Fusion of Similarity Measures to Characterize Sample Matrix Effects

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

Idaho State UNIVERSITY

Abstract

Multivariate calibration applied to spectroscopic data is firmly rooted in the field of analytical chemistry. Over the past several decades, numerous methods have been developed to deduce a calibration model to predict new analyte values with sufficient accuracy and precision. These calibration models produce good results when calibration (primary) and new prediction (secondary) samples are measured under similar conditions. However, inherent sample matrix effects and measurement conditions for the secondary samples are often dissimilar to calibration samples resulting in inaccurate and imprecise predictions. To combat this issue, calibration maintenance by model updating can be used to manipulate the calibration model to adapt to the secondary conditions.

Currently, evaluations of traditional and new calibration maintenance methods by researchers are performed without any consideration for the degree of difference between the primary and secondary data sets. Needed is a method that assesses the degree of difference between primary and secondary data sets for a robust evaluation of any model updating method. In order to solve this problem, multiple similarity measures are utilized in this presentation for a fusion consensus assessment of the degree of difference between the primary and secondary spectra assuming equal distributions of analyte values. Results will be shown for spectral data sets of varying similarity.

Objective

- Characterize the similarity between two data sets with the same prediction property using 15 similarity measures.
- Validate the method of using similarity measures by correlating the projected similarity to the relative prediction error between data sets.

Approach

Part 1: Calculating Indicator of Spectral Uniqueness (ISU)

Table 1. Similarity measure values for a single sample at a single eigenvector window.

$\mathbf{X}_p$	$\mathbf{X}_s$
d_{1p}	d_{1s}
d_{2p}	d_{2s}
$\vdots$	$\vdots$
d_{np}	d_{ns}

- For a single sample removed from secondary ($\mathbf{X}_s$), all similarity measures are calculated with respect to primary ($\mathbf{X}_p$) and the remaining sample spectra in $\mathbf{X}_s$
- d_{ip} and d_{is} represent the ith primary and secondary similarity measures respectively, where $1 \le i \le n$ (integer only) and n is the number of similarity measures

Table 2. Scaled similarity measure values for a sample at a single eigenvector window.

$\mathbf{X}_p$	$\mathbf{X}_s$
d'_{1p}	d'_{1s}
d'_{2p}	d'_{2s}
$\vdots$	$\vdots$
d'_{np}	d'_{ns}
$\frac{\sum_{i=1}^{n} s'_{ip}}{n}$	$\frac{\sum_{i=1}^{n} s'_{is}}{n}$

- Similarity measures scaled using the equations below
  $$d'_{is} = \frac{d_{is}}{d_{ip} + d_{is}} \qquad d'_{ip} = \frac{d_{ip}}{d_{ip} + d_{is}}$$
  where $0 \le d'_{is}, d'_{ip} \le 1$, and $d'_{is} + d'_{ip} = 1$
- Similarity between the removed sample and the respective space increases as the similarity measure value approaches 0
- Average similarity measure value is then calculated with respect to both spaces

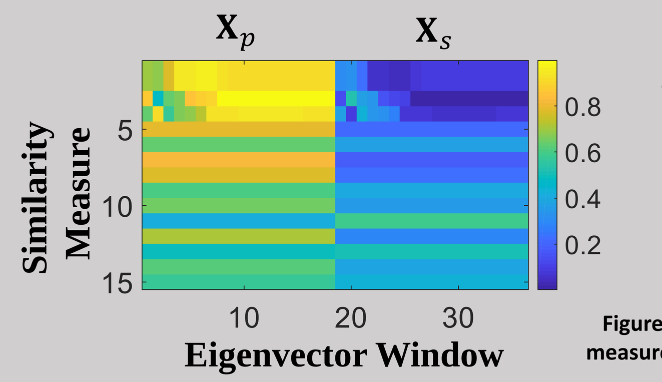


Figure 1. Image of scaled similarity measures for a single sample at 1-rank eigenvector windows.

