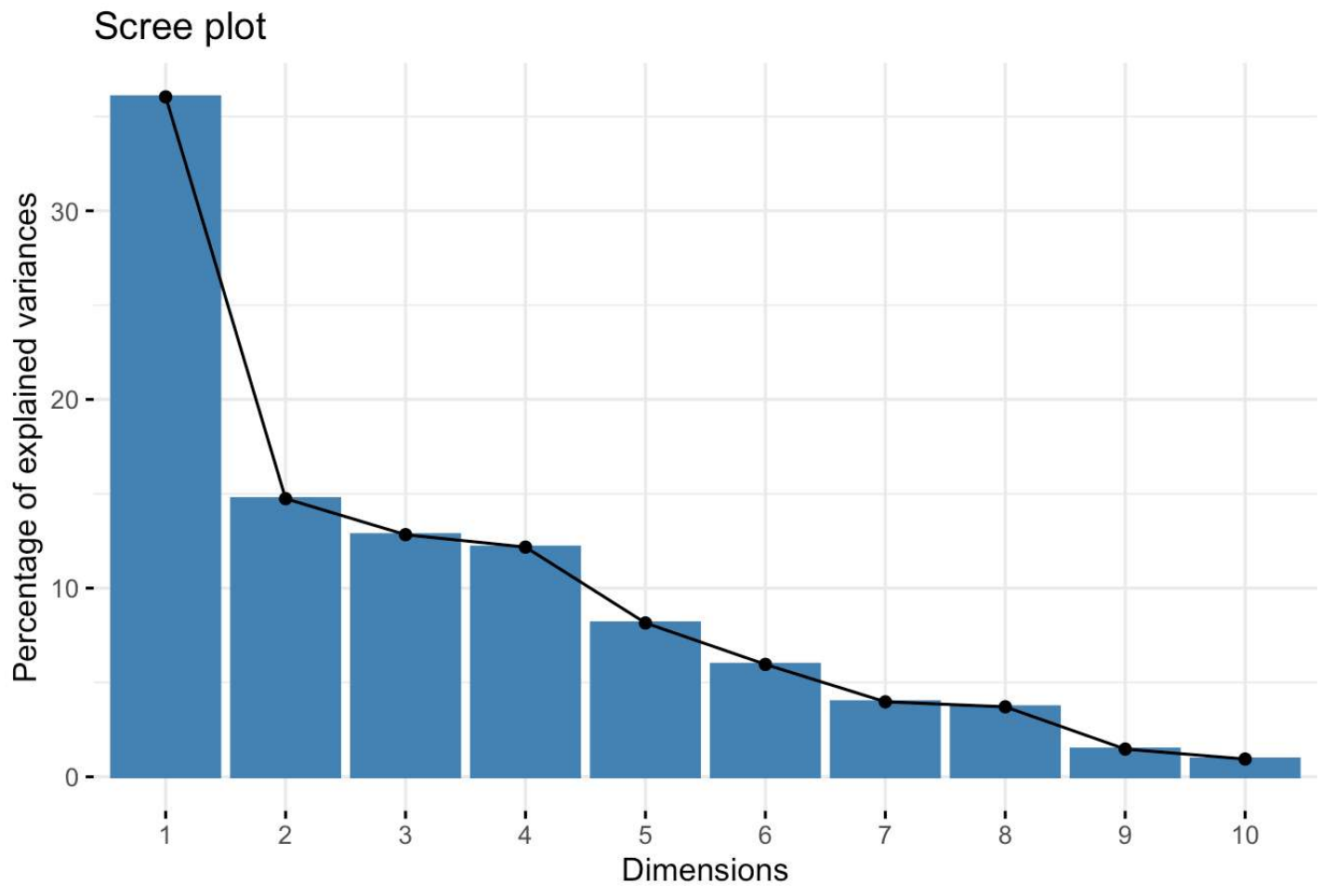
# PCA and Classification Solutions

1. (0 points) Please download the link file **trainingData.csv (https://northeastern.instructure.com/courses/193161/files/30491775?wrap=1)** ↓ **(https://northeastern.instructure.com/courses/193161/files/30491775/download?download_frd=1)** to your IDE. This data set contains 10-D training data for class 1 and class 2. Your tasks will include performing dimensionality reduction, classifier design and analysis.

