Putting Chemistry Back Into NIR Calibration Models with Gray Classical Least Squares

Barry M. Wise[1], Donal O'Sullivan[1] and Rasmus Bro[2]
[1]Eigenvector Research, Inc.
[2]University of Copenhagen

EIGENVECTOR
RESEARCH INCORPORATED

Abstract

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. In this work we examine a method for developing CLS models with predictive properties competitive with ILS formulations. This is done by using the CLS model "half-residuals" to develop pre-filters with Generalized Least Squares Weighting (GLSW) or External Parameter Orthogonalization (EPO). The result is calibration models that have chemically meaningful estimates of the pure component spectra, interpretable factors for non-idealities and minor components and good predictive ability. Gray CLS models are demonstrated with several NIR data sets and their performance is shown comparable to PLS models.

Outline

- The Classical Least Squares Model
  - Ways to generate spectral residuals
- Clutter Filters
  - EPO: External Parameter Orthogonalization
  - GLSW: Generalized Least Squares Weighting
- Example Data Sets
- Results
  - Choosing the Meta-parameters
  - Diagnostics
- Conclusions

Classical Least Squares

- CLS often used to develop spectroscopic calibration models
- The CLS assumes the data can be modeled as

$$\boldsymbol{X} = \boldsymbol{C}\boldsymbol{S}^T + \boldsymbol{E}$$

where:

- $\boldsymbol{X}$ $(M \times N_{\mathrm{x}})$ is the measured spectral response
- $\boldsymbol{S}$ $(N_{\mathrm{x}} \times K)$ is a matrix of pure spectral responses,
- $\boldsymbol{C}$ $(M \times K)$ is a matrix of concentrations and
- $\boldsymbol{E}$ $(M \times N_{\mathrm{x}})$ is noise or an error matrix.

Using the CLS Model

- If $\boldsymbol{S}$ is known, the $\widehat{\boldsymbol{C}}$ can be estimated from
  - $\widehat{\boldsymbol{C}} = \boldsymbol{X}\boldsymbol{S}(\boldsymbol{S}^{\boldsymbol{T}}\boldsymbol{S})^{-1}$
- If $\boldsymbol{S}$ is not known, it can be estimated from a (properly designed) calibration data set
  - $\hat{\boldsymbol{S}} = (\boldsymbol{C}^{T}\boldsymbol{C})^{-1}\boldsymbol{C}^{\boldsymbol{T}}\boldsymbol{X}$
- This is often the best way to estimate $\boldsymbol{S}$
  - Models $\boldsymbol{S}$ in the relevant sample matrix
  - Temperature, pressure, scattering effects, etc.

CLS Spectral Residuals

- Given $\hat{\boldsymbol{S}}$ there are two ways to get spectral residuals
- Conventional $\boldsymbol{R_c}$, estimate $\hat{\boldsymbol{C}}$ as above then
  - $\boldsymbol{R_c} = \boldsymbol{X} - \hat{\boldsymbol{C}}\hat{\boldsymbol{S}}^T$
- Half residuals $\boldsymbol{R_h}$, use original $\boldsymbol{C}$ instead of $\hat{\boldsymbol{C}}$ ← *CRUX*
  - $\boldsymbol{R_h} = \boldsymbol{X} - \boldsymbol{C}\hat{\boldsymbol{S}}^T$
- Note that $\boldsymbol{R_c}$ is orthogonal to $\hat{\boldsymbol{S}}$, whereas $\boldsymbol{R_h}$ is not
  - This will be important later!

EIGENVECTOR RESEARCH INCORPORATED

Main Problem with CLS

- As originally formulated, typically not competitive with PLS and other Inverse Least Squares (ILS) approaches on prediction error
- Can’t use with un-quantified unknown components
- Factor based methods (PLS, PCR) compensate for non-idealities by going beyond the number of known components
- With CLS you’re stuck. ??

Clutter Orthogonalization Filters

- Typically used as preprocessing in PLS or other ILS models
- Mitigate effect of large variations in spectra ***not*** related to property of interest
- Consider two popular orthogonalization filters here
  - External Parameter Orthogonalization (EPO)
  - Generalized Least Squares Weighting (GLSW)

EPO Filter

- Given a matrix $\boldsymbol{Z}$ which represents extraneous variation (matrix effects, clutter), decompose $\boldsymbol{Z}$ as
  - $\boldsymbol{Z} = \boldsymbol{USV}^T$