- Characterizes sample similarity to the primary and secondary data sets respectively

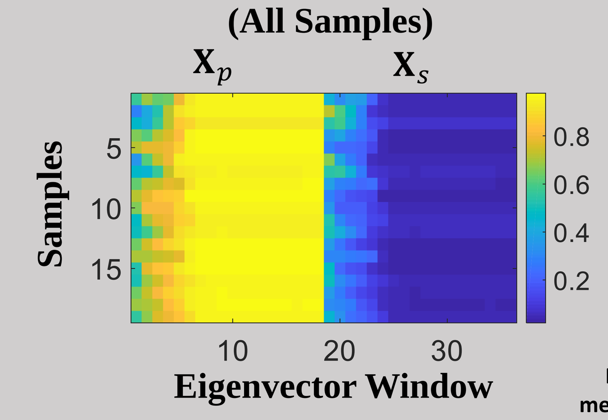


Figure 2. Image of average similarity measure value for every sample in $\mathbf{X}_s$ (21) at 1-rank eigenvector windows.

- Procedure repeated for all samples in $\mathbf{X}_s$
- Average value across all samples calculated

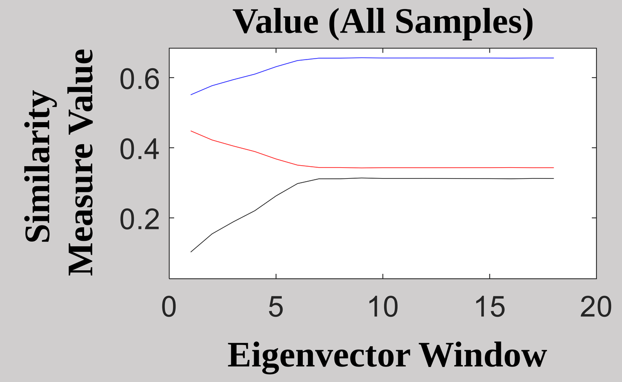
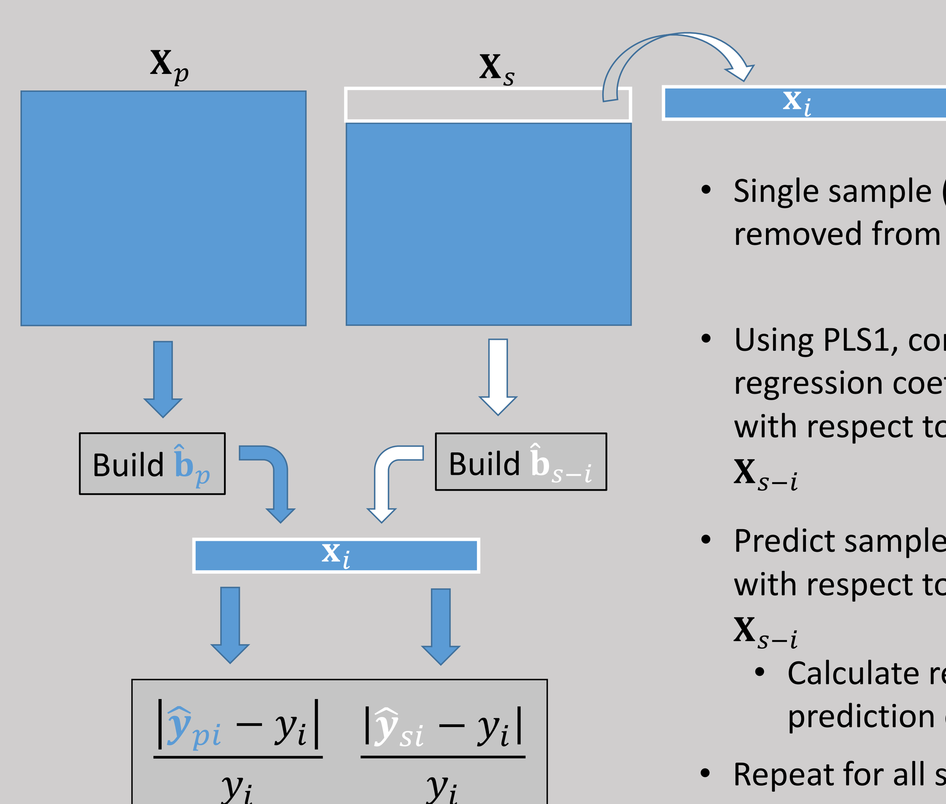


Figure 3. Average similarity measure value across all samples at 1-rank eigenvector windows.