2. (5 points) Our first task is to reduce the dimension of the dataset, if possible. Here, we use the notion of preserving scatter in doing so. Our hope is that MOST scatter is due to class differences and PCA will retain that. We begin by ignoring class labels, and find the per-column means and variances. Center each column by its mean and divide by the square root of its variance, in that order. Since scatter is sensitive to units of measurement, this step removes that sensitivity. You can find routines to do this automatically in your IDE by searching for "scaling".

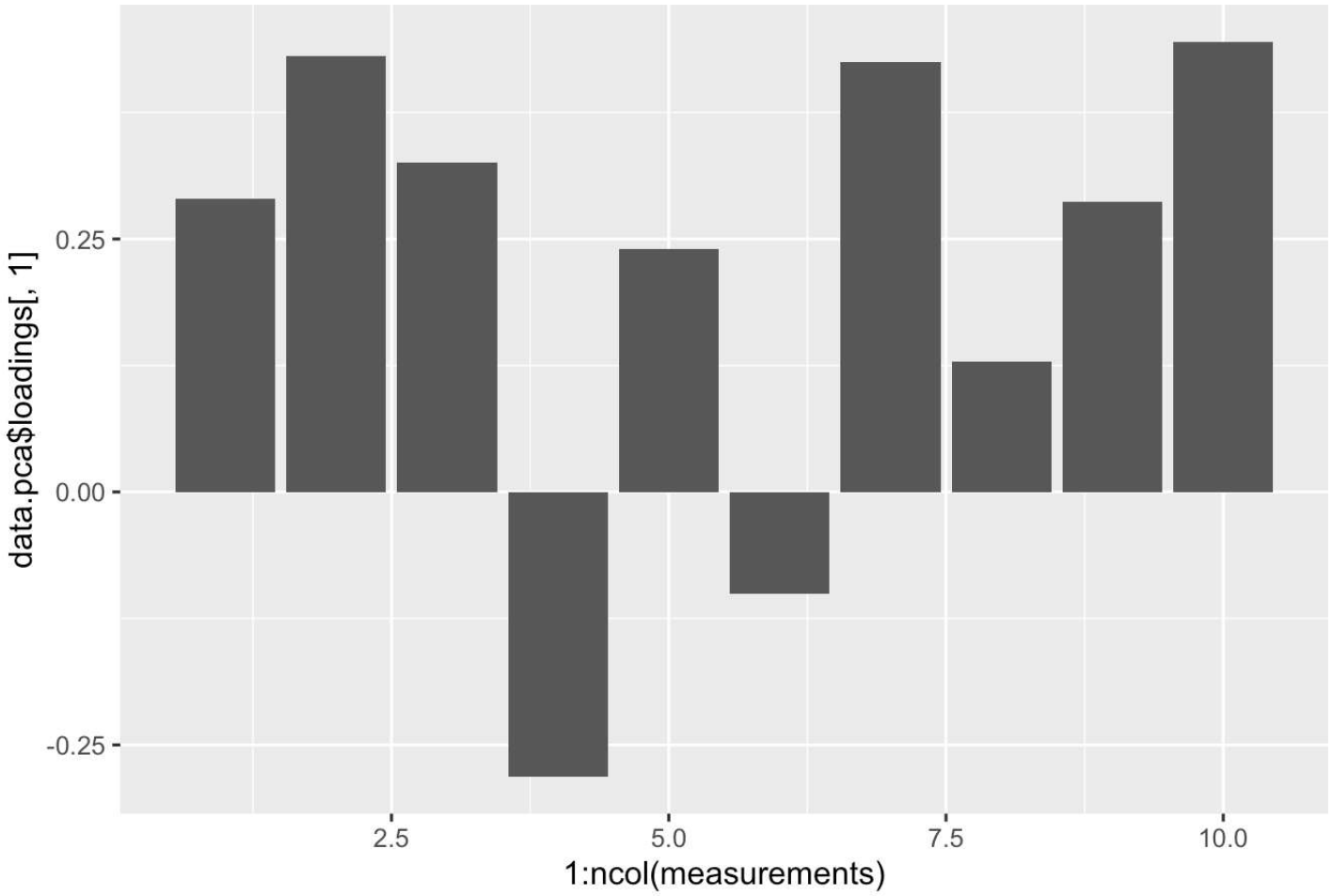
3. (15 points) Perform PCA on the scaled dataset, again ignoring class labels, and produce a scree plot. PCA produces a new set of 10 axes, ordered by the amount of scatter. What percentage of scatter is contained in the first principal component? The first two principal components? Interpret your results.



