Exploring Countries

Exploring Countries

Collected data on over 100 countries, including:

- GDP per capita
- Population
- Regime type
- Military expenditures
- ...

What can we do to understand our data better?

Dimensionality Reduction

Problem: Suppose that we wish to visualize *n* observations with measurements on a set of p features: $X_1, X_2, ..., X_p$ as part of exploratory data analysis. How do we do it?

Dimensionality Reduction

Problem: Suppose that we wish to visualize n observations with measurements on a set of p features: $X_1, X_2, ..., X_p$ as part of exploratory data analysis. How do we do it?

Scatterplots:

- When $p = 2$, create a two-dimensional scatterplot

Dimensionality Reduction

Problem: Suppose that we wish to visualize n observations with measurements on a set of p features: $X_1, X_2, ..., X_p$ as part of exploratory data analysis. How do we do it?

Scatterplots:

- When $p = 2$, create a two-dimensional scatterplot
- When $p = 3$: create 3 scatterplots (X_1 by X_2, X_1 by X_3, X_2 by X_3),

Dimensionality Reduction

Problem: Suppose that we wish to visualize n observations with measurements on a set of p features: $X_1, X_2, ..., X_p$ as part of exploratory data analysis. How do we do it?

Scatterplots:

- When $p = 2$, create a two-dimensional scatterplot
- When $p = 3$: create 3 scatterplots (X_1 by X_2, X_1 by X_3, X_2 by X_3),
- When $p = 10$: create 45 plots???

Dimensionality Reduction

Problem: Suppose that we wish to visualize *n* observations with measurements on a set of p features: $X_1, X_2, ..., X_p$ as part of exploratory data analysis. How do we do it?

Scatterplots:

- When $p = 2$, create a two-dimensional scatterplot
- When $p = 3$: create 3 scatterplots (X_1 by X_2, X_1 by X_3, X_2 by X_3),
- When $p = 10$: create 45 plots???
- When p is large $\rightsquigarrow$ impossible to look at all of them!

Dimensionality Reduction

Problem: Suppose that we wish to visualize n observations with measurements on a set of p features: $X_1, X_2, ..., X_p$ as part of exploratory data analysis. How do we do it?

Scatterplots:

- When $p = 2$, create a two-dimensional scatterplot
- When $p = 3$: create 3 scatterplots (X_1 by X_2, X_1 by X_3, X_2 by X_3),
- When $p = 10$: create 45 plots???
- When p is large $\rightsquigarrow$ impossible to look at all of them!

Solution: We want to find a low-dimensional representation of the data that captures as much information as possible. This is called dimensionality reduction.

Why Dimensionality Reduction?

Can be used for:

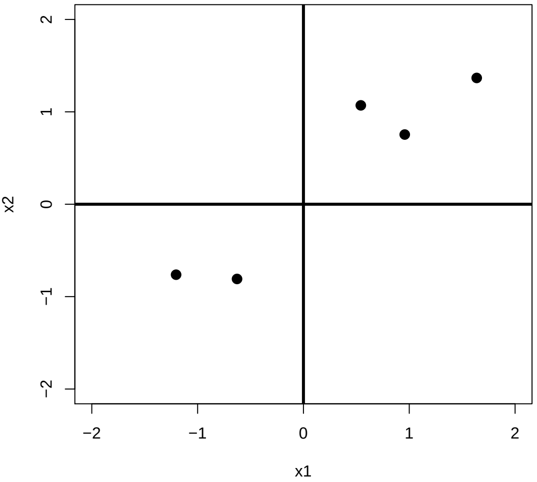
- data summary / exploration
- data visualization (of observations, or variables)
- producing features for use in supervised learning problems.
- data compression (use less memory)

Goal: summarize data in a lower-dimensional space
Method: Principal-Component Analysis
Key Terms:

- Dimensionality Reduction
- Projection
- Principle Component Analysis (PCA)
- Centered
- Scaled
- PC loadings
- PC scores
- Biplot
- Scree Plot
- Proportion of variance explained (PVE)
- Elbow

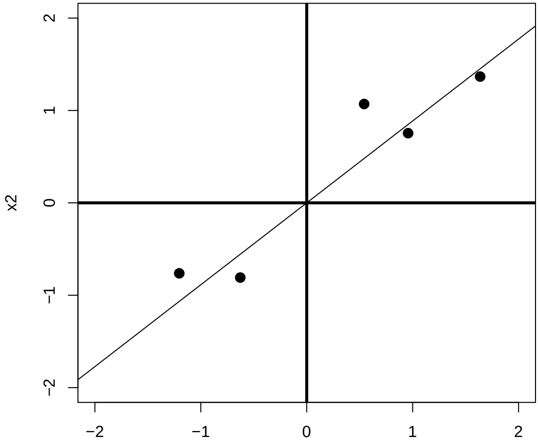
What does it mean to reduce data?

