Data 451
Principal Components Analysis

Hunter Glanz

February 19, 2019

Where does IQ come from?

Where does IQ come from?

- In 1904, Charles Spearman noted that children's performance across unrelated school subjects, like Classics, Math, and Music, were positively correlated.

Where does IQ come from?

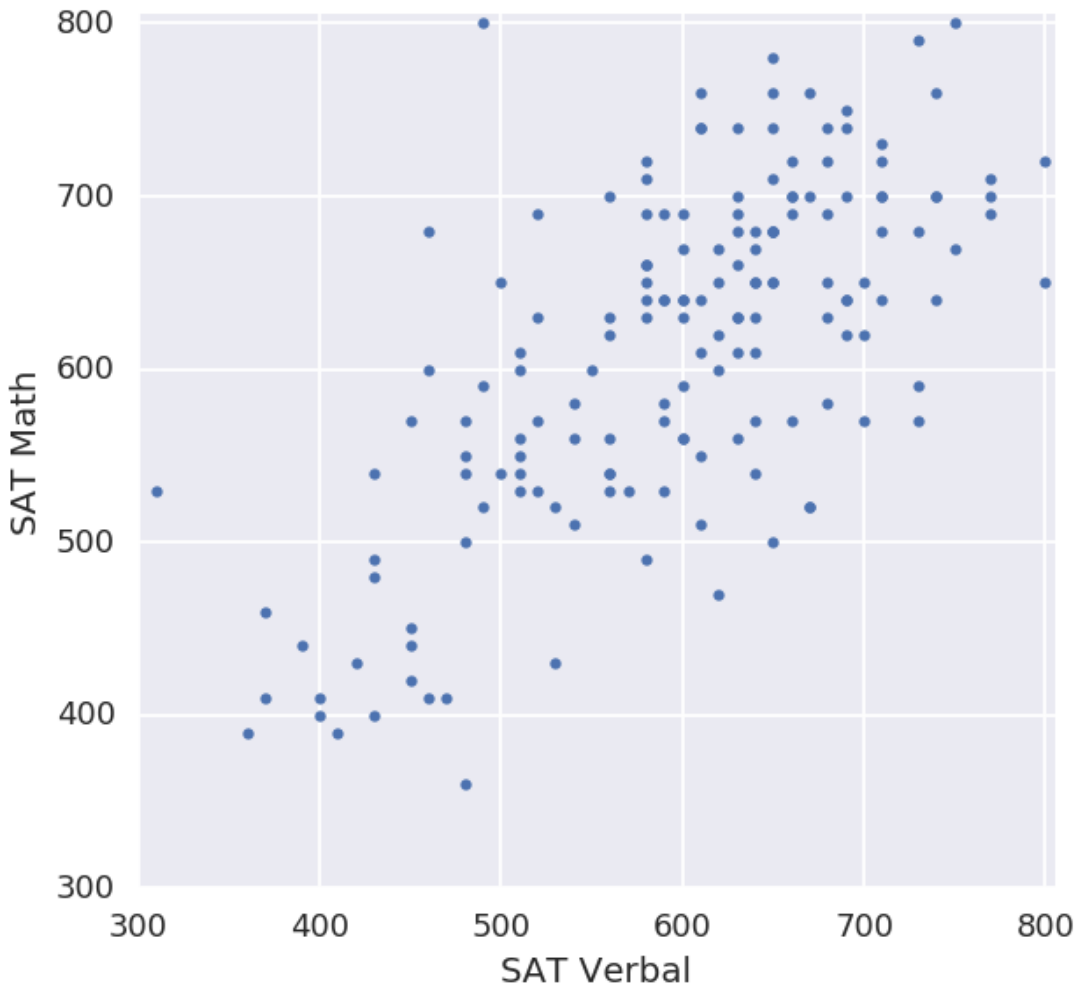
- In 1904, Charles Spearman noted that children's performance across unrelated school subjects, like Classics, Math, and Music, were positively correlated.
- He hypothesized that all cognitive ability could be traced to a single "general intelligence" factor, which he called the g factor.

Where does IQ come from?

