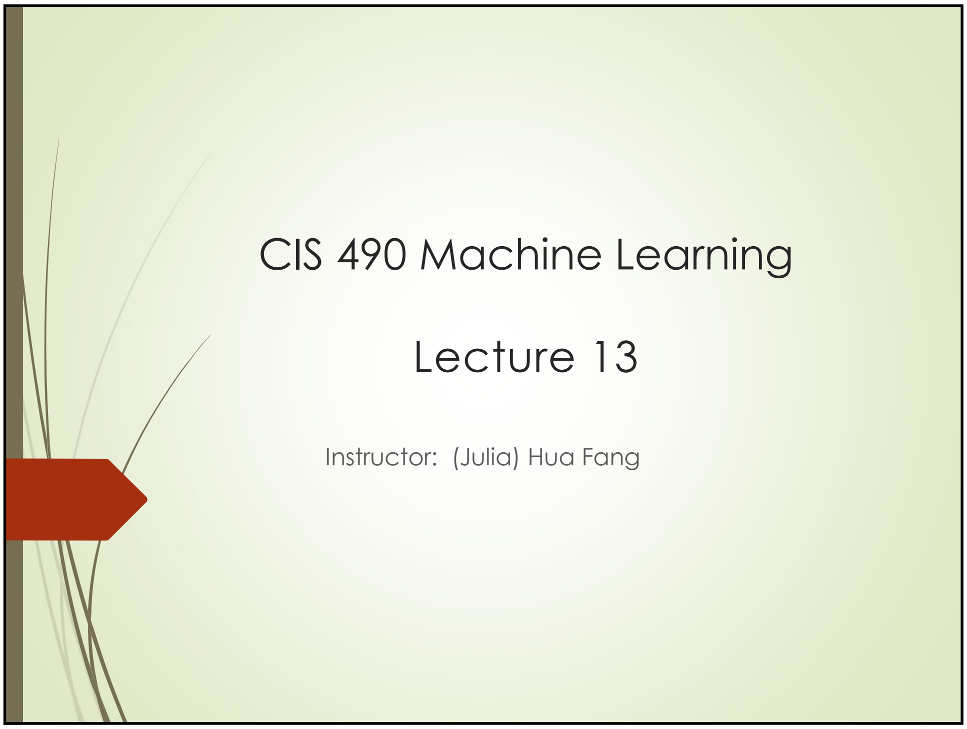
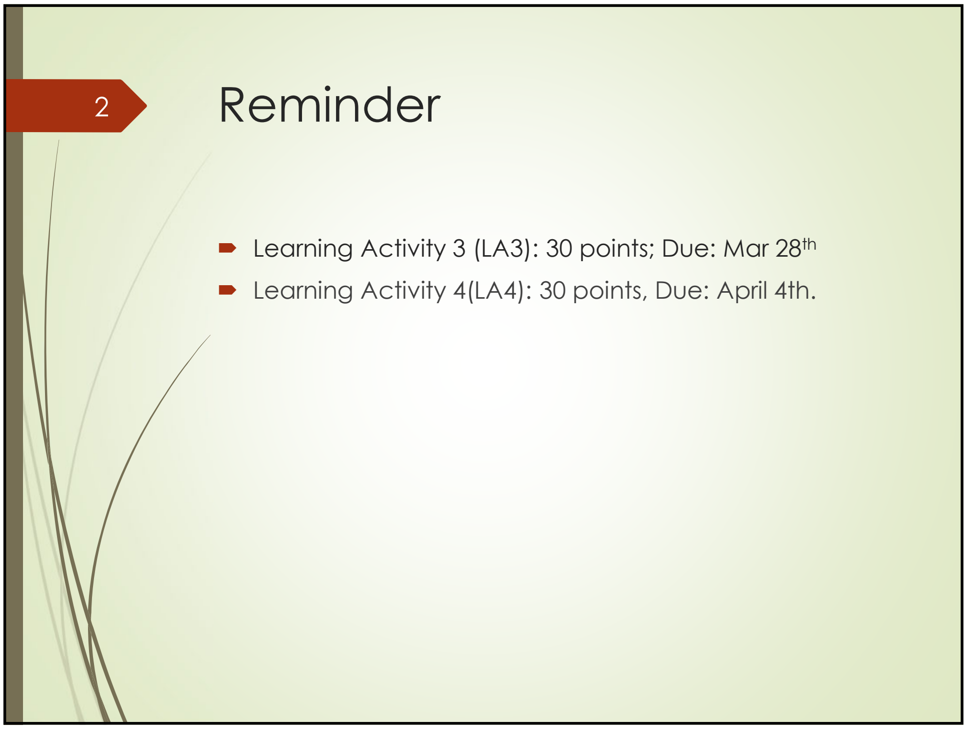