Roughly 36% of total scatter is retained in the first dimension, and just over 50% of scatter is retained in the first two dimensions? If the scatter were evenly distributed in all directions, then each dimension would have 10% of the scatter. The larger amounts of scatter indicate that the 10-D cloud of points is not "round", but "deflated" and PCA has put the flat dimensions last.

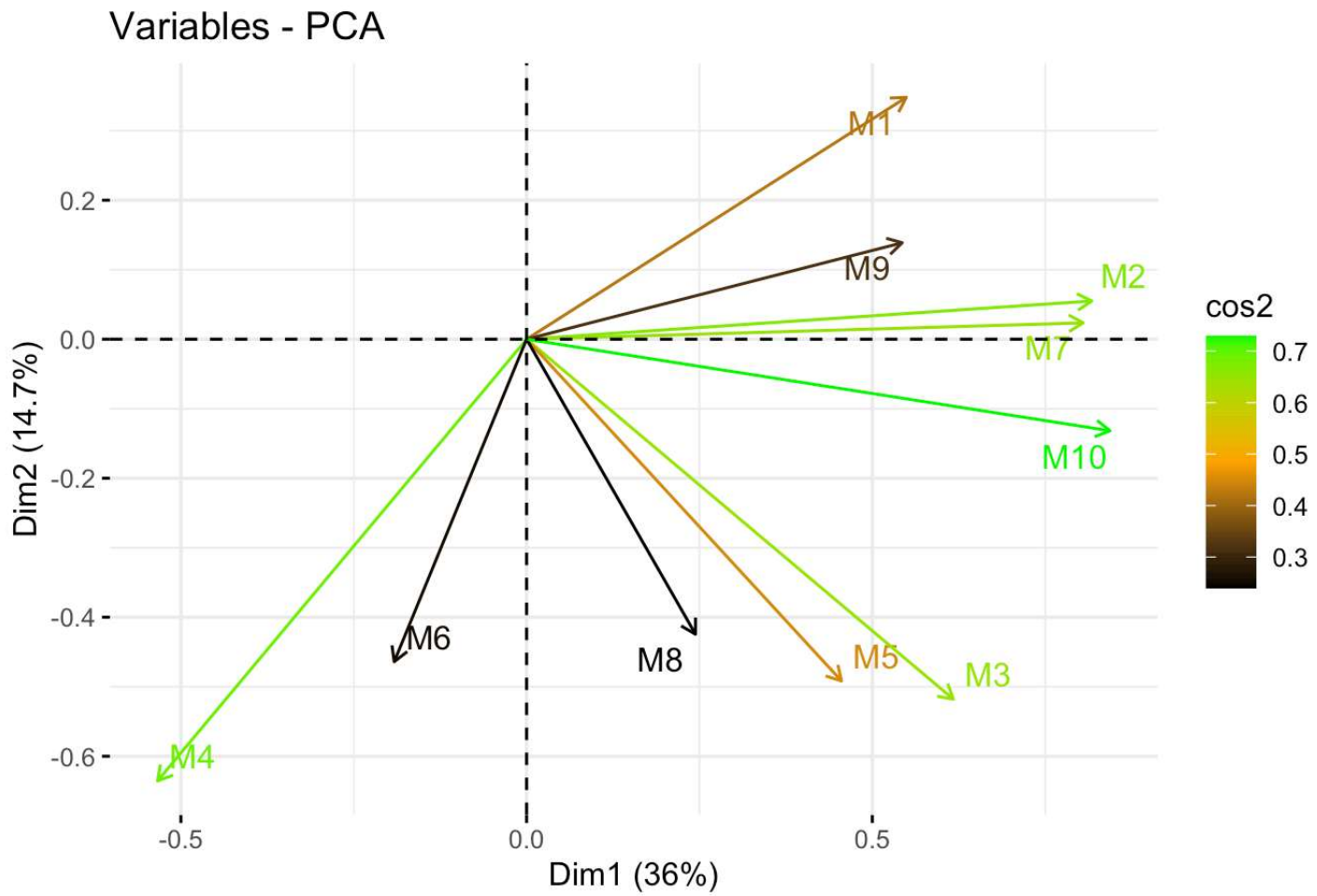
4. (10 points) PCA produces the principal components as a part of its output. Sometimes, these are called *loadings*. Produce a bar-chart of the first principal component elements (10 of them). Which components are negative and dominant? Which are positive and dominant?