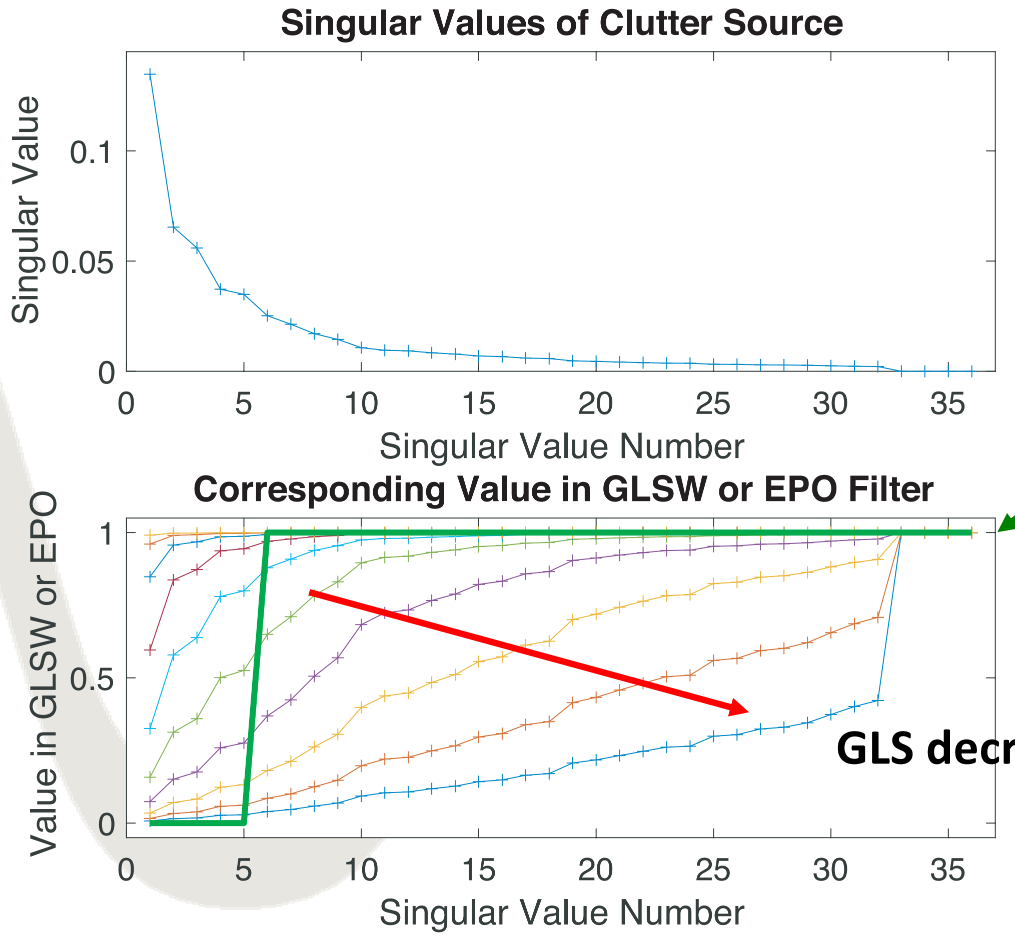
$\boldsymbol{Z} = ?$

- The number of filter factors $\boldsymbol{k}$ must be specified, then
  - $\boldsymbol{F}_{\boldsymbol{epo}} = \boldsymbol{I} - \boldsymbol{V}_{\boldsymbol{k}}\boldsymbol{V}_{\boldsymbol{k}}^{\boldsymbol{T}}$
- $\boldsymbol{F}_{\boldsymbol{epo}}$ is applied to $\boldsymbol{X}$ before calibration (and during prediction), and removes variations in the first $\boldsymbol{k}$ dimensions represented in $\boldsymbol{Z}$
- Equivalent to Extended Mixture Model (EMM)

GLSW Filter

- Similar to EPO filter except it shrinks dimensions rather than completely eliminating them
- Starting from the decomposition of $\boldsymbol{Z}$ above then
  - $\boldsymbol{F_{glsw}} = \boldsymbol{VD^{-1}V^{T}}$
- Where the diagonal elements of $\boldsymbol{D}$ are calculated as
  - $d_i = \sqrt{\left(\frac{s_i^2}{g^2}\right) + 1}$
- Where s_i is the ith diagonal element of $\boldsymbol{S}$ and g is a tune-able parameter which controls the shrinkage

Comparison of EPO & GLSW



Singular Values of Clutter Source

Corresponding Value in GLSW or EPO Filter

EPO 5 factors

GLS decreasing g

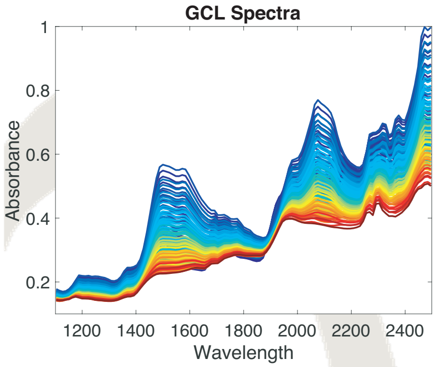
Combining CLS & Filters

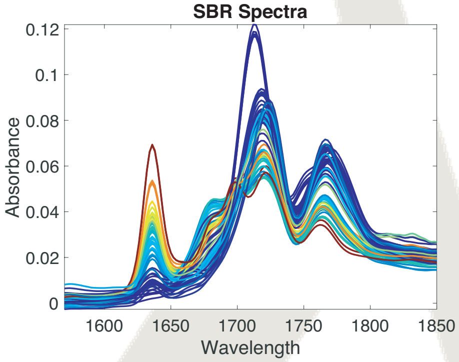
- Residuals from CLS models can be used as an estimate of the clutter, $\boldsymbol{Z}$. However,
  - Filter based on $\boldsymbol{R_c}$ has *no effect* as it is orthogonal to $\hat{\boldsymbol{S}}$.
  - $\boldsymbol{R_h}$, on the other hand, contains information about clutter *not* orthogonal to $\hat{\boldsymbol{S}}$.
- Filter based on $\boldsymbol{R_h}$ mitigates clutter *not* orthogonal to $\hat{\boldsymbol{S}}$ that would otherwise lead to additional error in $\hat{\boldsymbol{C}}$.

