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
spierob2@isu.edu, kalijohn@isu.edu

Idaho State UNIVERSITY

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 situations can 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 robusticity by referencing 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) $\mathbf{y} = \mathbf{X}\mathbf{b}$
  - Requires only a single tuning parameter
    - d = Number of Latent Variables $\mathbf{b} = \mathbf{X}_d^{+}\mathbf{y}$

Four model updating methods

- All require two tuning parameters
  - d Latent Variables and λ value

Labeled Secondary

- Local Mean Centering (LMC)

$$\begin{pmatrix}\mathbf{y}_P\\ \lambda\mathbf{y}_S\end{pmatrix}=\begin{pmatrix}\mathbf{X}_P\\ \lambda\mathbf{X}_S\end{pmatrix}\mathbf{b}$$

- Feature Augmentation 2A (FA-2A)

$$\begin{pmatrix}\mathbf{y}_P\\ \lambda\mathbf{y}_S\end{pmatrix}=\begin{pmatrix}\mathbf{X}_P & \mathbf{0}\\ \mathbf{0} & \mathbf{X}_P\\ \lambda\mathbf{X}_S & \lambda\mathbf{X}_S\end{pmatrix}\begin{pmatrix}\mathbf{b}_P\\ \mathbf{b}_S\end{pmatrix}$$

Unlabeled Secondary

Null Augmentation Regression (NAR)

$$\begin{pmatrix}\mathbf{y}_P\\ \mathbf{0}\end{pmatrix}=\begin{pmatrix}\mathbf{X}_P\\ \lambda\mathbf{R}\end{pmatrix}\mathbf{b}$$

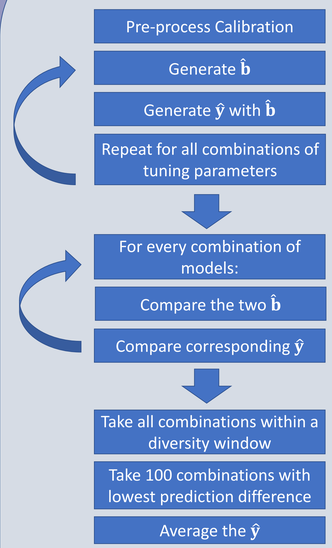
- NAR-Centroid (NAR-C) $\mathbf{R} = \left(\boldsymbol{\mu}_P - \boldsymbol{\mu}_S\right)^{\mathrm{T}}$
- NAR-Diagonal (NAR-D) $\mathbf{R} = diag\left(\boldsymbol{\mu}_P - \boldsymbol{\mu}_S\right)$

Validating Results:

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

$$RMSEV = \sqrt{\frac{\sum_{n=1}^{m}\left(y_n - y_n\right)^2}{m}}$$

Methodology



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

Gathering Statistics: Analyzing every combination of two models and the corresponding predictions to get their similarity measures

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

Figure 1. Flowchart for model selection using model diversity and prediction similarity measures

Similarity Measures

Model Similarity
Cosine of the angle between the i^{th} and j^{th} models

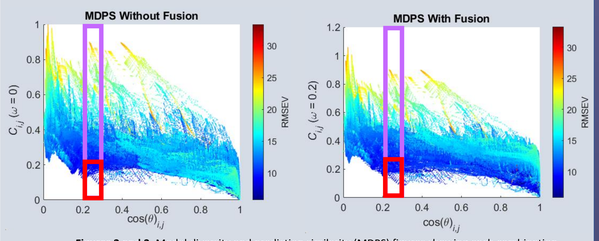
$$\cos\left(\theta\right)_{i,j} = \frac{\left(\mathbf{b}_i\right)^{\mathrm{T}}\left(\mathbf{b}_j\right)}{\left\|\mathbf{b}_i\right\|\left\|\mathbf{b}_j\right\|}$$

Prediction Similarity
Analyte prediction differences of the i^{th} and j^{th} models relative to the entire secondary spectra to obtain secondary prediction difference (SPD)

$$SPD_{i,j} = \sum_{n=1}^{m}\left|y_{n,i} - y_{n,j}\right|$$

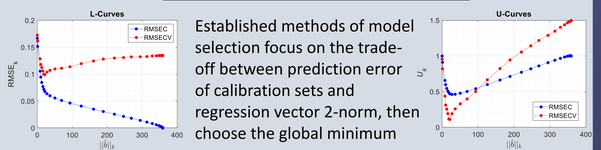
Range-Scaled Weighted Fusion (ω)
Weighting on regression vector 2-norm to prevent overfitting

$$C_{i,j} = \frac{SPD_{i,j} - \max\left(SPD\right)}{\max\left(SPD\right) - \min\left(SPD\right)} + \frac{\omega\left(\left\|\mathbf{b}\right\|_{i,j} - \max\left(\left\|\mathbf{b}\right\|\right)\right)}{\max\left(\left\|\mathbf{b}\right\|\right) - \min\left(\left\|\mathbf{b}\right\|\right)}$$



Figures 2 and 3. Model diversity and prediction similarity (MDPS) figures showing each combination of models generated by LMC organized by Cosine and SPD, with SPD appended with no fusion and 0.2 weighted fusion, respectively, for Figure 2 and Figure 3. The purple box indicates window of cosine selected, and the red box shows the lowest models ranked by SPD that are chosen

