CS6301: R For Data Scientists

LECTURE 6: PRINCIPAL COMPONENT ANALYSIS – PART 1

Background

PCA is a *dimension reduction* method – it is a way of viewing information with many dimensions

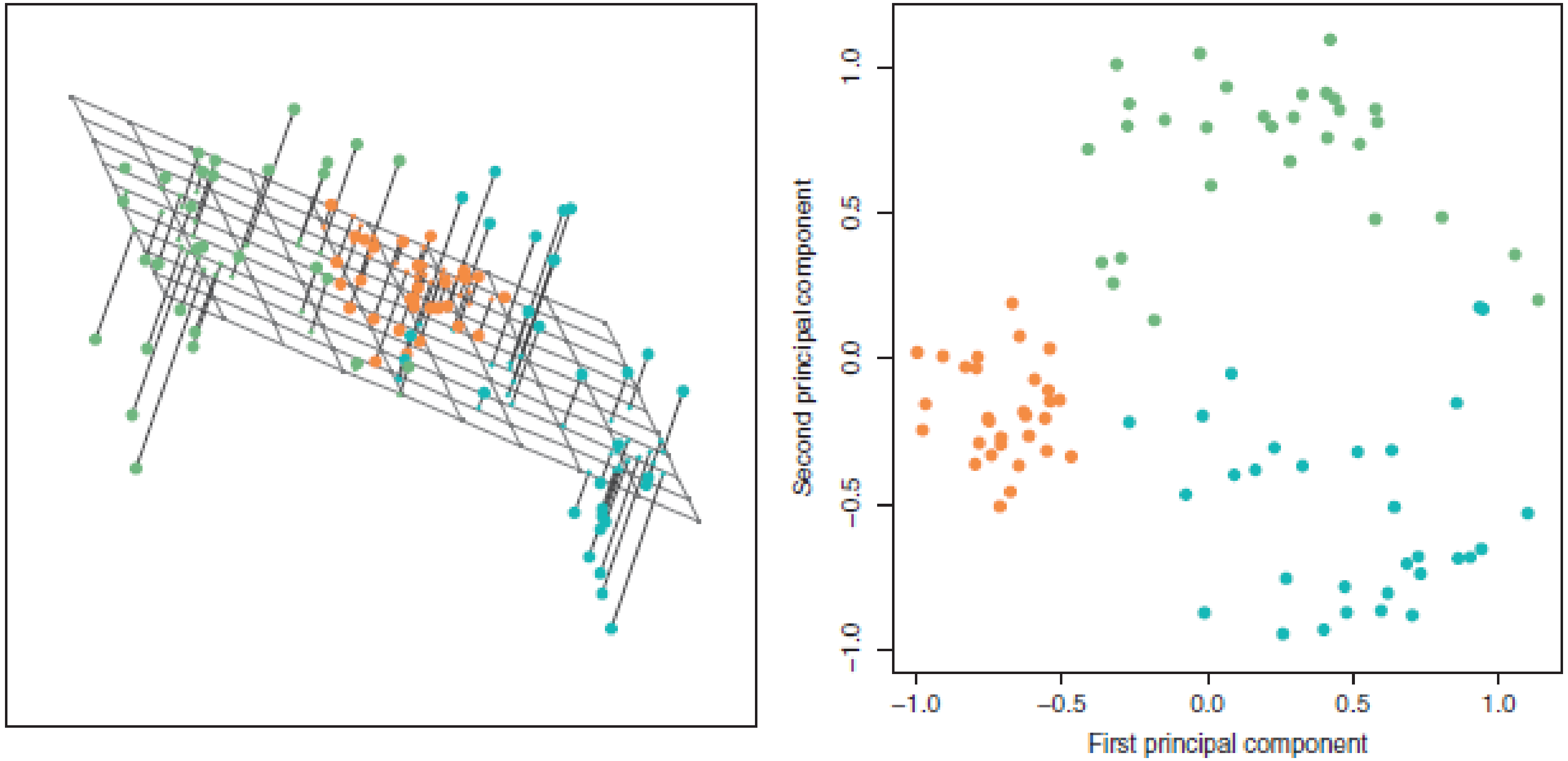
Also referred to as *feature extraction* – this considered a basic form

