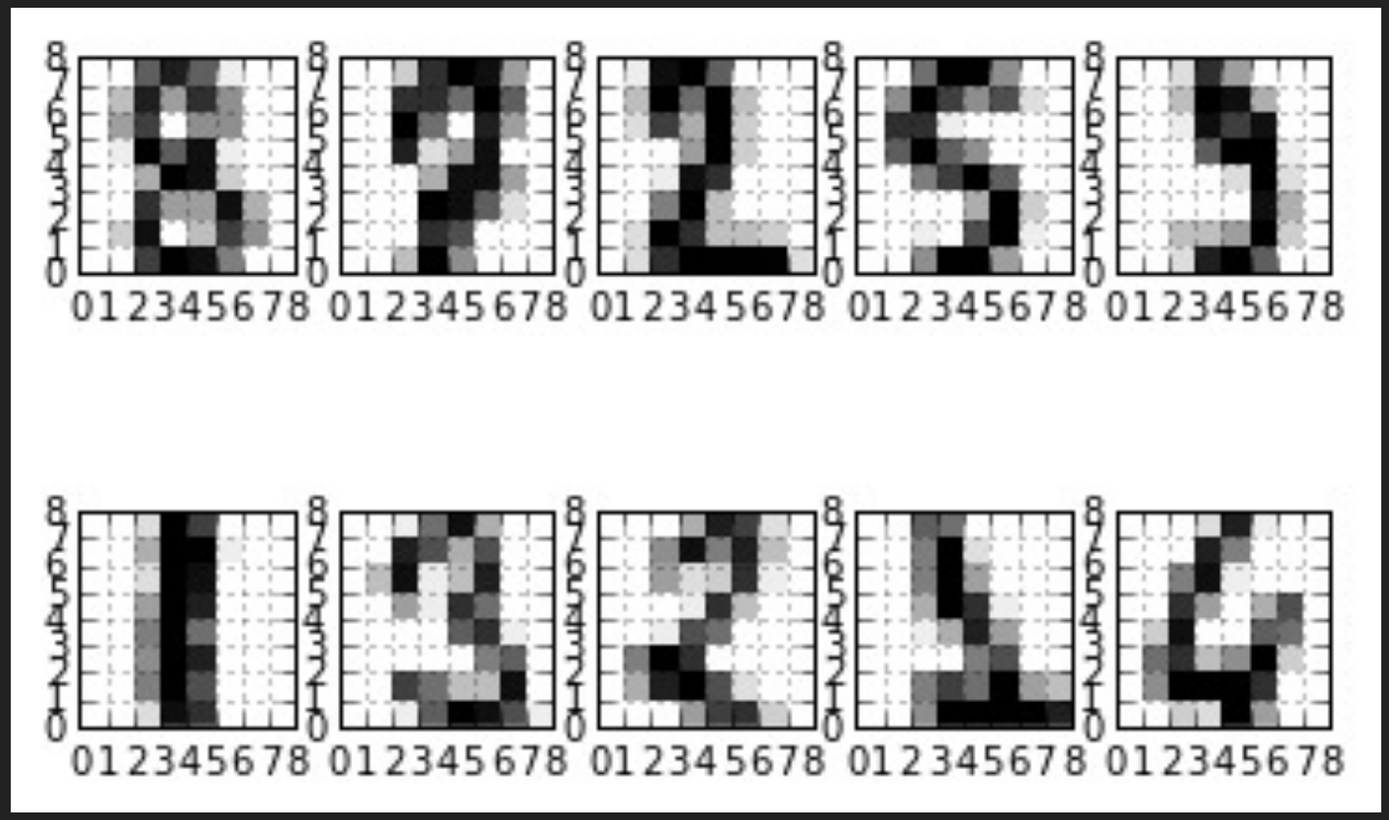
UNCHITTA KANJANASARATOOL

INTRO TO PRINCIPAL COMPONENT ANALYSIS

unchitta@ucla.edu

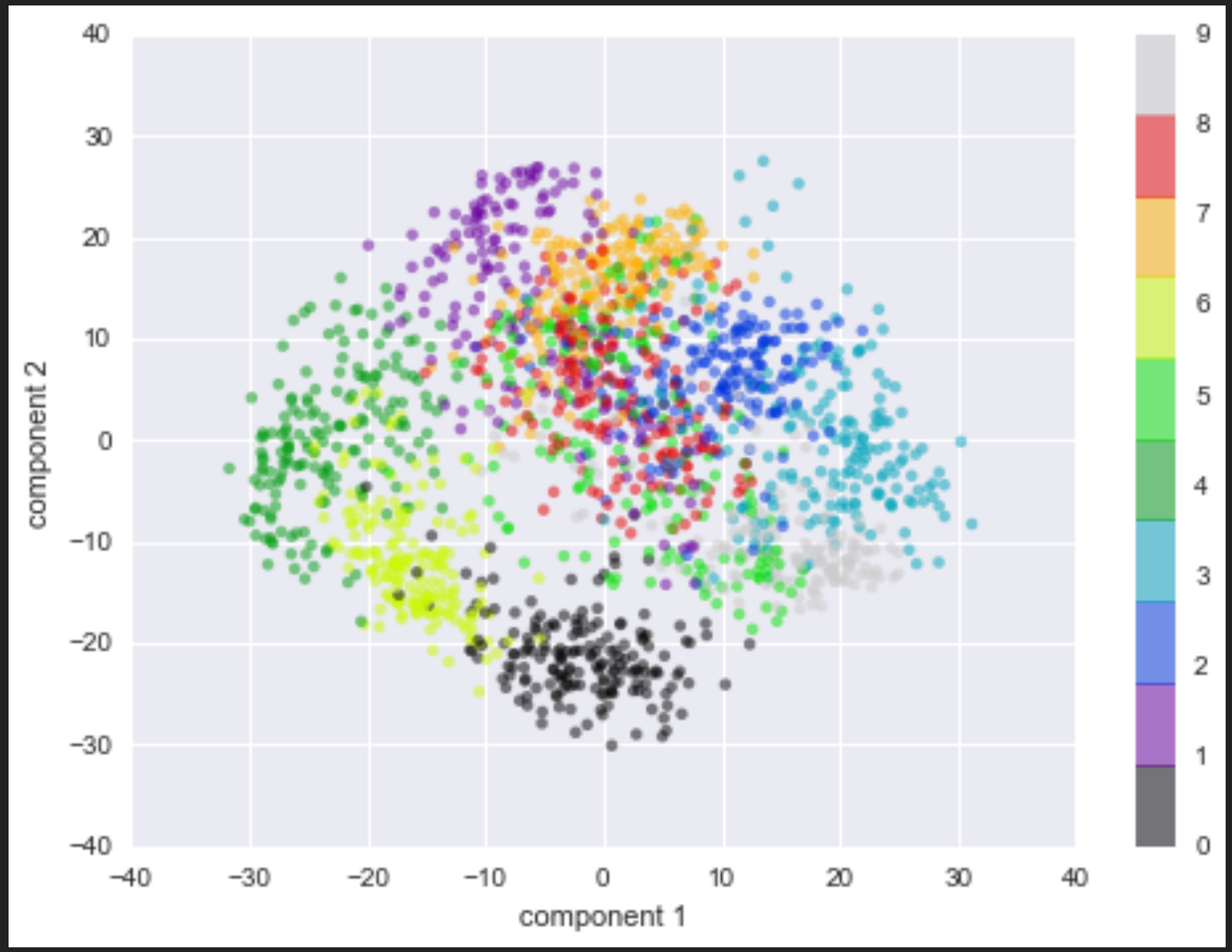
MOTIVATION: DIGITS DATA SET & PCA



Each data point is an 8x8 image. So, 64 dimensions.

n_x64 matrix…How do you visualize this data set?

unchitta@ucla.edu



Source: Python Data Science Handbook, Jake VanderPlas

unchitta@ucla.edu

WHY DIMENSION REDUCTION?

- Anything can go wrong in high dimensions
- Big Data => Data compression
- Visualization

CAUTIONS

- Data loss
- Some methods such as PCA tend to be linear
- Garbage in = Garbage out

unchitta@ucla.edu

DIMENSION REDUCTION METHODS

- **2 TYPES**
  - FEATURE SELECTION
  - FEATURE EXTRACTION
    - Linear & Generalized Discriminant Analysis (LDA & GDA)
    - **Principal Component Analysis**

unchitta@ucla.edu

PCA OUTLINE

- Variance & Covariance
- Intuition behind principle component analysis
- Eigenvectors; Covariance matrix
- Finding the principal components
- Examples
- Resources
- Questions

unchitta@ucla.edu

VARIANCE & COVARIANCE

- **VARIANCE:** Measures the variability or "spread" of the data. Squared of standard deviation.

