analytical chemistry

pubs.acs.org/ac Article

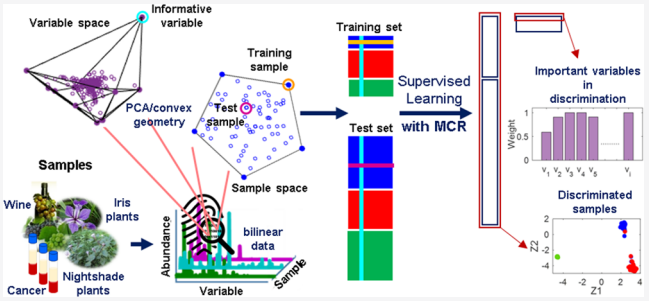
Workflow for the Supervised Learning of Chemical Data: Efficient Data Reduction-Multivariate Curve Resolution (EDR-MCR)

Samira Beyramysoltan, Hamid Abdollahi,* and Rabi A. Musah*

Cite This: *Anal. Chem.* 2021, 93, 5020−5027 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: A new method termed efficient data reduction-multivariate curve resolution (EDR-MCR) has been devised for classification of high-dimensional data. The method introduces the coupling of EDR and MCR as a new strategy for data splitting, variable selection, and supervised classification of high dimensionality data. The method reduces data dimensionality and selects the training set using principal component analysis (PCA) and convex geometry prior to data classification. Then, the reduced data are categorized using an MCR model, in which numerical constraints are imposed to resolve the data into classes and readily interpretable pure component signal weights. The performance of the EDR and supervised MCR methods were tested for their ability to enable discrimination between the constituents of two benchmark and two high-dimensional data sets. The results were compared with the output of the application of different data splitting methods including iterative random selection (IRS), Kennard−Stone (KS), and discrimination methods including partial least-squares-discriminant analysis (PLS-DA) and the ensemble-learning frameworks of linear discriminant analysis (LDA), k-nearest neighbors (KNN), classification and regression trees (CART), and support vector machine (SVM). Overall, EDR resulted in comparable results with other data splitting methods despite the small size of the training set samples that it created. The proposed MCR approach, in comparison with other commonly used supervised techniques, has the advantages of speed in implementation, tuning of fewer parameters, flexibility in the analysis of data characterized by low sample numbers and class imbalances, improved accuracy from the inclusion of additional system information in the form of numerical constraints, and the ability to resolve pure components signal weights.



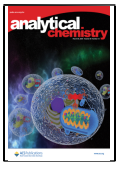
INTRODUCTION

A prevailing challenge in the ability to accurately classify and/or draw inferences from chemical data that enable prediction of trends and outcomes remains the determination of the most straightforward and accurate approach to accomplish the task.[1] A hallmark of this well-established field is the immensity of the range of algorithms that have been developed for this purpose. This reflects the truth of Wolpert's "no free lunch theorem",[2] in that there is no single approach that can be used to solve a broad range of classification problems.

In general, the development process for the establishment of a multivariate data analysis workflow that will accomplish classification and/or prediction for a given type of data is composed of two fundamental components: (1) data reduction methods (such as variable selection[3−5]) in which decisions are made about the subset of the full data set that is most important for revealing class distinctions and (2) splitting approaches which involve dividing the data into training and test sets and which can introduce biases into the model that influence the final result.[6] The process of determining the most appropriate data reduction and least biased splitting methods is critical and hinges on conclusions that are drawn from the results of supervised learning.[3−5,7,8] There is a copious body of literature that is solely devoted to the development of variable selection methods,[3−5] and numerous techniques have been developed for assessment of model accuracy. Some of the most widely used sampling approaches are Kennard−Stone, random selection, and stratified sampling.[7−14] The choice of the representative sampling method hinges on data set characters and on the user-defined proportion of samples (with different properties) within each split.[14]

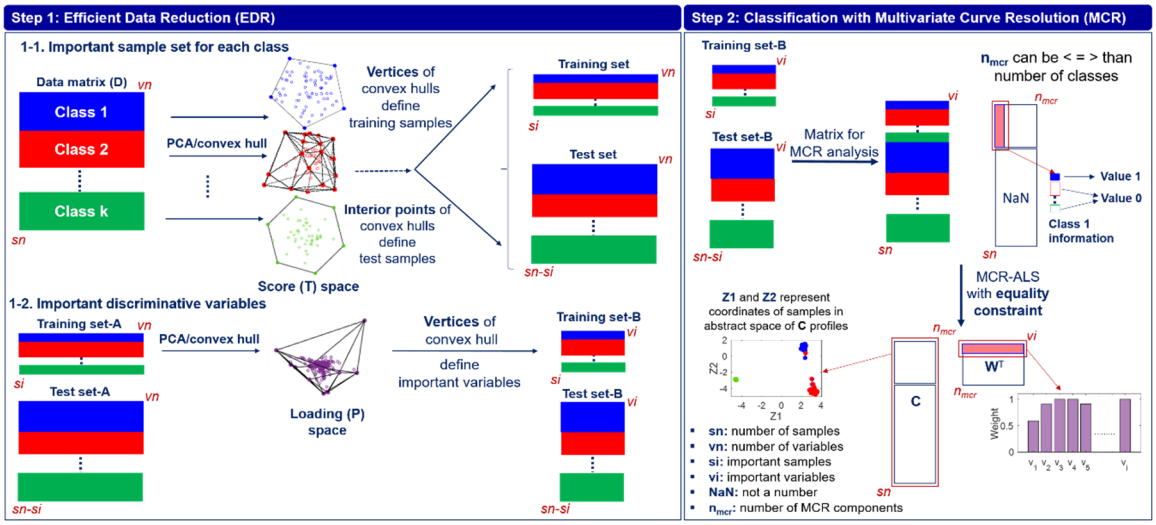
Here, PCA along with the convex geometry principle were adapted as a potential methodology for data reduction and data splitting in an approach termed efficient data reduction (EDR). According to convex hull intersection theory, there are some points (i.e., vertices of the convex hull) within the bilinear data set that are representative of all data. The system's information is stored within these points, and all the other

Received: April 2, 2020
Accepted: March 8, 2021
Published: March 19, 2021



https://doi.org/10.1021/acs.analchem.0c01427
Anal. Chem. 2021, 93, 5020−5027

Scheme 1. Workflow for Classification with EDR-MCR



points (inside the convex hull) can be considered to be linear combinations of these critical points.[15−20] Therefore, the critical points that define the vertices can in principle serve as the means by which to accomplish data reduction, extraction of information about the variables of greatest importance in facilitating learning, and determination of the training/test data split. EDR can be used for data splitting, variable selection, or both. The training and test matrices that result from the application of this approach can then serve as the input for machine learning methods to create supervised models. In this work, MCR methods, which already enjoy broad use in various fields,[21−27] were applied for supervised learning. MCR is a bilinear decomposition approach that supports inclusion of additional system information within the form of numerical constraints.[28−35] MCR is based on classical least-squares (CLS) models which are frequently used for target detection (when the target spectra are known) and for targeted anomaly detection with hyperspectral imaging data as well as with data of other types.[36−38] In order to develop an MCR algorithm for supervised learning, the class labels of the samples can be included as additional information in the form of soft or hard equality constraints within the MCR platform.