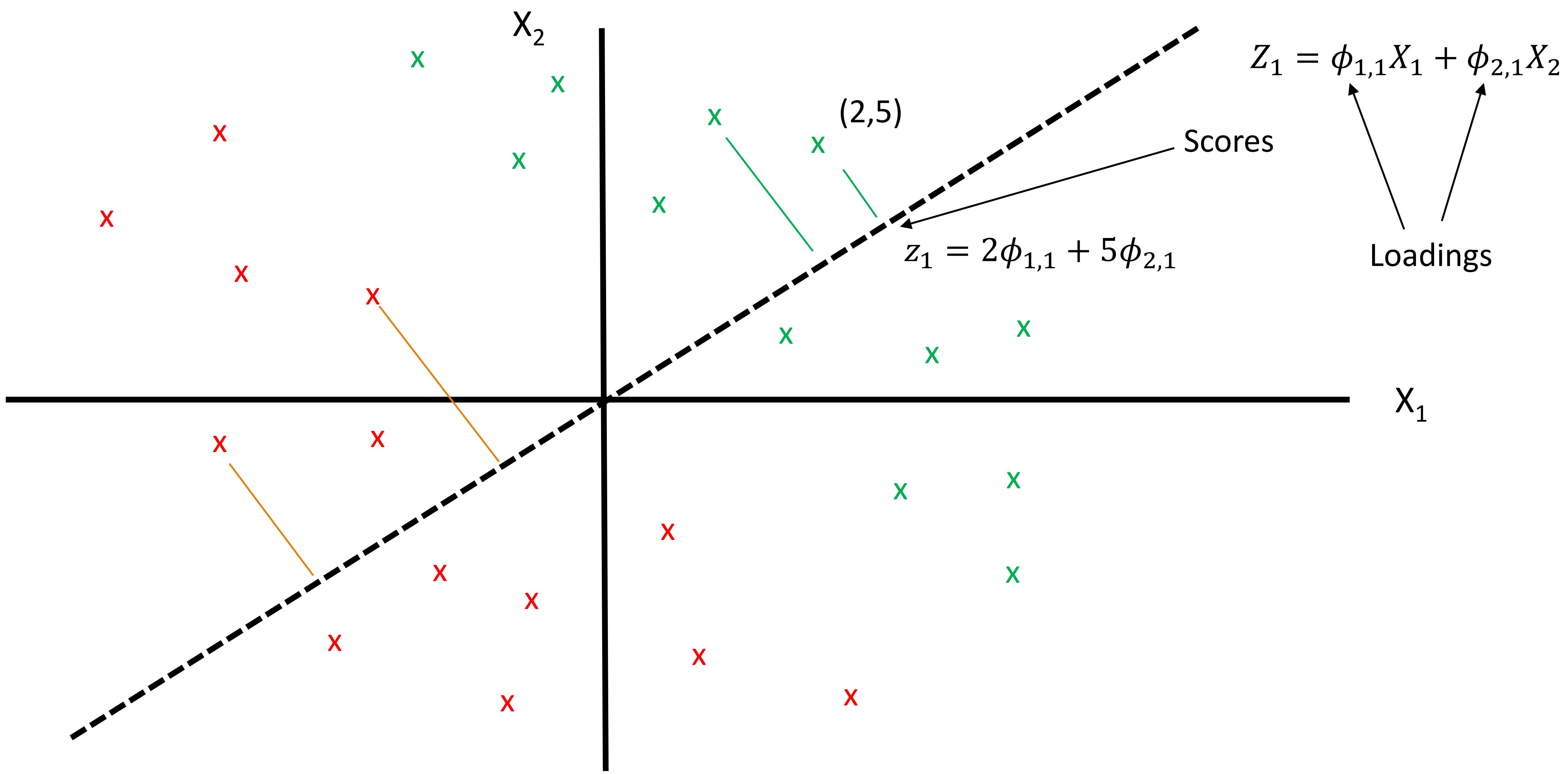
Main idea: We can see clusters most easily in a plane. Find the plane that "best represents" the high dimensional data, and look for patterns