CIS 490 Machine Learning

Lecture 13

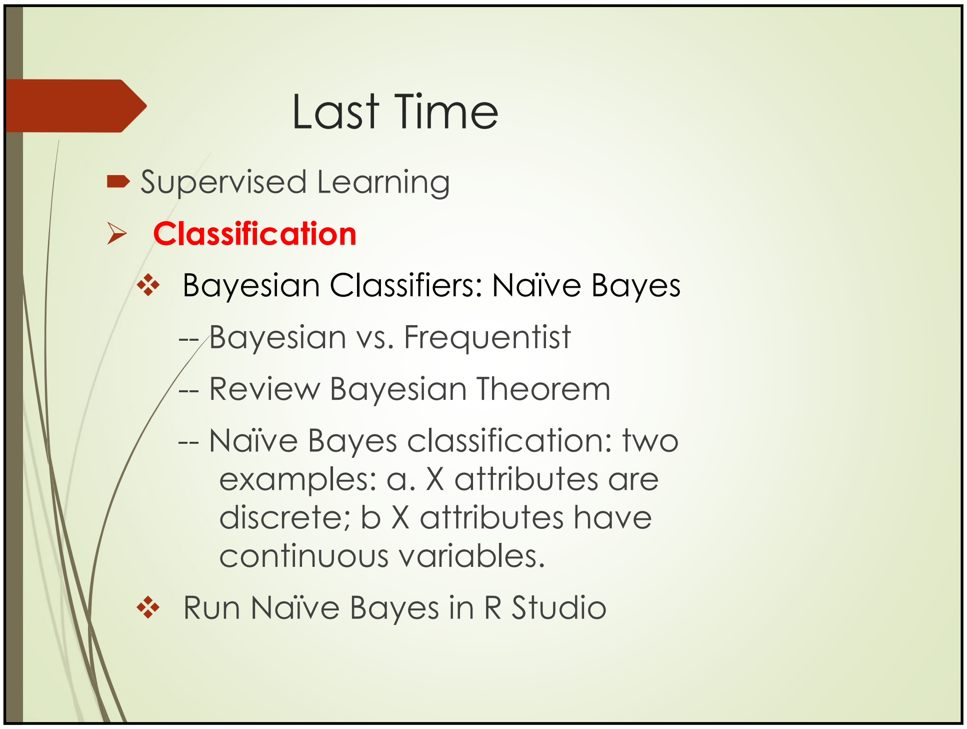
Instructor: (Julia) Hua Fang



2

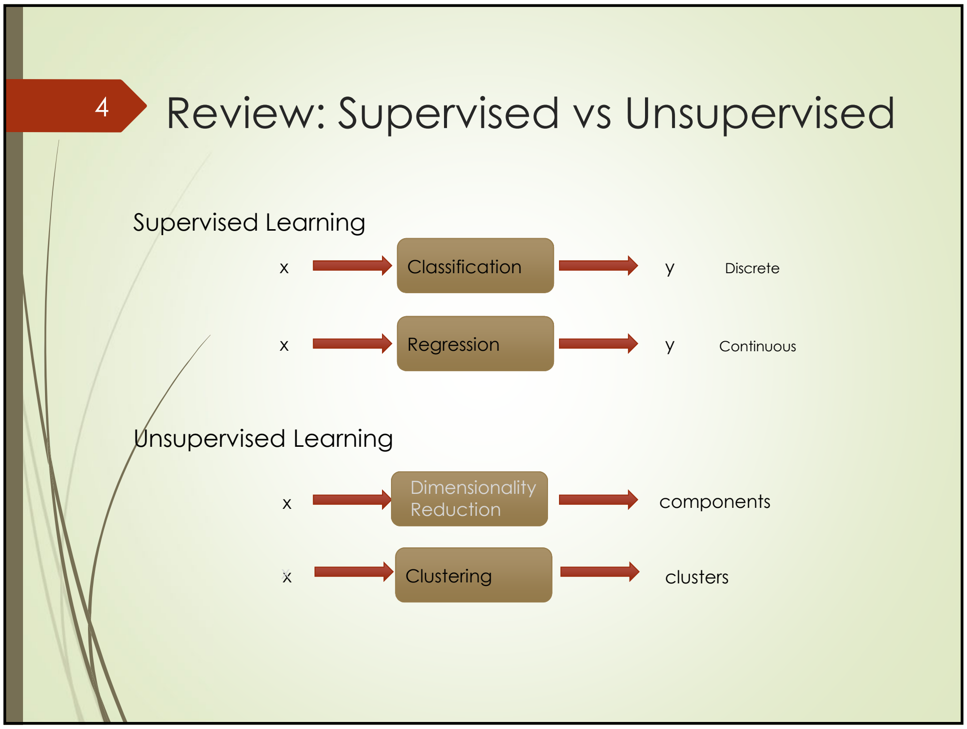
Reminder

- Learning Activity 3 (LA3): 30 points; Due: Mar 28th
- Learning Activity 4(LA4): 30 points, Due: April 4th.



Last Time

- Supervised Learning
- **Classification**
  - Bayesian Classifiers: Naïve Bayes
    - -- Bayesian vs. Frequentist
    - -- Review Bayesian Theorem
    - -- Naïve Bayes classification: two examples: a. X attributes are discrete; b X attributes have continuous variables.
  - Run Naïve Bayes in R Studio