We have coined the term "efficient data reduction-multivariate curve resolution" (EDR-MCR) to refer to the combination of EDR with MCR methods. To illustrate the utility of EDR and MCR in supervised learning, we show here the results of its application to four different types of data and compare and contrast them with the outcomes of previously reported data splitting and machine learning processing approaches. Two are publicly available benchmark data sets (composed of the physical characteristics of iris plants and the chemical composition of wines). The third is direct analysis in real time-high resolution mass spectrometry (DART-HRMS) data derived from analysis of seeds of plants from the nightshade plant family representing 24 species.[39] The forth is NMR data representative of human plasma from colorectal cancer and nonmalignant cases.[40]

Theory of the Proposed Method. The new EDR-MCR approach is founded on the in tandem implementation of two main steps: (1) data reduction using EDR, which intelligently reveals the training/test split as well as the most heavily weighted variables and (2) discrimination with MCR, which creates a model using a training set and predicts test samples. The general workflow of the method is presented in Scheme 1 and is described below for a set of experiments, $\mathbf{x}_1,\cdots, \mathbf{x}_{sn}$ (where "sn" refers to sample number) representing the measured features of samples in the form of a row-wise data matrix, **D**. **D** and the labels of the samples, $y_1, ..., y_{sn}$ serve as the input for the EDR-MCR method.

EDR. Implementation of the EDR strategy (step 1 in Scheme 1) has two components: data splitting, variable selection, or both. For selection of the training/test split, EDR can be performed as indicated in Scheme 1 step 1-1, and for selection of the discriminating variables, steps 1-1 and 1-2 should be accomplished consecutively. In Scheme 1 step 1-1, a disjoint classes reduction method was used for selection of training samples. Thus, samples belonging to each class ($\mathbf{x}_i \in$ class k) were disjointedly analyzed by PCA decomposition (as defined in eq 1), and the significant number (npc) of scores ($\mathbf{T}$) were used for creating the convex space after normalization. Normalization to the first eigenvector as a type of Borgen norm[41] was used. Therefore, the score vector of each sample was multiplied by the inverse of its first score value. With this normalization, the dimensionality of the space is reduced by 1, and the points are bounded in a simplex whose vertices are considered as corresponding with the samples most important in defining the data space. Equation 2 shows the expression used for the computation of the convex hull of the normalized scores of class k, where nk and vn are the number of samples belonging to class k and the number of variables, respectively.

$$\mathbf{D}(nk \times vn) = \mathbf{T}(nk \times npc)\mathbf{P}^{\mathrm{T}}(npc \times vn) + \mathbf{E}_{\mathrm{PCA}}(\mathrm{residual}) \quad (1)$$

$$S^k = \left\{ \sum_{i=1}^{nk} \lambda_i \mathbf{T}_i^n : \lambda_i \geq 0 \text{ for all } i \text{ and } \sum_{i=1}^{nk} \lambda_i = 1 \right\} \quad (2)$$

The samples of class k that correspond to the vertices of the computed convex hull (S^k) define the training set for class k. Conversely, samples related to the interior points of the convex hulls, which are linear combinations of the vertices of the convex hulls, define the test set.

https://doi.org/10.1021/acs.analchem.0c01427
Anal. Chem. 2021, 93, 5020−5027

In step 1-2 in Scheme 1, the matrices containing training samples that emerged from the classes in step 1-1 were merged to create a single matrix which was analyzed by PCA based on eq 3 to create a shared space of classes from which the informative variables could be extracted. This enables consideration of the discriminative between-class information in the selection of variables. After normalization, the loadings ($\mathbf{P}$) corresponding with the normalized scores were used for computation of the convex hull, V, consistent with eq 4. The vertices of the resulting convex hull reveal the variables important for reducing the dimensionality of the data. The reduced training and test matrices serve as the input for the application of MCR (step 2 in Scheme 1).

$$\mathbf{D}_{\text{training}}(si \times vn) = \mathbf{T}(si \times npc)\mathbf{P}^{\text{T}}(npc \times vn) + \mathbf{E}_{\text{PCA}}(\text{residual}) \quad (3)$$

$$V = \left\{ \sum_{j=1}^{vn} \lambda_j \mathbf{P}_j^{\text{T}} : \lambda_j \geq 0 \text{ for all } j \text{ and } \sum_{j}^{vn} \lambda_j = 1 \right\} \quad (4)$$

EDR is sensitive to outliers, which should be detected using outlier detection measures. The decision on whether a point is an outlier or a convex hull vertex depends on the accuracy of the outlier detection measure used. In this work, Hotelling's T-squared and Qres were used for outlier detection. However, simultaneous evaluation of multiple outlier measures can simplify outlier determination and provide improved detection.[42]

MCR. The matrix containing the training and test data ($\mathbf{D}_{\text{EDR}}$) was introduced to the MCR in step 2 in Scheme 1. MCR resolves the data matrix[30] to $\mathbf{C}$ and $\mathbf{W}^{\text{T}}$ according to eq 5.

$$\mathbf{D}_{\text{EDR}}(sn \times vi) = \mathbf{C}(sn \times n_{mcr})\mathrm{W}^{\text{T}}(n_{mcr} \times vi) + \mathbf{E}_{\text{mcr}}(\text{residual}) \quad (5)$$

To create a supervised learning model with the MCR method, the class labels of the samples were included as additional information in the form of soft or hard equality constraints within the MCR platform. To apply the equality constraint, one *not a number* "NaN" matrix composed of the number of samples × the number of MCR components (n_{mcr})[43] was defined, and the class information on the samples was included in the corresponding coordinates in the matrix. The labels of the training samples that were associated with each class were vectors that were assigned a binary code of 0 and 1 which designated the assignment or nonassignment of a sample to a class, respectively. MCR is initialized with estimates of $\mathbf{C}$ or $\mathbf{W}^{\text{T}}$, and $\mathbf{C}$ and $\mathbf{W}^{\text{T}}$ are optimized iteratively using an alternating least-squares (ALS) algorithm until convergence is reached and the constraints are fulfilled. The optimized $\mathbf{C}$ contains the samples' class information (i.e., sample similarities and differences). $\mathbf{W}^{\text{T}}$ aligns with the weights of the projected data in the class space and therefore reflects the relative weight contributions of the variables and their impact on the class structure. In implementing the hard equality constraint, $\mathbf{C}$ coordinates that corresponded with the training set were confined to values of 0 and 1 for assignment or nonassignment of a sample to a class, respectively, while for the soft equality constraint implementation, deviation from the constrained value was allowed. As detailed systematically in the literature,[44,45] implementation of soft constraints and specifically soft equality constraints has significant advantages, including improving prediction accuracy.

