John H. Kalivas Research

Our research focus is an area of analytical chemistry termed chemometrics that is concerned with the mathematical and statistical analysis of chemical data. We concentrate on writing computer algorithms to determine mathematical relationships between chemical data and the properties desired such as a concentration estimate of a substance in a sample. In todays world, the general terminology used is machine learning and artificial intelligence where algorithms are trained to recognize data patterns and predict properties for new situations. For example, training an algorithm to recognize what a cancer cell looks like for prostrate biopsies. Thus, chemometrics could be considered data science using machine learning with a focus on chemical data. Our mission is creating new algorithms to solve difficult analytical chemistry problems by leveraging inter- and intra-molecular interactions as information. Some of our research interests follow.

Multivariate calibration (modeling) is central to many disciplines including food analysis, food adulteration detection and authentication of product origin, environmental monitoring, industrial process analysis, medical diagnosis such as disease detection, pharmaceutical analysis, forensic analysis, detection of hidden radioactive material, and the list goes on. Work in our laboratory consists of developing new mathematical process and computer algorithm implementation in order to improve calibration quality and eliminate user decisions making the calibration and subsequent prediction (analysis) automatic. Another issue with multivariate modeling is maintenance. That is, developing a calibration model in one set of environmental, instrumental, physical, and chemical conditions (the primary conditions) and then updating the calibration to now work in new secondary conditions is a persistent universal problem (including machine learning). Our laboratory also works on methodology developments to solve this issue. In particular, model updating to new conditions without reference values (unlabeled data).

With the ever-growing availability of data, e.g., near infrared (NIR), Raman, fluorescence, etc., spectra), there is a high need for computer algorithms to mine through such libraries of data and identify those spectra similar to a spectrum just measured for a new sample. The user is interested in using the new spectrum for quantitative analysis of an analyte, perhaps glucose content for diabetic, and a model must first be formed using calibration spectra chemically matrix matched to the new sample spectrum. Because the analyte amount in the new sample is not known, identifying matrix matched calibration samples from a library is not a simple task. We are currently working on an algorithm to accomplish this objective. We call the algorithm local adaptive fusion regression (LAFR).

Classification is another major work area in our analytical chemistry research focus and like calibration for quantitative analysis, it is a multidiscipline problem. Classification is concerned with determining which class (category) a sample belongs to such as type of plastic for recycling or if a cell is cancerous. Our laboratory also works on developing new classification methodologies as well as applications such as food authentication or

adulteration detection with the goal to improve prediction quality and remove the user from the decision process.

Linear algebra and Matlab programing are the tools used in our research.

Here a few links on new projects and past students.

https://www.isu.edu/news/2019-fall/isu-researcher-john-kalivas-to-use-big-data-to-refine-cutting-edge-chemical-analysis-methods.html

https://www.isu.edu/newsandnotes/issues/2018/news-and-notes-3-15-2018/idaho-state-university-chemistry-researchers-use-thermal-imaging-to-retrieve-serial-numbers-from-stolen-property-in-a-non-destructive-manner.html

http://headlines.isu.edu/?p=4431

http://headlines.isu.edu/?p=3642

https://www.isu.edu/bengaltracks/issues/october-2014/isu-undergrad-presents-poster-at-analytical-chemistry-conferences-wins-first-place-prize.html

1) *Multivariate Calibration*

Multivariate calibration involves developing a mathematical relationship between a dependent variable and measured independent variables. For example, analyte concentration is the dependent variable (the *m* calibration sample values in the column vector **y**) and measured independent variables are readings over a series of wavelengths (the respective calibration spectra measured over *n* wavelengths as rows in **X**). Mathematically, this is expressed as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \tag{1}$$

Estimation of **b**, the regression or model vector, is obtained by

$$\hat{\mathbf{b}} = \mathbf{X}^{+}\mathbf{y} \tag{2}$$

where the superscript + indicates a generalized inverse of **X**. Once an acceptable regression vector has been obtained, it can be used for analysis of a future sample **x**, the measured spectrum of new sample, by

