analytical chemistry

pubs.acs.org/ac Article

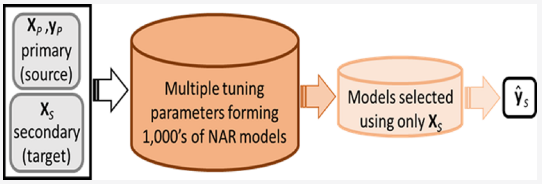
Calibration Model Updating to Novel Sample and Measurement Conditions without Reference Values

Robert C. Spiers and John H. Kalivas*

Cite This: *Anal. Chem.* 2021, 93, 9688−9696 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Updating a calibration model formed in original (*primary*) sample and spectral measurement conditions to predict analyte values in novel (*secondary*) conditions is an essential activity in analytical chemistry in order to avoid a complete recalibration. Established model updating methods require sample analyte reference values for a small set of secondary domain samples (labeled data) to be used in updating processes. Because obtaining reference values is time consuming and is the costly part of any calibration, methods are needed that do not require labeled secondary samples, thereby allowing on demand model updating. This paper compares model updating methods with and without labeled secondary samples. A hybrid model updating approach is also developed and evaluated. Unfortunately, a major impediment to adapting a model without secondary analyte reference values has been model selection. Because multiple tuning parameters are commonly involved in model updating methods, thousands of models are formed, making model selection complex. A recently developed framework is evaluated for automatic model selection of several two to three tuning parameter-based model updating methods without secondary analyte reference values (labels). The model selection method is based on model diversity and prediction similarity (MDPS) of the unlabeled samples to be predicted. The new secondary samples to be predicted can be used to form the updated models and again to select the final predicting models. Because models are formed and selected on demand to directly predict target samples, complicated cross-validation processes are not needed. Four near-infrared data sets covering 40 model updating situations are evaluated showing that MDPS can select reliable updated models outperforming or rivaling prediction errors from total recalibrations with secondary reference values.

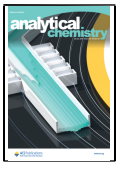


A prominent concern in analytical chemistry is to decrease the time and cost required to provide accurate analyses of sample compositions. Multivariate calibration is a partial solution where a calibration model is formed relative to a large number of reference samples spanning expected measurement and sample variances (matrix effects), including both analyte and interferent amounts such as concentrations. However, regardless of how prudently the expected variances are spanned, circumstances arise that invalidate new sample analyte predictions obtained from the original calibration model. For example, measurement conditions change, such as temperature or instrument components, or new samples are measured on different instruments. Sample conditions can also deviate such as new chemical processing batches, agriculture varieties, or growing seasons affecting spectra and analyte concentration ranges. The original and new sample and measurement matrix effects are respectively termed primary and secondary conditions. Other terminology sometimes used are a source for primary and a target for secondary.

Mathematically, the primary calibration situation is expressed by $\mathbf{y} = \mathbf{Xb} + \boldsymbol{e}$, where $\mathbf{y}$ denotes an $m \times 1$ column vector of analyte values for m samples, $\mathbf{X}$ symbolizes the $m \times n$ matrix of measured sample responses at n sensors, e.g., spectral wavelengths as used in this paper, $\mathbf{b}$ designates the model vector characterizing the current matrix effects spanned by the primary samples making up the $\mathbf{y}$ and $\mathbf{X}$ arrays, and $\boldsymbol{e}$ represents random normally distributed noise with mean zero. An estimated model regression vector ($\hat{\mathbf{b}}$) can be obtained in multiple ways, such as using partial least-squares (PLS), ridge regression (RR), or principal component regression (PCR). With a model regression vector estimated, it is then used to predict new sample analyte values. The primary calibration maintains prediction accuracy as long as the new secondary conditions are similar to the original primary calibration samples (matrix matched by the span of both $\mathbf{X}$ and $\mathbf{y}$). When the secondary conditions change enough, the original $\hat{\mathbf{b}}$ is no longer effective.

Various approaches exist to adapt a model to new current secondary conditions and maintain accurate predictions.[1] In machine learning terminology, these methods would be categorized as transfer learning.[2−4] Transfer learning is a broad-based term used for dealing with situations where the primary and secondary conditions differ. The difference can be any combination of dissimilarities between spectral shape and

Received: February 7, 2021
Accepted: June 28, 2021
Published: July 8, 2021



Downloaded via IDAHO STATE UNIV on July 21, 2021 at 23:49:20 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

ACS Publications © 2021 American Chemical Society

https://doi.org/10.1021/acs.analchem.1c00578
Anal. Chem. 2021, 93, 9688−9696

location changes[5−8] for **X** differences and the analyte and interferent distributions (concentration ranges) for **Y** disparities. Transfer learning by domain adaptation is restricted to cases where the only difference between primary and secondary conditions is due to **X** spectral changes. Said in another way, only the domains of the primary and secondary **X** distributions have shifted. A key point lacking from the analytical chemistry literature is a measure to identify which situation is occurring in order to guide the analyst to the proper model updating method. While our laboratory is currently working on such a measure (not reported on here), the paper's focus is domain adaptation for model updating.

Traditional methods of domain adaptation require a few measured secondary samples with known analyte (labeled samples) to reorient $\hat{\mathbf{b}}$ in its direction and magnitude, thereby spanning the new secondary conditions. A common studied approach for domain adaptation is local mean centering (LMC)[9,10] and is further described in the following Updating Methods section. Because secondary samples with labels are used, LMC may be useful in the transfer learning situations where analyte amounts have also changed in addition to spectral **X** shifts or when only the analyte amounts have shifted, e.g., new secondary samples have lower or higher analyte amounts than the primary samples.

While LMC only requires a few reference secondary samples, model updating needs to become more practical by not requiring any secondary reference values. Advancing model updating without secondary analyte values is becoming especially important as handheld devices improve and consumer diagnostics with a smartphone becomes more possible. The time-consuming and expensive part in any calibration is obtaining analyte values (labels) for **y**. It is relatively inexpensive to quickly amass a large set of unlabeled new secondary spectra. Also, critical to advancing model updating is the ability to select accurate prediction models from the collection of models formed across the multiple model-tuning parameters commonly involved in updating processes.