4 Review: Supervised vs Unsupervised

Supervised Learning

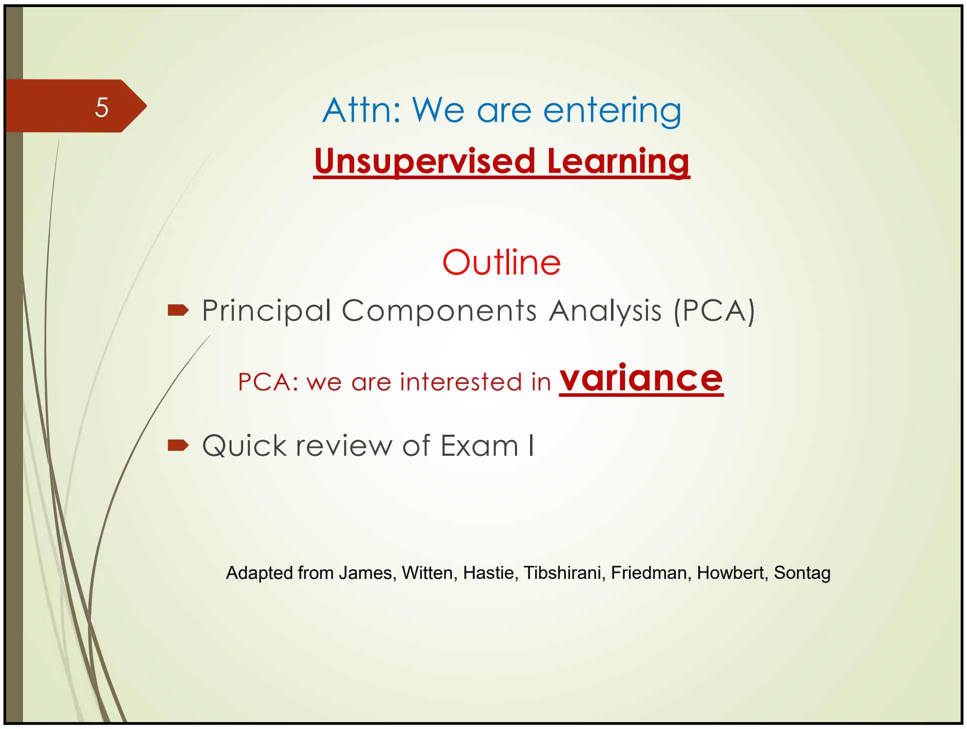
x → Classification → y Discrete

x → Regression → y Continuous

Unsupervised Learning

x → Dimensionality Reduction → components

x → Clustering → clusters



5

Attn: We are entering **Unsupervised Learning**

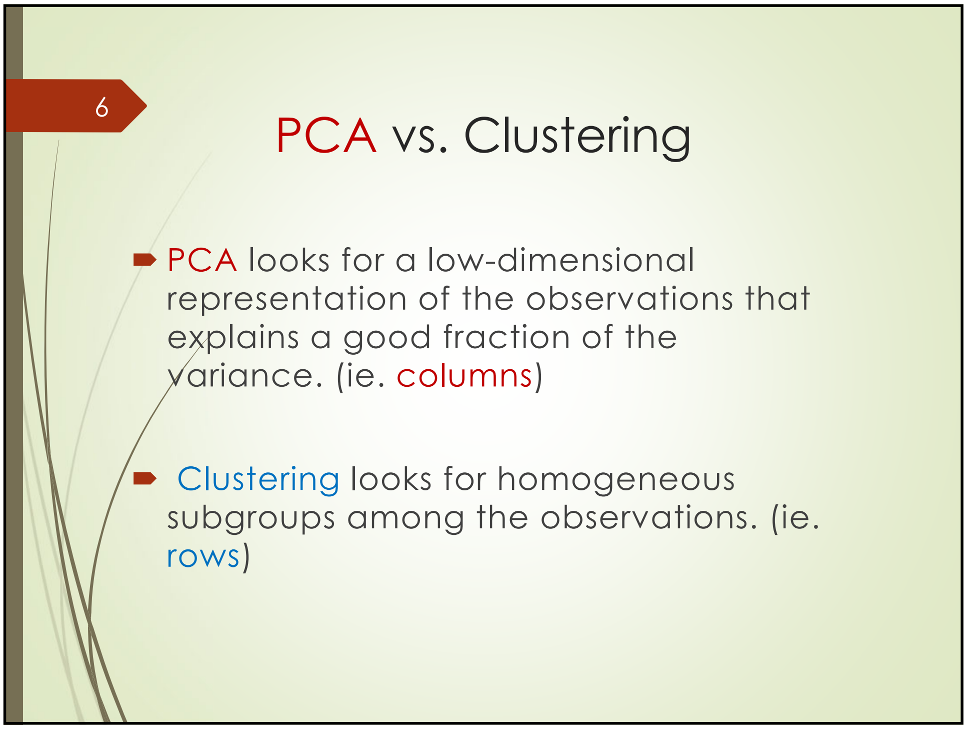
Outline

- Principal Components Analysis (PCA)

PCA: we are interested in **variance**

- Quick review of Exam I

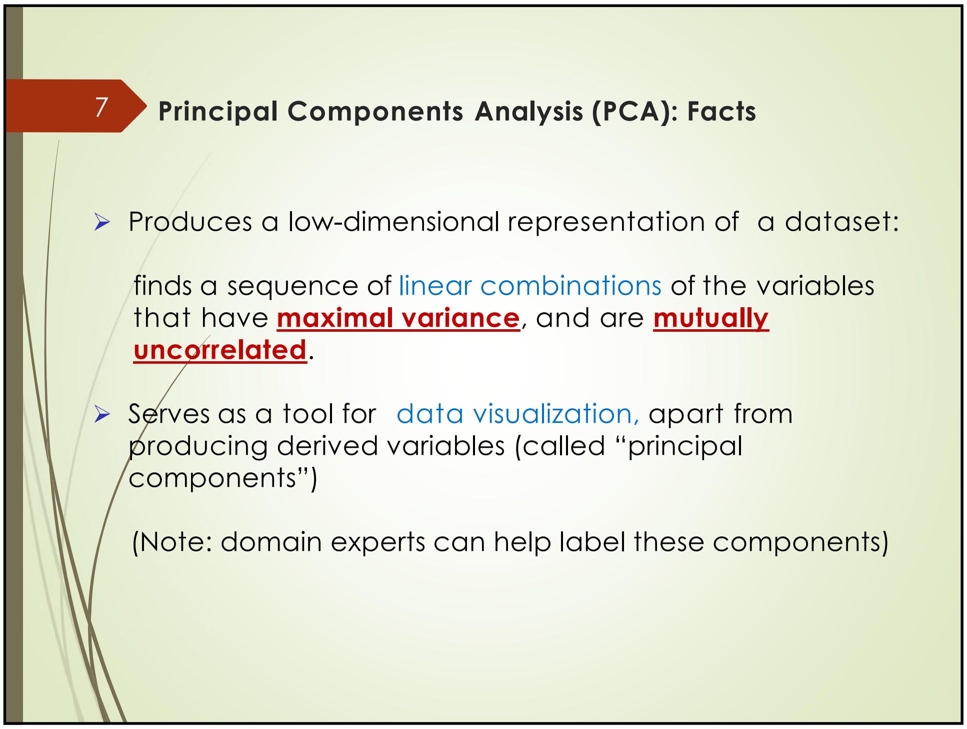
Adapted from James, Witten, Hastie, Tibshirani, Friedman, Howbert, Sontag



6

PCA vs. Clustering

- PCA looks for a low-dimensional representation of the observations that explains a good fraction of the variance. (ie. columns)
- Clustering looks for homogeneous subgroups among the observations. (ie. rows)



7

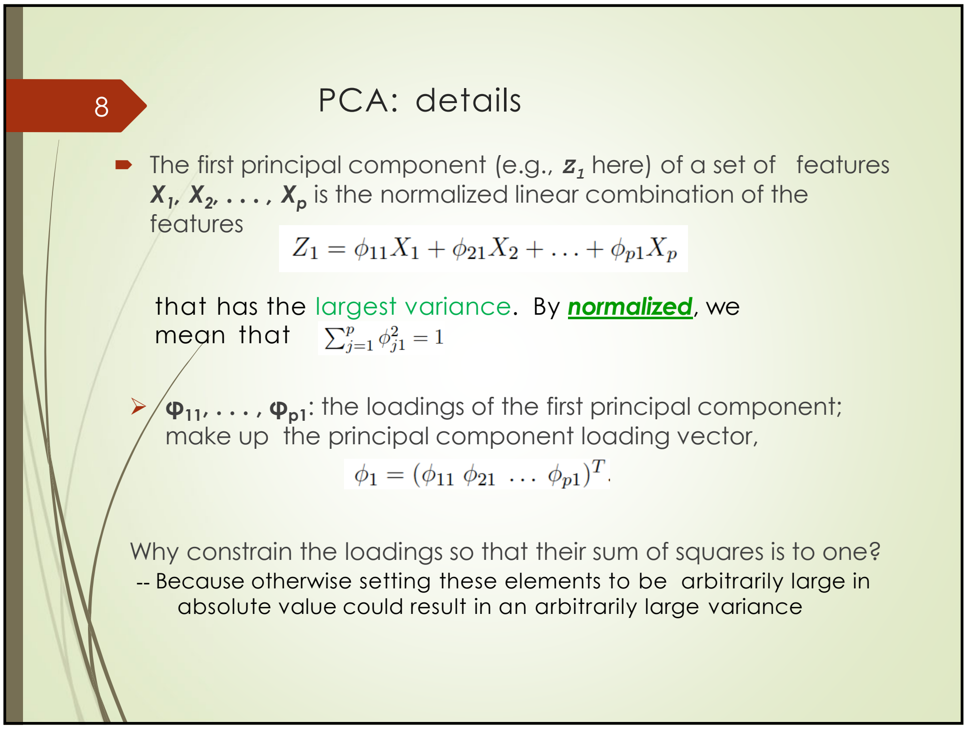
Principal Components Analysis (PCA): Facts

- Produces a low-dimensional representation of a dataset:

  finds a sequence of linear combinations of the variables that have **maximal variance**, and are **mutually uncorrelated**.

