Use Authorization

In presenting this dissertation in partial fulfillment of the requirements for an advanced degree at Idaho State University, I agree that the Library shall make it freely available for inspection. I further state that permission to download and/or print my dissertation for scholarly purposes may be granted by the Dean of the Graduate School, Dean of my academic division, or by the University Librarian. It is understood that any copying or publication of this dissertation for financial gain shall not be allowed without my written permission.

Signature _____

Date _____

Local Adaptive Fusion Regression: Local Calibration with Matrix Matched Samples

by

Rachel Emerson

A thesis

submitted in partial fulfillment

of the requirements for the degree of

Master of Science in the Department of Chemistry

Idaho State University

Fall 2016

Committee Approval

To the Graduate Faculty:

The members of the committee appointed to examine the thesis of Rachel Emerson find it satisfactory and recommend that it be accepted.

_____
Dr. John Kalivas
Major Advisor

_____
Dr. Rene Rodriguez
Committee Member

_____
Dr. Shu-Chuan Chen
Graduate Faculty Representative

Acknowledgements

I would like to thank Dr. John Kalivas for helping me through this research project and continuing to push me to think about the impact of my research more in depth. I gained a completely new and valuable skillset through this research, and I will always be grateful for that.

I would also like to thank Dr. Rene Rodriguez for his valuable input to my research project. His comments and questions regarding my thesis were very helpful for adding a broader perspective to this research.

Finally, I would like to thank my husband for being supportive of me while pursuing my degree and working full-time. This accomplishment would not have been possible with his love, help, and encouragement.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CARNAC	comparison analysis using restructured near-infrared and constituent data
CRRN	comparison of ranks by random numbers
EISC	extended inverted scatter correction
FD	Freedman-Diaconis
IR	infrared
JIT	just in time
LAFR	local adaptive fusion regression
LV	latent variables
LWR	locally weighted regression
NIR	near infrared
NMR	nuclear magnetic resonance
OP	orthogonal projections to a regression vector
PA	Procrustes analysis
PC	principal components
PCA	principal component analysis
PLS	partial least squares
RMSEC	root mean square error of calibration
RMSECV	root mean square error of cross validation
RMSEV	root mean square error of validation
RR	ridge regression
SD	standard deviation
SRD	sum of ranking differences

SVD	singular value decomposition
UV-Vis	ultraviolet-visible

Abstract

Spectroscopic data paired with chemometric modeling methods has become a powerful, cost effective, and rapid analytical tool over the last decade. However, large global spectral libraries spanning numerous sample matrix differences and instrument conditions are often nonlinear in relation to the measured chemical prediction property of interest. These differences result in lower model prediction accuracies. One solution to overcoming nonlinear relationships is to use local modeling techniques. In local modeling, a unique subset of calibration samples are selected from a global library for each specific target sample. Many local modeling algorithms rely on one or two spectral similarity measures for selecting calibration samples while overlooking similarities based on chemical properties. Current local modeling methods also require predetermined selections of specific variables including similarity merits, number of samples, and regression model tuning parameters. This work explores techniques for selecting local calibration samples that are both spectrally and chemically similar to the target sample while reducing the number of predetermined variables required. The process of local adaptive fusion regression (LAFR) employs many unique aspects, including data fusion and cross modeling, to select matrix matched calibration samples. Aspects of the local adaptive fusion regression process are first used to demonstrate why data fusion and cross modeling techniques are successful for identifying matrix matched calibration sets. The automated LAFR process, using these same techniques, then demonstrates how matrix matched local calibration sets are consistently formed and selected.

Chapter 1: Introduction to Local Modeling

1. Multivariate Calibration Models

Multivariate calibration models using a number of spectral measurement techniques, such as infrared (IR)[1-4], near-infrared (NIR)[5-7], and ultraviolet-visible (UV-Vis)[8-10] spectroscopy, have been developed by many industries and institutions. These techniques are commonly used as rapid screening tools for analysis of specific chemical properties because they are typically rapid, non-destructive, and cost effective. Multivariate modeling techniques are often required to relate spectral regions to a chemical analyte based on a linear relationship (Eq. 1.1).

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e} \tag{1.1}$$

The analyte measurement for m samples is represented as $\boldsymbol{y}_{(m \text{ x } 1)}$, $\boldsymbol{X}_{(m \text{ x } n)}$ is the spectral data over n variables, $\boldsymbol{b}_{(n \text{ x } 1)}$ is the regression vector relating the spectra to the analyte, and $\boldsymbol{e}_{(m \text{ x } 1)}$ is the normally distributed error. To use a multivariate inverse linear regression for predicting the modeled analyte for new samples, $\boldsymbol{b}$ is estimated as

$$\widehat{\boldsymbol{b}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{1.2}$$

However, for equation 1.2 to be true $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ must exist. For many multivariate models where the number of variables (n) is greater than the number of samples (m) a biased estimate regression vector, $\widehat{\boldsymbol{b}}_{(n \text{ x } 1)}$, is necessary (Eq. 1.3).

$$\widehat{\boldsymbol{b}} = \boldsymbol{X}^{+}\boldsymbol{y} \tag{1.3}$$

In biased multivariate regression methods, $\boldsymbol{X}^{+}_{(m \text{ x } n)}$ is referred to as a pseudoinverse as $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ does not mathematically exist. Common biased regression methods used to estimate $\widehat{\boldsymbol{b}}$ include principal component regression (PCR), partial least squares (PLS), and ridge regression (RR)[11-12]. For these biased regression methods a number of possible

estimated regression vectors are formed depending on the basis vectors calculated through each regression algorithm. The selection of a number of basis vectors that have a good bias/variance trade-off is a challenge in biased regression methods[12].

The main disadvantage to using these multivariate analysis techniques is developing the initial calibration model to predict the specific chemical reference values. In order to build accurate calibration models, the reference values must be determined by an accurate primary method and the calibration set must contain a large amount of spectral variability over a large chemical range[13]. The use of large spectral libraries with previously measured reference values can be one solution to limit the number of primary analyses required for initial model development. Disadvantages to using large spectral libraries are potentially non-linear relationships, increases in prediction errors, and sample inhomogeneity[14].

2. Local Modeling

To use large libraries of spectra effectively for developing accurate predicting calibration models, local calibration methods can be applied to select calibration samples that have similar spectral and chemical properties for a specific target sample[15]. In local modeling approaches, a unique model is developed for each target sample based on various measures of similarity. A few of the well known methods of local modeling include comparison analysis using restructured near-infrared and constituent data (CARNAC)[16-17], a local algorithm known as LOCAL[15, 18], and locally weighted regression (LWR)[19-22]. Locally weighted regression (LWR) or modifications of LWR, often referred to as just in time (JIT) modeling, are used in recent local modeling literature[23-26].

The general steps involved in LWR processes include: (1) selection of relevant samples from the library based on a similarity criteria with the target sample; (2) building a local model using these relevant library samples; (3) and predicting the target sample with the local model. Many of the similarity criteria reported are based on spectral similarity measures such as distance measurements[27-30], angle comparisons, or a combination of both distance and angle[31]. When combining two measures of similarity for a similarity criteria, a trade-off parameter is used (Eq. 1.4)[31].

$$s_i = \gamma e^{-d_i} + (1 - \gamma) \cos \theta_i \tag{1.4}$$

The overall similarity criteria (s_i) is calculated by combining the distance (d_i) and the cosine of the angle ($\cos \theta_i$) between the target spectrum and library spectrum (i) with a trade-off parameter (γ). The trade-off parameter is set between 0 and 1. In this equation, the closer s_i is to 1 the more similar the library spectrum is to the target spectrum. The Euclidean distance measurement (d_i) can be calculated as

$$d_i = \sqrt{(\boldsymbol{x}_t - \boldsymbol{x}_i)^T (\boldsymbol{x}_t - \boldsymbol{x}_i)} \tag{1.5}$$

and the cosine of the angle ($\cos \theta_i$) calculated as

$$\cos \theta_i = \frac{|\boldsymbol{x}_t^T \boldsymbol{x}_i|}{\|\boldsymbol{x}_t\|_2 \|\boldsymbol{x}_i\|_2} \tag{1.6}$$

In equations 1.5 and 1.6, $\boldsymbol{x}_{t_{(n \times 1)}}$ is the target spectrum and $\boldsymbol{x}_{i_{(n \times 1)}}$ is one spectrum from the library. These methods using a distance, angle, or a combination of both to determine spectrally similar samples have two disadvantages. The first disadvantage is the selection of the trade-off parameter variable. The second disadvantage is the limitations of using only spectral matching information.

The trade-off parameter can greatly influence the samples selected for the local model; discussed in further detail in Chapter 3. Figure 1.1 illustrates a hypothetical

situation of distance and angle comparisons between two vectors. For this illustration if the similarity measurement in equation 1.4 had trade-off parameter of $\gamma = 1$, the vectors *a* and *b* would have equal similarity indices compared to the target vector, *t*. However, if the trade-off parameter was set to 0 vector *b* would be considered more similar to the target sample.

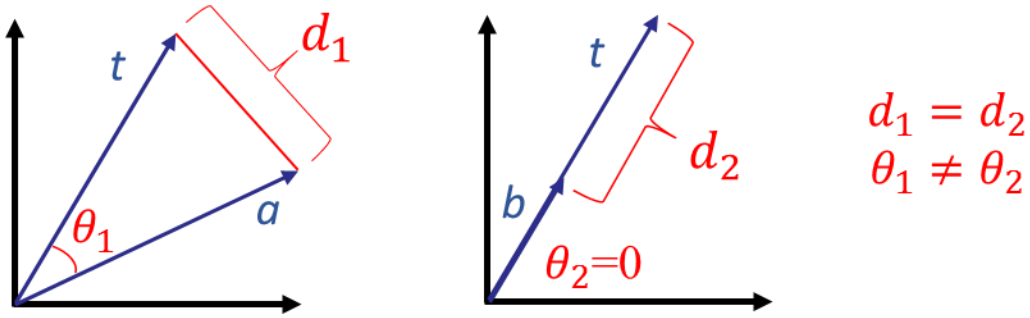


Figure 1.1. Illustration of angle, θ_i, and distance measurements, d_i, between the target vector (*t*) and vectors *a* and *b*.

One option to avoid selecting a trade-off parameter or selecting a single similarity measure is to use a wide variety of spectral similarity measures simultaneously. Equation 1.4 is one proposed method for the fusion of two similarity measures for the calculation of a single similarity index. However, there are other methods available for data fusion that are not limited by the number of measures included and do not require a trade-off parameter selection.

3. Data Fusion

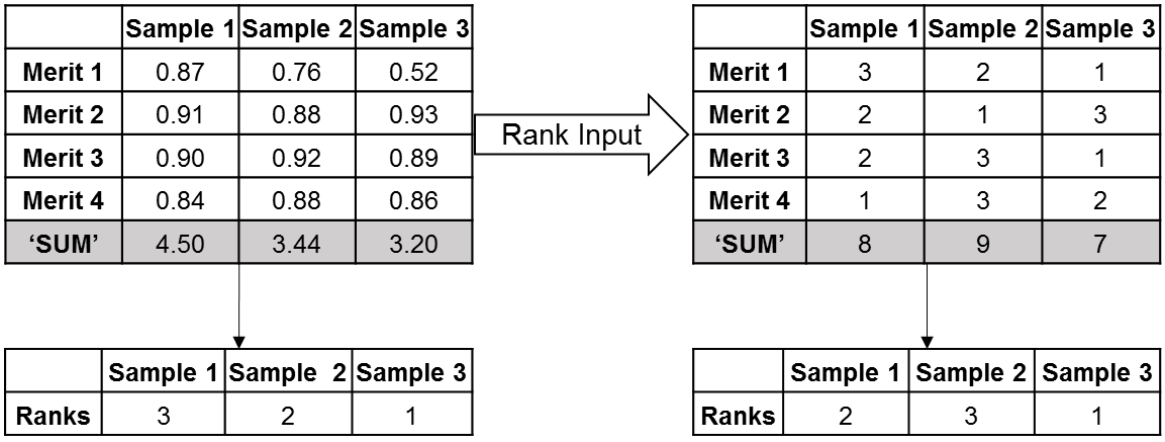
Data fusion is a common technique of combining multiple data inputs, such as distance or angle measurements between two spectra, into a single output that can be

more informative than the individual sources[32]. Data fusion has been used to combine different types of sensor data since the 1980s[33] and, more recently, has been shown as a method for combining similarity measures for identifying matching molecular structures from a database[34-35]. The benefit of data fusion is an increased level of confidence in selection, in this case, of a similar local calibration sample. If most of the similarity merits agree on the selection of one sample then the decision of selection becomes easier with the consensus of the merits. The use of data fusion methods also allows for a systematic ranking to determine the degree of similarity or dissimilarity across the samples being compared. Two methods for data fusion are discussed below.

3.1. *Fusion Rules*

Willet reviewed a number of different arithmetic based fusion methods[36]. For these fusion rules the different similarity merits for each sample being compared to a specific target sample are fused with an arithmetic function (e.g. maximum, minimum, or sum) resulting in a single value representing the overall similarity of the sample. These values are then used to rank the samples from most similar to least similar to a specified target sample. The comparison merit inputs used for the calculation of these functions can either be the raw comparison merit values or the rank comparison merit values. For the input of rank values, each individual similarity merit would be ranked from 1 to m for each of the m samples being compared. The fusion rules would then be applied to rank values of the all the similarity merits for a sample. Figure 1.2 shows how the sum fusion rule, which is the calculated sum of all the merits for each sample, is used for both the raw value input of similarity merits and the rank value inputs of the same similarity merits. For this example, the similarity merits with the highest values indicate similarity;

therefore, the highest ranks would also indicate the greatest similarity. If the raw similarity merits are used to calculate the ranks then Sample 3 is least similar to the target, and Sample 1 is most similar to the target. If the rank values of the similarity merits are used as inputs then Sample 3 is again least similar to the target sample and Sample 2 is most similar to the target sample.



	Sample 1	Sample 2	Sample 3
Merit 1	0.87	0.76	0.52
Merit 2	0.91	0.88	0.93
Merit 3	0.90	0.92	0.89
Merit 4	0.84	0.88	0.86
'SUM'	4.50	3.44	3.20

	Sample 1	Sample 2	Sample 3
Merit 1	3	2	1
Merit 2	2	1	3
Merit 3	2	3	1
Merit 4	1	3	2
'SUM'	8	9	7

	Sample 1	Sample 2	Sample 3
Ranks	3	2	1

	Sample 1	Sample 2	Sample 3
Ranks	2	3	1

Figure 1.2. Illustration of sum fusion rule for raw and rank inputs comparing three samples using four similarity merits.

3.2. *Sum of Ranking Differences*

Sum of ranking differences (SRD) is another data fusion method [37-38]. Sum of ranking differences has been shown in multiple analytical applications[37-40]. The SRD process uses a general algorithmic method for comparison and can be applied to many applications. The input data for SRD is a matrix with columns of variables for comparison and rows of objects used for the comparison. The SRD procedure involves ranking the input variables (columns) across the objects (rows) relative to a 'target' vector. The 'target' vector can be assigned as the minimum, maximum, mean, or median

for each corresponding row of objects. A hypothetical illustration is shown in Figure 1.3 for calculating SRD ranks. The first step shows the SRD input matrix of three samples compared using four similarity merits. The ‘target’ vector is the maximum value for each of the merits (highlighted in bold). In the case of these similarity measures the highest values would indicate similarity. The second step is to reorder the merits based on the target vector from minimum to maximum. The third step is to calculate the ranks from minimum to maximum for each of the variables and subtract the variable ranks from the target rank. The sum of these differences gives the sum of ranking differences. Sample 3 is the most similar to the target spectrum with a sum of ranking differences rank of 0.

	Sample 1	Sample 2	Sample 3	Target 'Max'
Merit 1	**0.87**	0.76	0.52	**0.87**
Merit 2	0.91	0.88	**0.93**	**0.93**
Merit 3	0.90	**0.92**	0.89	**0.92**
Merit 4	0.84	**0.89**	0.86	**0.89**

	Sample 1	Sample 2	Sample 3	Target 'Max'	Row Index
Merit 1	**0.87**	0.76	0.52	**0.87**	1
Merit 4	0.84	**0.89**	0.86	**0.88**	4
Merit 3	0.90	**0.92**	0.89	**0.92**	3
Merit 2	0.91	0.88	**0.93**	**0.93**	2

Row	Target 'Max'	Target rank	Sample1	Rank 1	Diff. 1	Sample 2	Rank 2	Diff. 2	Sample 3	Rank 3	Diff. 3
Merit 1	**0.87**	1	**0.87**	2	1	0.76	1	0	0.52	1	0
Merit 4	**0.88**	2	0.84	1	1	0.89	3	1	0.86	2	0
Merit 3	**0.92**	3	0.90	3	0	**0.92**	4	1	0.89	3	0
Merit 2	**0.93**	4	0.91	4	0	**0.88**	2	2	**0.93**	4	0
Sum					**2**			**4**			**0**

Figure 1.3. Illustration for calculating sum of ranking differences (SRD) ranks for three sample using four similarity merits.

The use of data fusion methods, such as fusion rules and sum of ranking differences, allows for a combination of spectral similarity merits, including distance and angle comparisons, to be used simultaneously without the need to select a trade-off parameter. However, selecting samples based on only spectral similarity is not ideal for local modeling.

4. Matrix Matching

Even with the advantages of data fusion to combine multiple types of spectral similarity merits, the information for the chemical data is still not taken into account. In

localization there can be situations where the range in the selected spectral matching samples is small but the range in the chemical information is large resulting in a poor local model (Fig. 1.4).

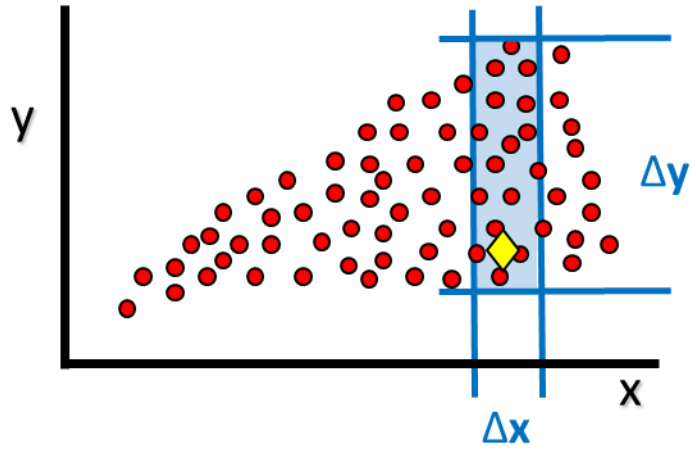


Figure 1.4. Illustration of localization around target sample (◊) based on spectral data (x) alone without considering changes in the chemical (y) ranges.

This chemical information includes the analyte component of interest along with all of the other species within the target and library samples. The total chemical profile of each sample represents the matrix of the samples. Both the spectral information and the different chemical profiles in a calibration sample and the target sample need to be considered for matrix matching of local calibration sets. The focus of this research is to show how matrix matched local calibration samples can be selected from global libraries of samples.

This work will demonstrate through the fusion of multiple similarity measures, including measures that assess chemical matching, a set of samples can be selected that

are not only spectrally matched but are chemically matched. Chapter 2 will focus on the definition and evaluation of matrix matching. This chapter will demonstrate how the fusion of proposed matrix matching similarity measures can identify known matrix matched calibration sets. Chapter 3 will focus on the incorporation of the matrix matching data fusion methods from Chapter 2 into the local adaptive fusion regression (LAFR) process. This chapter will demonstrate how the automated LAFR process mitigates and/or solves multiple identified challenges associated with current local modeling methods. The purpose of this chapter is to show methods for consistently forming and selecting local calibration sets with matrix matched samples for each of the target samples assessed.

5. References

1. Balabin, R. M.; Smirnov, S. V., Melamine detection by mid- and near-infrared (MIR/NIR) spectroscopy: A quick and sensitive method for dairy products analysis including liquid milk, infant formula, and milk powder. *Talanta* **2011,** *85* (1), 562-568.
2. Wang, L.; Mizaikoff, B., Application of multivariate data-analysis techniques to biomedical diagnostics based on mid-infrared spectroscopy. *Analytical and Bioanalytical Chemistry* **2008,** *391* (5), 1641-1654.
3. Rabenarivo, M.; Chapuis-Lardy, L.; Brunet, D.; Chotte, J.-L.; Rabeharisoa, L.; Barthès, B., Comparing near and mid-infrared reflectance spectroscopy for determining properties of Malagasy soils, using global or LOCAL calibration. *Journal of Near Infrared Spectroscopy* **2013,** *21* (6), 495-509.
4. Meza-Márquez, O. G.; Gallardo-Velázquez, T.; Osorio-Revilla, G., Application of mid-infrared spectroscopy with multivariate analysis and soft independent modeling of class analogies (SIMCA) for the detection of adulterants in minced beef. *Meat Science* **2010,** *86* (2), 511-519.
5. Aastveit, A. H.; Marum, P., Near-Infrared Reflectance Spectroscopy: Different Strategies for Local Calibrations in Analyses of Forage Quality. *Applied Spectroscopy* **1993,** *47* (4), 463-469.
6. Balabin, R. M.; Safieva, R. Z.; Lomakina, E. I., Near-infrared (NIR) spectroscopy for motor oil classification: From discriminant analysis to support vector machines. *Microchemical Journal* **2011,** *98* (1), 121-128.
7. Blanco, M.; Villarroya, I., NIR spectroscopy: a rapid-response analytical tool. *TrAC Trends in Analytical Chemistry* **2002,** *21* (4), 240-250.
8. Langergraber, G.; Fleischmann, N.; Hofstädter, F., A multivariate calibration procedure for UV/VIS spectrometric quantification of organic matter and nitrate in wastewater. *Water Science and Technology* **2003,** *47* (2), 63-71.
9. Viscarra Rossel, R. A.; Walvoort, D. J. J.; McBratney, A. B.; Janik, L. J.; Skjemstad, J. O., Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma* **2006,** *131* (1–2), 59-75.
10. Thanasoulias, N. C.; Piliouris, E. T.; Kotti, M.-S. E.; Evmiridis, N. P., Application of multivariate chemometrics in forensic soil discrimination based on the UV-Vis spectrum of the acid fraction of humus. *Forensic Science International* **2002,** *130* (2–3), 73-82.
11. Naes, T.; Isaksson, T.; Fearn, T.; Davies, T., *A User-Friendly Guid to Mutivariate Calibration and Classification*. NIR Publications: Chichester, 2002.
12. Kalivas, J. H., Comprehensive Chemometrics. In *Calibration Methodologies*, Brown, S.; Tauler, R.; R, W., Eds. Elsevier: Oxford, 2009; Vol. 3, pp 1-32.
13. Berzaghi, P.; Shenk, J.; Westerhaus, M., LOCAL prediction with near infrared multi-product databases. *Journal of Near Infrared Spectroscopy* **2000,** *8* (1), 1-9.
14. Sinnaeve, G.; Dardenne, P.; Agneessens, R., Global or local? A choice for NIR calibrations in analyses of forage quality. *Journal of Near Infrared Spectroscopy* **1994,** *2* (3), 163-175.
15. Berzahi, P.; Shenk, J.; Westerhaus, M., LOCAL prediction with near infrared multi-product databases. *Journal of Near Infrared Spectroscopy* **2000,** *8* (1), 1-9.

16. Davies, A. C.; Britcher, H.; Franklin, J.; Ring, S.; Grant, A.; McClure, W., The application of fourier-transformed near-infrared spectra to quantitative analysis by comparison of similarity indices (CARNAC). *Mikrochim Acta* **1988,** *94* (1-6), 61-64.
17. Davies, T.; Fearn, T., Local methods and CARNAC-D. In *Handbook of Near-Infrared Analysis*, 2008; pp 781-795.
18. Pérez-Marín, D.; Garrido-Varo, A.; Guerrero, J. E., Implementation of LOCAL Algorithm with Near-Infrared Spectroscopy for Compliance Assurance in Compound Feedingstuffs. *Applied Spectroscopy* **2005,** *59* (1), 69-77.
19. Atkeson, C. G.; Moore, A. W.; Schaal, S., Locally Weighted Learning. *Artif. Intell. Rev.* **1997,** *11* (1-5), 11-73.
20. Centner, V.; Massart, D. L., Optimization in Locally Weighted Regression. *Analytical Chemistry* **1998,** *70* (19), 4206-4211.
21. Chang, S.-Y.; Baughman, E. H.; McIntosh, B. C., Implementation of Locally Weighted Regression to Maintain Calibrations on FT-NIR Analyzers for Industrial Processes. *Applied Spectroscopy* **2001,** *55* (9), 1199-1206.
22. Cleveland, W. S.; Devlin, S. J., Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association* **1988,** *83* (403), 596-610.
23. He, K.; Cheng, H.; Du, W.; Qian, F., Online updating of NIR model and its industrial application via adaptive wavelength selection and local regression strategy. *Chemometrics and Intelligent Laboratory Systems* **2014,** *134*, 79-88.
24. Chen, K.; Ji, J.; Wang, H.; Liu, Y.; Song, Z., Adaptive local kernel-based learning for soft sensor modeling of nonlinear processes. *Chemical Engineering Research and Design* **2011,** *89* (10), 2117-2124.
25. Hazama, K.; Kano, M., Covariance-based locally weighted partial least squares for high-performance adaptive modeling. *Chemometrics and Intelligent Laboratory Systems* **2015,** *146*, 55-62.
26. Dahlbacka, J.; Lillhonga, T., Quantitative measurements of anaerobic digestion process parameters using near infrared spectroscopy and local calibration models. *Journal of Near Infrared Spectroscopy* **2013,** *21* (1), 23-33.
27. Naes, T.; Isaksson, T.; Kowalski, B. R., Locally Weighted Regression and Scatter Correction for Near-Infrared Reflectance Data. *Analytical Chemistry* **1990,** (62), 664-673.
28. Wang, Z.; Isaksson, T.; Kowalski, B. R., New approach for distance measurement in locally weighted regression. *Analytical Chemistry* **1994,** *66* (2), 249-260.
29. Lee, D. E.; Song, J.-H.; Song, S.-O.; Yoon, E. S., Weighted Support Vector Machine for Quality Estimation in the Polymerization Process. *Industrial & Engineering Chemistry Research* **2005,** *44* (7), 2101-2105.
30. Quan, T.; Liu, X.; Liu, Q., Weighted least squares support vector machine local region method for nonlinear time series prediction. *Applied Soft Computing* **2010,** *10* (2), 562-566.
31. Cheng, C.; Chiu, M.-S., A new data-based methodology for nonlinear process modeling. *Chemical Engineering Science* **2004,** *59* (13), 2801-2810.
32. Hall, D. L.; McMullen, S. A., *Mathematical techniques in multisensor data fusion*. Artech House: 2004.

33. Hall, D. L.; Llinas, J., An introduction to multisensor data fusion. *Proceedings of the IEEE* **1997,** *85* (1), 6-23.
34. Ginn, C. M. R.; Willett, P.; Bradshaw, J., Combination of molecular similarity measures using data fusion. *Perspectives in Drug Discovery and Design* **2000,** *20* (1), 1-16.
35. Ginn, C. M. R.; Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. W., Similarity Searching in Files of Three-Dimensional Chemical Structures: Evaluation of the EVA Descriptor and Combination of Rankings Using Data Fusion. *Journal of Chemical Information and Computer Sciences* **1997,** *37* (1), 23-37.
36. Willett, P., Combination of Similarity Rankings Using Data Fusion. *Journal of Chemical Information and Modeling* **2013,** *53* (1), 1-10.
37. Héberger, K., Sum of ranking differences compares methods or models fairly. *Trends in Analytical Chemistry* **2010,** *29* (1), 101-109.
38. Héberger, K.; Kollár-Hunek, K., Sum of ranking differences for method discrimination and its validation: comparison of ranks with random numbers. *Journal of Chemometrics* **2011,** *25* (4), 151-158.
39. Škrbić, B.; Héberger, K.; Đurišić-Mladenović, N., Comparison of multianalyte proficiency test results by sum of ranking differences, principal component analysis, and hierarchical cluster analysis. *Analytical and Bioanalytical Chemistry* **2013,** *405* (25), 8363-8375.
40. Kollár-Hunek, K.; Héberger, K., Method and model comparison by sum of ranking differences in cases of repeated observations (ties). *Chemometrics and Intelligent Laboratory Systems* **2013,** *127* (0), 139-146.

Chapter 2: Matrix Matching

1. Theory of Matrix Matching

The primary goal of local modeling is to select calibration samples that are both spectrally and chemically matched to the target sample[1]. An analyte spectrum, $\boldsymbol{x}_{(n \text{ x } 1)}$, measured over multiple wavelengths (n) can be written according to the Beer-Lambert Law (Eq. 2.1).

$$\boldsymbol{x}^T = y\boldsymbol{\varepsilon}^T\boldsymbol{P} + \boldsymbol{e}^T = y\tilde{\boldsymbol{s}}^T + \boldsymbol{e}^T \tag{2.1}$$

The concentration of the analyte is represented as y, $\boldsymbol{\varepsilon}_{(n \text{ x } 1)}$ are the molar absorptivities of the analyte over each wavelength, $\boldsymbol{P}_{(n \text{ x } n)}$ is a diagonal matrix of wavelength dependent pathlengths, and $\boldsymbol{e}_{(n \text{ x } 1)}$ accounts for random noise. The molar absorptivities and wavelength dependent pathlengths can be combined to form the pathlength corrected isolated pure component spectrum $\tilde{\boldsymbol{s}}_{(n \text{ x } 1)}$. When more than one molecular species is present in a sample, equation 2.1 becomes more complex. The matrix effects that each species, both analyte and interferent species, have on the spectrum ($\boldsymbol{x}$) at each wavelength must be taken into account (Eq. 2.2).

$$\begin{aligned} \boldsymbol{x}^T &= y_a\boldsymbol{\varepsilon}^T\boldsymbol{P}\boldsymbol{M}_a + y_{i_1}\boldsymbol{\varepsilon}^T\boldsymbol{P}\boldsymbol{M}_{i_1} + \cdots + y_{i_1}\boldsymbol{\varepsilon}^T\boldsymbol{P}\boldsymbol{M}_{i_p} + \boldsymbol{e}^T \\ &= y_a\boldsymbol{s}_a^T + y_{i_1}\boldsymbol{s}_{i_1}^T + \cdots + y_{i_p}\boldsymbol{s}_{i_p}^T + \boldsymbol{e}^T \\ &= \boldsymbol{y}^T\boldsymbol{S} + \boldsymbol{e}^T \end{aligned} \tag{2.2}$$

The diagonal matrices of wavelength dependent matrix effect perturbations for the analyte and each interferent species are represented as $\boldsymbol{M}_{a_{(n \text{ x } n)}}$ and $\boldsymbol{M}_{i_{p_{(n \text{ x } n)}}}$, where p is the number of interferent species. These matrices are dependent on the analyte (y_a) and interferent (y_{i_p}) concentrations due to the effects of intermolecular forces[2-5]. For this

multi-species sample, $\boldsymbol{s}_{a_{(n\,\mathrm{x}\,1)}}$ represents the matrix effected pure component analyte spectrum with pathlength correction, and $\boldsymbol{s}_{i_p\,(n\,\mathrm{x}\,1)}$ represents a matrix effected pure component interferent spectrum with pathlength correction. The resulting matrix, $\boldsymbol{S}_{((p+1)\,\mathrm{x}\,n)}$, is a combination of the matrix effected pure component spectra for all species present. The concentrations of all the species in the sample, $\boldsymbol{y}_{((p+1)\mathrm{x}\,1)}$, contain the analyte and interferent concentrations.

Each sample within a library of spectra can have unique physical and chemical matrix effects ($\boldsymbol{S}$). Equation 2.3 shows the resulting unique physical and chemical matrix equations for a library of spectra, $\boldsymbol{X}_{(m\,\mathrm{x}\,n)}$, of m samples.

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{y}_1^T & \cdots & 0^T \\ \vdots & \ddots & \vdots \\ 0^T & \cdots & \boldsymbol{y}_m^T \end{pmatrix} \begin{pmatrix} \boldsymbol{S}_1 \\ \vdots \\ \boldsymbol{S}_m \end{pmatrix} + \boldsymbol{E} \tag{2.3}$$

In this representation of library samples, $\boldsymbol{y}_m$ represents the concentrations of all the chemical species in each library sample, $\boldsymbol{S}_m$ represents the unique matrix effects for each sample in the library, and $\boldsymbol{E}_{(m\,\mathrm{x}\,n)}$ is the random noise. A new sample outside of the library of spectra, a target sample, is represented by equation 2.4.

$$\boldsymbol{x}_t^T = \boldsymbol{y}_t^T \boldsymbol{S}_t + \boldsymbol{e}_t^T \tag{2.4}$$

In this equation, $\boldsymbol{x}_{t_{(n\,\mathrm{x}\,1)}}$ is a target spectrum, $\boldsymbol{S}_{t_{((b+1)\mathrm{x}\,n)}}$ is a matrix of matrix effected pure component spectra for all species present, where b represents the number of interferent species in the target sample, $\boldsymbol{y}_{t_{((b+1)\,\mathrm{x}\,1)}}$ are the concentrations for all species in the target sample, and $\boldsymbol{e}_{t_{(n\,\mathrm{x}\,1)}}$ is random noise. When $\boldsymbol{S}_t \approx \boldsymbol{S}_{(1{:}m)}$ and $\boldsymbol{y}_t \approx \boldsymbol{y}_{(1{:}m)}$ then the target sample is considered matrix matched to the library samples.

The samples in a library set, as represented in equation 2.3, can be used as calibration samples to form a linear regression function for relating the library spectra to a single analyte species (Eq. 2.5),

$$\boldsymbol{y}_a = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e} \tag{2.5}$$

where $\boldsymbol{y}_{a_{(m\text{ x }1)}}$ are the analyte concentrations for an m number of calibration samples and $\boldsymbol{b}_{(n\text{ x }1)}$ is the regression vector. An estimated regression vector, $\widehat{\boldsymbol{b}}_{(n\text{ x }1)}$, calculated from this linear regression function, can be used to predict the analyte concentrations of the target sample (Eq. 2.6).

$$\hat{y}_{a,t} = \boldsymbol{x}_t^T\widehat{\boldsymbol{b}} \tag{2.6}$$

In this equation, $\hat{y}_{a,t}$ is the predicted analyte concentration of the target sample. The estimated regression vector can also predict the calibration samples used to form the regression model (Eq. 2.7).

$$\hat{y}_{a,m} = \boldsymbol{x}_m^T\widehat{\boldsymbol{b}} \tag{2.7}$$

The spectrum of one of the samples used to form the model (Eq. 2.5) is represented as $\boldsymbol{x}_m$ and $\hat{y}_{a,m}$ is the predicted analyte concentration of the calibration sample. For determining if the target sample is matrix matched to the calibration samples, $\hat{y}_{a,t}$ can be compared to $\hat{y}_{a,m}$. When $\hat{y}_{a,t} = \hat{y}_{a,m}$ then $\boldsymbol{x}_t$ should be matrix matched to $\boldsymbol{x}_m$. Unfortunately, $\hat{y}_t$ can be equivalent to $\hat{y}_m$ by chance without the samples being matrix matched. Equation 2.8 represents a situation where $\boldsymbol{x}_t = \boldsymbol{x}_m$ in the form of the linear regression relationships for predicting $\hat{y}_{a,t}$ and $\hat{y}_{a,m}$,

$$y_a\boldsymbol{s}_{a,t}^T\widehat{\boldsymbol{b}} + y_i\boldsymbol{s}_{i,t}^T\widehat{\boldsymbol{b}} + \cdots + \boldsymbol{e}^T\widehat{\boldsymbol{b}} = y_a\boldsymbol{s}_{a,m}^T\widehat{\boldsymbol{b}} + y_i\boldsymbol{s}_{i,m}^T\widehat{\boldsymbol{b}} + \cdots + \boldsymbol{e}^T\widehat{\boldsymbol{b}} \tag{2.8}$$

The species concentrations (y_a and y_i) and the matrix and pathlength corrected pure component spectrum ($\boldsymbol{s}_a$ and $\boldsymbol{s}_i$) interact with the regression vector ($\widehat{\boldsymbol{b}}$) to result in

possible chance equivalency of $\hat{y}_{a,t}$ and $\hat{y}_{a,m}$. As sample matrices can be complicated and the individual species are often unknown, it is hard to determine how likely this chance equivalency is to occur for comparing two specific samples.

2. Matrix Matching Assessment

The exact sample matrix, with all species identified and quantified along with the interactions between these species, is rarely known. However, there are proxy methods that can help identify and visualize matrix matching (Eq. 2.9).

$$\hat{y}_j = \boldsymbol{x}^T \widehat{\boldsymbol{b}} \alpha_j = \hat{y} \alpha_j \tag{2.9}$$

The interaction between $\widehat{\boldsymbol{b}}_{(n \text{ x } 1)}$, an estimated regression vector, scaled by α_j in equation 2.9 can help determine the degree of matrix matching between samples. In this equation, $\boldsymbol{x}_{(n \text{ x } 1)}$ is a sample spectrum, $\hat{y}$ is the prediction of the analyte, and $\hat{y}_j$ is the prediction of the analyte scaled by α_j. Samples that are matrix matched should be influenced similarly by α_j.

Two measurements based on equation 2.9 can be used to visualize the degree of matrix matching between multiple samples. The first measurement (merit) is the prediction error (Eq. 2.10).

$$|\hat{y}_j - y| \tag{2.10}$$

The differences between the predicted reference value ($\hat{y}_j$), with the influence of α_j, and a measured reference value (y) can be plotted against their respective α_j's. Calibration samples and a target sample can be compared using the prediction error merit. Each sample in the calibration set is removed one at a time and predicted by a linear regression model formed by the remaining samples from the calibration set. The target sample is also predicted by this model. This process of removing one calibration sample and

forming a model is repeated for each calibration sample in the set resulting in multiple predictions for the target sample. Figure 2.1 A shows an illustration of three situations of sample prediction errors for calibration samples and a target sample plotted with respect to their respective α_j's. A set of calibration sample scaled prediction errors are shown in blue for each plot and the target sample scaled prediction errors are shown in red. The α_j's represented in the plot for each individual sample correspond with $|\hat{y}_j - y| = 0$ and $|\hat{y}_j - y| = 1$ for each prediction error resulting in a "V" as there are two α_j solutions for $|\hat{y}_j - y| = 1$. When samples are matrix matched, all of the α_j values are similar when $|\hat{y}_j - y| = 0$ and $|\hat{y}_j - y| = 1$. The left most example in Figure 2.1 A shows samples with similar α_j's at $|\hat{y}_j - y| = 0$ and $|\hat{y}_j - y| = 1$ for both the calibration samples' prediction errors and the target sample prediction errors. The other two prediction error plots show situations where the target sample is not matrix matched to the calibration samples by either α_j at $|\hat{y}_j - y| = 0$ or $|\hat{y}_j - y| = 1$ discrepancies.

