LAB #2(A)

Goal: practice the basic visualization tools used in visual analytics
- use data from lab #1 (or other)
- use python for analyses and then upload results to browser app

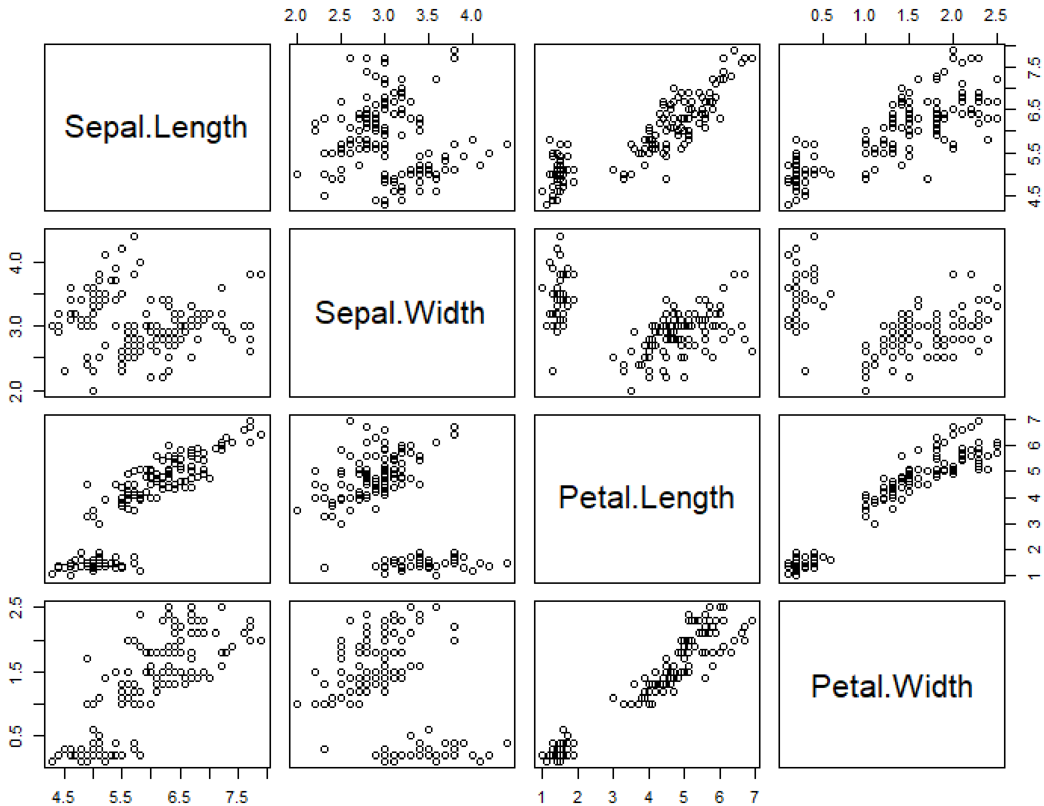
Task 1: basic dimension reduction and data visualization with PCA
- use PCA to compute the Eigenvectors of the data and visualize the Eigenvalues as a scree plot
- add an interaction element into the scree plot that allows the user to mark and select the intrinsic dimensionality index (d_i)
- plot the data into a PCA-based biplot

Task 2: visualize the data with a scatterplot matrix
- use the PCA components $\leq d_i$ to obtain the 4 attributes with the highest squared sum of PCA loadings and list them in a table on the webpage
- use these four attributes and construct a scatterplot matrix

Task 3: use k-means to find clusters and color the points by cluster ID
- use the elbow method to find the best k (visualize the function on k)

What's a Scatterplot Matrix?

All possible bivariate scatterplots arranged into a matrix



A Note on Biplots

Need to know

- How to project N-D points onto a 2D plane defined by two orthogonal N-D axis vectors
- How to project the data axis vectors onto this 2D plane
- The next few slides cover that

Projection Operations

How does 2D projection work in practice?

- N-dimensional point $x = \{x_1. x_2, x_3, \dots x_N\}$
- a basis of two orthogonal axis vectors defined in N-D space

