ldots, z_{ni})$.

- The eigendecomposition of $\mathbf{X}^T\mathbf{X}$:

$$\mathbf{X}^T\mathbf{X} = \mathbf{\Phi}\mathbf{\Sigma}^2\mathbf{\Phi}^T$$

PCA in practice: The biplot

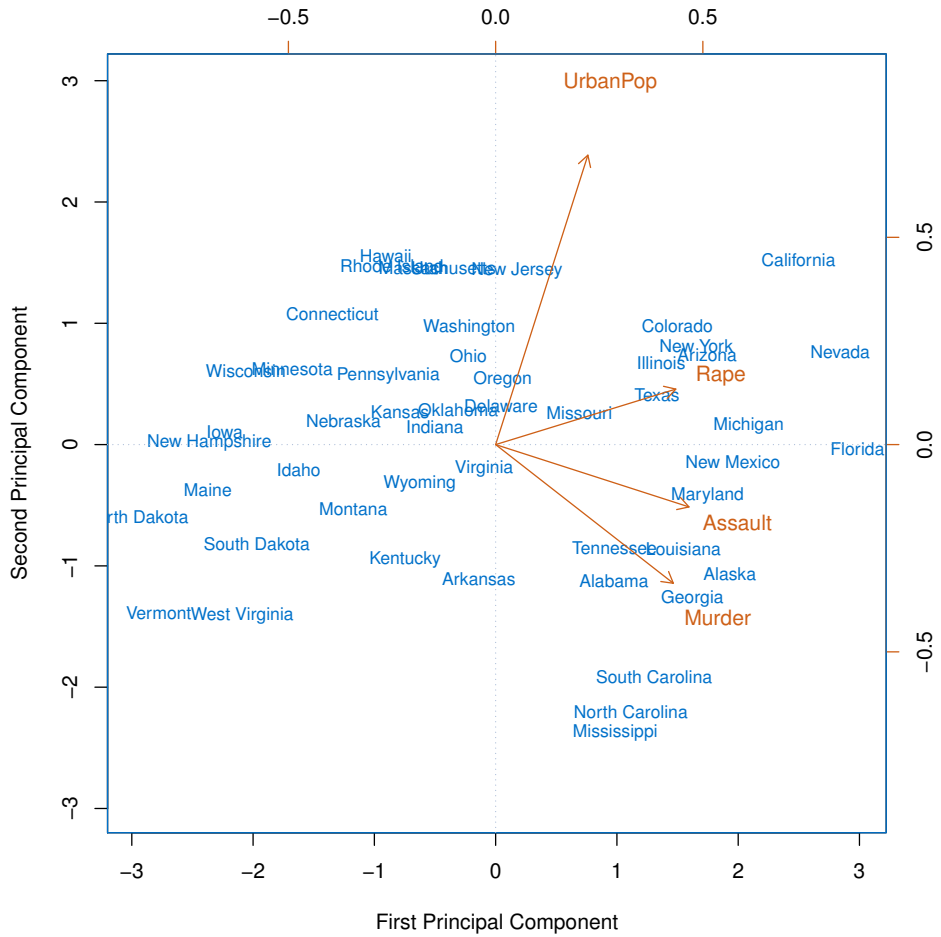


Figure 10.1

Scaling the variables

Most of the time, we don't care about the absolute numerical value of a variable. We care about the value relative to the spread observed in the sample.

In that case, before PCA, in addition to **centering** each variable, we also multiply it times a constant to make its variance equal to 1, i.e. we **standardize** each variable.