For these prediction errors plots, $|\hat{y}_j - y|$ can be represented in a couple of different scenarios to identify matrix matching. In the first scenario, explained by the description above, $\hat{y}_j$ are the scaled prediction error values for the calibration samples and the target sample for each leave-one-out model formed, and y are the corresponding reference values for each of the calibration samples and target sample. In this scenario, the target sample reference value must be known to identify matrix matching between target and calibration. In the second scenario, $\hat{y}_j$ are again the scaled prediction error values for the calibration samples and the target sample; however, the y values all correspond to each of the individual calibration samples' reference values for both the

target and calibration scaled prediction error calculations. In this scenario, if $\alpha_j \approx 1$ at $|\hat{y}_j - y| = 0$ when $\hat{y}_j$ are the scaled target predictions and y are the calibration samples, indicates matrix matching. This specific matrix matching indicator is also shown in the left most plot of Figure 2.1 A. For both scenarios, when $\alpha_j \approx 1$ for $|\hat{y}_j - y| = 0$ then $y \approx \hat{y}$ based on equations 2.9 and 2.10. In the second scenario, $y \approx \hat{y}$ demonstrates that there are calibration samples the have the same true reference value as the predicted target sample value.

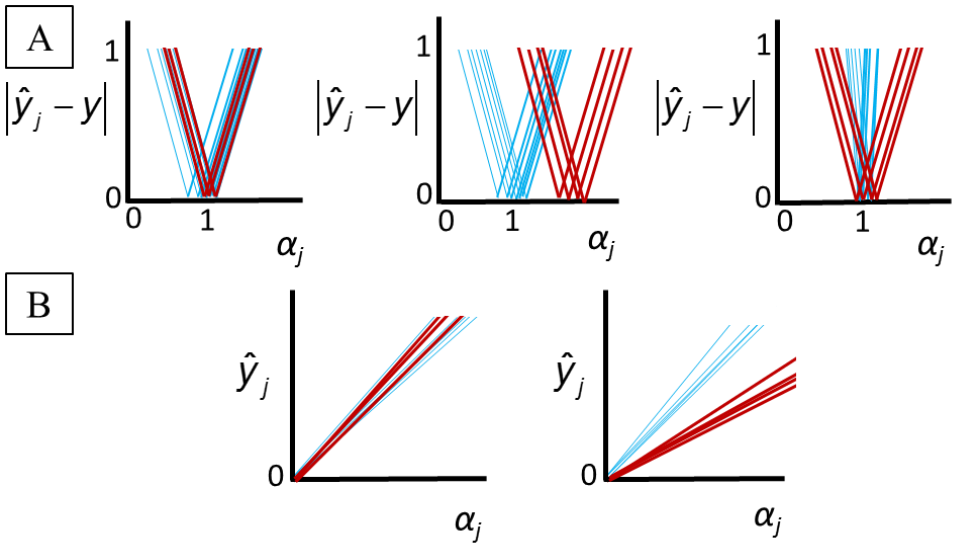


Figure 2.1. Illustration of (A) prediction error matching and (B) prediction slope matching. Calibration samples' scaled prediction errors are shown in blue and the target sample scaled prediction errors are shown red.

The second merit for matrix matching visualization is a prediction match or slope match. When respective $\hat{y}_j$ for the calibration samples' scaled predictions and target sample scaled predictions are plotted against their respective α_j's, the slopes of these

vectors can help indicate the degree in which samples are matrix matched. Figure 2.1 B shows an example, on the left, of the target sample (red) prediction slopes with respect to α_j similar to the calibration sample prediction slopes (blue). The figure on the right (Fig. 2.1 B) shows prediction slopes between the calibration samples and target sample predictions that do not have similar slopes and are therefore not matrix matched.

The scaled prediction errors and scaled prediction slopes are methods for visualization of matrix matching between samples; however, methods for automatically selecting matrix matched samples while avoiding prediction equivalencies are also necessary. Three techniques proposed, in this work, can help determine matrix matched calibration sets for a target sample based on the individual samples within each calibration set. The first technique is data fusion, as described in Chapter 1. The use of data fusion allows for multiple comparison merits, including prediction error based merits, to be used simultaneously. The second technique, made possible by data fusion, is the use of multiple models for prediction error merits. The matrix matching assessment plots (Fig. 2.1) above represent the scaled predictions from one model tuning parameter. Assuming the estimated regression model vector ($\widehat{\boldsymbol{b}}$) is calculated using biased regression techniques, multiple models are formed and the prediction errors from each model can be represented. The third technique is referred to as cross modeling, which is also made possible through data fusion. The prediction errors, selection of multiple models, and cross modeling process are described below.

2.1. *Regression Model Prediction Error Merits*

The merits discussed are based on predictions of calibration samples and target samples from a leave-one-out method for each of the calibration samples in a calibration

set (as described above). Regression model prediction errors are the absolute difference between the true value and the predicted value (Eq. 2.11).

$$e_{22} = |\hat{y}_2 - y_2| \tag{2.11}$$

where y_2 is the true reference value and $\hat{y}_2$ is the predicted reference value for a calibration sample predicted by a specific model. This merit is based solely on calibration samples and is used to ensure that the calibration set predicts each sample within the set accurately. A similar measurement to e_{22} is the absolute difference in predictions for the calibration sample ($\hat{y}_2$) and the target sample ($\hat{y}_1$), shown in equation 2.12.

$$e_{12} = |\hat{y}_1 - \hat{y}_2| \tag{2.12}$$

This merit is used to assess how closely two samples are predicted by the same model. These two prediction merits, referred to as Y merits, are listed in Table 2.1 with an assigned Merit ID for reference for discussion in section 6.2.

Table 2.1. Prediction merits (Y) for calibration set comparisons. (Notations indicated in footnotes.

Category	Merit	Input Assignments	Equation	Merit ID
Y	e_{22}	$y_2 = y_o$; $\hat{y}_2 = \hat{y}_o$	2.11	H1
Y	e_{12}	$\hat{y}_1 = \hat{y}_t$; $\hat{y}_2 = \hat{y}_o$	2.12	H2

y_o **:** calibration sample reference value removed from calibration set

$\hat{y}_o$ **:** predicted calibration sample reference value removed from calibration set

$\hat{y}_t$ **:** predicted target (target) sample reference value

2.2. *Multiple Model Tuning Parameters*

Prediction error merits, e_{22} (Eq. 2.11) and e_{12} (Eq. 2.12), require a selection of a tuning parameter. The selection process described here is based on PLS models with latent variables (LV's) as tuning parameters. The number of LV's possible is dependent on the rank of the system. As it is not reasonable or meaningful to use all possible LV's, a method for specifying a select number of LV's is necessary. One option for selecting a number of meaningful LV's is based on equation 2.13[6].

$$UM_k = \left(\frac{\overline{\|\hat{\boldsymbol{b}}_k\|} - \overline{\|\hat{\boldsymbol{b}}\|}_{min}}{\overline{\|\hat{\boldsymbol{b}}\|}_{max} - \overline{\|\hat{\boldsymbol{b}}\|}_{min}}\right) + \left(\frac{RMSECV_k - RMSECV_{min}}{RMSECV_{max} - RMSECV_{min}}\right) \tag{2.13}$$

The variable UM_k represents a method for combining a measure of model prediction variance, the Euclidean norm of the regression vector ($\|\hat{\boldsymbol{b}}\|$), and a measure of the model prediction bias, $RMSECV$. The average $\|\hat{\boldsymbol{b}}_k\|$ and $RMSECV_k$ are calculated from the prediction models formed by each calibration sample removed across all possible LV's (k). In equation 2.13 $RMSECV_k$ is calculated as

$$RMSECV_k = \sqrt{\frac{\sum_{i=1}^{m}(\hat{y}_i - y_i)^2}{m}} \tag{2.14}$$

The $RMSECV_k$ is the root-mean square error of cross validation and represents the prediction error for each calibration sample removed for m calibration samples. In equation 2.13, the $RMSECV_k$ is range scaled using the minimum prediction error ($RMSECV_{min}$) and maximum prediction error ($RMSECV_{max}$) across all the LV's. In equation 2.13, $\overline{\|\hat{\boldsymbol{b}}_k\|}$ represents the average Euclidean norm of each of the estimated regression vectors by the leave-one-out models formed for each LV. The $\overline{\|\hat{\boldsymbol{b}}_k\|}$ is range scaled by the minimum Euclidean norm estimated regression vector ($\overline{\|\hat{\boldsymbol{b}}\|}_{min}$) and the

maximum Euclidean norm estimated regression vector ($\overline{\|\widehat{\boldsymbol{b}}\|}_{max}$) across all the LV's. For equation 2.13, $RMSECV_k$ can also be represented by the root-mean-square error of calibration ($RMSEC_k$). However $RMSECV_k$ is used to avoid model selections that are over fitted to the calibration samples.

When UM_k is plotted against the LV index, a "U-curve" is formed. The minimum point in the "U-curve" can be used to automatically select an optimal tuning parameter for a balanced bias/variance tradeoff [6-8]. The identified tuning parameter at the minimum point in the "U-curve" is referred to as LV'. A set number of surrounding tuning parameters on each side of LV' can be used for representing multiple models for the prediction error merits and for other merits discussed that are based on PLS algorithms.

2.3. *Cross Modeling*

The prediction error merit, e_{12}, assesses how well a single target sample is predicted compared to the prediction of one calibration sample removed from a calibration set. In this work, prediction error merits are also used to assess how the target sample and a calibration sample removed from a calibration set are predicted by, what is referred to in this work as, an evaluating calibration set. This technique is referred to as cross modeling. Figure 2.2 shows an example of this cross modeling technique. The evaluating calibration sets used are the same sets as the calibration sets being compared to one another but can be any set of samples used to form a calibration model. The columns from this figure represent the calibration sets that are being compared and the rows represent the evaluating sets used to build the models, five models are represented for each evaluating set, for predicting the calibration samples from each calibration set and the target sample. For example, box 1 represents the mean prediction differences

between each sample in calibration set 1 and the target sample predicted by the remaining samples in calibration set 1 over five LV's. In box 1, the evaluating calibration set is represented by the remaining samples in calibration set 1. Box 2 shows the mean prediction differences between each calibration sample in calibration set 5 and the target sample predicted by calibration set 2, the evaluating set, over five LV's. This figure clearly shows that the mean prediction differences for the target sample and the samples in calibration set 5 are similar regardless of the evaluating set used to build a model. This indicates that the samples in calibration set 5 are most likely matrix matched to the target sample. The method of cross modeling can help further reduce the possibility of a chance matrix match between two spectra as many different models are used to assess the target sample and calibration samples.

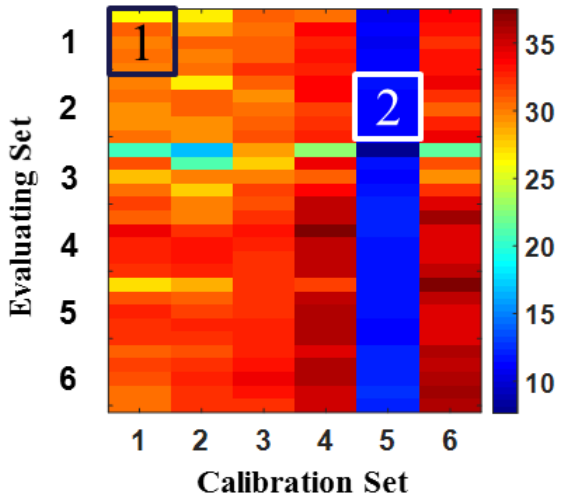


Figure 2.2. Example of cross modeling for prediction difference merit (e_{12}) for six calibration sets.

3. Calibration Set Spectral Comparison Methods

Prediction merits are important for determining matrix matching; however, with the use of the data fusion techniques, there is an opportunity to use spectral matching merits simultaneously with the prediction based merits to assist in the identification of matrix matched calibration sets. As previously discussed, the cosine of the angle between vectors ($cos\,\theta$) and the distance between vectors (Euclidean distance) are a couple common examples of spectral comparisons. Transformation merits, such as Procrustes analysis (PA)[9] and extended inverted scatter correction (EISC)[10], that determine the magnitude of the transformation required to make one spectrum resemble another spectrum, can also be measures of similarity. There are many other mathematical measurements for similarity assessment. Each of these merits can bring out unique aspects of spectral similarity.

All of the merits presented in this section are represented in a generalized format of $\boldsymbol{x}_{1_{(n\text{ x }1)}}$ and $\boldsymbol{x}_{2_{(n\text{ x }1)}}$. This generalization is necessary as these same merits are represented in multiple formats for Chapter 2 and Chapter 3. Detailed descriptions of the use of each merit are given within Chapters 2 and 3 for specific merit sections. In this chapter the objective of these merits is to compare calibration samples from specified calibration sets based on their matrix matching potential to a target sample. The merits are used for two different spectral conditions: raw spectra and as vectors from orthogonal projections to an estimated regression vector ($\widehat{\boldsymbol{b}}$). The definitions for $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ for each of these conditions is explained below.

For calibration set comparisons using raw spectral data, $\boldsymbol{x}_1$ is always equal to the spectrum of the calibration sample removed from a calibration set or the target sample

spectrum, and $\boldsymbol{x}_2$ is always equal to the mean of the calibration spectra in an evaluating calibration set. As discussed for cross modeling, the evaluating set mean can either be the calibration samples remaining from a set after one sample is removed or it can be another calibration set unrelated to the calibration sample being assessed. These merits, based on raw spectral data, are referred to as Spectral merits throughout the work. The objective of the Spectral merits is to measure the differences between the calibration spectrum removed from a calibration set and the target sample each compared to the evaluating set. Each merit is calculated for the sample removed from the calibration set (M_o) and calculated again for the target sample (M_t). The actual measurements used for assessing matrix matching potentials of calibration sets are the absolute differences between M_o and M_t (Eq. 2.15).

$$M = |M_o - M_t| \tag{2.15}$$

For example, for the Euclidean distance comparison merit, the Euclidean distance is calculated between the calibration sample spectrum removed from the calibration set and the mean of the evaluating set spectra (M_o) and calculated between the target sample spectrum the mean of the evaluating set spectra (M_t). The Euclidean distance merit used for calibration set comparisons would be the absolute difference between these two Euclidean distance measurements. The final Euclidean distance measurements are averaged across each calibration sample in a set as shown in Figure 2.2 and explained in the cross modeling procedure. The purpose of these merit calculations is to show that if the target sample and calibration samples in a calibration set are truly matrix matched then their spectral comparison differences will be similar for each evaluating calibration space.

For merits based on orthogonal projections to an estimated regression vector, referred to as OP merits, the orthogonal projections are calculated by equation 2.16.

$$\boldsymbol{x}^{\perp} = \left(\mathbf{I} - \hat{\mathbf{b}}_r \hat{\mathbf{b}}_r^T\right)\boldsymbol{x}_1 \tag{2.16}$$

where $\hat{\mathbf{b}}_r = \boldsymbol{X}_r^{+}\boldsymbol{y}_r$ is calculated from a PLS regression and **I** is an identity matrix. The PLS regression is formed by the samples in an evaluation set ($\boldsymbol{X}_r$ and corresponding reference values, $\boldsymbol{y}_r$). Again, the evaluation set is represented as the samples remaining after one sample is removed or as another calibration set. The orthogonal projections for the target spectrum ($\boldsymbol{x}_t^{\perp}$) and the individual calibration spectrum from each proposed calibration set ($\boldsymbol{x}_o^{\perp}$) are calculated for each model formed by the evaluation set following the cross modeling procedure. For these OP merits, $\boldsymbol{x}_1$ can be represented by either $\boldsymbol{x}_t^{\perp}$ or $\boldsymbol{x}_o^{\perp}$ and $\boldsymbol{x}_2$ can also be represented by either $\boldsymbol{x}_t^{\perp}$ or $\boldsymbol{x}_o^{\perp}$. The representations of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ for each of the individual OP merits is described in the merit summary sections (3.5.9 and 3.6.4).

The orthogonally projected spectrum represent the variability not accounted for by the regression model as the regression model only accounts for the variability related to the analyte species. This orthogonal variability can contain information related to interferent species. For samples that are matrix matched, the orthogonal projections from the regression model should be similar. Like the prediction based merits previously described, the orthogonal projection merits can be calculated for multiple estimated regression vectors. The tuning parameter selection methods described in section 2.2 are used to determine the estimated regression vectors used.

3.4. *Spectral Based Matrix Matching Merits*

The spectral based merits (both Spectral and OP) can be broken into three main categories; sample vector to calibration vector, sample domain to calibration domain and sample vector to calibration domain. Each of these categories is explained in detail below. For sample vector to calibration vector comparisons, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ spectra are treated as vectors. For sample domain to calibration domain comparisons, outer products are calculated for $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. For example, $\boldsymbol{x}_1\boldsymbol{x}_1^T = \boldsymbol{X}_{1_{(n \text{ x } n)}}$, where $\boldsymbol{X}_1$ is the outer product calculated for $\boldsymbol{x}_1$. These outer products result in a pseudo spectral domain. The benefits of using outer products are explained in section 3.6. For the sample vector to calibration domain merits, $\boldsymbol{x}_1$ is a vector of either the target sample spectrum or the calibration sample spectrum removed from a calibration set and the calibration domain represents the evaluating set, the remaining samples in the calibration set or another calibration set. The calibration domain is always referred to as $\boldsymbol{X}_r$ for the sample vector to calibration domain merits.

The solutions for some of the merits presented require the calculation of singular value compositions (SVD) (Eq. 2.17).

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{2.17}$$

where $\boldsymbol{A}_{(m \text{ x } n)}$ is a generic matrix of m rows and n columns. The eigenvector matrices are represented by $\boldsymbol{U}_{(m \text{ x m})}$ and $\boldsymbol{V}_{(n \text{ x } n)}$, and $\boldsymbol{\Sigma}_{(m \text{ x } n)}$ is a matrix of zeros except for the singular values**.** This decomposition is typically truncated to $\boldsymbol{U}_{(m \text{ x } l)}$**,** $\boldsymbol{V}_{(n \text{ x } l)}$, and $\boldsymbol{\Sigma}_{(l \text{ x } l)}$ where l represents the mathematical rank. For the sample vector to calibration vector and sample domain to calibration domain merits the SVD will have a rank of one ($l = 1$) where only the first eigenvectors and singular value are used. Each of these rank one

SVD calculations can be calculated in more computationally simplistic ways than solving for the SVD. In the following equations the SVD solutions along with simplifications are shown. For some merits SVD's are calculated for rank one matrices and in other merits SVD's are calculated for rank one vectors. As mentioned for the sample domain to calibration domain merits, the sample domains can be calculated by the outer product of two vectors resulting a rank one matrix. Rank one SVD of matrices are represented by equation 2.18.

$$\boldsymbol{A} = \boldsymbol{u}\sigma\boldsymbol{v}^T = \frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} \times \|\boldsymbol{a}\|^2 \times \frac{\boldsymbol{a}^T}{\|\boldsymbol{a}\|} \tag{2.18}$$

Where $\boldsymbol{u}$ and $\boldsymbol{v}$ are n x 1 eigenvectors and σ is the first singular value. The mathematical equivalents to these eigenvectors and singular value are shown using the vector used to create the matrix ($\boldsymbol{a}$) and Euclidean norm of the vector ($\|\boldsymbol{a}\|$). Rank one SVD's of vectors are represented by equations 2.19 and 2.20 for both column and row vector representations.

$$\boldsymbol{a} = \boldsymbol{u}\sigma v^T = \frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} \times \|\boldsymbol{a}\| \times 1 \tag{2.19}$$

In equation 2.19, $\boldsymbol{a}$ is a column vector of n x 1, $\boldsymbol{u}$ is an eigenvector of n x 1, σ is the singular value and v is equivalent to 1.

$$\boldsymbol{a}^T = u\sigma\boldsymbol{v}^T = 1 \times \|\boldsymbol{a}\| \times \frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} \tag{2.20}$$

In equation 2.20, $\boldsymbol{a}^T$ is a row vector of 1 x n, u is equivalent to 1, σ is the singular value and $\boldsymbol{v}$ is an eigenvector of n x 1. The sample vector to calibration vector and sample domain to calibration domain merits described below show both solutions based on SVD and the simplified calculations.

The sample vector to calibration domain merits will have matrices with ranks greater than one. Multiple eigenvectors and singular values can be used for the SVD

calculations of these merits. The specified number of eigenvectors and singular values are referred to as principal components (PC's). The number of PC's selected vary for the sample vector to calibration domain merits depending on how these merits are used. The specific number of PC's used in this chapter is described in section 3.7.

3.5. *Sample Vector to Calibration Vector Comparison Merits*

The following merits are comparison measurements of two vectors. Some of the merits presented can be calculated with column vectors $\boldsymbol{x}_{(n \text{ x } 1)}$ or row vectors $\boldsymbol{x}^T_{(1 \text{ x } n)}$ of n variables. The merits where the outcome is dependent on column vector or row vector input are noted in the discussion of the specific merit. Different combinations of row and column vectors for each merit were assessed; however, these results are not reported here. Only the row and column vector combinations that resulted in unique solutions are displayed in this section.

3.5.1 Cos θ

The measurement of *cos θ* is the cosine of the angle between two vectors (Eq. 2.21).

$$1 - cos\ \theta = 1 - \frac{|\boldsymbol{x}_1^T \boldsymbol{x}_2|}{\|\boldsymbol{x}_1\| \|\boldsymbol{x}_2\|} \tag{2.21}$$

This merit determines if two vectors ($\boldsymbol{x}_1$ and $\boldsymbol{x}_2$) have similar shapes. In this work, *cos θ* is subtracted from 1 so that a minimum values represent higher degrees of similarity for all merits.

The outer products of vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ can be compared based on the second order limited method for *cos θ* [11-12]. However, as previously mentioned, the rank for these outer products is one. The simplification of calculating *cos θ* for two outer product arrays can be simplified to the cosine of the angle between the two original vectors squared. Due

to this simplification, the merit $1 - cos^2\,\theta$ is expressed in the sample vector to calibration vector comparison merits section as opposed to the samples domain to calibration domain section (3.6) whose merits are based on outer product arrays (Eq. 2.22).

$$1 - cos\,\theta_{SOL} = 1 - M_1 M_2 = 1 - cos^2\,\theta \tag{2.22}$$

In this method, the SVD for each of the outer products $\boldsymbol{x}_1\boldsymbol{x}_1^T$ and $\boldsymbol{x}_2\boldsymbol{x}_2^T$ can be used to calculate M_1 and M_2 using only the first eigenvectors ($\boldsymbol{u}_1$, $\boldsymbol{u}_2$, $\boldsymbol{v}_1$, and $\boldsymbol{v}_2$). The expressions for M_1 and M_2 are represented as

$$M_1 = |\boldsymbol{u}_1^T\boldsymbol{u}_2| = \frac{\boldsymbol{x}_1^T}{\|\boldsymbol{x_1}\|}\frac{\boldsymbol{x}_2}{\|\boldsymbol{x}_2\|} = cos\,\theta$$

$$M_2 = |\boldsymbol{v}_1^T\boldsymbol{v}_2| = \frac{\boldsymbol{x}_1^T}{\|\boldsymbol{x_1}\|}\frac{\boldsymbol{x}_2}{\|\boldsymbol{x}_2\|} = cos\,\theta$$

These equations show that M_1 and M_2 are in fact equal to *cos* θ between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Like *cos* θ, the final merit ($cos^2\,\theta$) is subtracted from one.

3.5.2 Euclidean Distance

The Euclidean distance is the measure of distance, a measure of magnitude, between two vectors, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ (Eq. 2.23).

$$Euc = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^{\mathrm{T}}(\mathbf{x}_1 - \mathbf{x}_2)} \tag{2.23}$$

3.5.3 Determinant

The determinant is a measurement of the space formed between two vectors (Eq. 2.24).

$$Det = \left|\begin{pmatrix}\boldsymbol{x}_1^T\\ \boldsymbol{x}_2^T\end{pmatrix}(\boldsymbol{x}_1 \quad \boldsymbol{x}_2)\right| = (\|\boldsymbol{x}_1\|\|\boldsymbol{x}_2\|sin\,\theta_1)^2 \tag{2.24}$$

The two vectors are concatenated together to form rectangular matrices of 2 x n and n x 2. In this form the determinant can also be calculated by taking the square of the

product of the Euclidean normalizations ($\|\cdot\|$) of each $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ and the $sin\,\theta$ between the two vectors.

3.5.4 Procrustes Analysis

Procrustes analysis (PA) merits are used in this work to provide a measurement of similarity between two vectors by assessing mathematical transformations required to make one vector resemble a second vector through the degree of translation, dilation, and rotation[9]. The first merit, considered the unconstrained PA transformation merit, measures the degree of rotation and stretching. Row vectors ($\boldsymbol{x}^T_{(1\,\mathrm{x}\,n)}$) are used to calculate F_{21} (Eq. 2.25).

$$\boldsymbol{x}_1^T = \boldsymbol{x}_2^T\,\boldsymbol{F}_{21} \tag{2.25}$$

To solve for $\boldsymbol{F}_{21_{(n\,\mathrm{x}\,n)}}$, the pseudoinverse of $\boldsymbol{x}_2^T$ is required. An SVD can be used to calculate the pseudoinverse of $\boldsymbol{x}_2^T$. The equations are shown using SVD, but also as simplified equivalent representations based on equation 2.20.

$$\boldsymbol{F}_{21} = (\boldsymbol{x}_2^T)^+\boldsymbol{x}_1^T \tag{2.26}$$

In equation 2.26

$$(\boldsymbol{x}_2^T)^+ = u\sigma^{-1}\boldsymbol{v}^T = \frac{\boldsymbol{x}_2}{\|\boldsymbol{x}_2\|^2}$$

The final merit (F) is the Frobenius norm of the difference between $\boldsymbol{F}_{21}$ and $\boldsymbol{F}_{22}$ (Eq. 2.27).

$$F = \|\boldsymbol{F}_{21} - \boldsymbol{F}_{22}\|_F \tag{2.27}$$

The Forbenius normalization method is used in order to represent the transformation as a scalar value (Eq. 2.28)[13].

$$\|\cdot\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2} \tag{2.28}$$

For the Frobenius norm, m is the number of row elements, n is the number column elements, and a_{ij} is each individual element in the matrix. Other normalization techniques can be used in place of the Frobenius norm.

For the final merit in equation 2.27, $\boldsymbol{F}_{22}$ is calculated the same as $\boldsymbol{F}_{21}$ except that $\boldsymbol{x}_1^T$ is replaced with $\boldsymbol{x}_2^T$ in equation 2.25 to measure the unconstrained degree of transformation between $\boldsymbol{x}_2^T$ and $\boldsymbol{x}_2^T$. The difference between $\boldsymbol{F}_{21}$ and $\boldsymbol{F}_{22}$ in equation 2.27 determines the degree of transformation required between vector $\boldsymbol{x}_1^T$ and vector $\boldsymbol{x}_2^T$ in comparison to the degree of transformation required between vector $\boldsymbol{x}_2^T$ and vector $\boldsymbol{x}_2^T$. This comparison helps to account for the full degree of difference between the two vectors.

The constrained PA merits, ρ_{21} and H_{21}, measure the degree of dilation and rotation respectively between two vectors. For the sample vector to calibration vector comparisons, H_{21} is not used as the solution is always one. The dilation merit, ρ_{21}, is calculated using equation 2.29.

$$\boldsymbol{x}_1 = \rho_{21}\boldsymbol{x}_2 H_{21} \tag{2.29}$$

Solving for ρ_{21} is represented by equation 2.30.

$$\rho_{21} = \frac{\boldsymbol{x}_2^T \boldsymbol{x}_1}{tr(\boldsymbol{x}_2 \boldsymbol{x}_2{}^T)} \tag{2.30}$$

where the dilation transformation, ρ_{21}, is calculated by the inner product of $\boldsymbol{x}_2{}^T \boldsymbol{x}_1$ divided by the trace, or sum of diagonal elements, of the outer product of $\boldsymbol{x}_2 \boldsymbol{x}_2{}^T$. The final merit, ρ, is calculated as

$$\rho = |\rho_{21} - \rho_{22}| \tag{2.31}$$

In this equation, ρ is the absolute difference between the degree of transformation between $\boldsymbol{x}_2$ and $\boldsymbol{x}_1$ for ρ_{21} and the degree of transformation between $\boldsymbol{x}_2$ and $\boldsymbol{x}_2$ for ρ_{22}.

In this final merit calculation, ρ_{22} is calculated the same as ρ_{21} except that $\boldsymbol{x}_1$ is replaced with $\boldsymbol{x}_2$ in equation 2.29. However, in the rank one system ρ_{22} is always equivalent to 1. Like the unconstrained Procrustes analysis merit, a Frobenius norm could be used, but is not necessary as both ρ_{21} and ρ_{22} are scalar values.

3.5.5 Extended Inverted Scatter Correction

Extended inverted scatter correction (EISC) has been used in applications for analytical chemistry as a signal correction method[14] and calibration model transfer[10]. In this work EISC is used to measure spectral similarity. The EISC correction function, or transformation, is expressed in equation 2.32.

$$\boldsymbol{x}_1 = b_0\mathbf{1} + b_1\boldsymbol{x}_2 + b_2\boldsymbol{x}_2^2 + b_3\boldsymbol{x}_2^3 + b_4\frac{d\boldsymbol{x}_2}{d\lambda} + b_5\frac{d^2\boldsymbol{x}_2}{d\lambda^2} + b_6\lambda + b_7\lambda^2 + \mathrm{b}_8\ln\lambda \quad (2.32)$$

The correction terms include sources of spectral differences such as, wavelength scale shifts and bandwidth differences. This function is not limited to the correction terms listed above, but were terms selected for this work. The correction terms can be represented in matrix notation as $\boldsymbol{X}_{c_{(n\,\mathrm{x}\,f)}}$, where f is the number of correction terms (1, $\boldsymbol{x}_2, \boldsymbol{x}_2^2, \ldots$), and the correction coefficients can be represented as $\boldsymbol{b}_{21_{(f\,\mathrm{x}\,1)}}$. The transformation function is then rewritten as

$$\boldsymbol{x}_1 = \boldsymbol{X}_c\boldsymbol{b}_{21}$$

Solving for $\widehat{\boldsymbol{b}}_{21_{(f\,\mathrm{x}\,1)}}$ gives

$$\widehat{\boldsymbol{b}}_{21} = (\boldsymbol{X}_c^T\boldsymbol{X}_c)^{-1}\boldsymbol{X}_c^T\boldsymbol{x}_1$$

The Euclidean normalization of $\widehat{\boldsymbol{b}}_{21}$ ($\|\cdot\|$) can be used as a measure of similarity represented by $EISC\ b$ (Eq. 2.33).

$$EISC\ b = \left\|\widehat{\boldsymbol{b}}_{21}\right\| \quad (2.33)$$

In extended scatter correction by difference, an alternative calculation for EISC[10], $\boldsymbol{x}_2$ is added as a variable into the transformation function (Eq. 2.34).

$$\boldsymbol{x}_1 = \boldsymbol{x}_2 + b_0\mathbf{1} + b_1\boldsymbol{x}_2 + b_2\boldsymbol{x}_2^2 + b_3\boldsymbol{x}_2^3 + b_4\frac{d\boldsymbol{x}_2}{d\lambda} + b_5\frac{d^2\boldsymbol{x}_2}{d\lambda^2} + b_6\lambda + b_7\lambda^2 + \mathrm{b}_8\ln\lambda \quad (2.34)$$

The difference between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is then represented as $\boldsymbol{d}$ (Eq. 2.35).

$$\boldsymbol{x}_1 - \boldsymbol{x}_2 = \boldsymbol{d} = \boldsymbol{X}_c\boldsymbol{b}_d \quad (2.35)$$

The correction terms and correction coefficients are again represented in matrix notation as $\boldsymbol{X}_c$ and $\boldsymbol{b}_d$. Similarly to merit $EISC\ b$ (Eq. 2.33), $\boldsymbol{b}_d$ is solved for as

$$\widehat{\boldsymbol{b}}_d = (\boldsymbol{X}_c^T\boldsymbol{X}_c)^{-\mathbf{1}}\boldsymbol{X}_c^T\boldsymbol{d}$$

and the Euclidean normalization of this measure is another similarity measure, $EISC\ b_d$ (Eq. 2.36).

$$EISC\ b_d = \left\|\widehat{\boldsymbol{b}}_d\right\| \quad (2.36)$$

When this transformation is calculated with the difference between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$**,** the product of the correction terms ($\boldsymbol{X}_c$) and coefficients ($\widehat{\boldsymbol{b}}_d$) themselves can represent a measure of similarity between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ (Eq. 2.37). .

$$EISC\ Xb_d = \left\|\boldsymbol{X}_c\widehat{\boldsymbol{b}}_{\boldsymbol{d}}\right\| \quad (2.37)$$

The Euclidean normalization of $\boldsymbol{X}_c\boldsymbol{b}_d$ represents the merit $EISC\ Xb_d$.

3.5.6 Mahalanobis Distance

The following two merits, Mahalanobis distance and pooled Mahalanobis distance, are special cases of sample vector to calibration vector comparison merits. These merits use the outer product arrays of the individual vectors as a part of the calculation along with the vectors themselves. Merits that only use the outer product arrays for vectors comparisons are discussed in the section 3.6.

The Mahalanobis distance is used to determine the distance of one sample to the space of a group of samples[15]. This measure is commonly used for outlier determinations for linear regression models, but can also represent a measure of similarity. The original Mahalanobis distance calculation is manipulated for the calculation between two vectors (Eq. 2.38).

$$MD^V = \sqrt{(\boldsymbol{x}_2 - \boldsymbol{x}_1)^T \boldsymbol{X}_1^+ (\boldsymbol{x}_2 - \boldsymbol{x}_1)} \tag{2.38}$$

The typical covariance matrix used in the standard calculation of Mahalanobis distance is represented as an outer product array of $\boldsymbol{x}_{1_{(n \text{ x } 1)}}$ (Eq. 2.39).

$$\boldsymbol{X}_1 = \boldsymbol{x}_1 \boldsymbol{x}_1^T \tag{2.39}$$

The pseudoinverse of $\boldsymbol{X}_{1_{(n\,x\,n)}}$ is calculated through SVD

$$\boldsymbol{X}_1^+ = \boldsymbol{v}_1 \sigma_1^{-1} \boldsymbol{u}_1^T = \frac{\boldsymbol{x}_1}{\|\boldsymbol{x}_1\|} \frac{1}{\|\boldsymbol{x}_1\|^2} \frac{\boldsymbol{x}_1^T}{\|\boldsymbol{x}_1\|}$$

where only the first basis vector is used. The simplified calculation is also shown.

3.5.7 Pooled Mahalanobis Distance

The pooled Mahalanobis distance is a manipulation of the standard Mahalanobis distance used to compare the structural similarities between two multidimensional datasets[16]. In this work the pooled Mahalanobis distance is further manipulated to compare two vectors (Eq. 2.40).

$$MD_p^V = \sqrt{(\boldsymbol{x}_1 - \boldsymbol{x}_2)^T \boldsymbol{S}^+ (\boldsymbol{x}_1 - \boldsymbol{x}_2)} \tag{2.40}$$

In this merit, $\boldsymbol{S}^+$ is the pseudoinverse

$$\boldsymbol{S}^+ = \boldsymbol{v} \sigma^{-1} \boldsymbol{u}^T$$

of $\boldsymbol{S}$ (Eq. 2.41) calculated for one basis vector.

$$\boldsymbol{S} = \frac{\boldsymbol{X}_1 + \boldsymbol{X}_2}{2} \tag{2.41}$$

In equation 2.41, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are the outer product arrays of both $\boldsymbol{x}_{1_{(n \text{ x } 1)}}$ (Eq. 2.39) and $\boldsymbol{x}_{2_{(n \text{ x } 1)}}$ defined as

$$\boldsymbol{X}_2 = \boldsymbol{x}_2 \boldsymbol{x}_2^T \tag{2.42}$$

3.5.8 Merit Summary

Table 2.2 has a complete list of the sample vector to calibration vector merits used for comparing calibration sets. Both orthogonal projections to an estimated regression model vector (OP) and spectral methods (Spectral) are used to calculate these merits. All of the merits listed are used with the cross modeling techniques.

Table 2.2. Sample vector to calibration vector merits for calibration set comparisons for both Spectral and OP merits. (Notations indicated in footnotes).