How do we reduce this data from 2 dimensions to 1 dimension?



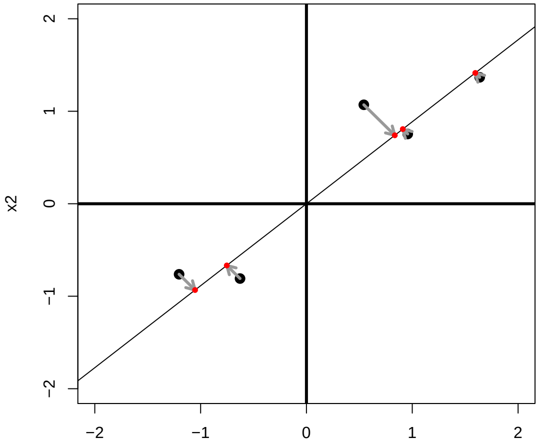
What does it mean to reduce data?

Identify a new dimension on which to project (aka "map" or "embed") the data.



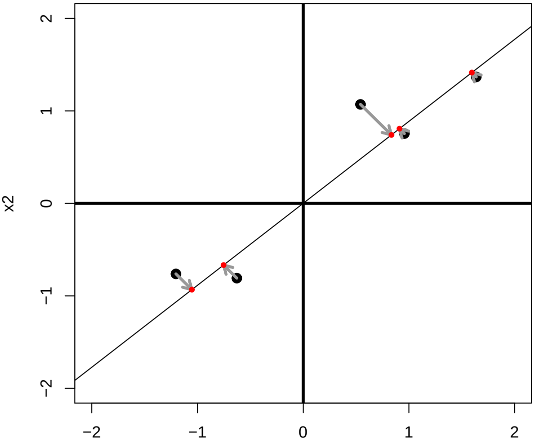
What does it mean to reduce data?

Projecting a point simply means finding the location on the line which is closest to the point.

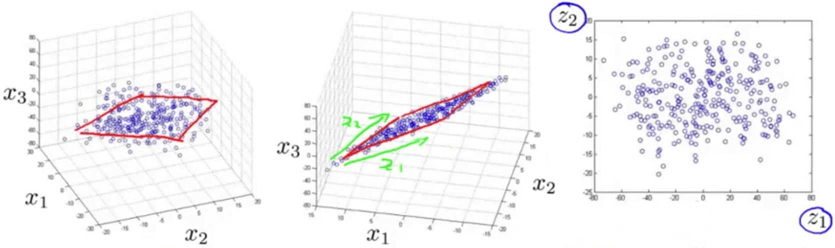


What does it mean to reduce data?

We can now represent each point with this new feature (z_1).
2 dimensions (on x_1 and x_2) $\rightarrow$ 1 dimension (on z_1)



From 3D to 2D



Principle Component Analysis

Principle Component Analysis (PCA) is a popular way to reduce dimensionality.

Principle Component Analysis

Principle Component Analysis (PCA) is a popular way to reduce dimensionality.

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

Principle Component Analysis

Principle Component Analysis (PCA) is a popular way to reduce dimensionality.

- 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.

Principle Component Analysis

Principle Component Analysis (PCA) is a popular way to reduce dimensionality.

- 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.
- By "interesting," we mean the amount that the observations vary along each dimension.

What are principle components?

PCA finds directions of maximal variance.

What are principle components?

PCA finds directions of maximal variance.

- **First** principal component: direction along which the observations vary the most.

What are principle components?

PCA finds directions of maximal variance.

- **First** principal component: direction along which the observations vary the most.
- **Second** component: direction of second greatest variance, and is orthogonal (perpendicular) to the first component.

What are principle components?

PCA finds directions of maximal variance.

- **First** principal component: direction along which the observations vary the most.
- **Second** component: direction of second greatest variance, and is orthogonal (perpendicular) to the first component.
- **Each succeeding component** in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components.

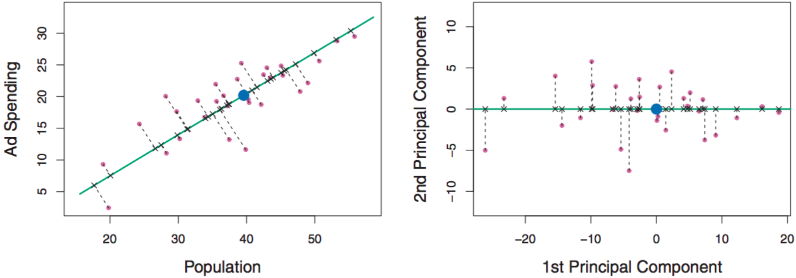
What are principle components?

