Special Issue Article

Journal of CHEMOMETRICS

Received: 21 October 2011, Revised: 20 January 2012, Accepted: 24 January 2012, Published online in Wiley Online Library: 2012

(wileyonlinelibrary.com) DOI: 10.1002/cem.2419

Evaluation of target factor analysis and net analyte signal as processes for classification purposes with application to benchmark data sets and extra virgin olive oil adulterant identification

Kevin Higgins[a], John H. Kalivas[a]* and Erik Andries[b,c]

Classifying samples into known categories is a common problem in analytical chemistry and other fields. For example, with spectroscopic data, samples are measured and the corresponding spectra are compared with existing spectral data sets of known classification (library sets) to determine the appropriate classification. Presented in this paper is a study of the simple and well known data analysis processes target factor analysis (TFA) and net analyte signal (NAS). Although TFA and NAS were originally derived for different purposes in analytical chemistry, they are based on the same calculation. The library set with the smallest TFA residual (smallest NAS) for a test sample spectrum can be used for classification purposes. Alternatively and equivalently, this paper uses the smallest angle (poorest selectivity in NAS terminology) between a new sample spectrum vector and the space spanned by each library loading vector basis set. The angle classification is compared with classifications by the Mahalanobis distance and *k*-nearest neighbors. The measures are evaluated with three spectroscopic data sets consisting of benchmark identification of plastic type (Raman) and gasoil plant source (ultraviolet) and a new extra virgin olive oil adulterant identification (fluorescence) data set. A fourth data set is the benchmark archeological data set. The Mahalanobis distance and *k*-nearest neighbors generally classify with 2%–40% and 0%–20% decreases in correct classifications, respectively, compared with TFA (NAS). Results from this study indicate that the simple TFA and NAS processes are useful underutilized classification and library searching tools. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: classification; target factor analysis; net analyte signal; extra virgin olive oil adulteration

1. INTRODUCTION

Part of analytical chemistry deals with classifying samples. For example, identifying container plastic types on the basis of spectral measurements is important for recycling purposes. Numerous approaches are available for classification problems [1–4]. As the number of potential classes increases for identifying a sample, the more difficult classification can become. Presented in this paper is a study evaluating the simple process of orthogonal projection analysis (OPA) used in target factor analysis (TFA) [5,6] and net analyte signal (NAS) [7,8] for the new purpose of classification. The OPA process can be angular-based and compares a sample spectrum to a library set of spectra and is further described in Section 2. Briefly, the approach is to collect a set of measurements made on a test sample and treat the set as a vector of measurements, for example, a spectrum or a series of chemical concentrations contained in the sample. Respective angles are calculated between this test vector and each corresponding basis set of vectors spanning each library space of a known class. The class with the smallest angle is the test sample identity. The OPA process is applicable in traditional library searching where only one spectrum is used to represent a chemical for a chemical class; that is, each chemical library set has only one pure component spectrum of the respective chemical.

Orthogonal projections are common in analytical chemistry. In addition to TFA and NAS, examples using OPA include preprocessing spectral data to remove nonanalyte information [9], peak purity assessment [10,11], identification of interferences [12,13], and quality control assessment [14]. To date, the authors are not aware of the OPA process used as a single classification merit. The method of soft independent modeling of class analogies [1,15] as described in [1] includes OPA in conjunction with the Mahalanobis distance (MD).

* *Correspondence to: J. H. Kalivas, Department of Chemistry, Idaho State University, Pocatello, Idaho 83209, USA.*
E-mail: kalijohn@isu.edu

a *K. Higgins, J. H. Kalivas*
Department of Chemistry, Idaho State University, Pocatello, Idaho 83209, USA

b *E. Andries*
Center for Advanced Research Computing, University of New Mexico, Albuquerque, New Mexico 87106, USA

c *E. Andries*
Department of Mathematics, Central New Mexico Community College, Albuquerque, New Mexico 87106, USA

The OPA format of comparing a sample spectrum vector with a matrix of spectra is related to another angular approach that compares a matrix of spectra to another matrix of spectra [16]. This other method was originally applied to studies of educational achievements in different student groups and has since been applied to compare different sampling seasons [17] and library searching second-order data sets, for example, spectrochromatograms [18]. Presented in this paper is proof that the angle is equivalent to OPA angle when the test matrix of spectra is replaced with a test sample vector.

In this paper, the OPA angle merit is compared with the MD classification approach [1–3], also commonly used for outlier detection [1,2,19]. Classification by *k*-nearest neighbors (KNN) [1–4] is additionally compared. The three classification methods are tested on four data sets consisting of three benchmark data sets: archeological [20], plastic [21], and gasoil plant identifications [22]. An extra virgin olive oil (EVOO) adulterant identification data set [23] is also evaluated. These data sets range from well-defined clusters in principal component analysis score plots to situations with nonunique overlapped clusters.

2. MATHEMATICS

A test sample vector is denoted as the $w \times 1$ vector $\mathbf{y}$, for w measured values, for example, a spectrum measured at w wavelengths. A class library set is symbolized by the $m \times w$ matrix $\mathbf{X}$ composed of m samples measured across the w variables. The transpose operation is indicated by a superscript t.