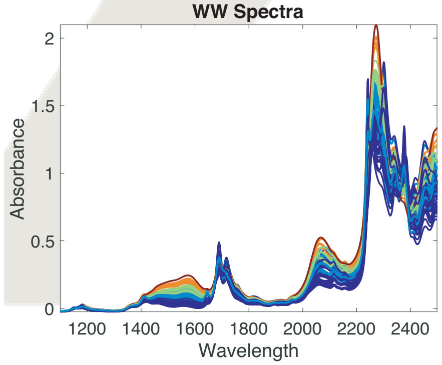
CRUX

The CLS Gray Model

- We refer to this combination of a CLS model with a filter based on the half residuals $\boldsymbol{R_h}$ as a “gray model”
- Incorporates aspects of both CLS and ILS models.
  - Based on a first principles model, the CLS “white” part
  - Includes tunable empirical part, EPO or GLSW filter “black” part
  - This model has a single adjustable parameter ($\boldsymbol{k}$ or g)
- GLS generally outperforms EPO, so only GLS results shown





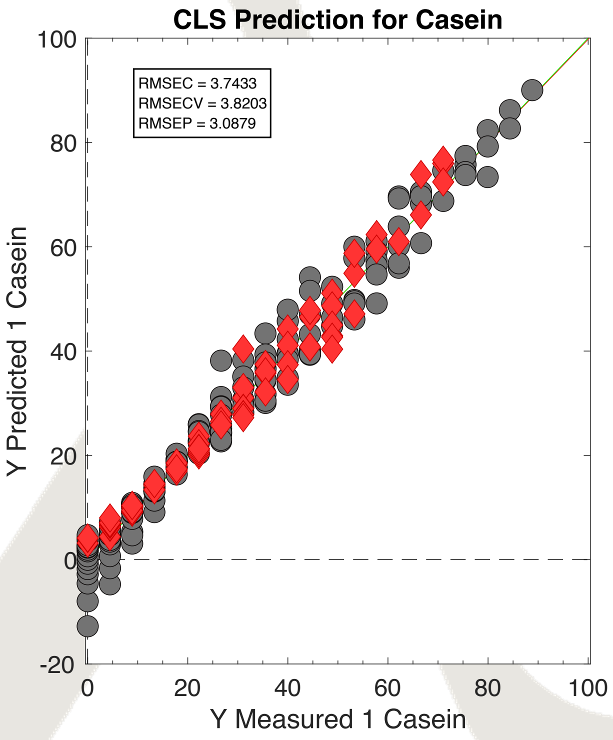


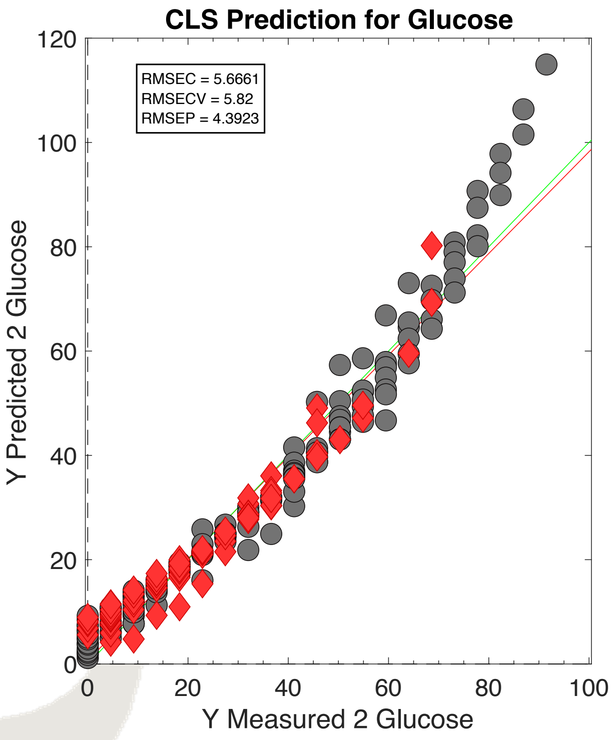
NIR Data Sets

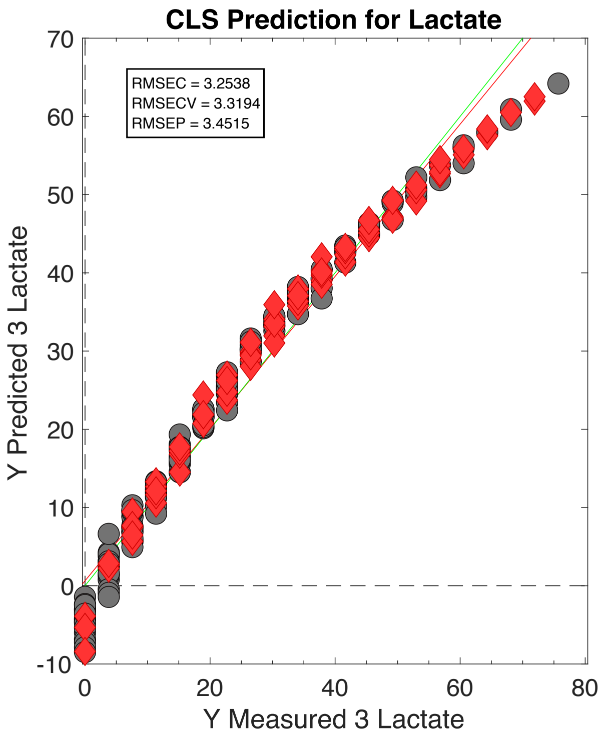
- Grain protein from Tormod Naes/Tomas Isaakson (CGL)
  - Casein, glucose, lactate and moisture, 231 samples (split 153/78), 117 wavelengths, full 3 component mixture design,
- Styrene-butadiene from Dupont/Chuck Miller (SBR)
  - Styrene, cis-, trans- and 1,2-butadiene, 70 samples (split 60/10), 141 wavelengths.
- Hydrocarbon mixture from Willem Windig (WW)
  - Butanol, dichloromethane, methanol, dichloropropane and acetone, 140 samples (split 93/47), 700 wavelengths, full design.