Category	Merit	Input Assignments	Equation	Merit ID
Spectral	$1 - cos^2\,\theta$	$\boldsymbol{x}_1 = \boldsymbol{x}_{o/t}\,;\, \boldsymbol{x}_2 = \overline{\boldsymbol{x}}_r$	2.22	H3
OP	Euc	$\boldsymbol{x}_1 = \boldsymbol{x}_o^{\perp}\,;\, \boldsymbol{x}_2 = \boldsymbol{x}_t^{\perp}$	2.23	H4
Spectral	Euc	$\boldsymbol{x}_1 = \boldsymbol{x}_{o/t}\,;\, \boldsymbol{x}_2 = \overline{\boldsymbol{x}}_r$	2.23	H5
OP	Det	$\boldsymbol{x}_1 = \boldsymbol{x}_{\boldsymbol{o}}^{\perp}\,;\, \boldsymbol{x}_2 = \boldsymbol{x}_t^{\perp}$	2.24	H6
Spectral	Det	$\boldsymbol{x}_1 = \boldsymbol{x}_{o/t}\,;\, \boldsymbol{x}_2 = \overline{\boldsymbol{x}}_r$	2.24	H7
OP	F	$\boldsymbol{x}_1^T = \boldsymbol{x}_t^{\perp T}\,;\, \boldsymbol{x}_2^T = \boldsymbol{x}_o^{\perp T}$	2.27	H8
OP	ρ	$\boldsymbol{x}_1 = \boldsymbol{x}_o^{\perp}\,;\, \boldsymbol{x}_2 = \boldsymbol{x}_t^{\perp}$	2.31	H9
OP	ρ	$\boldsymbol{x}_1 = \boldsymbol{x}_t^{\perp}\,;\, \boldsymbol{x}_2 = \boldsymbol{x}_o^{\perp}$	2.31	H10
OP	$EISC\ b$	$\boldsymbol{x}_1 = \boldsymbol{x}_o^{\perp}\,;\, \boldsymbol{x}_2 = \boldsymbol{x}_t^{\perp}$	2.33	H11
OP	$EISC\ b$	$\boldsymbol{x}_1 = \boldsymbol{x}_t^{\perp}\,;\, \boldsymbol{x}_2 = \boldsymbol{x}_o^{\perp}$	2.33	H12
OP	$EISC\ Xb_d$	$\boldsymbol{x}_1 = \boldsymbol{x}_o^{\perp}\,;\, \boldsymbol{x}_2 = \boldsymbol{x}_t^{\perp}$	2.37	H13
OP	$EISC\ Xb_d$	$\boldsymbol{x}_1 = \boldsymbol{x}_t^{\perp}\,;\, \boldsymbol{x}_2 = \boldsymbol{x}_o^{\perp}$	2.37	H14
Spectral	$EISC\ Xb_d$	$\boldsymbol{x}_1 = \boldsymbol{x}_{o/t}\,;\, \boldsymbol{x}_2 = \overline{\boldsymbol{x}}_r$	2.37	H15
Spectral	$EISC\ Xb_d$	$\boldsymbol{x}_1 = \overline{\boldsymbol{x}}_r;\, \boldsymbol{x}_2 = \boldsymbol{x}_{o/t}$	2.37	H16
OP	$EISC\ b_d$	$\boldsymbol{x}_1 = \boldsymbol{x}_o^{\perp}\,;\, \boldsymbol{x}_2 = \boldsymbol{x}_t^{\perp}$	2.36	H17
OP	$EISC\ b_d$	$\boldsymbol{x}_1 = \boldsymbol{x}_t^{\perp}\,;\, \boldsymbol{x}_2 = \boldsymbol{x}_o^{\perp}$	2.36	H18
Spectral	$EISC\ b_d$	$\boldsymbol{x}_1 = \boldsymbol{x}_{o/t}\,;\, \boldsymbol{x}_2 = \overline{\boldsymbol{x}}_r$	2.36	H19
Spectral	$EISC\ b_d$	$\boldsymbol{x}_1 = \overline{\boldsymbol{x}}_r;\, \boldsymbol{x}_2 = \boldsymbol{x}_{o/t}$	2.36	H20
OP	MD^v	$\boldsymbol{x}_1 = \boldsymbol{x}_o^{\perp}\,;\, \boldsymbol{x}_2 = \boldsymbol{x}_t^{\perp}$	2.38	H21
OP	MD^v	$\boldsymbol{x}_1 = \boldsymbol{x}_t^{\perp}\,;\, \boldsymbol{x}_2 = \boldsymbol{x}_o^{\perp}$	2.38	H22
OP	MD_p^v	$\boldsymbol{x}_1 = \boldsymbol{x}_o^{\perp}\,;\, \boldsymbol{x}_2 = \boldsymbol{x}_t^{\perp}$	2.40	H23

$\overline{\boldsymbol{x}}_r$: mean spectrum for the evaluating calibration set spectra

$\boldsymbol{x}_{o/t}$: sample removed from calibration set ($\boldsymbol{x}_o$) or target sample ($\boldsymbol{x}_t$)

$\boldsymbol{x}_o^{\perp}$: orthogonal projection to an estimated regression vector of removed calibration sample

$\boldsymbol{x}_t^{\perp}$: orthogonal projection to an estimated regression vector of target sample

3.6. *Sample Domain to Calibration Domain Comparison Merits*

The calculations for all of the merits described below are for single vectors that are represented as outer products. Vectors can form arrays by either calculating an outer product of two of the same vector (Eqs. 2.43 and 2.44) or two different vectors (Eqs. 2.45 and 2.46).

$$\boldsymbol{x}_{1_{(n\,\mathrm{x}\,1)}}\boldsymbol{x}_{1\,(1\,\mathrm{x}\,n)}^{T} = \boldsymbol{X}_{1_{(n\,\mathrm{x}\,n)}} \tag{2.43}$$

$$\boldsymbol{x}_{\boldsymbol{2}_{(n\,\mathrm{x}\,1)}}\boldsymbol{x}_{\boldsymbol{2}\,(1\,\mathrm{x}\,n)}^{T} = \boldsymbol{X}_{2_{(n\,\mathrm{x}\,n)}} \tag{2.44}$$

$$\boldsymbol{x}_{1_{(n\,\mathrm{x}\,1)}}\boldsymbol{x}_{2\,(1\,\mathrm{x}\,n)}^{T} = \boldsymbol{X}_{12_{(n\,\mathrm{x}\,n)}} \tag{2.45}$$

$$\boldsymbol{x}_{2_{(n\,\mathrm{x}\,1)}}\boldsymbol{x}_{1\,(1\,\mathrm{x}\,n)}^{T} = \boldsymbol{X}_{21_{(n\,\mathrm{x}\,n)}} \tag{2.46}$$

In outer product analysis, introduced by Barros et al.[17], the outer products are first calculated for a pair of vectors. The resulting array is then unfolded by concatenating each row of the array to form one long vector (Fig. 2.3).

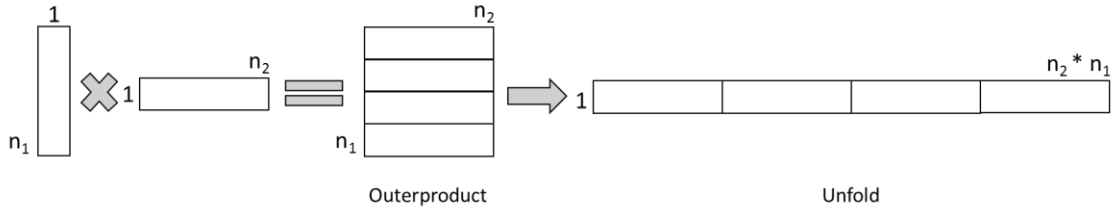


Figure 2.3. Schematic for outer product analysis.

This resulting unfolded vector (1 x $(n_2 * n_1)$) contains all possible product combinations between the two individual vectors. This technique has been used in literature as a data fusion method where two signals can be combined together to demonstrate, not only the individual signal outputs, but the product combinations of the two signals resulting in more robust multivariate models[18]. For this work, the use of outer product analysis can be used to represent vector data as sample domains. Outer products allows for vector data to

be viewed differently using the same types of comparison merits. For the following merits, the determinant calculation is the only merit where the outer products are unfolded as described in the Figure 2.3 schematic. All of the other merits described in this section retain the outer products for the calculations without unfolding. Like the sample vector to calibration vector comparison merits, when calculating SVD's for these outer product arrays only the first eigenvectors and eigenvalue are used as the rank is typically one.

3.6.1 Determinant

The determinant between two outer product arrays for this work uses the determinant calculation in equation 2.24. The outer products for $\boldsymbol{X}_{1_{(n \text{ x } n)}}$ (Eq. 2.43) and $\boldsymbol{X}_{21_{(n \text{ x } n)}}$ (Eq. 2.46) are unfolded to represent vectors $\boldsymbol{x}_{1_{(n*n \; x \; 1)}}$ and $\boldsymbol{x}_{2_{(n*n \text{ x } 1)}}$ respectively in calculating the determinant.

3.6.2 Euclidean Distance

To calculate the Euclidean distance between two outer product arrays each array can be unfolded for the typical Euclidean distance calculation (Eq. 2.23) or the Frobenius norm of the difference between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ can be calculated for the same results (Eq. 2.47).

$$\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|_F \tag{2.47}$$

3.6.3 Procrustes Analysis

The outer products, $\boldsymbol{X}_1$ (Eq. 2.43) and $\boldsymbol{X}_2$ (Eq. 2.44), are used for calculation of the unconstrained PA transformation merit ($\boldsymbol{F}_{21}$).

$$\boldsymbol{X}_1 = \boldsymbol{X}_2 \, \boldsymbol{F}_{21} \tag{2.48}$$

Solving for $\boldsymbol{F}_{21}$ is as follows

$$\boldsymbol{F}_{21} = \boldsymbol{X}_2^{+}(\boldsymbol{X}_1)$$

The pseudoinverse $\boldsymbol{X}_2^{+}$ and $\boldsymbol{X}_1$ can be calculated and represented using SVD, but are also shown in simplified formats in equations 2.49 and 2.50 respectively.

$$\boldsymbol{X}_2^{+} = \boldsymbol{v}\sigma^{-1}\boldsymbol{u}^T = \frac{\boldsymbol{x}_2}{\|\boldsymbol{x}_2\|}\frac{1}{\|\boldsymbol{x}_2\|^2}\frac{\boldsymbol{x}_2^T}{\|\boldsymbol{x}_2\|} = \frac{\boldsymbol{x}_2\boldsymbol{x}_2^T}{\|\boldsymbol{x}_2\|^4} \tag{2.49}$$

$$\boldsymbol{X}_1 = \boldsymbol{u}\sigma\boldsymbol{v}^T = \frac{\boldsymbol{x}_1}{\|\boldsymbol{x}_1\|}\|\boldsymbol{x}_1\|^2\frac{\boldsymbol{x}_1^T}{\|\boldsymbol{x}_1\|} \tag{2.50}$$

The final merit is calculated with the Frobenius norm (2.28) of the difference between $\boldsymbol{F}_{21}$ and $\boldsymbol{F}_{22}$ (Eq. 2.51).

$$F = \|\boldsymbol{F}_{21} - \boldsymbol{F}_{22}\|_F \tag{2.51}$$

where $\boldsymbol{F}_{22}$ is calculated the same as $\boldsymbol{F}_{21}$ except that $\boldsymbol{X}_1$ is replaced with $\boldsymbol{X}_2$ in equation 2.48.

The constrained PA merits ρ_{21} and H_{21} are calculated as

$$\boldsymbol{X}_1 = \rho_{21}\boldsymbol{X}_2\boldsymbol{H}_{21} \tag{2.52}$$

The solutions for ρ_{21} and H_{21} are shown in equations 2.53 and 2.54 respectively.

$$\rho_{21} = \frac{tr(\sigma_{21})}{tr(\boldsymbol{X}_2\boldsymbol{X}_2{}^T)} \tag{2.53}$$

$$\boldsymbol{H}_{21} = \boldsymbol{u}_{21}\boldsymbol{v}_{21}^T \tag{2.54}$$

An SVD is used for the product of the two outer product arrays, $\boldsymbol{X}_2$ and $\boldsymbol{X}_1$ to calculate σ_{21}, $\boldsymbol{u}_{21}$, and $\boldsymbol{v}_{21}^T$ using the first eigenvectors an eigenvalue (Eq. 2.55).

$$\boldsymbol{X}_2{}^T\boldsymbol{X}_1 = \boldsymbol{u}_{21}\sigma_{21}\boldsymbol{v}_{21}{}^T \tag{2.55}$$

The final merit calculations for ρ (Eq. 2.56) and H (Eq. 2.57) are calculated with the Frobenius norm.

$$\rho = \|\rho_{21} - \rho_{22}\|_F \tag{2.56}$$

$$H = \|\boldsymbol{H}_{21} - \boldsymbol{H}_{22}\|_F \tag{2.57}$$

For the calculations of ρ_{22} and $\boldsymbol{H}_{22}$ in the final equations, $\boldsymbol{X}_1$ is replaced with $\boldsymbol{X}_2$ in equation 2.52.

3.6.4 Merit Summary

Table 2.3 has a complete list of the sample domain to calibration domain merits used to select the matrix matched calibration sets. Both orthogonal projections to an estimated regression model vector (OP) and spectral methods (Spectral) are represented. All of the merits listed are used in the cross modeling process.

Table 2.3. Sample domain to calibration domain merits for calibration set comparisons for both Spectral and OP merits. (Notations indicated in footnotes).

Category	Merit	Input Assignments	Equations	Merit ID
OP	Det	$\boldsymbol{X}_1 = \boldsymbol{x}_o^{\perp}\boldsymbol{x}^{\perp T}_o$; $\boldsymbol{X}_2 = \boldsymbol{x}_t^{\perp}\boldsymbol{x}^{\perp T}_t$	2.24	H24
OP	Det	$\boldsymbol{X}_1 = \boldsymbol{x}_t^{\perp}\boldsymbol{x}^{\perp T}_o$; $\boldsymbol{X}_2 = \boldsymbol{x}_t^{\perp}\boldsymbol{x}^{\perp T}_t$	2.24	H25
OP	Euc	$\boldsymbol{X}_1 = \boldsymbol{x}_o^{\perp}\boldsymbol{x}^{\perp T}_o$; $\boldsymbol{X}_2 = \boldsymbol{x}_t^{\perp}\boldsymbol{x}^{\perp T}_t$	2.47	H26
OP	Euc	$\boldsymbol{X}_1 = \boldsymbol{x}_t^{\perp}\boldsymbol{x}^{\perp T}_o$; $\boldsymbol{X}_2 = \boldsymbol{x}_t^{\perp}\boldsymbol{x}^{\perp T}_t$	2.47	H27
OP	F	$\boldsymbol{X}_1 = \boldsymbol{x}_o^{\perp}\boldsymbol{x}^{\perp T}_o$; $\boldsymbol{X}_2 = \boldsymbol{x}_t^{\perp}\boldsymbol{x}^{\perp T}_t$	2.51	H28
OP	F	$\boldsymbol{X}_1 = \boldsymbol{x}_t^{\perp}\boldsymbol{x}^{\perp T}_t$; $\boldsymbol{X}_2 = \boldsymbol{x}_o^{\perp}\boldsymbol{x}^{\perp T}_o$	2.51	H29
Spectral	F	$\boldsymbol{X}_1 = \boldsymbol{x}_{o/t}\,\boldsymbol{x}^{T}_{o/t}$; $\boldsymbol{X}_2 = \overline{\boldsymbol{x}}_r\overline{\boldsymbol{x}}^{T}_r$	2.51	H30
OP	ρ	$\boldsymbol{X}_1 = \boldsymbol{x}_t^{\perp}\boldsymbol{x}^{\perp T}_t$; $\boldsymbol{X}_2 = \boldsymbol{x}_o^{\perp}\boldsymbol{x}^{\perp T}_o$	2.56	H31
OP	ρ	$\boldsymbol{X}_1 = \boldsymbol{x}_o^{\perp}\boldsymbol{x}^{\perp T}_o$; $\boldsymbol{X}_2 = \boldsymbol{x}_t^{\perp}\boldsymbol{x}^{\perp T}_t$	2.56	H32
Spectral	ρ	$\boldsymbol{X}_1 = \boldsymbol{x}_{o/t}\,\boldsymbol{x}^{T}_{o/t}$; $\boldsymbol{X}_2 = \overline{\boldsymbol{x}}_r\overline{\boldsymbol{x}}^{T}_r$	2.56	H33
Spectral	H	$\boldsymbol{X}_1 = \boldsymbol{x}_{o/t}\,\boldsymbol{x}^{T}_{o/t}$; $\boldsymbol{X}_2 = \overline{\boldsymbol{x}}_r\overline{\boldsymbol{x}}^{T}_r$	2.57	H34

$\overline{\boldsymbol{x}}_r$: mean spectrum for the evaluating calibration set spectra

$\boldsymbol{x}_{o/t}$: sample removed from calibration set ($\boldsymbol{x}_o$) or target sample ($\boldsymbol{x}_t$)

$\boldsymbol{x}_o^{\perp}$: orthogonal projection to a regression vector of sample removed from calibration set

$\boldsymbol{x}_t^{\perp}$: orthogonal projection to a regression vector of target sample

3.7. *Sample Vector to Calibration Domain Comparison Merits*

The following merits are used to compare one sample spectrum ($\boldsymbol{x}_1$), either a target sample or calibration sample removed from a calibration set, to an evaluation set ($\boldsymbol{X}_r$). Unlike the sample vector to calibration vector and sample domain to calibration domain comparison merits, the SVD calculations will include multiple basis vectors.

3.7.1 Mahalanobis Distance

Mahalanobis distance is calculated as follows

$$MD = \sqrt{(\boldsymbol{x}_1 - \overline{\boldsymbol{x}}_r)^T \widetilde{\boldsymbol{C}}_r^+ (\boldsymbol{x}_1 - \overline{\boldsymbol{x}}_r)} \tag{2.58}$$

where $\widetilde{\boldsymbol{C}}_{r\,(n\,\mathrm{x}\,n)}^+$ is the pseudoinverse of the covariance matrix $\widetilde{\boldsymbol{C}}_r$, $\overline{\boldsymbol{x}}_{r\,(n\,\mathrm{x}\,1)}$ is the average of the spectra from the evaluation set ($\boldsymbol{X}_{r\,(m\,\mathrm{x}\,n)}$), and $\boldsymbol{x}_{1\,(n\,\mathrm{x}\,1)}$ is the sample spectrum being compared to the evaluation set. The covariance matrix, $\widetilde{\boldsymbol{C}}_r$ is calculated in equation 2.59 and $\widetilde{\boldsymbol{C}}_r^+$ in equation 2.60.

$$\widetilde{\boldsymbol{C}}_r = \frac{\widetilde{\boldsymbol{X}}_r{}^T \widetilde{\boldsymbol{X}}_r}{n-1} \tag{2.59}$$

$$\widetilde{\boldsymbol{C}}_r^+ = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^T \tag{2.60}$$

where $\widetilde{\boldsymbol{X}}_r$ are the mean-centered spectra of the evaluation set. Mean centering implies that the column wise averages for each variable n are subtracted from each individual spectrum's corresponding n variables in the in the evaluation set.

3.7.2 Inner Product Correlation

The inner product correlation is a technique for explaining the correlation between two matrices based on their inner products with the purposes of matrix transformation[19]. This inner product correlation equation is manipulated to compare a vector ($\boldsymbol{x}_1$) to an evaluation set ($\boldsymbol{X}_r$) (2.61).

$$1 - r_1 = 1 - \frac{tr(\boldsymbol{X}_1^{+}\boldsymbol{U}_C\boldsymbol{\Sigma}_C\boldsymbol{V}_C^{T})}{\sqrt{(\sigma^{-1})^2\, tr[(\boldsymbol{\Sigma}_C^{-1})^2]}} \tag{2.61}$$

For variables $\boldsymbol{U}_C$, $\boldsymbol{\Sigma}_C$, and $\boldsymbol{V}_C$, an SVD of the altered covariance matrix ($\boldsymbol{C}_2$), where $\boldsymbol{X}_r$ are not mean-center, is calculated as

$$\boldsymbol{C}_r = \frac{\boldsymbol{X}_r{}^T\boldsymbol{X}_r}{m-1} = \boldsymbol{U}_C\boldsymbol{\Sigma}_C\boldsymbol{V}_C^T$$

The pseudoinverse, $\boldsymbol{C}_r^{+}$, is calculated with SVD as follows

$$\boldsymbol{C}_r^{+} = \boldsymbol{V}_C\boldsymbol{\Sigma}_C^{-1}\boldsymbol{U}_C^T$$

The trace of the singular value diagonal matrix, $\boldsymbol{\Sigma}_C^{-1}$, is used in equation 2.61. The vector ($\boldsymbol{x}_1$) is represented as an outer product array $\boldsymbol{X}_1$ (Eq. 2.43). The variable σ is calculated by the pseudoinverse of the outer product of $\boldsymbol{x}_1$. Unlike the covariance array of the evaluation set, $\boldsymbol{x}_1$ is only rank one and is shown in a more simplistic form.

$$\boldsymbol{X}_1^{+} = \boldsymbol{v}\sigma^{-1}\boldsymbol{u}^T = \frac{\boldsymbol{x}_1\boldsymbol{x}_1^T}{\|\boldsymbol{x}_1\|^2}$$

The final merit (Eq. 2.61) is subtracted from one so the lower values can represent a higher degree of similarity.

3.7.3 Divergence Criteria

The divergence criteria was originally introduced to measure the difference between two probability distributions[20] but, in this case, can be used to compare the differences between two calibration spaces. This equation, like the inner product correlation, can be manipulated to compare a vector ($\boldsymbol{x}_1$) and an evaluation set ($\boldsymbol{X}_r$) (Eq. 2.62).

$$Div = \left| \begin{matrix} \frac{\mathbf{1}}{\mathbf{2}} tr\big((\boldsymbol{X}_1 - \boldsymbol{C}_r)(\boldsymbol{X}_1^{+} - \boldsymbol{C}_r^{+})\big) \\ + \frac{\mathbf{1}}{\mathbf{2}} tr\big((\boldsymbol{X}_1^{+} + \boldsymbol{C}_r^{+})(\boldsymbol{x}_1 - \overline{\boldsymbol{x}}_r)(\boldsymbol{x}_1 - \overline{\boldsymbol{x}}_r)^T\big) \end{matrix} \right| \tag{2.62}$$

The divergence criteria uses the non-mean centered covariance array ($\boldsymbol{C}_r$), the pseudoinverse of the covariance array ($\boldsymbol{C}_2^+$), the outer product array of the spectrum $\boldsymbol{x}_1$ ($\boldsymbol{X}_1$), the pseudoinverse of $\boldsymbol{X}_1$ ($\boldsymbol{X}_1^+$), and the mean of the evaluation set $\overline{\boldsymbol{x}}_r$ spectra all previously defined.

3.7.1 Q Residual and Projection Angle

The Q residual measurement is a common outlier determination merit. In this work, the magnitude of the Q residual between a sample and an evaluation set domain can be used to determine similarity (Eq. 2.63).

$$Q = \left\| \boldsymbol{x}_1^q \right\|^2 \tag{2.63}$$

For the calculation of $\boldsymbol{x}_1^q$ (Eq. 2.64), an SVD is calculated for the evaluation set ($\boldsymbol{X}_r$) (Eq. 2.65).

$$\boldsymbol{x}_1^q = (\boldsymbol{I} - \boldsymbol{V}_r \boldsymbol{V}_r^T)\boldsymbol{x}_1 \tag{2.64}$$

$$\boldsymbol{X}_r = \boldsymbol{U}_r \boldsymbol{\Sigma}_r \boldsymbol{V}_r^T \tag{2.65}$$

The outer product of the eigenvectors ($\boldsymbol{V}_r$) along with an identity matrix ($\boldsymbol{I}$) are used to orthogonally project $\boldsymbol{x}_1$ into the evaluation set spectral space. The sine of the angle between the orthogonally projected spectrum, $\boldsymbol{x}_1^q$, and the original spectrum, $\boldsymbol{x}_1$, are then used as an additional measure of similarity (Eq. 2.66).

$$sin\ \theta = \frac{\|\boldsymbol{x}_1^q\|}{\|\boldsymbol{x}_1\|} \tag{2.66}$$

3.7.2 Merit Summary

Table 2.4 has a complete list of the sample vector to calibration domain merits used to select the matrix matched calibration sets. Only the raw spectral methods (Spectral) are used to calculate these merits. All of the merits listed are represented in the

calibration set methods by the cross modeling techniques. For the merits requiring SVD calculations, the number of PC's required to represent up to 99% of the cumulative variation for each calibration set was initially calculated. The average number of required PC's across all of the calibration sets being compared was the number of PC's used for the merits requiring SVD calculations.

Table 2.4. Sample vector to calibration domain merits for calibration set comparisons for Spectral merits. (Notations indicated in footnotes).

Category	Merit	Input Assignments	Equation	Merit ID
Spectral	MD	$\boldsymbol{x}_1 = \boldsymbol{x}_{o/t}$	2.58	H35
Spectral	$1 - r_1$	$\boldsymbol{x}_1 = \boldsymbol{x}_{o/t}$	2.61	H36
Spectral	Div	$\boldsymbol{x}_1 = \boldsymbol{x}_{o/t}$	2.62	H37
Spectral	Q	$\boldsymbol{x}_1 = \boldsymbol{x}_{o/t}$	2.63	H38
Spectral	$sin\,\theta$	$\boldsymbol{x}_1 = \boldsymbol{x}_{o/t}$	2.66	H39

$\boldsymbol{x}_{o/t}$: sample removed from calibration set ($\boldsymbol{x}_o$) or target sample ($\boldsymbol{x}_t$)

4. Methods for Selecting Matrix Matched Calibration Sets

4.1. *General Process Description for Calibration Set Comparison*

Figure 2.4 shows the process used to identify the appropriate matrix matched calibration set for a target sample. This schematic identifies specifically where and how the cross modeling, calibration set comparison merits, and data fusion are brought together.

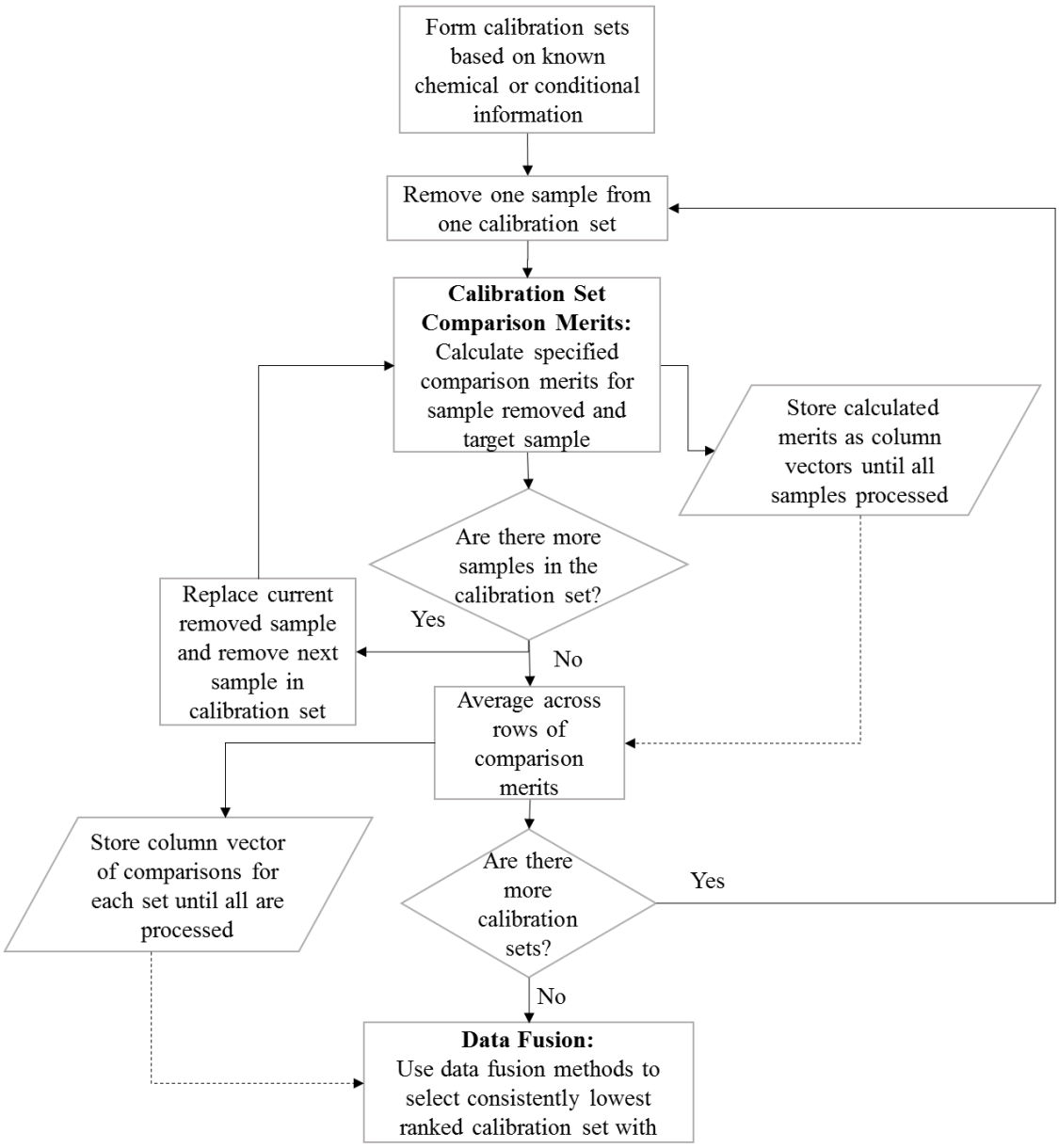


Figure 2.4. Schematic for calibration set selection process

For this work, each of the local calibration sets is manually formed based on the known sample matrix information. After determining the local calibration sets, one sample is removed from the first calibration set. This calibration sample, the target sample, and evaluation sets are used to calculate the matrix matching merits described above along with the cross modeling procedure. This process repeats until all of the

calibration samples for the first calibration set are processed. A data matrix is formed with rows of calibration set comparison merits, including the cross modeling generated merits, and columns of the merits calculated for each of the calibration samples removed from the calibration set. The individual rows of calibration set comparison merits for all of the calibration samples in the calibration set are averaged together into one column vector of comparison merits representing the overall degree of matrix matching for the first calibration set. This process repeats for each calibration set until there is one column of comparison merit data for each calibration set. Figure 2.2 shows this concept for a single comparison merit, e_{21}. Data fusion methods are then used to select which of the calibration sets has the highest degree of matrix matching for the target sample.

4.2. *Fusion Rules*

Multiple fusion methods are used in this work, specifically *SUM* (Eq. 2.67), which calculates the sum of the similarity measures for each column, *MED* (Eq. 2.68), which calculates the median across all similarity measures, and *L2* (Eq. 2.69), which is the Euclidean norm of the vector of similarity measures[21].

$$SUM_g = \sum_{i=1}^{f} S_i \tag{2.67}$$

$$MED_g = \text{med}\{S_1, \ldots S_f\} \tag{2.68}$$

$$L2_g = \sqrt{\sum_{i=1}^{f} (S_i)^2} \tag{2.69}$$

In these equations, S_i denotes each similarity measurement, where f is the number of similarity measures and g is the calibration set number. The raw comparison values (raw) after unit length normalization and the rank values (rank) for each comparison measure across calibration sets are used with each of the fusion methods, resulting in six final rankings. The raw and rank inputs are described in detail in the

Chapter 1 section 3. The rankings from each of these six methods are used in determining the final calibration set. The calibration set that is most consistently (four of the six rules) ranked lowest across the fusion rules is selected. All figures for the fusion ranking methods are shown in the order of SUM_g (raw), SUM_g (rank), MED_g (raw), MED_g (rank), $L2_g$ (raw), and $L2_g$ (rank). As each of the six rules can rank the calibration sets slightly differently, the consensus across all six allows for a confident decision about the matrix matching potential of the calibration set selected.

Each row of merits in the comparison merit matrix is normalized to unit length prior to analysis with the fusion rules. As all of the comparison merits presented have the potential to be different levels of magnitude, normalization is used to weight all of the comparison merits evenly so that one merit does not have more influence than another.

4.3. *Cross Modeling*

The cross modeling method is used for 38 of the 39 merits described above in Tables 2.1-2.4. The merit e_{22} is excluded from the cross modeling process as the purpose of this merit is to assess how well each calibration set predicts its own samples. There is no benefit to using evaluation sets to predict calibration samples without the target sample being included.

4.4. *Latent Variables Selection*

The prediction based merits and orthogonal projection to an estimated regression vector (OP) merits all require tuning parameter selection as these are based on PLS algorithms. Five LV's are used for each calibration set based on equation 2.13. It was determined through preliminary local models that five LV's covered a meaningful amount of information for these datasets. The identified tuning parameter (LV') at the

bottom of the "U-curve" along with four surrounding tuning parameters (two on each side of LV') are how the LV's are selected for all merits requiring LV selection. In situations where the minimum LV (LV') is less than 3, LV's are added on the right for a total of five LV's. For example, if $LV' = 2$ then LV's 1-5 would be used. Using this method of LV model selection, the same five LV's are not required for each of the calibration and evaluation sets.

4.5. *Spectral Preprocessing*

In every instance in this work where a PLS algorithm is used the spectra and the reference value data are mean-centered based on the mean of the calibration spectra and calibration reference values.

4.6. *Matrix Matching Assessment*

For prediction error matching (Eq. 2.10), α_j was calculated for $\hat{y}_j - y$ equivalent to -1, 0 and 1. For the slope matching plots α_j was set to 0, 1, and 2 for predictions of both the individual leave-one-out calibration samples and target sample.

5. Datasets

5.1. *NMR*

^{1}H-NMR spectra ranging from 0.6425-3.8403 ppm were collected for 231 mixtures of three alcohols (propanol, butanol, and pentanol). Each alcohol component had a concentration of 50 mM and was represented at 21 concentration levels from 0-100% in 5% increments[22]. For this work, only the mixtures containing at least 5% of each alcohol component are used (171 samples). Six calibration sets were formed keeping propanol and butanol components within 25% ranges as interferent species using pentanol as the analyte (Table 2.5). One sample from each calibration set was selected at

random as a target sample. The concentrations of the target samples for each species is noted in parenthesis in Table 2.5. The spectral range used was 1.45-1.65 ppm as the peak shifts in this range had the most visible variability. The spectra over this range of peak shifts for the six calibration sets and average spectrum for each calibration set are shown in Figure 2.5.

Table 2.5. NMR spectra alcohol concentrations ranging from 5-90% for each of the three alcohols (pentanol, propanol, and butanol) for six calibration sets.The target sample alcohol concentrations from each set are noted in parentheses.

Cal Set	%Pentanol*	%Propanol	%Butanol	# Samples
1	5-30 (5)	5-30 (15)	65-90 (80)	20
2	5-30 (10)	35-60 (55)	35-60 (35)	20
3	5-30 (30)	65-90 (65)	5-30 (5)	20
4	10-60 (30)	5-30 (25)	35-60 (45)	35
5	40-90 (70)	5-30 (10)	5-30 (20)	35
6	10-60 (25)	35-60 (55)	5-30 (20)	35

*Pentanol is the analyte for this work.

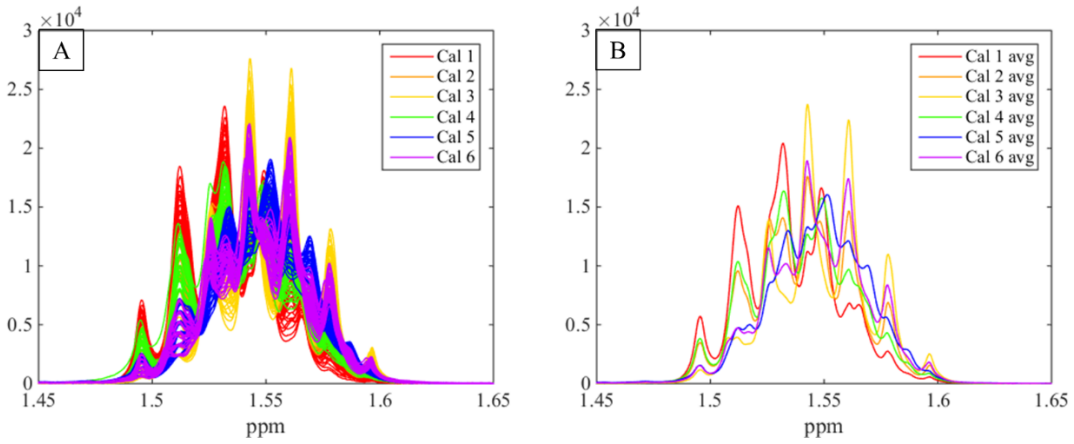


Figure 2.5. NMR calibration spectra for six calibration sets. Individual spectra for each calibration set from 1.45-1.65 ppm 875 chemical shift measurements (A); average calibration for each calibration set from 1.45-1.65 ppm 875 chemical shift measurements (B).

5.2. *Corn*

Near infrared spectra (NIR) for corn ranged from 1100 to 2498 nm at 2 nm intervals for 80 corn samples using three spectrometers (M5, Mp5, and Mp6)[23] (Fig. 2.6). To decrease computation time, 4 nm intervals for the spectra are used in this work, resulting in 350 variables. Reference values were provided for moisture, oil, protein, and starch compositions. Only the moisture and oil values are used. The same 10 samples selected at random from each instrument (30 samples total) are used as target samples. Distributions for the two reference values of the 70 calibration samples and 10 target samples are shown in Figure 2.7. The remaining 210 spectra from all three instruments are used as three spectral calibration spaces to demonstrate matrix matching.

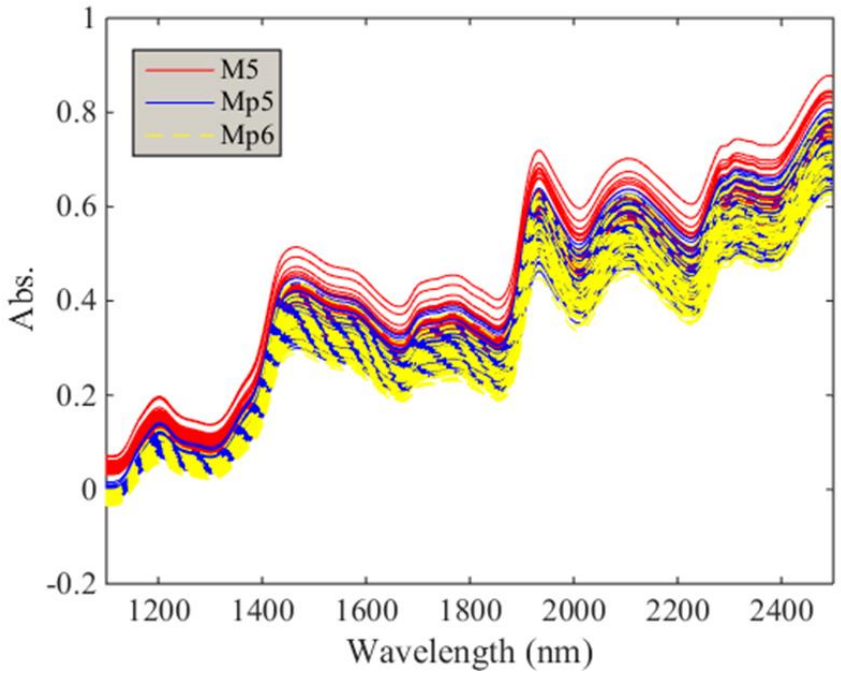


Figure 2.6. Corn instruments M5, Mp5, and Mp6 calibration spectra. Ranging from 1100-2400 nm wavelengths over 350 variables.

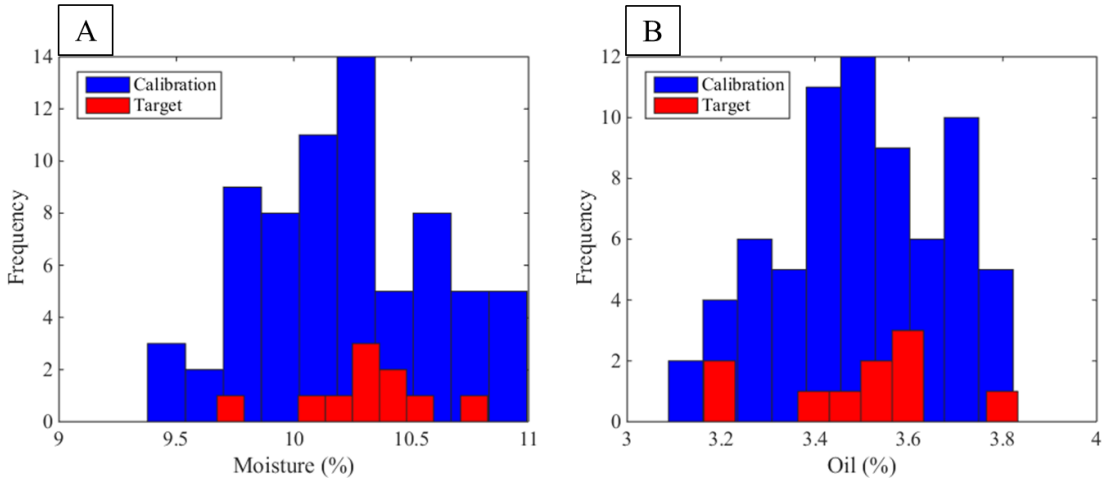


Figure 2.7. Corn reference value calibration and target distributions. Moisture (%) (A), oil (%) (B), distributions for 70 calibration samples (blue) and 10 target samples (red).