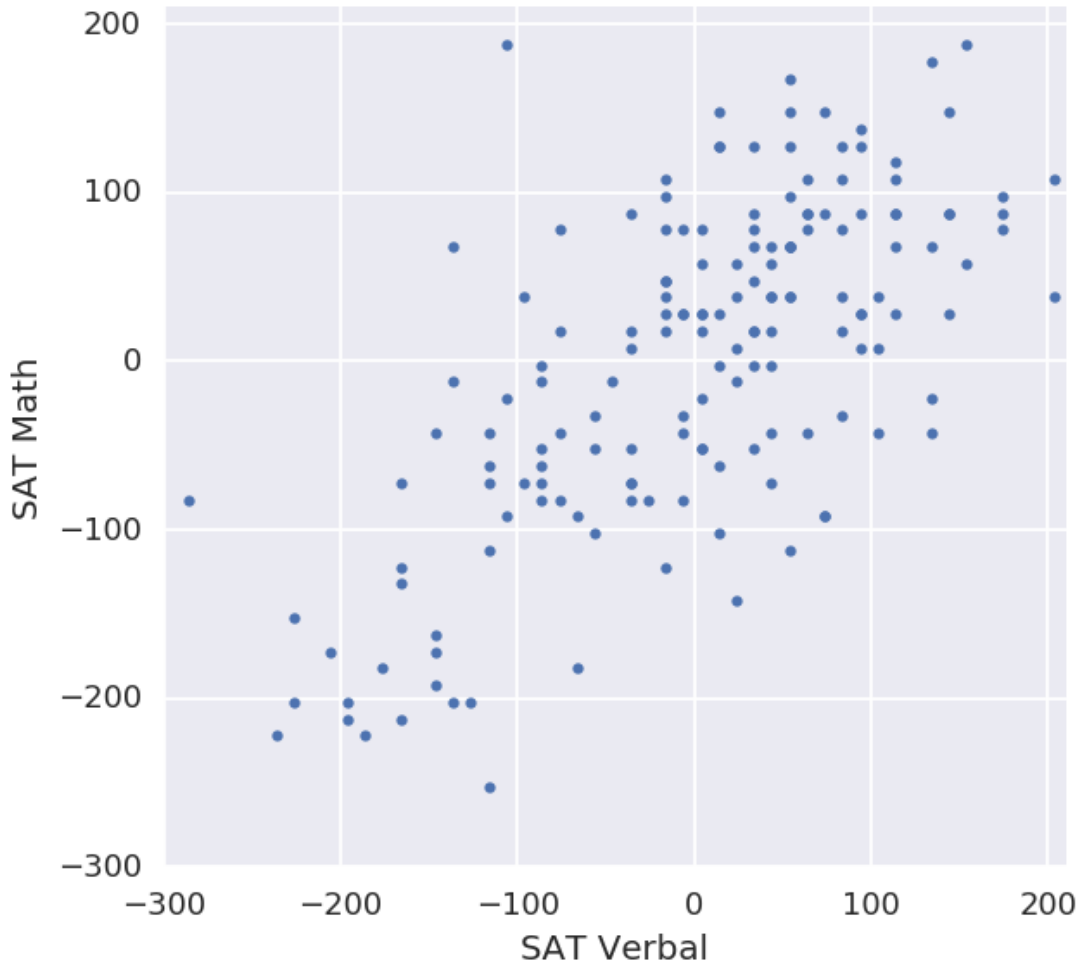
- In 1904, Charles Spearman noted that children's performance across unrelated school subjects, like Classics, Math, and Music, were positively correlated.
- He hypothesized that all cognitive ability could be traced to a single "general intelligence" factor, which he called the g factor.
- Later, IQ tests were designed to try to measure this g factor. It attempts to quantify intelligence along a single dimension.