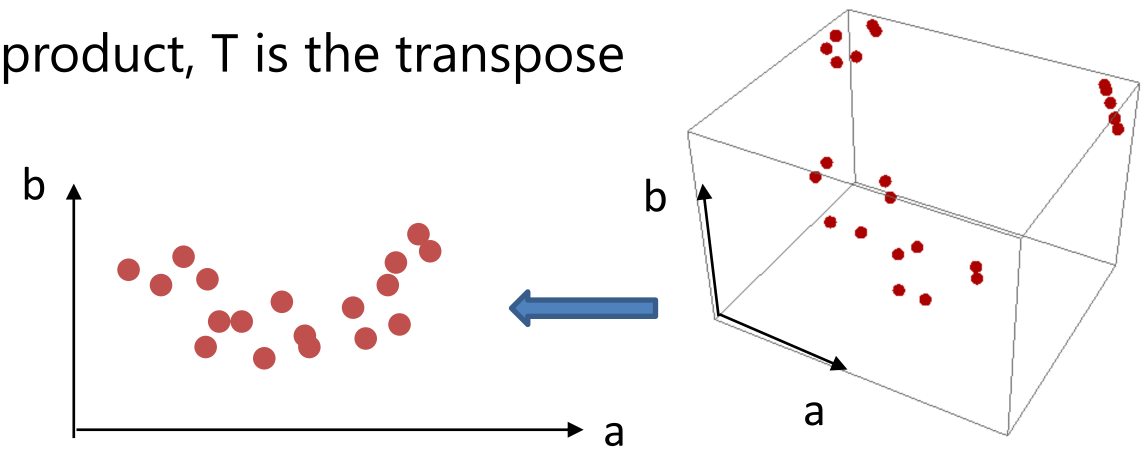
  $a = \{a_1. a_2, a_3, \dots a_N\}$

  $b = \{b_1. b_2, b_3, \dots b_N\}$
- a projection $\{x_a, x_b\}$ of x into the 2D basis spanned by $\{a, b\}$ is:

  $x_a = a \cdot x^T$

  $x_b = b \cdot x^T$

  where $\cdot$ is the dot product, T is the transpose



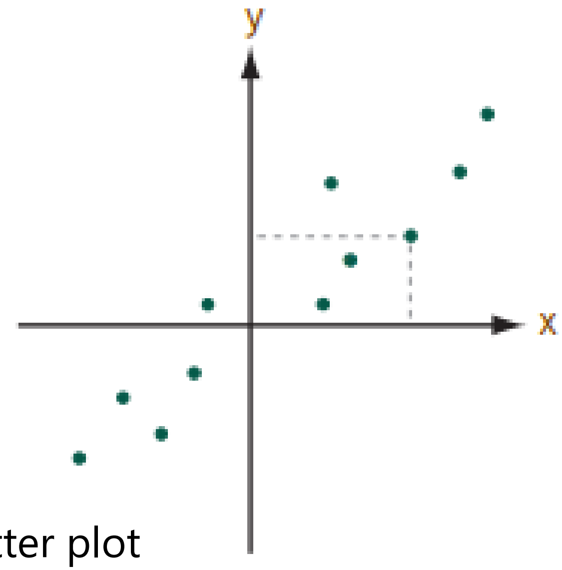
Creating the Biplot

Plots data points and dimension axes into a single visualization

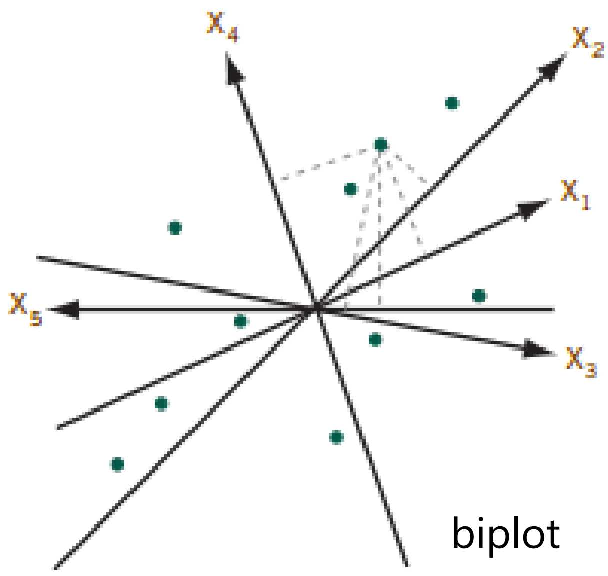
- uses first two PCA vectors as the basis to project into
- find plot coordinates [x] [y]

  for data points: [$PCA_1 \cdot$ data vector] [$PCA_2 \cdot$ data vector]

  for dimension axes: [PCA_1[dimension]] [PCA_2[dimension]]