6. Results for Selecting Matrix Matched Calibration Sets

6.1. *Matrix Matching*

As mentioned, the nuclear magnetic resonance (NMR) calibration sets were formed by limiting the concentrations of the interferent components (propanol and butanol) to a 25% range. This design is specifically used to demonstrate that, not only does the analyte concentration in the calibration set need to be similar to the target sample, but the interferent concentrations need to be similar for a sample to be truly matrix matched. The NMR spectra of each of the calibration sets in Figure 2.5 show this trend. This is especially true in calibration sets 1-3 where the analyte concentrations are identical while the interferent concentrations vary widely. Each of the chemical shift peaks, because of the nature of the molecular species and NMR itself, is affected by the concentrations of the other alcohol species in the solution.

The target sample from set 5 is used to show matrix matching effects as the analyte value is unique to the set 5 analyte distribution (Table 2.5). The set dependency of target sample 5 makes the matrix matching differences more distinct for the purposes of this work. Figure 2.8 shows both the prediction error and the prediction (slope) plots for target sample 5 predicted by each of the other calibration sets for a single latent variable (LV). For these matrix matching assessment plots, calibration sets 1, 2, 3, and 5 use LV 3, calibration set 4 uses LV 5, and calibration set 6 uses LV 4. These LV's correspond to the LV' discussion in the LV selection methods (section 2.2).

As discussed, the prediction error plots identify matrix matched samples through convergence of α_j around 1 when $|\hat{y} - y|$ is equivalent to 0. The spread between α_j at $|\hat{y} - y| = 1$, giving the shape of the "V", are also important for identifying matrix

matched samples. In Figure 2.8 A5, representing calibration set 5, is the only prediction error plot that meets both of these criteria. All of the target prediction errors, shown in red, have similar α_j's at $|\hat{y}_{j,t} - y_t| = 0$ compared to the calibration prediction error α_j's at $|\hat{y}_{j,o} - y_o| = 0$ that are close to $\alpha_j = 1$. The α_j's for $|\hat{y}_{j,t} - y_t|$=1 in A5 are only similar to the α_j's for $\left|\hat{y}_{j,o} - y_o\right| = 1$ for this one plot. There are a few samples within calibration set 5 that do not follow these matrix matching trends, but the majority of the calibration samples' scaled prediction errors are similar to the target samples prediction errors. These divergent samples would possibly be identified as outliers. No efforts were made for this dataset to remove outliers beforehand, as the purpose was to show general consensus of matrix matching between calibration sets. For calibration set 4, Figure 2.8 A4, there was a general convergence of $\alpha_j's$ around 1 for both $|\hat{y}_{j,o} - y_o| = 0$ and $|\hat{y}_{j,t} - y_t| = 0$. However, the general shape of the "V" (α_j at $\left|\hat{y}_j - y\right| = 1$) does not match between the calibration samples and target sample prediction errors.

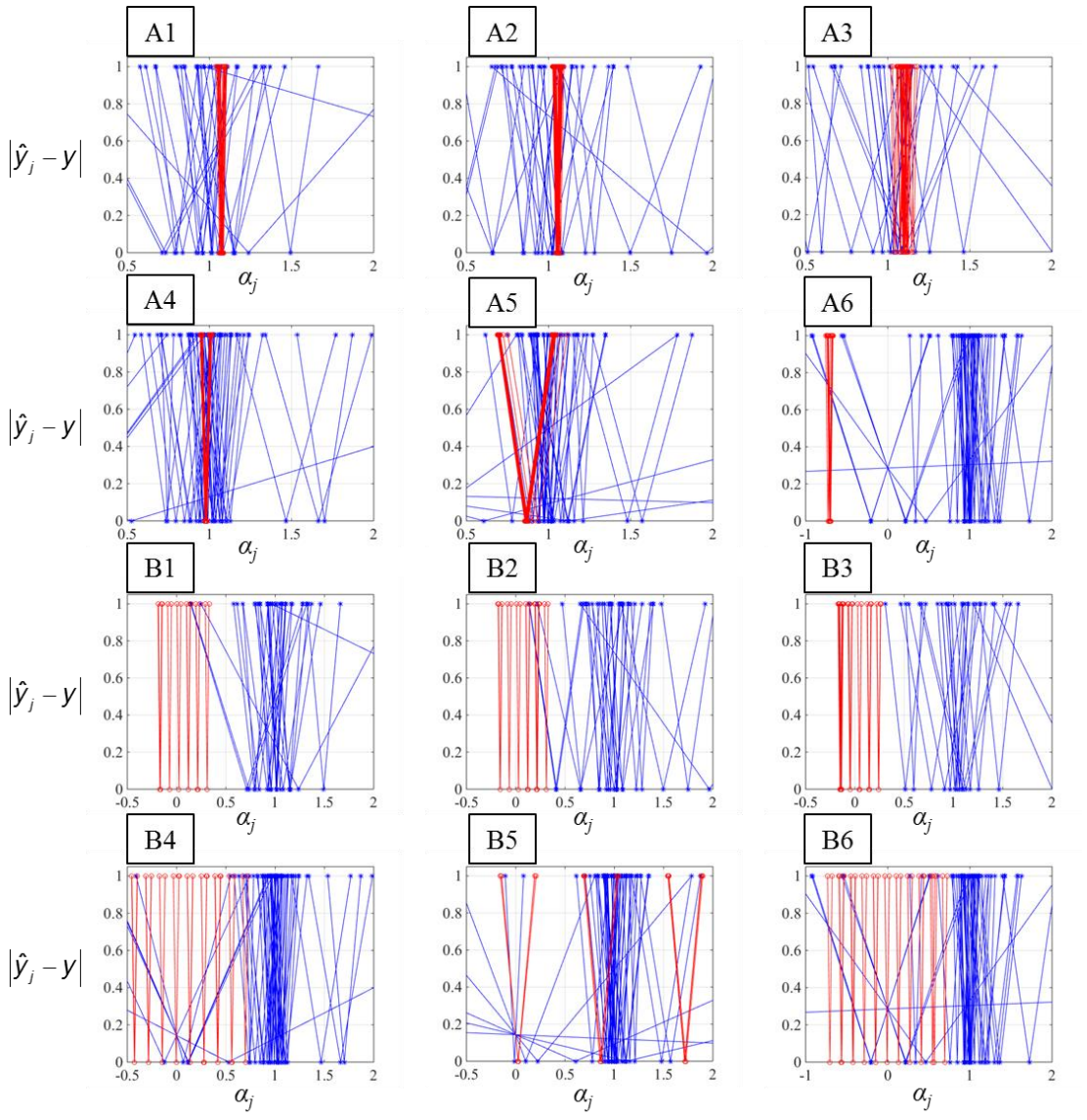


Figure 2.8. NMR prediction error matches for $|\hat{y}_{j,o} - y_o|$ (blue) and $|\hat{y}_{j,t} - y_t|$ (red) (A1-6) and prediction error matches for $|\hat{y}_{j,o} - y_o|$ (blue) and $|\hat{y}_{j,t} - y_o|$ (red) (B1-6) for target (t) sample 5 for the six calibration sets.

Also shown in Figure 2.8 B1-6 are the prediction errors for $|\hat{y}_{j,o} - y_o|$ (blue) and $|\hat{y}_{j,t} - y_o|$ (red). For this set of plots, the true value of each individual calibration sample in the set (y_o) is represented as a proxy for the true value of the target sample in the

$|\hat{y}_{j,t} - y_o|$ scaled prediction errors. Again for these plots, consistent $\alpha_j's$ at $|\hat{y} - y| = 1$ help identify matrix matching. The other indicating factor for matrix matching here is if the $\alpha_j's$ at $|\hat{y}_{j,t} - y_o| = 0$ surround $\alpha_j = 1$ or are near $\alpha_j = 1$. This situation indicates that $\hat{y}_{j,t} \approx y_o$. In these plots, the only case where $\alpha_j's$ at $|\hat{y}_{j,t} - y_o| = 0$ surround $\alpha_j = 1$ is for plot B5, corresponding to calibration set 5.

The prediction slope plots, Figure 2.9, distinctly show that calibration set 5 (A5) is the best set for target sample 5 as all the sample predictions for calibration and target have the same slopes.

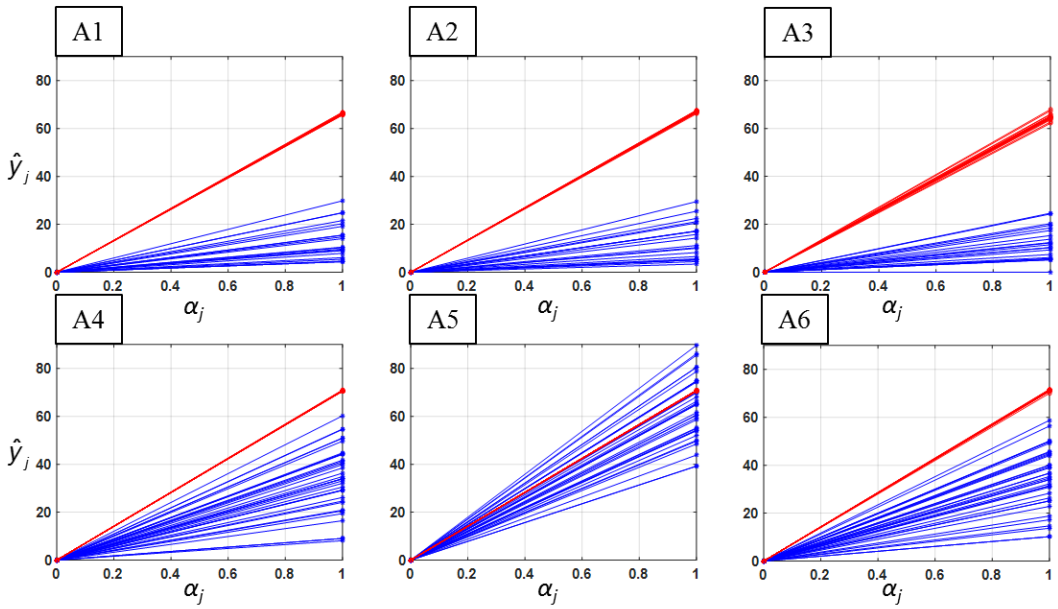


Figure 2.9. NMR prediction slopes for $\hat{y}_{j,o}$ (blue) and $\hat{y}_{j,t}$ (red) (A1-6) target (t) sample 5 for the six calibration sets.

The corn calibration sets are formed based on different instruments. Unlike the NMR, the matrix matching plots for the corn have the same chemical compositions.

However, the instrumental difference, such as wavelength dependent pathlength shifts can cause physical matrix perturbations (P) represented in equation 2.2. Figure 2.10 are the prediction error matrix matching plots for matching the scaled predictions of target sample 1 (measured on instrument M5) to the calibration sets formed by each instrument, M5, Mp5, and Mp6. The LV's represented for these plots are LV 7, 10, and 11 for instruments M5, Mp5, and Mp6 respectively.

The same types of matrix matching trends as the NMR results are seen with the corn dataset prediction errors (Fig. 2.10) and slope plots (Fig. 2.11). The matrix matched calibration set, M5 (Fig. 2.10 A1), has a similar convergence for all α_j's when $|\hat{y}_j - y| = 0$ that are around an $\alpha_j = 1$. Also, this plot shows similar "V" shapes as many of α_j's when $|\hat{y} - y| = 1$ follow the same trends for both $|\hat{y}_{j,o} - y_o|$ (blue) and $|\hat{y}_{j,t} - y_t|$ (red). In the other two plots (A2 and A3) $|\hat{y}_{j,t} - y_t|$ does not resemble the calibration prediction errors, ($|\hat{y}_{j,o} - y_o|$).

In plots B1-3 the prediction differences between the predicted target sample and the individual calibration samples' true values are shown. The M5 calibration set (B1) is the only plot where the α_j's at $|\hat{y}_{j,t} - y_o| = 0$ are around $\alpha_j = 1$.

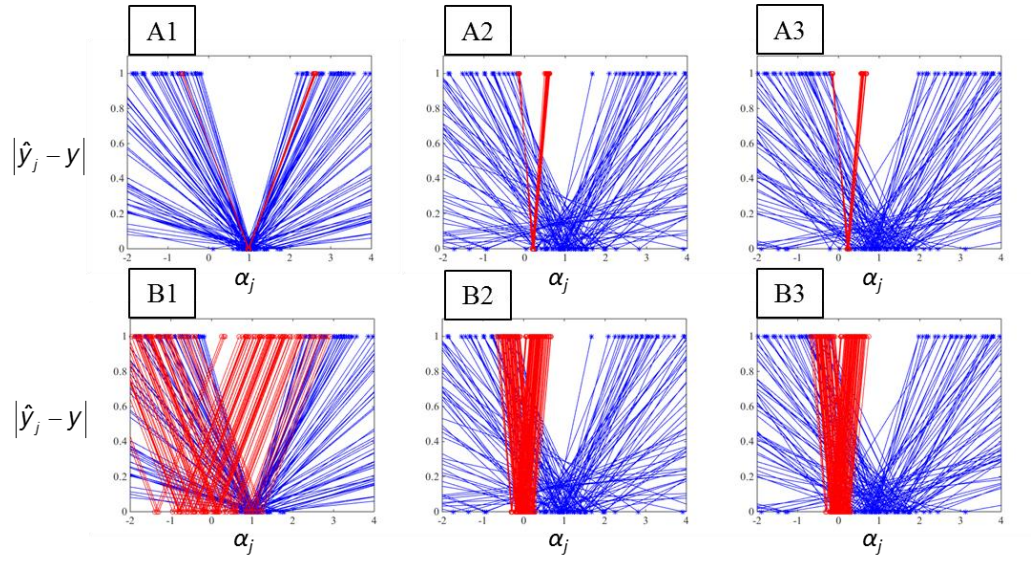


Figure 2.10. Corn prediction error matches $|\hat{y}_{j,o} - y_o|$ (blue) and $|\hat{y}_{j,t} - y_t|$ (red) (A1-3) and prediction error matches for $|\hat{y}_{j,o} - y_o|$ (blue) and $|\hat{y}_{j,t} - y_o|$ (red) (B1-3) for target (t) sample 1 for the instrumental calibration sets (M5 (1), Mp5 (2), and Mp6 (3)).

Instrument M5 is also identified as the best matrix matched calibration set by the slopes of the predictions between the calibration samples and target sample (Figure 2.11 A1).

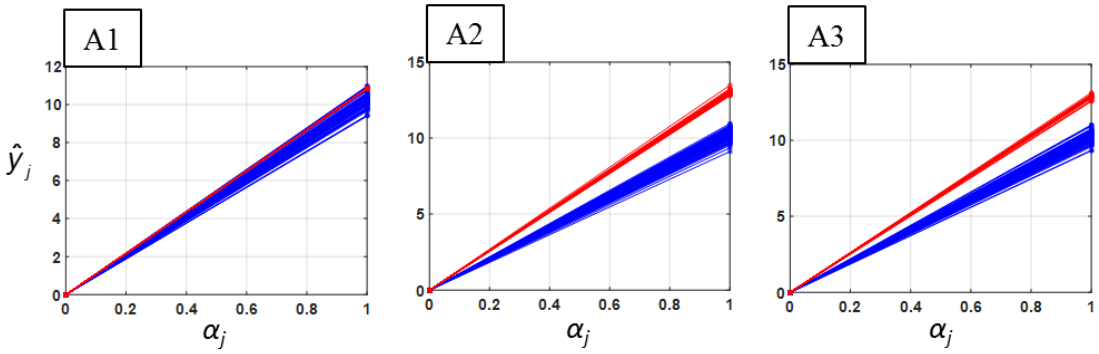


Figure 2.11. Corn prediction slopes for $\hat{y}_{j,o}$ (blue) and $\hat{y}_{j,t}$ (red) (A1-3) target (t) sample 1 for the instrumental calibration sets (M5 (1), Mp5 (2), and Mp6 (3)).

The prediction errors with scalar influence for both the prediction errors and slope matrix matching plots very clearly show which calibration set is matrix matched for both the NMR dataset and the corn dataset selected target samples. The combination of the other merits proposed in this work should be able to similarly identify the same calibration sets.

6.2. *Calibration Set Comparison Merits Calibration Set Selection*

The calibration set comparison merits calculated for NMR target sample 5 and each calibration sample from each calibration set with the cross modeling method are included in Table 2.1-2.4. The calibration set comparison merits are normalized to unit length across the rows for each of the six calibration sets are shown in Figure 2.12. Corresponding rows identifying the individual merits for this figure are listed in Table 2.6. The Y merits span from rows 1-35, OP merits from 36-695, and Spectral merits from 696-1295. The 5 LV's shown for the Y and OP merits ranged from 1-5 for calibration sets 1, 2, 3, and 5, 3-7 for calibration set 4, and 2-6 for calibration set 6. Principal components 1 through 18 are used in order to represent on average up to 99% of the cumulative spectral variability for the five sample vector to calibration domain Spectral merits requiring principal component selection.

For Figure 2.12, the minimum values for each merit represents a higher degree of matrix matching. In general, most of the merits agree that calibration set 5 is the best matrix matched calibration set for target sample 5. The fusion ranking methods (Fig. 2.13 A) for these similarity merits agree with this visual interpretation that calibration set 5 is the lowest rank calibration set for each of the six data fusion ranking methods. There are instances throughout the merits where specific LV's, PC's, or merit representations do

not identify calibration set 5 as the best calibration set. For example from rows 366-425, calibration set 2 appears to have the lowest merit values. These rows correspond to the Procrustes analysis dilation merits, ρ, calculated for the sample domain to calibration domain comparison for the OP merits (Table 2.6). These merits show that the dilation required on average for the orthogonal projections to the corresponding model regression vector of each calibration sample in calibration set 2 and calibration set 5 resemble dilation measurements for the orthogonal projections of target sample 5 into each of these calibration spaces respectively. This merit is an example of why only using one similarity merit gives a limited perspective for determining similar calibration sets/samples for matrix matching purposes. Using all 39 merits with cross modeling (for 38 of the 39 merits) over multiple tuning parameters helps to provide a full picture of how target sample 5 is matrix matched to each calibration set.

These merits can also show which calibration sets are similar to one another. Calibration sets 2 and 6 appear to have similar merit trends. Looking at both the spectra (Fig. 2.5) and the chemical profiles (Table 2.5), sets 2 and 6 have overlapping peak shifts and the pentanol (analyte) and propanol concentration ranges overlap.

Figure 2.13 B and C show the prediction errors of the target sample 5 for each calibration set across the 5 selected LV's. Figure 2.13 B1 are the prediction errors shown on a logarithmic scale to help with visual interpretation. Calibration set 4 shows that specific models for the LV's represented predict the target sample slightly better in comparison to the predictions errors for the 5 LV's represented by the calibration set 5 models. These models from calibration set 4 might represent chance predictions. In (Fig. 2.8 A4) the predictions error for the target samples calibration set 4 were similar to the

calibration samples at $|\hat{y}_j - y| = 0$; however, the prediction errors between the target and the calibration samples did not scale the same with different α_j's. The prediction errors for calibration set 5 are still relatively low in comparison to the other four calibration sets.

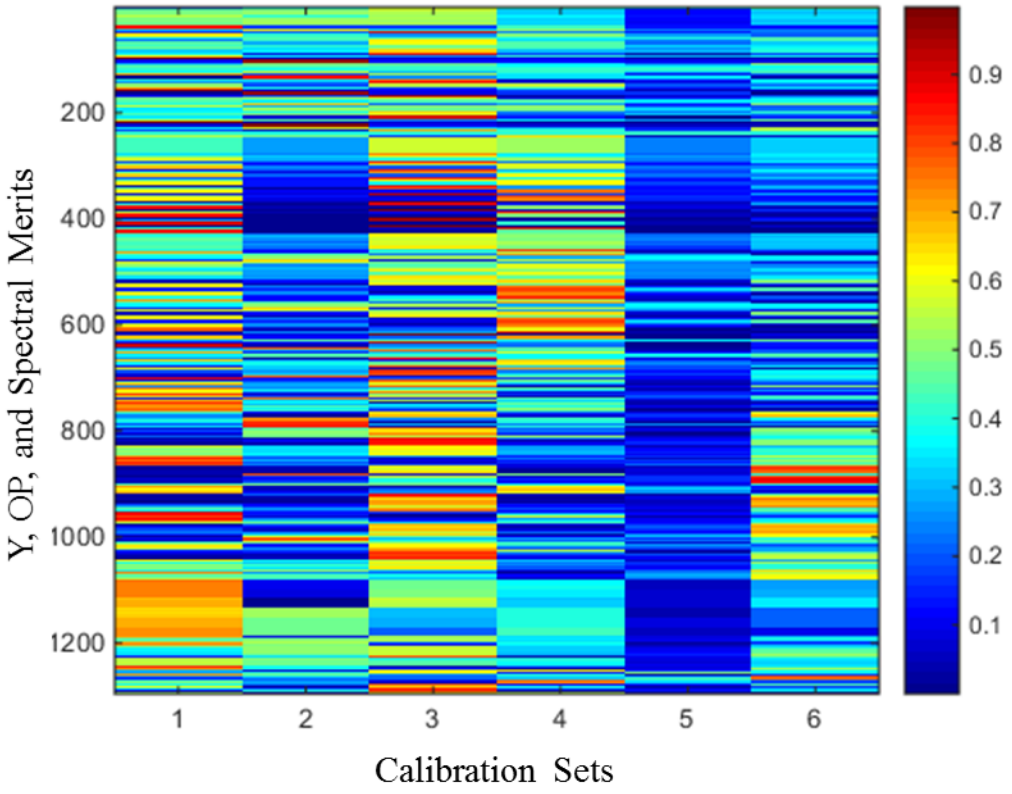


Figure 2.12. NMR Y, OP, and Spectral comparison merits of target sample 5 for each calibration set.

Table 2.6. Calibration set comparison merits and corresponding rows for all corn and NMR merit target samples.

Merit Category	Merit	Merit ID[a]	Rows (NMR)	Rows (Corn)
Y	e_{22}	H1	1-5	1-5
	e_{12}	H2	6-35	6-20
OP	Euc	H4	36-65	21-35
	Euc	H26	66-95	36-50
	Euc	H27	96-125	51-65
	Det	H6	126-155	66-80
	Det	H24	156-185	81-95
	Det	H25	186-215	96-110
	F	H8	216-245	111-125
	F	H28	246-275	122-140
	F	H29	276-305	141-155
	ρ	H9	306-335	156-170
	ρ	H10	336-365	171-185
	ρ	H31	366-395	186-200
	ρ	H32	396-425	201-215
	$EISC\ Xb_d$	H13	426-455	216-230
	$EISC\ Xb_d$	H14	456-485	231-245
	$EISC\ b_d$	H17	486-515	246-260
	$EISC\ b_d$	H18	516-545	261-275
	$EISC\ b$	H11	546-575	276-290
	$EISC\ b$	H12	576-605	291-305
	MD_p^v	H23	606-635	306-320
	MD^v	H21	636-665	321-335
	MD^v	H22	666-695	336-350
Spectral	$1 - cos^2\theta$	H3	696-701	351-353
	Euc	H5	702-707	354-356
	Det	H7	708-713	357-359
	F	H30	714-719	360-362
	ρ	H33	720-725	363-365
	H	H34	726-731	366-368
	$EISC\ Xb_d$	H15	732-737	369-371
	$EISC\ Xb_d$	H16	738-743	372-374
	$EISC\ b_d$	H19	744-749	375-377
	$EISC\ b_d$	H20	750-755	378-380
	MD	H35	756-863	381-422
	Q	H38	864-971	423-464
	$sin\,\theta$	H39	972-1079	465-506
	$1 - r_1$	H36	1080-1187	507-548
	Div	H37	1188-1295	249-590

[a]Refer to Tables 2.1, 2.2, 2.3, and 2.4 for equations relative to each Merit ID

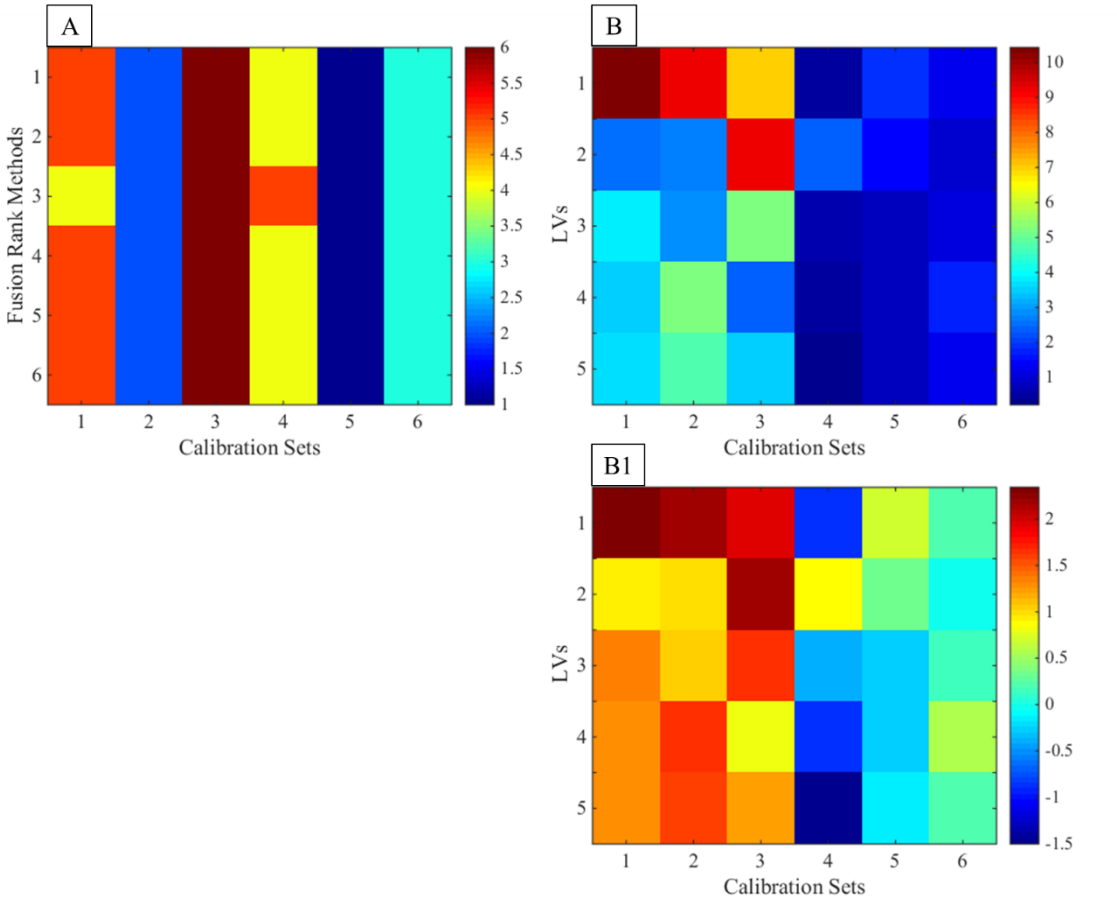


Figure 2.13. NMR fusion rankings and model target prediction errors for target 5. Six fusion ranking method ranks for each of the six calibration sets for target sample 5 (A); RMSEV's for target sample 5 across 5 selected LV's for PLS models formed by each of the 6 calibration sets (B); plot B shown on a logarithmic scale (B1).

The calibration set comparison merits, fusion ranks, and model prediction errors for the six target samples from the NMR dataset are shown in Figure 2.14. Target samples 1, 2, 3, 5, and 6 all consistently had a majority of the comparison merits agree on the correct respective calibration set as being the best matrix matched set. From the RMSEV's (Fig. 2.14 C and C1) many of the calibration sets result in similar predictions

of the target samples regardless of a matrix matched calibration set, except for target sample 5, where sets 4-6 had noticeably lower prediction errors.

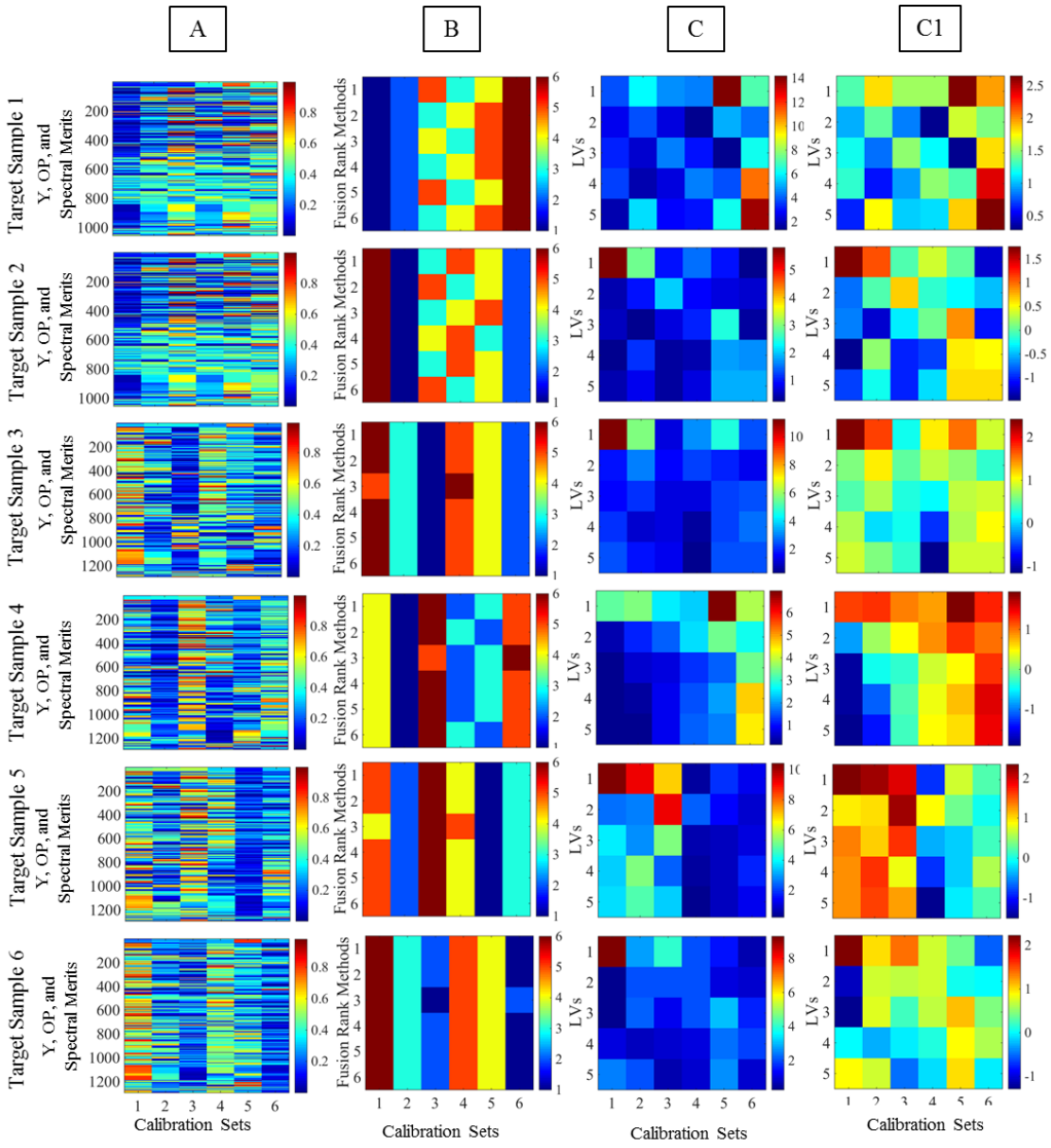


Figure 2.14. NMR calibration set comparison merits (A), fusion rankings (B), RMSEV's (C) for all 6 target samples for each calibration set, and plot C on a logarithmic scale (C1).

Target sample 4 similarity merits did not identify calibration set 4 as the correct set (it was ranked 2nd). Looking at the matrix alcohol concentrations between sets 2 and 4 (Table 2.5), both the analyte (pentanol) and butanol concentration ranges overlap. The propanol concentrations do not overlap, but target sample 4 is on the threshold between the concentration levels of propanol for these two sets at 25%, which is between 5-30% in set 4 and close to 35-60% in set 2. The analyte range was also wider for calibration set 4 than for set 2. As mentioned in Chapter 1, a pair of matrix matched samples should have both spectral and chemical similarities for true matching[24]. The definition of chemical similarity consists of having limited chemical ranges for all species in the matrix. As calibration set 2 has a smaller chemical range, it is more matrix matched at some degree to target sample 4 than the larger analyte range in calibration set 4.

For the corn dataset, the 39 comparison merits are calculated for the three instrument calibration sets for target sample 1 measured on the M5 instrument. The 2 Y merits (rows 1-20), 22 OP merits (rows 21-350), and 15 X merits (rows 351-590) from Table 2.6 are represented in Figure 2.15.

The LV's for the Y and OP merits range from 5-9 for M5, 8-12 for Mp5 and 9-13 for Mp6. Principal components represented are 1 to 14 for the five Spectral merits requiring PC selection. For target sample 1 the merits appear to consistently identify calibration set M5 as the best matrix matched set. This visual identification is supported by the consistency across the six fusion rankings and justified by the target prediction errors in (Fig. 2.16 A and B).

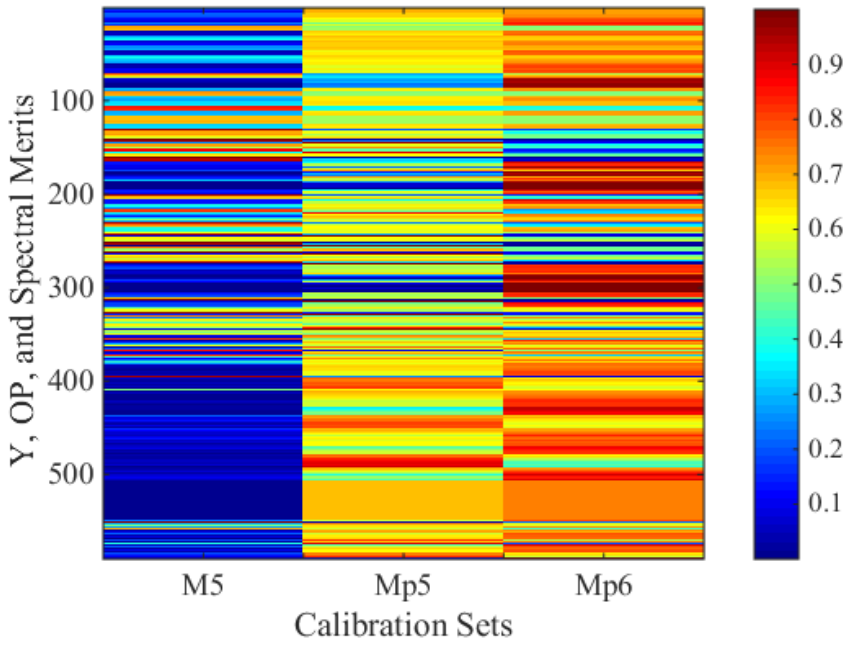


Figure 2.15. Y, OP, and Spectral merits of target sample 1 for each calibration set M5, Mp5, and Mp6.

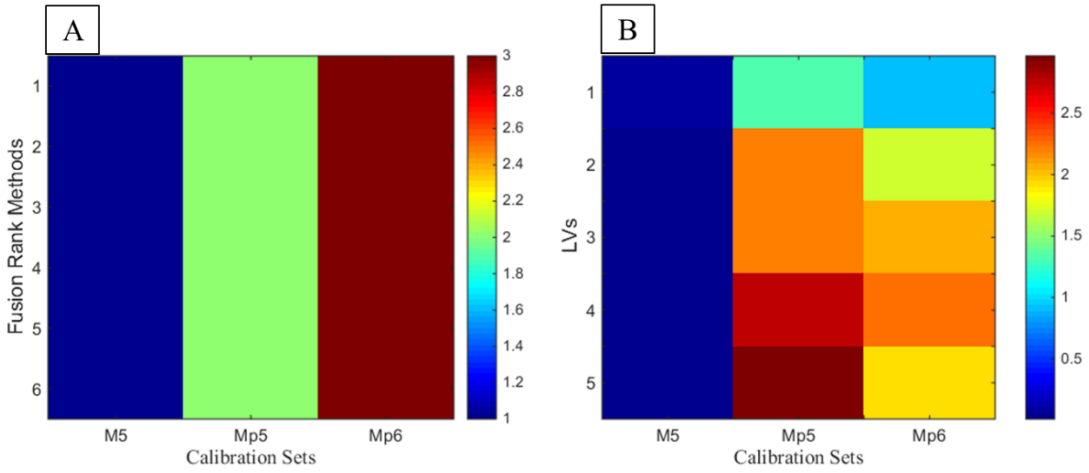


Figure 2.16. Corn fusion rankings and model prediction errors for target sample 1. Six fusion ranking method ranks for each of the six calibration sets for target sample 1 (A); RMSEV for target sample 1 across 5 selected LV's for PLS models formed by each of the three instrumental calibration sets (B).

There are a few merits (or specific tuning parameters calculated for a single merit) where Mp5 or Mp6 has lower values than M5. However these rows do not correspond to the Procrustes analysis dilation merit, ρ, for the OP merits as for the NMR dataset. This indicates that this merit is still valid for assessing calibration set matrix matching potential.

Figures 2.17 and 2.18 show the calibration set comparison merits for all 30 target samples from each of the three instruments (10 of the same samples removed from each instrument dataset) for the corn dataset moisture (%) reference values. Table 2.7 shows the resulting fusion rank calibration set selections based on consistently minimum rankings for the six fusion rules. The target samples from Mp5 and Mp6 instruments show that this same dilation merit, ρ, has lower values for the Mp5 and Mp6 instrument calibration samples as would be expected.