Traditional Model Selection



Established methods of model selection focus on the trade-off between prediction error of calibration sets and regression vector 2-norm, then choose the global minimum

Figure 4. Trade-off between RMSEC, RMSECV, and 2-norm as number of latent variables changes in PLS. U-curve figure demonstrates automatic model selection for PLS.

$$U_k = \frac{RMSE_k - \max\left(RMSE\right)}{\max\left(RMSE\right) - \min\left(RMSE\right)} + \frac{\left\|\mathbf{b}\right\|_k - \max\left(\left\|\mathbf{b}\right\|\right)}{\max\left(\left\|\mathbf{b}\right\|\right) - \min\left(\left\|\mathbf{b}\right\|\right)}$$

Data Description

Corn dataset: 700 NIR wavelength absorbances with four analyte values, moisture, oil, protein, and starch for 80 samples of corn measured on three instruments: m5, mp5, mp6. 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 BBar (Montana). Global is analyzed as primary, with 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

Division of samples for updating

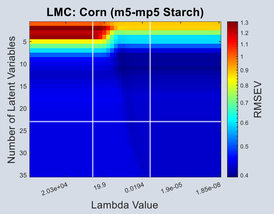
	Primary	Secondary	Validation
Corn	40	5	20
Tablet	60	6	24
Soil	4184	10	22

Table 1. Primary and secondary sample sizes for model updating. Secondary and Validation are combined for NAR.

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



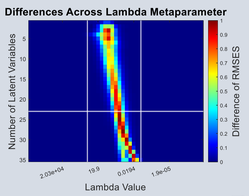


Figure 5 (above). Heatmap of RMSEV for all models generated by LMC with white lines indicating the truncation of tuning parameter ranges after convergence is assured

Figure 6 (left). Heatmap of successive differences of RMSES showing generation of the lambda convergence range

Results

Single Parameter: PLS Latent Variable

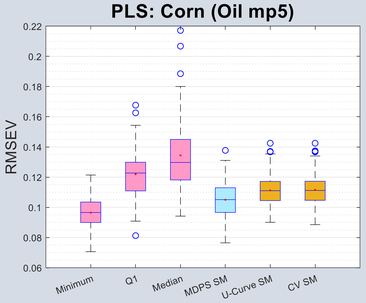


Figure 7. RMSEV of selected models for the novel MDPS (blue) against each quartile of total models generated (pink) and against the older methods of fusion U-curves and cross-validation U-curves for PLS (orange)

Multiple Parameter: Model Updating

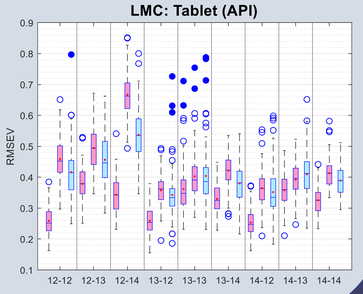


Figure 8. RMSEV of selected models using MDPS (blue) in every primary/secondary combination in Tablet, compared against the minimum and first quartile of all models generated (pink)

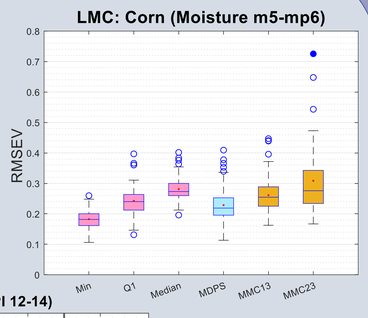


Figure 9. Boxplot of RMSEV for models selected out of models generated in LMC, using MDPS, against the traditional method of multiparameter model selection using U-curve sum raw fusion merits

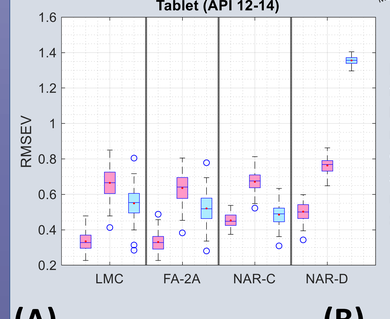


Figure 10. Boxplot of RMSEV for models selected using MDPS (blue) against minimum and first quartile of all models possible to be selected (pink) across each model updating method (LMC, FA-2A, NAR-C, NAR-D).

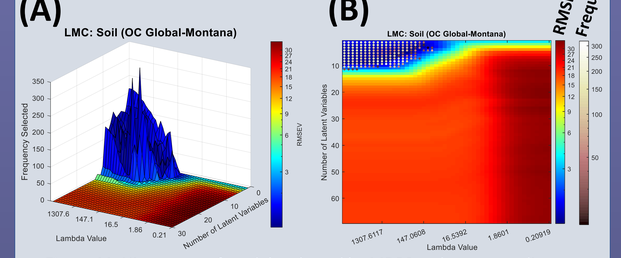


Figure 11. Histograms of models selected by MDPS and corresponding RMSEV relative to both metaparameters. (**A**) is color-coded to RMSEV and shows frequency on the z-axis. (**B**) is color-coded RMSEV for the overall image, and each of the circles is color coded to the frequency that corresponding model is selected

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 consistently selects models with low RMSEV
  - 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 robusticity of metaparameter convergence algorithm

Acknowledgements:

Work supported by the National Science Foundation under grant No. CHE-1506417 (co-funded by CDS and E Programs) and is gratefully acknowledged by the authors.