You should either get this graph, or its inverted version:




Dimensions 4 and 6 are negative and dominant. Dimensions 2, 7 and 10 are positive and dominant. (If you plot is the inverse, then swap "positive" and "negative".

5. (15 points) A biplot is a useful visualization of the PCA output. It displays the original, scaled measurements M1,...M10 as 2-D vectors with axes given by the first two principal components. Produce a biplot of your PCA output. Which measurements have vector close in angle with each other? Produce 4 such groups. The original scaled measurements within each group are positively correlated. (Positively correlated = measurements tend to go up together and down together.)




Group: M2, M7, and M10

Group: M1 and M9

Group: M4 and M6

Group M3, M5, and M8

6. (10 points) The length of each vector in the biplot help us understand its role in retaining scatter. Original, scaled measurements with shortest vectors play the least role in retaining scatter. The longest vectors indicate measurements that play the greatest role in retaining scatter. Find the labels of 2 original, scales measurements which had the least scatter. Find the labels of the 2 original, scaled measurements which had the most scatter.

Least scatter: pick 2 of M6, M8, M9

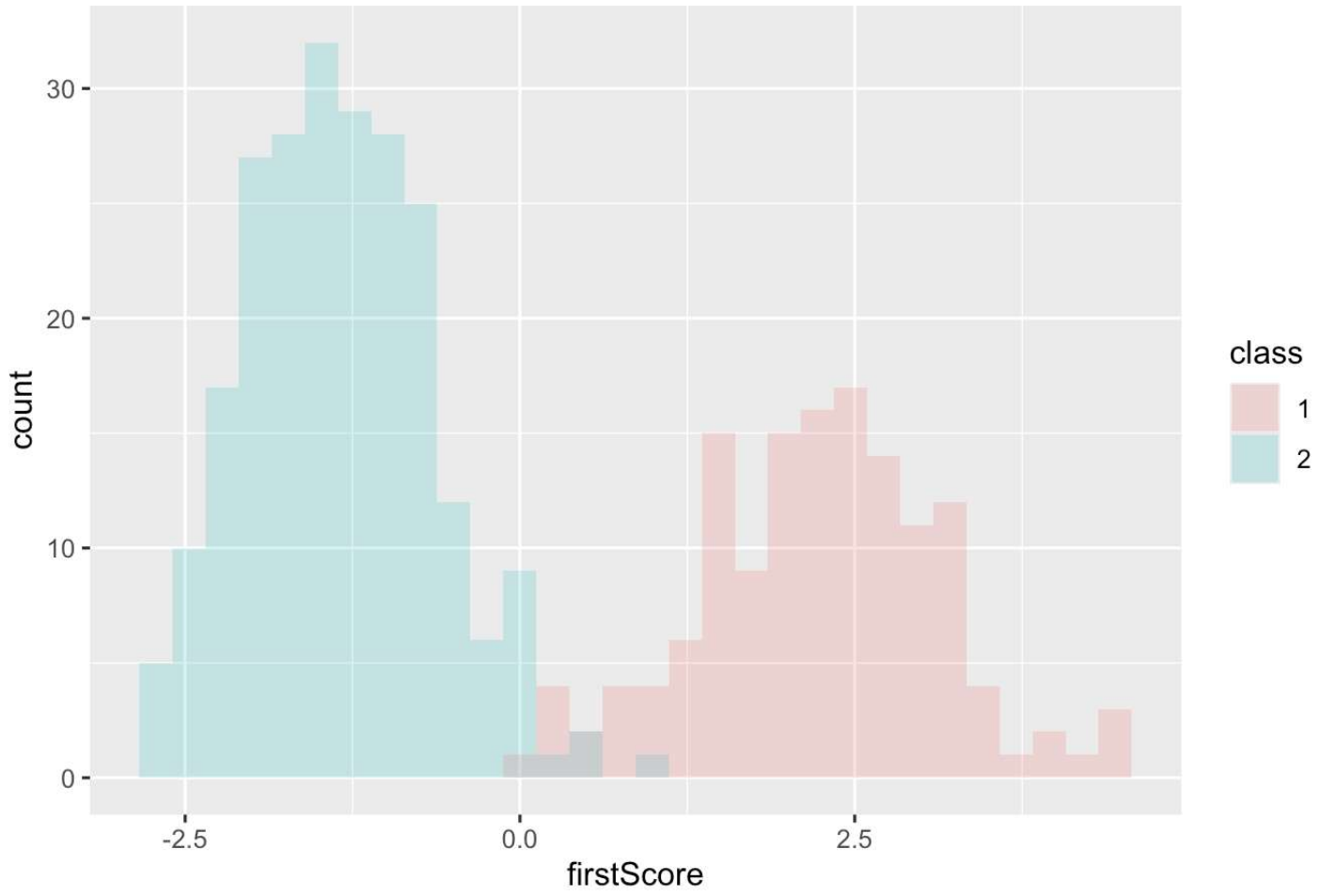
Most scatter: pick 2 of M2, M3, M10

7. (10 points) Two opposing vectors in the biplot indicate original, scaled measurements which are negatively correlated. What is the measurement most negatively correlated with M4?