- Serves as a tool for data visualization, apart from producing derived variables (called "principal components")

  (Note: domain experts can help label these components)



8

PCA: details

- The first principal component (e.g., z_1 here) of a set of features $X_1, X_2, \ldots, X_p$ is the normalized linear combination of the features

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \ldots + \phi_{p1}X_p$$

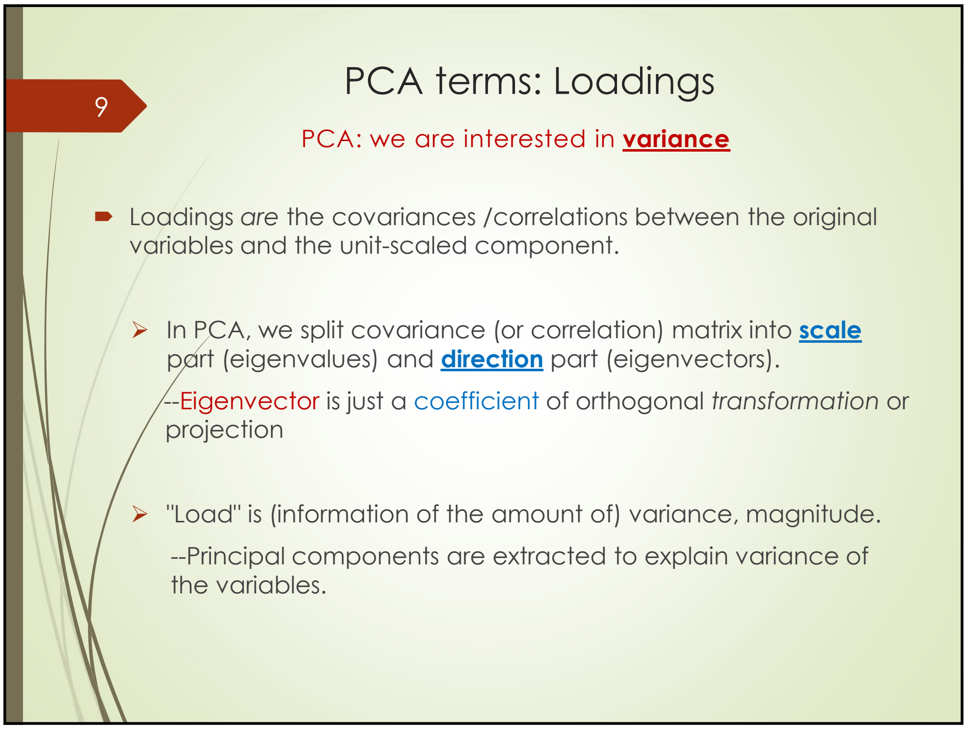
that has the largest variance. By ***normalized***, we mean that $\sum_{j=1}^{p} \phi_{j1}^2 = 1$

- $\varphi_{11}, \ldots, \varphi_{p1}$: the loadings of the first principal component; make up the principal component loading vector,

$$\phi_1 = (\phi_{11}\ \phi_{21}\ \ldots\ \phi_{p1})^T.$$

Why constrain the loadings so that their sum of squares is to one?

-- Because otherwise setting these elements to be arbitrarily large in absolute value could result in an arbitrarily large variance



9

PCA terms: Loadings

PCA: we are interested in **variance**

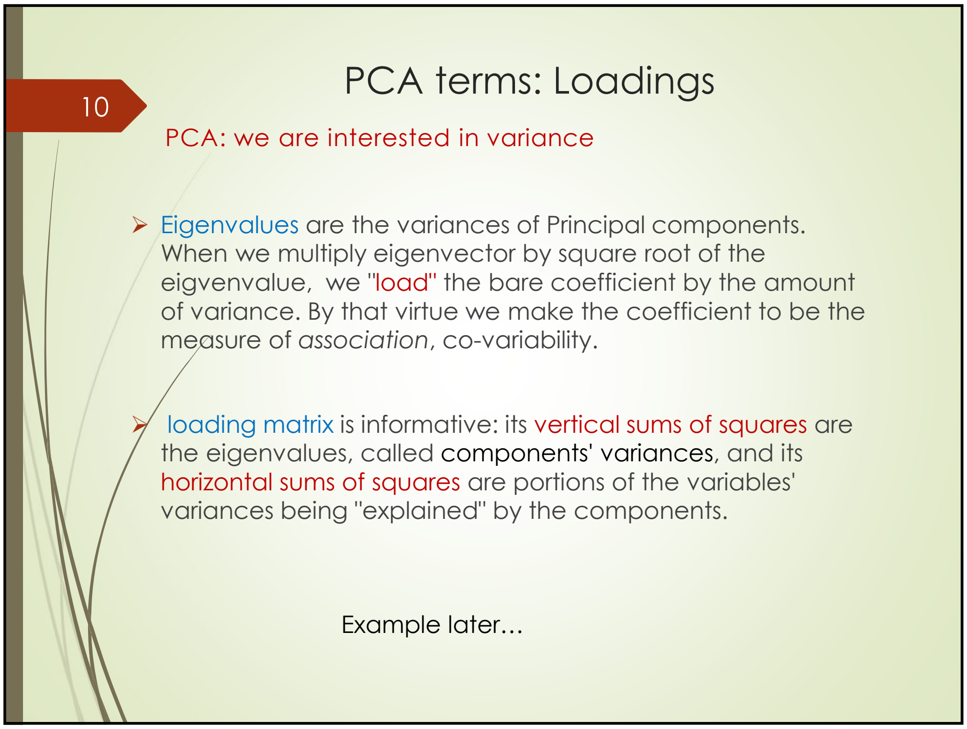
- Loadings *are* the covariances /correlations between the original variables and the unit-scaled component.

  - In PCA, we split covariance (or correlation) matrix into **scale** part (eigenvalues) and **direction** part (eigenvectors).

    --Eigenvector is just a coefficient of orthogonal *transformation* or projection

  - "Load" is (information of the amount of) variance, magnitude.

    --Principal components are extracted to explain variance of the variables.



