Reducing Spectral Analyte Prediction Error with Penalties on Interferents
William Spence, John H. Kalivas

Department of Chemistry, Idaho State University, Pocatello, ID 83209, USA
spenwill@isu.edu; kalijo@isu.edu

Introduction
A goal of chemometric multivariate calibration (modeling) is to predict analyte concentration in a sample using spectral data. Multiple types of modeling methods have been used to predict analyte concentration. However, the samples contain interferents that influence the model and if not fully corrected by the model, analyte concentration prediction errors occur. To reduce the prediction errors caused by interferent species in the system, two new methods were designed to incorporate interferent information. One of the methods uses interferent spectra to require the model to be orthogonal to the interferents. The other method uses interferent spectra to form an orthogonal or oblique model to the interferents.

The methods are compared to ridge regression (RR) and partial least squares (PLS) using a near infrared data set. Sum of ranking differences (SRD) is used to select models. The new methods have better analytic prediction errors and robustness, but more data sets need to be tested to confirm that new methods are more effective.

Objectives
1. Determine if prediction error improves with augmenting and weighting interferent samples containing no analyte
2. Determine the best method of using samples non-analyses
3. Develop a selection method that picks the best prediction models

Beer-Lambert Law
- \( x = y, k, \ldots \) is a part of the calibration solution
- \( y \) is the concentration of the analyte
- \( k \) is the pure component spectra of the analyte
- \( \lambda \) is the concentration of the interferent species
- \( r \) is a vector of random noise that always affects the model

Multivariate Modeling
- \( y = Xb \): model calibration coefficient vector
- \( b = X^T \hat{y} \): estimated model calibration coefficient vector
- \( T \): pseudo inverse
- \( \hat{y} \) is the vector of random noise that always affects the model

Methods (continued)
2. \( b_k \) method: (PLSb & TRb)
- \( b_k \) method uses a projection matrix of the vector \( b_k \) that is made from the decomposition of the matrix \( N \)
- \( k \): set of eigenvectors
- \( u \): sample vectors of \( N \)
- \( N = X^T N \)

Experimental (continued)
- Cross validation:
  - 70% calibration picked out of 16 samples
  - 1000 splits
- Tuning parameter ranges:
  - \( k \): 2LV through 11LVs are used
  - \( \eta \) & \( \lambda \): 0 to 5
- Validation:
  - \( \eta \) & \( \lambda \): 0.0001 to 0.1
- Selection of \( \eta \) & \( \lambda \):
  - 1: higher weight of variance and bias
  - 2: higher weight of bias and variance

Conclusion
- Compared to PLS and RR:
  - Augmenting and weighting non-analyte spectra reduces prediction error
  - TRN reduces the prediction error and variance the most
  - New models in new methods improved prediction error but increased variance

Continuing Work
- \( b_k \) methods will be added to list of methods used in SRD
- \( b_k \) methods will influence SRD to pick a higher weight of \( \eta \)
- Higher \( \eta \) have more improvement in prediction and variance

Acknowledgements
This project is supported through the National Science Foundation under Grant No. CHE-1506847 (co-funded by CDS and E Programs), and Idaho State University.