EIGENVECTOR
RESEARCH INCORPORATED

CLS Predictions for CGL

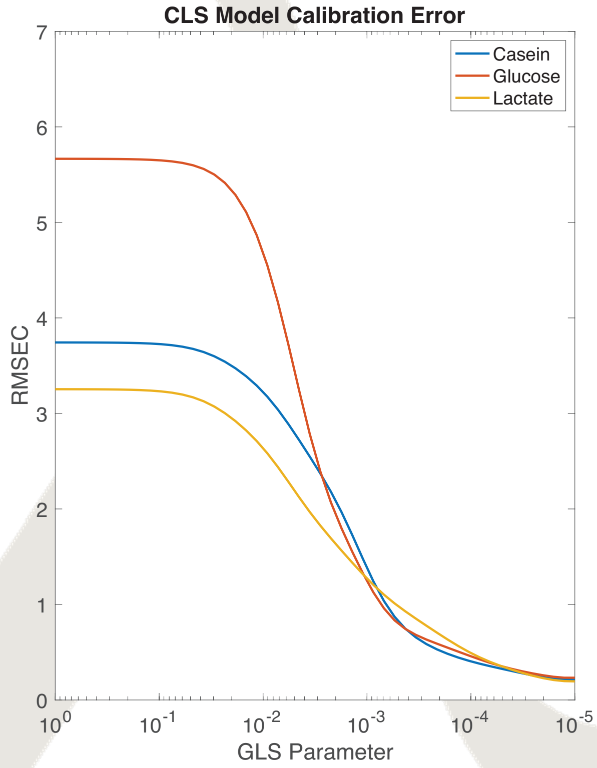


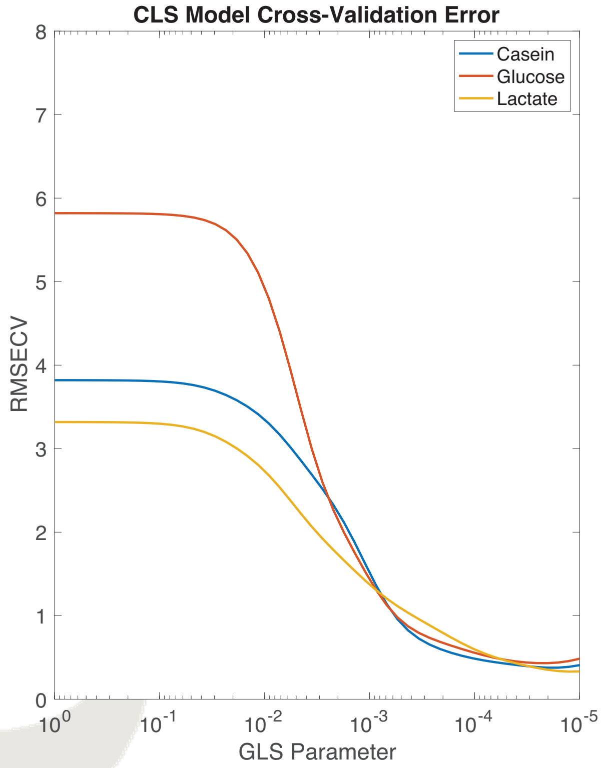


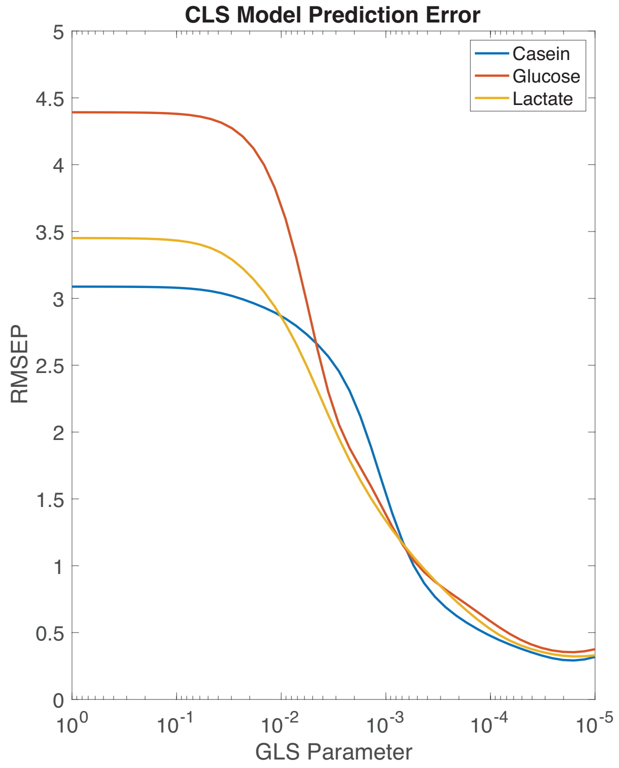


EIGENVECTOR RESEARCH INCORPORATED

CGL Model Error – RMSEC/CV/P

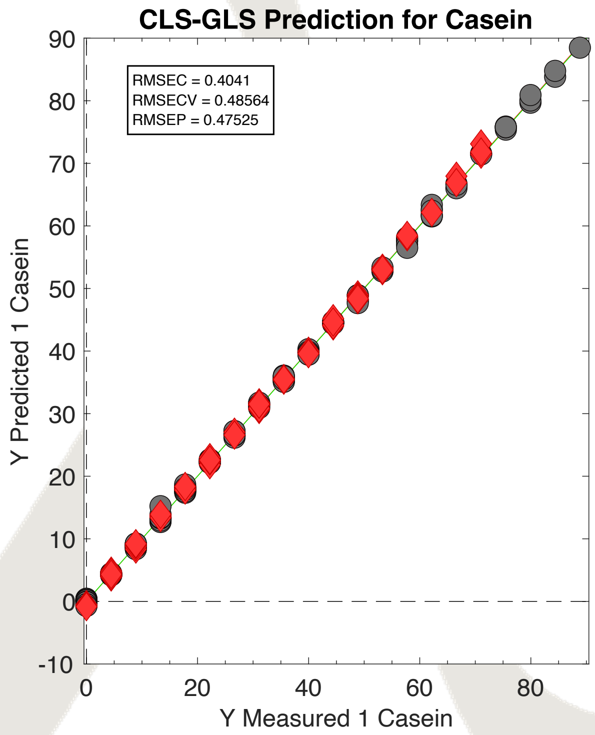

CLS Model Calibration Error

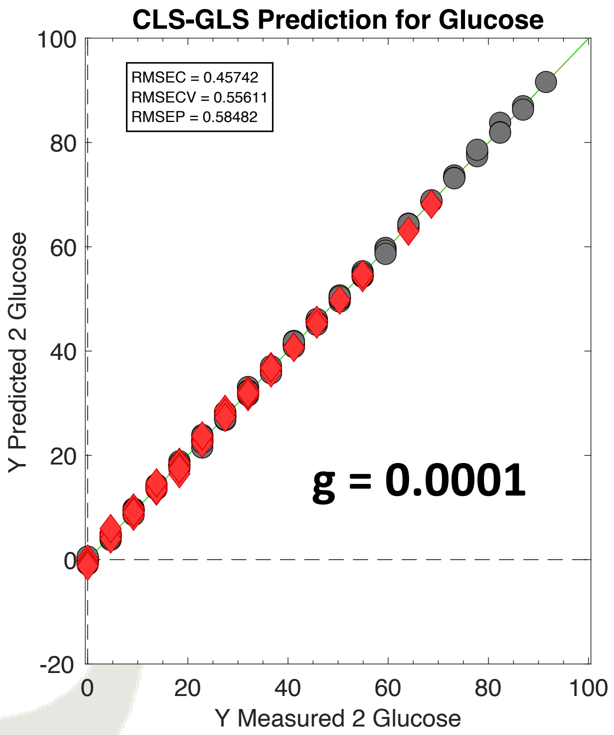

CLS Model Cross-Validation Error