Although each library set does not typically have the same number of samples, the samples making up a library set need to span the variances making up the class. For example, spectra measured on samples of a specific plastic type (essentially pure component spectra) should capture the instrument profile as well as perhaps temperature effects. As another spectral example, if the goal is to identify an impurity present in a product, then each impurity spectral library set could span a concentration variance of that impurity. Described following is OPA in the TFA and NAS frameworks. The reader is referred to [1–4,19] for information on MD and KNN.

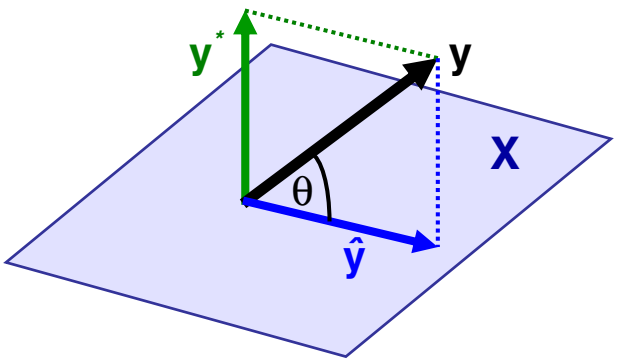

Figure 1. Orthogonal projection geometry for orthogonal projection analysis (target factor analysis and net analyte signal) where $\hat{\boldsymbol{y}}$ represents the projection of $\boldsymbol{y}$ (the test vector) into a particular library set $\mathbf{X}$, $\boldsymbol{y}^*$ denotes the orthogonal projection of $\boldsymbol{y}$, and the angle between $\boldsymbol{y}$ and the space spanned by $\mathbf{X}$ is symbolized by θ.

2.1. Orthogonal projection analysis

The orthogonal projection of the test sample vector $\boldsymbol{y}$ onto a space spanned by a library matrix $\mathbf{X}$ is obtained by

$$\boldsymbol{y}* = (\mathbf{I} - \mathbf{P})\boldsymbol{y} \tag{1}$$

where $\mathbf{I}$ is the $w \times w$ identity matrix, $\mathbf{P}$ represents a class projection matrix that projects onto the corresponding $\mathbf{X}$, $(\mathbf{I}-\mathbf{P})$ denotes the projection orthogonal to the span of $\mathbf{X}$, and $\boldsymbol{y}^*$denotes the resultant vector from the orthogonal projection of $\boldsymbol{y}$. The vector $\boldsymbol{y}^*$ can also be considered the residual vector after removing that part of $\boldsymbol{y}$ described by $\mathbf{X}$. Plotted in Figure 1 is a characterization of the orthogonal projection operation.

To obtain respective library class projection matrices $\mathbf{P}$, each particular library class matrix $\mathbf{X}$ is decomposed by a singular value decomposition (SVD) $\mathbf{X} = \mathbf{USV}^t$ where $\mathbf{U}$ represents the $m \times k$ matrix of left singular vectors (eigenvectors of $\mathbf{XX}^t$) with k being the mathematical rank of $\mathbf{X}$ (min(m,w)), $\mathbf{S}$ symbolizes the $k \times k$ diagonal matrix of singular values on the diagonal, and $\mathbf{V}$ denotes the $w \times k$ matrix of right singular vectors (eigenvectors of $\mathbf{X}^t\mathbf{X}$). Henceforth, the vectors in $\mathbf{V}$ shall be referred to as loading vectors. The loading vectors are used to calculate a projection matrix by $\mathbf{P} = \mathbf{VV}^t$. Because there is a total of k loading vectors for a particular $\mathbf{X}$, there are up to k projection matrices for that $\mathbf{X}$. Therefore, the success of OPA depends on the number of loading vectors used to form $\mathbf{P}$. It should be noted that the success of MD also depends on the number of loading vectors used in the MD calculation. Section 4 describes the effect of the number of loading vectors.

Traditional application of TFA uses a pure component spectrum of a chemical for $\boldsymbol{y}$ to target test its presence in one mixture set $\mathbf{X}$ where $\mathbf{X}$ is commonly obtained for a sample by processing the sample through a chromatographic system hyphenated with a spectral instrument. In TFA, the orthogonal projection is typically not used but the projection of $\boldsymbol{y}$ into $\mathbf{X}$ computed by $\hat{\boldsymbol{y}} = \mathbf{P}\boldsymbol{y}$ instead. If $\hat{\boldsymbol{y}}$ matches $\boldsymbol{y}$, that is, the Euclidian norm (two-norm) $\|\hat{\boldsymbol{y}} - \boldsymbol{y}\|_2$ (two-norm of the residual vector) is small, then the chemical is determined to be present in $\mathbf{X}$. Equivalently, using the orthogonal projection, the two-norm of $\boldsymbol{y}^*$ ($\|\boldsymbol{y}^*\|_2$) is small if the chemical is present in the sample. To date, the authors are not aware of the projection geometry of TFA being used as a simple single classification tool for multiclass situations.

