Directing a calibration to a useful model in the deployment domain

J.H.Kalivas[1] R. Spiers[2]

[1] Department of Chemistry, Idaho State University, Pocatello, ID 83209 USA, johnkalivas1@isu.edu

[2] Department of Chemistry, Idaho State University, Pocatello, ID 83209 USA, robertspiers@isu.edu

Keywords: Physicochemical matrix matching, local modeling, domain adaptation, Rashomon effect.

1 Introduction

Chemometric calibration (machine learning training) with spectral data can form accurate prediction models. However, once such a “good” chemical model has been obtained relative to the source training data, it often fails to correctly predict analyte amounts for samples from a target deployment domain. Model failure results because the training set does not fully *match* the target domains in terms of all hidden matrix effects influencing measured spectra. These non-trackable matrix effects stem from a large variety of sources and are sample dependent. Some example sources that alter spectra include humidity, temperature, instrument drift, sample composition (analyte and other species amounts), and *physico*chemical properties such as inter- and intra-molecular interactions. For biological specimens, *physio*chemical properties affect spectra sample-wise due to cellular microenvironments variances. Thus, it is difficult to form a global calibration to predict all potential target samples and model performance degrades as the target deployment domain conditions deviate from the original source calibration conditions. Useful models for the deployment domain with accurate predictions can be obtained by directed adjustments to source models.

Model updating based approaches can be used to solve the matrix matching problem by augmenting target domain spectra to calibration spectra. The original source calibration model is reorientated to a useful direction and magnitude that makes the model invariant to the condition differences between the source and target sample domain. Various approaches exist including only adding unlabeled target domain spectra (samples without reference values) [1]. Due to thousands of models generated with data augmentation, selection of a useful model is difficult. In fact, the Rashomon effect of modeling characterizes the situation and model interpretability is questionable. Model diversity and prediction similarity (MDPS) was recently developed for model selection [2] and is briefly presented. Unique is that no target deployment domain sample reference values are used.

Another tactic to solve the sample-wise matrix matching problem is local modeling where a library of sample spectra is mined for calibration set of samples matrix matched to each new target sample spectrum. A model is then formed to predict the analyte in the target sample. Because the hidden matrix effects are sample-wise unique, each target sample requires mining for its particular matrix matched training set. Confounding chemical spectral based library searches is that libraries contain thousands of samples with conflicting hidden matrix effects making it difficult to identify a matched training set for each target sample spectrum simultaneously with the particular analyte prediction property amount. Further constraining the methods is the abundance of hyperparameters requiring manual optimization. Presented is local adaptive fusion regression (LAFR) that forms hundreds of linear local models from a reference database where each calibration set focuses on distinct and consistent hidden matrix effects. Developed for LAFR is a measure termed the physicochemical responding integrated similarity metric (PRISM) that is a hybrid fusion algorithm [3,4] based on a consensus of similarity measures using a novel cross-modeling procedure. With PRISM, LAFR is directed to select only useful models. All LAFR hyperparameters are self-optimized making LAFR autonomous.

2 Material and methods

All algorithms were developed by the authors using MATLAB. A suite of model updating algorithms by data augmentation as well as the MDPS model selection algorithm can be downloaded [5]. Results for LAFR are presented using a difficult NIR soil library with nearly 100,000 reference samples and spectra from across the United States.

3 Results and discussion

Figure 1 shows a graphical characterization of the model updating and selection process. References 1 and 2 contain results for several NIR data sets. Shown in Figure 2 are typical LAFR results demonstrating calibration localization to closely bracket the analyte amount in each target sample.

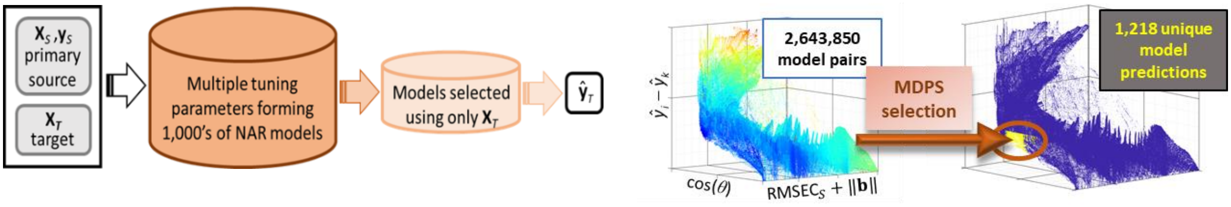
X_S, y_S primary source
X_T target
Multiple tuning parameters forming 1,000's of NAR models
Models selected using only X_T
$\hat{y}_T$
2,643,850 model pairs
MDPS selection
1,218 unique model predictions
$\hat{y}_i - \hat{y}_k$
$\cos(\theta)$
$RMSEC_S + \|\mathbf{b}\|$

Figure 1 – Data augmentation by null augmented regression (NAR) with unlabeled samples (left) and (right) the model selection approach with the color scheme varying from the best RMSEV (blue) values to the worst (red). Models are selected in a region balancing model dissimilarly and prediction similarity.

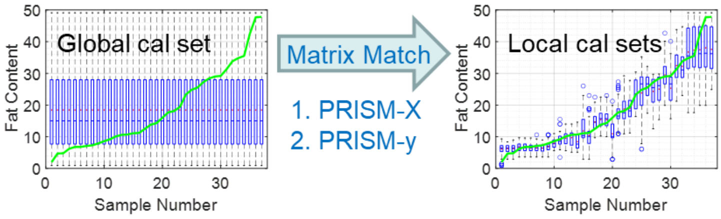

Figure 2 – Typical LAFR results where the green line is the target sample amount for a NIR meat data set.

4 Conclusion

Matrix matching a calibration set to the deployment domain is difficult at best. Overviewed was a model updating approach with a model selection method that provides a solution to this problem under certain conditions. For the more generic situation, the local modeling method LAFR was shown to successfully mine a large library identifying calibration sets closely matched to each target sample in terms of the analyte content. Key to the success of LAFR is PRISM that uses a consensus of similarity measures to assess each library sample for its degree of matrix matching to a target prediction sample. Ongoing work with PRISM includes using it to identify how similar two datasets are. Lastly, the Rashomon effect was introduced bringing in the question of model interpretability.

5 References

[1] R.C. Spiers, J.H. Kalivas: Calibration model updating to novel sample and measurement conditions without reference values. *Anal. Chem.*, 93:9688-9696, 2021.

[2] R.C. Spiers, J.H. Kalivas: Reliable model selection with reference values by utilizing model diversity with prediction similarity. *CJ. Chem. Info. Model.*, 61:2220-22230, 2021.

[3] Brownfield, T. Lemos, J.H. Kalivas: Consensus classification using non-optimized classifiers. *Anal. Chem.*, 90:4429-4437, 2018.

[4] T. Lemos, R. Emerson, J.H. Kalivas: Identifying chemical, physical, and instrumental matrix matched samples by leveraging spectral model regression vectors. *Anal. Chem.*, 92:815-823, 2020.

[5] Model updating and selection software link: https://www.isu.edu/chem/faculty/staffdirectoryentries/kalivas-john.html.