- **COVARIANCE:** Measures the strength of the relationship between 2 or more variables (can be *positive*, *negative*, or *no relationship*).

$$\sigma^2 = \sum \frac{(X - \mu)^2}{N}$$

$$Cov(X,Y) = \frac{\sum (X_i - \bar{X}) * (Y_i - \bar{Y})}{n}$$

unchitta@ucla.edu

INTUITION BEHIND PRINCIPAL COMPONENT ANALYSIS

- Example from Penn State STAT 505: Imagine a data set where observations are ratings of communities on 9 different criteria:
  - Housing, Climate, Health, Crime, Transportation, Education, Arts, Recreation, Economics
- **Can we construct perhaps 4 new features from the existing ones that would still contain relatively as much information?**

unchitta@ucla.edu

INTUITION BEHIND PRINCIPAL COMPONENT ANALYSIS

- PCA: geometrically, linear transformation from one coordinate system to another
- The bases for the new subspace are called ***principal components (PCs)***, constructed using linear combination of the variables of the original data
- New variables created by PCA should represent the same amount of information as the original variables (i.e. same total variance, just a change of basis)

unchitta@ucla.edu

INTUITION BEHIND PRINCIPAL COMPONENT ANALYSIS

- *m*-features data can produce *m* different principal components.
- Reduce dimension by eliminating later PCs that have low variances.
  - Projection of *p*-dimensional data onto lower *k*-dimensional subspace (k < p)