It should be noted that both training and test samples simultaneously contribute to the building of the MCR model. This permits inclusion of test set information in the creation of the model. In fact, it was shown that this approach yielded better quantitative results for MCR-ALS in comparison with partial least-squares (PLS), when a limited number of calibration samples were available.[28]

For evaluation of unknown test samples, data can be analyzed and classified in step 2. Therefore, the samples were reduced using detected variables (in step 1-2) and added to the training set for MCR analysis. The EDR part of the algorithm was written in house in MATLAB, and MCR was run using MCR-ALS GUI 2.0.[46] All calculations were performed using MATLAB 9.3.0, R2019a software (The MathWorks, Inc., Natick, MA, USA). Practical notes on the application of EDR-MCR are presented in the Supporting Information.

EXPERIMENTAL DATA

Four data sets, two benchmarks (https://archive.ics.uci.edu/ml/datasets), and two high-dimensional chemical data sets were considered in our assessment of the abilities of the proposed method in contrast to others.

Benchmark-I: Iris Flower Data Set. The data set represents the three species *I. setosa*, *I. virginica*, and *I. versicolor* with consideration of four features, namely, sepal length, sepal width, petal length, and petal width.

Benchmark-II. Wine Data Set. The data set contains the chemical constituents of wines grown in the same region in Italy but derived from three different cultivars. The 13 quantified attributes for characterization of the wines are alcohol, maleic acid, ash, ash alkalinity, magnesium, total phenols, flavonoids, nonflavonoid phenols, proanthocyanins, color intensity, hue, the OD280/OD315 of diluted wines, and proline.

Hallucinogenic Solanaceae (Nightshade) Species Data Set. This data represents a 24-class problem composed of DART-HRMS of the seeds of 24 plant species (detailed in the Supporting Information, Nightshade species section). These plants are taxonomically related and are members of five genera in the Solanaceae family. The details of the sample preparation, instrumentation, and DART-HRMS analyses are as described previously.[39] Figure S1 displays representative mass spectra of the 24 species, in the mass range m/z 40−700. The data matrix had 219 rows (i.e., number of samples) and 2976 columns (i.e., number of m/z values).

Colorectal Cancer (CRC) Data Set. The data represent human plasma samples from a verified CRC group and a group with other nonmalignant findings and were reported as part of a study on patients undergoing large bowel endoscopy due to symptoms which could be ascribed to CRC.[40] These samples were analyzed using both fluorescence and ^{1}H NMR spectroscopy (CPMG and NOESY-Presat). In this report, we used the NMR data which were composed of 94 samples (47 cancer samples and 47 adenoma samples) and 455 peaks. The first 201 peaks were from CPMG, and the remaining 254 were from the NOESY data.

RESULTS AND DISCUSSION

With the aim of assessing the adequacy of the EDR-MCR approach relative to other supervised learning processes, the

https://doi.org/10.1021/acs.analchem.0c01427
Anal. Chem. 2021, 93, 5020−5027

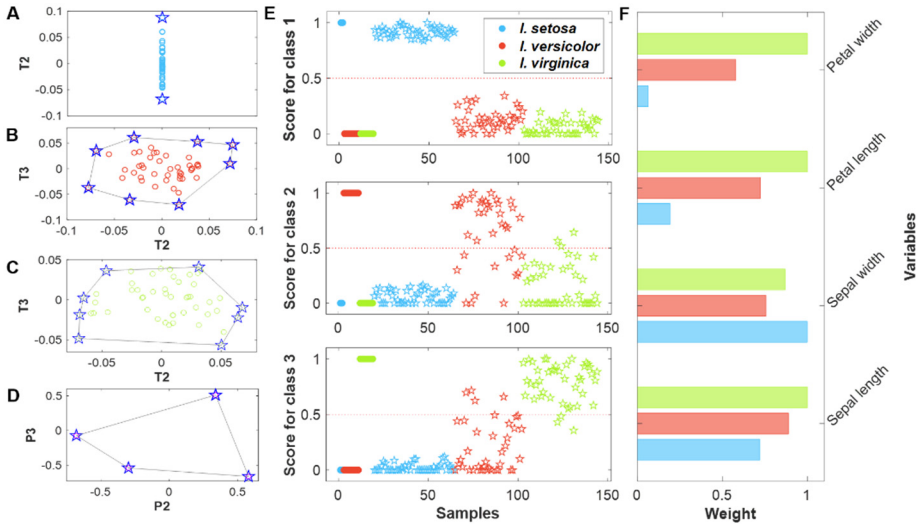


Figure 1. Results of the application of EDR (A–D) and MCR (E, F) in the analysis of the iris flower data set. The blue stars in panels A–C define the extreme points of the convex line in case A and the vertices of the convex hull in cases B and C and represent the coordinates of the samples that were selected as important and as training samples. The circles, color coded by species, represent those samples at the interior of the convex space that were selected as test samples. The blue stars in panel D represent the vertices of the convex hull in the loading space and therefore the coordinates of the important variables. (A) Score plot illustrating the convex line in normalized 1-D space generated by EDR for the selection of the optimal training and test samples for *I. setosa*. Normalization of the two PCs that explained 99% of the data variance resulted in the 1-D plot. (B) Score space obtained from PCA analysis showing the convex hull which, following normalization, revealed the training/test samples for *I. versicolor*. (C) Score space resulting from PCA analysis showing the computed convex hull which, following normalization, revealed the test/training samples for *I. virginica*. (D) Loading space derived from PCA analysis showing the convex hull which, following normalization, revealed the important discriminative variables for the whole iris flower data set. (E) Three profiles that were revealed by the application of MCR on the reduced data which resulted from EDR. The top plot defines the class profile for *I. setosa*. The middle plot defines the class profile for *I. versicolor*. The bottom plot defines the class for *I. virginica*. The red dashed lines in the plots display the threshold (at 0.5) that was used for class assignment. (F) Illustration of the weight profiles resulting from MCR that were normalized to show the relative differences between the weights of each variable for each class.

four data sets were subjected to other data splitting and discrimination methods. The data sets were divided into training and test sets using independent IRS (100 iterations), KS, and EDR step 1-1 methods and then discriminated by PLS-DA and error-correcting output code (ECOC) multiclass models[47] using LDA, KNN, CART, and SVM. Details on the strategies taken for data set preprocessing, parameter setting, and methods comparison are discussed in the strategies for comparison of different methods section in the Supporting Information.

