Multivariate Data Analysis
Prof. Dr. Christina Andersson

Exercise Sheet 2: Partial Least Squares Regression

Computer Problems:

1. This example, from Umetrics (1995), demonstrates different ways to examine a PLS model. The data come from the field of drug discovery. New drugs are developed from chemicals that are biologically active. Testing a compound for biological activity is an expensive procedure, so it is useful to be able to predict biological activity from cheaper chemical measurements. In fact, computational chemistry makes it possible to calculate certain chemical measurements without even making the compound. These measurements include size, lipophilicity, and polarity at various sites on the molecule. The data set named *pentaTrain*, contains these data.

   (a) Which variables are contained in the data set?

   (b) Perform a standard PLS! Use log_RAI as response variable and $S1 - S5, L1 - L5, P1 - P5$ as explanatory variables.

   (c) How much predictor and response variation is explained by each PLS factor?

   (d) Concentrate on the first two extracted factors: How much predictor and response variation is explained?

   (e) Create a correlation loading plot for the first two factors.

   (f) Interpret the correlation loading plot.

2. *Gasoline* is a data set with NIR spectra and octane numbers of 60 gasoline samples. The NIR spectra were measured using diffuse reflectance as log(1/R) from 900 nm to 1700 nm in 2 nm intervals, giving 401 wavelengths.

   (a) Which variables are contained in the data set?

   (b) Perform a standard PLS! Use *Octane* as response variable and the NIR-variables as explanatory variables.

   (c) How much predictor and response variation is explained by each PLS factor?

   (d) Concentrate on the first two extracted factors: How much predictor and response variation is explained?

(e) Create a correlation loading plot for the first two factors.

(f) Interpret the correlation loading plot.