We will do this by projecting the higher dimensional data onto axes, or *principal components,* in a way that spreads the data out as much as possible

We look at lower dimensional problems first, to try to understand how the method works

PCA – Simple Example





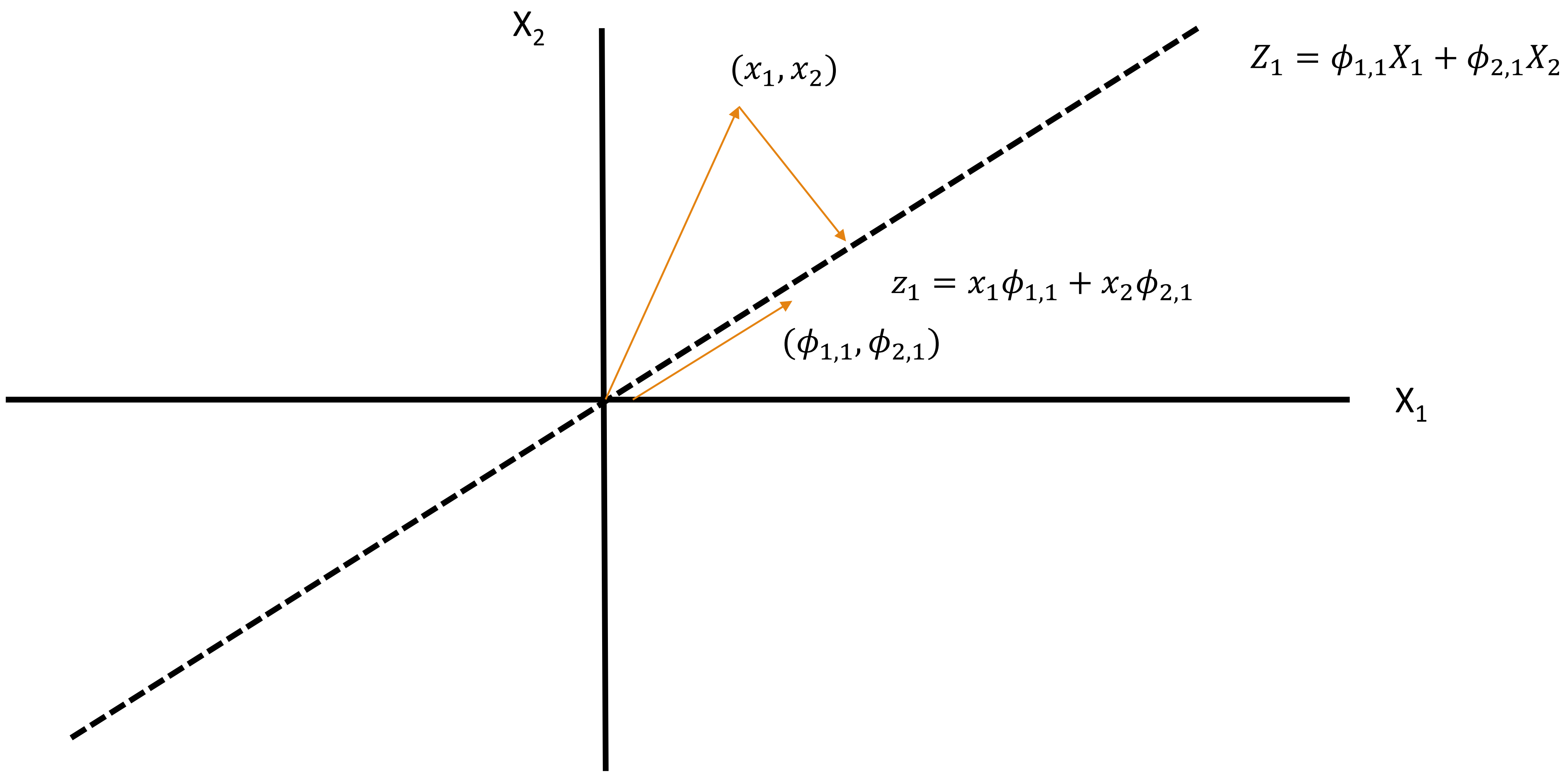
X2
X1
(2,5)
Scores
$Z_1 = \phi_{1,1}X_1 + \phi_{2,1}X_2$
$z_1 = 2\phi_{1,1} + 5\phi_{2,1}$
Loadings