The OPA process, and hence, TFA, can also be obtained by the NAS approach. Here, the angle (θ) between $\boldsymbol{y}$ and the $\mathbf{V}$ basis set spanning a particular $\mathbf{X}$ is obtained from

$$\sin\theta = \frac{\|\mathbf{y}*\|_2}{\|\mathbf{y}\|_2} \tag{2}$$

In the NAS literature, this ratio is known as the selectivity when $\mathbf{X}$ denotes a set of nonanalyte spectra and $\boldsymbol{y}$ represents a sample spectrum. As used here, the ratio denotes the fraction of the test sample remaining after the orthogonal projection and varies from 0 to 1. Ideally, if the test sample belongs to the class, the ratio is 0 (poor selectivity) and if the test sample does not belong to the class, the ratio is 1 (the best selectivity and $\boldsymbol{y}$ is unique compared to $\mathbf{X}$). For classification or library searching purposes, the ratio for a test sample is determined for each library set and the ratio closest to 0 would identify the class it belongs to. Rather than using closeness to 0, the approach here is to use the angle computed by

$$\theta = \sin^{-1}\left(\frac{\|\mathbf{y}*\|_2}{\|\mathbf{y}\|_2}\right) \quad (3)$$

The class with the smallest θ indicates the class membership for $\boldsymbol{y}$.

Equivalently, the cosine of the angle θ in Figure 1 can be computed by

$$\cos\theta = \frac{\|\hat{\boldsymbol{y}}\|_2}{\|\boldsymbol{y}\|_2} \quad (4)$$

followed by the appropriate mathematics to obtain θ. The cos θ value varies from 1 to 0. Ideally, if the test sample belongs to the class, the ratio is 1, and if the test sample does not belong to the class, the ratio is 0. The cos θ value is commonly used in traditional spectral library searching. In this approach, a test sample spectrum is sequentially compared with individual pure component library spectra. The OPA approach described here is not the same. Specifically, the OPA angle is obtained between the test sample spectrum and the space spanned by a library set, for example, a set of spectra measured for a pure component substance. In traditional library searching, the angle is obtained between the test sample spectrum and one pure component spectrum. The advantage of using a library set of spectra for a pure component substance is exemplified in the plastic data set.

2.2. Equivalency of OPA to another angular measure

Details of another angle measure are described in [16], and a brief outline is provided here. Let k be the rank of $\mathbf{X}$, and the rank of $\boldsymbol{y}$ is 1. When the angular relationship between two data sets (two matrices of respective data) is sought, the normal process involves computing individual SVDs of the two spaces being compared. In this paper, one of the data sets, $\boldsymbol{y}$, is a vector, not a matrix. Writing $\boldsymbol{y}$ as a row vector, the SVD $\boldsymbol{y}^t = u_y s_y \mathbf{v}_y^t$ results in u_y, s_y, and $\mathbf{v}_y$ having dimensions 1×1, 1×1, and $w \times 1$, respectively, with $\mathbf{v}_y$ being $\boldsymbol{y}$ normalized to unit length and $s_y = \|\boldsymbol{y}\|_2$. For clarification, the SVD of $\mathbf{X}$ is notated as $\mathbf{X} = \mathbf{U}_\mathbf{X}\mathbf{S}_\mathbf{X}\mathbf{V}_\mathbf{X}^t$ where, as with OPA, $\mathbf{U}_\mathbf{X}$, $\mathbf{S}_\mathbf{X}$, and $\mathbf{V}_\mathbf{X}$ are $w \times k$, $k \times k$, and $w \times k$, respectively. In the original development, angles between respective $\mathbf{U}$ and $\mathbf{V}$ spaces from the two SVDs could be computed, for example, angles between the respective $\mathbf{U}$ and $\mathbf{V}$ spaces of two spectrochromatograms $\mathbf{Y}$ and $\mathbf{X}$ [18]. Such an analysis between spectrochromatograms provides two angular relationships, one each for the chromatographic ($\mathbf{U}$) space and the other for the spectral ($\mathbf{V}$) space. Because $\boldsymbol{y}$ is a vector in this paper, there is only one angle to compute, the angle between $\mathbf{v}_y$ ($\boldsymbol{y}$ normalized to unit length) and the space spanned by $\mathbf{V}_\mathbf{X}$. Because up to k loading vectors can be used for $\mathbf{V}_\mathbf{X}$, there are up to k angles that can be determined.

An angle is obtained by first computing the $k \times 1$ vector $\mathbf{m} = \mathbf{V}_\mathbf{X}^t\mathbf{v}_\mathbf{y}$. Because $\mathbf{v}_y$ and the vectors in $\mathbf{V}_\mathbf{X}$ have unit length, then $\mathbf{m}$ contains the cos θ between $\mathbf{v}_y$ and each vector in $\mathbf{V}_\mathbf{X}$. An SVD is now performed on $\mathbf{m}$ giving $\mathbf{m} = \mathbf{u}_\mathbf{m} s_\mathbf{m} v_\mathbf{m}^t$ with the one singular value $s_\mathbf{m}$, representing the cosine of the angle between $\mathbf{v}_y$ and the space spanned by $\mathbf{V}_\mathbf{X}$, that is, cos $\theta = s_\mathbf{m}$. The angle is then obtained from $\theta = \cos^{-1}(s_\mathbf{m})$. Because $\mathbf{m}$ is a vector, $s_\mathbf{m} = \|\mathbf{m}\|_2$, and hence, cos $\theta = \|\mathbf{m}\|_2$. When used for classification of a test vector $\mathbf{y}$ instead of a test matrix $\mathbf{Y}$, the approach collapses to that of OPA. This is shown in the following.