An emphasis in recent works developing domain adaptation methodologies for model updating using unlabeled secondary samples has been to orthogonalize $\hat{\mathbf{b}}$ to representative spectral differences between primary and secondary conditions.[9−13] These model updating methods to-date have shown to be viable approaches but model selection is still lacking. Specifically, multiple tuning parameters are involved ranging from two[9−11] to four[12] and wavelet optimizations are required.[13] Thus, thousands of models can be made by each method. Model selection and optimization have relied on prediction errors for primary and/or secondary analyte reference samples, and hence, the methods are not yet fully free of requiring new secondary analyte reference values.

Recently, a new approach to model selection with one or more tuning parameters was developed and evaluated.[14] It is based on model diversity and prediction similarity (MDPS). The MDPS approach selects models for only the specific secondary samples being predicted.

Developed and evaluated in this paper are domain adaptation model updating methods not requiring secondary samples with reference values that may be useful for other transfer learning situations. Model selection by MDPS is shown to be effective with up to three tuning parameters.

Because the model updating methods presented involve unlabeled secondary samples without analyte reference values in conjunction with known primary analyte samples, some comments are in order regarding the literature referring to this situation as semisupervised learning. There are two basic approaches to semisupervised domain adaptation and other transfer learning methods. One process is termed inductive and the other is transductive.[3,4] In the primary/secondary context, inductive involves using a secondary sample set without analyte reference values to form models that are then used to predict other new secondary samples. Transductive refers to also using this unlabeled secondary sample set to form models, but the models are now used to predict the same secondary samples used to form the models. Thus, the transductive process probably holds an advantage in prediction accuracy over inductive approaches because the same samples used in forming the updated models are the same samples being predicted. This paper only involves the transductive use of new secondary samples without analyte values. The MDPS model selection further leverages these unlabeled secondary samples to be predicted.

With the advent of handheld spectral devices, on demand field analysis becomes more applicable with transductive approaches. Inductive updating is also useful, but as noted above, transductive maintains a small advantage. Applications range across disciplines such as updating a primary source model predicting tree pulp content for one species in a geographical area to predict pulp content in a new region (and perhaps a new related species too) out in the field negating the need to obtain samples with follow-up laboratory reference analysis. Similarly, a primary model could be updated online in a manufacturing process to predict the analyte content in a new batch.

UPDATING METHODS

Labeled Updating. The local mean centering (LMC) approach to model updating requires labeled samples (samples with analyte reference values) from the same secondary conditions as the samples to be predicted. The augmented regression equation for LMC is

$$\begin{pmatrix} \mathbf{y}_P \\ \tau \mathbf{y}_S \end{pmatrix} = \begin{pmatrix} \mathbf{X}_P \\ \tau \mathbf{X}_S \end{pmatrix} \mathbf{b} \tag{1}$$

where respective primary (P) and secondary (S) calibration **y** vectors contain reference analyte values, **X** matrices contain the corresponding measured spectra, τ denotes a scalar tuning parameter ranging from zero to infinity, which weights the augmented secondary samples, and the vectors and matrices are locally mean centered with respect to the corresponding condition. The primary array sizes are as defined earlier and the secondary $\mathbf{y}_S$ and $\mathbf{X}_S$ are $l \times 1$ and $l \times n$ for l secondary samples. The goal of LMC is to use as few secondary samples as possible to avoid a full recalibration. Listed in Table S1 of the Supporting Information (SI) is the penalty expression for LMC. The augmented equation for LMC is solved using PLS introducing latent variables (LVs) as a second tuning parameter.

Unlabeled Updating. Being able to update a primary model to new secondary conditions using unlabeled secondary spectra can substantially reduce costs and shorten analysis times. With LMC or a full recalibration, secondary samples have to be obtained, transported, and ultimately analyzed by a reference method in a laboratory setting. However, unlabeled spectra can be measured in the field.

https://doi.org/10.1021/acs.analchem.1c00578
Anal. Chem. 2021, 93, 9688−9696

A family of null augmented regression (NAR) equations can be formed using the general augmented equation:

$$\begin{pmatrix} \mathbf{y}_P \\ 0 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_P \\ \lambda \mathbf{R} \end{pmatrix} \mathbf{b} \tag{2}$$

where $\mathbf{R}$ refers to spectral differences between primary and secondary conditions.[9] The λ value is used to weight the samples as with LMC and penalize the degree of orthogonality between $\mathbf{b}$ and spectral differences. The λ values vary across the same range as τ in LMC, and models are generated by PLS requiring optimization of two tuning parameters. As previously noted, the transductive approach to unlabeled data is used, and hence, the unlabeled secondary samples forming $\mathbf{R}$ are the same samples to be predicted. The goal is to identify a model vector that is properly nulled to the domain shift of the particular prediction samples making up $\mathbf{R}$. The new predictiong sample spectra are centered to the primary samples prior to prediction.

Recently, an approach to construct $\mathbf{R}$ was presented as an NAR eigenvalue (NARE)[9] that uses centroid differences between primary and unlabeled secondary samples. Using centroid differences makes $\mathbf{R}$ a row vector $\mathbf{R} = (\boldsymbol{\mu}_P - \boldsymbol{\mu}_{SU})$ where $\boldsymbol{\mu}$ represents the corresponding subscripted mean spectra and *SU* denotes secondary unlabeled spectra.

An approach related to a method termed domain invariant PLS[11] is one that uses for $\mathbf{R}$ the difference between locally mean-centered covariance primary and unlabeled secondary spectral matrices. The method adapted in this paper is referred to as NAR-Cov1 with $\mathbf{R}$ calculated by $\mathbf{R} = \frac{1}{m_P}(\mathbf{X}_P^T\mathbf{X}_P) - \frac{1}{m_{SU}}(\mathbf{X}_{SU}^T\mathbf{X}_{SU})$, where m_P and m_{SU} indicate the number of samples in primary and secondary locally mean-centered arrays, respectively, and the superscript T symbolizes the matrix algebra transpose operation.

While local centering is effective at removing spectral biases in primary and secondary samples as with LMC, local centering to form covariance matrices could remove key spectroscopic information needed to improve the model orthogonalization against spectral differences. In conjunction with NAR-Cov1, no centering (NAR-Cov2) is also evaluated. Because there is no centering, $\mathbf{R}$ is now a difference of outer product matrices, but the Cov2 is used for consistency.

Discussed in the Supporting Information with Table S1 are other NAR approaches. One method is related to a linear joint trained framework for model updating that combines both centroid and covariate shifts requiring four tuning parameters.[12] A similar mixed approach using three tuning parameters is noted as NARE-Cov in Table S1.