Benchmark Data Set. Two publicly available and well-established benchmark data sets which have been used previously to test machine learning approaches were analyzed. The benchmark data sets have dimensions of 150 × 4 for the iris and 178 × 13 for the wine. The FreeVis plots[48,49] in Figure S2 enable visualization of the results of projection of iris (panel A) and wine (panel B) multivariate data in 2D space. The blue, red, and green colors, respectively, correspond to *I. setosa*, *I. versicolor*, and *I. virginica* classes in the iris data set and to the Group 1, Group 2, and Group 3 classes in the wine data set, respectively. According to Figure S2-A, sepal length has less of an impact on discrimination of iris plant species when compared with the other features. In Figure S2-B, the ash and magnesium features have low weights in terms of discrimination of the wine data set. As expected, total phenols and proanthocyanins are highly correlated with flavonoids and hue, respectively.

Iris Data Set Data Reduction. Six of the samples were identified to be outliers based on three principal components.

Sample Selection. After removing outliers, the data were categorized into training/test samples (19/125, 99/45, and 101/43 using EDR, KS, and IRS, respectively). The details for the samples in each class are displayed in Tables S1–S3.

Two, three, and three PCs were found to be important for defining the PCA space of samples belonging to *I. setosa*, *I. versicolor*, and *I. virginica*, respectively. The dimensions of the normalized PCA space used for calculation of the convex space were 1, 2, and 2 for *I. setosa*, *I. versicolor*, and *I. virginica*, respectively, as shown in Figure 1A–C. For class *I. setosa* (Figure 1A), the data points merged along a straight line in which the two extreme data points (indicated by stars in Figure 1A) defined the training samples. Thus, only two samples defined the set for this class. The vertices of the two resulting 2-D convex spaces which defined the training samples for classes *I. versicolor* and *I. virginica* (shown in Figure 1B and C, respectively, are illustrated with stars).

Variable Selection. All four variables were selected to be important using step 1-2 of the EDR method, as they define the vertices of the convex hull in the normalized 2D space (shown in Figure 1D), which was obtained from implementation of PCA on the training sets of all three classes resulting from EDR.

Wine Data Set Data Reduction. The results of the application of EDR to the wine data set are presented in Figure S3. Of the 178 samples (based on the three principal components), 13 were found to be outliers and were removed before analysis.

Sample Selection. On application of EDR, 59 samples (7 group 1, 32 group 2, and 20 group 3) and 106 samples (51

https://doi.org/10.1021/acs.analchem.0c01427
Anal. Chem. 2021, 93, 5020–5027

Table 1. Performance Merits of Various Discrimination Methods Including Accuracy, Overall Sensitivity, Specificity, and Overall Precision and F1-Score for Discrimination of Iris, Wine, Nightshade, and CRC Data Sets[a]

	Accuracy	Sensitivity	Specificity	Precision	F1-score	Accuracy	Sensitivity	Specificity	Precision	F1-score
	Iris flower data set					DART-HRMS data set-Nightshade species				
EDR-KNN	0.86	0.86	0.93	0.89	0.86	0.62	0.69	0.98	0.67	0.60
EDR-LDA	0.88	0.88	0.94	0.88	0.87	0.80	0.86	0.99	0.89	0.80
EDR-CART	0.83	0.83	0.92	0.85	0.83	0.61	0.72	0.98	0.69	0.60
EDR-SVM	0.92	0.92	0.96	0.93	0.92	0.80	0.84	1.00	0.82	0.79
EDR-PLS-DA	0.91	0.91	0.96	0.91	0.91	0.82	0.84	0.99	0.84	0.82
EDR-MCR	0.94	0.93	0.97	0.93	0.94	0.96	0.95	1.00	0.96	0.96
KS-KNN	1.00	1.00	1.00	1.00	1.00	0.90	0.90	1.00	0.89	0.88
KS-LDA	1.00	1.00	1.00	1.00	1.00	0.92	0.92	1.00	0.94	0.92
KS-CART	0.96	0.96	0.98	0.96	0.96	0.78	0.77	0.99	0.79	0.75
KS-SVM	1.00	1.00	1.00	1.00	1.00	0.92	0.92	1.00	0.92	0.92
KS-PLS-DA	0.98	0.98	0.99	0.98	0.98	0.94	0.94	1.00	0.94	0.94
KS-MCR	0.95	0.95	0.98	0.96	0.95	0.90	0.92	1.00	0.93	0.91
IRS-KNN	0.98	0.98	0.99	0.98	0.98	0.81	0.80	0.99	0.82	0.80
IRS-LDA	0.98	0.98	0.99	0.98	0.98	0.83	0.82	0.99	0.84	0.82
IRS-CART	0.95	0.95	0.97	0.95	0.95	0.83	0.82	0.99	0.83	0.82
IRS-SVM	0.96	0.96	0.98	0.96	0.96	0.87	0.87	0.99	0.87	0.86
IRS-PLS-DA	0.96	0.96	0.98	0.96	0.96	0.79	0.77	0.99	0.78	0.77
IRS-MCR	0.90	0.90	0.95	0.92	0.90	0.87	0.86	0.99	0.86	0.86
	Wine data set					NMR data set-CRC				
EDR-KNN	0.97	0.98	0.99	0.97	0.97	0.51	0.26	0.93	0.86	0.40
EDR-LDA	1.00	1.00	1.00	1.00	1.00	0.57	0.39	0.86	0.82	0.53
EDR-CART	0.90	0.93	0.95	0.91	0.91	0.68	0.65	0.71	0.79	0.71
EDR-SVM	0.99	0.99	1.00	0.99	0.99	0.62	0.43	0.93	0.91	0.59
EDR-PLS-DA	0.98	0.99	0.99	0.98	0.98	0.65	0.83	0.36	0.68	0.75
EDR-MCR	0.97	0.98	0.99	0.97	0.97	0.78	0.77	0.77	0.77	0.77
KS-KNN	1.00	1.00	1.00	1.00	1.00	0.72	0.72	0.72	0.72	0.72
KS-LDA	1.00	1.00	1.00	1.00	1.00	0.71	0.70	0.70	0.71	0.70
KS-CART	0.96	0.96	0.98	0.96	0.96	0.68	0.68	0.68	0.67	0.67
KS-SVM	1.00	1.00	1.00	1.00	1.00	0.71	0.70	0.70	0.71	0.70
KS-PLS-DA	0.98	0.98	0.99	0.98	0.98	0.50	0.46	0.46	0.44	0.43
KS-MCR	0.98	0.98	0.99	0.98	0.98	0.61	0.59	0.59	0.60	0.59
IRS-KNN	0.98	0.98	0.99	0.98	0.98	0.62	0.62	0.62	0.62	0.62
IRS-LDA	0.97	0.97	0.99	0.97	0.97	0.62	0.62	0.62	0.62	0.62
IRS-CART	0.94	0.94	0.97	0.94	0.94	0.55	0.55	0.55	0.55	0.55
IRS-SVM	0.98	0.98	0.99	0.98	0.98	0.64	0.64	0.64	0.65	0.64
IRS-PLS-DA	0.94	0.95	0.97	0.94	0.94	0.61	0.61	0.61	0.63	0.59
IRS-MCR	0.93	0.94	0.97	0.93	0.93	0.67	0.67	0.67	0.67	0.66