10

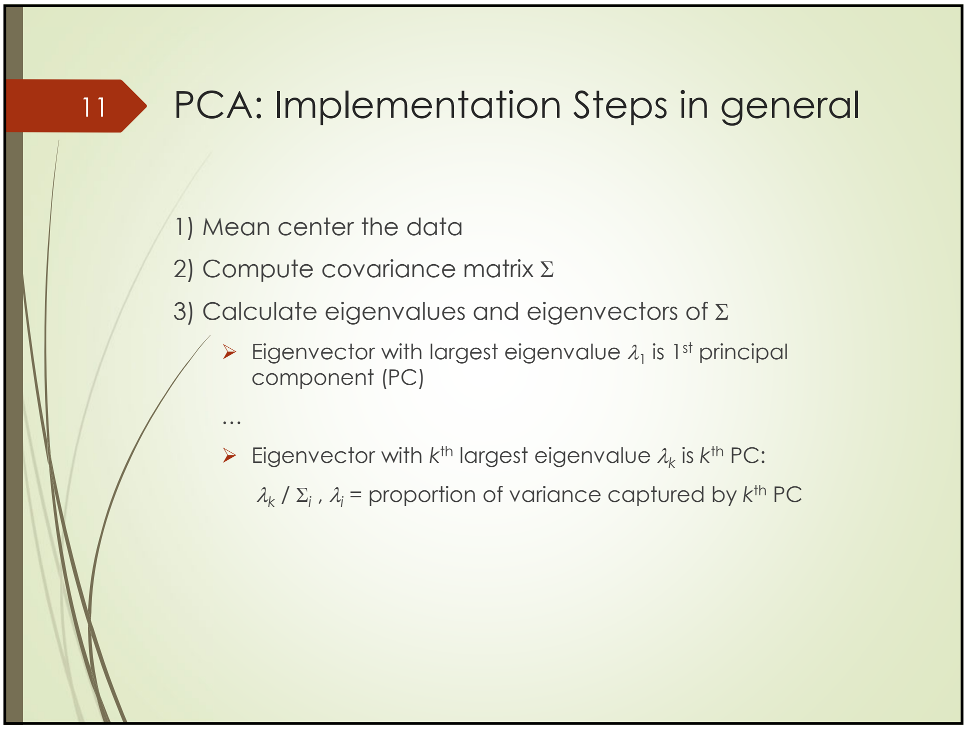
PCA terms: Loadings

PCA: we are interested in variance

- Eigenvalues are the variances of Principal components. When we multiply eigenvector by square root of the eigvenvalue, we "load" the bare coefficient by the amount of variance. By that virtue we make the coefficient to be the measure of *association*, co-variability.

- loading matrix is informative: its vertical sums of squares are the eigenvalues, called components' variances, and its horizontal sums of squares are portions of the variables' variances being "explained" by the components.

Example later…



11

PCA: Implementation Steps in general

1) Mean center the data

2) Compute covariance matrix Σ

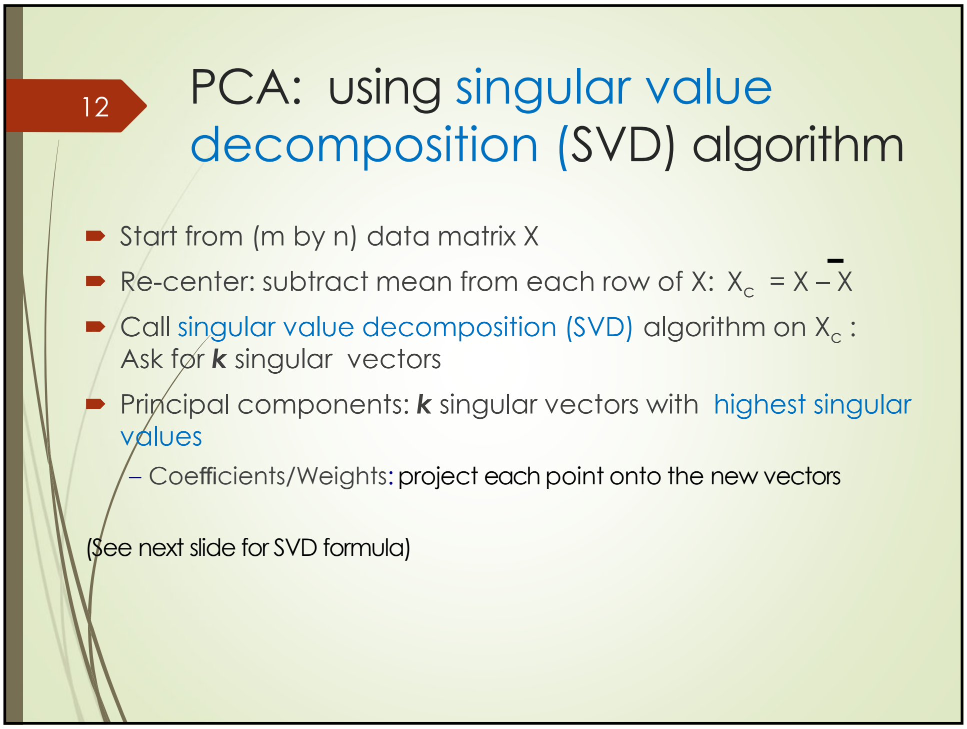
3) Calculate eigenvalues and eigenvectors of Σ

- Eigenvector with largest eigenvalue λ_1 is 1st principal component (PC)

…

- Eigenvector with k^{th} largest eigenvalue λ_k is k^{th} PC:

  $\lambda_k / \Sigma_i , \lambda_i$ = proportion of variance captured by k^{th} PC



12

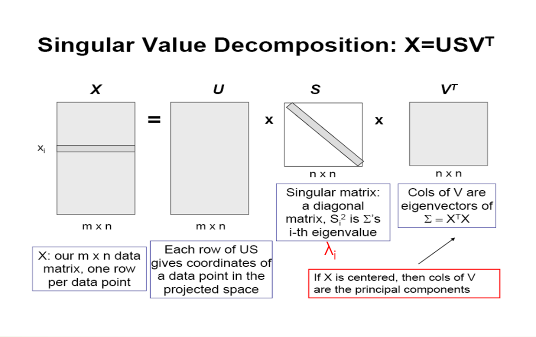
PCA: using singular value decomposition (SVD) algorithm

- Start from (m by n) data matrix X
- Re-center: subtract mean from each row of X: $X_C = X - \bar{X}$
- Call singular value decomposition (SVD) algorithm on X_C : Ask for ***k*** singular vectors
- Principal components: ***k*** singular vectors with highest singular values
  - – Coefficients/Weights: project each point onto the new vectors

(See next slide for SVD formula)

13

Supplementary: singular value decomposition (SVD)



Singular Value Decomposition: X=USV^T

Singular matrix: a diagonal matrix, S_i^2 is Σ's i-th eigenvalue

Cols of V are eigenvectors of $\Sigma = X^TX$

λ_i

X: our m x n data matrix, one row per data point

Each row of US gives coordinates of a data point in the projected space

If X is centered, then cols of V are the principal components

- Write **X=USV^T** or **X=W SV^T**
  - ➢ **X** data matrix, one row per data point
  - ➢ **W/U** weight matrix, one row per data point – coordinate of **x**^{i} in eigenspace, ie, in the projected space.
  - ➢ **S** singular value matrix, a diagonal matrix: eigenvalue λ
  - ➢ **V**^{T} singular vector matrix: eigenvector **v**$_j$

14

PCA: Illustrative example

15

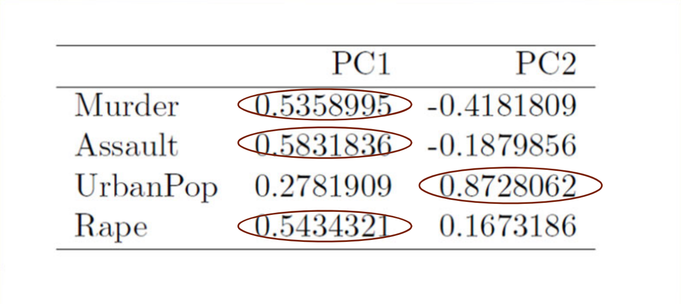
PCA: Illustrative Example

- USAarrests data: For each of the 50 states in the United States, the data set contains the number of arrests per 100, 000 residents for each of three crimes: Assault, Murder and Rape.

