PCR and PSL

Erik Kuitunen

October 1, 2024

Introduction, data partitioning and pre-treatment

Data used in these practical activities is house prices data from 1990. It contains 10 columns, one of which is categorical and thus dismissed. Of the nine remaining columns, eight are used as feature variables (X) and one, specifically median house values, as the target (Y).

Variable "total bedrooms" is the only variable that contains missing data, and those data points are discarded. Since the data is not time series data, regular division into calibration and test sets is made, with the test partition containing 25% of the data and without stratification. Partitions are then scaled and centered using standardization. Boxplot of scaled variables can be seen in Figure 1.

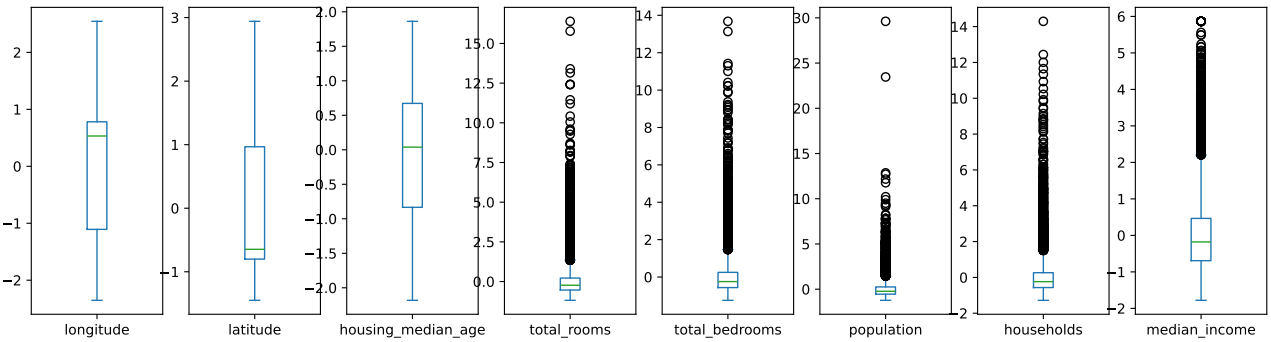

Figure 1: Boxplot of scaled variables. Numerous potential outliers can be seen in variables "housing_median_age", "total_bedrooms", "population", "households" and "median_income".

First models

PCR and PLS models are first calibrated using all available variables and PC's. A full 100% explained variance is captured in PCR, since all PC's are in use. This of course is not desirable. This PCR model achieved R^2 score of 0.64. Cumulative explained variance and the R^2 score can be seen in Figure 2. Strangely enough, identical R^2 score of 0.64 was obtained using PLS.

I was not able to understand if the explained variance of Y is indeed the R^2 score, or something else. Likewise, the explained variances of PLS method was left as an unsolved problem for me.

