STA 437/1005, Fall 2009 — Assignment #3

Due at the start of the lecture on November 30. (Note that due to lack of time, I've had to abandon the idea of an assignment with a two-part solution/critique form that I discussed earlier.)

Please hand this assignment in on 8 1/2 by 11 inch paper, stapled in the upper-left corner, without any folder or other packaging around it.

This assignment is worth 12% of the course grade. It is to be done by each student individually. You may discuss this assignment in general terms with other students, but the work you hand in should be your own. In particular, you should not leave any discussion of this assignment with any written notes or other recordings, nor receive any written or other material from anyone else by other means such as email.

For this assignment, you will look again at the two data sets you used for the second assignment. This time, you will apply principal component analysis (PCA) and factor analysis (FA) to the data, comparing the results you obtain with these methods, and seeing the effect of reducing dimensionality on use of the data for regression or classification. Another purpose of the assignment is to give general insight into properties of PCA and FA.

This handout, the data sets, some hints about useful R commands, and a solution to the second assignment are (or will shortly be) available from the course web page, at `http://www.utstat.utoronto.ca/~radford/sta437/`

Data set 1:

I have provided a version of this data set with nine observations deleted, because they have values for one or more of the variables that are of doubtful accuracy. This version also omits the "density" variable (retaining "pcfat") and the "age" variable. You should use this version for this assignment, and not remove any additional observations as outliers. You should read this data with the "head=TRUE" option. The data frame will contain column names that are the indexes of the observations in the original data set (some indexes are therefore skipped).

You should perform PCA using only the 12 variables other than "pcfat". You should try PCA with and without scaling these variables to all have standard deviation one, and discuss whether scaling or not scaling (or something else) seems most appropriate. You should look at scree plots of the variances in successive principal components, and on this basis comment on how many components should (perhaps) be used.

You should also perform factor analysis on this data, using the 12 variables other than "pcfat". You should try models with one and with two factors.

You should try to interpret the coefficients found by FA and PCA in terms of the meaning of the variables, using your common sense knowledge of the variability of human body proportions.

You should compare the two-factor model with the first two principal components (both with scaling and without scaling). For this, you should look at the coefficients of the linear combinations of the original 12 variables (after centering) that are used to project onto the components found with PCA, and the linear combinations that are used to predict the values of the factors with the FA model. You should take into account the non-uniqueness of the FA solution when doing this, as well as the non-uniqueness of the sign for the principal components.

You should also look at predicting "pcfat" from the other 12 variables, and compare linear regression on all 12 variables with linear regression on a smaller number of variables found using PCA (projections on the components) and with a smaller number of variables found using FA (from the regression estimates of the unobserved factors). You can judge how well "pcfat" can be predicted using the adjusted R-squared value output by the "summary" command applied to the output of "lm". You can also consider adding BMI (weight divided by height squared) as an additional predictor in the regression, and see whether this helps.

Data set 2:

For this data set, you should use the same data file as for the second assignment. It should be read with the "head=TRUE" option. You should not remove any observations as outliers.

You should perform PCA using observations in all classes, and using the 36 variables in this data set with the class variable omitted. You should try PCA with and without scaling these variables to all have standard deviation one, and discuss whether scaling or not scaling (or something else) seems most appropriate. You should also look at scree plots of the variances in successive principal component directions, and on this basis comment on how many components should (perhaps) be used.

You should also perform factor analysis on this data (again, for all classes, 36 variables), using two factors. You should compare the results of PCA and FA. For this, you should look at the coefficients of the linear combinations of the original 36 variables (after centering) that are used to project onto the components found with PCA, and the linear combinations that are used to predict the values of the factors with the FA model. You should take into account the non-uniqueness of the FA solution when doing this, as well as the non-uniqueness of the sign for the principal components.

You should try to interpret the coefficients found by FA and PCA (including how many components seem to be needed) in terms of the meaning of the 36 variables, as observations on 4 spectral bands for 9 pixels, and of the context that observations come from three classes.

You should also look at how effective reducing the 36 variables to only 2 variable with PCA or FA would be if the goal is to classify future observations into one of these three classes. You can do this informally by just looking at scatterplots with the class indicated by colour.