- ISU calculated by subtracting average similarity measure value with respect to secondary from average similarity measure value with respect to primary
- Subtraction performed at last eigenvector window
  - Negates need for model selection

Part 2: Correlating ISU with Relative Prediction Error



- Single sample ($\mathbf{x}_i$) removed from $\mathbf{X}_s$
- Using PLS1, construct regression coefficients with respect to $\mathbf{X}_p$ and $\mathbf{X}_{s-i}$
- Predict sample out with respect to $\mathbf{X}_p$ and $\mathbf{X}_{s-i}$
  - Calculate relative prediction errors
- Repeat for all samples in $\mathbf{X}_s$

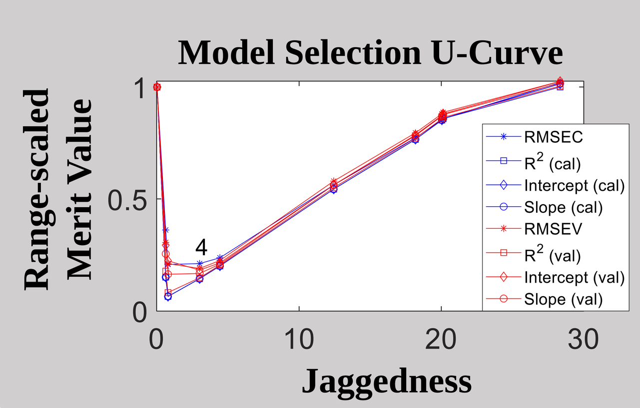


Figure 4. U-curve with all model selection merits at all eigenvector windows

- Select an optimal latent variable (LV) model for $\mathbf{X}_p$
- Consensus selection is 4 LV model

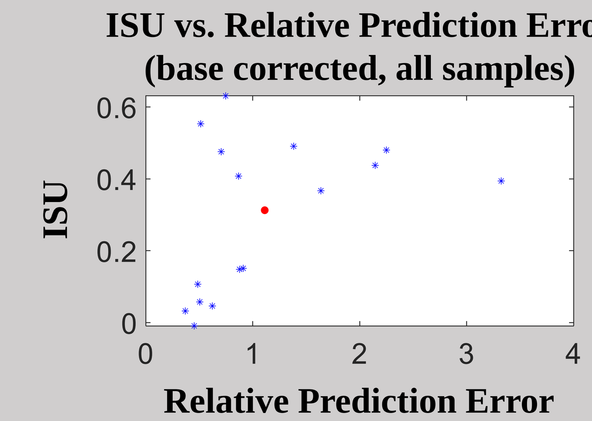


Figure 5. Correlation of relative prediction error to ISU value for all samples / overall average correlation

- Average relative prediction error (base corrected) calculated at selected LV with respect to $\mathbf{X}_p$ and $\mathbf{X}_{s-i}$

$$\text{Average Relative Prediction Error (base corrected)} = \frac{\sum_{i=1}^{m_p} \frac{|\hat{y}_{pi} - y_i|}{y_i}}{m_p} - \frac{\sum_{i=1}^{m_s} \frac{|\hat{y}_{si} - y_i|}{y_i}}{m_s}$$

- m_p and m_s denote the number of samples in primary and secondary respectively

Data Set

Temperature
19 samples containing mixtures of ethanol, water, and isopropanol measured at 5 different temperatures via NIR spectrometer.