Finding Principal Components - Example

Problem: Find $(\phi_{1,1}, \phi_{2,1})$ such that

$$\max_{\phi_{1,1},\phi_{2,1}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\phi_{1,1} x_{i,1} + \phi_{2,1} x_{i,2} \right)^2 \right\}$$

subject to $\phi_{1,1}^2 + \phi_{2,1}^2 = 1$



X_2
(x_1, x_2)
$Z_1 = \phi_{1,1}X_1 + \phi_{2,1}X_2$
$z_1 = x_1\phi_{1,1} + x_2\phi_{2,1}$
$(\phi_{1,1}, \phi_{2,1})$
X_1

First Principal Component

Suppose we have a set of data points in p-dimensional space:

$$\{(x_{i,1}, x_{i,2}, \dots, x_{i,p})\} \;\; i = 1, \dots, n$$

So each data point is a p-tuple.

The *first principal component* is defined to be

$$Z_1 = \varphi_{1,1}X_1 + \varphi_{2,1}X_2 + \cdots + \varphi_{p,1}X_p$$

where the coefficients $\varphi_{i,1}$ are found in a way that maximizes the variance of the projections of the data points onto this coordinate, and are also normalized:

$$\sum_{i=1}^{p} \varphi_{i,1}^2 = 1$$

First Principal Component

The *loadings* make up the First Principal Component Vector,

$$\boldsymbol{\varphi_1} = (\varphi_{1,1}, \varphi_{2,1}, \dots, \varphi_{p,1})^t$$

This is a vector in p-dimensional space that represents the direction of the first principal component

The *scores* are the scalars obtained by taking the dot product of each data point with this vector. For example, the score of the i^{th} data point is:

$$z_{i,1} = \varphi_{1,1}x_{i,1} + \varphi_{2,1}x_{i,2} + \cdots + \varphi_{p,1}x_{i,p}$$

First Principal Component

Putting it all together: The loading vector is found by maximizing the scores over all normalized possible loading vectors:

$$\max\left\{\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\varphi_{j,1}x_{i,j}\right)^{2}\right\} = \max\left\{\frac{1}{n}\sum_{i=1}^{n}(z_{i1})^{2}\right\}$$

subject to $\sum_{j=1}^{p}\varphi_{j,1}^{2} = 1$.

This is the vector that spreads the data out the most in this one direction.