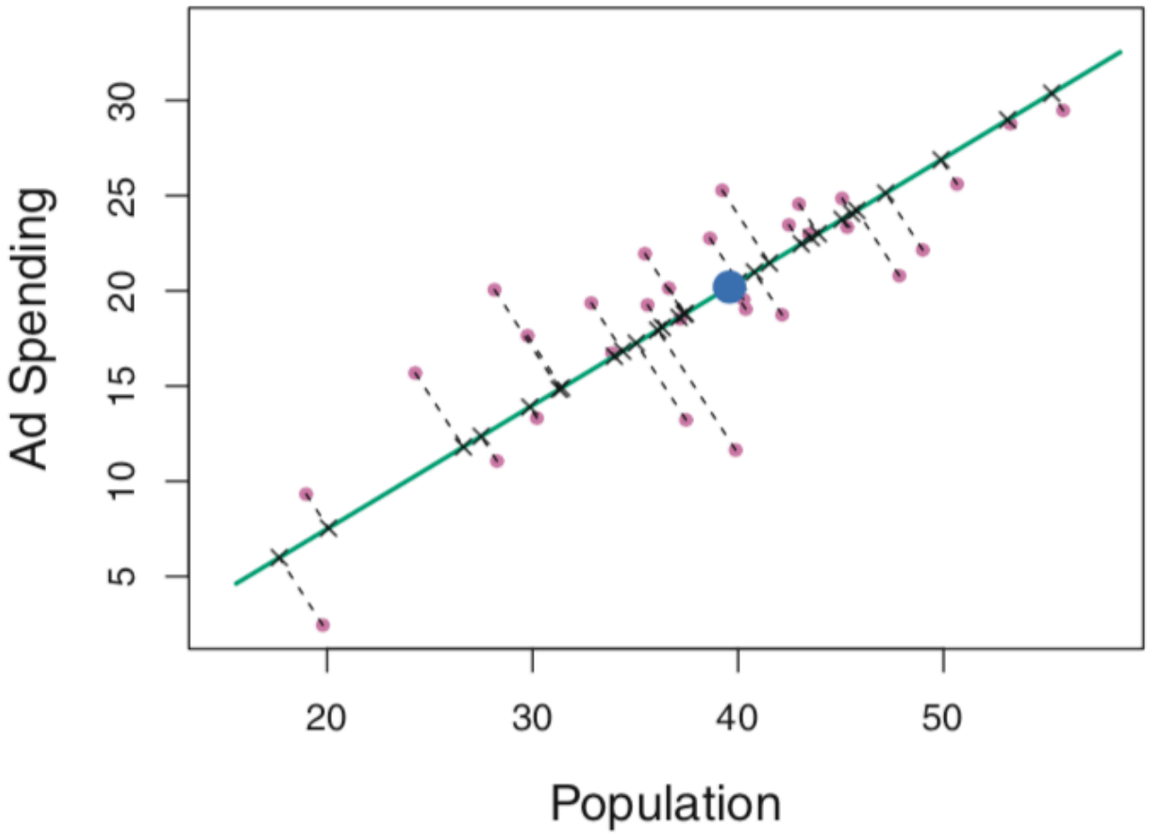
unchitta@ucla.edu

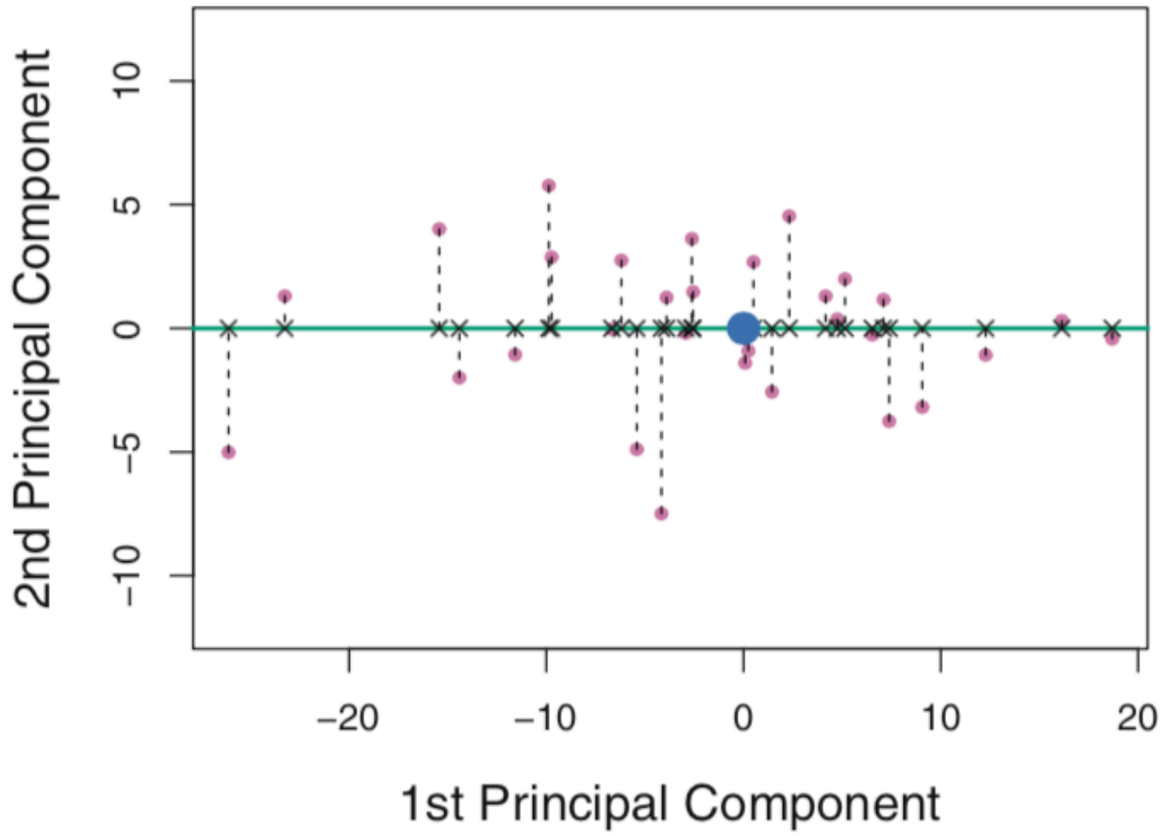
DIGGING DEEPER

- PC1 is chosen to be the direction in which observations vary the most (highest variance).
    - Why? Similar observations don't reveal much information and pattern, do they?
    - Another interpretation
- PC2 is required to be uncorrelated to PC1 (i.e., they are *orthogonal*) and by constraint points in direction with second highest variance, etc.

unchitta@ucla.edu

DIGGING DEEPER





Source: Introduction to Statistical Learning with Applications in R (p.232)
Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

unchitta@ucla.edu

REVIEW: EIGENVECTORS & EIGENVALUES

- The eigenvector of a linear transformation matrix X is a vector $\boldsymbol{v}$ such that when multiplied by the transformation gives back the vector itself, scaled by a scalar (its corresponding eigenvalue).

$$X\boldsymbol{v} = \boldsymbol{\lambda}\boldsymbol{v}$$

- Symmetric matrices have some exceptional properties...

unchitta@ucla.edu

COVARIANCE MATRIX

$$\text{Cov}(\boldsymbol{X}) = \begin{bmatrix} \text{Cov}(x_1,x_1) & \text{Cov}(x_1,x_2) & \text{Cov}(x_1,x_3) \\ \text{Cov}(x_2,x_1) & \text{Cov}(x_2,x_2) & \text{Cov}(x_2,x_3) \\ \text{Cov}(x_3,x_1) & \text{Cov}(x_3,x_2) & \text{Cov}(x_3,x_3) \end{bmatrix}$$