Identifying the g factor in data

Identifying the g factor in data

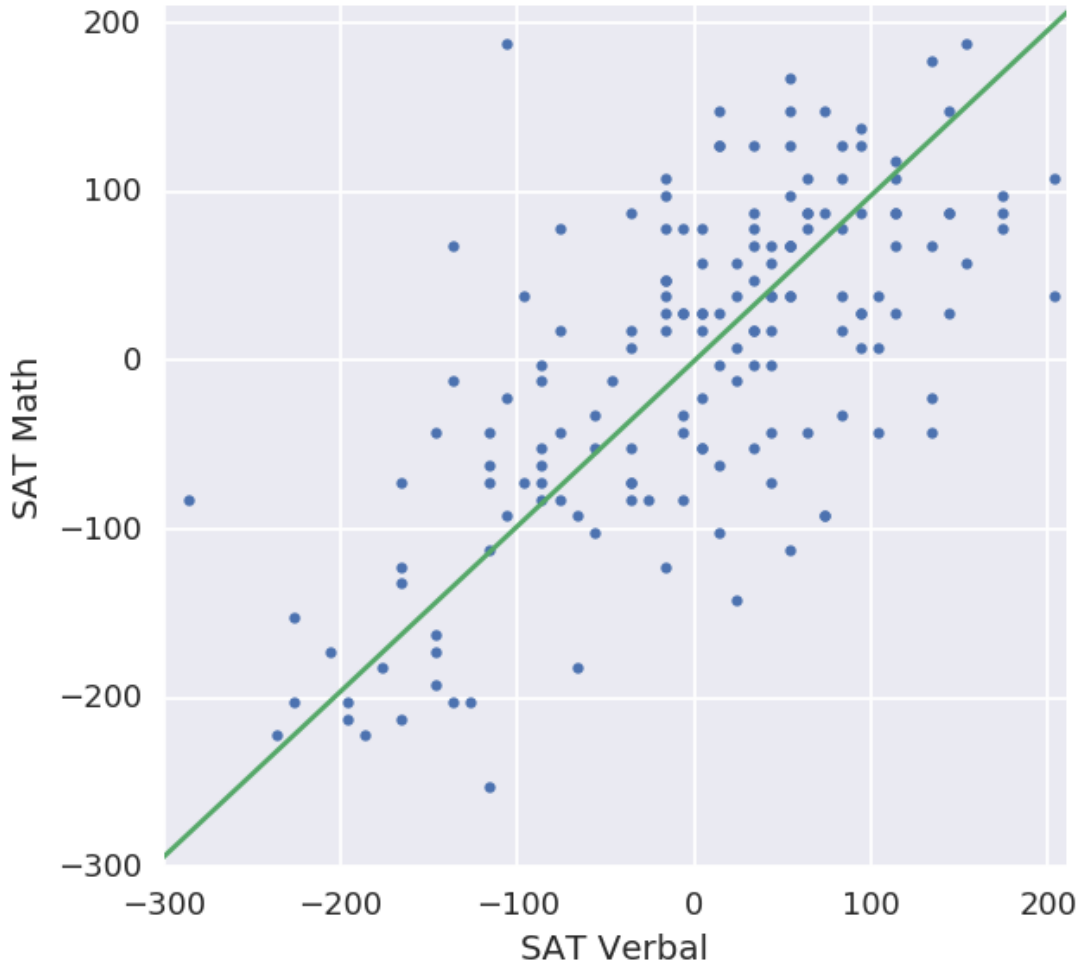


Identifying the g factor in data



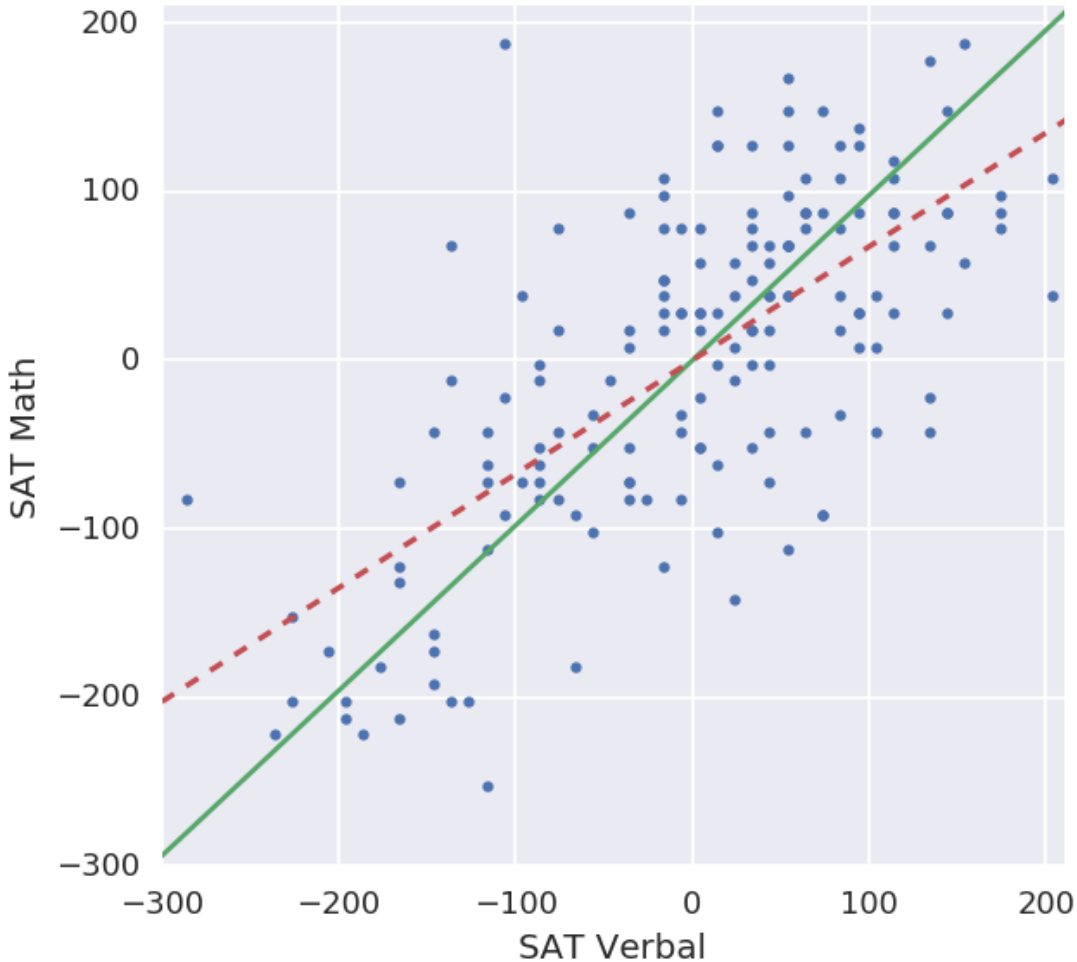
It is typical to first **center** the variables so that they have mean 0.

Identifying the g factor in data



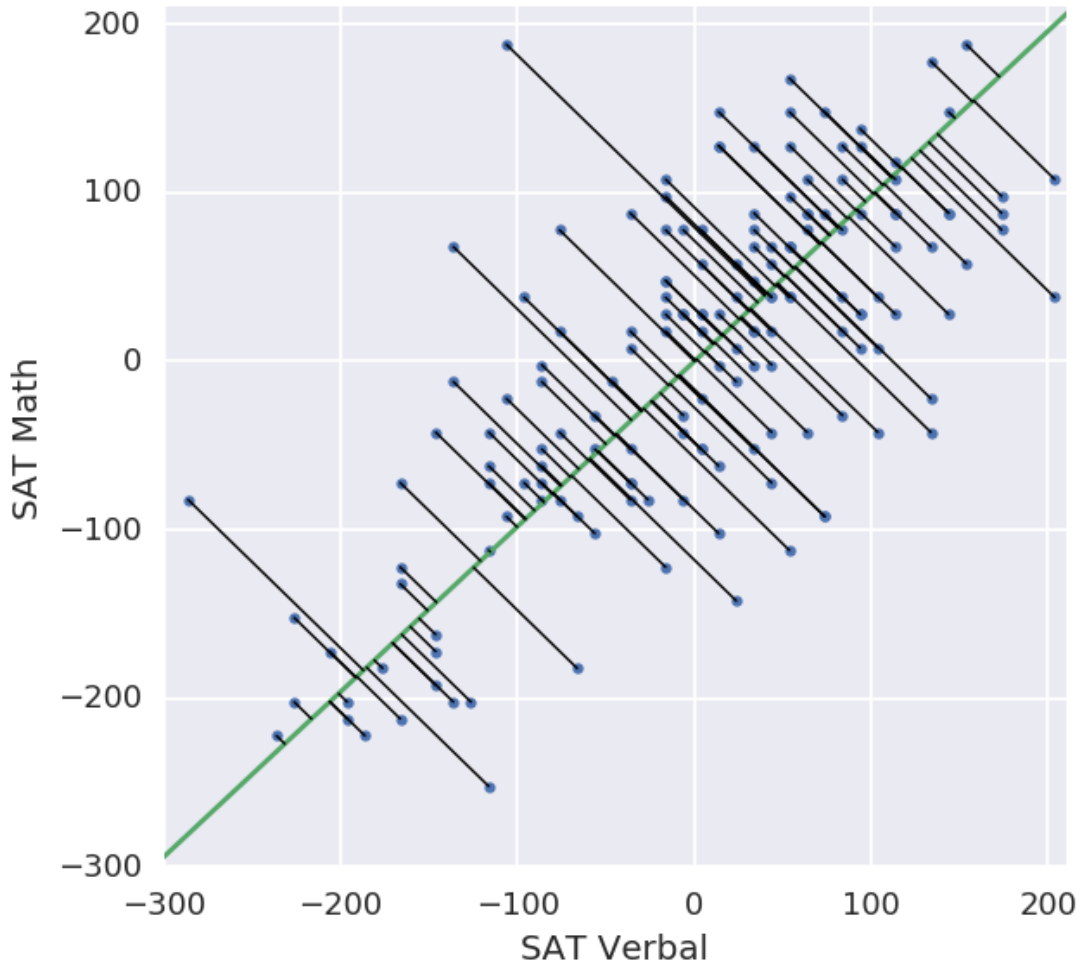
The g factor is a combination of math and verbal skills. It is the direction in the data of greatest variability. This is called the **first principal component**.