Example: scaled vs. unscaled PCA

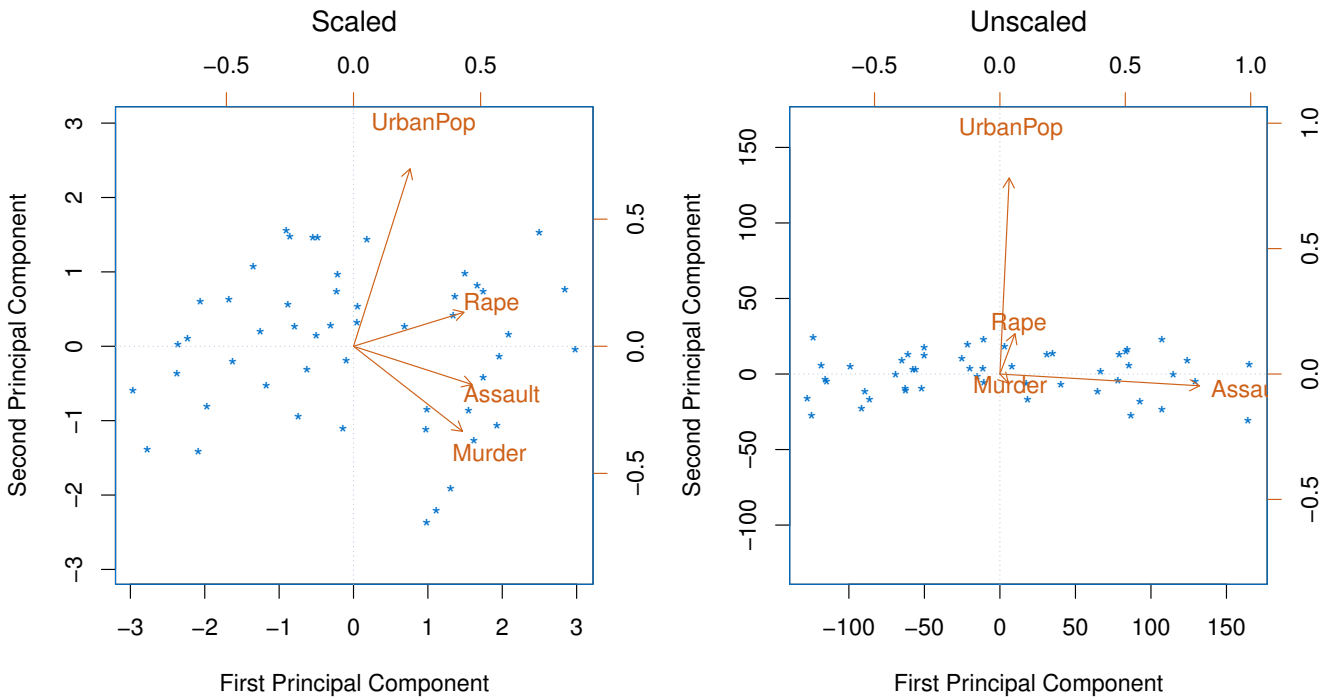


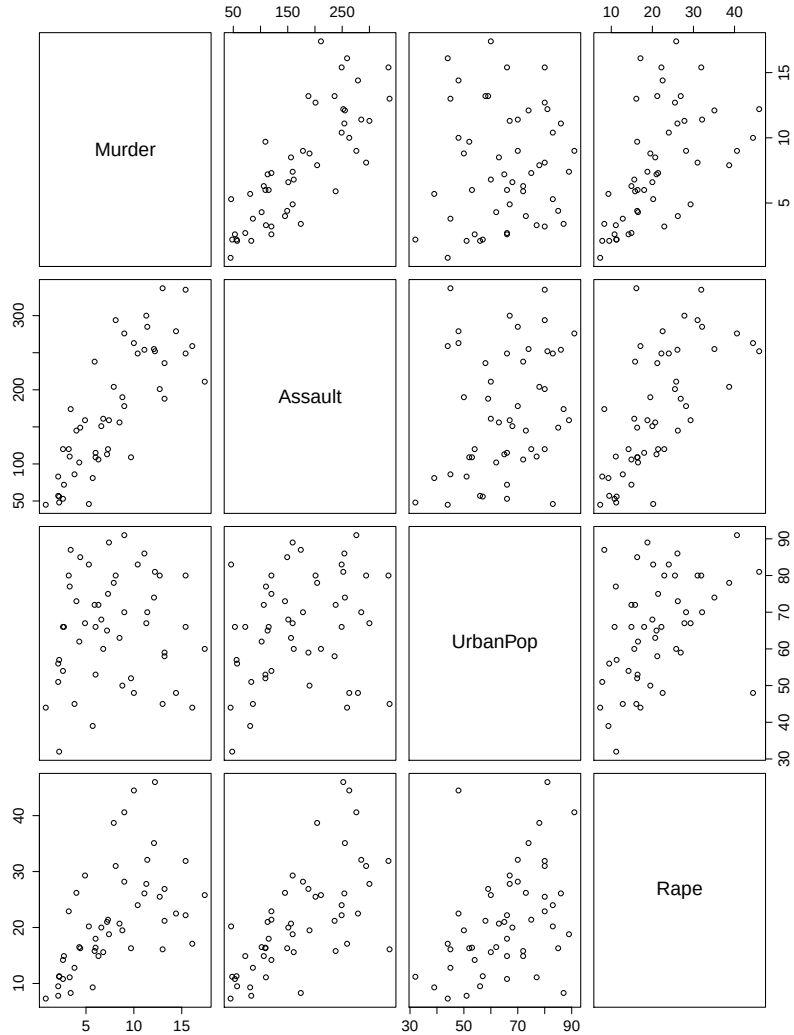
Figure 10.3

Scaling the variables

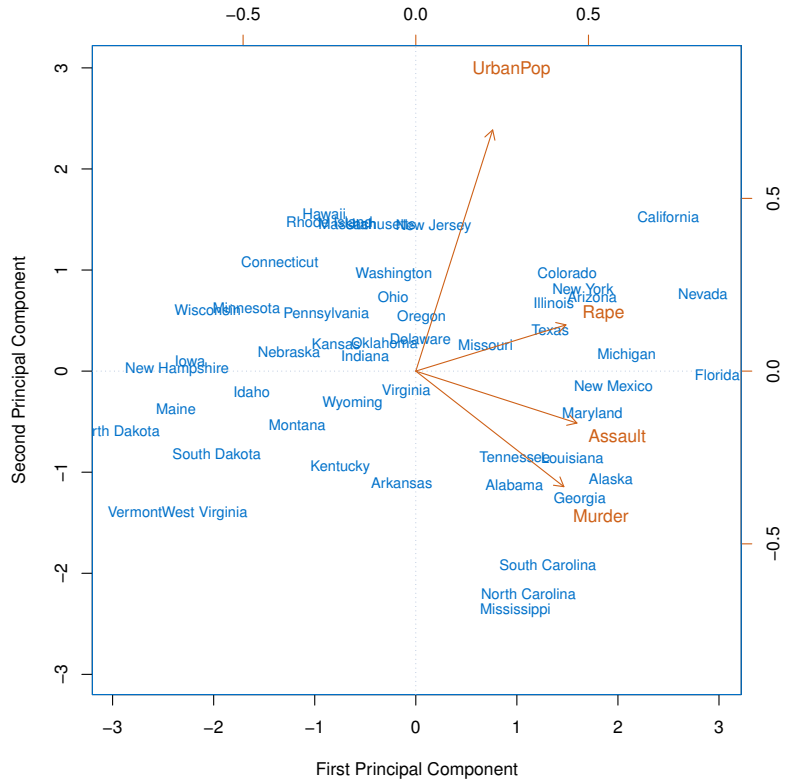
In special cases, we have variables measured in the same unit; e.g. gene expression levels for different genes.

Therefore, we care about the absolute value of the variables and we can perform PCA without scaling.

How many principal components are enough?



How many principal components are enough?



How can we tell if 2 principal components capture most of the relevant information?

The proportion of variance explained

We can think of the top **principal components** as directions in space in which the data vary the most.