Presented are results of MDPS selected models for the unlabeled methods NARE and NAR-Cov1 and -Cov2 in comparison to the LMC labeled method. Four NIR datasets are studied. The result's focus is relative to prediction errors.

Hybrid Updating by Adding Labeled to Unlabeled Samples. Results are presented for a new hybrid version that couples NARE with LMC. In addition to augmenting the primary reference samples with the same unlabeled samples to be predicted later with NARE, the hybrid approach also includes a few labeled secondary samples. The method is referred to as NARE-LMC. It is later shown that fewer labeled samples are needed compared to LMC alone. The specific NARE-LMC hybrid augmented equation becomes

$$\begin{pmatrix} \mathbf{y}_P \\ 0 \\ \tau \mathbf{y}_S \end{pmatrix} = \begin{pmatrix} \mathbf{X}_P \\ \lambda(\boldsymbol{\mu}_P - \boldsymbol{\mu}_{SU}) \\ \tau \mathbf{X}_S \end{pmatrix} \mathbf{b} \tag{3}$$

Estimation of $\mathbf{b}$ is by PLS and three tuning parameters that need to be optimized. In the hybrid appraoch, the new predcition samples are centered to the labeled secondary sample mean used in the eq. 4. Other NAR hybrid models can be formed as noted in Table S1.

MODEL SELECTION

It was recently shown that the degree of difference between primary and secondary $\mathbf{X}$ domains dictates whether primary or secondary prediction errors can be used to select updated models.[15] If the differences are small, then primary prediction errors can be used to select models. If the difference is greater, selecting models relative to primary prediction errors produces unsatisfactory results requiring some secondary reference samples for proper model selection. However, for transductive unlabeled model updating methods, model selection cannot include secondary reference values.

Used here is a recently developed consensus model selection approach that leverages model diversity with prediction similarity (MDPS)[14] to identify appropriate models. The MDPS protocol does not use secondary analyte reference values in selecting models, and hence, MDPS is compatible with unlabeled secondary situations. Additionally, because models are selected for unlabeled secondary spectra, massive amounts of unlabeled data can be included to better span the secondary domain that may improve prediction accuracy depending on the degree of secondary matrix effects.

The consensus modeling concept behind MDPS is that if two models are sufficiently different from one another but generate similar predictions, then these two models may contain robust predictions that can be extrapolated to imply accurate predictions. The MDPS method includes a measure to guard against selecting over- and under-fitted models with similar predictions.

As shown in the following, the final MDPS selected models for LMC and all NAR methods are those explicitly selected to predict a specific set of new secondary samples. Additionally, these new secondary samples are the same samples used to form the NAR models. Thus, the NAR methods maintain an advantage over LMC. However, LMC can leverage the augmented secondary analyte reference values to presumably form model vectors better characterizing the linear relationship between analyte content and spectral responses. However, LMC requires laboratory reference analysis of some representative secondary samples. The hybrid approach in eq 3 has the NAR and LMC advantages and the LMC secondary disadvantage. Regardless of the model updating method, because models are specifically selected to predict a particular new sample set, there is a small bias (overfit) toward the new prediction samples. However, the models formed and selected are only meant to predict these particular samples. If new samples require prediction, then new models should be formed and selected. Because models are formed and selected on demand to directly predict target samples, a complex cross-validation is not needed.

Model Diversity and Prediction Similarity (MDPS). The MDPS method is briefly described and further details are provided in ref 14. After the entire model set is generated over

https://doi.org/10.1021/acs.analchem.1c00578
Anal. Chem. 2021, 93, 9688−9696

a range of tuning parameter values, all possible combinations of two models are compared for diversity and prediction similarities. The cosine of the angle between two models ($\cos(\theta)$) is used to assess model diversity with values ranging from 1, indicating total similarity, to 0, signifying complete orthogonality (complete dissimilarity). The composite prediction similarity for the ith and jth models is expressed by

$$C_{i,j} = \mathrm{SPD}_{i,j}^{\mathrm{RS}} + \omega(\overline{\mathrm{RMSEC}}_{i,j}^{\mathrm{RS}} + \overline{\|\hat{\mathbf{b}}\|}_{i,j}^{\mathrm{RS}}) \quad (4)$$

where $\mathrm{SPD}_{i,j} = \sum_{k=1}^{t} |\hat{y}_{k,i} - \hat{y}_{k,j}|$ is the secondary prediction differences (SPD) for all t unlabeled target samples to be predicted (for the NAR methods, these would be the same samples used to form $\boldsymbol{R}$), $\hat{y}$ is the respective model predictions, $\overline{\mathrm{RMSEC}}_{i,j}$ corresponds to the average of the two primary root-mean-square error of predictions for the two models being compared, $\overline{\|\hat{\mathbf{b}}\|}_{i,j}$ signifies the mean vector 2-norm, superscript RS denotes range-scaled values between 0 and 1, inclusive, and ω (≥ 0) weights the bias-variance trade-off determined by the U-curve formed using the respective bias and variance terms in the parenthesis. This trade-off is captured by the 2-norm avoiding over-fitted models and the primary prediction error shielding against under-fitting.

In a recent work, it was shown that using model diversity thresholds of $0.3 < \cos(\theta) < 0.5$ in combination with $\omega = 0.4$ work well for MDPS to select accurately predicting models with LMC.[14] The weighted mean predictions from the lower 10% of the models in the model diversity range are then used as the single prediction value for each secondary sample, i.e., each model prediction of a secondary sample in the modeling updating method is weighted by the frequency in which the model lies in the lower 10%.

The MDPS concept with two ω values is shown in Figure 1 for the Goat data using NARE where the red box highlights all

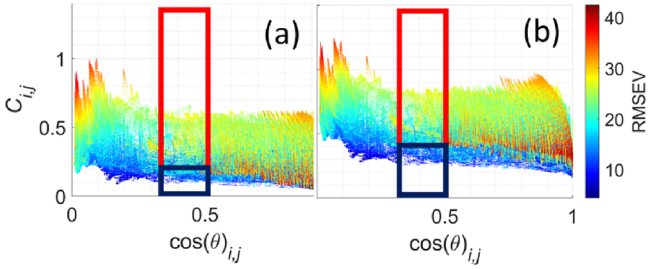


Figure 1. Goat dataset scatter plot of composite prediction similarities against model diversities for all possible NARE model combinations. Points are color coded according to average RMSEV values between models. Panel (a) shows $\omega = 0$ and (b) has $\omega = 0.4$. The red box indicates the model diversity threshold of $0.3 < \cos(\theta) < 0.5$, and the dark blue box shows the lowest 10% of models selected in the diversity region.

model pairs in the model diversity range $0.3 < \cos(\theta) < 0.5$. A zoomed view of the highlighted model diversity range is provided in Figure S1 of the Supporting Information. The dark blue box highlights those models selected in the lower 10% range of the red box. Model pairs are color coded to the RMSE validation (RMSEV) values for the secondary samples. From Figure 1b with $\omega = 0.4$, it is observed that by increasing ω from 0 in Figure 1a, the model selection region improves. Specifically, poor predicting models corresponding to the RMSEV color-coded points light blue and red are pushed out of the selection region with $\omega = 0.4$. Another characterization of this effect is shown in ref 14.