Identifying the g factor in data



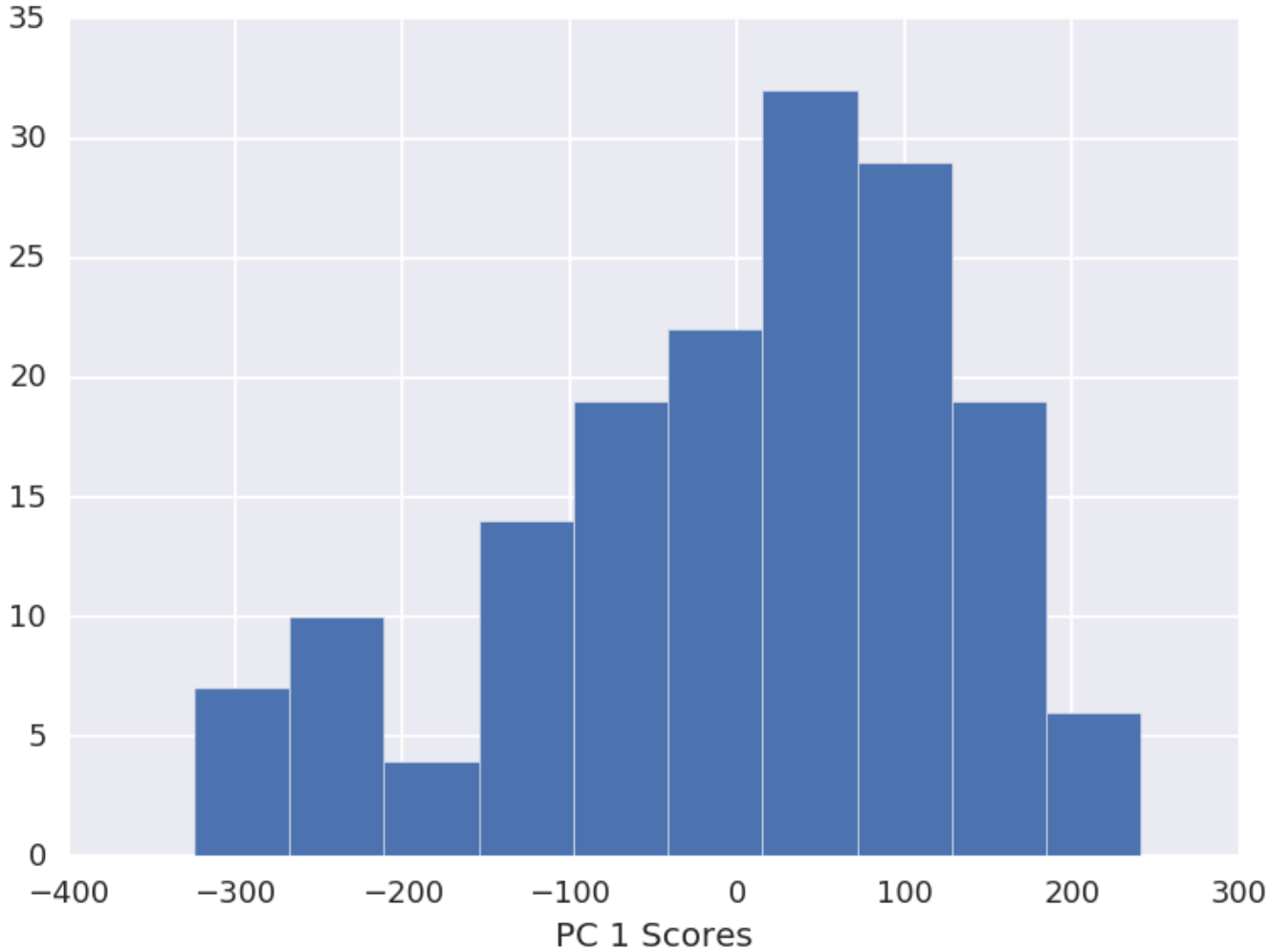
The direction of greatest variability is *different* from the linear regression line.

Identifying the g factor in data



We can **project** the points onto this direction to obtain **scores**.

Identifying the g factor in data



We've now reduced our two-dimensional data to just a single dimension.

So What Is PCA?

- Suppose we have variables $X_1, \ldots, X_p$
- *Principal components analysis* (PCA) is a **zero correlation, rotational transform** of these variables

So What Is PCA?