The angle relationship given in Equation (4) for Figure 1 can be expanded to

$$\cos\theta = \frac{\sqrt{\hat{\boldsymbol{y}}^t\hat{\boldsymbol{y}}}}{\|\boldsymbol{y}\|_2} = \frac{\sqrt{\boldsymbol{y}^t\mathbf{V}_\mathbf{X}\mathbf{V}_\mathbf{X}^t\mathbf{V}_\mathbf{X}\mathbf{V}_\mathbf{X}^t\boldsymbol{y}}}{\|\boldsymbol{y}\|_2} = \frac{\sqrt{\boldsymbol{y}^t\mathbf{V}_\mathbf{X}\mathbf{V}_\mathbf{X}^t\boldsymbol{y}}}{\|\boldsymbol{y}\|_2} = \frac{\sqrt{\|\mathbf{V}_\mathbf{X}^t\boldsymbol{y}\|_2^2}}{\|\boldsymbol{y}\|_2} = \frac{\|\mathbf{V}_\mathbf{X}^t\boldsymbol{y}\|_2}{\|\boldsymbol{y}\|_2} = \|\mathbf{V}_\mathbf{X}^t\mathbf{v}_\mathbf{y}\|_2 = \|\mathbf{m}\|_2 \quad (5)$$

resulting in

$$\theta = \cos^{-1}(\|\mathbf{m}\|_2) \quad (6)$$

If the test vector $\boldsymbol{y}$ is normalized to unit length for OPA, then Equation (3) becomes

$$\theta = \sin^{-1}(\|\boldsymbol{y}*\|_2) \quad (7)$$

and the two angles are equal.

3.7. Determining the number of loading vectors

The accuracy of OPA and MD depend on the number of loading vectors. Numerous approaches have been developed to select the number of loading vectors [24,25]. The focus of this paper is not to compare these methods. Instead, the same process is used for OPA and MD. The procedure used in this study is based on a newly developed method named determination of rank by augmentation (DRAUG) [26]. The process determines the minimum number of loading vectors needed to span a space, that is, the number needed to properly characterize a library set $\mathbf{X}$. The DRAUG methodology distinguishes primary loading vectors (chemical, instrumental, etc.) from secondary loading vectors (experimental errors) independent of the distribution of experimental uncertainties. Reference [26] has the details and Matlab code.

3. EXPERIMENTAL

3.1. Software

Programs for OPA and MD were written by the authors using MATLAB 2010b (The MathWorks, Natick, MA). The published program DRAUG was used for determining the number of loading vectors [26]. The MATLAB Statistics Toolbox was used for KNN with the Euclidean distance.

3.2. Plastic

The plastic identification data set consists of six classes that are six of the seven commercial plastic types (numbers 1–6) [21]. Samples were measured using Raman spectroscopy over the wavelength range 850–1800 cm^{-1} consisting of 1093 wavelengths per spectrum. Classes one through six have 30, 29, 13, 22, 23, and 29 samples, respectively, corresponding to plastic types polyethylene terephthalate, high-density polyethylene, polyvinyl chloride, low-density polyethylene, polypropylene, and polystyrene. Data was used as measured without any preprocessing.

3.3. Archeological

The archeological data set has four classes, a class for a different obsidian source [20]. This benchmark data set is often used in classification studies. Samples are measured using X-ray fluorescence spectroscopy for the analysis of 10 trace metals. The 10 metals are Fe, Ti, Ba, Ca, K, Mn, Rb, Sr, Y, and Zr. Concentrations of each

metal ranged from 40 to 1000 ppm. The classes have 10, 9, 23, and 21 samples. Data was used as measured without any preprocessing.

3.4. Gasoil

The gasoil data set has three classes corresponding to three gasoil sources [22]. Samples were measured over the wavelength range of 200–400 nm for a total of 572 wavelengths per spectrum. The classes (sources) have 59, 25, and 30 samples, respectively. Data was used as measured without any preprocessing.

3.5. Extra virgin olive oil

The EVOO data set consists of six classes [23]. Each class is a set of EVOO samples that has been adulterated with different oils. The oils are corn, olive-pomace, soybean, sunflower, rapeseed, and walnut oil. In each of the classes, adulterant concentrations range from 0.5% to 95% with 31 samples measured in each class except for the sunflower oil class that has 30 samples. Samples were measured using synchronous fluorescence spectroscopy across the wavelength range 250–400 nm at Δ20-nm difference. Each spectrum is measured over 151 wavelengths. Data was used as measured without any preprocessing.