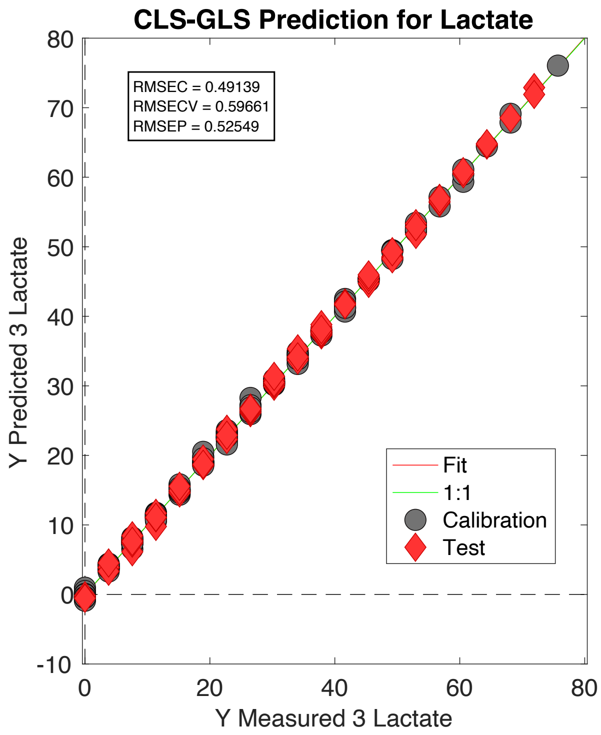

CLS Model Prediction Error

EIGENVECTOR RESEARCH INCORPORATED

CLS-GLS Predictions for CGL

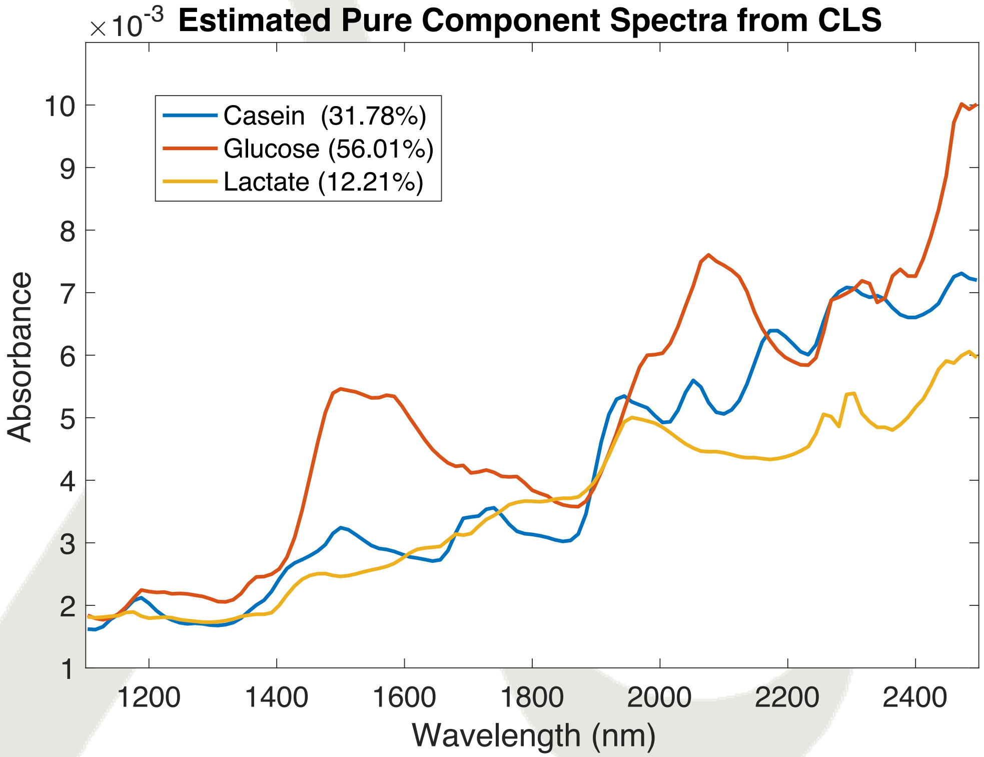


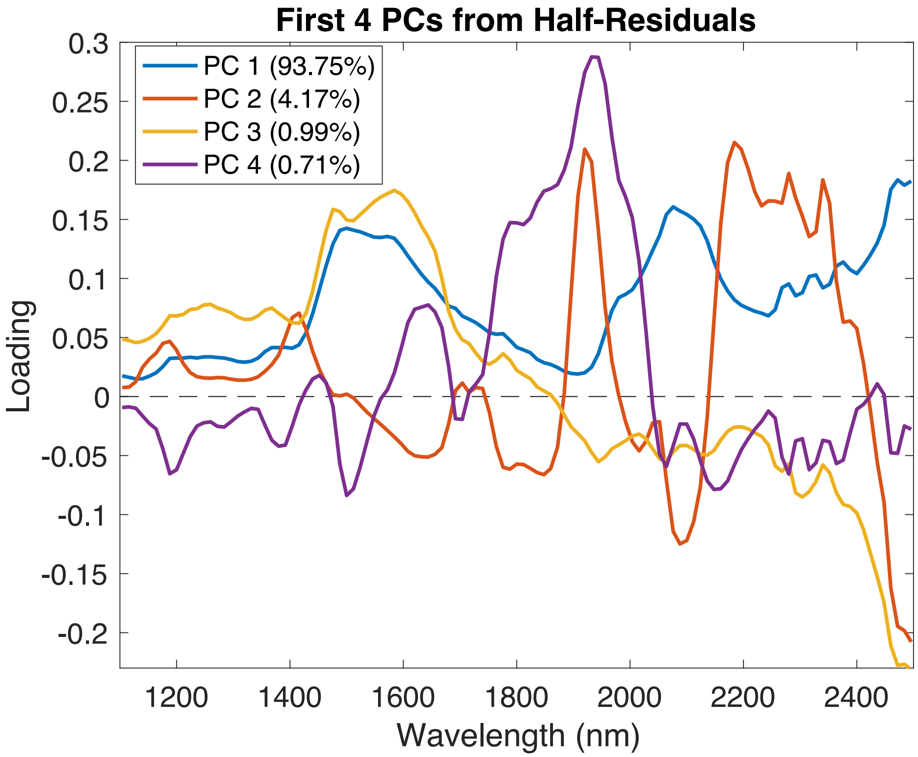




CLS	0.47	0.58	0.53	g=0.0001
PLS	0.52	0.68	0.59	12 LVs

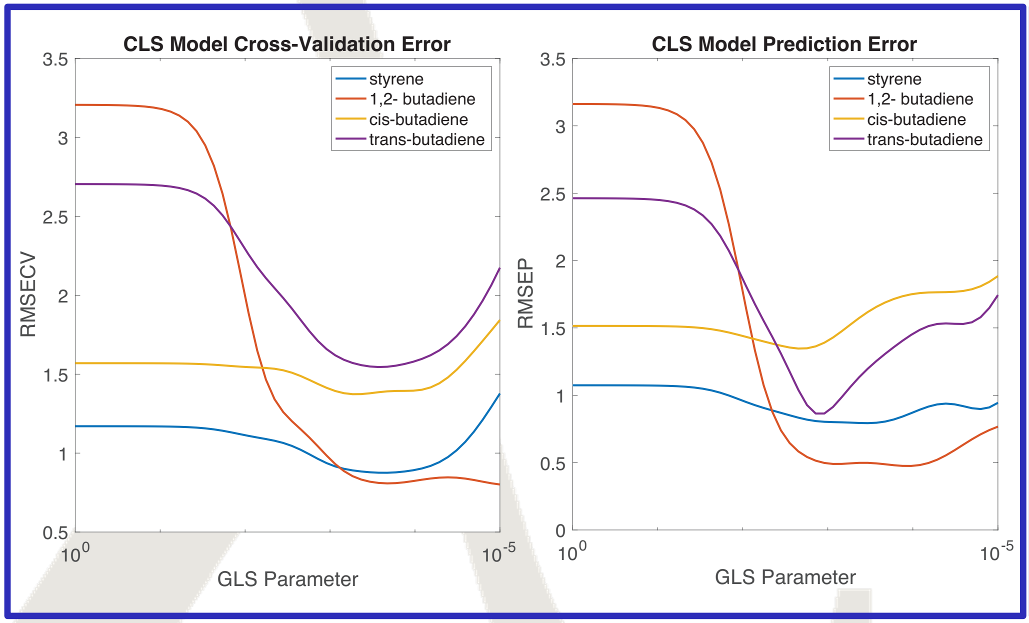
Diagnostic Information





EIGENVECTOR RESEARCH INCORPORATED

