Stat 8932: Envelope Methods

Project Outlines

This project will focus on the application of envelope methods in Chemometrics. I will take a two-way approach: the application of PLS and sparse PLS in analyzing high-dimensional chemometrics data, as well as application of envelopes.

For each chemical compounds, there are five kinds of predictors: Topostructural, Topochemical, 3-dimensional, Quantum Chemical and Atom-pairs. Data on these predictors can be used to build models that predict certain properties of a chemical compound, like mutagenicity, toxicity etc. The dataset that I am planning to analyze has 508 compounds, each with ~2500 descriptors as well as its mutagenicity status as found by the Ames mutagenicity test. This dataset is heterogenous in nature, which means it consists of chemical compounds from different structural classes.

A common way of analyzing chemometrics data is to take a hierarchical approach: sequentially building models based on the different classes of predictors. My goal is to build hierarchical envelope models for the prediction of mutagenicity, as well as their interpretation from a chemometrics perspective. Since the dataset is heterogenous, performance of envelope models built on the full data as well as different classes of compounds can potentially differ by a considerable amount.

Following the lines of the Sparse PLS paper by Chun and Keles, models will be built based on original PLS and the sparse PLS method developed in the paper. The R-package `spls` will be used to implement sparse PLS. Excluding atom-pairs data there are about 400 predictors, so two sets of predictions based on models with and without the atom-pairs data can be used to compare the performance of PLS and S-PLS in normal and high-dimensional setup. Prediction from these models will also be compared with the envelope predictions.

References

Chun, H.; and Keles, S. Sparse Partial Least Squares Regression for Simultaneous Dimension Reduction and Variable Selection. *J. Royal Stat. Soc. Ser. B*, **2010**, 72 (1), 3–25.