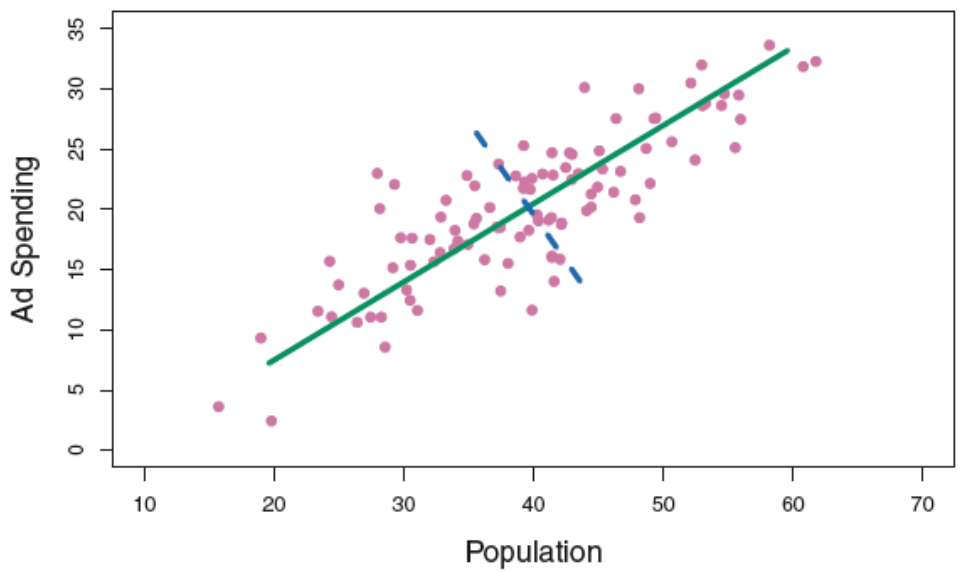
- Suppose we have variables $X_1, \ldots, X_p$
- *Principal components analysis* (PCA) is a **zero correlation, rotational transform** of these variables
    - ...with some bonus features and properties

So What Is PCA?

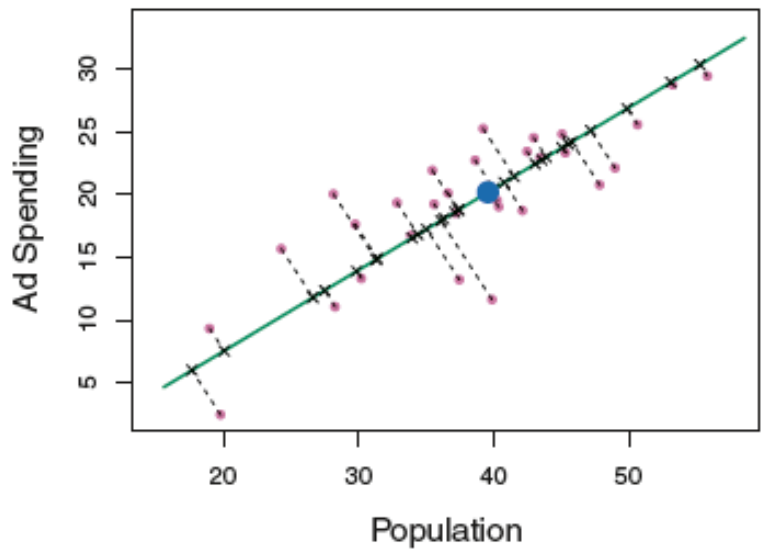
- Suppose we have variables $X_1, \ldots, X_p$
- *Principal components analysis* (PCA) is a **zero correlation, rotational transform** of these variables
  - ...with some bonus features and properties
- PCA finds a low-dimensional representation of a data set that contains as much of the variation as possible:

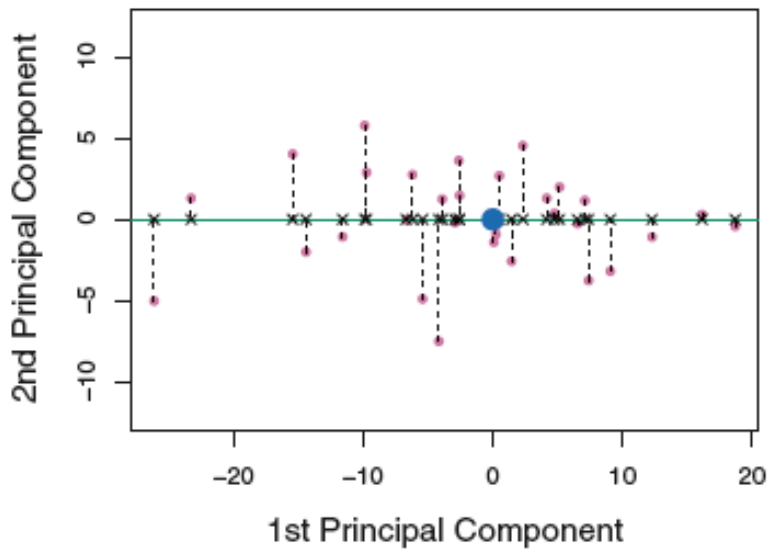
So What Is PCA?