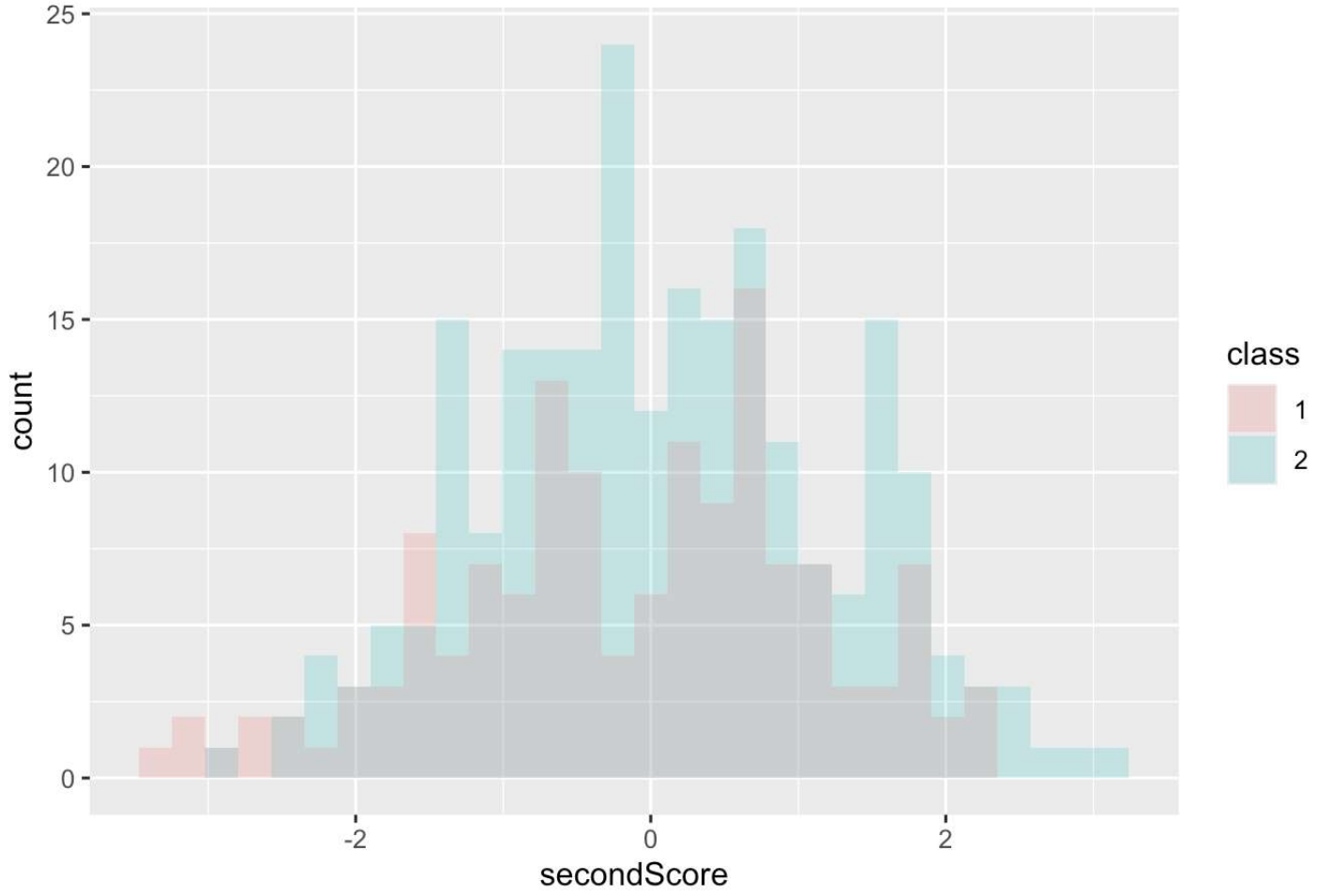
M1

8. (0 point) Another useful output of PCA are *scores*. Scores are the new coordinates of each original, scaled measurement in the new coordinate system (the principal components). For example, the first score for each measurement represents the coordinate along the first principal component axis. Attach two new columns to the data frame loaded in 1., called "score1" and "score2". Our plan is to develop classifiers which only use these scores, not the original data, so what follows may seem familiar.

9. (20 points) One the the benefits of the scores is that they are *uncorrelated*. For Gaussian scores, this indicates *independence*. Plot the two histograms of score1 on the same graph, one for each class. Are those histograms bell-shaped, approximately? The amount of overlap between class histograms indicates the difficulty of good class decisions. How is the overlap of these histograms? Does the first score retain the class differences in the original measurerments? Perform the same for score 2, and answer the above questions again.