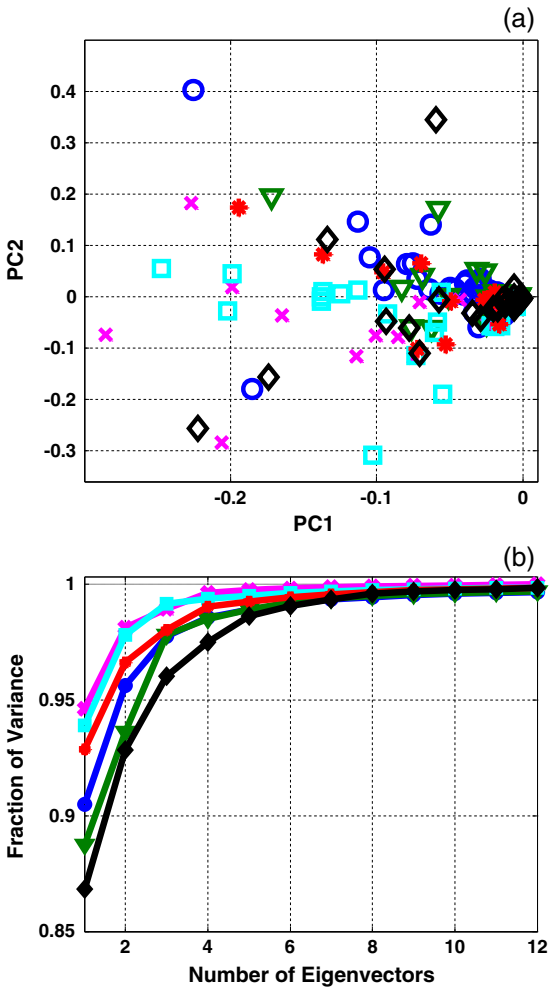

Figure 2. The principal component analysis characterization of the plastic data set. (a) Score plot using the first two principal components and (b) scree plot showing the cumulative fraction of total variance explained for each plastic-type library. Plastic types are (blue, circle) type 1, (green, upside down triangle) type 2, (magenta, x) type 3, (cyan, square) type 4, (red, asterisk) type 5, and (black, diamond) type 6. PC, principal component.

3.6. Cross-validation classification process

Leave one out cross-validation (LOOCV) was used to test each of the three methods [1,2]. Briefly, for a particular data set, a test sample is removed from a library set. The OPA angle and MD are computed for the removed sample relative to the library set it belongs to and all other library sets in the particular data set. The angle and MD values are obtained from one loading vector to the minimum library rank defined by the library set of the corresponding data set with the smallest rank. The sample is replaced, and the process repeats until each sample in a library set has acted as the test sample. The process is then repeated for each library set in the data set. The same LOOCV process was used for KNN using the Euclidean distance with majority vote and varying the number of neighbors from 1 to 11.

It is important to note that LOOCV is not used to determine the number of loading vectors for OPA and MD. The DRAUG approach is used for this purpose. The LOOCV is also not used to determine the best number of neighbors to use in KNN. In this case, the best number of neighbors is not determined, and only classification trends are studied by varying the number of

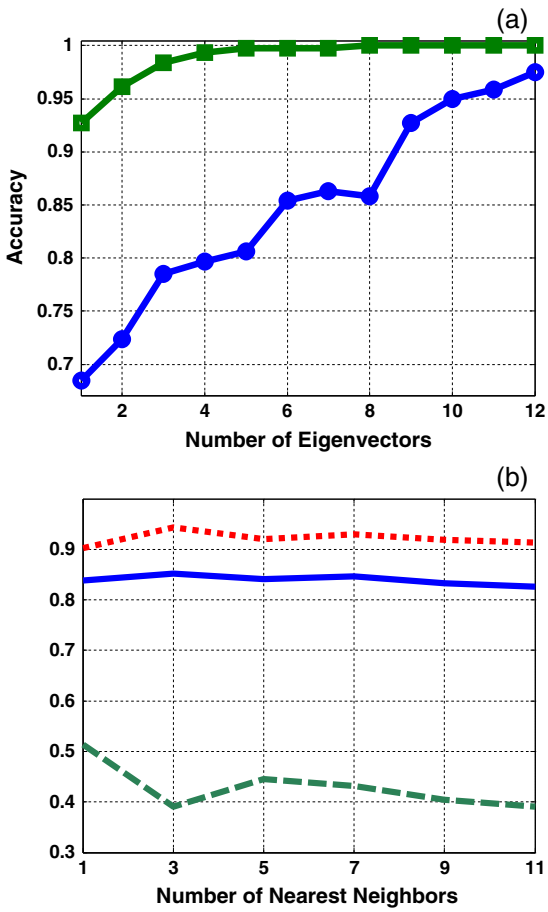

Figure 3. (a) Overall accuracy values for all plastic types using classification methods (green, circles) orthogonal projection analysis and (blue, circles) Mahalanobis distance. (b) Overall *k*-nearest neighbors accuracy (blue), sensitivity (dash green), and specificity (dot red).