[a]Sensitivity, specificity, and precision of each class are displayed in Tables S1−S6, S11−S13, S18−S20, and S25−S30. Confusion matrices for test set prediction are shown in Appendixes S1−S4. Error rate: 1, accuracy. False positive rate: 1, specificity. False negative rate: 1, sensitivity.

group 1, 29 group 2, and 26 group 3) were extracted for the training and test sets, respectively, using the first three PCs resulting from PCA on the logarithm-transformed data in each class. The data were divided into training/test samples for each class (114/51 and 116/49 using KS and IRS, respectively) and are shown in Tables S4−S6.

Variable Selection. The three PCs corresponding to PCA analysis of the training set derived from step1-1 of EDR were used for variable selection. Eight variables (i.e., malic acid, ash alkalinity, magnesium, flavonoids, color intensity, hue, the OD280/OD315 of diluted wines, and proline) coincided with the vertices of the convex hull, while the total phenols and nonflavonoid phenols and the proanthocyanins, which are correlated with flavonoids and hue, respectively, did not. This is consistent with the FreeVis plot results (Figure S2).

Classification. MCR was performed to resolve the iris and wine data sets (using three and four MCR components, respectively). Figure 1E and F, along with Figure S3 illustrate the scaled class and weight profiles which resulted from the application of MCR to the iris and wine data sets, respectively. The samples belonging to the training and test sets are shown as color-coded circles and stars. Each row of the **C** matrix corresponds to a sample. The column position of values >0.5 in a row of matrix **C** was considered as the criterion that defined the class of samples corresponding with that row. The weight profiles show different weights for petal width and petal length variables on the class profiles for each species and illustrate their impact on the classification of each species. Table 1 contains the accuracy, overall sensitivity, specificity, and precision and F1-score merits of the trained EDR-, KS-, and IRS- MCR, PLS-DA, and the following multiclass ECOC models using the classifiers: regularized LDA, CART, KNN, and SVM for the iris and wine data sets. The details regarding the parameters and the sensitivity, specificity, and precision of each class can be found in Tables S1−S3 for the iris data set and in Tables S4−S6 for the wine data. The confusion matrices for prediction of test samples are displayed in Appendices S1A-C and S2A-C for the iris and wine data sets, respectively. The results show that EDR and MCR compare favorably with other powerful and well-accepted data splitting and pattern recognition methods. If an accuracy of the model of ≥0.9 and a true positive rate (sensitivity) for each class of ≥0.8 are considered as the criteria for assessment of model performance, then all the methods except for EDR-CART perform well for discrimination of the wine data set. IRS-MCR, EDR-SVM, CART, and LDA do not generate well-fitting models for the iris data set.

Nightshade Data Set. In addition to the benchmark data sets, the EDR-MCR approach was also applied to high-dimensional DART-HRMS data that was acquired from analysis of the seeds of 24 psychoactive plant species from the nightshade family. Species-level discrimination of these plants is of importance in agricultural, medicinal, and forensic contexts. In our previous work,[39] a two-level hierarchical

https://doi.org/10.1021/acs.analchem.0c01427
Anal. Chem. 2021, 93, 5020−5027

classification tree, inspired by the known taxonomic relationships between the represented plant species, was designed to categorize the plant samples first by genus and subsequently by species. In each node of the classification tree, subwindow permutation analysis (SPA) and PLS-DA were applied to the data for variable selection and discrimination, respectively. The applied method increased the accuracy of 100 × bootstrap validation to 95% compared to 84% for a flat 24-classs problem. The number of variables used for discrimination of the plant samples by genus was identified to be 170, and the number of variables for categorizing the *Atropa*, *Brugmansia*, *Datura*, *Hyoscyamus*, and *Mandragora* genera into the represented species were 40, 50, 80, 30, and 15, respectively. In this study, the data set was analyzed as a 24-class problem, without considering class hierarchy. Representative training sets were selected using EDR, KS, and IRS and were determined to be 109, 170, and 176 in number, respectively. The training/test sets for each class are shown in Tables S7, S14, and S21. For EDR, all of the samples for species *B. versicolor* and *H. Pusillus* were recognized as training sets, and there were no test samples designated for these two species. From step 1-2 of EDR, 13 PCs (which explained 90% of the data variance) were used to create the shared space between classes, and, based on the created convex hull vertices, 56 m/z values were sufficient to define the bounded subspace. The m/z values displayed in Table S7, which included the biomarkers m/z 290.1726 (i.e., atropine) and 304.1571 (i.e., scopolamine) were among those selected. Twenty-three of these were identical to the high-impact variables identified in the previous study.[39] Therefore, the variables of the training and test sets were reduced to 56 and subjected to machine learning methods for generation of the classification models. Table 1 contains the figures of merit for the trained EDR-, KS-, and IRS-PLS-DA, MCR, and the following multiclass ECOC models for test prediction: regularized LDA, CART, KNN, and SVM. For the MCR method, samples were assigned to classes, with a class profile threshold of <0.4.