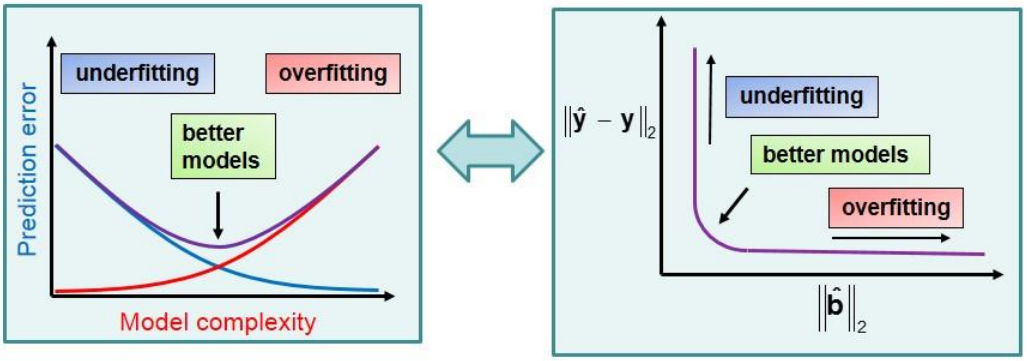
$$\hat{y} = \mathbf{x}^{T}\hat{\mathbf{b}} \tag{3}$$

Estimating **b** is regularly described in analytical chemistry and other fields as a minimization problem in the L_2 norm (2-norm, Euclidean norm, $\|\bullet\|_2$) of

$$\min\left(\left\| \mathbf{Xb} - \mathbf{y} \right\|_2^2 \right) \quad (4)$$

In essence, expression (4) is a minimization of the accuracy error (bias). However, not only is minimizing bias a crucial component in forming a calibration model, but another essential part of calibration is minimizing the variance associated with estimates of *y* for new samples.

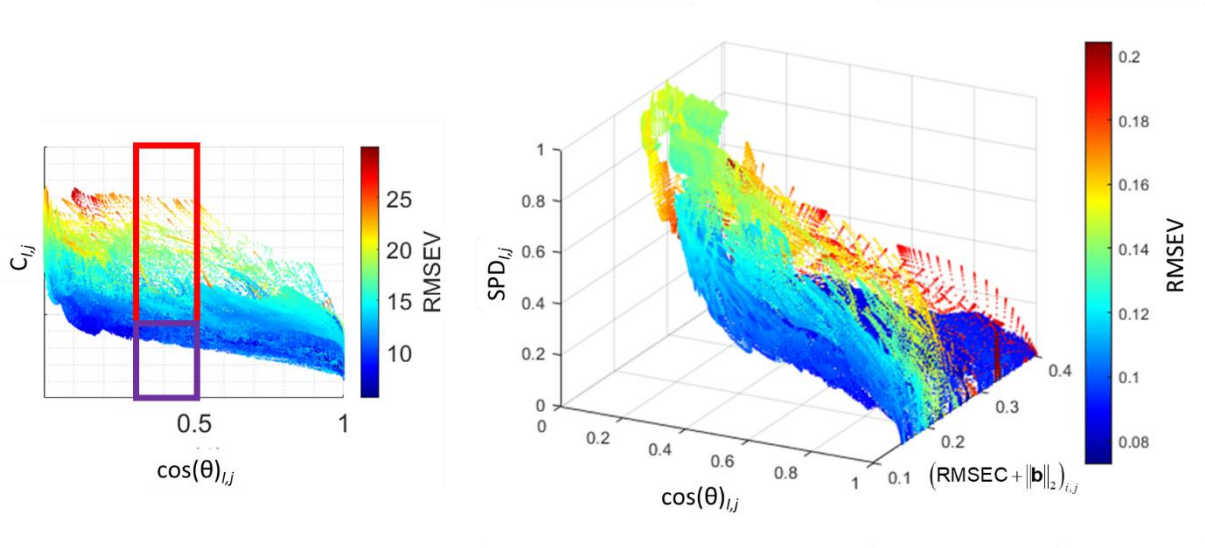
Bias and variance are complementary measures in the sense that a decrease in bias results in an increase in variance for prediction of a sample. As the ***following figure*** shows, there is a tradeoff of bias (prediction error) for variance (model complexity). As respective tuning parameters for calibration methods vary to generate different regression vectors from eq. (2), the bias decreases at a sacrifice to variance increasing and vice-versa as characterized below.