While models with minimum RMSEV values occur in the $\cos(\theta)$ region around 0.75 in Figure 1, the diversity range $0.3 < \cos(\theta) < 0.5$ is better suited. This range was previously determined suitable[14] and it was again an effective compromise functioning across the many datasets and modeling methods in this study. Specifically, MDPS generally selects models in this range with RMSEV values below or at the first quartile of all potential models. In Figure 1b, there are 2100 Goat models creating 2,203,950 model pair combinations with 285,374 model pairs in the red box and 28,537 model pairs selected in the dark blue box for a total of 825 unique models selected by MDPS with low RMSEV values. Heatmaps of the RMSEV and selected model histogram are shown in the Supporting Information demonstrating that the most frequently selected models do maintain low RMSEV values.

Another point to consider when setting $\cos(\theta)$ model diversity value measures and ω in the prediction similarity measure (balancing the amount of under- and over-fitting) is the number of tuning parameters involved in the calibration process. Generally, the more tuning parameters to be determined, the larger the size of the model space spanning all the generated models and the greater the model diversity existing in this model space, e.g., a PLS set of models for a primary calibration with one tuning parameter (number of LVs) compared to the NAR family of models with two to three tuning parameters. These points are further discussed in ref 14 in the framework of the Rashomon effect where there is not one best accurate predicting model, but a collection of models exists with similar accurate analyte predictions.[16,17] Additionally, the degree of domain difference between the primary and secondary conditions and the density of acceptable models across the tuning parameter values may affect the $\cos(\theta)$ and ω values.

EXPERIMENTAL SECTION

Software. All algorithms were developed by the authors using MATLAB R2019b. An NAR suite of algorithms and the MDPS model selection algorithm can be downloaded.[18] The code can be easily altered to form other NAR methods.

Data Descriptions. Four NIR datasets were studied: Corn,[19] Soy,[20] Goat,[21] and Tablet.[22] Considering all possible combinations of primary and secondary conditions, 40 model updating situations are produced.

Corn. The same 80 cornmeal samples were measured across three instruments, m5, mp5, and mp6, over 1100−2498 nm at 2 nm increments for 700 wavelengths. Analyte prediction properties are moisture (9.377−10.993%), oil (3.088−3.832%), protein (7.654−9.711%), and starch (62.826−66.472%). All 24 possible updating conditions were analyzed covering the four analytes across each of the six instrument updating combinations. Spectra and a principal component (PC) score plot are shown in Figure S2.

Soy. The same 60 soy seed samples with moisture (5.9−18.4%), oil (29.0−43.4%), and protein (14.7−22.9%) analyte reference values were measured on instruments R1 and R2 from 1100−2500 nm with 4 nm increments for a total of 300 wavelengths. All six updating scenarios were studied between R1 and R2 for each of the three analytes. Spectra and a PC score plot are presented in Figure S3.

Goat. Feces goat samples were analyzed for juniper berry content in 1999 (61 samples) and 2002 (48 samples). Samples

https://doi.org/10.1021/acs.analchem.1c00578
Anal. Chem. 2021, 93, 9688−9696

were measured from 400−2500 nm at even wavelengths (1050 total wavelengths). For this study, 1999 and 2002 correspond respectively to primary and secondary conditions for one updating situation. Spectra, a PC score plot, and histograms of the **y** juniper berry analyte content are displayed in Figure S4. The histograms show similar distributions between 1999 and 2002.

Tablet. Pharmaceutical tablets were produced with an active pharmaceutical ingredient (API) escitolopram in four nominal tablet weight categories (types 1, 2, 3, and 4) with respective total tablet weights of 90, 125, 188, and 250 mg. The different total weights make unique sizes with respective tablet thicknesses ranging from 2.9 to 4.3 mm. Tablets were produced in two settings: laboratory for primary and full for secondary batches. There are 30 tablets for each batch of tablet type making 120 tablets for each batch. The spectra plotted in Figure S5 were measured from 7400−10,500 cm^{-1} for a total of 404 wavelengths. The corresponding PC score plots are also in Figure S5. Histograms of the API analyte content for primary and secondary batches presented in Figure S5 characterize small differences in the **y** distributions and the full API content are essentially spanned by the lab samples. Past studies have shown that updating is best when tablet type 1 samples are always involved. Including type 1 produces 9 possible updating situations with two tablet types in each of the primary and secondary batches: 1&2−1&2, 1&2−1&3, 1&2−1&4, 1&3−1&2, 1&3−1&3, 1&3−1&4, 1&4−1&2, 1&4−1&3, and 1&4−1&4, with the first condition listed being primary.

Parameter Values for Model Formation and Selection. The number of LVs for PLS ranged from 1 through the mathematical rank of each primary spectral matrix $\mathbf{X}_P$. There are 50 values each for the τ and λ tuning parameters except for NARE-LMC with 30 values each to reduce computation time. All values exponentially decrease from the highest to the lowest singular values of each $\mathbf{X}_P$ spectral matrix.

As previously noted, only models in the diversity range $0.3 < \cos(\theta) < 0.5$ at $\omega = 0.4$ were evaluated for possible model selection. This diversity range and the ω value were empirically determined optimal in a previous work involving two tuning parameters and small deviations from these values are acceptable.[14]

Data Splitting for Validation. In order to evaluate model updating prediction accuracies in conjunction with model selection by MDPS, 100 random sample splits were used. The sample division sizes for primary source calibration (PRI), augmented secondary calibration (SCAL) for LMC and NARE-LMC, and secondary validation (SVAL) are shown in Table S2.