- Suppose we have variables $X_1, \ldots, X_p$
- *Principal components analysis* (PCA) is a **zero correlation, rotational transform** of these variables
  - ...with some bonus features and properties
- PCA finds a low-dimensional representation of a data set that contains as much of the variation as possible:

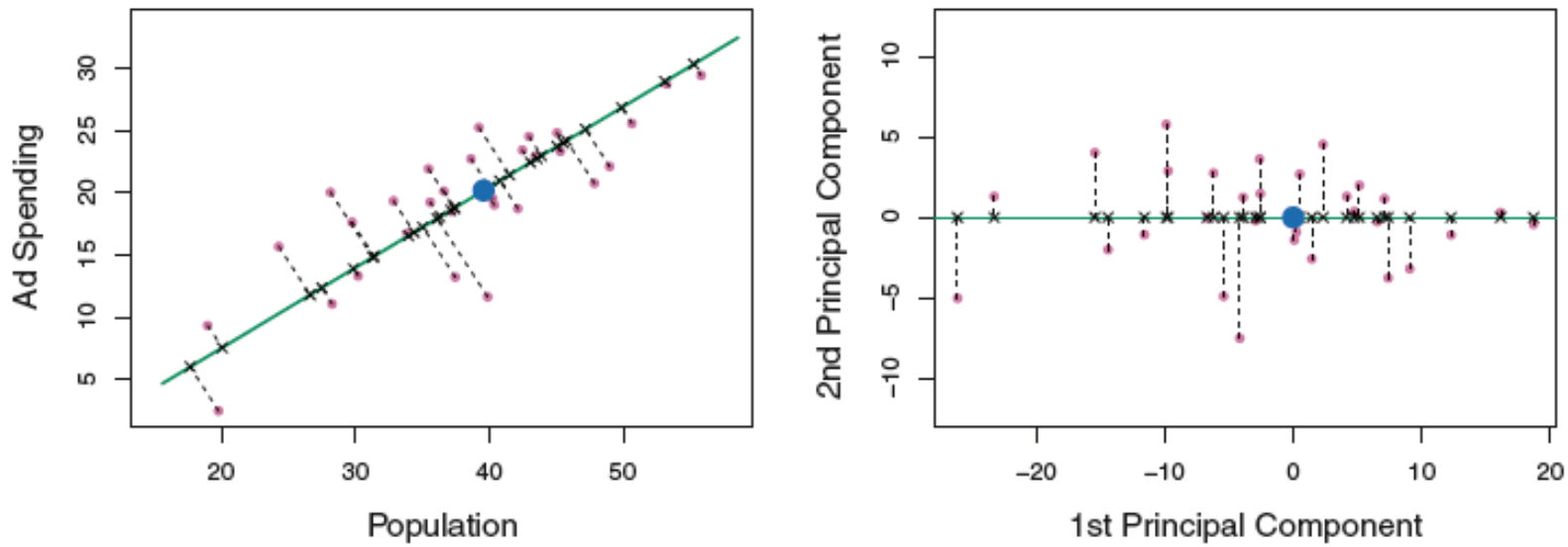


Zero Correlation Rotational Transform





Zero Correlation Rotational Transform



- **How are the principal components defined in terms of the original variables?**

Performing PCA

- We get $min(n-1, p)$ principal components in total

Performing PCA

- We get $min(n-1, p)$ principal components in total
- Each principal component is a linear combination of the original variables

Performing PCA

- We get $min(n-1, p)$ principal components in total
- Each principal component is a linear combination of the original variables
    - We call the converted values *scores*

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \cdots + \phi_{pi}x_{ip}$$

where z_{i1} is the **score** of the first principal component for the ith observation and the ϕ's are the principal component **loadings**

Performing PCA

- We get $min(n-1, p)$ principal components in total
- Each principal component is a linear combination of the original variables
    - We call the converted values *scores*

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \cdots + \phi_{pi}x_{ip}$$

where z_{i1} is the **score** of the first principal component for the ith observation and the ϕ's are the principal component **loadings**

- First PC has most variation; second PC is the linear combination that has maximal variance and is *uncorrelated* with the first PC and so on...

Performing PCA

- We get $min(n-1, p)$ principal components in total
- Each principal component is a linear combination of the original variables
  - We call the converted values *scores*

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \cdots + \phi_{pi}x_{ip}$$

where z_{i1} is the **score** of the first principal component for the ith observation and the ϕ's are the principal component **loadings**

- First PC has most variation; second PC is the linear combination that has maximal variance and is *uncorrelated* with the first PC and so on...
- **How are the PCs determined?**

Performing PCA

- We get $min(n-1, p)$ principal components in total
- Each principal component is a linear combination of the original variables
  - We call the converted values *scores*

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \cdots + \phi_{pi}x_{ip}$$

where z_{i1} is the **score** of the first principal component for the ith observation and the ϕ's are the principal component **loadings**

- First PC has most variation; second PC is the linear combination that has maximal variance and is *uncorrelated* with the first PC and so on...
- **How are the PCs determined?**
  - Eigen decomposition!

Another Interpretation of PCA

- Principal components provide low-dimensional linear surfaces that are *closest* to the observations

Another Interpretation of PCA

- Principal components provide low-dimensional linear surfaces that are *closest* to the observations
  - First PC: the **line** in p-dimensional space that is *closest* to the n observations

Another Interpretation of PCA

- Principal components provide low-dimensional linear surfaces that are *closest* to the observations
  - First PC: the **line** in p-dimensional space that is *closest* to the n observations
  - First 2 PCs: span the **plane** that is closest to the n observations
  - and so on...

More Details of PCA

- Since PCs are constructed to capture maximal variance in the original data, we want to ensure the process is “fair"

More Details of PCA

- Since PCs are constructed to capture maximal variance in the original data, we want to ensure the process is “fair"

If $\text{Var}(X) = 10$, then $\text{Var}(2.54X)$ will be _____.

More Details of PCA

- Since PCs are constructed to capture maximal variance in the original data, we want to ensure the process is “fair"

If $\text{Var}(X) = 10$, then $\text{Var}(2.54X)$ will be _____.

