REGULARIZATION ADAPTION PROCESSES FOR MULTIVARIATE CALIBRATION MAINTENANCE

Anit Gurung[a], John H. Kalivas[a], Erik Andries[b]

[a]*Department of Chemistry, Idaho State University, 921 S. 8th Avenue, Pocatello, ID 83209, USA*
[b]*Center for Advanced Research Computing, University of New Mexico, Albuquerque, NM 87106, USA*
[b]*Department of Mathematics, Central New Mexico Community College, Albuquerque, NM 87106, USA*
[a]*guruanit@isu.edu*, [a]*kalijohn@isu.edu*

Abstract

In the field of chemometrics, an important issue in multivariate calibration is model updating. Model updating is the adaption process in which a model obtained for a given set of samples and measurement conditions (primary) is updated to predict the analyte in new samples and measurement conditions (secondary). The calibration method partial least squares is applied with two new updating approaches. In one approach, only one updated model is obtained to predict the analyte amount in both primary and secondary conditions. The other approach forms two updated models in which one model is used to predict in primary conditions and second model based on the first model is used to predict in secondary conditions. Both approaches are evaluated with near-infrared spectral datasets. Datasets include spectra of soil, corn, olive oil adulterated with sunflower and pharmaceutical tablets. Fusion process and single merits are used to select models. Model selection methods are evaluated based on prediction errors using selected models.

Objective

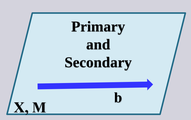
- Develop a new effective modal updating approach.

Model Updating Approach

Partial Least Squares (PLS)

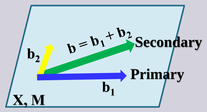
1b-PLS: 1 updating model

$$\begin{pmatrix} \mathbf{y} \\ \lambda \mathbf{y}_M \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \lambda \mathbf{M} \end{pmatrix} \mathbf{b}$$



2b-PLS : 2 updating models

$$\begin{pmatrix} \mathbf{y} \\ \lambda \mathbf{y}_M \end{pmatrix} = \begin{pmatrix} \mathbf{X} & \mathbf{0} \\ \lambda \mathbf{M} & \lambda \mathbf{M} \end{pmatrix} \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix}$$



- **X :** calibration samples in primary condition.
- **M :** calibration samples in secondary condition weighted with λ values.
- **y** and $\mathbf{y}_M$ **:** actual analyte concentration.
- $\hat{\mathbf{y}}$ and $\hat{\mathbf{y}}_M$ **:** analyte concentration prediction.
- $\hat{\mathbf{b}}$, $\hat{\mathbf{b}}_1$ **and** $\hat{\mathbf{b}}_2$ **:** estimated model regression vectors.
- Validation samples in secondary condition are considered to validate the model's accuracy and precision.

Model Measures

- **Bias :**
  - R^2, Slope (m), y-intercept (c)
    - Secondary Calibration (M)
    - Secondary Validation (V)
  - Root Mean Square Error (RMSE)
    - RMSEM, RMSEV

$$\mathbf{y}_i = m_i \mathbf{y}_i + c_i$$

i^{th} model

$$RMSE_i = \sqrt{\frac{\sum_{k=1}^{n} \left(y_k - y_k\right)^2}{n}}$$

- **Variance :**
  - Euclidean 2-norm ($\|\hat{\mathbf{b}}\|_2$)
  - Jaggedness (J)

$$J_i = \sum_{j=2}^{n} \sqrt{\left(\hat{b}_{ij} - \hat{b}_{ij-1}\right)^2}$$

j^{th} value in regression vector of a model

- **U-Curves :**

Bias-variance trade-off

$$M1_i = \left(\frac{\|\mathbf{b}\|_i - \|\mathbf{b}\|_{min}}{\|\mathbf{b}\|_{max} - \|\mathbf{b}\|_{min}}\right) + \left(\frac{\mathbf{RMSEM}_i - \mathbf{RMSEM}_{min}}{\mathbf{RMSEM}_{max} - \mathbf{RMSEM}_{min}}\right)$$

$$M2_i = \left(\frac{\mathbf{J}_i - \mathbf{J}_{min}}{\mathbf{J}_{max} - \mathbf{J}_{min}}\right) + \left(\frac{\mathbf{RMSEM}_i - \mathbf{RMSEM}_{min}}{\mathbf{RMSEM}_{max} - \mathbf{RMSEM}_{min}}\right)$$

- M1
- M2

Data Centering

- **Local mean centering**
  - **X, M,** y and $\mathbf{y}_M$ are centered across samples to respective means.
  - Mean of **M** and **y** are used to center validation samples in secondary condition.

