

MAST90138 Week 4 Lab

Problems:

The iris data contain various measurements (sepal length, sepal width, petal length and petal width) of 50 flowers from each of 3 species of iris flowers. Type `help(iris)` to learn about the format of these data.

1. Load the `iris` data in R (they are already in R).
2. Print on the screen the name of each column of this data set.
3. Draw a scatterplot of each pair of numerical variable of this dataset.
4. Draw a scatterplot of each pair of numerical variable of this dataset, this time using a different colour per species (in other words, use 3 colours, one per species). Add a title to the main graph, well above the tickmarks of the pairwise plots. Do the three species appear to be well separated in these graphs?

Note, using `main` to put the title produces the title well above the tickmarks, which is not the case if you use `title(Iris data')`.

5. Compute the eigenvectors and eigenvalues of the covariance matrix of these data (using the four numerical variables) and store them in, respectively, a matrix and a vector called `G` and `lambda`.
6. Compute the four principal components of these data (using the four numerical variables). What is the variance of each of the four PCs? Which fraction of the sum of the variances of the four components of the original data do each of the PCs explain? What does this suggest?
7. Plot the first PCs along a horizontal line, using a different colour for each species. Label by PC1 the x axis and remove entirely the second axis from the figure (keep only the vertical lines but remove all tickmarks etc). Does the first PC separate well all 3 species?
8. Do a scatterplot of the first two PCs, using a different colour for each species and putting x and y labels to the axes (call them PC1 and PC2). Do the first two PCs together separate well all 3 species?