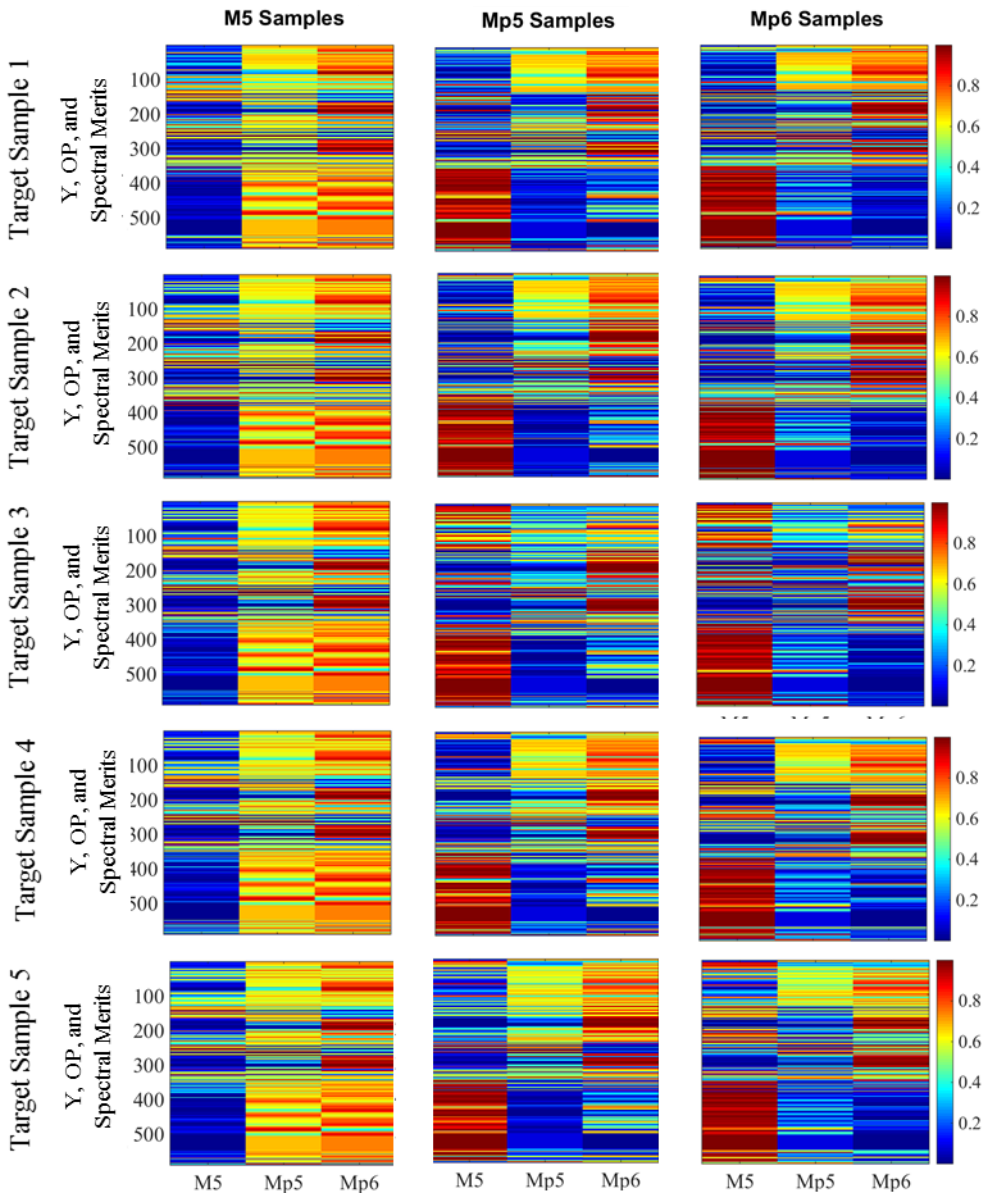


Figure 2.17. Corn calibration set comparison merits for target samples 1-5 from each instrument for moisture (%) reference value.

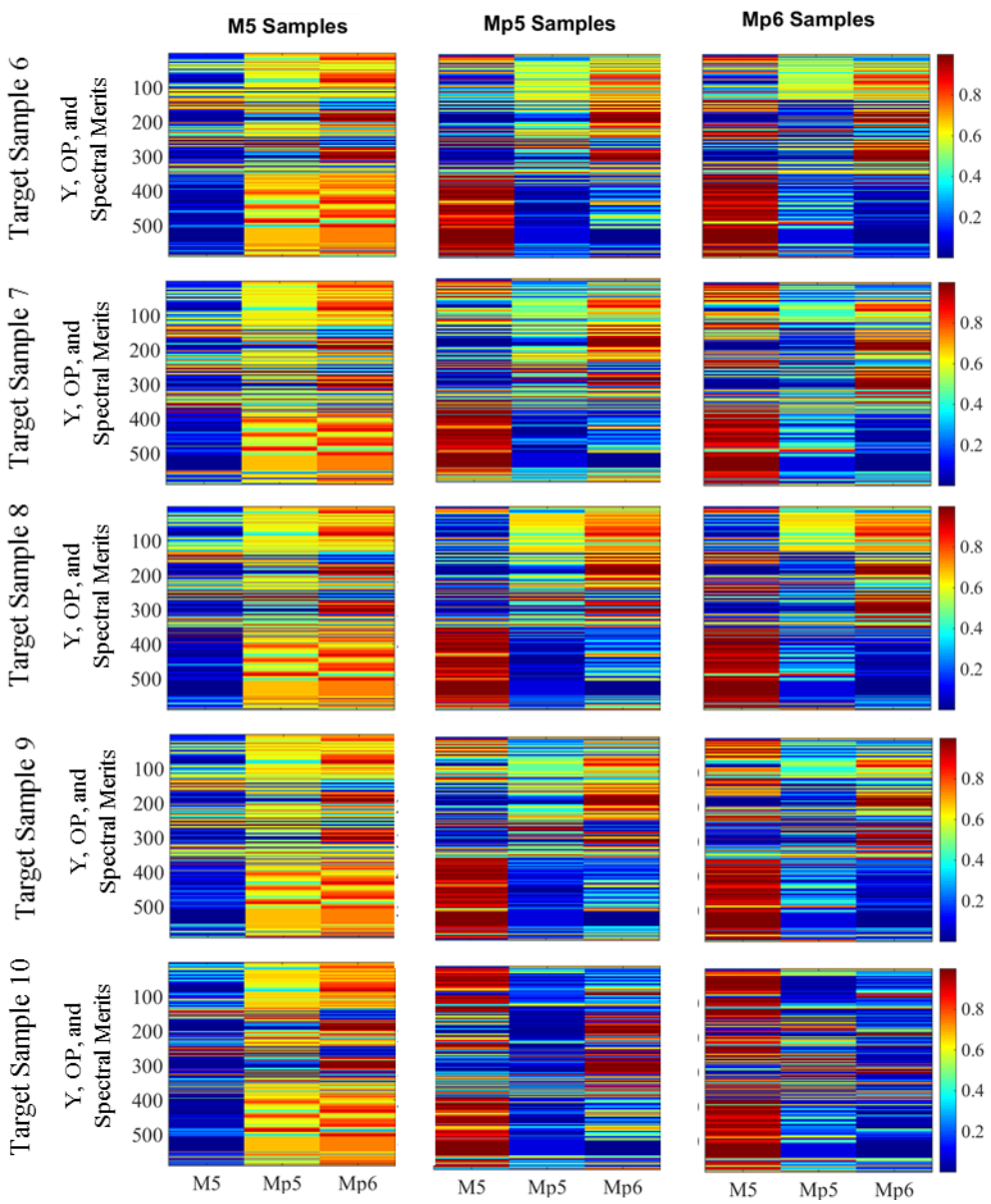


Figure 2.18. Corn calibration set comparison merits for target samples 6-10 from each instrument for moisture (%) reference value.

The target samples from M5 and Mp5 are all consistently identified to matrix match their correct corresponding instrument based on Table 2.7. The target samples from Mp6, using the calibration set comparison and fusion ranking rules, have calibration set Mp5 selected as the better matrix matched calibration set. Because of the nature of the

ranking process, when two calibration sets have the same or very similar the calibration set comparison merit input values they can mathematically be assigned the same rank; however, the first calibration set numerically is assigned the lower rank value as there can only be one calibration set this assigned a rank of 1. As the Mp5 and Mp6 instruments are very closely related, seen spectrally in Figure 2.6, the regression prediction based merits (Y merits and the OP merits) were often very similar to one another for the two instruments. This trend is seen in the Mp5 and Mp6 target sample merit inputs for many of the target samples (Fig. 2.17 and 2.18). The Spectral merits in the case of Mp5 and Mp6 samples were much more efficient at identifying the unique differences between the Mp5 and Mp6 calibration samples in comparison to the corresponding target samples. This indicates that the prediction models between Mp5 and Mp6 calibration sets are very similar even with the slight spectral discrepancies.

The combination of spectral and prediction based merits need to be used together to assess the best matrix matched calibration set to represent both spectral and chemical similarities. As the correct calibration set is not selected for Mp6 this could indicate that there is more weight representing the Y and OP merits due to the number of merits. For this dataset around 60% of the merits are based on the predictive regression models and 40% are spectrally based. As the number of spectral merits fluctuates with the number of PC's required to account for up to 99% of the cumulative variation, it can be difficult to maintain an even balance of merit types for each dataset. Methods for balancing these merits is not investigated in this work but should be a considered in future work.

Table 2.7. Fusion rank calibration set selection for each of the 10 target samples from each instrument for moisture (%) reference value.

Sample	M5	Mp5	Mp6
1	M5	Mp5	Mp5
2	M5	Mp5	Mp5
3	M5	Mp5	Mp5
4	M5	Mp5	Mp5
5	M5	Mp5	Mp5
6	M5	Mp5	Mp5
7	M5	Mp5	Mp5
8	M5	Mp5	Mp5
9	M5	Mp5	Mp5
10	M5	Mp5	Mp6
Correct/total	10/10	10/10	1/10

Figures 2.19 and 2.20 and Table 2.8 contain the calibration set comparison merits and calibration set selection results for all ten target samples measured on the three instruments for the reference property oil (%) for the corn dataset. As already discussed, the merits that most clearly identify the correct calibration set were the Spectral merits (rows 351-590). These merits remain constant for all reference values and therefore show the same trends as Figures 2.17 and 2.18. Even though M5 instrument target samples all have the correct calibration set selected, there seemed to be more overlap between the Y merits and orthogonal project merits for calibration sets M5 and Mp5 than seen with the moisture (%) reference values. This could indicate that the spectral regions that are most influential to predicting oil (%) do not vary as greatly between these two instruments. The correct calibration sets are ultimately identified and selected for all M5 and Mp5 target samples. The ability of this process to select matrix matched calibration sets across

multiple reference values indicates the adaptable nature of the algorithm. Though this particular dataset could be correctly identified with the Spectral merits alone, this will not be the case for every dataset.

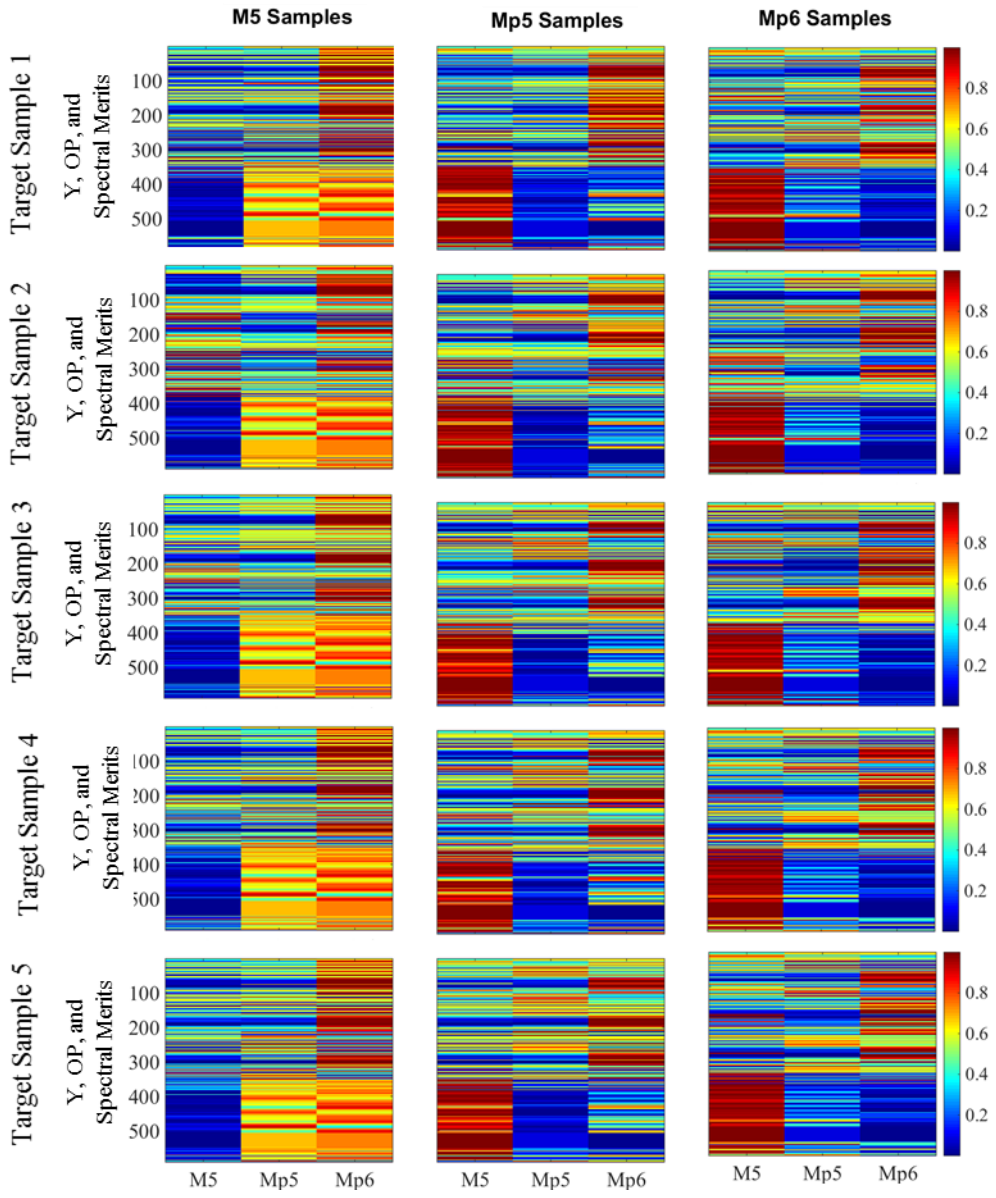


Figure 2.19. Corn calibration set comparison merits for all samples 1-5 from each instrument for oil (%) reference value.

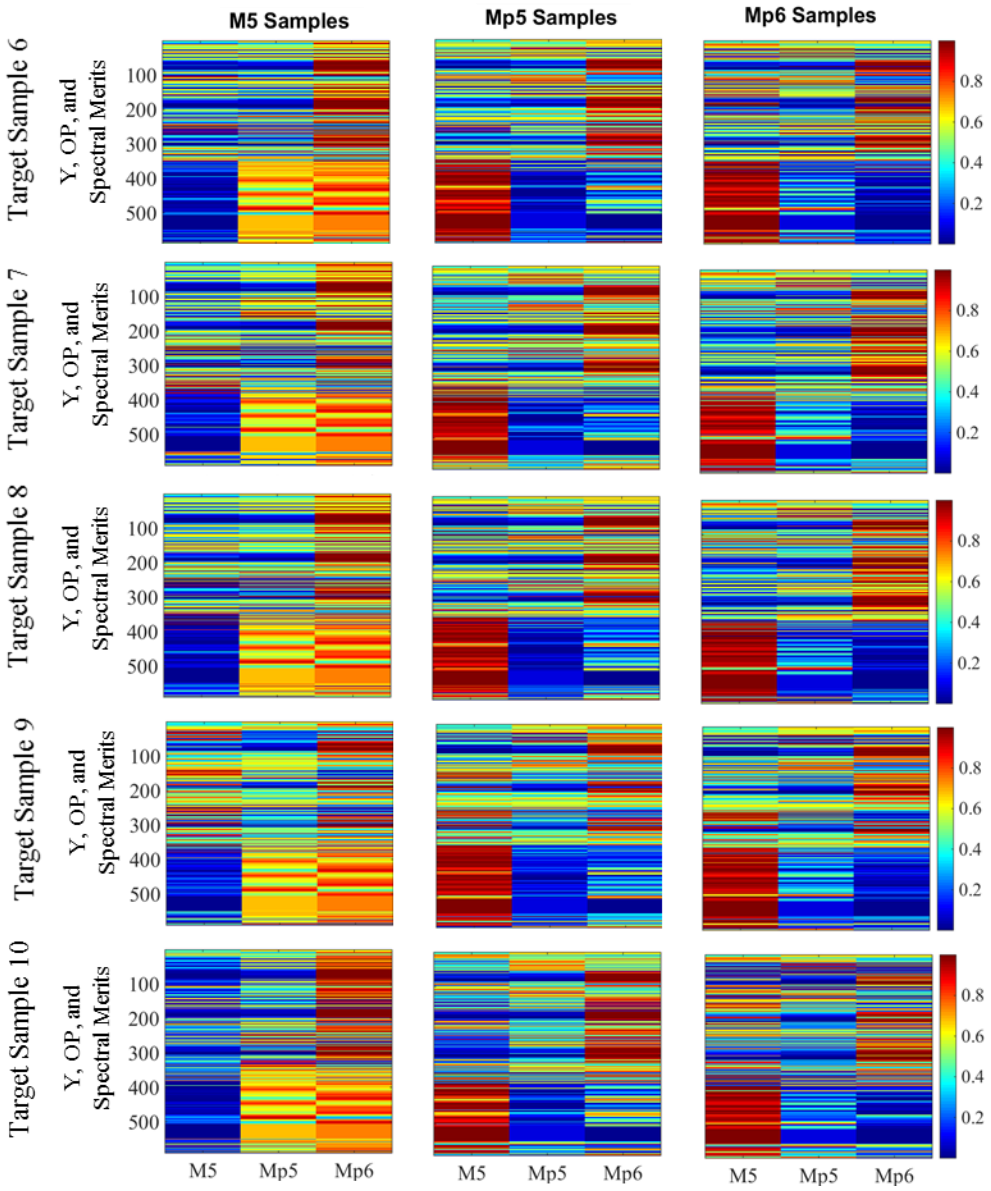


Figure 2.20. Corn calibration set comparison merits for all samples 6-10 from each instrument for oil (%) reference value.

Table 2.8. Fusion rank calibration set selection for each of the 10 target samples from each instrument for oil (%) reference value.

Sample	M5	Mp5	Mp6
1	M5	Mp5	Mp5
2	M5	Mp5	Mp5
3	M5	Mp5	Mp5
4	M5	Mp5	Mp5
5	M5	Mp5	Mp5
6	M5	Mp5	Mp5
7	M5	Mp5	Mp5
8	M5	Mp5	Mp5
9	M5	Mp5	Mp5
10	M5	Mp5	Mp5
Correct/total	10/10	10/10	0/10

Tables 2.7 and 2.8 indicate that calibration set Mp5 is a better matrix match to the Mp6 target samples than the Mp6 calibration set. If both the Mp5 and Mp6 instruments predict the target samples from Mp6 similarly based on similar linear regressions than it does not matter if the Mp5 instrument is the selected matrix match calibration set. Figure 2.21 shows the predictions versus the true values for the Mp6 target samples for each of the three instruments for moisture and oil prediction properties. Table 2.9 has the regression merits for these predictions. The LV's represented for each of the models are based on equation 2.13 for LV selection of LV'. For the moisture property the prediction error is much greater for the Mp5 predictions than the Mp6 predictions (0.65 vs. 0.16); however, the R^2 values, slope and intercept between linear regressions formed by these two sets is similar. For the property oil, all three regressions from Figure 2.21 and Table 2.9 are similar to one another for all three regression models. The current prediction based merits for calibration set comparison were not able to differentiate these true

prediction differences between the Mp5 and Mp6 calibration regressions for the 10 target samples for either prediction property.

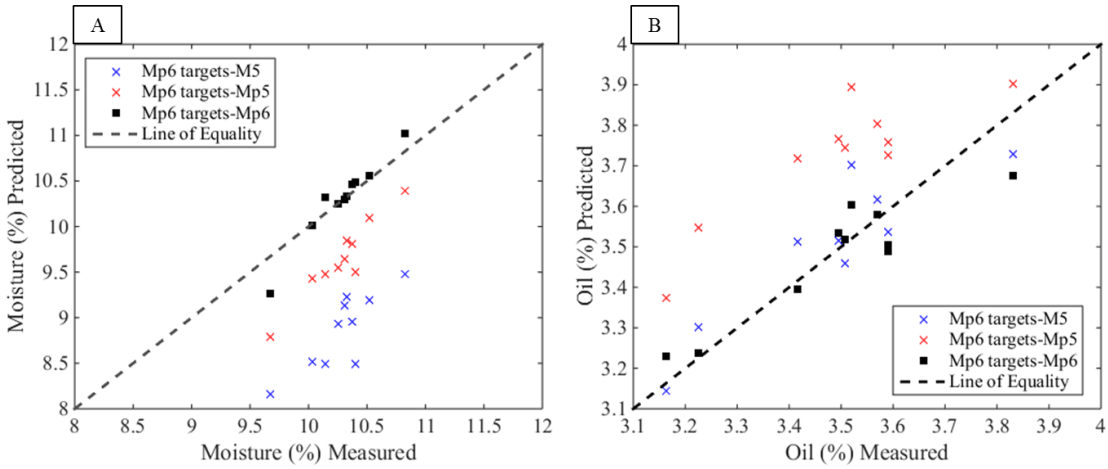


Figure 2.21. Predicted moisture (%) values and oil (%) values for Mp6 target samples versus true measured moisture (%) (A) and oil (%) (B) values using regression model built with each of the three instrument calibration sets.

Table 2.9. Regression statistics for Mp6 target sample moisture (%) and oil (%) predictions versus the true measured moisture (%) and oil (%) values for each of the three instrument calibration sets.

Property	Model	LV	RMSEV	R^2	Intercept	Slope
Moisture	M5	7	1.44	0.70	-3.13	1.17
	Mp5	10	0.65	0.91	-4.25	1.35
	Mp6	11	0.16	0.95	-4.51	1.44
Oil	M5	10	0.10	0.77	0.68	0.81
	Mp5	10	0.25	0.77	1.17	0.73
	Mp6	12	0.08	0.86	0.97	0.71

7. Conclusion

The 39 merits, made up of regression model prediction based merits, orthogonal projections to estimate regression vectors, and raw spectral comparisons, in combination with cross modeling and multiple tuning parameters, are sufficient in identifying matrix matched calibration sets for fixed calibration sets with known chemical and physical matrices, such as the NMR and corn data sets. The matrix matched calibration set is correctly identified for many of the target samples presented in this study. Data fusion methods are used to automatically identify matrix matched calibration sets for the target spectrum for matrix difference due to both analyte and interferent compositions and instrumental differences. The corn data set demonstrates that the balance between prediction based (Y and OP) and raw spectral (Spectral) merits is important depending on the matrix differences between the different calibration sets. The type of matrix effects (instrumental versus chemical) plays a role in determining which calibration set comparison merits should be used in this process. If the differences in matrix effects are known to only be based on instrumental differences and not chemical, then spectral comparison merits would be able to identify the matrix matched calibration set, at least based on the corn dataset results.

As a result of the data fusion methods used, this process is not limited by how many rows of comparison data can be used, the comparison techniques and merits represented for these two data sets can be further expanded or limited as needed. The main limitation is computational time. The increase in array sizes and numbers of comparison merits does increase the time to calculate all of the cross modeling steps and fusion method calculations.

Another set of comparison merits, not currently included in the process, would be to apply variable (e.g. wavelength) selection techniques. Each merit would be calculated for each of the wavelength ranges selected. In the data shown in this work, the full wavelengths and chemical shift ranges are used. Additional rows for each merit would result from using multiple varying sizes of variable windows for subsets of the calibration set. Multiple sets of random variables (wavelengths) would also be an option. Another method of expanding or changing the ratios of the different merit types, would be to use another type of regression model vector ($\widehat{\boldsymbol{b}}$) algorithm. The examples above were based on $\widehat{\boldsymbol{b}}$ calculated from a PLS regression models. Principal component regression, Tikhonov regularization[25], or ridge regression[26] could also be used instead of or in addition to PLS.

Overall, this work demonstrates the utility of data fusion with various types of comparison merits for automatically identifying matrix matched calibration sets without the use of reference values for target samples. The next step would be to use these same comparison merits to evaluate calibration sets where the matrix differences are unknown.

8. References

1. Anderssen R, S.; Osborne B, G.; Wesley I, J., The application of localisation to near infrared calibration and prediction through partial least squares regression. *Journal of Near Infrared Spectroscopy* **2003,** *11* (1), 39-48.
2. Cho, T.; Kida, I.; Ninomiya, J.; Ikawa, S.-i., Intramolecular hydrogen bond and molecular conformation. Part 2.—Effect of pressure and temperature on the IR spectra of some hydroxy ketones. *J. Chem. Soc., Faraday Trans.* **1994,** *90* (1), 103-107.
3. Hildrum, K. I., *Near infra-red spectroscopy: Bridging the gap between data analysis and NIR applications*. Ellis Horwood Ltd: 1992.
4. Kamiya, N.; Sekigawa, T.; Ikawa, S.-I., Intramolecular hydrogen bond and molecular conformation. Part 1.—Effect of pressure and temperature on the infrared spectra of 4-hydroxy-4-methylpentan-2-one (diacetone alcohol) in dilute solutions. *J. Chem. Soc., Faraday Trans.* **1993,** *89* (3), 489-493.
5. Wülfert, F.; Kok, W. T.; Smilde, A. K., Influence of Temperature on Vibrational Spectra and Consequences for the Predictive Ability of Multivariate Models. *Analytical Chemistry* **1998,** *70* (9), 1761-1767.
6. Kalivas, J. H.; Palmer, J., Characterizing multivariate calibration tradeoffs (bias, variance, selectivity, and sensitivity) to select model tuning parameters. *Journal of Chemometrics* **2014,** *28* (5), 347-357.
7. Kalivas, J. H.; Héberger, K.; Andries, E., Sum of ranking differences (SRD) to ensemble multivariate calibration model merits for tuning parameter selection and comparing calibration methods. *Analytica Chimica Acta* **2015,** *869*, 21-33.
8. Tencate, A. J.; Kalivas, J. H.; White, A. J., Fusion strategies for selecting multiple tuning parameters for multivariate calibration and other penalty based processes: A model updating application for pharmaceutical analysis. *Analytica Chimica Acta* **2016,** *921*, 28-37.
9. Carlosena, A.; Andrade, J. M.; Kubista, M.; Prada, D., Procrustes Rotation as a Way To Compare Different Sampling Seasons in Soils. *Analytical Chemistry* **1995,** *67* (14), 2373-2378.
10. Ottaway, J.; Kalivas, J. H., Feasibility Study for Transforming Spectral and Instrumental Artifacts for Multivariate Calibration Maintenance. *Applied Spectroscopy* **2015,** *69* (3), 407-416.
11. Krzanowski, W. J., Between-Groups Comparison of Principal Components. *Journal of the American Statistical Association* **1979,** *74* (367), 703-707.
12. Anderson, C. E.; Nieves, R. G.; Kalivas, J. H., Orthogonality considerations for library searching Nth-order data. *Chemometrics and Intelligent Laboratory Systems* **1998,** *41* (1), 115-125.
13. Horn, R. A.; Johnson, C. R., Norms for Vectors and Matrices. In *Matrix Analysis*, Cambridge University Press: Cambridge, England, 1990.
14. Pedersen, D. K.; Martens, H.; Nielsen, J. P.; Engelsen, S. B., Near-infrared absorption and scattering separated by extended inverted signal correction (EISC): analysis of near-infrared transmittance spectra of single wheat seeds. *Applied spectroscopy* **2002,** *56* (9), 1206-1214.

15. Jouan-Rimbaud, D.; Massart, D. L.; Saby, C. A.; Puel, C., Characterisation of the representativity of selected sets of samples in multivariate calibration and pattern recognition. *Analytica Chimica Acta* **1997,** *350* (1–2), 149-161.
16. Jouan-Rimbaud, D.; Massart, D. L.; Saby, C. A.; Puel, C., Determination of the representativity between two multidimensional data sets by a comparison of their structure. *Chemometrics and Intelligent Laboratory Systems* **1998,** *40* (2), 129-144.
17. Barros, A. S.; Safar, M.; Devaux, M. F.; Robert, P.; Bertrand, D.; Rutledge, D. N., Relations between mid-nfrared spectra detected by analysis of variance of an intervariable data matrix. *Applied Spectroscopy* **1997,** *51* (9), 1384-1393.
18. Maalouly, J.; Eveleigh, L.; Rutledge, D. N.; Ducauze, C. J., Application of 2D correlation spectroscopy and outer product analysis to infrared spectra of sugar beets. *Vibrational Spectroscopy* **2004,** *36* (2), 279-285.
19. Ramsay, J. O.; Berge, J.; Styan, G. P. H., Matrix correlation. *Psychometrika* **1984,** *49* (3), 403-423.
20. Kailath, T., The Divergence and Bhattacharyya Distance Measures in Signal Selection. *IEEE Transactions on Communication Technology* **1967,** *15* (1), 52-60.
21. Willett, P., Combination of Similarity Rankings Using Data Fusion. *Journal of Chemical Information and Modeling* **2013,** *53* (1), 1-10.
22. Winning, H.; Larsen, F. H.; Bro, R.; Engelsen, S. B., Quantitative analysis of NMR spectra with chemometrics. *Journal of Magnetic Resonance* **2008,** *190* (1), 26-32.
23. Eigenvector Research Inc, http://www.eigenvector.com/data/Corn/. Wenatchee, Washington.
24. Berzaghi, P.; Shenk, J.; Westerhaus, M., LOCAL prediction with near infrared multi-product databases. *Journal of Near Infrared Spectroscopy* **2000,** *8* (1), 1-9.
25. Tikhonov, A. In *Solution of incorrectly formulated problems and the regularization method*, Soviet Math. Dokl., 1963; p 1035/1038.
26. Kalivas, J. H., Comprehensive Chemometrics. In *Calibration Methodologies*, Brown, S.; Tauler, R.; R, W., Eds. Elsevier: Oxford, 2009; Vol. 3, pp 1-32.

Chapter 3: Local Adaptive Fusion Regression (LAFR) Process

1. Introduction

Local modeling methods have been widely proposed and used for many different applications in industrial processes[1-9]. A few of these applications include the pharmaceutical industry for active ingredient estimation[9], the petrochemical industry as a real time quality control monitoring system[1], and the agricultural industry for assessing corn grain hardness[2]. The benefit that local modeling techniques provide is the ability to assess important process properties in real-time without requiring offline methodology to measure these same properties. The main objective for these local modeling processes is to select a number of samples from a global sample set that are similar to a target sample in order to build a predictive model for the target sample. Local modeling approaches described for each of these applications recommend multiple types of local modeling parameters necessary for selecting local calibration samples. Local modeling parameters that can vary greatly between applications include similarity merits used to select samples from the global sample set that are similar to the target sample and methods for selecting the number of global samples to include in the local calibration sets. The differences in the local modeling parameters proposed are typically process and dataset dependent making it difficult to apply the same parameters across different types of applications.

1.1. *Similarity Measures*

There are many combinations of similarity measures proposed for various local modeling methods. Reported spectral similarity measures include Euclidean distance[7, 10-11], Mahalanobis distance[12-13], combinations of distance and angle measurements[14], and merits such Q residuals and Hotelling's T^2 statistic[5]. There are also studies that propose a

variety of similarity measures based on reference value data, primarily prediction results[6, 13, 15-19]. These types of similarity merits require a preliminary prediction model to be formed, either by the global dataset or a subset of the global samples. For any biased multivariate regression prediction models used in these methods, a tuning parameter must also be selected.

Variations in the similarity measure(s) can affect which samples are selected for the local model and ultimately the performance of the model. In one study, Mahalanobis distances were compared to Euclidean distances as similarity measurements[20]. Of the four datasets that were assessed, the difference in the prediction error results from PLS models, between the local samples selected by Mahalanobis distance and Euclidean distance, changed depending on the dataset and the number of LV's selected. Another study, using a multi-layer approach to local modeling, assessed nine different similarity measures including Euclidean distances of the predictions of local calibration models formed from each previous layer (a subset of the global samples) and the Euclidean distances of the scores of layer of samples calculated from PLS[17]. This study concluded that the best similarity measures for sample selection were dependent on the type of spectral pre-treatments used and the dataset itself. In another study, Mahalanobis distances calculated in three different ways were compared as similarity measures[13]. The Mahalanobis distances calculations included a typical spectral based distance, a modified Mahalanobis distance of principal component analysis (PCA) scores that were weighted for each principal component (PC) depending on the global sample predictions at that PC, and a modified Mahalanobis distance combining the spectral distance comparisons with distances of the global and target predictions based on a global calibration model. For

these three similarity measurement scenarios, the third method typically selected local calibration samples resulting in the lowest prediction errors for the target samples. However, using this method there is still a reliance on the selection of principal components for each Mahalanobis distance measurement and the predictive ability of the global calibration model for predicting the target samples initially. Another complicating factor for the third Mahalanobis distance merit described is the selection of a method for combining the spectral Mahalanobis distances with the prediction based Mahalanobis distances.

Multiple local modeling methods propose combinations of two similarity measures[5, 7, 13-14, 18]. When more than one similarity measure is used, a method for combining these measures into a single similarity measure or similarity index is necessary. Equation 3.1 shows a generic form for calculating a similarity index (SI_i) by combining two similarity merits together (SI_1 and SI_2) .

$$SI_i = \gamma SI_1 + (1-\gamma)SI_2 \tag{3.1}$$

In this equation, γ is referred to as a trade-off parameter and is set between 0 and 1 in order to give equal or varying weights to each of the individual similarity measures. This trade-off parameter has been shown to impact the prediction ability of the final local model. In a process using local modeling to track catalyst deactivation and recovery, two similarity merits, Hotelling's T^2 and Q statistic, are represented as SI_1 and SI_2 respectively in equation 3.1[5]. It was determined that for the process conditions of the catalyst system if $\gamma = 0.01$, as opposed to $\gamma = 0$, the correlation coefficients between measured and estimated results increased 15%, and the model target sample prediction errors decreased 10%. However, for these same experimental conditions if $\gamma > 0.01$ the

model performance tended to decrease. This indicates that the combination of both measurements helped the local model but only if the weight of the Hotelling T^2 merit was much smaller than the Q residual merit.

In another study, where a spectral similarity merit, Euclidean distance, was combined with chemical similarity information, based on a proposed adaptive algorithm for obtaining target prediction values, multiple trade-off parameters were assessed ranging from 0 to 1 in 0.1 increments[21]. It was determined that a trade-off of 0 or 1 resulted in the highest prediction errors while trade-offs ranging between 0.1 and 0.9 resulted in lower prediction errors.

These two studies indicate local model improvement with the incorporation of more than one similarity merit; however, the selection of a trade-off parameter is an important factor in optimizing the final local model when only two similarity measures are represented. Overall, in terms of similarity merits, there is no consensus of which similarity measures are superior to others for local modeling as many different similarity merits and combinations of similarity merits are proposed in the literature discussed. One important conclusion from the various similarity measures proposed is that the incorporation of chemical information generally resulted in improvements in local model performance. This conclusion supports that fact that localization should be carried out with respect to both spectral matching and chemical matching[22]. However, tuning parameter selection is often necessary for incorporation of prediction based chemical information.

1.2. *Selection of Number of Samples*

Along with selecting similarity measures for local calibration models, the number of samples must also be selected. There are multiple proposed methods for this selection process; however, the selection methods are mainly determined by trial and error and are dependent on the local modeling results for the specific process or dataset used in the study[5-7, 14, 23]. This trial and error method is not ideal as it requires development time and is not adaptable to new process conditions requiring similar local modeling methods.

The simplest method for selecting the number of samples is to set an integer based on known information about the dataset. In one study, where the objective was to monitor a catalytic process over time, the number of local calibration samples was set to window sizes of 10 and 20 days with a change of one day for each local calibration set[5]. A local calibration set was selected based on similarity of a target sample to the data in one of the local sets of 10 or 20 day data ranges. The problem with this type of method for setting the number local calibration samples is that, depending on day to day changes observed in the measurements, 10 or 20 days might capture too much or too little variability for accurately predicting a new sample.

The number of samples selected can also be based on thresholds of the similarity measures. In a study focused on developing an adaptive local modeling algorithm for near infrared (NIR) data, many different local modeling parameters were proposed, including one to assist in selecting the number of local calibration samples[7]. The sample selection similarity merit used was spectral Euclidean distances. Local samples were selected if the Euclidean distance between the target and the global sample measured was less than 0.9. This 0.9 threshold was set based on trial and error methodology in order to

achieve a minimum of 50 samples in each local calibration set. Again, this method was reliant on the global dataset properties and based on one minimum sample number requirement.

There are recent studies that have proposed adaptive local modeling techniques for updating the number of samples selected for each local model built[6, 14, 23]. For these adaptive techniques the global samples are still ranked based on a specific similarity measurement, and the number of samples included in the local model is based on building multiple local calibration models. In two of these studies, the optimal number of samples was selected by minimizing the prediction errors for the local calibration samples through cross-validation techniques[14, 22]. One drawback to this method is that low prediction errors for the calibration samples does not guarantee a low prediction error for the target sample. Another drawback is that tuning parameters still need to be selected for each model formed in order to compare the calibration prediction errors.

The identified challenges in the local modeling processes discussed above are that there is no consensus on which similarity measures should be used, for similarity measures based on predictions, a tuning parameter must be selected, and methods for selecting the appropriate number of samples for the local calibration set are dataset dependent or dependent on calibration sample prediction errors. The algorithm proposed in this work, termed local adaptive fusion regression (LAFR), provides methodologies to alleviate some of these challenges. Chapter 2 described one step of the LAFR process for selecting a matrix matched local calibration set from manually generated calibration sets. In this process multiple similarity comparison merits, including both spectral and prediction based, multiple tuning parameters, a new technique proposed called cross

modeling, and data fusion methods are used. This step in the LAFR algorithm for comparing multiple local calibration sets allows for a wide variety of calibration sets with different sample sizes and reference value ranges to be formed and compared in order to address the challenges in the local modeling methods listed above.

2. LAFR Algorithm

The overall objective of the LAFR algorithm is to select a local calibration set of samples that have a reduced analyte chemical range and are matrix matched to the individual target samples. The key steps to achieving this goal include the formation of many local calibration sets with limited chemical ranges using an iterative multi-parameter process. These local calibration sets are then compared to each other using a number of matrix matching assessment merits (described in Chapter 2). Figure 3.1 contains the algorithm process in its most general format. Specific steps from the algorithm are described in detail in the following sections but are briefly described here.

