Harnessing Model Diversity and Prediction Similarity for Selecting Multivariate Calibration Tuning Parameters

Robert Spiers, John H. Kalivas
Department of Chemistry
Idaho State University
921 S. 8th Ave., STOP 8023 Pocatello, ID 83209, USA
spieryb2@isu.edu, kalijohn@isu.edu

Abstract
Multivariate calibration offers a more cost-effective mechanism to obtain sample analyte values of a substance (e.g., protein, moisture). However, the calibration process requires variation of certain tuning parameters in order to obtain the most accurate model, which requires an optimal model to be selected from the given options. Model selection is especially important in the case of model updating, where models are calibrated from spectral and reference information in both the original (primary) conditions and new (secondary) conditions in order to better predict new spectra generated in secondary conditions. Secondary conditions can include new instruments, temperatures, or any other condition affecting the shape and magnitude of the spectra relative to analyte values. The difficulty of model selection is exacerbated as the number of tuning parameters increases relative to the model. In contrast with other model selection techniques, this poster prioritizes model diversity while maintaining similar analyte prediction values to choose a set of acceptable models. Selection is achieved by comparing every combination of two models and the generated predictions. This model selection technique is tested across the calibration method partial least squares (PLS) and four model updating methods: two require a small set of secondary samples with analyte values and two do not require the secondary analyte values (unlabeled data). This novel approach of model selection was assessed using different weighted combinations of model diversity and prediction similarity measures in order to determine the combination with the lowest prediction error of new secondary samples across a variety of datasets and conditions. Results are presented showing the cosine of the angle between models in combination with model vector 2-norms and prediction differences are key to selecting models.

Objective
- Develop and analyze a new model selection method based on model diversity and prediction similarity (MDPS).
- Confirm feasibility by benchmarking against the first quartile of all models in the calibration or updating sets.

Approach
Five model generation methods are used:

One multivariate calibration method

- Partial Least Squares (PLS)
  - Requires only a single tuning parameter

Four model updating methods

- All require two tuning parameters
- d = Number of Latent Variables

Labeled Secondary

- Local Mean Centering (LMC)
- Null Augmentation Regression (NAR
- Feature Augmentation 2A (FA-2A)
- Null-Centroid (NAR-C)
- Null-Diagonal (NAR-D)

Validating Results:
Selected models are validated by using additional spectra from the secondary sample set that were not included in forming the model

Methodology
Pre-process Calibration
Generate b
Generate y with b
Repeat for all combinations of tuning parameters

For every combination of models:
  - Compare the two b
  - Compare corresponding y
  - Take all combinations within a diversity window

Take 100 combinations with lowest prediction difference

Model Generation:
Iterating through every combination of tuning parameters to create a set of total models

Model Selection:
Using model diversity and prediction similarity (MDPS) measures to select a subset of models from the total

Data Description
Corn dataset: 700 NIR wavelength absorbance with four analyte values, moisture, oil, protein, and starch for 80 samples of corn measured on three instruments, m5, m7, m9. Each combination of analyte with instrument are analyzed as primary and secondary

Soil dataset: Spectra of soil samples and their corresponding concentrations of organic content are divided into two sets: Global and Bilan (Montana). Global is analyzed as primary, and Montana as secondary

Tablet dataset: Four batches of pharmaceutical tablets sorted by active pharmaceutical ingredient (API) are split into a laboratory subset and full production subset, with 30 samples in each batch and subset. Lab is always analyzed as primary, and full as secondary

Metaparameter Convergence
An algorithm was developed to automatically find the region of interest to perform model selection in this method confirms the first quartile and median of all possible models by excluding repetitive models

Conclusion
- Robust, dataset independent model selection can be performed using model diversity and prediction similarity measures
- Cosine of the angles between the two models is most effective
- Using sum weighted fusion between 2-norm and secondary prediction differences solves the problem of overfitting
- MDPS model selection works well with models low RMSE
- Nearly universally performs at or below the first quartile
- Can outperform existing methods of model selection
- Using NAR methods, MDPS provides the first method of harnessing entirely unlabeled secondary data for model updating and selection
- NAR-C with entirely unlabeled secondary data is shown to often produce similar prediction error as labeled secondary methods

Future Work
- Apply Tikhonov Regularization methods instead of PLS
- Further analyze robustness of metaparameter convergence algorithm

Acknowledgements:
Work supported by the National Science Foundation under grant No. CHE-1504417 (co-funded by CTS and E-Programs) and is gratefully acknowledged by the authors.