1. smaller
2. larger
3. Not enough information to tell

More Details of PCA

- Since PCs are constructed to capture maximal variance in the original data, we want to ensure the process is "fair"

If $\text{Var}(X) = 10$, then $\text{Var}(2.54X)$ will be _____.

1. smaller
2. larger
3. Not enough information to tell

- For this reason, it's common practice to individually scale the variables to have mean 0 and standard deviation 1

More Details of PCA contd.

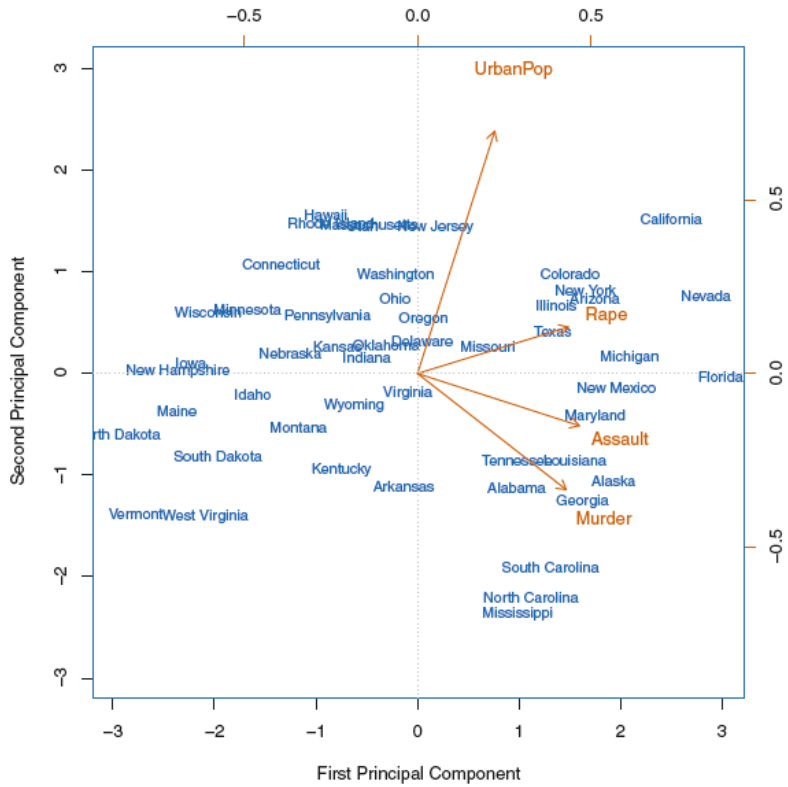
- Principal component loadings are unique up to a sign flip

More Details of PCA contd.

- Principal component loadings are unique up to a sign flip
  - The loading vector as a whole is usually what's interpreted
  - The interpretation won't change with a sign flip

More Details of PCA contd.

- Principal component loadings are unique up to a sign flip
  - The loading vector as a whole is usually what's interpreted
  - The interpretation won't change with a sign flip



	PC1	PC2
Murder	0.5358995	−0.4181809
Assault	0.5831836	−0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

More Details of PCA contd.

- The eigen decomposition process is actually able to tell us the *proportion of variance explained* (PVE) by each principal component!

More Details of PCA contd.

- The eigen decomposition process is actually able to tell us the *proportion of variance explained* (PVE) by each principal component!
- This is akin to telling us how much information from our original data is captured in a lower dimensional space (i.e. some number of PCs less than p)

More Details of PCA contd.

- The eigen decomposition process is actually able to tell us the *proportion of variance explained* (PVE) by each principal component!
- This is akin to telling us how much information from our original data is captured in a lower dimensional space (i.e. some number of PCs less than p)
- For the `Arrests` example:
  - 62.01% variation explained by the first PC (Z_1)
  - 24.74% variation explained by the second PC (Z_2)
  - 8.91% variation explained by the third PC (Z_3)
  - 4.34% variation explained by the fourth PC (Z_4)

Main Theme of PCA

Consolidate information present in data, into a lower dimensional space (i.e. using less than p variables)

Main Theme of PCA

Consolidate information present in data, into a lower dimensional space (i.e. using less than p variables)

- Popular uses:

Main Theme of PCA

Consolidate information present in data, into a lower dimensional space (i.e. using less than p variables)

- Popular uses:
  - Visualization

Main Theme of PCA

Consolidate information present in data, into a lower dimensional space (i.e. using less than p variables)

- Popular uses:
  - Visualization
  - Dimension reduction (i.e. use PCs instead of raw variables)

Main Theme of PCA

Consolidate information present in data, into a lower dimensional space (i.e. using less than p variables)

- Popular uses:
  - Visualization
  - Dimension reduction (i.e. use PCs instead of raw variables)
    - Principal components **regression**

Main Theme of PCA

Consolidate information present in data, into a lower dimensional space (i.e. using less than p variables)

- Popular uses:
  - Visualization
  - Dimension reduction (i.e. use PCs instead of raw variables)
    - Principal components **regression**
    - Really any statistical technique or learning method!

Main Theme of PCA

Consolidate information present in data, into a lower dimensional space (i.e. using less than p variables)

- Popular uses:
  - Visualization
  - Dimension reduction (i.e. use PCs instead of raw variables)
    - Principal components **regression**
    - Really any statistical technique or learning method!
    - Remember that each PC involves *all* of the original variables
      $\Longrightarrow$ PCA is not a variable selection procedure!