Thus, both issues need to be examined when determining an acceptable $\hat{\mathbf{b}}$. Our lab has recently shown that the underlying bias/variance tradeoff is the corresponding balance of the model selectivity/sensitivity tradeoff and we are leveraging the bias/variance tradeoff for automatic model selection.

2) *Model Selection*

Fundamental to most modeling methods is selection of the best model with a solid bias/variance tradeoff just previously noted. Calibration typically involves one or more tuning parameters and optimizing to the best value(s) is challenging. Typically, selection ends up being either empirical by the user or biased by a selection algorithm towards the merit such a cross validation prediction error. Our lab also works on data fusion processes for model selection incorporating more than one measure of model quality. Recently, we have a developed a generic process that is automatic. The method uses a consensus approach where model diversity and prediction similarity are emphasized. Thus, the selection method is data set independent as well as modeling method independent. The ***following figure*** graphically characterizes our consensus approach where each point is a model pair for over a million points. The darker the blue color, the better the model pair. Our algorithm finds a suitable point (model).



3) *Calibration Maintenance*

Once a multivariate model is estimated, the duration of the model usefulness becomes relevant. The current *primary* calibration model can fail due to a host of reasons such as uncalibrated spectral features appearing in new *secondary* samples later in time, the calibrated analyte is now lower or higher than the primary calibration concentrations in **y**, or other new secondary effects resulting from changes in the instrument and sample type. Our research lab works on developing new mathematical processes to maintain the current primary model to deal with new chemical, physical, environmental, and/or instrumental effects not in the current calibration domain. Some of these new methods are listed in the ***Table below***.

Models	Terms and Notes
$\begin{pmatrix}\mathbf{y}_P \\ \lambda\mathbf{y}_S\end{pmatrix}=\begin{pmatrix}\mathbf{X}_P \\ \lambda\mathbf{X}_S\end{pmatrix}\mathbf{b}$ $\begin{pmatrix}\mathbf{y}_P \\ \mathbf{0} \\ \lambda\mathbf{y}_S\end{pmatrix}=\begin{pmatrix}\mathbf{X}_P & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_P \\ \lambda\mathbf{X}_S & \beta\mathbf{X}_S\end{pmatrix}\begin{pmatrix}\mathbf{b}_P \\ \mathbf{b}_S\end{pmatrix}$ $\begin{pmatrix}\mathbf{y}_P \\ 0\end{pmatrix}=\begin{pmatrix}\mathbf{X}_P \\ \lambda(\bar{\mathbf{x}}_P-\bar{\mathbf{x}}_S)\end{pmatrix}\mathbf{b}$ $\begin{pmatrix}1 \\ \mathbf{0}\end{pmatrix}=\begin{pmatrix}\mathbf{k}_P^T \\ \lambda\mathbf{N}\end{pmatrix}\mathbf{b}$ $\begin{pmatrix}\mathbf{y}_P \\ \mathbf{0}\end{pmatrix}=\begin{pmatrix}\mathbf{X}_P \\ \lambda\mathbf{X}_{P(S)}\left(\mathbf{I}-\mathbf{V}_{S(P)}\mathbf{V}_{S(P)}^T\right)\end{pmatrix}\mathbf{b}$	$\mathbf{X}_P$ and $\mathbf{X}_S$ are primary and secondary measured data with corresponding reference values $\mathbf{y}_P$ and $\mathbf{y}_S$ λ and β are adjustable tuning parameters **k** is a pure component analyte spectrum and **N** are the non-analyte spectra **V** eigenvectors from the SVD of respective matrices Note: Models are solved by PLS or Tikhonov regularization processes. One norm (L_1) variations are also used for sparse models.