In step 1, the global samples are sorted based on the reference values of interest. Step 2 calculates the spectral similarity between the target spectrum and each global spectrum. In this process 26 similarity merits are used. These merits and the process of sample selection are described in section 2.1. A set of calibration parameters are then selected in step 3. There are nine calibration parameters that are considered and are further described in section 2.2. In step 4, the library size is determined by the parameters set in step 3. If the library size is set to “Global” then the process goes directly to step 6. If the library size is not set to “Global” then the 26 spectral similarity merits are used to determine which global spectra are most similar to the target spectrum (step 5) described in section 2.1. In step 6, outliers are iteratively removed from the library space and the

target sample is tested as an outlier to the outlier free library space. This outlier check process is further discussed in section 2.3. If the target sample is deemed an outlier to the library space (step 7) this calibration parameter option is discarded, and the process moves on to the next calibration parameter option back to step 3. If the target sample is not an outlier to the library space then it is used to form "y-windows" of local calibration sets based on the calibration parameter options for determining "y-window" ranges (step 8). This process is further described in section 2.4. The outlier check process is also performed on each of the local calibration sets formed in step 9. Step 10 calculates the calibration set comparison merits, described in Chapter 2, for each of the local calibration sets. Using fusion rules, a single local calibration set is selected and stored (step 11). If there are more calibration set parameter options then the process repeats from step 3. After all of the calibration parameter set options have been assessed, the local calibration sets selected for each parameter option are compared to one another using the same calibration set comparison merits from Chapter 2 (step 14). Note that the global calibration set is added in as one of the final calibration sets to be compared. The fusion rules, again, are used to select a final calibration set (step 15) and a prediction model is formed to predict the target sample.

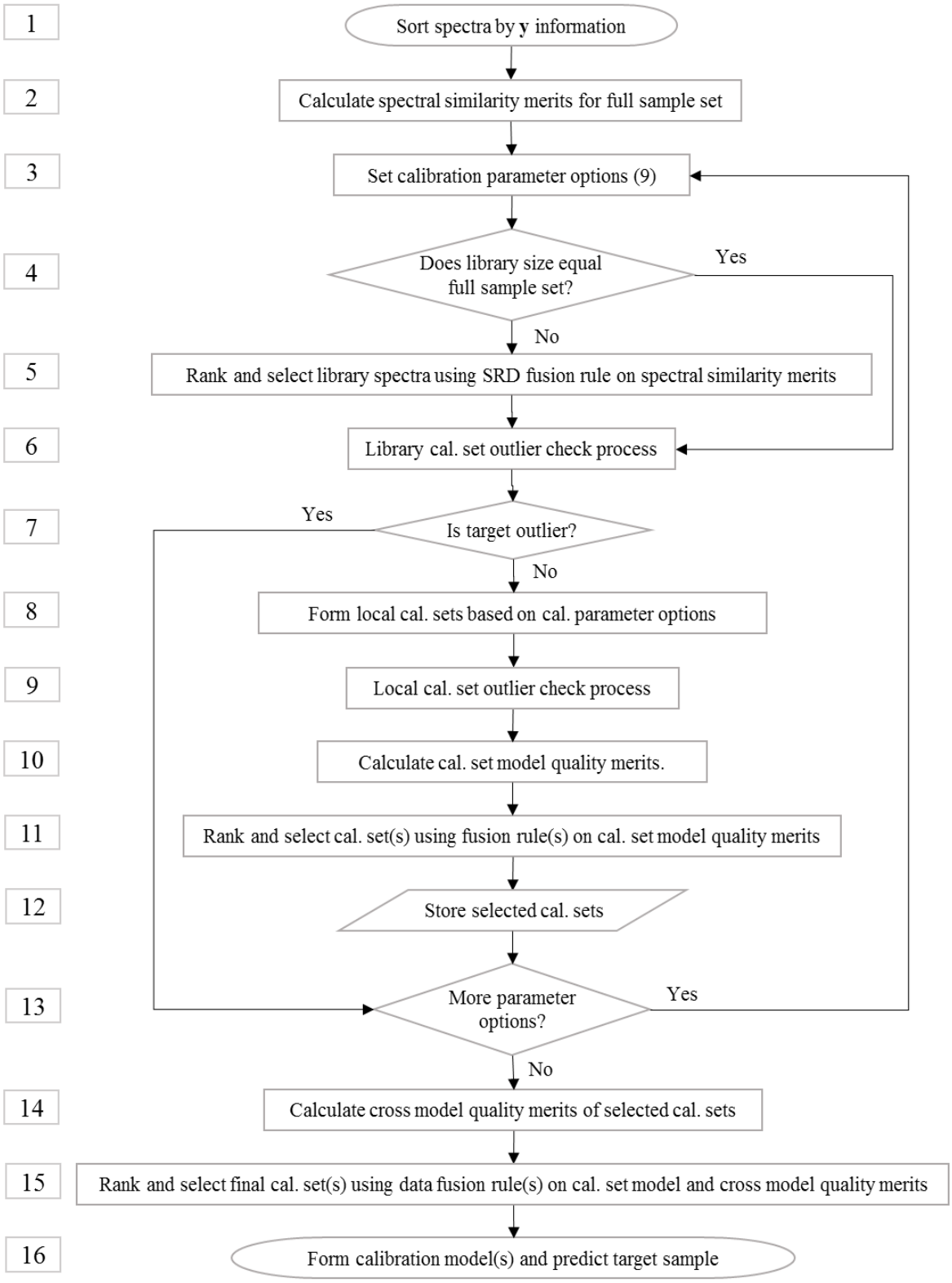


Figure 3.1. Flowchart for local adaptive fusion regression algorithm.

2.1. *Determining Spectrally Similar Library Spaces*

This process corresponds to steps 2-5 in the LAFR algorithm (Fig. 3.1). Many of the local modeling processes describe how one or two spectral similarity merits are used to select the samples for a local model. In LAFR, 26 spectral similarity merits are used to select spectrally similar samples that will then be separated into a collection of local calibrations sets for further evaluation. The spectral similarity merits are listed in Tables 3.1 and 3.2. The calculations for most of these merits are described in detail in Chapter 2. For these merits, the sample vector to calibration vector and sample domain to calibration domain merit calculation methods are used. For the sample vector to calibration vector merits, $\boldsymbol{x}_1$can either be equal to the target spectrum ($\boldsymbol{x}_t$) or the global spectrum ($\boldsymbol{x}_l$) and the same is true for $\boldsymbol{x}_2$. For the sample domain to calibration domain comparison merits, as described in Chapter 2, outer products are calculated for $\boldsymbol{x}_1$ and $\boldsymbol{x}_t$ vectors. The specific outer products are listed in Table 3.2.

One additional merit, Bartlett's statistic, is not used in the calibration set comparison methods and is presented in this section. Bartlett's statistic is calculated as a sample domain to calibration domain merit. Bartlett's statistic is a method to compare the variance-covariance matrices of two datasets to determine similarities in both magnitude and direction[24-25]. This merit ranges from 0 to 1, where 1 is the most similar and 0 is the least similar. For this work, Bartlett's statistic merit is manipulated to assess similarity between two vectors (Eq. 3.2).

$$1 - C_{bart} = 1 - \exp(\frac{-c_1}{m_1+m_2}) \tag{3.2}$$

For this calculation, m_1 and m_2 are the number of samples for each samples sets being compared, in this case both m_1 and m_2 are equal to one. The variable c_1 is calculated as

$$c_1 = c_2[(m_1)\ln(|\sigma_1^{-1}\sigma_s|) + (m_2)\ln(|\sigma_2^{-1}\sigma_s|)] \quad (3.3)$$

Where σ_1^{-1} is the inverse of the first PC eigenvalue from the pseudoinverse $\boldsymbol{X}_1^{+}$, and σ_2^{-1} is the inverse of the first PC eigenvalue from the pseudoinverse $\boldsymbol{X}_2^{+}$. For this equation, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are the outer products of $\boldsymbol{x}_1$ (Eq. 2.43) and $\boldsymbol{x}_2$ (Eq. 2.44)**.** In this equation, $|\sigma_1^{-1}\sigma_s|$and $|\sigma_2^{-1}\sigma_s|$ represent the determinants of the products. Here, σ_s is the first eigenvalue calculated from the SVD of equation 2.41 calculating $\boldsymbol{S}$.

$$\boldsymbol{S} = \frac{\boldsymbol{X}_1\boldsymbol{X}_2}{2}$$

Also in equation 3.3, c_2 is calculated as

$$c_2 = \left(\frac{2n^2 + 3n - 1}{6(n+1)}\left[\frac{-1}{m_1 + m_2}\right]\right)$$

Similarly to $\cos\theta$, the final merit (2.43) is subtracted from one so that lower values could be associated with higher degrees of similarity.

These 26 merits are used to calculate spectral similarity rankings of the global samples using the fusion rule sum of ranking differences (SRD)[26-29]. The SRD process is described in Chapter 1 (Fig. 1.3). For the selection of similar spectra for this work, the SRD ‘target’ vector is set to maximum. This indicates that samples with lower ranks are more dissimilar to the target sample.

Table 3.1. Sample vector to calibration vector merits for selection of spectrally similar library samples. (Notations indicated in footnotes).

Category	Merit	Input Assignments	Equation
Spectral	$1 - cos\,\theta$	$\boldsymbol{x}_1 = \boldsymbol{x}_l\ ;\ \boldsymbol{x}_2 = \boldsymbol{x}_t$	2.21
Spectral	$1 - cos^2\,\theta$	$\boldsymbol{x}_1 = \boldsymbol{x}_l\ ;\ \boldsymbol{x}_2 = \boldsymbol{x}_t$	2.22
Spectral	Euc	$\boldsymbol{x}_1 = \boldsymbol{x}_l\ ;\ \mathbf{x}_2 = \boldsymbol{x}_t$	2.23
Spectral	Det	$\boldsymbol{x}_1 = \boldsymbol{x}_l\ ;\ \boldsymbol{x}_2 = \boldsymbol{x}_t$	2.24
Spectral	F	$\boldsymbol{x}_1^T = \boldsymbol{x}_t^T\ ;\ \boldsymbol{x}_2^T = \boldsymbol{x}_l^T$	2.27
Spectral	ρ	$\boldsymbol{x}_1 = \boldsymbol{x}_l\ ;\ \boldsymbol{x}_2 = \boldsymbol{x}_t$	2.31
Spectral	ρ	$\boldsymbol{x}_1 = \boldsymbol{x}_t\ ;\ \boldsymbol{x}_2 = \boldsymbol{x}_l$	2.31
Spectral	$EISC\ Xb_d$	$\boldsymbol{x}_1 = \boldsymbol{x}_l\ ;\ \boldsymbol{x}_2 = \boldsymbol{x}_t$	2.37
Spectral	$EISC\ Xb_d$	$\boldsymbol{x}_1 = \boldsymbol{x}_t\ ;\ \boldsymbol{x}_2 = \boldsymbol{x}_l$	2.37
Spectral	$EISC\ b_d$	$\boldsymbol{x}_1 = \boldsymbol{x}_l\ ;\ \boldsymbol{x}_2 = \boldsymbol{x}_t$	2.36
Spectral	$EISC\ b_d$	$\boldsymbol{x}_1 = \boldsymbol{x}_t\ ;\ \boldsymbol{x}_2 = \boldsymbol{x}_l$	2.36
Spectral	$EISC\ b$	$\boldsymbol{x}_1 = \boldsymbol{x}_l\ ;\ \boldsymbol{x}_2 = \boldsymbol{x}_t$	2.33
Spectral	$EISC\ b$	$\boldsymbol{x}_1 = \boldsymbol{x}_t\ ;\ \boldsymbol{x}_2 = \boldsymbol{x}_l$	2.33
Spectral	MD_p^v	$\boldsymbol{x}_1 = \boldsymbol{x}_l\ ;\ \boldsymbol{x}_2 = \boldsymbol{x}_t$	2.40
Spectral	MD^v	$\boldsymbol{x}_1 = \boldsymbol{x}_l\ ;\ \boldsymbol{x}_2 = \boldsymbol{x}_t$	2.38
Spectral	MD^v	$\boldsymbol{x}_1 = \boldsymbol{x}_t\ ;\ \boldsymbol{x}_2 = \boldsymbol{x}_l$	2.38

$\boldsymbol{x}_l$: global sample

$\boldsymbol{x}_t$: target sample

Table 3.2. Sample domain to calibration domain merits for selection of spectrally similar library samples. (Notations indicated in footnotes).

Category	Merit	Input Assignments	Equation
Spectral	Euc	$\boldsymbol{X}_1 = \boldsymbol{x}_l\boldsymbol{x}_l^T$; $\boldsymbol{X}_2 = \boldsymbol{x}_t\boldsymbol{x}_t^T$	2.47
Spectral	Euc	$\boldsymbol{X}_1 = \boldsymbol{x}_t\boldsymbol{x}_l^T$; $\boldsymbol{X}_2 = \boldsymbol{x}_t\boldsymbol{x}_t^T$	2.47
Spectral	Det	$\boldsymbol{X}_1 = \boldsymbol{x}_l\boldsymbol{x}_l^T$; $\boldsymbol{X}_2 = \boldsymbol{x}_t\boldsymbol{x}_t^T$	2.24
Spectral	Det	$\boldsymbol{X}_1 = \boldsymbol{x}_t\boldsymbol{x}_l^T$; $\boldsymbol{X}_2 = \boldsymbol{x}_t\boldsymbol{x}_t^T$	2.24
Spectral	F	$\boldsymbol{X}_1 = \boldsymbol{x}_l\boldsymbol{x}_l^T$; $\boldsymbol{X}_2 = \boldsymbol{x}_t\boldsymbol{x}_t^T$	2.51
Spectral	F	$\boldsymbol{X}_1 = \boldsymbol{x}_t\boldsymbol{x}_t^T$; $\boldsymbol{X}_2 = \boldsymbol{x}_l\boldsymbol{x}_l^T$	2.51
Spectral	ρ	$\boldsymbol{X}_1 = \boldsymbol{x}_t\boldsymbol{x}_t^T$; $\boldsymbol{X}_2 = \boldsymbol{x}_l\boldsymbol{x}_l^T$	2.56
Spectral	ρ	$\boldsymbol{X}_1 = \boldsymbol{x}_l\boldsymbol{x}_l^T$; $\boldsymbol{X}_2 = \boldsymbol{x}_t\boldsymbol{x}_t^T$	2.56
Spectral	H	$\boldsymbol{X}_1 = \boldsymbol{x}_t\boldsymbol{x}_t^T$; $\boldsymbol{X}_2 = \boldsymbol{x}_l\boldsymbol{x}_l^T$	2.57
Spectral	$1 - C_{bart}$	$\boldsymbol{X}_1 = \boldsymbol{x}_l\boldsymbol{x}_l^T$; $\boldsymbol{X}_2 = \boldsymbol{x}_t\boldsymbol{x}_t^T$	3.2

$\boldsymbol{x}_l$: global sample

$\boldsymbol{x}_t$: target sample

One useful feature of SRD is a ranking validation method used to assess the probability that the calculated SRD ranks are statistically different from randomly assigned rankings[27]. This process is referred to as comparison of ranks by random number (CRRN). In CRRN a normalized probability distribution is generated from 100,000 iterations of SRD ranks calculated from randomly generated numbers. Using this normalized random rank distribution feature, a threshold can be set to exclude samples that have rankings one, two, or *n* standard deviations to the left of the mean of the random rank distribution (Fig. 3.2). These rankings would indicate that these samples are more dissimilar to the target sample for some level of certainty.

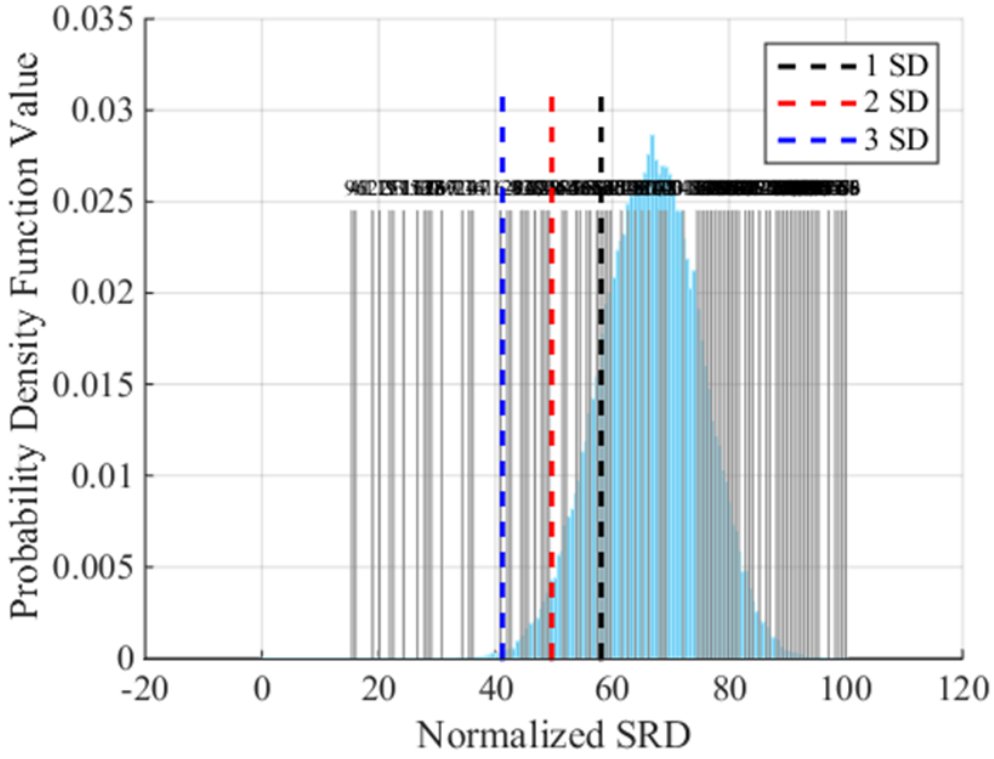


Figure 3.2. Sum of ranking differences probability density function values for comparison of ranks by random numbers (CRRN) process versus normalized SRD rankings. Standard deviations (SD) 1-3 are listed. The normalized ranks for a set of samples are plotted (solid black lines). The CRRN probability distribution is shown in light blue.

After a number of samples have been excluded from the library space initially based on *n* standard deviations from the CRRN distribution, the remaining samples, referred to as '1st SRD', can be ranked again through the SRD process using the same 26 spectral similarity measurements. The sample rankings will change as the number of samples and relationships are altered. Again, the samples that are to the left of the specified *n* standard deviations from the mean of the CRRN distribution are excluded to

form a second library space, referred to as '2nd SRD'. This library space selection process repeats until no samples are to the left of the specified *n* standard deviations. All of the library spaces formed, including the global space prior to SRD selection, are used in the LAFR process. In this work a standard deviation of 3 is used; however, setting this standard deviation is considered one of the adjustable parameter options discussed in the following section.

2.2. *Local Calibration Set Formation Parameters*

The local calibration set parameters correspond to step 3 in the LAFR algorithm. All of the parameters considered in this process are listed in Table 3.3 along with a short description and examples. Some of the examples listed for each parameter are not used for the dataset presented in this work. These examples are shown to indicate the flexibility and possibilities for this process. For example, in parameter 4 for setting the range of the reference value data for each local calibration set, one of the examples listed is based on using the known analytical error of the primary analysis method as a guideline to setting how large or small the reference value ranges should be. As the analytical error is not known for the datasets used in this work this method is not employed. The ability to vary these parameter options is a key factor in differentiating LAFR from other local modeling methods.

Table 3.3. Nine adjustable parameters used for LAFR process.

Parameter ID	Parameter	Description	Examples
1	Spectral similarity merits	Combination of vector shape and magnitude comparisons	-angle between vectors -Euclidean distance
2	Library size	Number of samples in the library space for forming local cal. sets	- global - sample selected by SRD iterations (n* SD)
3	Local cal. set min size	Minimum number of samples needed for each local cal. set formed	- no size requirement -integers -percentage of samples in library
4	Local cal. set y range max	Largest acceptable range for y data in each local cal. set formed	- no range requirement - percentage of y library range -based on error of primary analytical method
5	Local cal. set y window overlap	Amount of overlap between y data for each local cal. set window	-percentage of previous calibration set -number of samples
6	Outlier merits	Outlier determination merits used for library space and the local cal. sets	-Studentized residuals -Mahalanobis distance
7	Local cal. set comparison merits	Merits used to compare local cal. sets	-spectral -prediction -orthogonal projections to regression vector
8	Tuning parameter selection	Process for selecting tuning parameters for PLS or SVD based merits and setting number of tuning paramameters	-specified number of LV -PC's accounting for percentage of variability
9	Fusion rules	Fusion rules used for data fusion based selections	-SRD -SUM rule -combination of fusion rules

*n is the specified standard deviation left of the SRD random ranking probability distribution used to determine which samples should be included or excluded

Adjusting each of the nine LAFR process parameter options can result in an extremely large number of parameter set combinations. For instance, if each of the nine parameter options had three levels, such as 10, 15, and 20 samples for parameter 3, then there would be approximately 20,000 unique parameter option set combinations created increasing the computational time. For this work the parameters that are adjusted were limited to parameters 3 and 4, the local calibration set minimum size and the local calibration set y range. The specific parameter used for the dataset presented are described in section 3.

2.3. *Outlier Determination*

The outlier check process is a two-part process. The first part of the outlier check process is to remove all samples from the library spaces or local calibration set, both are referred to in this section as the calibration space, that are identified as outliers. Like many of the steps for this algorithm, the outlier check process employs the use of multiple merits and data fusion methods to avoid selecting a single outlier determination merit.

For removing outliers from the calibration space, each sample in the calibration set is removed one at a time and the merits in Tables 3.4-3.7 are calculated. For the prediction merits (Table 3.4) a PLS algorithm is used to form a model to calculate the predictions for each calibration sample ($\hat{y}_o$) across a specified number of tuning parameters (LV's). The number of LV's used is set to five (LAFR parameter 8) and the process for determining which five LV's to use is described in Chapter 2 section 2.2. All of the calculations for these merits are described in Chapter 2 sections 2 and 3. All three categories of spectral merits calculations are represented in the outlier determination

merits: sample vector to calibration vector, sample domain to calibration domain, and sample vector to calibration domain. Only the Spectral based approach is used for these merits (described in Chapter 2). For the sample vector to calibration vector merits (Table 3.5), $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ can both represent either the sample removed from the calibration space ($\boldsymbol{x}_o$) or the average spectrum of the calibration space with the sample removed ($\overline{\boldsymbol{x}}_r$). For the sample domain to calibration domain merits (Table 3.6), $\boldsymbol{X}_1$ is always equal to the outer product of the sample removed $\boldsymbol{x}_o$ (or the target sample ($\boldsymbol{x}_t$) for the 2nd part of the outlier process), and $\boldsymbol{X}_2$ is the outer product of the averaged spectra of the remaining calibration space ($\overline{\boldsymbol{x}}_r$). In the sample vector to calibration domain merits (Table 3.7), $\boldsymbol{X}_r$ are the samples remaining in the calibration set. These sample vector to calibration domain merits use a set of principal components (PC's). The PC's used are determined by the number necessary to account for up to 99% of the cumulative variability for each individual calibration space

There are two additional prediction merits (Y) used for the outlier check process not described in Chapter 2; the Studentized residual and the matrix match ratio merit. These two merits are described below.

2.3.1 Studentized Residual

A Studentized residual (Eq. 3.4) is the residual measurement of a prediction normalized by its estimated standard deviation and is often used for outlier detection in linear regression[30].

$$t_2 = \left|\frac{y_2 - \hat{y}_2}{\sigma\sqrt{1-h}}\right| \tag{3.4}$$

As explained for outlier detection, each sample in a library space or local calibration set is removed one at a time for outlier determination. In this equation, y_2 is the reference

value of the sample removed, $\hat{y}_2$ is the predicted reference value using a model formed by the remaining samples, and σ (Eq. 3.5) and h (Eq. 3.6) are calculated below.

$$\sigma = \sqrt{\frac{\sum_{s=1}^{m-1}(y_s - \hat{y}_s)^2}{m-2}} \tag{3.5}$$

$$h = \frac{1}{m} + \boldsymbol{x}_2^T(\boldsymbol{X}_r^T\boldsymbol{X}_r)^+\boldsymbol{x}_2 \tag{3.6}$$

In equation 3.5, σ represents the standardized error of the samples used to build the model, where m is the total number of samples prior to removing one sample, and y_s and $\hat{y}_s$ represent the individual reference values and predicted reference values for each of the samples remaining in the library space of calibration set. In equation 3.6, the pseudoinverse of $\boldsymbol{X}_r^T\boldsymbol{X}_r$ **(**$\boldsymbol{X}_r^T\boldsymbol{X}_r^+$**)** for the calculation of h, often referred to as leverage, is calculated from a PLS algorithm, where $\boldsymbol{X}_{r_{((m-1)\,x\,n)}}$ are the spectra of the samples remaining in the calibration space, and $\boldsymbol{x}_{2_{(n\,x\,1)}}$ is the spectrum of the sample removed.

2.3.2 Matrix Match Ratio

The matrix match ratio (MMR) is based on equations 2.9 and 2.10 for the matrix matching assessment measurement when $|\hat{y}_j - y| = 0$ (Eq. 3.7).

$$MMR = |\frac{y_2}{\hat{y}_2} - \bar{R}| \tag{3.7}$$

The calculation for $\frac{y_2}{\hat{y}_2}$ is based on equation 2.9. When $|\hat{y}_{j,2} - y_2| = 0$ then α_j can be calculated by $\frac{y_2}{\hat{y}_2}$. Here, $\hat{y}_{j,2}$ represents the scaled predication error of the sample removed from the calibration space. In this equation, $\bar{R}$ is calculated as

$$\bar{R} = \frac{\sum_{i=1}^{m}\left(\frac{y_i}{\hat{y}_i}\right)}{m}$$

The variable $\bar{R}$ is the mean value of all α_j's for each sample in the calibration space when $|\hat{y}_{j,i} - y_i| = 0$. In this calculation, y_i are the true reference value for m samples in the

calibration space, $\hat{y}_i$ are the predicted reference values, and $\hat{y}_{j,i}$ are the scaled predicted reference values. Both the Studentized residuals and matrix match ratio were added to Table 3.4 along with the prediction error (e_{22}) described in Chapter 2.

Table 3.4. Prediction merits for outlier determination. (Notations indicated in footnotes)

Category	Merit	Input Assignments	Equation
Y	e_{22}	$y_2 = y_o; \hat{y}_2 = \hat{y}_o$	2.11
Y	t_2	$y_2 = y_o; \hat{y}_2 = \hat{y}_o;$ $\boldsymbol{x}_2 = \boldsymbol{x}_o$	3.4
Y	MMR	$y_2 = y_o; \hat{y}_2 = \hat{y}_o$	3.7

y_o **:** calibration sample reference value removed from calibration space

$\hat{y}_o$ **:** predicted calibration sample reference value removed from calibration space

$\boldsymbol{x}_o$: sample removed from calibration space

Table 3.5. Sample vector to calibration vector merits for outlier determination. (Notations indicated in footnotes)

Category	Merit	Input Assignments	Equation
Spectral	$1 - cos^2\,\theta$	$\boldsymbol{x}_1 = \boldsymbol{x}_{o/t}\,; \boldsymbol{x}_2 = \overline{\boldsymbol{x}}_r$	2.22
Spectral	Euc	$\boldsymbol{x}_1 = \boldsymbol{x}_{o/t}\,; \boldsymbol{x}_2 = \overline{\boldsymbol{x}}_r$	2.23
Spectral	Det	$\boldsymbol{x}_1 = \boldsymbol{x}_{o/t}\,; \boldsymbol{x}_2 = \overline{\boldsymbol{x}}_r$	2.24
Spectral	$EISC\ Xb_d$	$\boldsymbol{x}_1 = \overline{\boldsymbol{x}}_r\,; \boldsymbol{x}_2 = \boldsymbol{x}_{o/t}$	2.37
Spectral	$EISC\ Xb_d$	$\boldsymbol{x_1} = \boldsymbol{x}_{o/t}\,; \boldsymbol{x}_2 = \overline{\boldsymbol{x}}_r$	2.37
Spectral	$EISC\ b_d$	$\boldsymbol{x}_1 = \overline{\boldsymbol{x}}_r\,; \boldsymbol{x}_2 = \boldsymbol{x}_{o/t}$	2.36
Spectral	$EISC\ b_d$	$\boldsymbol{x}_1 = \boldsymbol{x}_{o/t}\,; \boldsymbol{x}_2 = \overline{\boldsymbol{x}}_r$	2.36

$\overline{\boldsymbol{x}}_r$: mean spectrum for the remaining calibration space spectra

$\boldsymbol{x}_{o/t}$: sample removed from calibration space ($\boldsymbol{x}_o$) or target sample ($\boldsymbol{x}_t$)

Table 3.6. Sample domain to calibration domain merits for outlier determination.

(Notations indicated in footnotes)

Category	Merit	Input Assignments	Equation
Spectral	F	$\boldsymbol{X}_1 = \boldsymbol{x}_{o/t}\,\boldsymbol{x}_{o/t}^T$; $\boldsymbol{X}_2 = \overline{\boldsymbol{x}}_r\overline{\boldsymbol{x}}_r^T$	2.51
Spectral	ρ	$\boldsymbol{X}_1 = \boldsymbol{x}_{o/t}\,\boldsymbol{x}_{o/t}^T$; $\boldsymbol{X}_2 = \overline{\boldsymbol{x}}_r\overline{\boldsymbol{x}}_r^T$	2.56
Spectral	H	$\boldsymbol{X}_1 = \boldsymbol{x}_{o/t}\,\boldsymbol{x}_{o/t}^T$; $\boldsymbol{X}_2 = \overline{\boldsymbol{x}}_r\overline{\boldsymbol{x}}_r^T$	2.57

$\overline{\boldsymbol{x}}_r$: mean spectrum for the remaining calibration space spectra

$\boldsymbol{x}_{o/t}$: sample removed from calibration space ($\boldsymbol{x}_o$) or target sample ($\boldsymbol{x}_t$)

Table 3.7. Sample vector to calibration domain merits for outlier determination.

(Notations indicated in footnotes)

Category	Merit	Input Assignments	Equation
Spectral	MD	$\boldsymbol{x}_1 = \boldsymbol{x}_{o/t}$	2.58
Spectral	Q	$\boldsymbol{x}_1 = \boldsymbol{x}_{o/t}$	2.63
Spectral	$sin\,\theta$	$\boldsymbol{x}_1 = \boldsymbol{x}_{o/t}$	2.66
Spectral	$1 - r_1$	$\boldsymbol{x}_1 = \boldsymbol{x}_{o/t}$	2.61
Spectral	Div	$\boldsymbol{x}_1 = \boldsymbol{x}_{o/t}$	2.62

$\boldsymbol{x}_{o/t}$: sample removed from calibration space ($\boldsymbol{x}_o$) or target sample ($\boldsymbol{x}_t$)

The merits assessing each sample within the library space or local calibration set are then used with the SRD data fusion method and a target set to “maximum”. From the SRD CRRN distribution, it can be determined that all samples left of the random ranking probability distribution at a specified standard deviation (SD) are dissimilar to the rest of the samples in the library space or local calibration set (Fig. 3.2). For this work, all samples left of 3 standard deviations are outliers and are removed. This outlier check

process is repeated until no samples have SRD normalized ranks left of 3 standard deviations from the mean of the CRRN distribution.

The second part of the outlier check process is to determine if the target sample is an outlier to the calibration space (the outlier-free library space or outlier-free local calibration set). The merits from Tables 3.5 to 3.7 are calculated for the target sample ($\boldsymbol{x}_t$) and the calibration space ($\boldsymbol{X}_r$). The prediction merits in Table 3.4 are not used for this part of the outlier check process as a reference value for the target sample would be required. The merits calculated from the target sample and the calibration space along with the merits calculated for each calibration sample using the leave-one-out methodology, previous described in the first part of the outlier check process, are combined together. The SRD fusion data method using 3 SD for CRRN distribution is used to determine if the target sample is an outlier to the calibration space. If the target sample is left of 3 SD from the mean of the CRRN distribution then it is considered an outlier.

2.4. *Formation of Local Calibration Sets*

Forming local calibration sets is step 8 in the LAFR algorithm. For the automated process of forming calibration sets the parameter set options, specifically parameters 3 and 4, must be simultaneously met. Currently, the following steps for formation of the local calibration sets are based on the assumption of a single Gaussian type distribution, with or without skewed sides, for the reference value data of each of the library spaces. If the global calibration set reference values, observed prior to the LAFR process, result in a combination of multi-peaked Gaussian distributions, then these distributions can be run as separate library spaces through the process for forming local calibration sets. This

multi-peak trend was not observed for any of the reference values in the dataset assessed in this work.

The formation of the local calibration sets for a specified library space is based on local calibration set parameters 3-5; local calibration set min size, local calibration set y range max, and local calibration set y window overlap. Figure 3.3 shows a flowchart of the automated process for forming multiple local calibration sets that meet the criteria established by the parameter options.

An initial reference value range is established using the Freedman-Diaconis (FD) rule (Eq. 3.8) for selecting range size for the library space reference value data.

$$y_{range} = 2\ (\frac{IQR}{\frac{1}{m^3}}) \tag{3.8}$$

In this equation, IQR is the interquartile range of the data, and m is the number of samples in the library space. The FD rule is typically less sensitive to data with outliers or "heavy" tails[31]. As the distributions are not individually viewed during this algorithm, the objective was to establish an initial range size that was not too broad as to encompass the possible outliers. The outliers are removed after the formation of local calibration sets.

If parameter 4 (local calibration set y range max) is greater than the FD determined range, then a set of local calibration sets are formed using the FD range and the specified "y-window" overlap (parameter 5). If parameter 4 is less than the FD range, then the local y range is set equal to parameter 4 and the local calibration sets are formed using parameter 5. The next step determines if any of the local calibration sets formed have less than the minimum required number of samples (parameter 3). If any of the local sets have less than the required number of samples, and the local y range is less than the local calibration set y range max (parameter 4), then the local y range is increased by

1.0% and the local sets are reformed. Otherwise, if any of the local sets have less than the required number of samples, and the local y range is greater than or equal to parameter 4, then samples are removed from the distribution.

For the removal of samples, the local sets are split into two halves. For example, if there were 30 sets formed then the first 15 would be the first half and 15 in the second half. If there is an uneven number of local sets then the first half has one more set than the second half. If there are local sets in the first half that have less than the required number of samples then the first sample in the library space is removed. If there are local calibration sets in the second half of sets that have less than the required number of samples then the last sample in the library space is removed. With this new library of samples formed, with either one or two less samples, a FD range is again calculated and the process starts over.

This iterative process of checking for a range and local calibration set sample number meeting both parameters 3 and 4 is repeated until both conditions are met by every local calibration set formed. If these parameters cannot be met, then no local calibration sets formed are used in the remainder of the LAFR process for that parameter set combination.

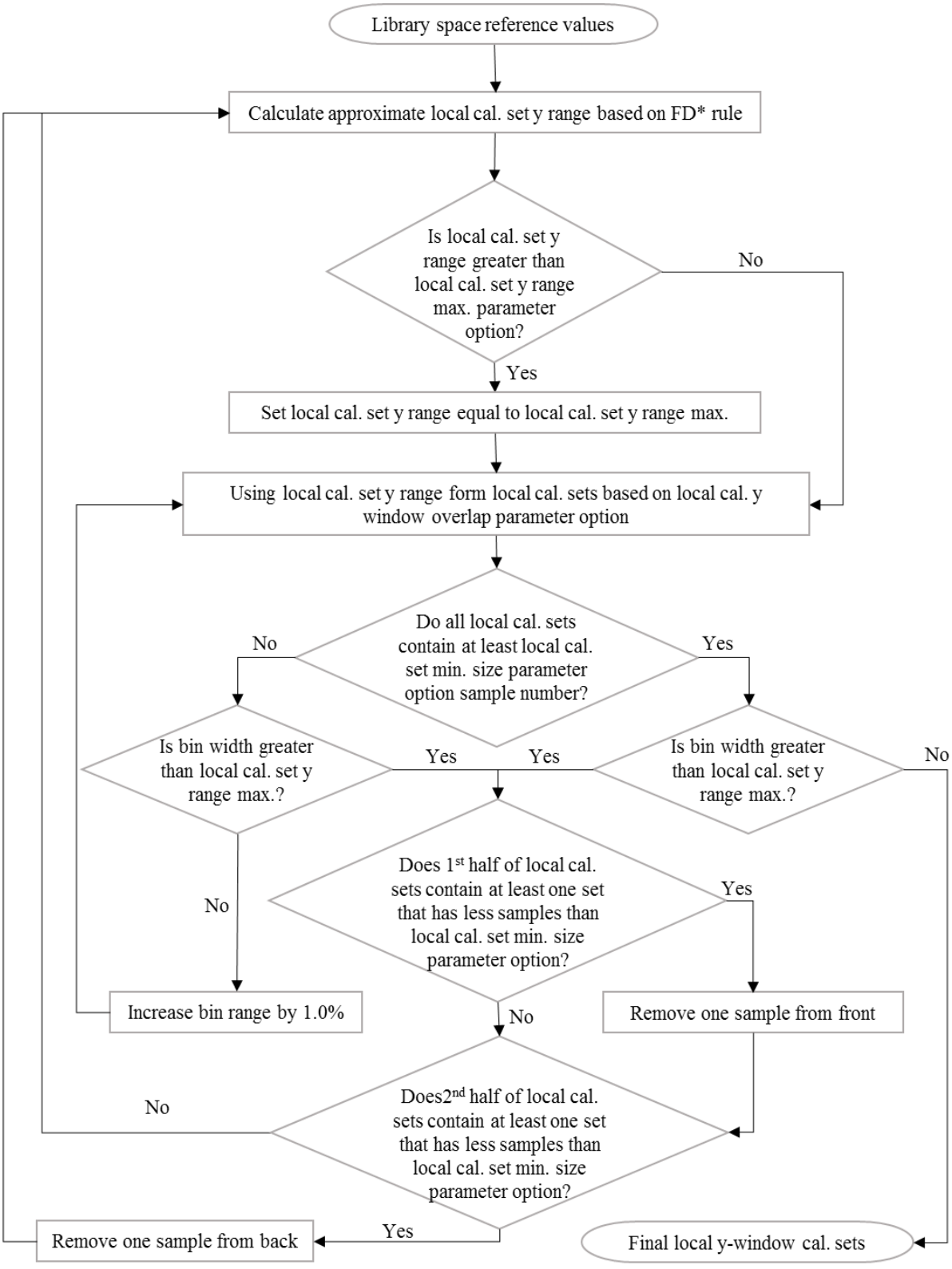


Figure 3.3. Flowchart for the formation of local calibration sets based on specified parameters. *FD refers to the Freedman-Diaconis rule.