A Personal Example – Tasseled Cap Transformation

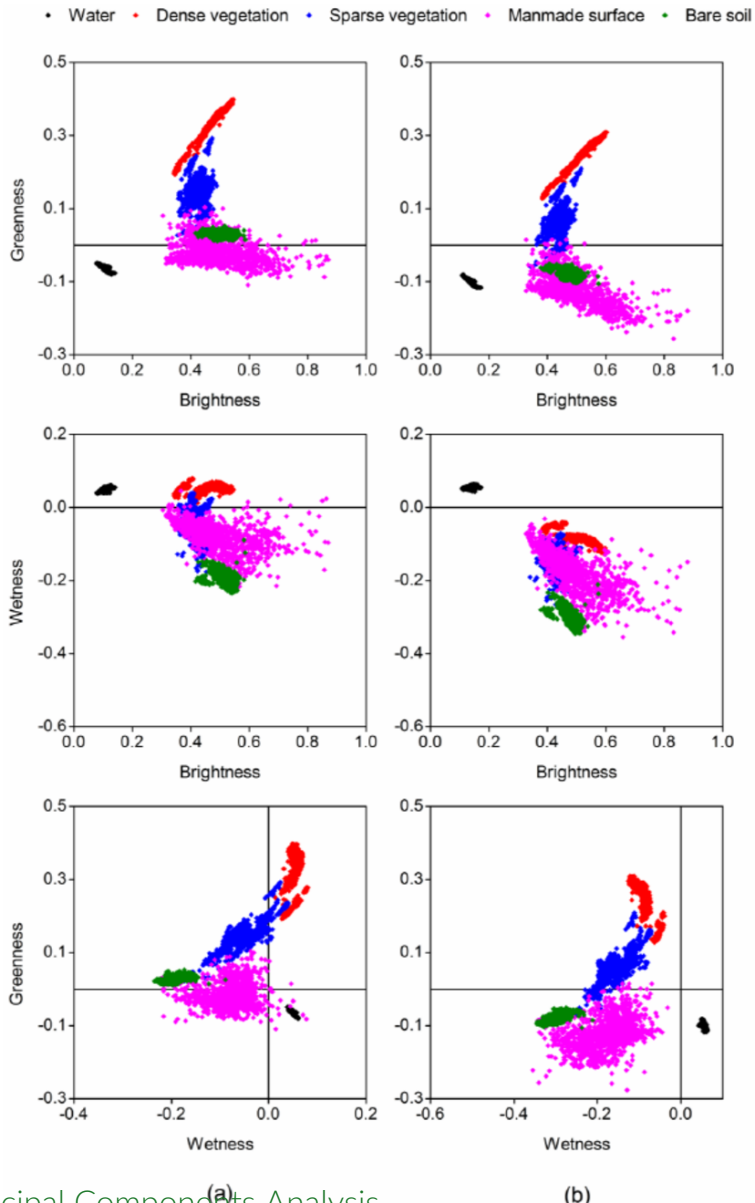


Table 2. TCT coefficients for Landsat 8 at-satellite reflectance.

Landsat 8 TCT	(Blue) Band 2	(Green) Band 3	(Red) Band 4	(NIR) Band 5	(SWIR1) Band 6	(SWIR2) Band 7
Brightness	0.3029	0.2786	0.4733	0.5599	0.508	0.1872
Greenness	−0.2941	−0.243	−0.5424	0.7276	0.0713	−0.1608
Wetness	0.1511	0.1973	0.3283	0.3407	−0.7117	−0.4559
TCT4	−0.8239	0.0849	0.4396	−0.058	0.2013	−0.2773
TCT5	−0.3294	0.0557	0.1056	0.1855	−0.4349	0.8085
TCT6	0.1079	−0.9023	0.4119	0.0575	−0.0259	0.0252

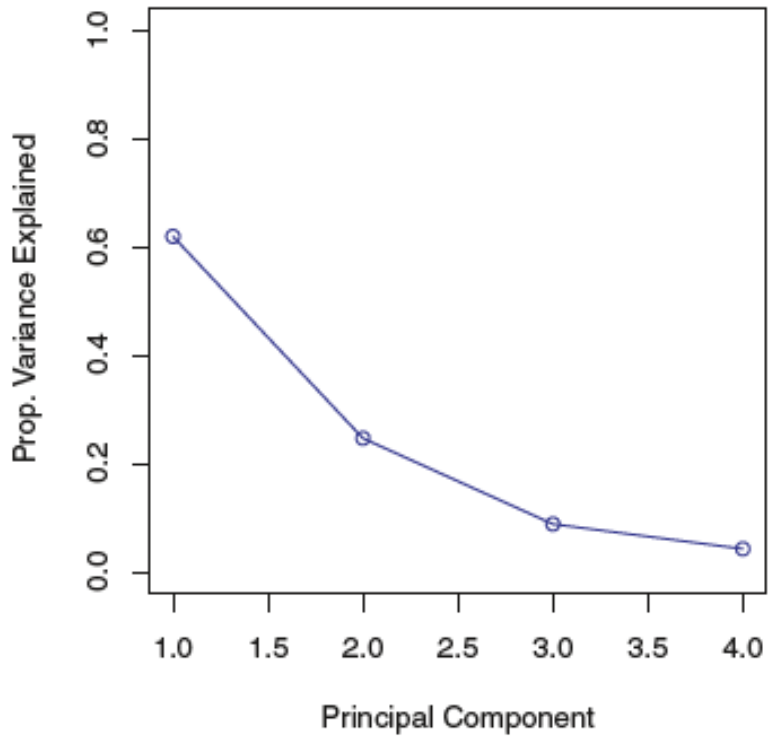
How Many Principal Components Do We Use?

How Many Principal Components Do We Use?

- **Cross-validation!**

How Many Principal Components Do We Use?

- **Cross-validation!**
- Visual inspection of a *scree plot*:



- Look for the *elbow*