Mean RMSEV and R^2 values (from plotting predicted analyte validation values against reference values) across the 100 random splits are reported. Note that for the Corn and Soy datasets, the same samples were measured in both primary and secondary conditions. Thus, care is taken to ensure the same samples do not appear in both primary and secondary sets on each random split.

To expand the number of samples forming **R** for NARE and NAR-Cov1 and -Cov2, the SCAL samples are appended to SVAL (removing reference analyte labels). This expanded set is also used in MDPS. However, the SCAL samples are not predicted for a fair comparison of respective SVAL RMSEV and R^2 values across all updating methods. For hybrid NARE-LMC, only SVAL was used to form **R**.

Benchmarks. Three baseline prediction errors are needed to evaluate and compare model updating methods. For all three baselines, PLS is used.

The first baseline is primary predicting secondary (PPS), which assesses the situation of the primary calibration model being used to predict the new secondary validation sample set SVAL in Table S2. The PPS models are PLS calibrations of the primary sample set PRI in Table S2. The PPS RMSEV and R^2 values show the necessity for model updating. Generally, the greater the matrix is affected by the differences between primary and secondary domains, PPS is expected to perform worse.

The next baseline is secondary predicting secondary (SPS) to characterize the instance when the time and expense is taken for a full recalibration. Thus, model updating methods should ideally be competitive with SPS. To ensure representative SPS models are used, each full secondary dataset is randomly split 100 times with 60% of the samples used for calibration and 40% for the validation.

The third baseline is the small secondary predicting secondary (SSPS). The SSPS result needs to show that using the small SCAL set is not practical for predicting the larger SVAL sample set and some primary data is needed. This baseline is only used with LMC.

Shown in figures are the boxplot trends of RMSEV and R^2 values at the minima and first two quartiles for PPS, SPS, and SSPS models.

RESULTS AND DISCUSSION

It was recently shown that if quartile boxplots of model quality measures, such as RMSEV and R^2 values, over a series of data splits are used to evaluate and compare tuning parameter-based methods, then it becomes critical to exclude from consideration those tuning parameter values where models have essentially converged.[14] For example, depending on the range of tuning parameter values, an excessive number of under- and over-fitted models can be formed. These subsets of respective models typically predict similarly, and hence, quartile boxplots of compiled model quality measures including these models misrepresent the methods being evaluated.

To objectively compare labeled and unlabeled updating methods using quartile boxplots over the 100 random splits, the same approach used previously for LMC[14] is used in this paper to identify tuning parameter regions that span nonconverged tuning parameter values. This region is referred to as the active bias-variance trade-off zone and covers the tuning parameter transition region where variation between models is important. Tuning parameter values beyond the active bias-variance zone are deemed over- or under-fitted and need to be removed before making quartile boxplots of model quality measures. The Supporting Information contains information on how the tuning parameter active bias-variance trade-off zones are identified with up to three tuning parameters, as well as showing the misrepresentation.

All results shown and discussed are based on only those models selected from the active bias-variance trade-off zones. However, it is important to note that MDPS model selection does not necessarily require a prior determination of tuning parameter active bias-variance trade-off zones. Convergences are determined for a fair comparison of method-specific boxplots. Results presented in the Supporting Information

https://doi.org/10.1021/acs.analchem.1c00578
Anal. Chem. 2021, 93, 9688−9696

demonstrate this fact and it was also shown in the previous MDPS paper.[14]

Number of Unlabeled Samples. A crucial issue with transductive model updating is model performance as the number of unlabeled secondary samples increases (or decreases). Such an analysis was undertaken with the Goat results for the 100 random splits presented in Figure 2. Shown

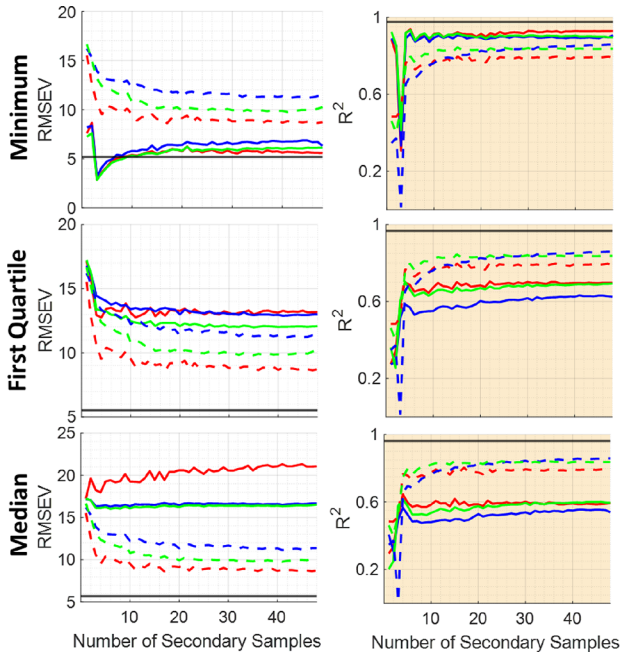


Figure 2. Goat mean minima, first quartiles, and medians of RMSEV and R^2 values for the following situations in the order: SPS (black solid line), NARE for all generated models (red solid line), $NARE_{MDPS}$ for selected models (red dash line), NAR-Cov1 (blue solid line), NAR-$Cov1_{MDPS}$ (blue dash line), NAR-Cov2 (green solid line), and NAR-$Cov2_{MDPS}$ (green dash).

are the mean minima, first quartiles, and medians (second quartiles) for baseline SPS, NARE, NAR-Cov1, and NAR-Cov2. Also, shown are the mean minima and quartiles associated with the MDPS selected models. Plots in Figure 2 are typical and another data situation is plotted in Figure S13.

Regardless of the model updating method, Figure 2 shows that a minimum number of samples is needed for $\mathbf{X}_{SU}$. For this dataset and others, it is observed that after five unlabeled secondary samples are augmented to the 61 labeled primary samples, the RMSEV values substantially decrease compared to augmenting with only one sample. The RMSEV values gradually further decrease up to 10 samples. The R^2 values level off after five samples are used and predicted. From Figure 2, it is also observed that while the RMSEV first quartiles and medians of the models evaluated tend to decrease and converge to single values after a few unlabeled samples are added, the minima decrease and then begin to rise after more than three unlabeled samples. The lowest minimum occurs at three unlabeled samples then rises again due to the over-fitted nature of minima RMSEV values with few samples. As additional samples are added, the minima RMSEV values increase because it becomes difficult for a model to closely fit the noise for all samples compared to when fewer samples are used in $\mathbf{X}_{SU}$, and hence, these models can be considered more robust. The MDPS selected models do not have this problem indicating that MDPS is not selecting the over-fitted models. The first quartiles and medians of evaluated models both continue decreasing as more unlabeled secondary samples are included due to the over-fitted models no longer being considered as with the minima trends.