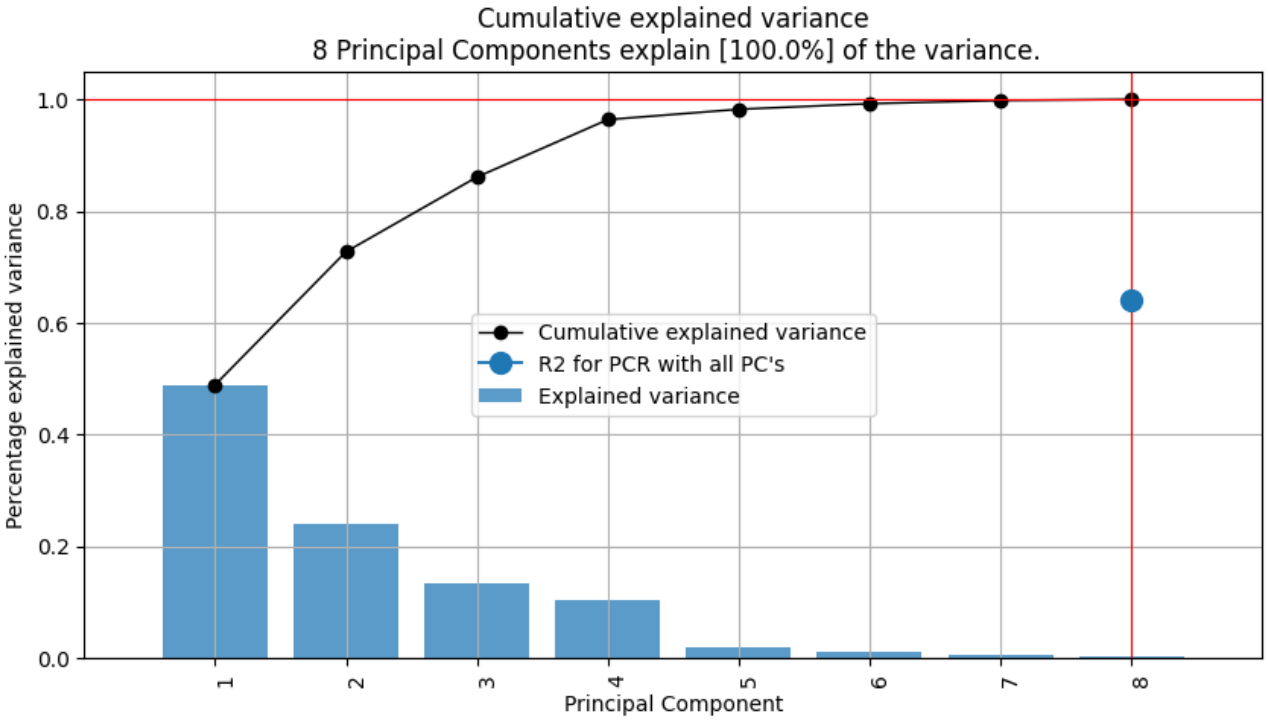

Figure 2: Cumulative explained variance of the PC's.

Number of latent variables

To determine the optimal amount of latent variables in the models, MSE- and R^2 plots were drawn, and they can be seen in Figure 3. Q^2 plot was excluded from the scope, as no cross-validation was utilized in the tasks. The author is quite confused about the fact that the plots