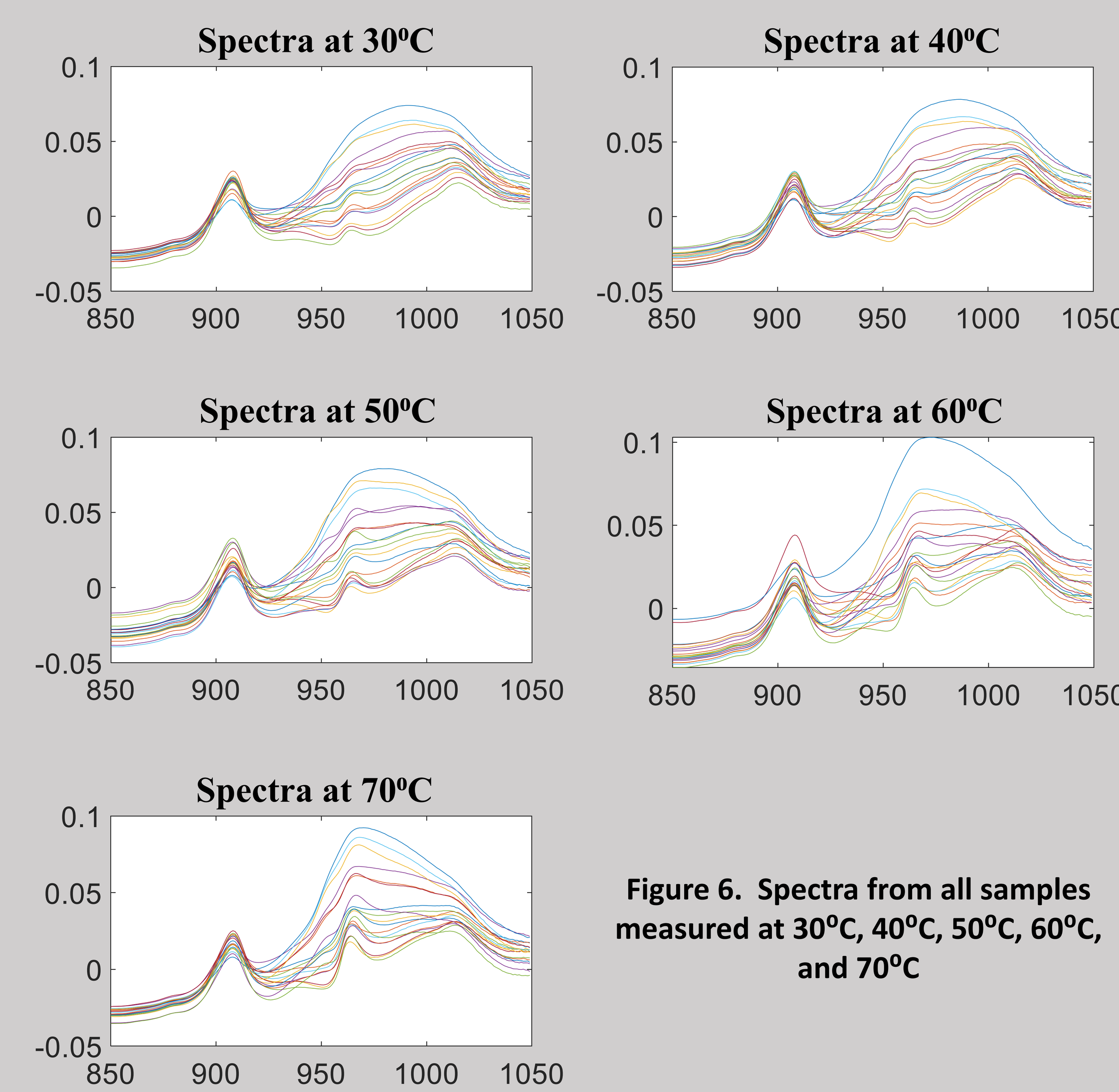


Figure 6. Spectra from all samples measured at 30°C, 40°C, 50°C, 60°C, and 70°C

Results

Temperature Data Set Results

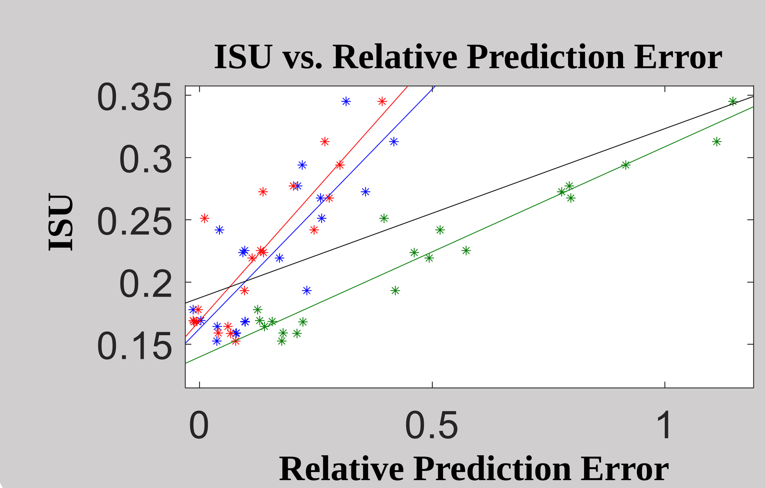


Figure 7. ISU correlation with relative prediction error for all prediction properties in Temperature

Table 3. ISU vs. relative prediction error correlation values for Temperature data set

Prediction Property	R^2	Intercept
Isopropanol	0.7383	0.1688
Water	0.6565	0.1625
Ethanol	0.9351	0.1398
Composite	0.4078	0.1872

- Separate trends for each prediction property
- Indicator that analyte information must be accounted for

Conclusion / Future Work

- ISU criterion is effective at assessing similarity between data sets
  - ISU correlation to prediction error is analyte dependent
    - Account for by including **y** measures
  - Add more **X** similarity measures
  - Evaluate preprocessing methods

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.

Math Appendix / Similarity Measures