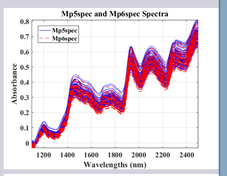
Model Selection

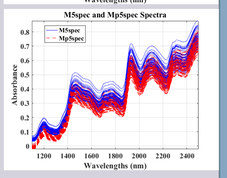
- There are large number of models with a unique combination of tuning parameter (λ) and latent variables (LV) for each.
- **Fusion Process**
  - Total of **7 Model Measures** :
    - RMSEM, R^2 M, Slope M, $\|\hat{\mathbf{b}}\|_2$, J, M1 and M2
  - **Sum Fusion** (**SF**) and **Median Fusion** (**MF**) are used to select models.

Experimental Design & Results

Corn Data

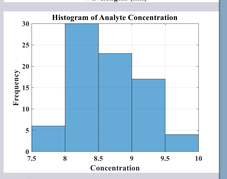
- 80 samples
- 3 different instruments :
  M5spec
  Mp5spec
  Mp6spec
- Analyte : **Protein**
- 350 wavelengths (1100 nm – 2500 nm)
- First 30 samples (constant) are used for primary calibration.
- 1000 random cross validation splits on remaining 50 samples used for secondary condition
  - Calibration set : 20 samples
  - Validation set: 30 samples
- 1 through 30 LV's





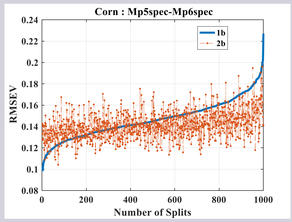
Set 1 (Mp5spec – Mp6spec)
- Primary: **Mp5spec**
- Secondary: **Mp6spec**
- 60 λ's ranging from 1000 to 0.0171

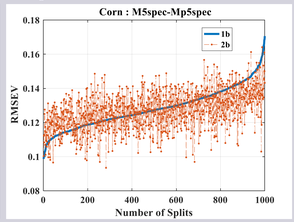
Set 2 (M5spec – Mp5spec)
- Primary: **M5spec**
- Secondary: **Mp5spec**
- 60 λ's ranging from 10 to 0.0027



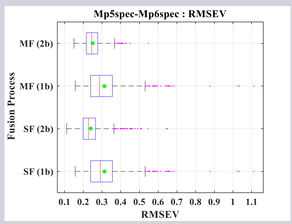
Set 1 : Results

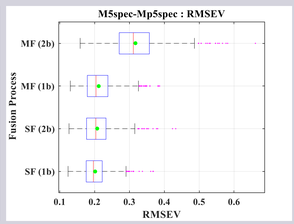
Result 1-1 : *Model with minimum RMSEV plot sorted based on* ***1b-PLS***

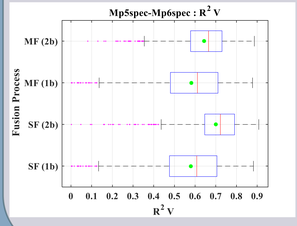


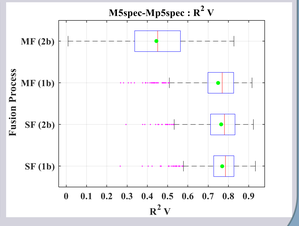


Result 1-2 : *Boxplot of Selected Models (Bias Measures)* - Mean

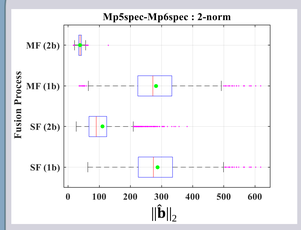


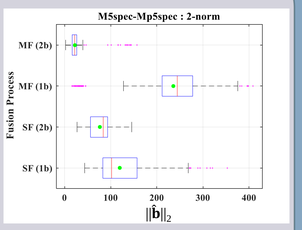






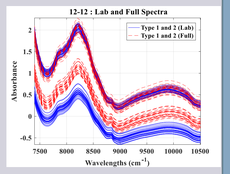
Result 1-3 : *Boxplot of Selected Models (Variance Measure)*

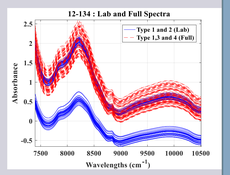




Pharmaceutical Tablet Data

- 310 Escitolopram tablets
  Type 1 (90 mg)
  Type 2 (125 mg),
  Type 3 (188 mg)
  Type 4 (250 mg)



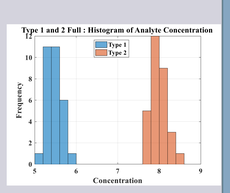


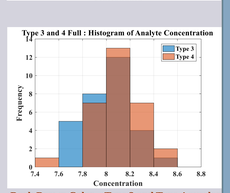
- Each tablet types are produced in three batches.
  - **Laboratory** (30 samples)
  - **Pilot** (10 samples)
  - **Full** (30 samples)
- Analyte : **Active Pharmaceutical Ingredient (API) content**
- **Set 1 (12-12)** and **Set 2 (12-134)**
  - Primary : Laboratory (Lab)
    - **Type 1**
    - **Type 2**
  - All 60 samples in total (30 each)
- 2 different tablet type combination produced in **Industry (Full)** are used for secondary condition.