Intrinsically, changing the number of secondary samples affects both the model updating and selection processes. Very few secondary samples and **R** is overfitted to the primary and secondary differences of the specific secondary samples and cannot effectively encode these differences. The model diversity requirement used in MDPS is set to only select robust models. Therefore, models at minima RMSEV values are too over-fitted and are not selected. In contrast, when many secondary samples are used, **R** is no longer over-fitted to the noise of the secondary samples, and therefore its models can be labeled as more robust. These models are now accessible to the MDPS diversity criterion since the robusticity requirement is now satisfied.

Results in Figure 2 do not fully answer the question on how many specific samples are needed for $\mathbf{X}_{SU}$ in order to effectively update a model. This number will most likely depend on several factors. One is the degree of difference between the primary and secondary domains in terms of the total matrix matching relative to spectral effects and analyte values. Said another way, the number of samples will depend on if the task is domain adaptation or transfer learning. Another factor is the degree of intrasimilarity between the samples used in $\mathbf{X}_{SU}$. Both of these factors will also probably affect how well MDPS works. Results presented in Figure 2 and Figure S12 indicate that large numbers of samples are not needed for the data sets studied. The number of samples listed in Table S1 are few and provide acceptable results.

Method Comparisons. It is impossible to show results for all 40 updating situations. Thus, a few characteristic results are presented in Figure 3 for each of the four main datasets. There is only one Goat dataset updating situation and numerous cases for the other three datasets. Thus, Figure 3 covers the Goat dataset and one model updating situation for the other three datasets (see the Figure 3 caption). Results for other updating cases are shown the Supporting Information.

Figure 3 has results for the baseline SPS, labeled LMC, unlabeled NARE, NAR-Cov1 and NAR-Cov2, and hybrid NARE-LMC. The boxplots in Figure S16 duplicate those in Figure 3 but also include the baselines PPS and SSPS. As expected, PPS and SSPS in Figure S16 perform poorly even at the generally over-fitted minima models, demonstrating the need to update the primary model. The small secondary reference set SSPS is also not able to predict the secondary samples and assistance from labeled primary samples is used with LMC.

From Figure 3, it can also be observed that the unlabeled model updating methods NARE, NAR-Cov1, and NAR-Cov2 essentially perform equivalently for three of the datasets (Figure 3a,c,d). More importantly, the methods generally perform as well as or better than LMC that requires secondary reference values.

For the three datasets where the analyte distributions are well matched, NAR-Cov1 and NAR-Cov2 provide acceptable results. Conversely, for Tablet 1&4−1&3 in Figure 3b, the primary and secondary datasets are not well matched in terms of **X** in combination with small differences in analyte ranges (see Figure S5). In this circumstance, the NAR-Cov methods do not perform as well, where NAR-Cov1 with local centering

https://doi.org/10.1021/acs.analchem.1c00578
Anal. Chem. 2021, 93, 9688−9696

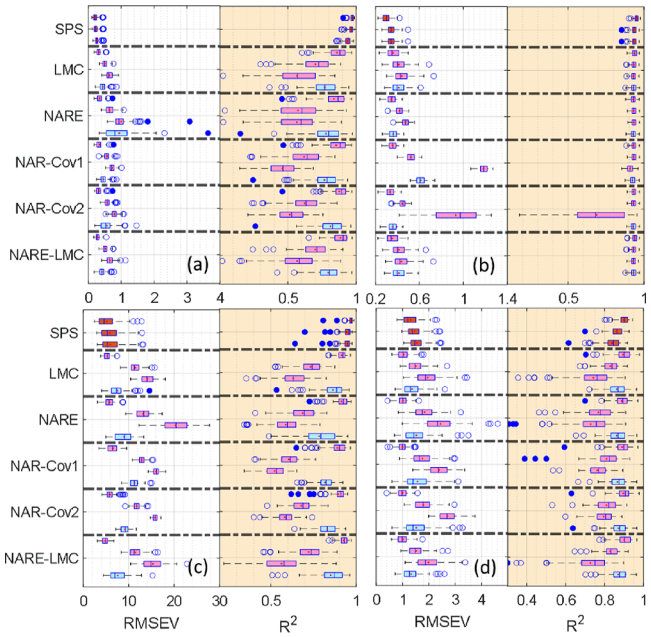


Figure 3. Boxplots comparing model updating and a baseline method. Shown are the RMSEV minima, first quartiles, and medians for models evaluated and the MDPS models selected from the evaluated models for datasets (a) Corn mp6−m5 starch, (b) Tablet 1&4−1&3, (c) Goat 99−02, and (d) Soy R1 and R2 moisture.

performs worse than NAR-Cov2 with no centering. Plots in Figure S5 of spectra and PC scores show that within each primary and secondary conditions there are unique spectral differences. Having categorical situations expressed in outer product arrays (multimodal) in combination with small differences in analyte ranges makes it more difficult to form accurate predicting models that are simultaneously nulled (orthogonal) to the matrices of categorical differences. Thus, too much predictive analyte information is probably lost with the NAR-Cov null penalty in this situation. For NARE, models only need to be orthogonal to differences between mean vectors. These trends across the four datasets in Figure 3 are similar to those observed with other dataset situations shown in Figure S17.

The hybrid NARE-LMC approach provides small improvements relative to LMC and NARE alone (Figure 3 and Figures S14 and S15). However, the small improvement from NARE-LMC over LMC indicates that by including unlabeled samples, fewer labeled samples are needed with NARE-LMC.