- We also record UrbanPop: the percent of the population in each state living in urban areas
  - The principal component score vectors have length $n = 50$, and the principal component loading vectors have length $p = 4$.
  - PCA was performed after standardizing each variable to have mean zero and standard deviation one.

16

PCA Loadings: in class exercise

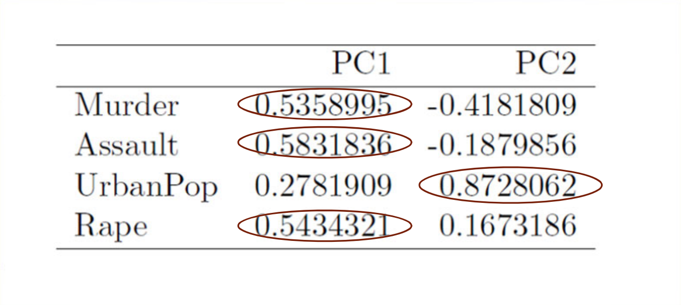


	PC1	PC2
Murder	0.5358995	-0.4181809
Assault	0.5831836	-0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

- The first loading vector places approximately equal weight/coefficient on which variables, with much less weight on which variable?
- How about the 2nd loading vector?

17

PCA Loadings: in class exercise

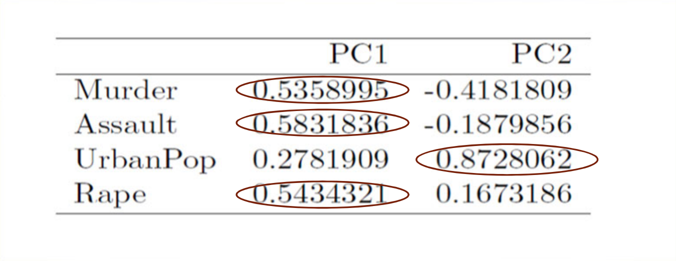


	PC1	PC2
Murder	0.5358995	-0.4181809
Assault	0.5831836	-0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

- The first loading vector places approximately equal weight/coefficient on which variables (answer: Murder, Assault and Rape), with much less weight on which variable (UrbanPop)?
- How about the 2nd loading vector? (Answer: has UrbanPop loaded)

18

PCA Loadings: in class exercise

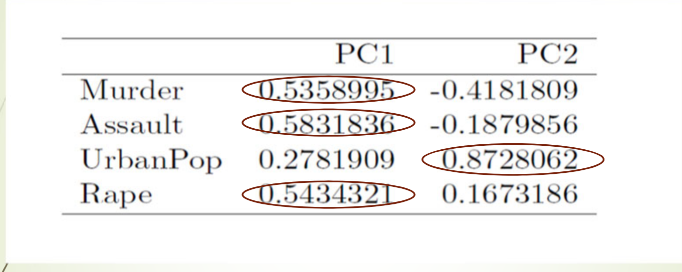


	PC1	PC2
Murder	0.5358995	-0.4181809
Assault	0.5831836	-0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

- What does PC1 measure?
- What does PC2 measure?

19

PCA Loadings: in class exercise



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

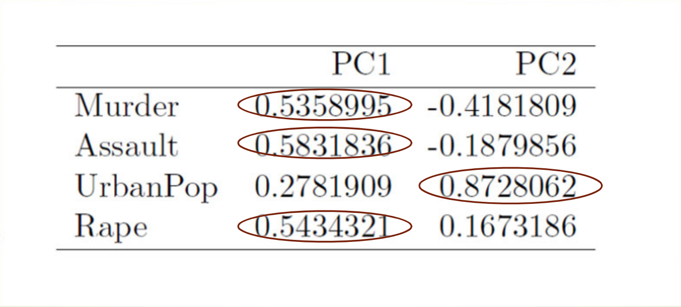
- What does PC1 measure?
  Answer: measure of overall rates of serious crimes; indicates that the crime-related variables are correlated with each other (i.e., states with high murder rates tend to have high assault and rape rates)

- What does PC2 measure?
  PC2 measures the level of urbanization of the state

20

PCA Loadings: in class exercise

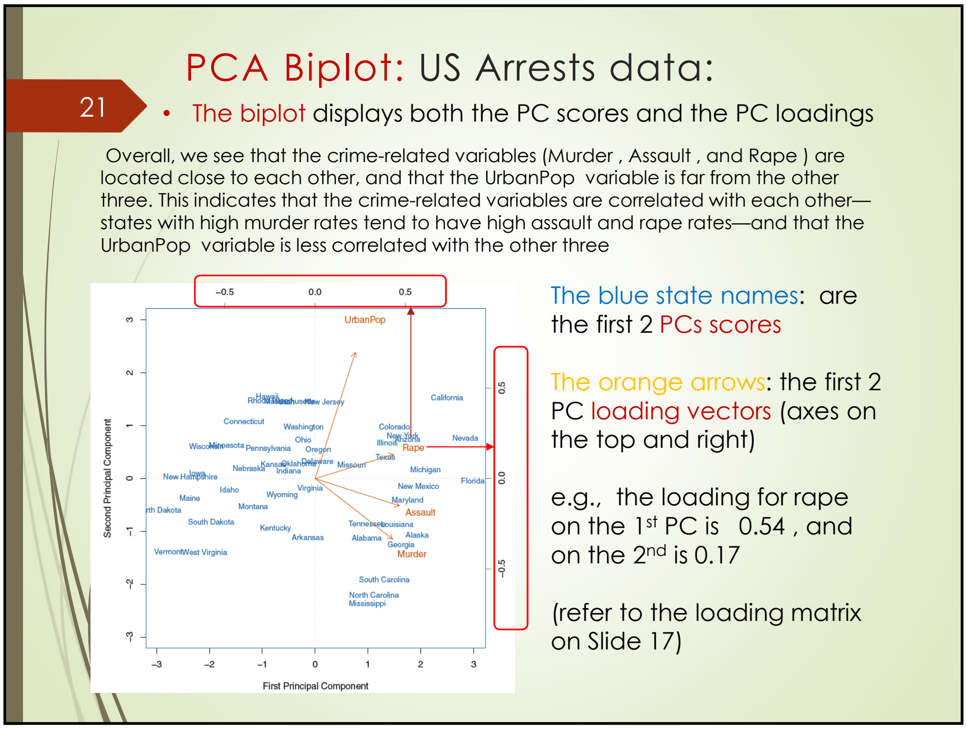
Loading matrix is informative



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

- its vertical sums of squares are the eigenvalues, components' variances being "explained" by the variables.