Set 1 (12-12)
- Secondary : Full
  - **Type 1**
  - **Type 2**

Set 2 (12-134)
- Secondary : Full
  - **Type 1**
  - **Type 3**
  - **Type 4**

- 1000 random cross validation splits are performed for secondary condition
  - **Set 1**
    - Calibration set : 10 samples (5 from each)
    - Validation set: 30 samples (15 from each)
  - **Set 2**
    - Calibration set : 15 samples (5 from each)
    - Validation set : 45 samples (15 from each)
- 404 wavelengths (4000 cm^{-1} – 14000 cm^{-1}) (700 nm – 2500 nm)
- 1 through 30 LV's
- 60 λ's ranging from 1000 to 0.0673.

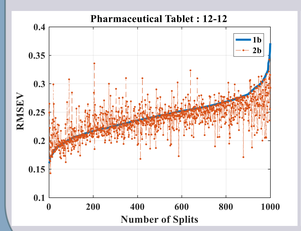


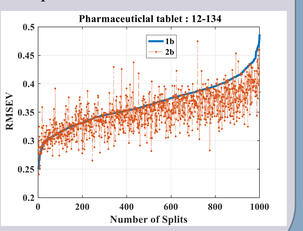


Dark Brown Color – Type 3 and Type 4 overlap

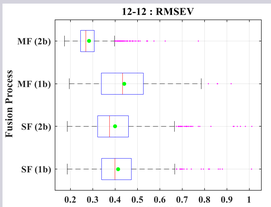
Set 1 : Results

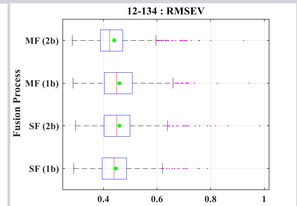
Results 2-1 : *Model with minimum RMSEV plot sorted based on* ***1b-PLS***

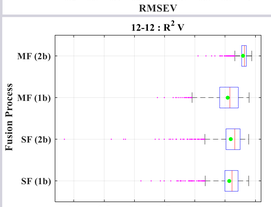


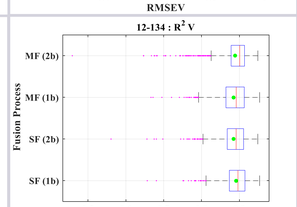


Result 2-2 : *Boxplot of Selected Models (Bias Measures)* - Mean

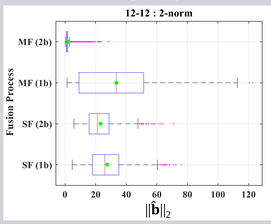


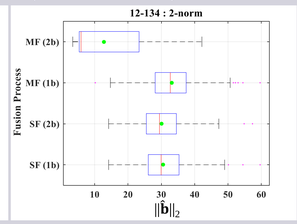






Result 2-3 : *Boxplot of Selected Models (Variance Measure)*





Observations

- **Result 1-1 and 2-1 :**
  - Cross validation splits where the prediction error is large using **1b-PLS**, generally **2b-PLS** lowers the prediction error significantly.
- **Result 1-2, 1-3, 2-2 and 2-3 :**
  - Both approaches have similar low prediction error for selected models
    - Corn (m5spec-mp5spec), Tablet (12-134)
    - Bias-Variance trade off is similar.
  - **2b-PLS** has lower prediction error of selected models than **1b-PLS** with lower 2-norm.
    - Bias-Variance trade off is better in **2b-PLS** than **1b-PLS.**
    - Corn (mp5spec-mp6spec), and Tablet (12-12)

Conclusions

- For modal updating, it seems to be clear that
  - **2b-PLS** works better than **1b-PLS** on the dataset where the differences in spectra of samples in primary and secondary conditions are minimal.
  - Both **1b-PLS** and **2b-PLS** give similar results when spectra are unique to each other like change in intensity.
    - In this situation, it suggests to look for different method for the approach using two updating models.
- Regarding the model selection method, based on the consistency of selecting the better model across most of the data, sum fusion seems to work best in picking the models.
- But, median fusion can not be disregarded as it can pick better models than sum fusion in some datasets.

Future Work

- Apply model updating approaches with unlabeled data.
  - Unlabeled data does not have analyte concentration values of samples.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. CHE-1506417 (co-funded by CDS&E Program) and is gratefully acknowledged by the authors.