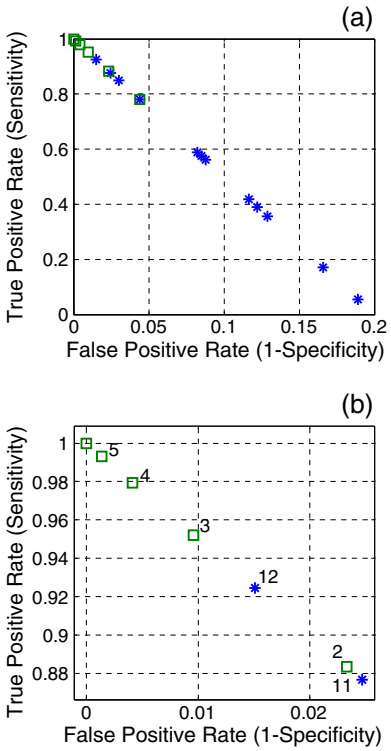

Figure 4. The receiver operator characteristic plot for the plastic data set. Classification methods are denoted by (green squares) orthogonal projection analysis and (blue asterisks) Mahalanobis distance. (a) All loading vectors and (b) zoom of (a). Numbers in (b) represent the number of loading vectors. The overall receiver operator characteristic plot across all plastic types for each method is shown.

neighbors. These trends are compared with classification results from OPA and MD as well as the trends obtained by varying the number of eigenvectors for OPA and MD. As described in the experimental section, some of the library sets are small (9, 10, and 13 samples), and hence, LOOCV is used.

3.7. Classification assessment

The classification performance [27,28] of each method was assessed on the basis of the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) as the number of loading vectors range from one to the minimum overall library rank for a particular data set. If a sample is classified belonging to a library set and it belongs to the class, it is a TP. If a sample is classified as not belonging to a library set and it does not belong to the class, it is a TN. If a sample is classified as belonging to a library set and it does not belong to that class, it is an FP. Lastly, if a sample is classified as not

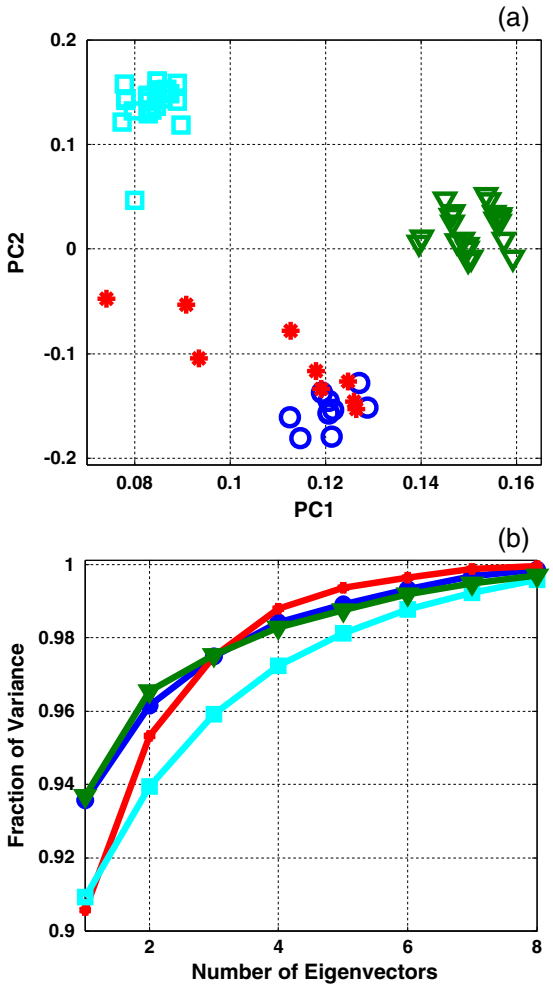

Figure 5. The principal component analysis characterization of the archeological data set. (a) Score plot using the first two principal components (PCs) and (b) scree plot showing the cumulative fraction of total variance explained for each source library. Sources are (blue, circle) source 1, (red, asterisk) source 2, (cyan, square) source 3, and (green upside down triangle) source 4.

Table 1. Accuracy, sensitivity, and specificity values for the plastic data set

Library plastic[1]	Accuracy (%)		Sensitivity (%)		Specificity (%)	
	OPA	MD	OPA	MD	OPA	MD
Type 1 (9)	100	94	100	83	100	97
Type 2 (9)	100	97	100	93	100	99
Type 3 (4)	100	85	100	54	100	91
Type 4 (6)	100	86	100	59	100	92
Type 5 (9)	100	78	100	35	100	87
Type 6 (11)	100	98	100	93	100	99

Values are broken down by each library set (plastic type).
OPA, orthogonal projection analysis; MD, Mahalanobis distance.
[1]Values in parentheses are determination of rank by augmentation loading vector number rounded to the nearest whole number.