2.5. *Calibration Set Comparison Merits*

As discussed in Chapter 2, the local calibration set comparison merits are based on a combination of model prediction errors (Y), spectral matching (Spectral), and orthogonal projections to a regression vector (OP). The purpose of these merits, along with the cross modeling procedures described in Chapter 2, is to select a matrix matched calibration set. This step is important in the LAFR process as it allows for a standardized method for comparing calibration sets formed using multiple sets of parameter options (described in section 2.4). In the LAFR process a calibration set is selected from each of the parameter set combinations first rather than collecting all of the calibration sets from every parameter set combination and selecting a final calibration set. It was determined that this method was more successful at selecting a good local calibration model. Examples of this are shown in the results section (section 5).

3. Methods for Local Adaptive Fusion Regression Setup

3.1. *Global Calibration Models*

All target samples from each dataset are predicted by the global calibration set in order to compare the global prediction results to the local model prediction results formed through the LAFR process. All global models are built using a PLS algorithm. In order to select a tuning parameter, the global dataset is randomly split into 80% calibration samples and 20% validation samples over 20 iterations. To select a model with a good bias/variance tradeoff, the average root mean square of calibration (RMSEC) and root mean square of cross-validation (RMSECV) are plotted against the average of the Euclidean norm of the estimated regression vector ($\|\widehat{\boldsymbol{b}}\|$)[32]. This results in a curve shape resembling an “L”, at least for the RMSEC versus $\|\widehat{\boldsymbol{b}}\|$ plots. The model in the corner

region of the "L-curve" for both the RMSEC and RMSECV is selected for each reference value prediction model. The R^2 values, of the predicted versus the measured reference values, for calibration and cross-validation are also plotted against the $\|\hat{\boldsymbol{b}}\|$, creating an inverted "L-curve" shape, to help inform model selection. The global model for each of the prediction properties can also be selected using the LV selection procedures described in Chapter 2 section 2.2. However, the intention of this LV selection procedure is to automatically select models without manually making a selection. As the global models do not need an automatic selection processes the "L-curve" method described above is used as the RMSEC, RMSEV, and R^2 values can all be used simultaneously for model selection.

3.2. *Data Preprocessing and Software*

All spectra and reference values for each regression model formed are mean centered. For the mean centering process the column mean of the calibration spectra ($\boldsymbol{x}_n$) for each variable n is subtracted from each of the respective variables in each of the individual calibration spectra and the spectra of any samples to be predicted by the calibration samples used to form the regression models. The mean reference values for the calibration samples are also subtracted from each calibration sample and reference values for validation samples not included in the calibration space.

The LAFR algorithm and all regression models formed use code generated in MATLAB R2014b (The MathWorks AB, Kista, Sweden).

3.3. *Selected Local Calibration Set Parameters*

Table 3.8 shows the local calibration parameters for the meat dataset (described in section 4). The parameters option are the same for all three of the reference values used

for the meat dataset. As stated in section 2.2, many of these parameters are held constant for this study. Only parameters 3, with four inputs, and 4, with three inputs, have multiple inputs listed in Table 3.8.

Table 3.8. Local calibration set parameters specified for meat dataset.

Parameter ID	Parameter	Meat
1	Spectral similarity merits	26 spectral
2	Library size[a]	Global, SRD iterations (3SD)
3	Local calibration set min size	**1)** 10 **2)** 15 **3)** 20 **4)** 30
4	Local calibration set y range max	**1)** None **2)** 1/5 global range **3)** 1/10 global range
5	Local calibration set y window overlap	33% of previous window sample number
6	Outlier merits	1st Part: 15 spectral and 3 Y 2nd Part: 15 spectral
7	Local calibration set comparison merits	2 Y, 15 Spectral, and 22 OP
8	Tuning parameters selection	Regression merits: 5 LV SVD merits: 99% cumulative variation
9	Fusion Rules	Library size and outlier detection: SRD Calibration set selection: 6 fusion rules

[a] For the library size based on an SRD iterative selection, the standard deviation(s) (SD) are specified.

4. Dataset

Near infrared spectra from 850-1050 nm over 100 channels (2 nm intervals) for 240 samples of meat were collected by Tecator on a Tecator Infratec Food and Feed Analyzer (FOSS, Tecator AB, Höganäs, Sweden)[33-34]. Absorbance measurements are

reported in $1/\log_{10}$ transmittance units. Reference values were provided for moisture, protein, and fat. All three reference values are used through the LAFR calibration set selection process. The distributions of these reference values is shown in Figure 3.4 (B-D). The established splits for this data set are; training 129, monitoring 43, testing 43. For this work the training set and monitoring set are combined for the global spectra (172 spectra) and the testing spectra were using as the target samples (43 spectra) (Fig. 3.4 A).

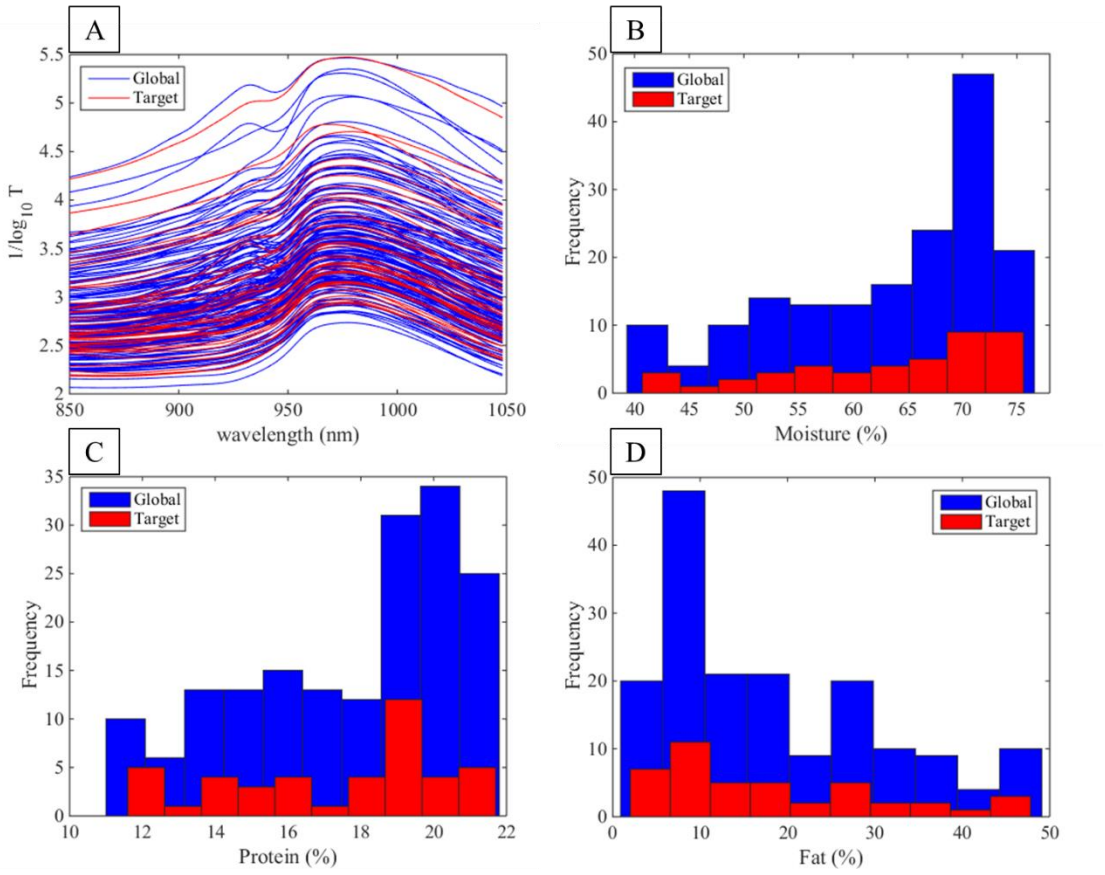


Figure 3.4. Global and target spectra (A) and distributions for moisture (%) (B), fat (%) (C), and protein (%) (D) for meat dataset.

5. Results

5.1. *Global Models*

For the global calibration models, the average prediction errors (RMSEC/CV), prediction versus measured R^2 values and Euclidean norm regression vectors ($\|\hat{\boldsymbol{b}}\|$) are used to select a single model (LV) for each of the three reference properties (Fig. 3.5). This figure also shows the prediction errors (RMSEV) and R^2 values for the target samples. The model prediction information from the target samples is not used to select the LV only to show how the target samples compare to the calibration samples. Though this dataset is typically considered non-linear, the global models selected all had relatively high R^2 values, >0.95, between the predictions and measured values for both calibration and cross-validation for the selected models (Table 3.9). The 43 target samples, on average, are predicted well by the selected models for each of the reference properties, resulting in similar prediction errors and prediction versus measured R^2 values to the calibration.

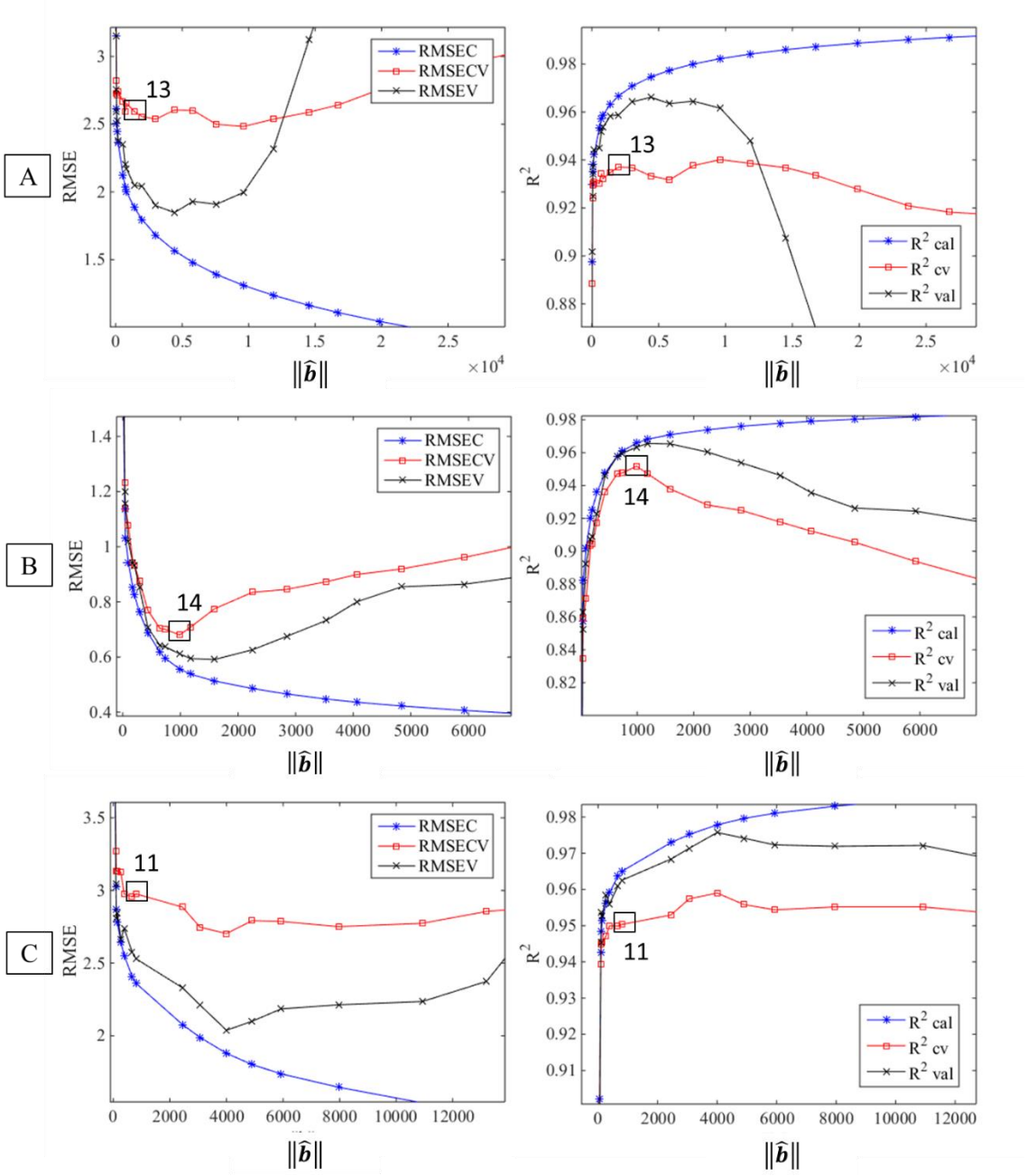


Figure 3.5. RMSE (C/CV/V) and R^2 (cal/cv/val) versus Euclidean norm of the regression vector ($\|\hat{b}\|$) of PLS prediction models for moisture (A), protein (B), and fat (C).

Table 3.9. Global model merits RMSE (C/CV/V) and R^2 (cal/cv/val) for each moisture, protein and fat PLS models.

Model	LV	RMSEC	RMSECV	RMSEV	R^2 cal	R^2 cv	R^2 val
Moisture	13	1.80	2.56	2.04	0.97	0.94	0.96
Protein	14	0.56	0.68	0.61	0.97	0.95	0.96
Fat	11	2.36	2.98	2.53	0.97	0.95	0.96

5.2. *LAFR Results-Moisture (%)*

For the reference value moisture, the LAFR local calibration models perform better than the global model. The final LAFR local model regression results for predictions of all 43 target samples are shown in Figure 3.6 and Table 3.10. For the LAFR local modeling process, a unique local calibration set is selected for each of the 43 samples displayed in this figure. For the local calibration sets selected, five models are built using the five selected LV's for each of the local calibration sets. The Local (max) and Local (min) for the target samples are the maximum and minimum prediction errors these five models formed. This figure shows the predicted target values from the global model, each local model with the highest prediction error, represented as Local (max), and each local model with the lowest prediction error, represented as Local (min), versus the true moisture values. Table 3.10 are the corresponding linear regression statistics for these three regressions including the regression formed by the averaging the predictions of the five LV's for each local calibration set, represented as Local (avg). For all of the proposed local models, the R^2 values between predicted and true values are higher than the global model, ranging from 0.98-1.00 compared to 0.96, and the prediction errors are lower for the local models compared to the global model, 0.44-1.09 versus 2.04. The

linear regression for each of these local models is also closer to the line of equality with slopes near 1 and intercepts close to 0.

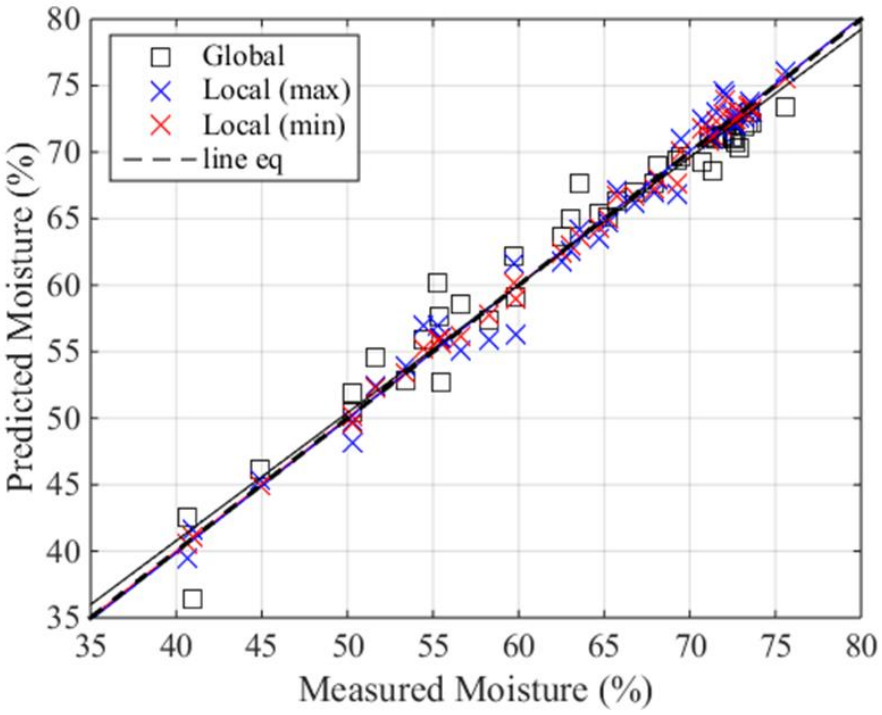


Figure 3.6. Regression of prediction versus measured moisture (%) for global (Global) and local model predictions with highest error (Local (max)) and lowest error (Local (min)).

Table 3.10. Regression statistics of predicted versus measured moisture (%) for global and local predictions for m target samples.

Model	LV	R^2	Slope	Intercept	$\frac{\sum(\lvert\hat{y}-y\rvert)}{m}$
Global	13	0.96	0.96	2.36	2.04
Local (max)	2-15	0.98	1.01	-0.42	1.09
Local (min)	1-15	1.00	1.00	-0.09	0.44
Local (avg)	1-15	0.99	1.01	-0.38	0.72

Figure 3.7 shows some of the results from each of the local calibration sets selected. Plots A, A1 and B show the regression error trends for each of the 43 target samples (sorted based on the reference values) compared to the global regression errors. Plot A1 is plot A in a logarithmic scale. Logarithmic plots are used for other figures throughout the discussion to provide better visual comparisons. In plot A (and A1), the local prediction errors, across the five selected LV's, along with the prediction errors for the global calibration model at LV 13 are shown. From this plot it appears that the global model predicts the samples in the middle of the reference value range (samples 20-35; corresponding to concentrations around 67-72%) better than the samples with lower moisture concentrations. This is probably due to the distribution in moisture concentrations for the global dataset (Fig. 3.4 B). This distribution shows that there are more samples with higher moisture concentrations. The LAFR selected models do not have this same trend of dependency on the distribution of the concentrations. There are target samples represented over the entire moisture range with low prediction errors.

For the LAFR process, there is currently not a method for selecting the final prediction model; hence, why there are five models represented for each local calibration set. However, if the five LV's for the selected models have consistent prediction results, it might indicate that a single model selection method is not required. In this case the average predictions across the five models would be sufficient for the final predictions of the target sample. The five LV's for each local model range from 1-15, as indicated in Table 3.10. This means that some of the local models are represented by LV's 1-5, some are represented by LV's 10-15, and any range of 5 LV's between 1 and 15. Many of the local models' prediction errors are relatively consistent across the five LV's. There are,

however, some LV's that are noticeably different. For example, the prediction error for target sample 16 for the 5th LV shown is not consistent and has the highest prediction error of all of the local models represented in this plot. This LV inconsistency does help to justify the need for more than one LV represented for these local models throughout the LAFR process. Selecting LV's based on the calibration samples with cross-validation methods will not always result in a good prediction model for a target sample. A range of LV does, however, allow for a greater probability of having at least one model that does predict the target sample sufficiently.

In plot B, the global target sample regression errors are subtracted from the local models with the highest target prediction errors (max) and lowest target prediction errors (min). Difference values below zero indicate an improvement in local model prediction error compared to the global model. Target samples 2, 3, 10 and 19 all show great model improvement for both the minimum and maximum target sample prediction errors compared to the global model. There are some instances where the highest error local model predicts worse than the global model while the lowest error local model predicts better than the global; as in target samples 9 and 14. Again, supporting the importance of having a range of LV's. Between samples 20-35, as mentioned, the global model predicts relatively well making it difficult for any local models to outperform the global model prediction.

Plot C shows the final number of calibration samples included in each of the selected local calibration sets. The global calibration model included 172 samples. The maximum number of calibration samples selected to build a local model is 73 (~42% of global). The average number of calibration samples for all of these local models is around

25 (~15% of global). Very few samples from the global dataset are required for building accurate prediction models for this reference property. This plot also shows that the number of samples required for each target sample does vary (10-73). A fixed calibration size, proposed by some of the local model methods described above, would not meet the needs for a local model for each of the target samples represented here. It was important, for this dataset, to allow for multiple minimum calibration sample threshold requirements so that the optimal number of samples was available and could be selected by the calibration set comparison merits.

Along with the local predictions errors and number of samples included in each of the local calibrations, it is important to understand the chemical ranges for these local calibrations sets. As one of the primary goals of the LAFR process is to form and select calibration sets that have matrix matching potential, the reference property range is an important factor. Plot D, for this figure, displays the moisture distributions for each local calibration set with the true reference values of the target samples superimposed. This plot demonstrates that the LAFR process does select an analyte matched local calibration set for each individual target sample using the LAFR process. Chemical matching of the analyte is one important indicator of matrix matching.

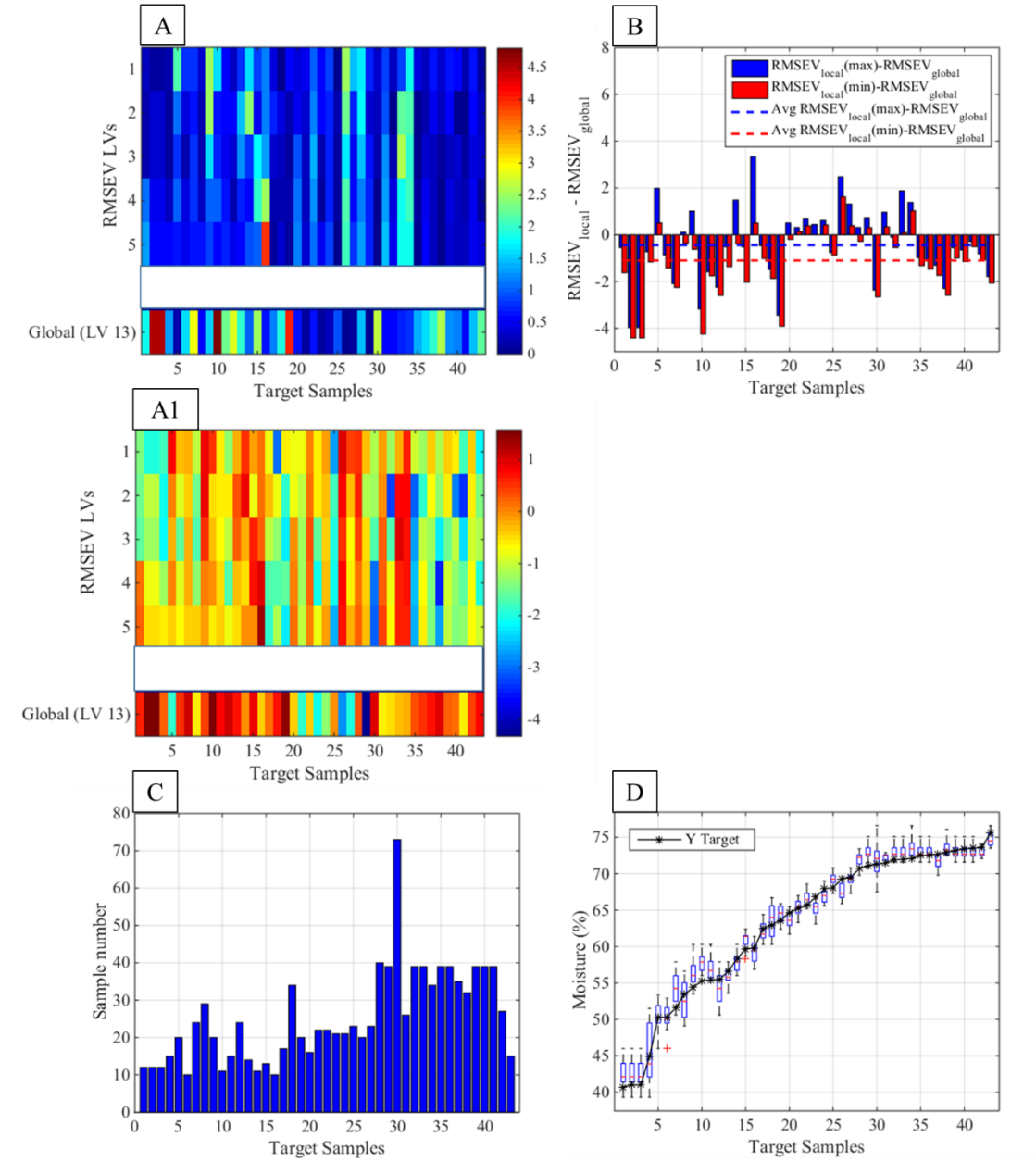


Figure 3.7. LAFR calibration set selection results for the moisture (%) reference value of meat. RMSEV over 5 LV's for LAFR model selections and global RMSEV for LV 13 (A); plot A on a logarithmic scale (A1); difference between RMSEV for LAFR (min and max) and global RMSEV (B); number of samples included in each LAFR selection (C); moisture distribution for each LAFR selection with the corresponding target reference values (D).

5.3. *LAFR Results-Protein (%)*

Figure 3.8 and Table 3.11 show the final local model results for each of the 43 target samples for the reference property protein. The local models for the protein reference values do not perform as well as the local models formed for the moisture reference values. The difficulty with the protein values for this dataset is the chemical range is relatively small in comparison to the other reference values. Also, the prediction errors for the global model are very small in comparison to the other reference values. These errors might be closer to the primary analytical method uncertainty for the global models; however, this unknown for the dataset. These factors seem to have a negative effect on the LAFR process results. The predictions for target samples with reference values between 12-16% protein tended to be predicted poorly by the local models. In general, the average prediction errors for all of the local models were higher than the global model prediction error (Table 3.11). The linear regression for the global model target predictions is also closer to the line of equality than any of the local models.

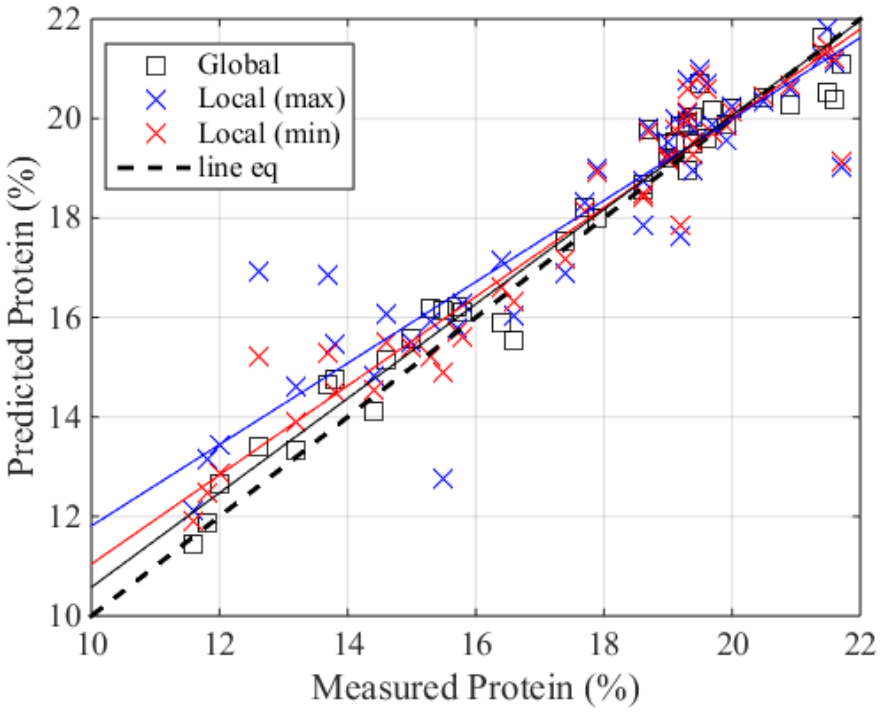


Figure 3.8. Regression of prediction versus measured protein (%) for global (Global) and local model predictions with highest error (Local (max)) and lowest error (Local (min)).

Table 3.11. Regression statistics of predicted versus measured protein (%) for global and local predictions for m target samples.

Model	LV	R^2	Slope	Intercept	$\frac{\sum(\lvert\hat{y}-y\rvert)}{m}$
Global	14	0.97	0.95	1.09	0.48
Local (max)	1-14	0.84	0.82	3.63	0.96
Local (min)	1-15	0.93	0.90	2.08	0.58
Local (avg)	1-15	0.89	0.85	2.98	0.66

Figure 3.9 shows the trends of each of the local models selected for the 43 target samples sorted based on the true protein concentrations. As seen with the regression plot in Figure 3.8, the target samples with the lower concentrations have higher prediction

errors (plots A, A1 and B). The regression error trends show that a few of the target samples have lower prediction errors compared the global (11, 12, 41, and 42). These same samples also have good matches for the protein ranges for the local calibration sets in plot D. The target samples that have the highest prediction errors tend to have local calibration set protein ranges that are not consistent with the target reference values; samples 5 and 43 for example. This demonstrates how important chemical ranges are for local modeling with the purpose of matrix matching.

Though the protein data do not result in better prediction models for the target samples, there is still evidence of the positive aspects of the LAFR process in general. The protein ranges for the selected local calibration sets do not match the target reference values as closely as the moisture local calibration sets did; however, many of the local calibration sets' protein ranges are still in general very similar to the target samples protein values. This shows the power of this algorithm in selecting chemically matched calibration sets. One possible explanation for the poor prediction errors is that only a few parameter set option combinations are assessed for this dataset. These parameter set options might not be ideal for forming the best predicting matrix matched local calibration sets for the protein reference values. Another possibility is that the entire wavelength range was used for this process. Additional wavelength selection methods might be beneficial in improving the prediction abilities of the local models formed for the protein reference value.

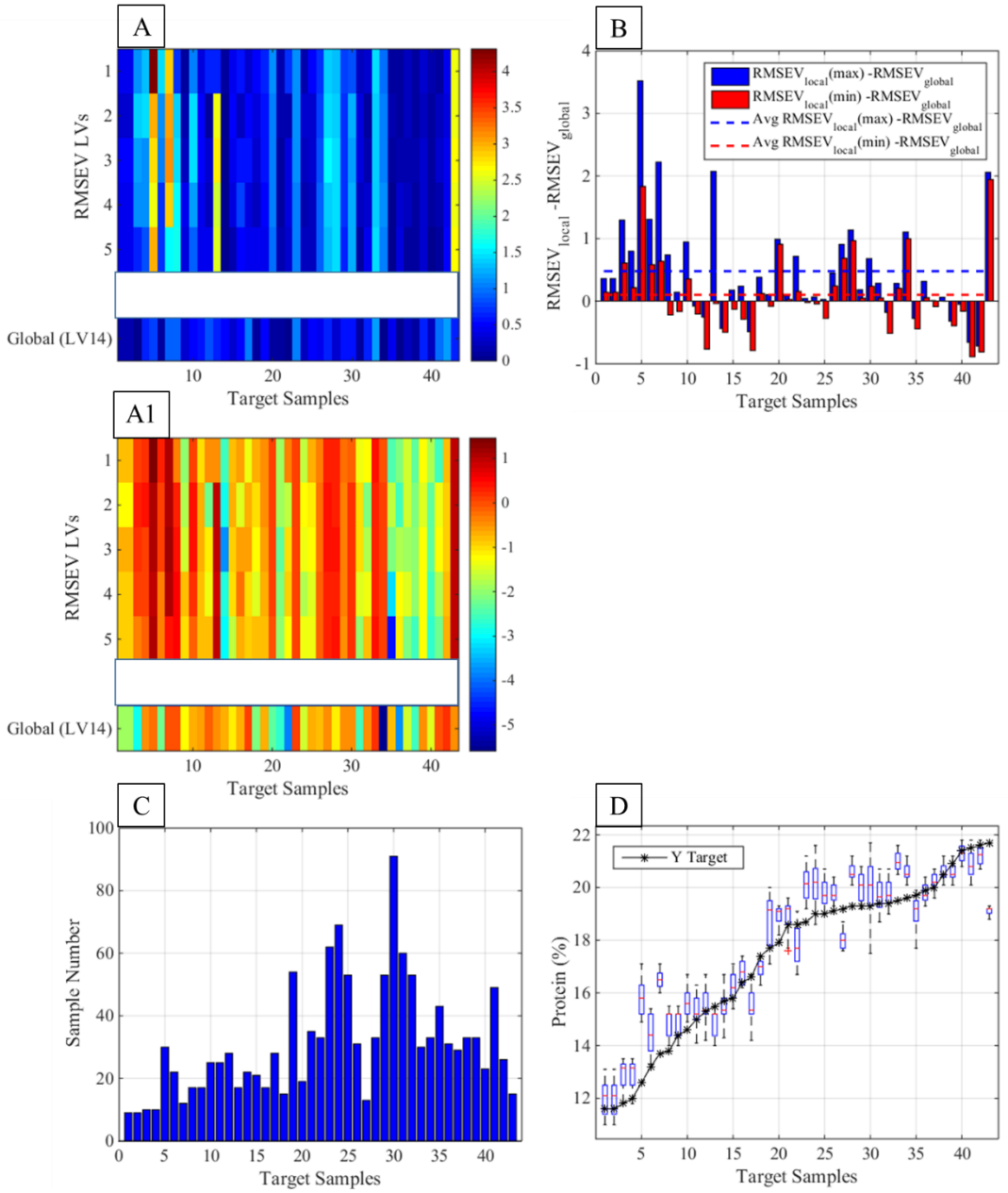


Figure 3.9. LAFR calibration set selection results for protein reference value of meat. RMSEV over 5 LV's for LAFR model selections and global RMSEV for LV 14 (A); plot A on a logarithm scale (A1); difference between RMSEV for LAFR (min and max) and global RMSEV (B); number of samples included in each LAFR selection (C); protein distribution for each LAFR selection with the corresponding target reference values (D).

5.4. *LAFR Results-Fat (%)*

The final LAFR results for the reference property of fat are similar to the moisture property local calibration results. Figure 3.10 and Table 3.12 show the regressions for the fat local models and the global model for each of the 43 target samples. The local models with the highest error (Local (max)) performed slightly worse than the global model. The prediction errors were higher (2.43 compared to 2.07). The local models with the lowest prediction errors (Local (min)), however, performed better than the global model with prediction errors around 0.70. This inconsistency with the local models formed for each target sample promote the need for a final LV selection process. For these target samples, the prediction errors from the models averaged across the five LV's are still lower than the global model prediction errors.

Unlike the moisture and protein reference values for the meat dataset, fat has multiple literature examples for calibration model forming processes, other than local modeling, used to form prediction models for the same 43 target samples[35-38]. Comparing these results to the LAFR results indicate how well the LAFR process works. In terms of prediction error, the LAFR process performed better than two of the studies. One study explored a modified penalized signal regression technique with a prediction error of 1.73[35]. A second study, using a stacked regression technique for the predictions from multiple spectral preprocessing models, reported a prediction error of 1.82[36]. There were also two studies found that did slightly better than the LAFR process if the minimum local regression error models were the final models selected. In one study, focused on variable selection applications, the reported error was 0.66[37], and in the second study, regarding an alternative support vector regression algorithm, reported an error of 0.48[38].

The LAFR process is comparable to the prediction errors reported for these methods with some identified room for prediction error improvement. However, it is important to look at the other properties of regression models, including correlation measurements for the predictions versus the measured values. The LAFR process for the fat property results in high R^2 values (0.95-0.99) for the predicted values versus the measured values. As R^2 values were not reported in these other studies, they cannot be compared to the LAFR results.

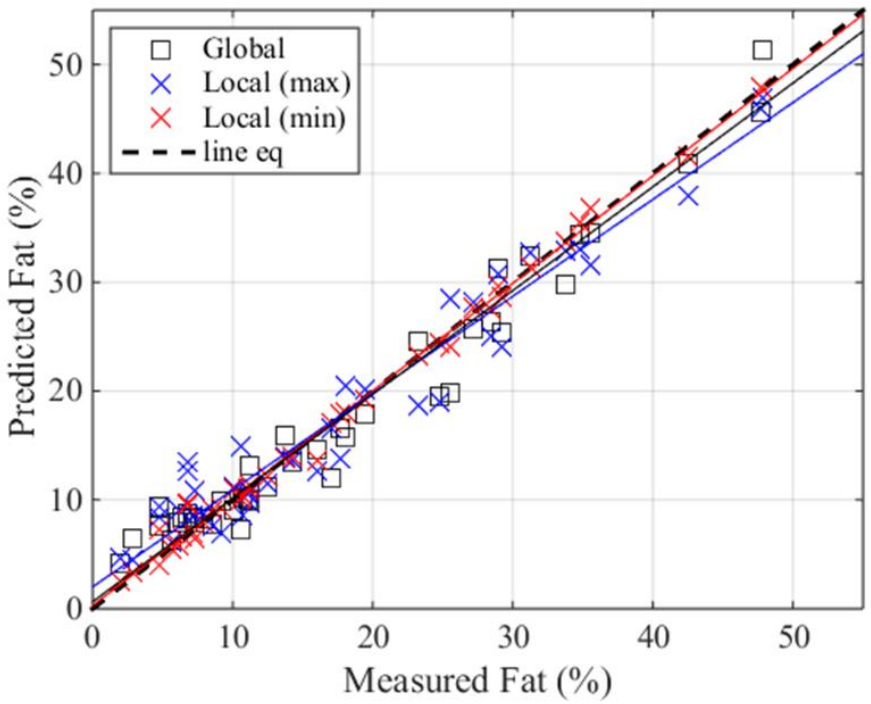


Figure 3.10. Regression of prediction versus measured fat (%) for global (Global) and local predictions with the highest error (Local (max)) and lowest error (Local (min)).

Table 3.12. Regression statistics of predicted versus measured fat (%) for global and local predictions for m target samples.

Model	LV	R^2	Slope	Intercept	$\frac{\sum(\lvert\hat{y}-y\rvert)}{m}$
Global	11	0.96	0.95	0.64	2.07
Local (max)	1-10	0.95	0.89	2.01	2.43
Local (min)	1-11	0.99	0.99	0.29	0.70
Local (avg)	1-11	0.97	0.96	1.56	1.52

Figure 3.11 reflects trends of inconsistency across the LV's for each local model seen in the Local (min) and Local (max) regression statistics (Table 3.12). In plot A, the ranges of prediction errors is very large for some of the target samples. For example, target sample 9 ranges 0.05 to 5.8. Plot B shows that the local models with the highest errors are worse than the global model for more than half of the target samples, while a majority of the local models with the lowest predictions errors perform better than the global model. One reason for the wide spread in LV's could be due to the number of selected samples for each of the local models (plot C). The average number of samples is around 12 (~7% of global). A few of the models formed have only 6-7 samples. Note that even though the minimum sample number is set to 10 as one of the parameter options, after the outlier check process, the number of samples left in each local model can be less than this minimum set value. For this reference value property, as most of the local models had very few samples included, five LV's might be too high. The range represented by the LV's for each model and comparison merits dependent on PLS algorithms needs to be meaningful. Further investigation is required for optimizing the LV selection process and number of LV's to include. For these results the average

predictions across the five LV's is not ideal for final predictions for each of the 43 target samples.