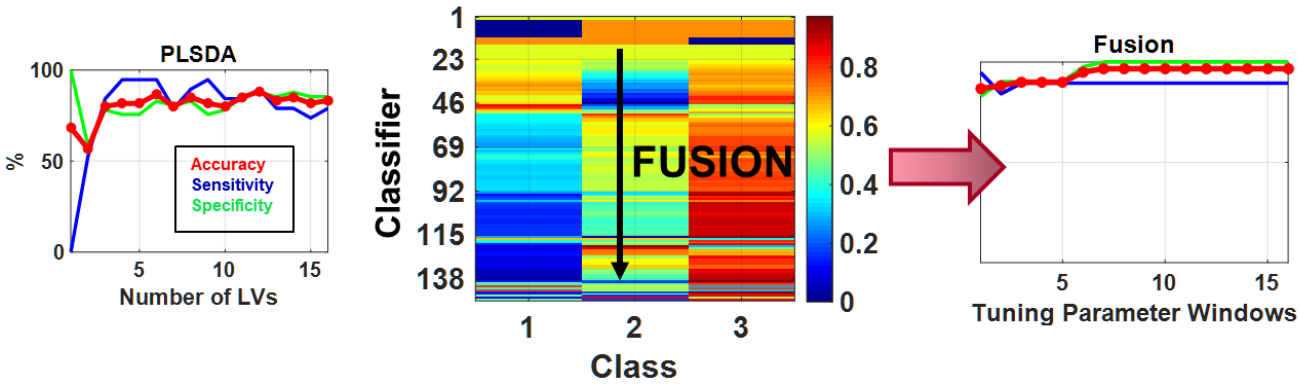
Crucial to useful calibration maintenance is being able to update (tune the current model) to the new conditions with only unlabeled data from the new conditions. Specifically, obtaining the reference values for **y** is the time consuming and expensive step. Removing this step allows instantaneous model updating needed for on-site analysis.

Application of the methods presented in the table include using a single analyte pure component spectrum augmented with blank samples (samples without the analyte) such as analysis of food adulterants in extra virgin olive oil samples or the active pharmaceutical ingredient (API) content in medicine. Other situations involve model updating to a new instrument, new geographical region, a new year such as growing season, or a new patient for medical diagnostics.

A large part of calibration maintenance is assessing how different primary and secondary conditions are. That is, how well do the conditions matrix matched. Our lab also works on how best to determine this. This step is especially difficult with unlabeled secondary data because part of matrix matching is the analyte and interferent concentration profiles. Without knowing all reference values, this step is difficult. Our new process is being used for new global and local modeling methodologies. We named the methods global adaptive fusion regression (GAFR) and local AFR (LAFR).

4) *<u>Data Fusion for Outlier Detection and Classification without Model Optimization</u>*

Recently our research has broadened into data fusion processes to avoid model selection. Our approach has proven successful for outlier detection and classification problems including signal classifiers. ***Below is an example*** of the improvement we obtain with our method that uses fusion of non-optimized classifiers. Because the optimization step is removed, the training step is also removed.



Data sets we have worked with include determining product of origin (product authentication) such Argan oil, beer, fava beans, perfume, flower species, and purity of food products such as strawberry puree and meat (adulteration studies).

5) *Chemical Education*

We have on going chemical education projects in our research laboratory. These consist of developing new laboratory exercises for general chemistry, quantitative analysis, and instrumental analysis. In the past, our laboratory has developed new guided inquiry labs for general chemistry and quantitative analysis including chemical analysis of live trout for their fat and moisture contents. Currently, we are working on new labs for instrumental analysis that provide greener approaches to multivariate calibration using one of our newly developed model updating process.

Service-learning is becoming ever more important in the education of students to become responsible chemists. Service-learning involves students in thoughtfully organized service activities addressing community needs and complementing students' academic studies. Service-learning results from a curriculum that extends the classroom into the community combing education and service and includes class time to reflect on the service experience. Our research desires are to develop new service-learning components in general and analytical chemistry courses.