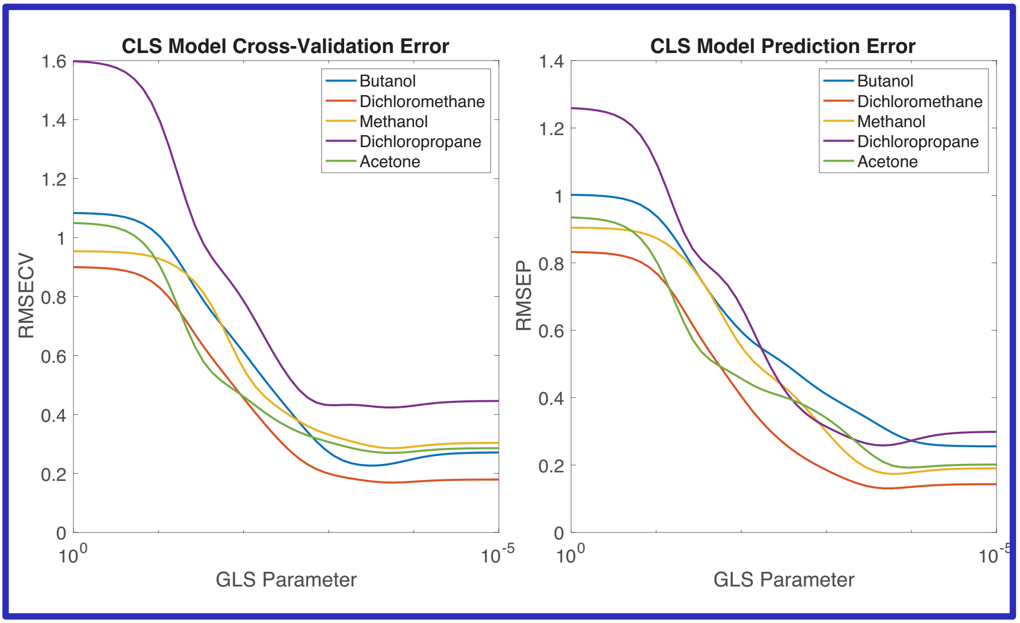
SBR Data Results



CLS Model Cross-Validation Error

CLS Model Prediction Error

WW Data Results



CLS Model Cross-Validation Error

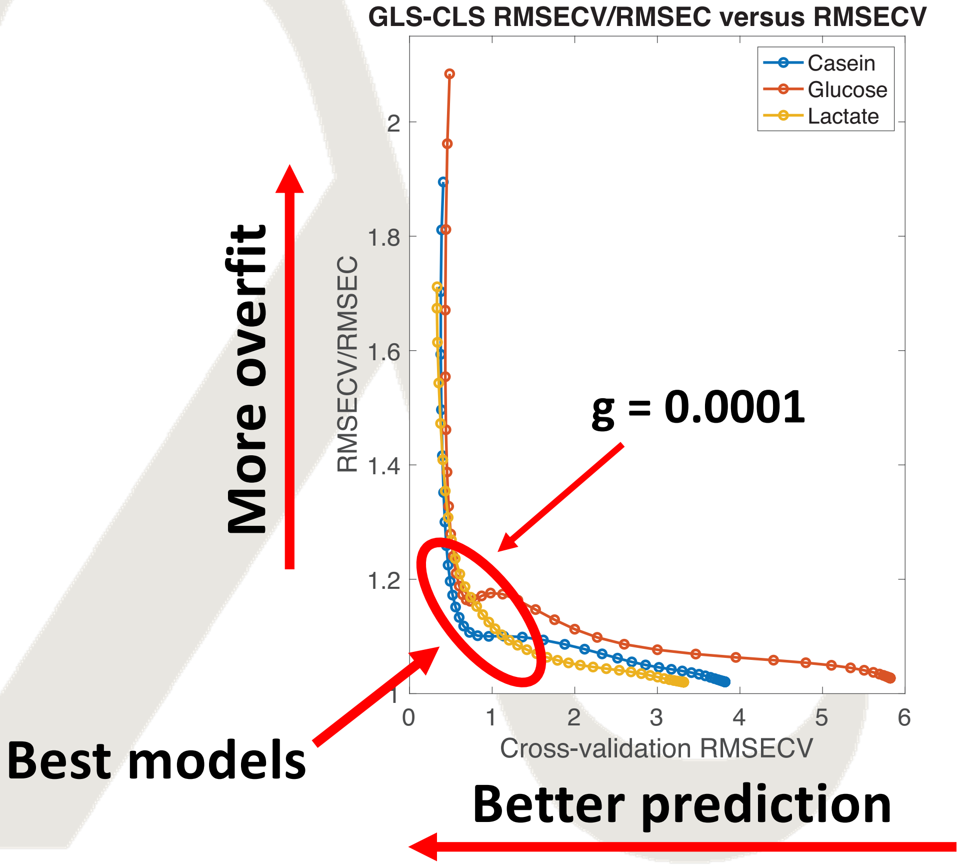
CLS Model Prediction Error

EIGENVECTOR
RESEARCH INCORPORATED

Choosing Meta-parameters

- Usual aspects of cross-validation apply