Table 4. Vector-to-space similarity measures with corresponding equations (require a tuning parameter window).

Mahalanobis Distance	Q-Residual	Sinθ	Divergence Criterion
$\tilde{\mathbf{C}} = \frac{\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}}}{n}$ $\tilde{\mathbf{C}} = \mathbf{USV}^{\mathrm{T}}$ $\tilde{\mathbf{C}}_k^{+} = \mathbf{U}_k\mathbf{S}_k^{-1}\mathbf{V}_k^{\mathrm{T}}$ $MD_i = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}}\tilde{\mathbf{C}}_k^{+}(\mathbf{x}_i - \bar{\mathbf{x}})}$	$\mathbf{x}_i^{\perp} = (\mathbf{I} - \mathbf{V}_k\mathbf{V}_k^{\mathrm{T}})\mathbf{x}_i$ $\mathbf{X} = \mathbf{USV}^{\mathrm{T}}$ $Q_i = \lVert\mathbf{x}_i^{\perp}\rVert$	$\sin\theta_i = \frac{\lVert\mathbf{x}_i^{\perp}\rVert}{\lVert\mathbf{x}_i\rVert}$	$DC_i = \left\lvert\frac{1}{2}tr((\mathbf{X}_i - \mathbf{C})(\mathbf{X}_i^{+} - \mathbf{C}_k^{+}))\right\rvert + \left\lvert\frac{1}{2}tr((\mathbf{X}_i^{+} - \mathbf{C}_k^{+})(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}})\right\rvert$

Table 5a. Vector-to-vector similarity measures with corresponding equations.