The details associated with the optimized parameters, confusion matrix, and performance merits, i.e., sensitivity, precision, and specificity of each class for predicting training and test sets, appears in Appendices S3A−S3C and Tables S7−S27. Figure S4 displays the projected features of the matrices (with the training and test sets indicated using circle and star symbols, respectively) in the 2-D space of a *t*-distributed stochastic neighbor embedding (*t*-SNE) plot.[50] Utilization of the neighbor-embedding technique preserves the pairwise similarities (by considering the "cosine" similarity metric) of the data points of neighbors by minimizing the divergence of similarity distributions between neighbors and embedding the points in a 2-D space. Figures S4A and B show projection plots for the raw data and EDR variable reduced data, respectively. Although they reveal the presence of clusters, there is no clear discrimination between the color-coded classes. Figure S4C illustrates the projection, on a *t*-SNE space, of the scaled class profiles that are resolved from application of MCR. Twenty-four MCR components were applied for resolving the nightshade data set. Visual assessment of the results reveals that the MCR model performs very well in resolving class profiles of nightshade family species, as demonstrated by the apparent separation of classes in Figure S4C. According to Table 1, when the data splitting is based on the EDR strategy, KNN and CART perform similarly in terms of F1-score, overall sensitivity, and accuracy, but their performance is lower than that for the other methods. By these same metrics, MCR outperforms LDA, PLS-DA, and SVM methods. With the exception of the CART method, the discrimination methods exhibit similar performance when the training set is selected using the KS method. For the training set selected by IRS, the MCR and SVM models have similar and better performances. However, since accuracy, overall sensitivity, and F1-score alone are typically not enough to decide on the suitability of a classifier, the other performance characteristics such as sensitivity, specificity, and precision of each class were also considered. The results for test set prediction are shown in Tables S11−S13, S18−S20, and S25−S27 and reveal that the true positive rate (sensitivity) of each class is zero for *A. komarovii*, *D. quercifolia*, and *H. niger* using EDR-KNN; *B. sanguinea* and *D. quercifolia* using EDR-CART; *D. inoxia* using KS-KNN; *B. suaveolens* and *D. discolor* using KS-CART; and *B. arborea* using KS-SVM. The true positive rate was also zero for *A. komarovii*, *D. quercifolia*, and *H. niger* in prediction of the training set using EDR-KNN. In addition, the fit was poor for some of the species when discriminated with EDR-SVM, LDA, and PLS-DA and IRS-KNN, CART, LDA, PLS-DA, SVM, and MCR, as the true positive rates of those classes in test set prediction were as low as <0.5. On the other hand, EDR-MCR properly predicted all of the species. This demonstrates that MCR exhibits good classification performance even with a small number of samples (i.e., two) and the presence of class imbalances.

We conclude that the utilization of both training and test sets to generate the MCR model, which allows consideration of test set information, confers on MCR the ability to better train classes with only two samples. The resolved and scaled weight profiles are demonstrated for 24 species in Figures S5−S9. Each weight profile is related to one species and shows the impact of the indicated m/z values on separation of that class from the others. For example, in the *Atropa* genus, m/z 304.1571 does impact the distinction of *A. belladonna* from the two other represented species *A. beatica* and *A. komarovii*. As another example featuring the *Brugmansia* genus, m/z 291.17 is heavily weighted for the *B. aurea* species, while it has no impact in the feature space of the four other species.

CRC Data Set. The EDR-MCR method was applied to high-dimensional two-class NMR data derived from analysis of 94 human blood plasma samples from donors with and without colorectal cancer.[40] In previous reports, this data was analyzed following its fusion with fluorescence spectroscopy data,[40,51] and in one study[51] in which the CRC samples were clustered using an unsupervised data fusion model, 71.4% accuracy with 63.6% sensitivity and 78.1% specificity were achieved. Using EDR, KS, and IRS, 57, 66, and 66 samples, respectively, were selected as training sets. In EDR, five and six principle components which explained 99.9% of the data variance for the cancer and noncancer samples, respectively, resulted in 33 and 24 training set samples for the cancer and noncancer samples, respectively. Six principal components defined the shared space of the training sets and revealed 10 important variables. The results for the evaluation of machine learning methods and MCR for distinguishing between cancerous and noncancerous samples are displayed in Table 1. The optimal parameters for the PLS-DA and ECOC models and performance specifications for each class are shown in Tables S28−S30. The confusion matrix associated with the prediction of test samples is shown in Appendices S4A−S4C.

https://doi.org/10.1021/acs.analchem.0c01427
Anal. Chem. 2021, 93, 5020−5027

Two MCR components were found to be optimal for sample discrimination.

The results in Table 1 and Figure S10 illustrate that the EDR-MCR model performed well in comparison with the other methods and published literature report.[51] The EDR-MCR discrimination model performed well and resulted in 78% accuracy with 71% sensitivity and 83% specificity in distinguishing between CRC samples. On the basis of Figure S10, two variables (i.e., Var 1 and Var 8) of the 10 selected variables have the most impact on the ability to distinguish between cancer and noncancer samples. According to Tables S28−S30, in comparison, EDR-CART, KS-KNN, KS-CART, and IRS-MCR show relatively good performance and have a sensitivity and specificity of >0.60.

The analysis results for the four data sets illustrate that the EDR-MCR method is highly suitable for supervised learning tasks. It is noteworthy that the results of the application of MCR are affected by rotational ambiguity, which means that there is a set of solutions that fit the data and fulfill the constraints equally.[52,53]

Generally, the non-negativity constraint is a robust approach that results in a reduced number of feasible solutions. As such, its application to several systems is a sufficient condition to resolve profiles uniquely.[54] In discriminative analysis in general, implementation of the non-negativity constraint for the MCR method involves the application of the accessible information about most of the chemical system, and this results in a reduced number of feasible solutions which have positive values in their class and weight profiles. Overall, the application of the non-negativity constraint has a significant effect in terms of reduction of the number of feasible solutions in cases where the application of the equality constraint does not result in a unique solution. Implementing equality, correlation, and known value constraints can result in unique solutions for specific conditions.[28,29,34] In the present work, implementation of the equality constraint (for introducing class label information into the MCR process) resulted in unique solutions relating to the similarity of the training and test sample spaces. Further investigation into the conditions necessary for obtaining unique solutions when the MCR method is used for discriminant analysis and classification are the subject of ongoing investigations in our laboratories.