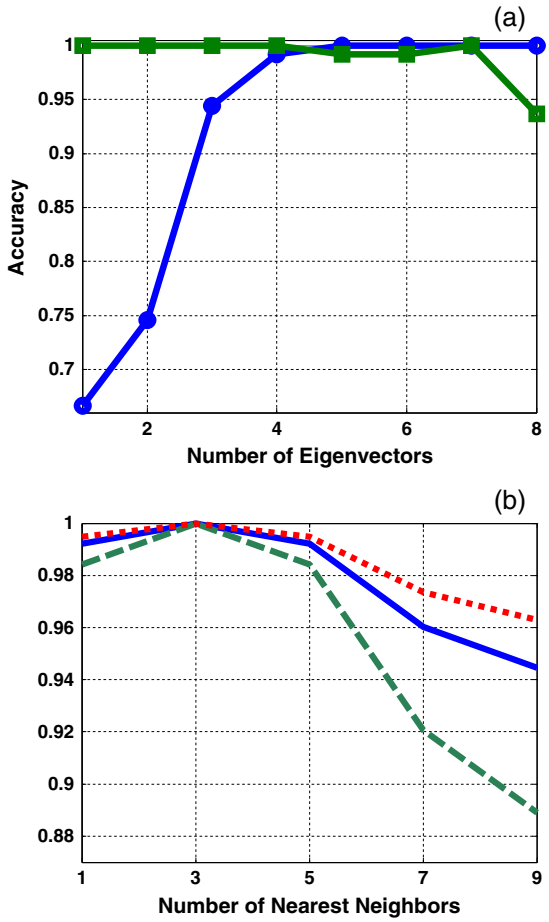

Figure 6. (a) Overall accuracy values for all archeological classes using classification methods (green, circles) orthogonal projection analysis and (blue, circles) Mahalanobis distance. (b) Overall *k*-nearest neighbors accuracy (blue), sensitivity (dash green), and specificity (dot red).

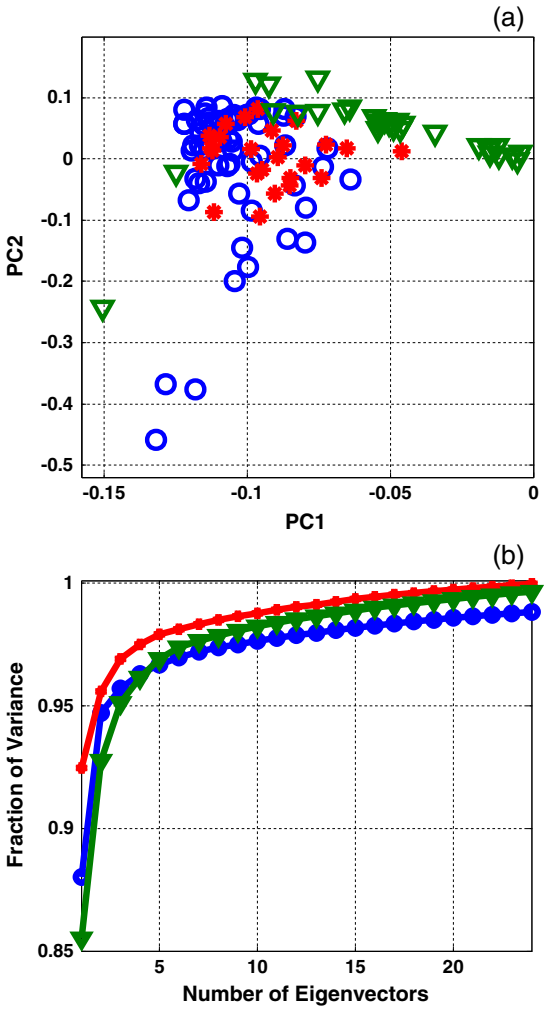

Figure 7. The principal component analysis characterization of the gasoil data set. (a) Score plot using the first two principal components (PCs) and (b) scree plot showing the cumulative fraction of total variance explained for each source library. Sources are (blue, circle) source 1, (red, asterisk) source 2, and (green, upside down triangle) source 3.

belonging to a library set and it does belong to that class, it is an FN. Classification performance is then evaluated by the accuracy term [28,29] computed by

$$accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (8)$$

for the respective number of loading vectors or neighbors.

In addition to plotting the accuracy as a function of the number of loading vectors or neighbors, the receiver operator characteristic (ROC) plot can be used to graphically present the classification behavior as the number of loading vectors or neighbors vary. The ROC plot shows the separation ability of a binary classifier by iteratively setting the classifier thresholds [30,31]. For the studies presented in this paper, an ROC plot is obtained by plotting the TP rate (sensitivity, SE) against the FP rate (1-specificity, SP) for each set of loading vectors or neighbors

Table 2. Accuracy, sensitivity, and specificity values for the archeological data set

Library source[1]	Accuracy (%)		Sensitivity (%)		Specificity (%)	
	OPA	MD	OPA	MD	OPA	MD
Source 1 (2)	100	80	100	60	100	87
Source 2 (4)	100	100	100	100	100	100
Source 3 (4)	100	98	100	96	100	99
Source 4 (3)	100	100	100	100	100	100

Values are broken down by each library set (source).
OPA, orthogonal projection analysis; MD, Mahalanobis distance.
[1]Values in parentheses are determination of rank by augmentation loading vector number rounded to the nearest whole number.