Prediction errors can be further reduced by optimizing the effective prediction region of $\mathbf{X}_{SU}$ such that the only samples predicted are those $\mathbf{X}_{SU}$ samples spectrally bracketed (spanned) by other $\mathbf{X}_{SU}$ samples. Such potential prediction samples can be detected by using a Kennard−Stone[23] sample split on $\mathbf{X}_{SU}$. A Kennard−Stone study based on the Euclidean distance was performed (not shown) where all of $\mathbf{X}_{SU}$ is used to form models, but only the innermost spectrally bracketed $\mathbf{X}_{SU}$ samples are used in MDPS to select models and be predicted. Prediction errors are always noticeably reduced for the most centroid $\mathbf{X}_{SU}$ samples. As additional samples are included in MDPS and predicted moving out from the centroid, prediction errors increase converging to values reported using the full $\mathbf{X}_{SU}$ in MDPS with corresponding predictions. Thus, if experimental design constraints are possible, results from the NAR family of model updating methods can be improved compared to predicting all samples in an $\mathbf{X}_{SU}$. It may be possible that if only one new sample prediction is desired, e.g., one sampling site location, then a collection of spectra could be measured in close proximity to form $\mathbf{R}$ and only the Kennard−Stone most centroid spectrum is predicted for the sample analyte value. This scheme assumes that some secondary spectral variance exists across the sampling sites.

It has been suggested to fully assess the success of a calibration transfer method; the relationship of the new samples to the updated model space should be evaluated with prediction outlier diagnostic.[24] However, these studies were not performed.

Histograms of Selected Models. Instead of the usual approach of selecting one model to predict new samples, the MDPS approach selects a collection of models ideally located around the minimum RMSEV value for a consensus prediction. The mean RMSEV heatmap across the 100 data splits is shown in Figure 4 in conjunction with the histogram

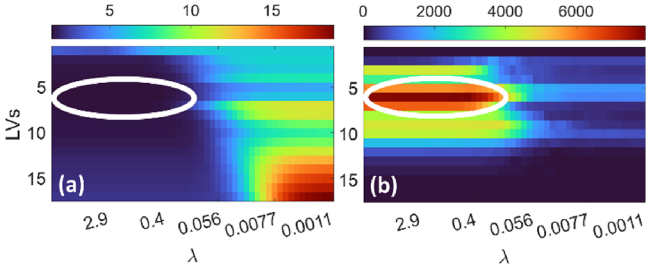


Figure 4. NARE heatmaps of (a) mean Soy R1-R2 moisture RMSEV values across the 100 outer data splits and (b) histogram of models selected by MDPS. Areas with the lowest RMSEV values and matched models on the histogram are circled.

heatmap of MDPS selected models for NARE using the Soy R1-R2 updating situation. From the heatmaps, it is observed that the most frequently selected models are indeed the same models with lower RMSEV values.

Displayed in Figure 5 are similar heatmaps for NAR-Cov2 using the same updating situation. It is again observed that

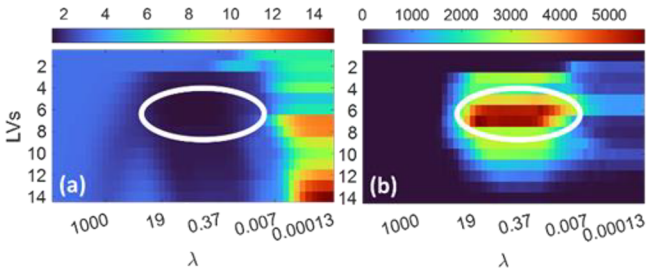


Figure 5. NAR-Cov2 heatmaps of (a) mean Soy R1-R2 moisture RMSEV values across the 100 data splits and (b) histogram of models selected by MDPS. Areas with the lowest RMSEV values and matched models on the histogram are circled.

MDPS mostly selects models correlated to low RMSEV values. Histogram and RMSEV heatmaps graphed for additional datasets in the Supporting Information including NAR-Cov1 show that these observations are general. Even in the distinctive Tablet situations where the NAR-Cov methods do not perform as well, MDPS largely selects models associated with lower RMSEV values.

https://doi.org/10.1021/acs.analchem.1c00578
Anal. Chem. 2021, 93, 9688−9696

The MDPS approach is also able to select acceptable models for NARE-LMC with three tuning parameters. For example, shown in Figure 6 is a heatmap of the RMSEV values and

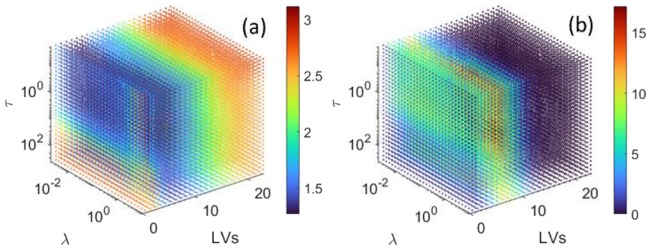


Figure 6. NARE-LMC heatmaps of (a) mean Soy R1-R2 moisture RMSEV values across the 100 data splits and (b) histogram of MDPS selected models. Log ruler for τ and λ.

corresponding histogram of selected models for the Soy R1-R2 moisture case using NARE-LMC. The most frequently selected models are the models with the lowest RMSEV values. Parallel results are shown in the Supporting Information for other datasets.

CONCLUSIONS

Results presented further confirm that model updating with unlabeled secondary data is efficient and accurate. The foundation of the unlabeled model updating methods presented can be extended to a family of updating methods with different penalty terms (some of which are noted in the Supporting Information). However, model selection for such methods has been missing, specifically selecting acceptable models to predict secondary samples without corresponding reference values. If secondary conditions are only moderately different than the primary conditions, prediction errors for the primary samples with reference values can possibly be used. However, without a measure to indicate the degree of difference, such an approach should only be used with caution. The MDPS process selects models specifically targeted to the new prediction samples. The efficacy of the MDPS approach is shown to not only be robust across datasets but also robust throughout each of the model updating methods. Predictions from the MDPS selected models are consistently at or below the first quartile of all models formed.

As noted, what is needed is a measure to characterize the degree of respective **X** and **y** differences between primary with reference values (labeled primary) and secondary samples without references (unlabeled secondary) in order to ascertain which situation is present to decide on an appropriate updating scheme. Our laboratory is currently working on a measure in conjunction with evaluating LMC, NARE, and NARE-LMC for the different possible scenarios. It is expected that as the degree of analyte content distributions differ, the more a method using some reference values such as LMC or NARE-LMC will be needed.

It was also discussed that if applicable, the unlabeled secondary samples used in the model updating can be Kennard−Stone sorted. Reduced predication errors are obtained for the centermost samples.

