Homework 5: Partial Least Squares Regression

Due October 9, 2018 at 2:10 PM

This homework will investigate PLS models to predict the percent body fatgiven a variety of information about body measurements and estimated body fat percentage.

The file hwkdataNEW.mat contains data related to body measurements and estimated body fat percentage. The matrix x contains 14 predictor variables:

1. Age (yrs)
2. Weight (lbs)
3. Height (inches)
4. Adiposity Index (kg/m^2)
5. Neck circumference (cm)
6. Chest circumference (cm)
7. Abdomen circumference (cm)
8. Hip circumference (cm)
9. Thigh circumference (cm)
10. Knee circumference (cm)
11. Ankle circumference (cm)
12. Extended bicep circumference (cm)
13. Forearm circumference (cm)
14. Wrist circumference (cm)

The column vector y contains the output variable, percent body fat. The correct citation for these data is: Penrose, K., Nelson, A., and Fisher, A. (1985), "Generalized Body Composition Prediction Equation for Men Using Simple Measurement Techniques", *Medicine and Science in Sports and Exercise*, 17(2), 189.

This assignment will look at the predictive ability of partial least squares regression (PLS) and compare it to the methods we've investigated previously.

1. Divide the data into training, test, and validation data sets. You **must** use the same training, test, and validation data sets that you used in Homework 3 and 4.

2. Use cross-validation to determine the appropriate number of latent variables for your PLS model. Be sure to describe the cross-validation method in the methodology section of your report.

3. Analyze the loadings of the first few LVs. What do these tell you about the relationships between the inputs and between the inputs and the output?

4. Compare the LV loadings to the PC loadings from homework 4. What similarities and differences are there in the LVs and PCs? Explain any similarities or differences in the context of PLS and PCA (the correlation of PCs to the output may be helpful here!).

5. Compare the validation performance of your PLS model with that of your *best* PCR and regression models. Comment on the results.

6. Is there evidence that a nonlinear PLS model would outperform the linear PLS? Explain your reasoning (but you don't need to develop the nonlinear PLS model).

For this homework, prepare a written report in IEEE format. Include any plots and tables that will support your findings. Make sure you correctly label your figures and tables and refer to them in the text. Include an appropriate citation for the data, both in the text and in the list of references. Your report should include **at a minimum** an abstract, introduction, methodology, results (and discussion!), conclusions, and references. Note that the methodology section of this report (and every report!) should describe the algorithm that you're using – not the implementation in MATLAB. Include all your code in an appendix (single column) at the end of the report. Convert your report to .pdf before submitting it through Canvas.