CONCLUSIONS

A novel approach termed EDR-MCR was developed for multiclass classification of high-dimensional data. The method introduces the coupling of EDR and MCR as a new strategy for data splitting, variable selection, and supervised classification of high dimensionality data. In comparison with investigated data splitting and classification methods, EDR-MCR exhibited better performance in the classification of the two high-dimensional data sets analyzed (i.e., the CRC and nighshade plant data sets). EDR-MCR has the potential to be used both as a one class classifier and also as a discrimination analyzer. EDR provides a simple method to reduce the data using the most dominant samples and variables. The results show that EDR exhibited results that were comparable to those of other data splitting methods, and in comparison with other data splitting methods, EDR in combination with MCR has better performance and results in a functional model even when used with a limited number of training samples. While multivariate curve resolution tasks are well known to analytical chemists, the approach presented here has not been reported for classification and discrimination tasks. Described here for the first time is a method for classification and discrimination that takes advantage of important features of the MCR technique in order to accomplish classification and discrimination. In comparison with other classification methods, MCR offers the added advantages of (1) speed, (2) tuning of fewer parameters and simplicity of tuning, (3) flexibility in the analysis of data sets that is characterized by low sample numbers and class imbalances, (4) improved accuracy from inclusion of additional information about the system in the form of a numerical constraint, and (5) pure component signal weights that result from bilinear decomposition and which are useful for revealing a given variables' impact. This approach can be readily applied to a broad range of data types. Future studies will explore additional examples of the application of the conceptual framework developed in this work to the classification of a diversity of types of data.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.0c01427.

Practical notes on the application of EDR-MCR (notes 1-3); Nightshade species; strategy for comparision of different methods; 10 additional figures, 30 tables, and 4 appendices referenced in the text: FreeViz plot of iris and wine data sets; DART-HRMS spectra representative of 24 hallucinogenic species from the nightshade family; scaled class and weight profiles resolved by EDR-MCR for wine and nightshade data sets; analysis results for the iris, wine, nightshade, and CRC data sets including selected variables and samples for training and test sets by EDR; performance results for prediction of training and test samples by MCR, PLS-DA, and ECOC multiclass models using LDA, KNN, CART, and SVM. (PDF)

AUTHOR INFORMATION

Corresponding Authors

Hamid Abdollahi − *Department of Chemistry, Institute for Advanced Studies in Basic Sciences (IASBS), Zanjan 45137-66731, Iran;* orcid.org/0000-0002-5994-6365; Phone: (+98) 24-3315-3122; Email: abd@iasbs.ac.ir

Rabi A. Musah − *Department of Chemistry, University at Albany, State University of New York, Albany, New York 12222, United States;* orcid.org/0000-0002-3135-4130; Phone: 518-437-3740; Email: rmusah@albany.edu

Author

Samira Beyramysoltan − *Department of Chemistry, University at Albany, State University of New York, Albany, New York 12222, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.0c01427

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The financial support of the U.S. National Institute of Justice to R.A.M. (Grants 2015-DN-BX-K057, 2018-R2-CX-0012, and 2019-DU-BX-0026), as well as the U.S. National Science

Foundation to R.A.M. (Grant 1429329), is gratefully acknowledged. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the U.S. Department of Justice.