- If the data has been centered to have mean zero, $\text{Cov}(\boldsymbol{X})$ is given conveniently by

$$\text{Cov}(\boldsymbol{X}) = \frac{1}{n\text{-}1}(\boldsymbol{X}^T\boldsymbol{X})$$

- Property: $\boldsymbol{X}^T\boldsymbol{X}$ is symmetric, diagonalizable, and **positive semi-definite** (so all eigenvalues positive)

unchitta@ucla.edu

REPRESENTING COV(X) WITH EIGENVECTORS

- Cov($\boldsymbol{X}$) reveals shape (spread and orientation) of the data.
- Represent Cov($\boldsymbol{X}$) with a vector: should point in the direction of the largest spread of the data, and magnitude should equal the spread (variance) in this direction.
- Turns out that the *largest eigenvector* of Cov($\boldsymbol{X}$) points in this direction, and its magnitude equals its corresponding eigenvalue.
    - Rayleigh Quotient

unchitta@ucla.edu

SO, FINDING PRINCIPAL COMPONENTS

Mathematically...

PCA: Eigendecomposition of ($\boldsymbol{X}^\top \boldsymbol{X}$)

&

Linear Transformation of $\boldsymbol{X}$

unchitta@ucla.edu

FINDING PRINCIPAL COMPONENTS

- GENERAL PROCEDURE:
  - Standardize data
  - Find the covariance matrix of the data
  - Find eigenvectors of the covariance matrix and corresponding eigenvalues, then normalize eigenvectors
  - Sort eigenvectors according to their eigenvalues in decreasing order
  - Choose first *k* vectors to be new *k* dimensions
  - Transform data into *k* dimensions (change of basis)

unchitta@ucla.edu

EXPLAINED VARIANCE

Perhaps the most important question of PCA:

How many PCs to keep?