Procrustes Analysis (unconstrained)	Procrustes Analysis (Constrained)	Extended Inverted Signal Correction Difference
$\bar{\mathbf{X}} = \rho\bar{\mathbf{X}}\mathbf{H}$, $\bar{\mathbf{X}} = \rho_i\mathbf{X}_i\mathbf{H}_i$ $\bar{\mathbf{X}}^{\mathrm{T}}\bar{\mathbf{X}} = \mathbf{USV}^{\mathrm{T}}$, $\mathbf{X}_i^{\mathrm{T}}\bar{\mathbf{X}} = \mathbf{U}_i\mathbf{S}_i\mathbf{V}_i^{\mathrm{T}}$ $\mathbf{H} = \mathbf{UV}^{\mathrm{T}}$, $\mathbf{H}_i = \mathbf{U}_i\mathbf{V}_i^{\mathrm{T}}$ $\rho = \frac{tr(\mathbf{S})}{tr(\bar{\mathbf{X}}\bar{\mathbf{X}}^{\mathrm{T}})}$, $\rho_i = \frac{tr(\mathbf{S}_i)}{tr(\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}})}$	$\bar{\mathbf{X}} = \mathbf{X}_i\mathbf{T}_i$ $\bar{\mathbf{X}} = \bar{\mathbf{X}}\mathbf{T}$ $\mathbf{T}_i = \mathbf{X}_i^{+}\bar{\mathbf{X}}$ $\mathbf{T} = \bar{\mathbf{X}}^{+}\bar{\mathbf{X}}$	$\bar{\mathbf{x}} = \mathbf{x}_i + \mathbf{X}_c\mathbf{b}$ $\mathbf{d} = \bar{\mathbf{x}} - \mathbf{x}_i = \mathbf{X}_c\mathbf{b}$ $\mathbf{X}_c = \left(1, \mathbf{x}_i, \mathbf{x}_i^2, \frac{d}{d\lambda}\mathbf{x}_i, \frac{d^2}{d\lambda^2}\mathbf{x}_i, \boldsymbol{\lambda}, \boldsymbol{\lambda}^2, \ln(\boldsymbol{\lambda})\right)$
$\Delta H_i = \lVert\mathbf{H}_i - \mathbf{H}\rVert_F$ $\Delta\rho_i = \lvert\rho_i - \rho\rvert$	$\Delta T_i = \lVert\mathbf{T}_i - \mathbf{T}\rVert_F$	$\mathrm{EISCD}_{\mathbf{b}} = \lVert\mathbf{b}\rVert$ $\mathrm{EISCD}_{\mathbf{X}_c\mathbf{b}} = \lVert\mathbf{X}_c\mathbf{b}\rVert$

Table 5b. Vector-to-vector similarity measures with corresponding equations.

Determinant	Inner Product Correlation	Euclidian Distance	$1 - \cos\theta$
$\det_i = Det\left(\begin{bmatrix}\mathbf{x}_i^{\mathrm{T}} \\ \bar{\mathbf{x}}^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\mathbf{x}_i & \bar{\mathbf{x}}\end{bmatrix}\right) = \left(\lVert\mathbf{x}_i\rVert\lVert\bar{\mathbf{x}}\rVert\sin\theta_i\right)^2$	$1 - r_i = 1 - \frac{tr(\mathbf{X}_i^{\mathrm{T}}\bar{\mathbf{X}})}{\sqrt{tr(\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i)tr(\bar{\mathbf{X}}^{\mathrm{T}}\bar{\mathbf{X}})}}$	$d_i = \lVert\mathbf{x}_i - \bar{\mathbf{x}}\rVert$	$1 - \cos\theta_i = 1 - \frac{\mathbf{x}_i^{\mathrm{T}}\bar{\mathbf{x}}}{\lVert\mathbf{x}_i\rVert\lVert\bar{\mathbf{x}}\rVert}$

Notation for **Table 1, 2a, 2b**

- k subscript denotes the arbitrary number of eigenvectors or latent variables selected
- $\bar{\mathbf{x}}$ is the column-wise mean vector of $\mathbf{X}$
- Outer product arrays $\bar{\mathbf{X}}$ and $\mathbf{X}_i$ are computed by $\bar{\mathbf{X}} = \bar{\mathbf{x}}\bar{\mathbf{x}}^{\mathrm{T}}$ and $\mathbf{X}_i = \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}$
- $\lVert\ \rVert_F$ denotes the Frobenius norm
- $\boldsymbol{\lambda}$ is the vector of wavelengths
- Four EISCD similarity measures are created by swapping $\mathbf{x}_i$ and $\bar{\mathbf{x}}$