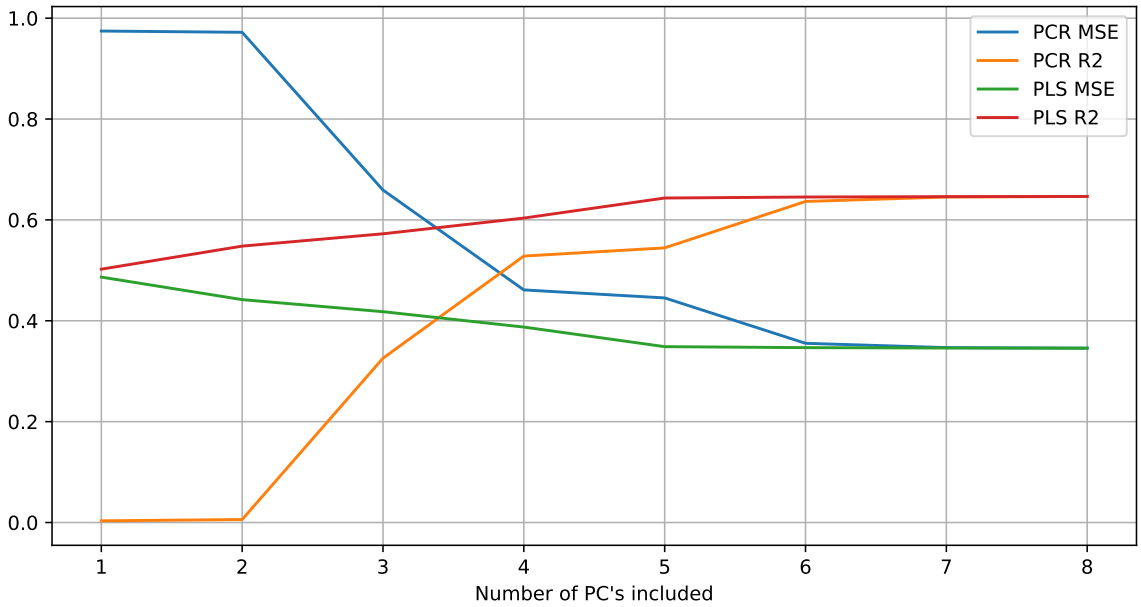

Figure 3: MSE and E^2 plots as function of number of PC's/variables included in the models.

are symmetric for the corresponding regression methods. There also seems to be some kind

of 64% cap in the R^2 score regardless of the model, which is once again something I have no answers to. Mistakes in the codes are of course always possible.

Continuing with the assumption that the plots are correct, it would seem that for PCR, four to six PC's could be optimal; there are no major changes in the R^2 score after the fourth PC, which would keep the model quite simple. Similar observations can be made for PLS model: four variables are a reasonable choice, achieving better scores than PCR while keeping the model complexity at the same level.

Regression coefficients

Regression coefficients for different models can be seen for PCR and PLS in Figures 4 and 5, respectively. Plot in the upper left corner corresponds to the model with one PC/latent variable, and lower right corner is for all eight PC's/variables.

For PCR, it would seem that "median income" is the most important variable, since it is quite large in every plot from the third plot onwards. Taking into account the scores presented in Figure 3, "housing median age" and "total rooms" are also quite important variables as they too clearly have roles in plots corresponding to PC numbers four to six. On the other hand, longitude and latitude have significant coefficients in PC number seven and eight, and cannot be fully neglected.