unchitta@ucla.edu

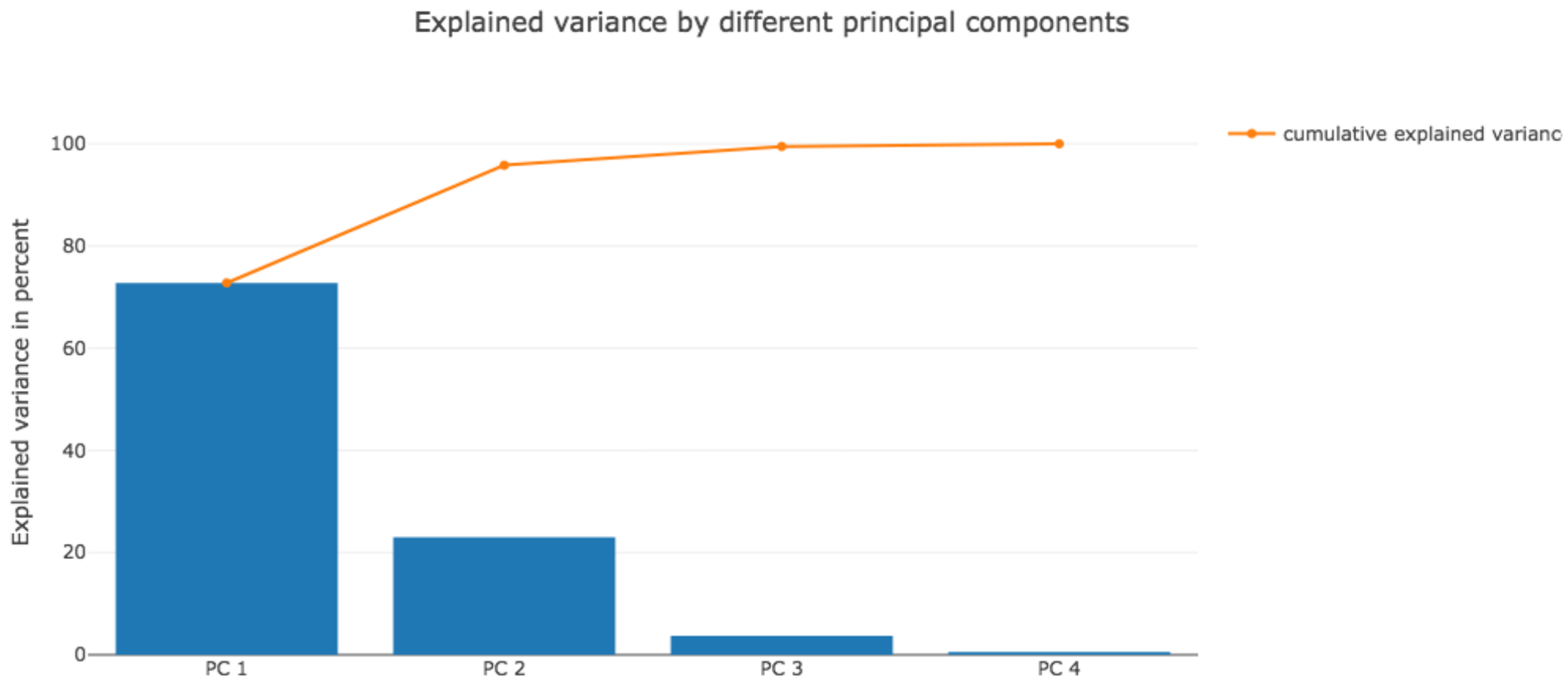
EXPLAINED VARIANCE

- **Total variance**: sum of all individual variances of either the original variables or the PCs (they are the same).
- **Explained variance** or the variance explained by a PC: the ratio between the variance of that PC and the total variance.
  - We want to maximize explained variance (i.e., sufficiently small loss of information) while at the same time also keep interpretability in mind

unchitta@ucla.edu

EXPLAINED VARIANCE

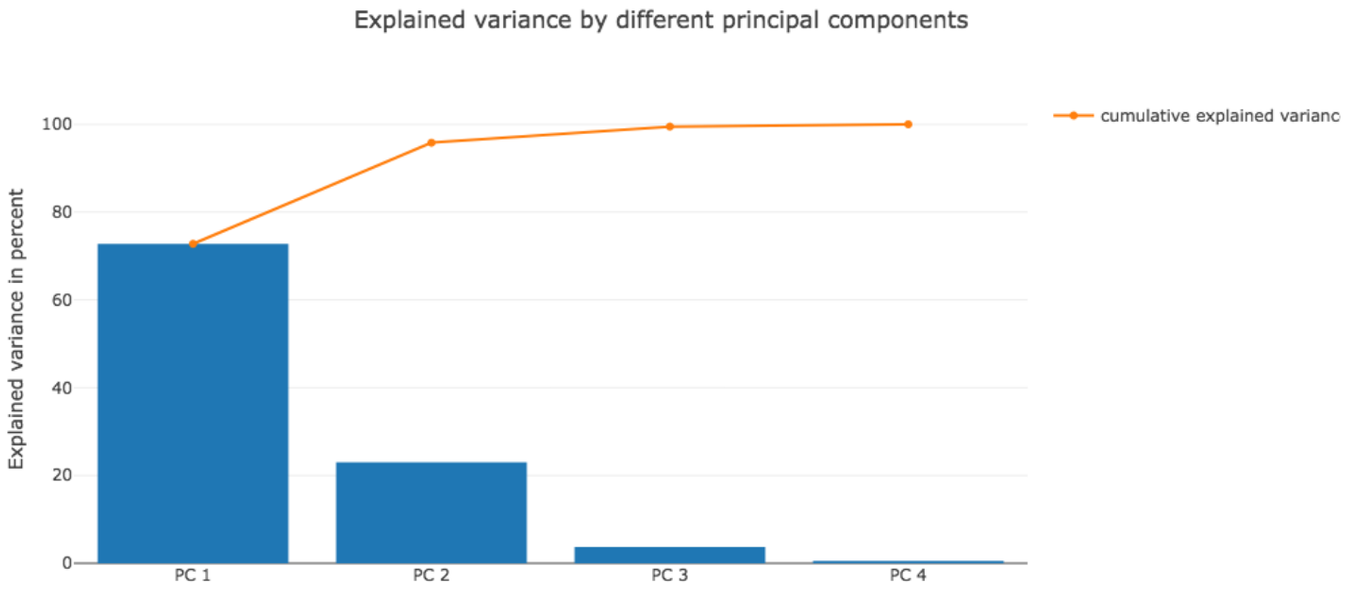
- In practice, we graph the number of principal components against the explained variance, like below:



Source: plot.ly/ipython-notebooks/principal-component-analysis/

unchitta@ucla.edu

EXPLAINED VARIANCE



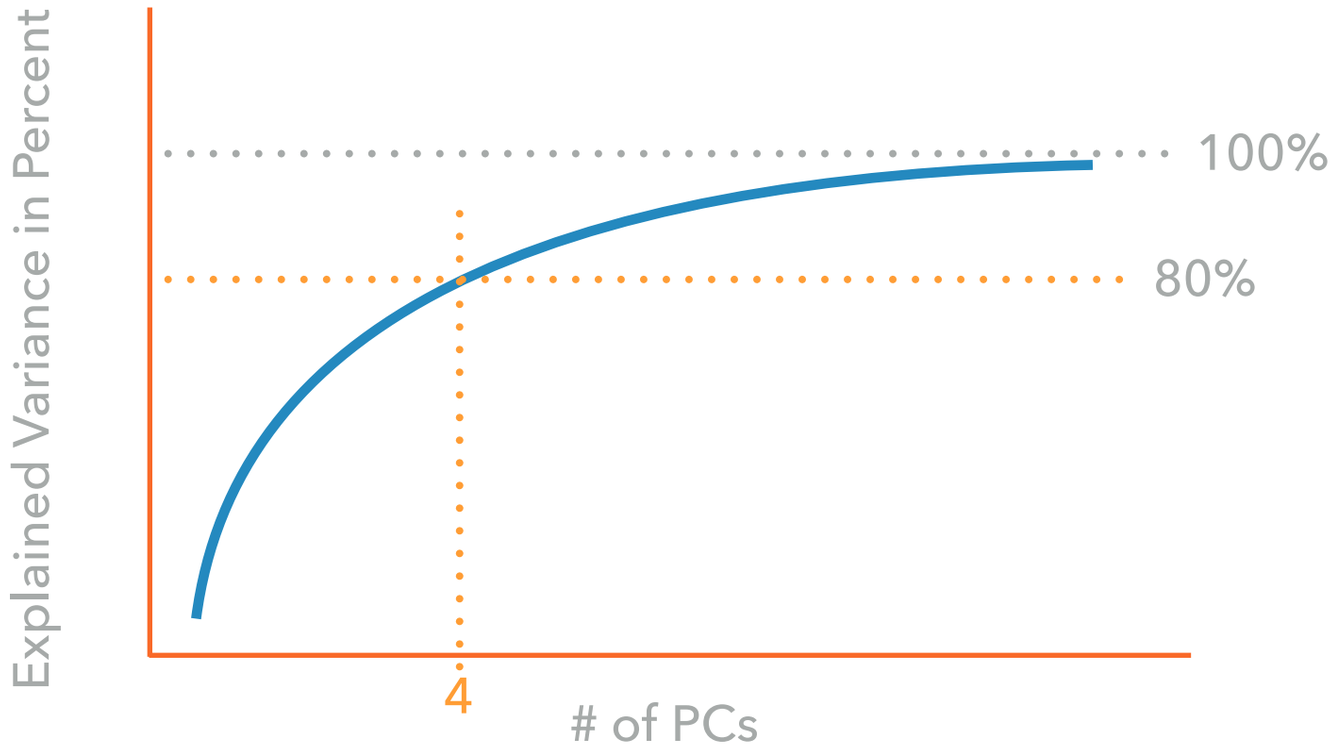
Source: plot.ly/ipython-notebooks/principal-component-analysis/

- From this plot, the first 2 principal components account for ~95% of the information
- Safe to discard last 2 components
- Went from 4 dimensions to 2 dimensions

unchitta@ucla.edu

EXPLAINED VARIANCE

- Shape of the explained variance curve is not always a sharp shoulder, in which case we may set a definite threshold (e.g. say, we want just 80% of the info)



unchitta@ucla.edu

- Example from Introduction to Statistical Learning with Applications in R (G. James): USArrests data set.
  - Data contains number of arrests per 100,000 residents in each of 50 states, for each of the 3 crimes: Assault, Murder, and Rape. Data also contain UrbanPop, percent of population in each state living in urban area.
- The authors performed PCA on data and computed 2 principal components, then transformed the data into 2-dimensional space