- Watch for overfit, *i.e.* when fit error (RMSEC) is much lower than prediction error (RMSECV)
  - Plot Overfit vs. Cross-validation error
    - Ratio RMSECV/RMSEC versus RMSECV
    - In units of predicted variable

RMSECV/C for CGL & SBR

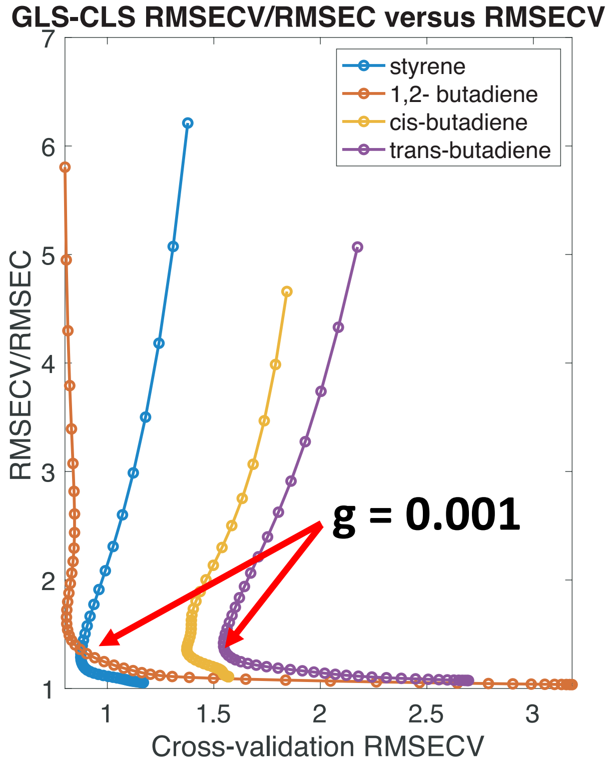


g = 0.0001

More overfit

Best models

Better prediction



g = 0.001

Conclusions

- EPO and GLSW filters can be used to improve CLS model predictive performance – **Gray Models**
  - Key is use of “half-residuals” $\boldsymbol{R_h}$
  - One adjustable parameter ($\boldsymbol{k}$ or g)
- Resulting models competitive with PLS in predictive ability
- Model selection criteria as usual
  - Overfit (RMSECV/C) vs. Prediction (RMSECV) very useful!
- Main advantages interpretability, explain-ability
- Available in PLS_Toolbox/Solo 9.3