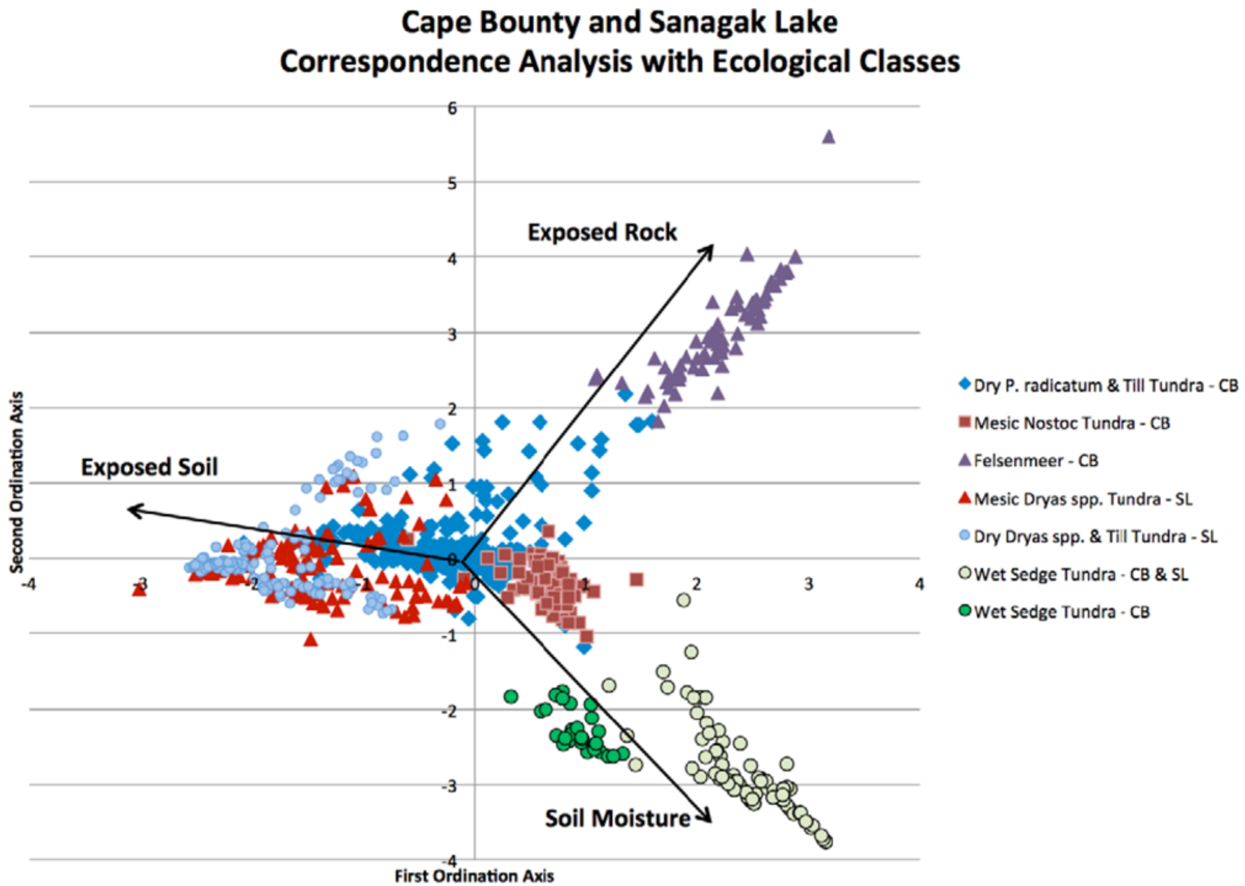
scatter plot



biplot

Biplot Example

They are a linear projection into the 2D PCA-vector basis



Possible Workflow

Upon loading of the dataset, using all numerical attributes

- Compute PCA and obtain Eigenvectors and Eigenvalues
- Compute clusters for k=1…10, and for each k (1) keep the MSE score and (2) store each point's cluster ID into a dedicated column

Construct initial displays

- Display PCA scree plot as a bar chart and set the initial intrinsic dimensionality as the elbow of the plot; color the respective bar
- Display k-means MSE plot as a bar chart; set the initial k as the elbow point and color the respective bar
- Display biplot and color the points by the set initial k
- Display scatterplot matrix where the attributes are chosen using the initial intrinsic dimensionality and color the points according to the initial k

Possible Workflow Continued

Support user interactions

- User can choose a different intrinsic dimensionality in the scree plot by coloring the respective bar
- User can choose a different k in the k-means MSE plot, again by coloring the respective bar

Make observations

- Observe what happens when you make these changes

Additional optional observations

- Observe what happens in the biplot when different PCA vectors are chosen as the projection basis
- You could choose them in the scree plot by coloring the respective two bars or choosing a different visual marking (bold bar outline)
- It will have an effect on the quality of the display (you can keep the k as chosen)

Scoring and Due Dates

Each (task) bullet item carries 10 points

- Adding what is mentioned on the slide "Possible Workflow Continued" earns up to 10 points of extra credit

Don't forget to

- label the axes and tick marks where appropriate
- show color legends where appropriate
- provide a meaningful header on each plot

Due date

- due Sept 28, end of day
- no group project, single only

Deliverables

Submit on Brightspace

- voice-narrated video file to show all features of your software in action
- in the video discuss any interesting observations you were able to make in the data
- report
  - describe interesting observations (beyond the video)
  - mention anything noteworthy about implementation (beyond the video)
- zip file with complete source code as well as the data
- submit video as an extra file

Grading

Grading

- TA will pick students at random for thorough code review sessions
- you better know your code !!!
- so, please do not just copy code beyond the D3 templates
- or even worse, videotape someone else's program