- its horizontal sums of squares are portions of the variables' variances being "explained" by the components.



21

PCA Biplot: US Arrests data:

- The biplot displays both the PC scores and the PC loadings

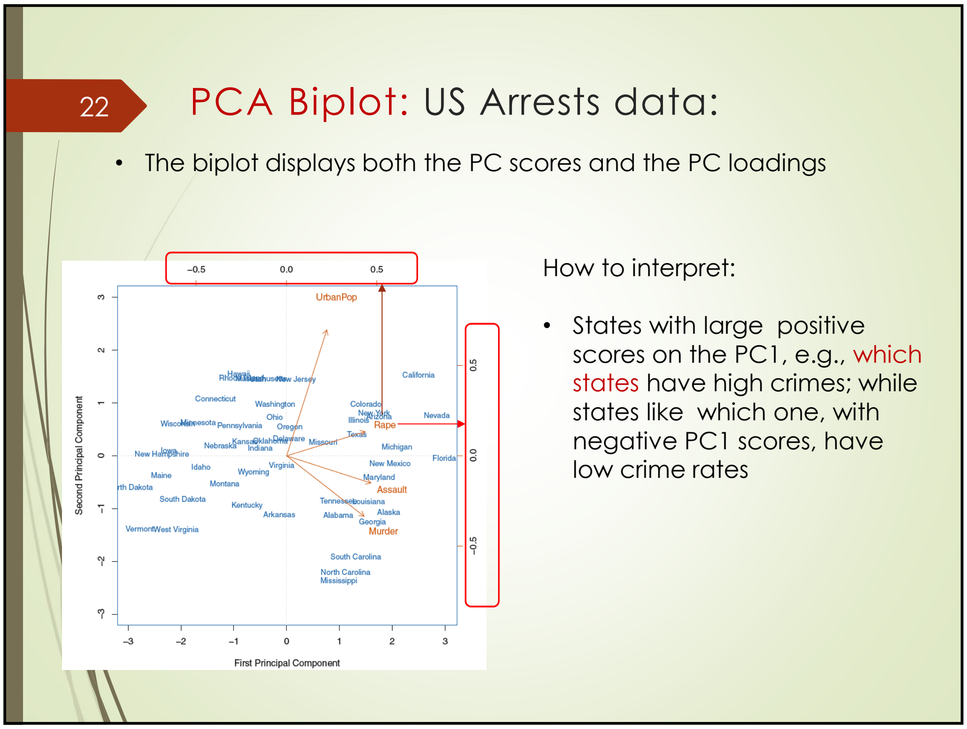
Overall, we see that the crime-related variables (Murder , Assault , and Rape) are located close to each other, and that the UrbanPop variable is far from the other three. This indicates that the crime-related variables are correlated with each other—states with high murder rates tend to have high assault and rape rates—and that the UrbanPop variable is less correlated with the other three

The blue state names: are the first 2 PCs scores

The orange arrows: the first 2 PC loading vectors (axes on the top and right)

e.g., the loading for rape on the 1st PC is 0.54 , and on the 2nd is 0.17

(refer to the loading matrix on Slide 17)



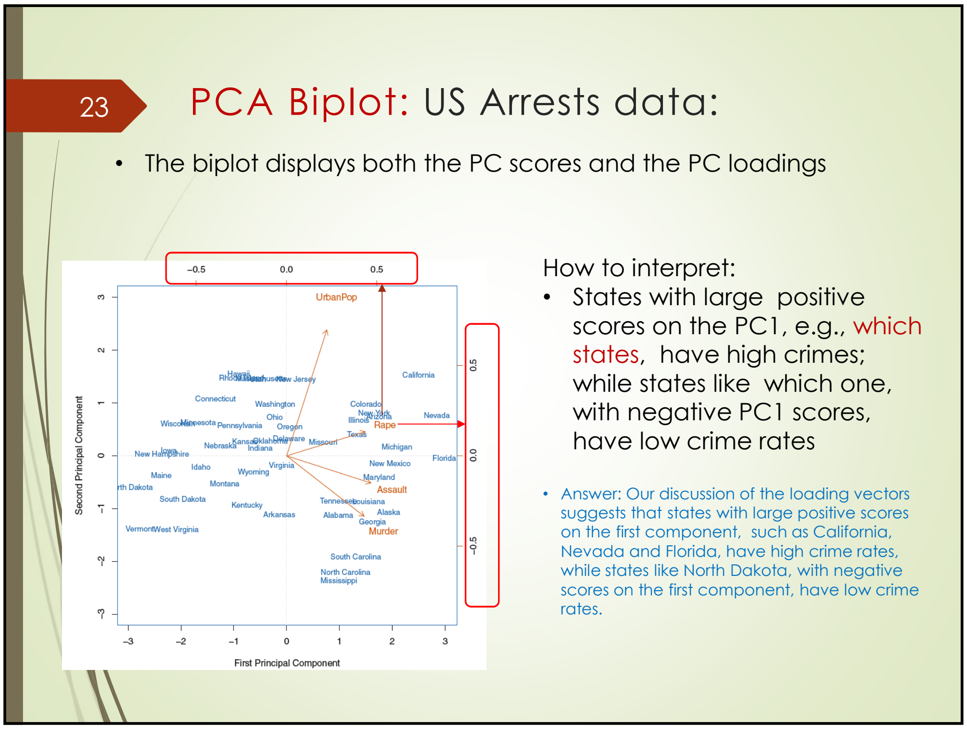
22

PCA Biplot: US Arrests data:

- The biplot displays both the PC scores and the PC loadings

How to interpret:

- States with large positive scores on the PC1, e.g., which states have high crimes; while states like which one, with negative PC1 scores, have low crime rates