Principal Components

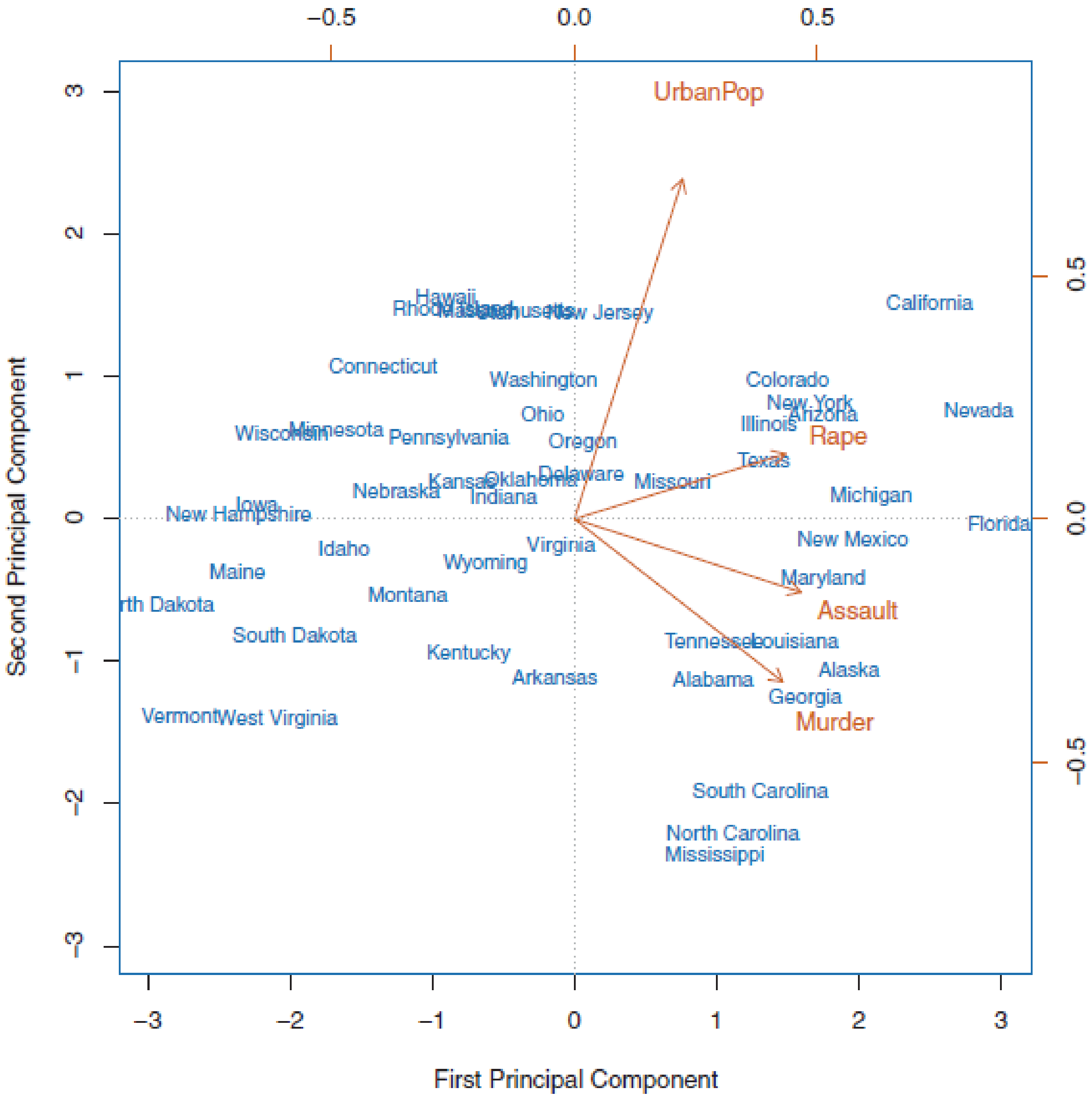
How do we find the remaining components?

To find the next component, we find a vector *orthogonal* to the first vector which explains most of the remaining variance … and continue until we have *p* vectors

This is a rotation of the original coordinate system to a new coordinate system, one in which the data is as spread out as possible along the axis (PCs)

Note we will always have *p* PCs, and in the end all of the variance is completely explained

This is usually done by finding eigenvectors of the correlation matrix


UrbanPop
Rape
Assault
Murder
First Principal Component
Second Principal Component
Hawaii
Rhode Island
Massachusetts
Utah
New Jersey
California
Connecticut
Washington
Colorado
New York
Arizona
Illinois
Nevada
Ohio
Wisconsin
Minnesota
Pennsylvania
Oregon
Texas
Delaware
Missouri
Kansas
Oklahoma
Nebraska
Indiana
Michigan
Iowa
New Hampshire
Florida
Idaho
Virginia
New Mexico
Maine
Wyoming
Maryland
rth Dakota
Montana
South Dakota
Kentucky
Tennessee
Louisiana
Arkansas
Alabama
Alaska
Georgia
Vermont
West Virginia
South Carolina
North Carolina
Mississippi

Scaling In PCA