Gray Classical Least Squares: Putting Chemistry Back into Calibration Models

Barry M. Wise[1], Donal O'Sullivan[1] and Rasmus Bro[2]
[1] Eigenvector Research, Inc. Manson, WA USA, bmw@eigenvector.com
[2] University of Copenhagen, Copenhagen, DENMARK, rb@food.ku.dk

Keywords: Regression, Classical Least Squares, Othogonalization Filters.

1 Introduction

There is a resurgence in the use of Classical Least Squares (CLS) models primarily due to their interpretability. When used with spectroscopic systems that follow the Lambert-Beer law CLS models follow naturally from first principles. Unfortunately, CLS models typically do not have the predictive ability of inverse least squares (ILS) models such as Partial Least Squares (PLS) regression: the prediction error of CLS models is usually higher, and often notably so. This is largely due to non-idealities in the data of interest along with the presence of unaccounted for minor components, *e.g.* scatter and baseline variations. PLS models handle these situations by adding components to the model that keep the resulting regression vector orthogonal to the non-ideal variations.

2 Theory

In this work we propose a method for developing CLS models[1] with predictive properties competitive with ILS formulations. This is done by first creating the CLS model "half-residuals." Typically, the residuals in CLS models are obtained with the estimated concentrations, $\widehat{\boldsymbol{C}}$. Given initial values for measured spectra $\boldsymbol{X}$ and measured concentrations $\boldsymbol{C}$, the normal equations are applied to obtain the estimated pure component spectra $\widehat{\boldsymbol{S}}$, and then $\widehat{\boldsymbol{S}}$ is used with $\boldsymbol{X}$ to obtain $\widehat{\boldsymbol{C}}$. These are used to reconstruct $\boldsymbol{X}$ and the residuals are calculated as follows:

$$\boldsymbol{R_c} = \boldsymbol{X} - \widehat{\boldsymbol{C}}\widehat{\boldsymbol{S}}^T \qquad (1)$$

The subscript c is added to denote that these are the conventional residuals, $\boldsymbol{R_c}$. On the other hand, residuals could be calculated using the estimated $\widehat{\boldsymbol{S}}$ and the original $\boldsymbol{C}$.

$$\boldsymbol{R_h} = \boldsymbol{X} - \boldsymbol{C}\widehat{\boldsymbol{S}}^T \qquad (2)$$

We refer to these as the "half residuals" $\boldsymbol{R_h}$. An important distinction between these two types of residuals is that $\boldsymbol{R_c}$ is completely orthogonal to $\widehat{\boldsymbol{S}}$, whereas $\boldsymbol{R_h}$ is not. Thus, any filter based on $\boldsymbol{R_c}$ would have no effect on predictions. $\boldsymbol{R_h}$ on the other hand can be used to develop pre-filters with Generalized Least Squares Weighting[2] (GLSW) or External Parameter Orthogonalization[3] (EPO). We refer to these as Gray CLS models. It can be shown for the EPO case that these models are equivalent to Extended Least Squares (ELS) formulations.

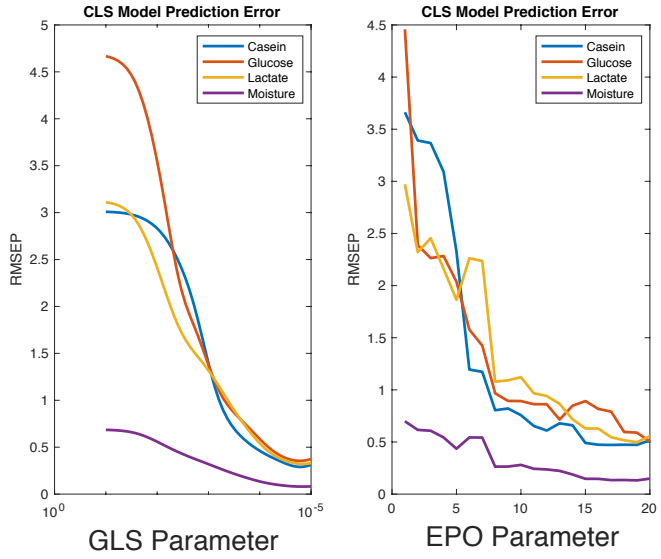
3 Material and methods

The proposed Gray CLS model formulation is tested on 5 publicly available data sets. Each data set was split into calibration and independent validation sets. Root-mean-square error of calibration

(RMSEC), error of cross-validation (RMSECV) and prediction on the separate test sets (RMSEP) were computed for all the data sets.

4 Results and discussion

Typical results for predictions from Gray CLS Model are shown below. RMSEC, RMSECV (not shown) and RMSEP (below) all decline substantially compared to the conventional CLS model, which starts from the values at the left side of each panel. The GLSW results are especially appealing because of the smoothness of the curves with the continuously adjustable *g* parameter. EPO results tend to have local minimums and in general do not achieve results as low as GLSW.



RMSEP on Validation Set for Casein-Glucose-Lactate Data.

These formulations retain the opportunity to learn from the analyte factors *(i.e.* pure component spectra) and from the GLSW or EPO model parameters.

5 Conclusion

The CLS gray model is yet another way to skin the multivariate calibration cat. Results using the method are competitive in terms of prediction error in most instances with PLS models. Like PLS and PCR models they use a single adjustable parameter but have the advantage that all analytes are predicted from a single model. They are also more firmly based in first principles than ILS models and as such may be more easily explainable to the chemometricly unwashed.

6 References

[1] J. Workman, "Classical Least Squares, Part I: Mathematical Theory" *Spectroscopy*, May 1, 2010.

[2] J.M. Roger, F. Chauchard, V. Bellon-Maurel, "EPO–PLS external parameter orthogonalisation of PLS application to temperature-independent measurement of sugar content of intact fruits" *Chemo. Intell. Lab Sys*., 66(2), pps 191-204, June 2003.

[3] H. Martens, M. Høy, B.M. Wise, R. Bro and P.B. Brockhoff, "Pre-whitening of data by covariance-weighted pre-processing," *Journal of Chemometrics*, **17**(3), pps 153-165, March 2003.