23

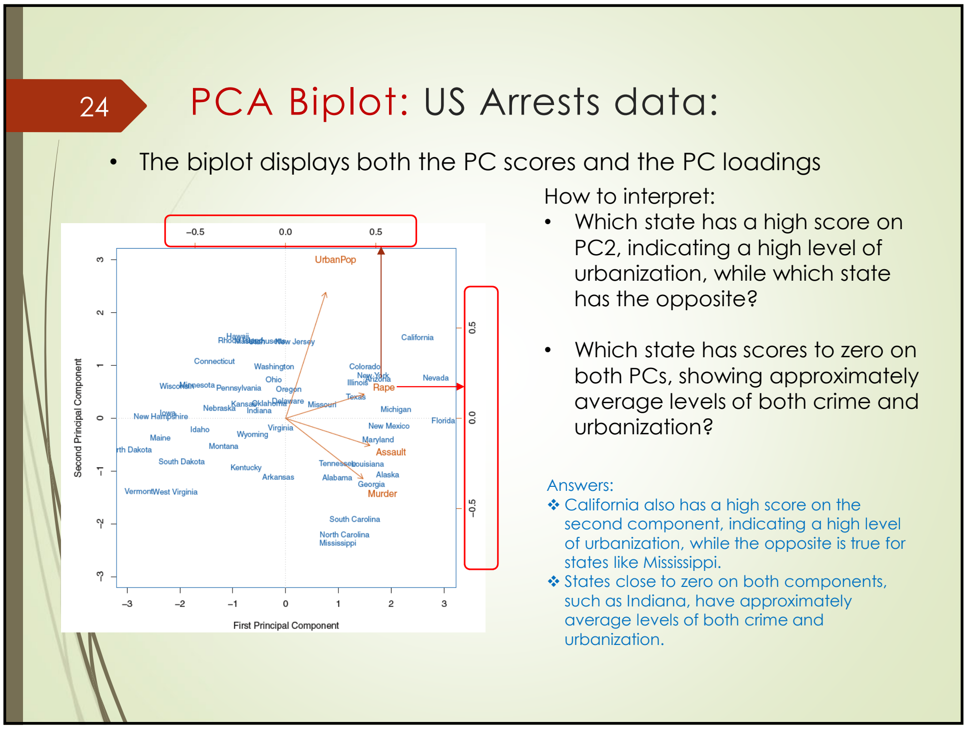
PCA Biplot: US Arrests data:

- The biplot displays both the PC scores and the PC loadings

How to interpret:

- States with large positive scores on the PC1, e.g., which states, have high crimes; while states like which one, with negative PC1 scores, have low crime rates

- Answer: Our discussion of the loading vectors suggests that states with large positive scores on the first component, such as California, Nevada and Florida, have high crime rates, while states like North Dakota, with negative scores on the first component, have low crime rates.



24

PCA Biplot: US Arrests data:

- The biplot displays both the PC scores and the PC loadings

How to interpret:

- Which state has a high score on PC2, indicating a high level of urbanization, while which state has the opposite?

- Which state has scores to zero on both PCs, showing approximately average levels of both crime and urbanization?

Answers:

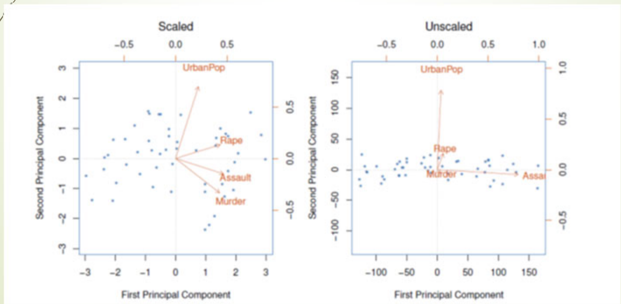
❖ California also has a high score on the second component, indicating a high level of urbanization, while the opposite is true for states like Mississippi.

❖ States close to zero on both components, such as Indiana, have approximately average levels of both crime and urbanization.

25

PCA: Scaling of the variables matters

- If the variables are in different units, scaling each to have standard deviation equal to one is recommended.
- If they are in the same units, you might or might not scale the variables.



26

PCA: Proportion Variance Explained (PVE)

- To understand the strength of each component, compute the proportion of variance explained (PVE) by each one.
- The *total variance* present in a data set (assuming that the variables have been centered to have mean zero) is defined as

$$\sum_{j=1}^{p} \text{Var}(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2,$$

and the variance explained by the m^{th} principal component is

$$\text{Var}(Z_m) = \frac{1}{n} \sum_{i=1}^{n} z_{im}^2.$$

It can be shown that with $M = \min(n - 1, p)$. $\sum_{j=1}^{p} \text{Var}(X_j) = \sum_{m=1}^{M} \text{Var}(Z_m),$

This tells us the variance in data is fully explained by the identified principal components.

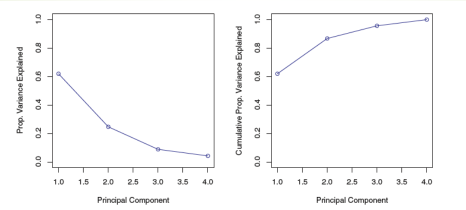
27

PCA: PVE & Scree Plots

- Therefore, the PVE of the ***m*th** principal component is given by the positive quantity between 0 and 1

$$\frac{\sum_{i=1}^{n} z_{im}^2}{\sum_{j=1}^{p}\sum_{i=1}^{n} x_{ij}^2}$$

The PVEs sum to one. Also use the cumulative PVEs



Scree Plots

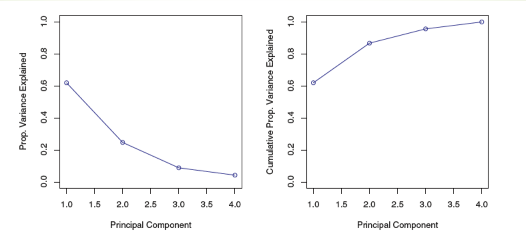
Left: a scree plot depicting the proportion of variance explained by each of the four principal components in the USArrests data.
Right: the cumulative proportion of variance explained by the four principal components in the USArrests data.

28

How many principal components should we use: Scree Plot

If we use principal components as a summary of our data, how many components are sufficient?

-- the "scree plot" can be used as a guide: we look for an "elbow".



29

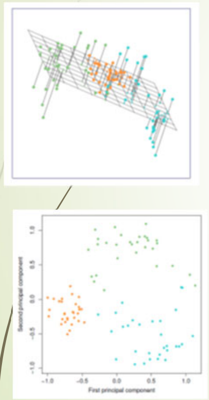
R: Running PCA in R studio.

R: Run Naïve Bayes in Rstudio using USArrests data

Let’s go through the instruction file “R_PCA_S22.docx” posted with LS13 slides at myCourses.

30

Suppl.: Another Interpretation of Principal Components (visualization)



PCA find the hyperplane closest to the observations:

- The first principal component loading vector has a very special property: it defines the line in p-dimensional space that is closest to the n observations (using average squared Euclidean distance as a measure of closeness)
- The notion of principal components as the dimensions that are closest to the n observations extends beyond just the first principal component.

For instance, the first two principal components of a data set span the plane that is closest to the n observations, in terms of average squared Euclidean distance.