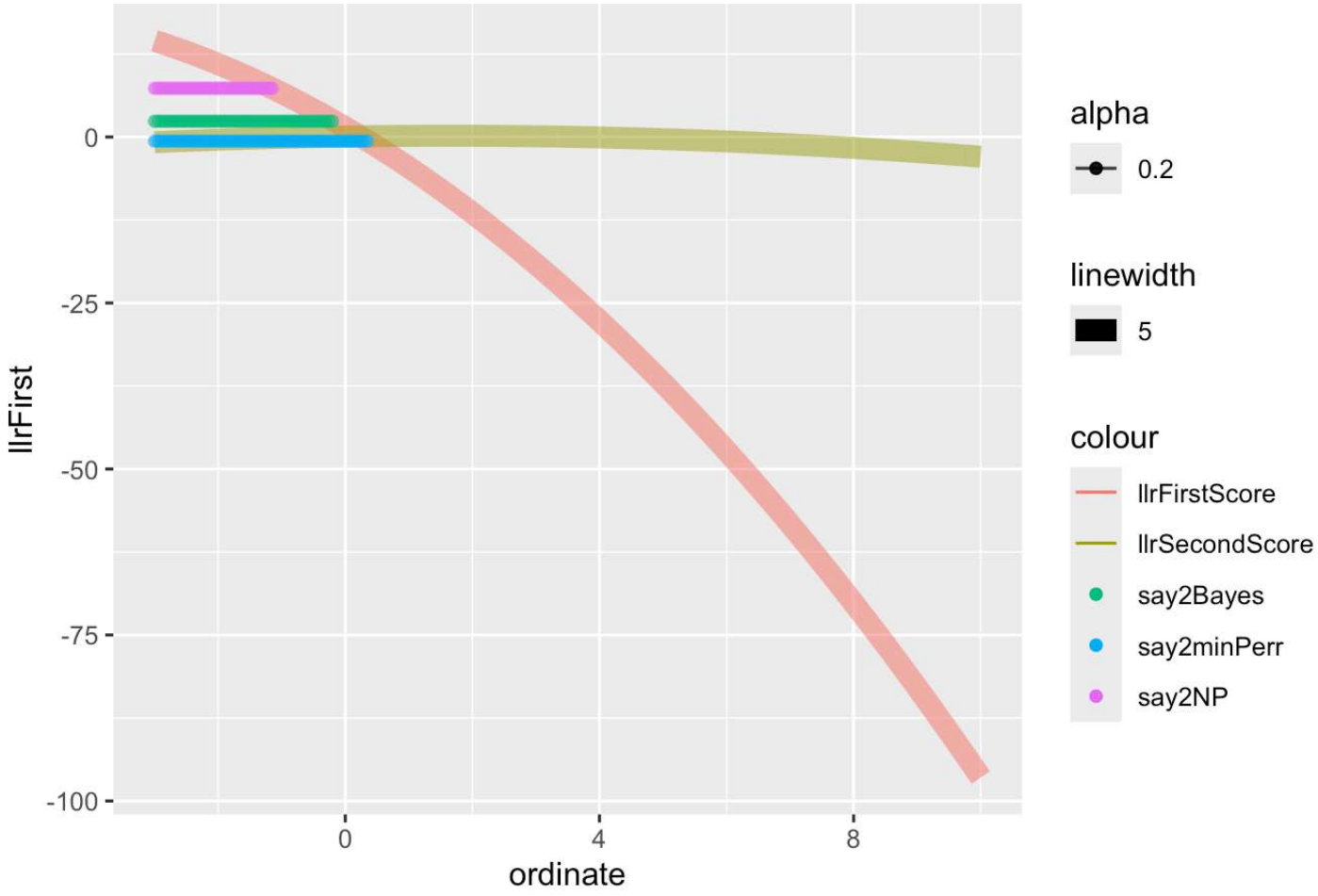



Score 1: both class histograms are bell-shaped, and the overlap is small. Score 1 retains alot of the class differences.




Score 2: both class histograms are bell-shaped, and the amount of overlap is huge. Class differences are not retained in score 2.

10. (15 points) Using only score 1, design the following classifiers for a single measurement: minimum risk, minimum probability of error, and Neyman-Pearson. Please make use of the following data: P{class 1}=0.35, cost of saying class 1 when wrong = 1, cost of saying 2 when wrong = 20, false alarm probability <= 1/10000. Plot the log likelihood ratio versus a range of scores. On the same graph provide the log-likelihood ratio of the second score (even through we do not use it here.) On the same graph, indicate the decision region for class 2 by horizontal lines at the log threshold heights. Provide the risk of the minimum risk test, the probabilty of error of the minimum probability of error test, and false alarm probability and the power of the NP test.




Bayes risk = 0.125

Minimum probability of error = 0.03

false alarm probabililty = 8.9e-4 < 1e-5

power = 0.62