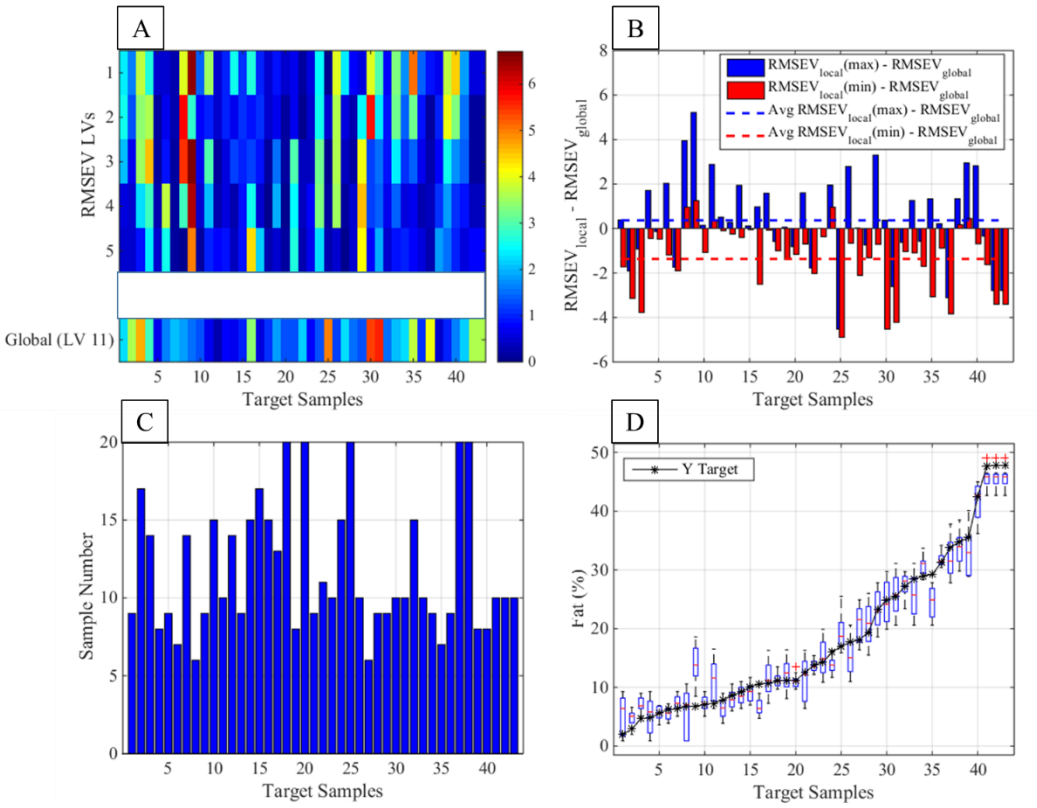


Figure 3.11. LAFR calibration set selection results for fat reference value of meat. RMSEV over 5 LV's for LAFR model selections and global RMSEV for LV 11 (A); difference between RMSEV for LAFR (min and max) and global RMSEV (B); number of samples included in each LAFR selection (C); fat distribution for each LAFR selection with the corresponding target reference values (D).

Though the LV's for the local models and calibration comparison merits might not have been ideal for the local models formed based on fat, the chemical matching ranges in plot D show the same positive trends as the moisture and protein selected local

models. A majority of the local calibration sets selected for the fat target reference values have good agreement in the fat ranges of the selected local calibration sets. Only three target sets, 9, 35, and 16, have ranges that do not encompass the target reference values based on the entire range represented by the boxplot (including outliers) but are still relatively close.

5.5. *Selection Calibration Sets for Moisture (%)*

For the meat spectra, the parameter option combinations, introduced in Table 3.8, result in anywhere between 12-108 unique combinations depending the on the number of SRD iterations that occur when forming the local library spaces; the same library spaces are formed regardless of the reference property (explained in section 2.1). The number of SRD iterations span from none to nine for this dataset. Each of these parameters options form multiple local calibration sets that are dependent on the reference values, as explained in section 2.4. The number of individual calibration sets formed and analyzed after the outlier check process for each of the 43 target samples range from 13 to 405. Meaning that 405 individual local calibration sets are assessed for at least one of the 43 target samples.

For the LAFR process, there are two calibration set selection steps (Fig. 3.1 step 11 and 15); the selection of a calibration set formed from each local calibration parameter set option combinations and the selection of the final calibration set from the parameter based calibration set selections. The resulting selections for each step are investigated for two of the target samples from the moisture property local model results; target sample 26, which has the highest average prediction error, and target sample 2, which has the lowest average prediction error (Figure 3.7 plots A and B). The parameter set

combinations shown for each target sample in the following section are the same local calibration sets represented by the final LAFR results in Figure 3.7. These results are used to demonstrate how the final local calibration set is selected using comparison merits and fusion rules in each of the calibration set selection steps.

5.5.1 Parameter Option Calibration Set Selection

The parameter option calibration set selection (Step 11 in Fig. 3.1) acts as a "rough" selection process selecting the best matrix matched local calibration set from a specific parameter option combination. Figure 3.12 shows data for the local calibration sets formed from the set parameters for target sample 26 (the target sample with the highest average prediction errors). The parameter set combination shown here is as follows; parameter 2: 'Global' library size; parameter 3: minimum of 15 samples; and parameter 4: 1/10 of the total global moisture range as the maximum y range threshold. Only parameters 2-4 are noted as the other parameters stay constant for the calibration set formation process. Parameters 3 and 4 are manually adjusted, and parameter 2 iteratively adjusts.

This parameter option combination forms 13 local calibration sets. Plot A are the normalized calibration comparison merits for each of the 13 calibration sets. Unlike the NMR and corn data comparison merit plots (Fig. 2.12 and 2.15) described in Chapter 2, there is not a single visually discernable matrix matched calibration set. Plot C shows the rankings for the six fusion methods for these 13 local calibration sets. Local calibration set 6 ranks lowest for 5 of the 6 fusion methods, calibration set 7 ranks the second lowest, and calibration set 8 ranks third lowest. These rankings match the local calibration set

ranges shown in Plot D as calibration sets 6-8 encompass the target sample reference value superimposed on this on plot.

The prediction errors for these calibration sets (plot B), do not necessarily reflect the same trends as plots C and D, where calibration sets 6 appears to be matrix matched. The LV's represented by calibration set 6 are 9-13, selected based on the methods described in Chapter 2 section 2.2. The RMSEV's for target sample 26 over LV range is 1.7-2.5. Looking at all of the LV's possible for calibration set 6 reveals that, LV's 5-7 have RMSEV values ranging from 0.2-0.8, outperforming the global model. This indicates that five LV's might be too limited in certain cases for selecting an accurate prediction model for a target sample or that the current method for selecting the five LV's does not select the best 5 LV's. One of the challenges identified through the adaptive local modeling literature presented[6, 13, 15-19] is the requirement to select a single LV within the local modeling algorithm in order to incorporate chemical information based on predictions. This example shows that LV selection based partially on the cross-validation prediction errors of the calibration set does not always identify models that can predict a potentially matrix matched target sample, even when a range of five LV's is used. The *RMSECV* in equation 2.13 for LV selection is currently based on a leave-one-out cross validation method, which can sometimes lead to overfitting to the calibration samples for prediction models[39]. For this sample, it appears that the algorithm did select a potentially matrix matched calibration set; however, the models selected were not sufficient for predicting the target sample.

The results for the parameter option set combination for target sample 2 are shown in Figure 3.13. The parameter options shown here include; parameter 2: 'Global'

library size; parameter 3: minimum of 10 samples; and parameter 4: no set range as the local calibration set y range max. This target sample parameter option combination only has 3 local calibration sets formed that were processed through the LAFR algorithm. Any calibration sets where the target sample is an outlier are not processed through the calibration set comparison methods in the LAFR process. As parameter 4 does not denote a maximum y range, the moisture ranges for these local sets tended to be larger than the local calibration sets formed for target sample 26 (plot D). For this target sample, the comparison merits (plot A) do visually identify calibration set 1 as the best matrix matched. This identification is supported by the fusion method rankings in plot C. The target sample prediction errors (plot B) and analyte range for calibration set 1 (plot D), also indicate that this set is the best matrix matched calibration set of the three. The LV range for local calibration set 1 is 3-7. This range also corresponds with the minimum RMSEV value possible for target sample 2. This calibration set only contains 12 samples suggesting that five LV's, in this particular case, are probably sufficient for covering a range of models where a good prediction model for both the calibration samples and the target sample is possible.

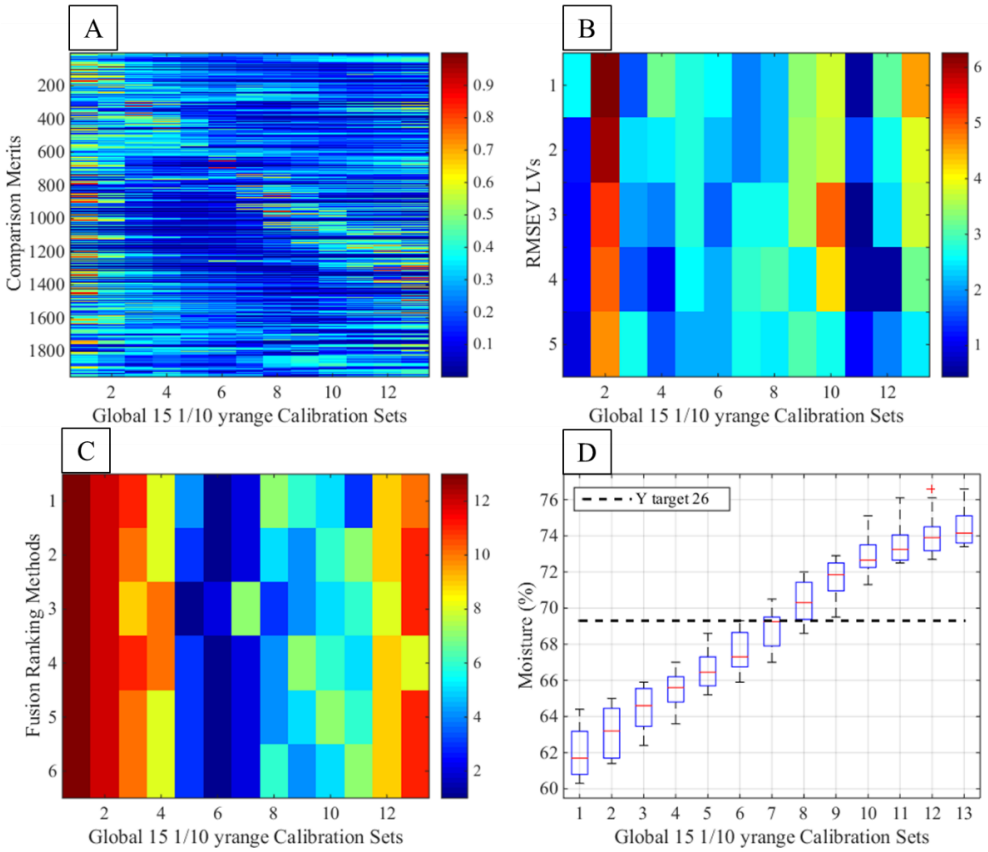


Figure 3.12. Parameter set combination 'Global; 15min; 1/10 y range' local calibration sets for target sample 26 for moisture (%) property. Comparison merits (A); RMSEV over 5 selected LV's (B); fusion ranking methods (C); local calibration set moisture distributions and target sample 26 reference value (--) (D).

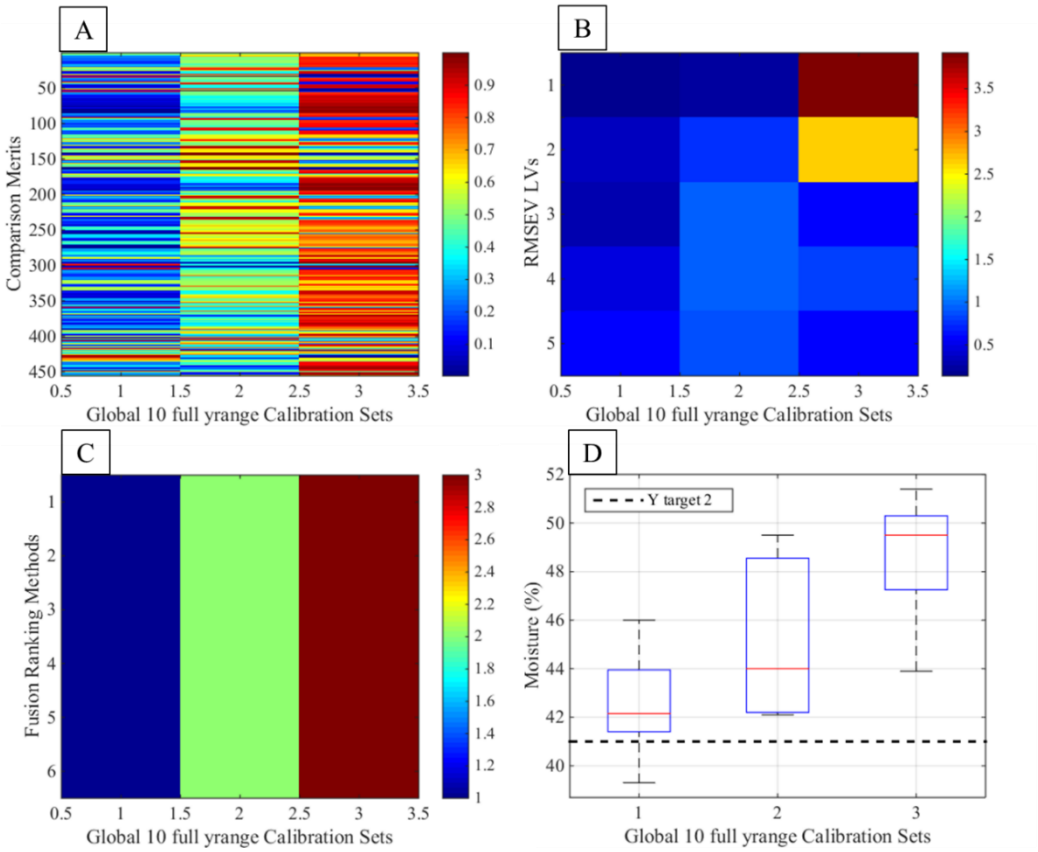


Figure 3.13. Parameter set combination 'Global; 10min; no y range' local calibration sets for target sample 2 for moisture (%) property. Comparison merits (A); RMSEV over 5 selected LV's (B); fusion ranking methods (C); local calibration set moisture distributions and target sample 2 reference value (--) (D).

5.5.2 Final Calibration Set Selection

The final calibration set selection (Step 15 in Fig. 3.1) acts as a "fine" selection method to select the best matrix matched local calibration set from all of the sets selected based on each unique parameter option combination. Most of the calibration sets selected from the parameter option combinations have analyte ranges that are relatively similar to

the target sample reference value; however, this is only the case if a local calibration set was formed that had an appropriate analyte range. In some cases, the global model, after an outlier removal process, can be the best matrix matched calibration set. To account for this, the outlier-cleaned global calibration set is included as a possible calibration set for the final selection process. The final calibration sets selected are shown for both target samples 26 and 2 in Figures 3.14 and 3.15 respectively.

The comparison merits (plot A), regression target sample prediction errors (plot B), fusion ranking results (plot C), and the ranges for the moisture data (plot D) for all of the parameter selected calibration sets for target sample 26 are shown in Figure 3.14. The first set for all four plots is the outlier-cleaned global calibration set. The comparison merits clearly show that the global calibration set is not the best matrix matched. Based on plot B, however, the global calibration set does have the lowest prediction errors. In plot D, almost all of the moisture ranges for the parameter based calibration sets are chemically matched to the target reference value; however, the prediction errors vary widely for each of these sets. The fusion ranking methods identify calibration set 3 (this is the same local calibration set as set 6 from Figure 3.12) as most closely matched to the target sample but the overall ranking orders for the remaining calibration sets are not consistent for the fusion methods.

Figure 3.15 are the plots for the final calibration set selection for target sample 2. There is a large difference between the calibration selection results for target samples 2 and 26. In target sample 2, the calibration set comparison merits (plot A), visually distinguish calibration set 1 as the best matrix matched with consensus from the fusion rankings in plot C. The global calibration set for these plots is set 10. The moisture range

for calibration set 1 is also the closest matched to the target sample. Unlike the final calibration sets for target sample 26, there is a wide range of the moisture local calibration ranges shown in plot D, most of which, do not correspond well to the target sample moisture value. This is most likely due to the sparse number of global samples available in the analyte range of target sample 2. The target reference value is around 41% moisture. From Figure 3.4 it appears that there are only 15 samples ranging from 39-46% moisture available in the global dataset.

There are multiple explanations for the result differences between target samples 2 and 26. All of the final local models selected from the parameter set combinations for target sample 26 had higher LV ranges (8-16) than target sample 2 with LV ranges of 2-11. As discussed with the parameter option calibration set selection for target sample 26, the higher LV's could indicate that the local calibration sets formed tend to be more fitted to the calibration samples. The inconsistencies in the fusion rankings for target 26 (Fig. 3.14 C) could also indicate that the comparison merits used are not as efficient for distinguishing between the calibration sets formed for this sample. Typically, when the fusion rankings show inconsistencies the comparison merits have inconsistent trends across the different merits for each calibration set being compared.

Another aspect of the local modeling effort is the selection of spectrally matched calibration sets. Figure 3.16 shows the spectra for the final local calibration sets selected for both target samples compared to the global spectra. For the final calibration sets selected for both target samples 2 and 26 in the library space used to form "y-window" calibration sets are based on the "Global" library space (parameter 2). It would be expected that a small analyte range would result in a small spectral range. This

expectation is seen for the final calibration selection spectra for target sample 2 (Fig. 3.16 B). However, in some cases, as for target sample 26, the small analyte range of the final selected calibration set did not result in a small spectral range. This spread in the spectral range could help explain the inconsistent comparison merit and data fusion results seen for this target sample.

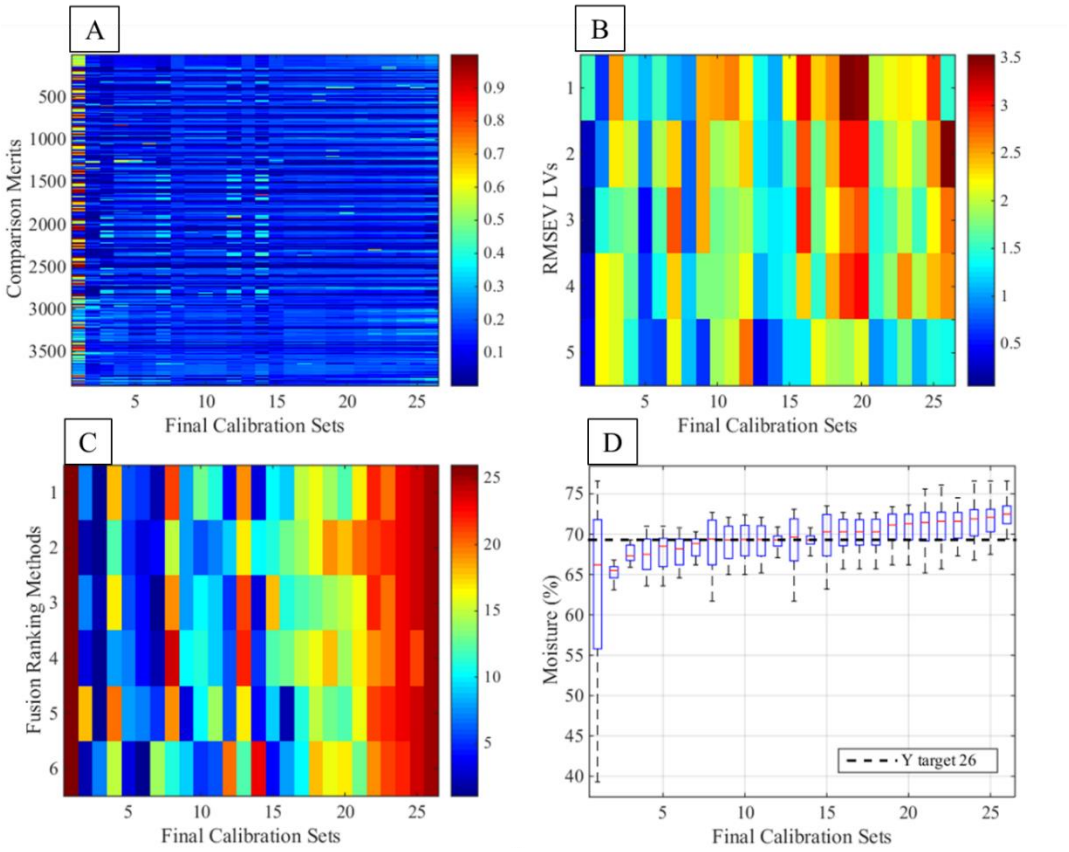


Figure 3.14. Final calibration set selection for target sample 26 and global calibration set as set 1 for moisture (%) property. Comparison merits (A); RMSEV over 5 selected LV's (B); fusion ranking methods (C); local calibration set moisture distributions and target sample 26 reference value (--) (D).

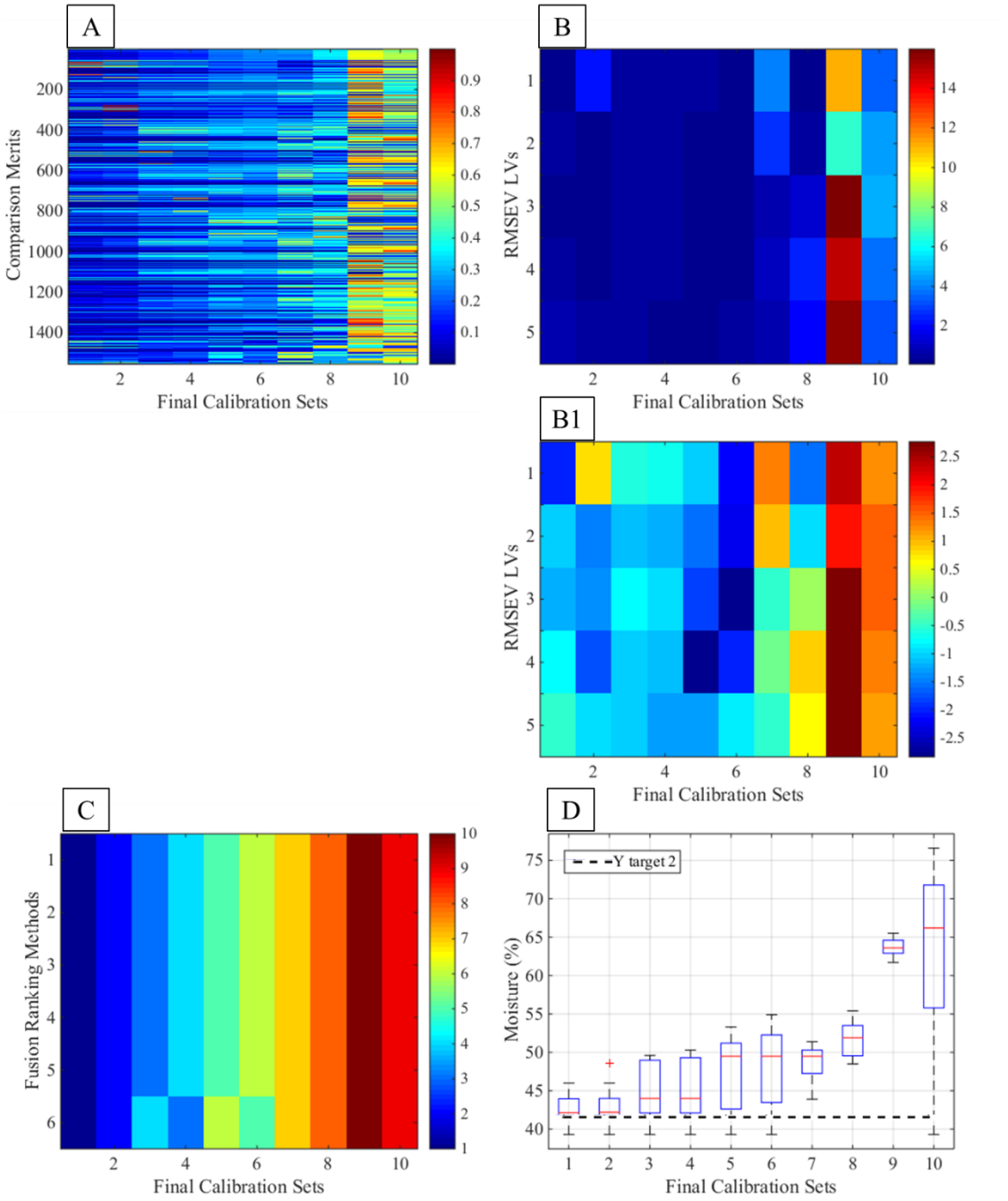


Figure 3.15. Final calibration set selection for target sample 2 and global calibration set as set 10 for moisture (%) property. Comparison merits (A); RMSEV over 5 selected LV's (B); plot B shown a logarithmic scale (B1); fusion ranking methods (C); local calibration set moisture distributions and target sample 2 reference value (--) (D).

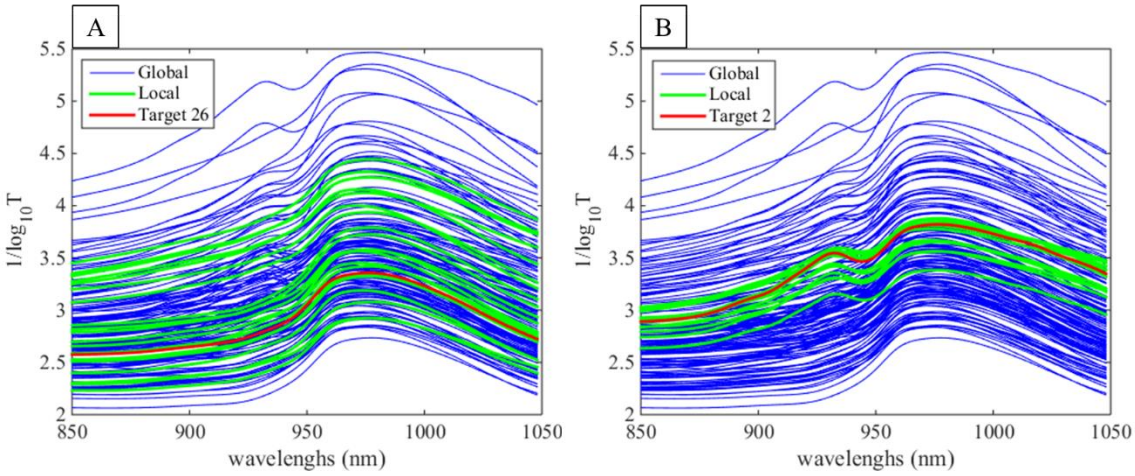


Figure 3.16. Final calibration set spectra comparison for target sample 26 (A) and target sample 2 (B) for moisture (%) property.

6. Conclusion

One of the main challenges identified through the discussion of the final local model results for each of the reference values and the calibration set selection process for two target samples from the moisture dataset is the importance of tuning parameter selection. In local modeling the reference values are unknown for the target samples; however, it is important to include the chemical information, as noted in many local modeling techniques [6-7, 14, 21, 23]. The methods discussed to include the chemical variability and information in these local modeling techniques were based on predictions of target samples compared to the predictions of true values of the calibration samples through iterative local modeling procedures or global models. All of these methods require tuning parameter selection. The LAFR process addressed the need for tuning parameter selection in order to incorporate chemical information into the local sample selection by providing means for the incorporation of multiple tuning parameters. This

multiple tuning parameter selection method was necessary, but still requires more effort for selecting the best tuning parameters available.

For the final local model selected for fat, the limited number of local calibration samples in the final calibration sets did not require the full five LV's used for the LAFR process. The LV number is an adaptable parameter (parameter 8; Table 3.8), but is kept constant for all the reference properties for this dataset. Looking at the algorithm process for calibration set selection, it was noted that target sample 26 for moisture had the potential for a lower prediction error if the range of LV had been expanded or if a different LV selection methods had been applied. As the LAFR process shows that, even with five tuning parameters, the ideal model for each target sample is not always identified. The inclusion of multiple tuning parameters is still likely more beneficial than considering one tuning parameter for each model formed throughout the algorithm.

The LAFR algorithm shows potential as an effective adaptable local modeling technique. Many unique aspects of the LAFR process are beneficial for local modeling. The data fusion approaches throughout the different steps in the LAFR algorithm provide the ability to use multiple merits simultaneously without having to select one or two, as is the case with many local modeling methods. Data fusion is used for the selection of unique spectrally similar spaces for each target sample using 26 spectral similarity merits. Data fusion also allows for comparison of the local analyte range limited calibration sets using both prediction data and spectral data across multiple tuning parameters for all merits in order to select the best matrix matched calibration sets. Additionally, data fusion is applied to the outlier check processes throughout the algorithm.

The other unique aspect of this algorithm is the concept of the flexible parameter options. Many of the local modeling methods proposed require the single selection of a set number of calibration samples, analyte range, or a set threshold for establishing each local calibration set. In the LAFR process, as there is a means for comparing local calibration sets to one another, many different parameter combinations are possible including the number of calibration samples and chemical range combinations included in this study. Because of this flexible parameter input aspect, the LAFR process can be applied to different types of datasets and applications. This study only shows the results for three reference values for a single dataset. The adjustable parameters were limited to the number of calibration samples and analyte range thresholds (e.g. four minimum calibration set sizes and three max chemical range thresholds). The only limiting factor for assessing more combinations of these parameter set options is computational time. As the purpose for many industrial local modeling efforts is to have real-time output information, the LAFR algorithm, in its current form, would not be able to provide these in process local model results if a large number parameter option combinations were required. However, the concepts and process provided by this algorithm do have the potential to be a powerful adaptable local modeling tool.

7. References

1. Kim, S.; Kano, M.; Hasebe, S.; Takinami, A.; Seki, T., Long-Term Industrial Applications of Inferential Control Based on Just-In-Time Soft-Sensors: Economical Impact and Challenges. *Industrial & Engineering Chemistry Research* **2013,** *52* (35), 12346-12356.
2. Guelpa, A.; Bevilacqua, M.; Marini, F.; O'Kennedy, K.; Geladi, P.; Manley, M., Application of Rapid Visco Analyser (RVA) viscograms and chemometrics for maize hardness characterisation. *Food Chemistry* **2015,** *173*, 1220-1227.
3. Fernández-Ahumada, E.; Fearn, T.; Gómez-Cabrera, A.; Guerrero-Ginel, J. E.; Pérez-Marín, D. C.; Garrido-Varo, A., Evaluation of Local Approaches to Obtain Accurate Near-Infrared (NIR) Equations for Prediction of Ingredient Composition of Compound Feeds. *Applied Spectroscopy* **2013,** *67* (8), 924-929.
4. Zamora-Rojas, E.; Garrido-Varo, A.; Van den Berg, F.; Guerrero-Ginel, J. E.; Pérez-Marín, D. C., Evaluation of a new local modelling approach for large and heterogeneous NIRS data sets. *Chemometrics and Intelligent Laboratory Systems* **2010,** *101* (2), 87-94.
5. Fujiwara, K.; Kano, M.; Hasebe, S.; Takinami, A., Soft-sensor development using correlation-based just-in-time modeling. *AIChE Journal* **2009,** *55* (7), 1754-1765.
6. Jin, H.; Chen, X.; Yang, J.; Wang, L.; Wu, L., Online local learning based adaptive soft sensor and its application to an industrial fed-batch chlortetracycline fermentation process. *Chemometrics and Intelligent Laboratory Systems* **2015,** *143*, 58-78.
7. He, K.; Cheng, H.; Du, W.; Qian, F., Online updating of NIR model and its industrial application via adaptive wavelength selection and local regression strategy. *Chemometrics and Intelligent Laboratory Systems* **2014,** *134*, 79-88.
8. Zhang, X.; Li, Y.; Lv, F., Complex batch processes quality prediction using non-Gaussian dissimilarity measure based just-in-time learning model. *IFAC-PapersOnLine* **2015,** *48* (21), 595-600.
9. Kim, S.; Kano, M.; Nakagawa, H.; Hasebe, S., Estimation of active pharmaceutical ingredients content using locally weighted partial least squares and statistical wavelength selection. *International Journal of Pharmaceutics* **2011,** *421* (2), 269-274.
10. Lee, D. E.; Song, J.-H.; Song, S.-O.; Yoon, E. S., Weighted Support Vector Machine for Quality Estimation in the Polymerization Process. *Industrial & Engineering Chemistry Research* **2005,** *44* (7), 2101-2105.
11. Quan, T.; Liu, X.; Liu, Q., Weighted least squares support vector machine local region method for nonlinear time series prediction. *Applied Soft Computing* **2010,** *10* (2), 562-566.
12. Naes, T.; Isaksson, T.; Kowalski, B. R., Locally Weighted Regression and Scatter Correction for Near-Infrared Reflectance Data. *Analytical Chemistry* **1990,** (62), 664-673.
13. Wang, Z.; Isaksson, T.; Kowalski, B. R., New approach for distance measurement in locally weighted regression. *Analytical Chemistry* **1994,** *66* (2), 249-260.
14. Cheng, C.; Chiu, M.-S., A new data-based methodology for nonlinear process modeling. *Chemical Engineering Science* **2004,** *59* (13), 2801-2810.

15. Næs, T.; Isaksson, T., Locally Weighted Regression in Diffuse Near-Infrared Transmittance Spectroscopy. *Applied Spectroscopy* **1992,** *46* (1), 34-43.
16. Dahlbacka, J.; Lillhonga, T., Moisture measurement in timber utilising a multi-layer partial least squares calibration approach. *Journal of Near Infrared Spectroscopy* **2010,** *18* (6), 425-433.
17. Dahlbacka, J.; Lillhonga, T., Quantitative measurements of anaerobic digestion process parameters using near infrared spectroscopy and local calibration models. *Journal of Near Infrared Spectroscopy* **2013,** *21* (1), 23-33.
18. Chang, H.; Zhu, L.; Lou, X.; Meng, X.; Guo, Y.; Wang, Z., A new local modelling approach based on predicted errors for near-infrared spectral analysis. *Journal of Analytical Methods in Chemistry* **2016**, 8.
19. Hazama, K.; Kano, M., Covariance-based locally weighted partial least squares for high-performance adaptive modeling. *Chemometrics and Intelligent Laboratory Systems* **2015,** *146*, 55-62.
20. Centner, V.; Massart, D. L., Optimization in Locally Weighted Regression. *Analytical Chemistry* **1998,** *70* (19), 4206-4211.
21. Chen, K.; Ji, J.; Wang, H.; Liu, Y.; Song, Z., Adaptive local kernel-based learning for soft sensor modeling of nonlinear processes. *Chemical Engineering Research and Design* **2011,** *89* (10), 2117-2124.
22. Anderssen R, S.; Osborne B, G.; Wesley I, J., The application of localisation to near infrared calibration and prediction through partial least squares regression. *Journal of Near Infrared Spectroscopy* **2003,** *11* (1), 39-48.
23. Jin, H.; Chen, X.; Wang, L.; Yang, K.; Wu, L., Adaptive Soft Sensor Development Based on Online Ensemble Gaussian Process Regression for Nonlinear Time-Varying Batch Processes. *Industrial & Engineering Chemistry Research* **2015,** *54* (30), 7320-7345.
24. Jouan-Rimbaud, D.; Massart, D. L.; Saby, C. A.; Puel, C., Characterisation of the representativity of selected sets of samples in multivariate calibration and pattern recognition. *Analytica Chimica Acta* **1997,** *350* (1–2), 149-161.
25. Jouan-Rimbaud, D.; Massart, D. L.; Saby, C. A.; Puel, C., Determination of the representativity between two multidimensional data sets by a comparison of their structure. *Chemometrics and Intelligent Laboratory Systems* **1998,** *40* (2), 129-144.
26. Héberger, K., Sum of ranking differences compares methods or models fairly. *Trends in Analytical Chemistry* **2010,** *29* (1), 101-109.
27. Héberger, K.; Kollár-Hunek, K., Sum of ranking differences for method discrimination and its validation: comparison of ranks with random numbers. *Journal of Chemometrics* **2011,** *25* (4), 151-158.
28. Kalivas, J. H.; Héberger, K.; Andries, E., Sum of ranking differences (SRD) to ensemble multivariate calibration model merits for tuning parameter selection and comparing calibration methods. *Analytica Chimica Acta* **2015,** *869*, 21-33.
29. Héberger, K.; Škrbić, B., Ranking and similarity for quantitative structure–retention relationship models in predicting Lee retention indices of polycyclic aromatic hydrocarbons. *Analytica Chimica Acta* **2012,** *716*, 92-100.
30. Cook, R. D., Detection of influential observation in linear regression. *Technometrics* **1977,** *19* (1), 15-18.

31. Freedman, D.; Diaconis, P., On the histogram as a density estimator: L2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **1981,** *57* (4), 453-476.
32. Kalivas, J. H., Comprehensive Chemometrics. In *Calibration Methodologies*, Brown, S.; Tauler, R.; R, W., Eds. Elsevier: Oxford, 2009; Vol. 3, pp 1-32.
33. Borggaard, C.; Thodberg, H. H., Optimal minimal neural interpretation of spectra. *Analytical Chemistry* **1992,** *64* (5), 545-551.
34. Thodberg, H. H., Ace of Bayes: Application of Neural Networks with Pruning. *The Danish Meat Research Institute* **1993,** *Manuscript 1132*, 1-40.
35. Eilers, P. H. C.; Li, B.; Marx, B. D., Multivariate calibration with single-index signal regression. *Chemometrics and Intelligent Laboratory Systems* **2009,** *96* (2), 196-202.
36. Xu, L.; Zhou, Y.-P.; Tang, L.-J.; Wu, H.-L.; Jiang, J.-H.; Shen, G.-L.; Yu, R.-Q., Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration. *Analytica Chimica Acta* **2008,** *616* (2), 138-143.
37. Rossi, F.; Lendasse, A.; François, D.; Wertz, V.; Verleysen, M., Mutual information for the selection of relevant variables in spectrometric nonlinear modelling. *Chemometrics and Intelligent Laboratory Systems* **2006,** *80* (2), 215-226.
38. Hernández, N.; Talavera, I.; Dago, A.; Biscay, R. J.; Ferreira, M. M. C.; Porro, D., Relevance vector machines for multivariate calibration purposes. *Journal of Chemometrics* **2008,** *22* (11-12), 686-694.
39. Xu, Q.-S.; Liang, Y.-Z., Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems* **2001,** *56* (1), 1-11.