Lastly, it is appealing to be able to interpret the final model weight values on labeled or unlabeled secondary samples used in forming models. However, there are several factors that go into the magnitude of these tuning parameter values. The key is the degree of difference between the spectral domains and how similar the analyte amounts and other spectral responding interferences are to the referenced primary calibration samples, i.e., how well the secondary samples are matrix matched by both **X** and **y**. Also important are the calibration sample density and respective error structures,[23] number of primary samples, and the number of secondary samples used in the model updating algorithm. The inability to interpret weight values is much like the situation of the inability to interpret model regression vectors due to the many model regression vectors that can accurately predict the same samples.[25−28]

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.1c00578.

> Other NAR model updating methods are listed and briefly discussed, included is a discussion of the importance of removing converged models to avoid quartile boxplots of model quality measures misrepresenting actual results, the procedure for automatic determination of converged models is presented, and additional dataset results are shown (PDF)

AUTHOR INFORMATION

Corresponding Author

John H. Kalivas − *Department of Chemistry, Idaho State University, Pocatello, Idaho 83209, United States;* orcid.org/0000-0001-7056-976X; Email: kalijohn@isu.edu

Author

Robert C. Spiers − *Department of Chemistry, Idaho State University, Pocatello, Idaho 83209, United States;* orcid.org/0000-0001-8947-7080

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.1c00578

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This material is based upon the work supported by the National Science Foundation under grant nos. CHE-1506417 (co-funded by CDS&E) and CHE-1904166 (co-funded by CDS&E and the Office of Investigative and Forensic Sciences in the National Institute of Justice) and is gratefully acknowledged by the authors. The authors are thankful to Erik Andries for his insightful discussions.

REFERENCES

(1) Brown, S. D. *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, 2nd Ed.; Brown, S.; Tauler, R.; Walczak, B. Eds.; Elsevier: Amsterdam, The Netherlands, 2020; Vol. 3, pp. 359−391.

(2) Kouw, W.M.; Loog, M. *An Introduction to Domain Adaptation and Transfer Learning*; Cornell Univeristy: Technical Report, arXiv:1812.11806v2, Jan. 14, 2019; https://arxiv.org/abs/1812.11806v2

(3) Arnold, A.; Nallapati, R.; Cohen, W. W. *Workshops Proceedings of the 7th {IEEE} International Conference on Data Mining {(ICDM)}, October 28−31, 2007, Omaha, Nebraska*, Tung, A. K. H.; Zhu, Q.; Ramakrishnan, N.; Zaïane, O. R.; Shi, Y.; Clifton, C. W.; Wu, X. Eds.;

IEEE Computer Society: 2007; pp. 77−82, DOI: 10.1109/ICDMW.2007.109
(4) Pan, S. J.; Yang, Q. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345−1359.
(5) Jouan-Rimbaud, D.; Massart, D. L.; Saby, C. A.; Puel, C. *Anal. Chim. Acta* **1997**, *350*, 149−161.
(6) Jouan-Rimbaud, D.; Massart, D. L.; Saby, C. A.; Puel, C. *Chemom. Intell. Lab. Syst.* **1998**, *40*, 129−144.
(7) Daszykowski, M.; Walczak, B.; Massart, D. L. *Anal. Chim. Acta* **2002**, *468*, 91−103.
(8) Smilde, A. K.; Timmerman, M. E.; Saccenti, E.; Jansen, J. J.; Hoefsloot, H. C. J. *J. Chemom.* **2015**, *29*, 277−288.
(9) Andries, E.; Kalivas, J. H.; Gurung, A. *J. Chemom.* **2019**, *33*, 1−20.
(10) Kalivas, J. H.; Siano, G.; Andries, E.; Goicoechea, H. *Appl. Spectrosc.* **2009**, *63*, 800−809.
(11) Nikzad-Langerodi, R.; Zellinger, W.; Lughofer, E.; Saminger-Platz, S. *Anal. Chem.* **2018**, *90*, 6693−6701.
(12) Larsen, J. S.; Clemmensen, L.; Stockmarr, A.; Skov, T.; Larsen, A.; Ersbøll, B. K. *J. Chemom.* **2020**, *34*, 1−22.
(13) Poerio, D. V.; Brown, S. D. *Appl. Spectrosc.* **2018**, *72*, 378−391.
(14) Spiers, R. C.; Kalivas, J. H. *J. Chem. Inf. Model.* **2021**, *61*, 2220−2230.
(15) Gurung, A.; Kalivas, J. H. *J. Chemom.* **2020**, *34*, 1−12.
(16) Semenova, L.; Rudin, C.; Parr, R. A Study in Rashomon Curves and Volumes: A New Perspective on Generalization and Model Simplicity in Machine Learning. *arXiv:1908.01755v2 [cs.LG]*, downloaded Feb. 2021.
(17) Breiman, L. *Stat. Sci.* **2001**, *16*, 199−231.
(18) *NAR and MDPS software link* : https://www.isu.edu/chem/faculty/staffdirectoryentries/kalivas-john.html (accessed on July 1, 2021).
(19) Wise, B. M.; Gallagher, N. B. *Eigenvector Research*, Manson, WA. http://www.eigenvector.com/data/index.htm
(20) Bouveresse, E.; Hartmann, C.; Massart, D. L.; Last, I. R.; Prebble, K. A. *Anal. Chem.* **1996**, *68*, 982−990.
(21) Walker, J. W.; Campbell, E. S.; Lupton, C. J.; Taylor, C. A., Jr.; Waldron, D. F.; Landau, S. Y. *J. Anim. Sci.* **2007**, *85*, 518−526.
(22) Dyrby, M.; Engelsen, S. B.; Nørgaard, L.; Bruhn, M.; Lundsberg-Nielsen, L. *Appl. Spectrosc.* **2002**, *56*, 579−585.
(23) Kennard, W.; Stone, L. A. *Technometrics* **1969**, *11*, 137−148.
(24) Guenard, R. D.; Wehlburg, C. M.; Pell, R. J.; Haaland, D. M. *Appl. Spectrosc.* **2007**, *61*, 747−754.
(25) Brown, C. D. *Anal. Chem.* **2004**, *76*, 4364−4373.
(26) Brown, C. D.; Green, R. L. *TrAC, Trends Anal. Chem.* **2009**, *28*, 506−514.
(27) Kunz, M. R.; Ottaway, J.; Kalivas, J. H.; Andries, E. *J. Chemom.* **2010**, *24*, 218−229.
(28) Kalivas, J. H.; Ferré, J.; Tencate, A. J. *J. Chemom.* **2017**, *31*, e2925.

https://doi.org/10.1021/acs.analchem.1c00578
Anal. Chem. 2021, 93, 9688−9696