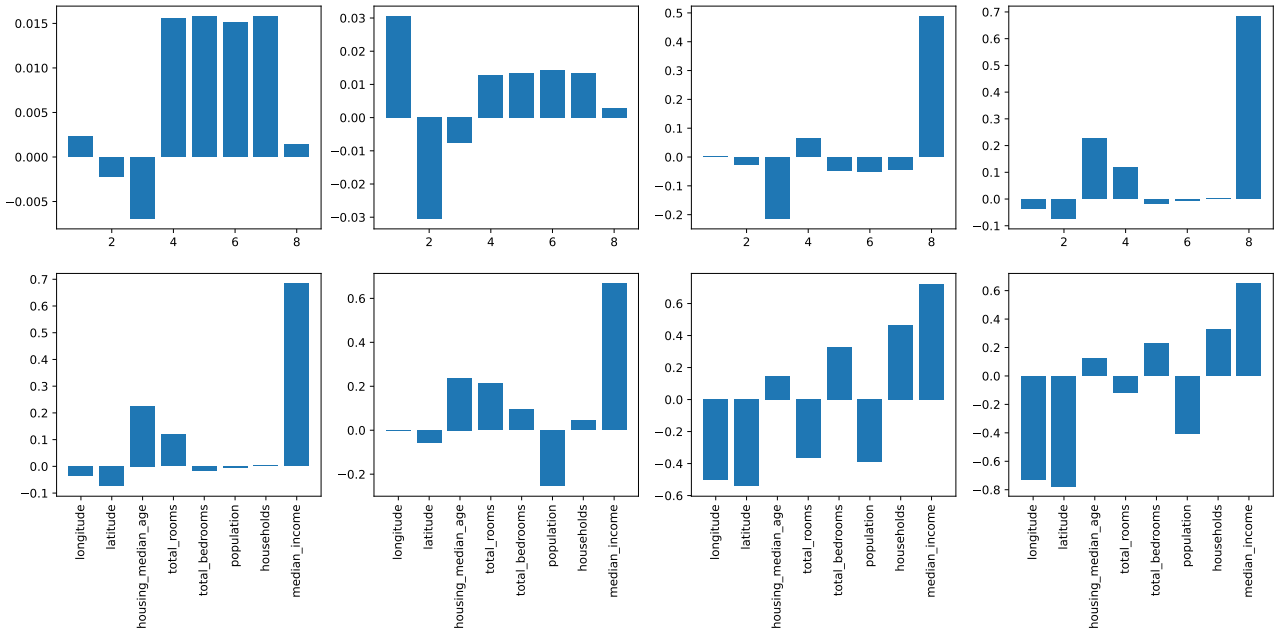

Figure 4: Regression coefficients for PCR. Number of PCs included in the model increase from left to right and up to down, starting from one at upper left corner.

Similar trend of "median income" having the most influential role continues in PLS models: it seems to have the greatest coefficient in almost all of the plots, save for the last two. Absolute values of the coefficients for "longitude" and "latitude" are also in top three in all but three first plots. The "housing median age", let alone the "number of rooms", do not seem to have quite as much impact in PLS as they had in PCR. On the other hand, "total bedrooms", "population" and "households" have coefficients with quite similar magnitudes from fourth plot onwards.

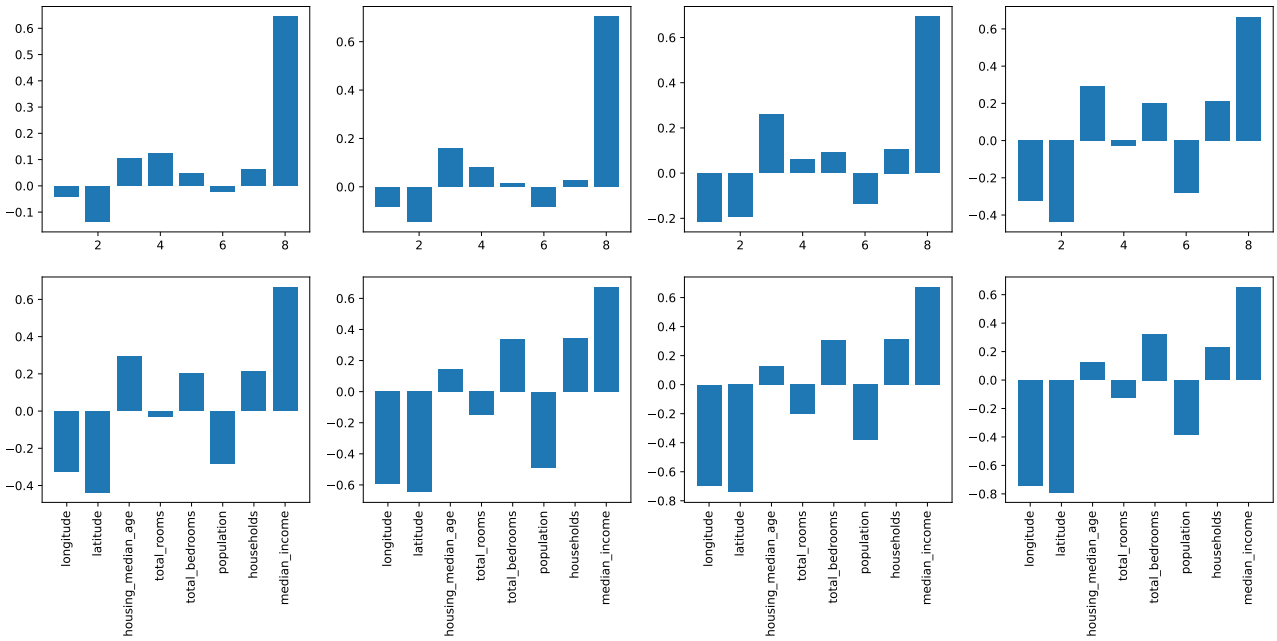

Figure 5: Regression coefficients for PLS. Number of variables included in the model increase from left to right and up to down, starting from one at upper left corner.