REFERENCES

(1) Zhang, J. M.; Harman, M.; Ma, L.; Liu, Y. *IEEE T. Software Eng.* **2020**, DOI: 10.1109/TSE.2019.2962027.
(2) Ho, Y. C.; Pepyne, D. L. *J. Optimiz. Theory Appl.* **2002**, *115* (3), 549−570.
(3) Ghanbari, N. A Review of Feature Selection Methods with the Applications in Pattern Recognition in the Last Decade. In *Fundamental Research in Electrical Engineering*; Shahram, M. K., Ed.; Springer Singapore: Singapore, 2019; pp 163−171.
(4) Heinze, G.; Wallisch, C.; Dunkler, D. *Biom. J.* **2018**, *60* (3), 431−449.
(5) Roffo, G., Feature selection library (MATLAB toolbox). *arXiv:1607.01327*, 2016.
(6) Nowotny, T. *Front. Robot. AI.* **2014**, *1* (5), na DOI: 10.3389/frobt.2014.00005.
(7) de Boves Harrington, P. *TrAC, Trends Anal. Chem.* **2006**, *25* (11), 1112−1124.
(8) Rao, C. R.; Wu, Y. *J. Stat. Plan. Inference* **2005**, *128* (1), 231−240.
(9) Kohav, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the14th International Joint Conference on Artificial Intelligence*; Morgan Kaufmann Publishers, Inc.: San Francisco, CA, USA, 1995; Vol. *2*, pp 1137−1145.
(10) Borovicka, T.; Jirina, M., Jr.; Kordik, P.; Jirina, M. Selecting Representative Data Sets. In *Advances in Data Mining Knowledge Discovery and Applications*; Karahoca, A., Ed.; InTech: Rijeka, Croatia, 2012; pp 43−70.
(11) Daszykowski, M.; Walczak, B.; Massart, D. L. *Anal. Chim. Acta* **2002**, *468* (1), 91−103.
(12) Morais, C. L. M.; Santos, M. C. D.; Lima, K. M. G.; Martin, F. L. *Bioinformatics* **2019**, *35* (24), 5257−5263.
(13) Lee, L. C.; Liong, C.-Y.; Jemain, A. A. *AIP Conf. Proc.* **2017**, *1940* (1), 020116.
(14) Reitermanová, Z., Data Splitting. In *WDS'10 Proceedings of Contributed Papers*, 2010; Vol. *10*, pp 31−36.
(15) Ghaffari, M.; Omidikia, N.; Ruckebusch, C. *Anal. Chem.* **2019**, *91* (17), 10943−10948.
(16) Böhm, C.; Kriegel, H.-P. In Determining the Convex Hull in Large Multidimensional Databases, Data Warehousing and Knowledge Discovery; In *Data Warehousing and Knowledge Discovery*; Kambayashi, Y., Winiwarter, W., Arikawa, M., Eds.; Springer: Berlin, Heidelberg, 2001; pp 294−306.
(17) Nemirko, A. P. *Pattern Recognit. Image Anal.* **2017**, *27* (3), 387−394.
(18) Zhong, J.; Tang, K.; Qin, A. K. Finding Convex Hull Vertices in Metric Space. In *International Joint Conference on Neural Networks (IJCNN)*, Beijing, China, 6−11 July, 2014; pp 1587−1592.
(19) Mavroforakis, M. E.; Theodoridis, S. *Neural Networ.* **2006**, *17* (3), 671−682.
(20) Thurau, C.; Kersting, K.; Wahabzada, M.; Bauckhage, C. *Knowl. Inf. Sys.* **2011**, *29* (2), 457−478.
(21) Jaumot, J.; Aviñó, A.; Eritja, R.; Tauler, R.; Gargallo, R. *J. Biomol. Struct. Dyn.* **2003**, *21* (2), 267−278.
(22) Wentzell, P. D.; Karakach, T. K.; Roy, S.; Martinez, M. J.; Allen, C. P.; Werner-Washburne, M. *BMC Bioinf.* **2006**, *7* (1), 343.
(23) Paatero, P.; Hopke, P. K.; Begum, B. A.; Biswas, S. K. *Atmos. Environ.* **2005**, *39* (1), 193−201.
(24) Navea, S.; De Juan, A.; Tauler, R. *Anal. Chem.* **2003**, *75* (20), 5592−5601.
(25) De Juan, A.; Jaumot, J.; Tauler, R. *Anal. Methods* **2014**, *6* (14), 4964−4976.
(26) Ruckebusch, C.; Blanchet, L. *Anal. Chim. Acta* **2013**, *765*, 28−36.
(27) Azzouz, T.; Tauler, R. *Talanta* **2008**, *74* (5), 1201−1210.
(28) Akbari Lakeh, M.; Abdollahi, H. *Anal. Chim. Acta* **2018**, *1030*, 42−51.
(29) Beyramysoltan, S.; Rajkó, R.; Abdollahi, H. *Anal. Chim. Acta* **2013**, *791*, 25−35.
(30) De Juan, A.; Tauler, R. *Crit. Rev. Anal. Chem.* **2006**, *36* (3−4), 163−176.
(31) Omidikia, N.; Beyramysoltan, S.; Mohammad Jafari, J.; Tavakkoli, E.; Akbari Lakeh, M.; Alinaghi, M.; Ghaffari, M.; Khodadadi Karimvand, S.; Rajkó, R.; Abdollahi, H. *J. Chemom.* **2018**, *32* (12), e2975.
(32) Tavakkoli, E.; Abdollahi, H.; Gemperline, P. J. *Analyst* **2020**, *145* (1), 223−232.
(33) Pomareda, V.; Guamán, A. V.; Mohammadnejad, M.; Calvo, D.; Pardo, A.; Marco, S. *Chemom. Intell. Lab. Syst.* **2012**, *118*, 219−229.
(34) Ahmadi, G.; Tauler, R.; Abdollahi, H. *Chemom. Intell. Lab. Syst.* **2015**, *142*, 143−150.
(35) Pomerantsev, A. L.; Zontov, Y. V.; Rodionova, O. Y. *J. Chemom.* **2014**, *28* (10), 740−748.
(36) Rajkó, R.; Beyramysoltan, S.; Abdollahi, H.; Eőri, J.; Pongor, G. *Anal. Chim. Acta* **2015**, *888*, 19−26.
(37) Gallagher, N. B., Classical least Squares for Detection and Classification. In *Data Handling in Science and Technology*; Amigo, J. M., Ed.; Elsevier, 2020; Vol. *32*, pp 231−246.
(38) Gallagher, N. B., Detection, Classification, and Quantification in Hyperspectral Images Using Classical Least Squares Models. In *Techniques and Applications of Hyperspectral Image Analysis*; Grahn, H. F., Geladi, P., Eds.; Wiley, **2007**; pp 181−202..
(39) Beyramysoltan, S.; Abdul-Rahman, N.-H.; Musah, R. A. *Talanta* **2019**, *204*, 739−746.
(40) Bro, R.; Nielsen, H. J.; Savorani, F.; Kjeldahl, K.; Christensen, I. J.; Brünner, N.; Lawaetz, A. J. *Metabolomics* **2013**, *9* (1), 3−8.
(41) Rajkó, R. *J. Chemom.* **2009**, *23* (6), 265−274.
(42) Brownfield, B.; Kalivas, J. H. *Anal. Chem.* **2017**, *89* (9), 5087−5094.
(43) Motegi, H.; Tsuboi, Y.; Saga, A.; Kagami, T.; Inoue, M.; Toki, H.; Minowa, O.; Noda, T.; Kikuchi, J. *Sci. Rep.* **2015**, *5* (1), 15710.
(44) Akbari Lakeh, M.; Abdollahi, H.; Gemperline, P. J. *Anal. Chim. Acta* **2020**, *1105*, 64−73.
(45) Gemperline, P. J.; Cash, E. *Anal. Chem.* **2003**, *75* (16), 4236−4243.
(46) Jaumot, J.; De Juan, A.; Tauler, R. *Chemom. Intell. Lab. Syst.* **2015**, *140*, 1−12.
(47) Caruana, R.; Niculescu-Mizil, A.; Crew, G.; Ksikes, A. Ensemble Selection from Libraries of Models. In *Twenty-First International Conference on Machine Learning*, Banff, Alberta, Canada, July 4−8, 2004.
(48) Demšar, J.; Curk, T.; Erjavec, A.; Gorup, Č.; Hočevar, T.; Milutinovič, M.; MoŽina, M.; Polajnar, M.; Toplak, M.; Staric̀, A.; Stajdohar, M.; Umek, L.; Žagar, L.; Žbontar, J.; Žitnik, M.; Zupan, B. *J. Mach. Learn. Res.* **2013**, *14*, 2349−2353.
(49) Demšar, J.; Leban, G.; Zupan, B. *J. Biomed. Inf.* **2007**, *40* (6), 661−671.
(50) van der Maaten, L.; Hinton, G. E. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.
(51) Acar, E.; Lawaetz, A. J.; Rasmussen, M. A.; Bro, R. Structure-Revealing Data Fusion Model with Applications in Metabolomics. In *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Osaka, Japan, 2013; pp 6023−6026.
(52) Golshan, A.; Abdollahi, H.; Beyramysoltan, S.; Maeder, M.; Neymeyr, K.; Rajkó, R.; Sawall, M.; Tauler, R. *Anal. Chim. Acta* **2016**, *911*, 1−13.
(53) Tauler, R.; Smilde, A.; Kowalski, B. *J. Chemom.* **1995**, *9* (1), 31−58.
(54) Rajkó, R.; Abdollahi, H.; Beyramysoltan, S.; Omidikia, N. *Anal. Chim. Acta* **2015**, *855*, 21−33.

https://doi.org/10.1021/acs.analchem.0c01427
Anal. Chem. 2021, 93, 5020−5027