PCA finds directions of maximal variance.

- **First** principal component: direction along which the observations vary the most.
- **Second** component: direction of second greatest variance, and is orthogonal (perpendicular) to the first component.
- **Each succeeding component** in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components.
- Finds up to p unique components.

Another interpretation of principle components

The first principal component defines the line that is as close as possible to the data (using average squared Euclidean distance).



PCA Inputs

For PCA to work, the features should be preprocessed in a particular way.

PCA Inputs

For PCA to work, the features should be preprocessed in a particular way.

1. Centered to have mean zero (i.e. de-meaning).
   - Compute mean of each feature.
   - Subtract each value of a feature by the mean of that feature.
   - `colMeans(feature) = 0`

PCA Inputs

For PCA to work, the features should be preprocessed in a particular way.

1. Centered to have mean zero (i.e. de-meaning).
   - Compute mean of each feature.
   - Subtract each value of a feature by the mean of that feature.
   - `colMeans(feature) = 0`
2. Scaled for comparability.
   - Different features on different scales or units of measurement (number of bedrooms versus price of house)
   - Rescale variables to have comparable ranges of values
   - `standard deviation(feature) = 1`

PCA Outputs

K principle components are computed via eigen decomposition.

PCA Outputs

K principle components are computed via eigen decomposition.

1. $\boldsymbol{w}_k$: Principal component loadings: Unit vector defining direction of given component (length of p). In other words: How much is given component weighted towards a each variable?

$$\boldsymbol{w}_1 = (w_1, w_2, \ldots, w_p)$$

PCA Outputs

K principle components are computed via eigen decomposition.

1. $\boldsymbol{w}_k$: Principal component loadings: Unit vector defining direction of given component (length of p). In other words: How much is given component weighted towards a each variable?

$$\begin{aligned} \boldsymbol{w}_1 &= (w_1, w_2, \ldots, w_p) \\ \boldsymbol{w}_k &= (w_{1k}, w_{2k}, \ldots, w_{pk}) \end{aligned}$$

PCA Outputs

K principle components are computed via eigen decomposition.

1. $\boldsymbol{w}_k$: Principal component loadings: Unit vector defining direction of given component (length of p). In other words: How much is given component weighted towards a each variable?

$$\begin{aligned} \boldsymbol{w}_1 &= (w_1, w_2, \ldots, w_p) \\ \boldsymbol{w}_k &= (w_{1k}, w_{2k}, \ldots, w_{pk}) \end{aligned}$$

2. $\boldsymbol{z}_i$: Principal component scores: projected values for each observation given component (length of n)

$$\boldsymbol{z}_1 = (z_1, z_2, \ldots, z_n)$$

PCA Outputs

K principle components are computed via eigen decomposition.

1. $\boldsymbol{w}_k$: Principal component loadings: Unit vector defining direction of given component (length of p). In other words: How much is given component weighted towards a each variable?

$$\begin{aligned} \boldsymbol{w}_1 &= (w_1, w_2, \ldots, w_p) \\ \boldsymbol{w}_k &= (w_{1k}, w_{2k}, \ldots, w_{pk}) \end{aligned}$$

2. $\boldsymbol{z}_i$: Principal component scores: projected values for each observation given component (length of n)

$$\begin{aligned} \boldsymbol{z}_1 &= (z_1, z_2, \ldots, z_n) \\ \boldsymbol{z}_k &= (z_{1k}, z_{2k}, \ldots, z_{nk}) \end{aligned}$$

PCA Outputs

K principle components are computed via eigen decomposition.

1. $\boldsymbol{w}_k$: Principal component loadings: Unit vector defining direction of given component (length of p). In other words: How much is given component weighted towards a each variable?

$$\begin{aligned} \boldsymbol{w}_1 &= (w_1, w_2, \ldots, w_p) \\ \boldsymbol{w}_k &= (w_{1k}, w_{2k}, \ldots, w_{pk}) \end{aligned}$$

2. $\boldsymbol{z}_i$: Principal component scores: projected values for each observation given component (length of n)

$$\begin{aligned} \boldsymbol{z}_1 &= (z_1, z_2, \ldots, z_n) \\ \boldsymbol{z}_k &= (z_{1k}, z_{2k}, \ldots, z_{nk}) \end{aligned}$$

