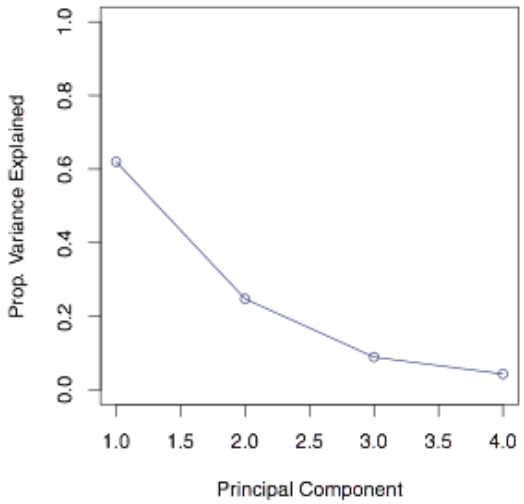
Cogs 109: Modeling and Data Analysis

Homework 7

Due **Thursday 11/29**

1. Two students each run PCA on the same dataset. The dataset has n=100 observations, with p=4 predictors.
    a. How many principal components (PCs) are there?
    b. Do you expect that the two students will arrive at the same result (i.e. the same PC coefficients, scores and variances)?
    c. Consider the scree plot (right). Based on this plot, how many PCs should you consider if you want to capture >60% of the total data variance?
    d. How many PCs do you need to capture >80% of the variance?



2. **Extra credit:** Consider a dataset with n=100 observations and p=30 predictors. What is the minimum fraction of the total data variance captured by the first PC?
3. Two of your fellow students each run k-means clustering on the same dataset. They both choose k=4.
    a. (2 points) Do you expect that they will both come up with the same clustering? Why or why not?
    b. (2 points) In your own words, define a *local optimum* and a *global optimum* of an objective function. Which of these two best describes the result of k-means clustering?
    c. (1 point) Name one strategy the students could use to reduce the random variance in their cluster results.
4. ISLR problem 10.3 (page 414).
5. **Hierarchical clustering.** Using the same dataset as in problem ISLR 10.3 (6 observations, 2 predictors), perform hierarchical clustering.
    a. (1 point) First use single-linkage clustering and plot the resulting dendrogram.
    b. (1 point) Plot the dendrogram using complete-linkage clustering.
    c. (2 points) Do these results generally agree with each other and with the results of k-means clustering? Why or why not?

Code hints for MATLAB: `kmeans, randi, linkage, dendrogram`

Python: `sklearn.cluster.KMeans`, `scipy.cluster.hierarchy.linkage`, `scipy.cluster.hierarchy.dendrogram`