unchitta@ucla.edu

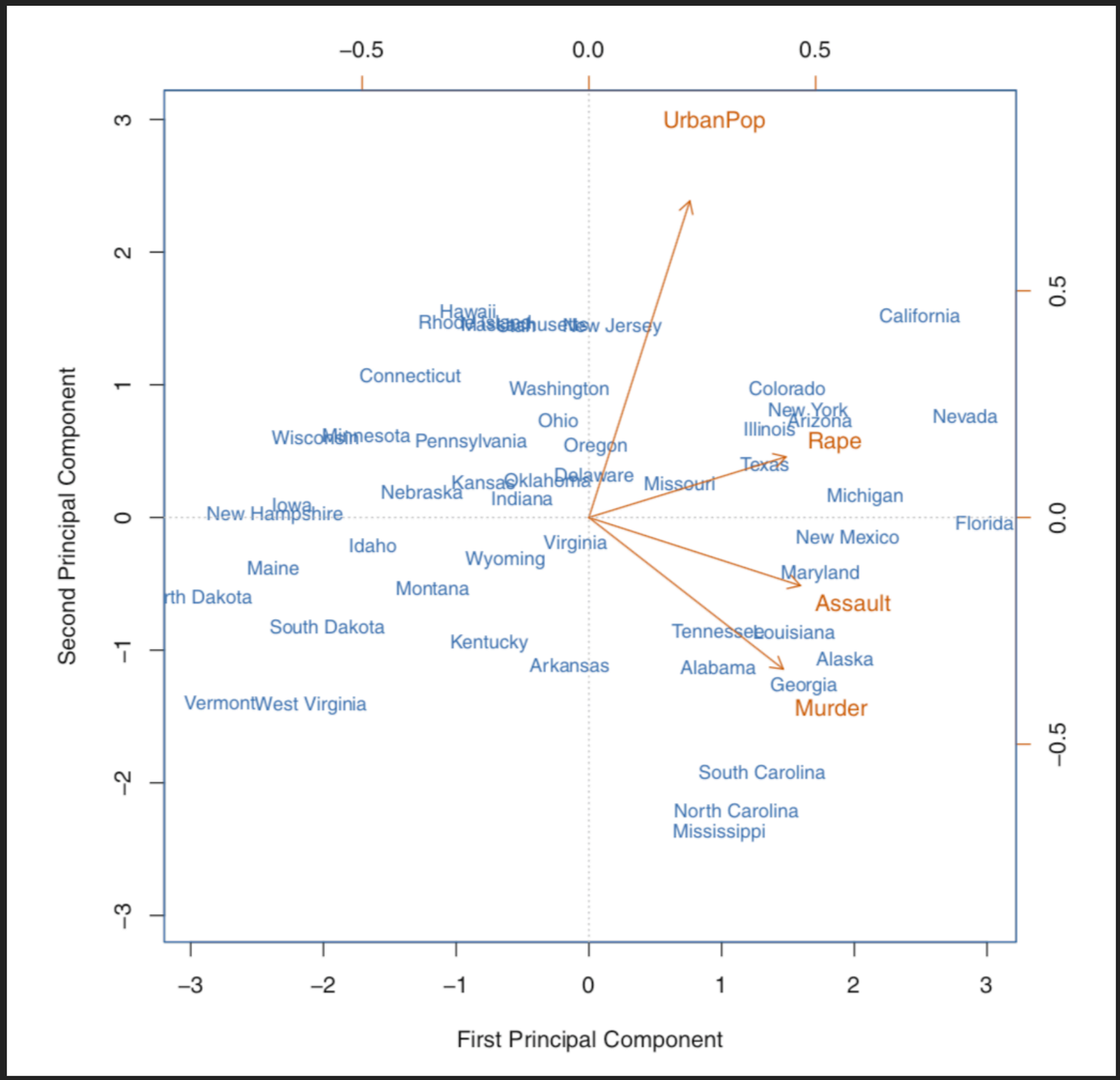
	PC1	PC2
Murder	0.5358995	−0.4181809
Assault	0.5831836	−0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

PC1 and PC2 are the principal component “loading vectors” aka the eigenvectors. The weights are the “loadings.”

PC1 is associated much more with Murder, Assault, and Rape than UrbanPop. PC2 is dominantly associated with UrbanPop. UrbanPop is much less correlated with the rest.

EXAMPLE

unchitta@ucla.edu



Biplot (scores + loadings) – Interpretation

unchitta@ucla.edu

OTHER REMARKS

- PCA does *NOT* change the data in any way
- Sometimes the patterns are not clear
- Messy data will give you messy principal components
- PCA tends to find linear relationships
- Supervised version of PCA: Principal Component Regression
- SVD is more common in practice and more stable numerically, but is analogous to eigendecomposition, but with a few other requirements

unchitta@ucla.edu

RESOURCES

pastebin.com/ADTbqHSX

QUESTIONS

THANK YOU FOR COMING!