3. Principal component eigenvalues: defines "interestness" of each component.

An Example: USArrests

	Murder	Assault	UrbanPop	Rape
Alabama	13.20	236	58	21.20
Alaska	10.00	263	48	44.50
Arizona	8.10	294	80	31.00
Arkansas	8.80	190	50	19.50
California	9.00	276	91	40.60
Colorado	7.90	204	78	38.70

An Example: USArrests

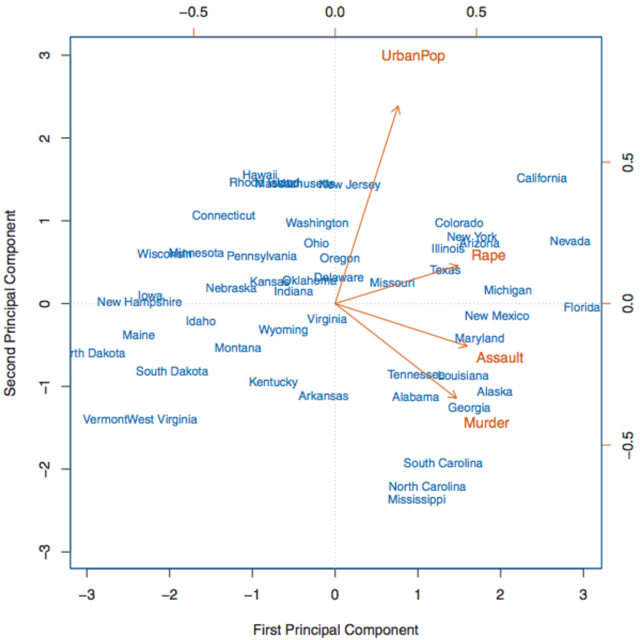
1. Standardize each variable to have mean of 0 and standard deviation 1.

An Example: USArrests

1. Standardize each variable to have mean of 0 and standard deviation 1.
2. Compute 2 principle components. For each PC:
   - Loading vector: length $p = 4$.
   - Score vector: length $n = 50$.

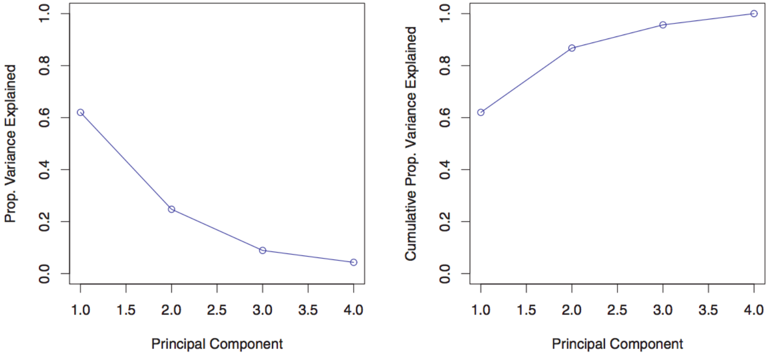
An Example: USArrests

1. Standardize each variable to have mean of 0 and standard deviation 1.
2. Compute 2 principle components. For each PC:
   - Loading vector: length p = 4.
   - Score vector: length n = 50.
3. Plot PC1 against PC2.


-0.5
0.0
0.5
UrbanPop
Rape
Assault
Murder
Hawaii
New Jersey
California
Connecticut
Washington
Colorado
New York
Arizona
Nevada
Ohio
Illinois
Pennsylvania
Oregon
Texas
Nebraska
Indiana
Delaware
Missouri
Michigan
New Hampshire
Florida
Idaho
Virginia
New Mexico
Maine
Wyoming
Maryland
rth Dakota
Montana
South Dakota
Kentucky
Tennessee
Louisiana
Arkansas
Alabama
Alaska
Georgia
VermontWest Virginia
South Carolina
North Carolina
Mississippi
Second Principal Component
First Principal Component
-3
-2
-1
0
1
2
3

How much variation is explained by each principal component?

Scree Plot: Visualize proportion of variance explained (PVE) by each component



How many components do we need?

We choose the smallest number of principal components required to explain a sizable amount of variation in the data.

How many components do we need?

We choose the smallest number of principal components required to explain a sizable amount of variation in the data.

Eyeball the scree plot, and look for the elbow (the point at which the proportion of variance explained by each subsequent principal component drops off.)

Goal: summarize data in a lower-dimensional space
Method: Principal-Component Analysis
Key Terms:

- Dimensionality Reduction
- Projection
- Principle Component Analysis (PCA)
- Centered
- Scaled
- PC loadings
- PC scores
- Biplot
- Scree Plot
- Proportion of variance explained (PVE)
- Elbow

To the R Code!