ETC3250 Lab 6

Di Cook

1 September 2015

Principal component analysis, and dimension reduction: SOLUTION

We will run PCA on the multiple test scores for Australian 15 year olds PISA test scores.

Assignment 6

Turn in two items: a .Rmd document, and the output .pdf from running it. No need to include the R output and plots in your pdf, but the code should be in the Rmd file.

Task 1

Read in the PISA data. How many students were tested? How many variables are included in the data set? Read the data dictionary to find out what the variables named ST08Q01 PV1MACC PV2MACC PV3MACC PV4MACC PV5MACC PV1MACQ PV2MACQ PV3MACQ PV4MACQ PV5MACQ PV1MACS PV2MACS PV3MACS PV4MACS PV5MACS PV1MACU PV2MACU PV3MACU PV4MACU PV5MACU PV1MAPE PV2MAPE PV3MAPE PV4MAPE PV5MAPE PV1MAPF PV2MAPF PV3MAPF PV4MAPF PV5MAPF PV1MAPI PV2MAPI PV3MAPI PV4MAPI PV5MAPI are. Write a couple of sentences describing them.

14,481 Australian students were tested in 2012. This data set has 80 variables.

ST08Q01 is the gender of the student

PV1MACC-PV5MACC are measuring understanding of change and relationships, PV1MACQ-PV5MACQ measure understanding od quantity, PV1MACS-PV5MACS, measure space and shape, PV1MACU-PV5MACU measure uncertainty and data, perhaps the closest to statistics, PV1MAPE-PV5MAPE measure employ, which we would guess to be run the ideas, PV1MAPF-PV5MAPF cover formulating the problems, and PV1MAPI-PV5MAPI tests interpretative skills.

Task 2

Compute a PCA on the variables PV1MACC through PV5MAPI. Make a scree plot, and examine the principal components for the first 4. What proportion of variation in the data is explained by the first principal component? Second, third and fourth?

The proportion of variation explained by PC1 is 0.88, and 0.02, 0.01 and 0.01, *for the second, third and fourth respectively.*

Task 3

Compute the average for each student for each of the different types of math tasks. Based on the PCA explain why this would be a reasonable thing to do. Make a scatterplot matrix of the average scores.

The main source of variation in the data is the sum of all types of tests. The second and third principal components suggest the secondary source of variation is in the type of math test, because the coefficients break

up into groups of 5 that match the different tests. This tells us that the test scores are pretty similar for type of test, and differ more across test types, so it is reasonable to average the scores for the type of test.

Task 4

Compute the average overall math score for each student (this means averaging PV1MATH-PV5MATH). Make a side-by-side boxplot of these scores by gender. Is there a difference in math scores for girls and boys? Write a few sentences explaining what you learn. (Note that the full range of math scores is 0-1000.)

The median math score for girls is a little lower than for boys, and the spread is a little smaller. The boys scores may be considered roughly to be shifted up by about 10 points from the girls scores. There is little difference in the two distributions because 10 points out of 1000 is very small. The top score was earned by a boy, but the second top by a girl, and the lowest score was earned by a boy.

Task 5

How many different schools were included in the survey? Compute the average math score (average the averages) and standard deviation for each school, and make an ordered dotplot (with bars indicating one standard deviation above and below the mean) of these averages. Write a couple of sentences that describe how math scores vary across schools.

There is a difference of about 600 points from the average of the top school to that of the bottom. The standard deviations are reasonably similar, on the order of 100 points. The data suggests that the school does matter, and that schools in Australia are heterogeneous in the math scores the students earn.