The ith **score vector** $(z_{1i}, \ldots, z_{ni})$ can be interpreted as a *new* variable. The variance of this variable decreases as we take i from 1 to p. However, the total variance of the score vectors is the same as the total variance of the original variables:

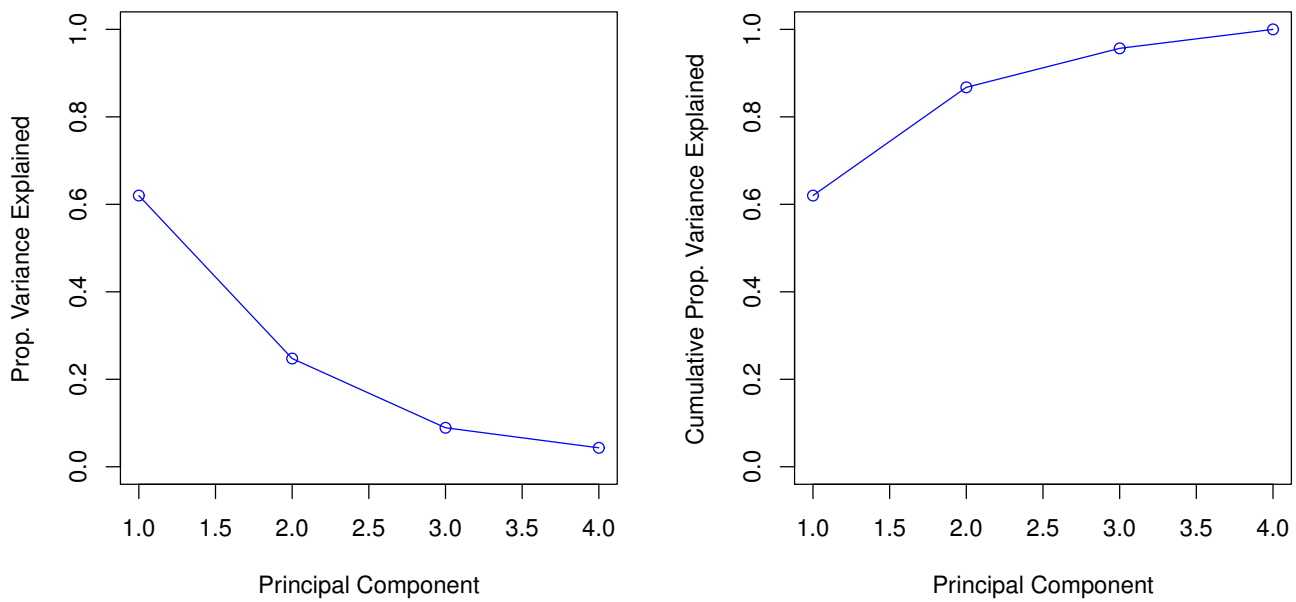
$$\sum_{i=1}^{p} \frac{1}{n} \sum_{j=1}^{n} z_{ji}^2 = \sum_{k=1}^{p} \mathsf{Var}(x_k).$$

We can quantify how much of the variance is captured by the first m principal components/score variables.

The proportion of variance explained

The variance of the mth score variable is:

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



Scree plot

Generalizations of PCA

PCA works under a Euclidean geometry in the space of variables. Often, the natural geometry is different:

- We expect some variables to be "closer" to each other than to other variables.
- Some correlations between variables would be more surprising than others.

Examples:

- Variables are pixel values, samples are different images of the brain. We expect neighboring pixels to have stronger correlations.
- Variables are rainfall measurements at different regions. We expect neighboring regions to have higher correlations.

Generalizations of PCA

There are ways to include this knowledge in a PCA. See:

1. Susan Holmes. *Multivariate Analysis, the French way.* (2006).
2. Omar de la Cruz and Susan Holmes. *An introduction to the duality diagram*. (2011).
3. Stéphane Dray and Thibaut Jombart. *Revisiting Guerry's data: Introducing spatial constraints in multivariate analysis.* (2011).
4. Genevera Allen, Logan Grosenick, and Jonathan Taylor. *A Generalized Least Squares Matrix Decomposition*. (2011).