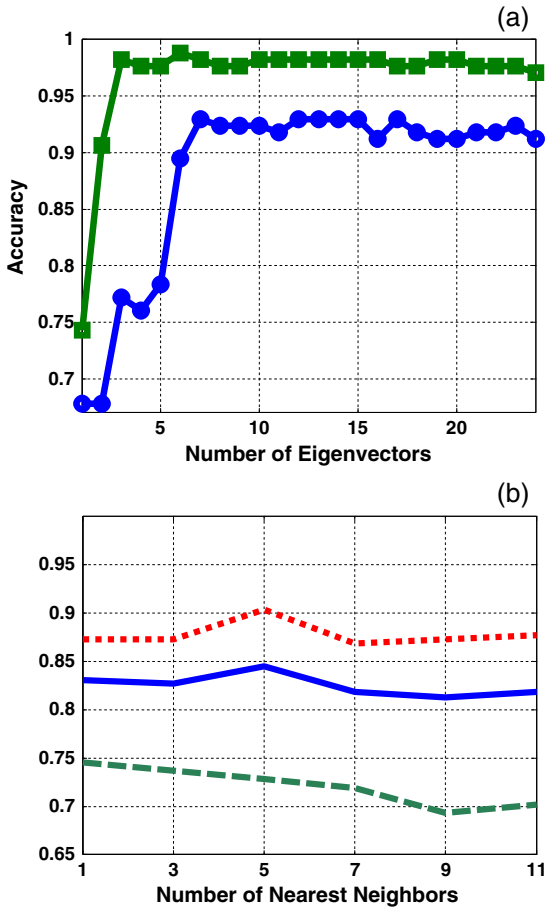

Figure 8. (a) Overall accuracy values for all gasoil classes using classification methods (green, circles) orthogonal projection analysis and (blue, circles) Mahalanobis distance. (b) Overall *k*-nearest neighbors accuracy (blue), sensitivity (dash green), and specificity (dot red).

where SE and SP are computed by

$$SE = TP/(TP + FN) \tag{9}$$

$$SP = TN/(TN + FP) \tag{10}$$

In this study, OPA angle and MD thresholds are not varied. As noted previously, respective classification of a sample in this paper is based on the library set with the smallest angle, smallest MD, and majority vote of nearest neighbors. Thus, threshold values for the OPA and MD ROC plots are the number of loading vectors.

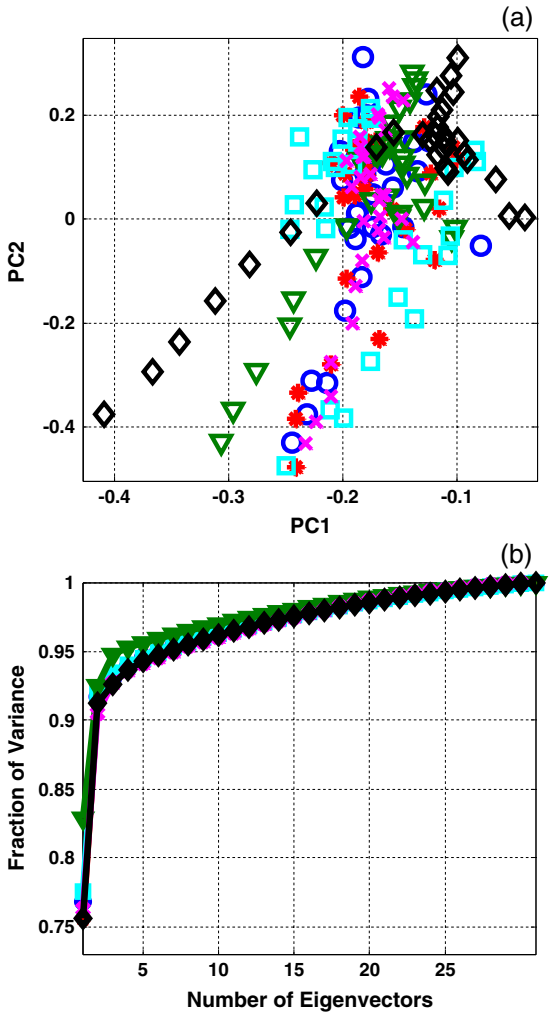

Figure 9. The principal component analysis characterization of the extra virgin olive oil data set. (a) Score plot using the first two principal components (PCs) and (b) scree plot showing the cumulative fraction of total variance explained for each source library. Adulterant oils are (blue, circle) corn, (red, asterisk) olive-pomace, (green, upside triangle) rapeseed, (cyan, square) soybean, (magenta, x) sunflower, and (black, diamond) walnut.

4. RESULTS AND DISCUSSION

Accuracy, SE, and SP are tabulated classwise for OPA and MD on the basis of the number of loading vectors determined by DRAUG.

Table 3. Accuracy, sensitivity, and specificity values for the gasoil data set

Library source[1]	Accuracy (%)		Sensitivity (%)		Specificity (%)	
	OPA	MD	OPA	MD	OPA	MD
Source 1 (11)	100	100	100	100	100	100
Source 2 (8)	95	89	92	84	96	92
Source 3 (11)	98	82	97	73	98	87

Values are broken down by each library set (source).
OPA, orthogonal projection analysis; MD, Mahalanobis distance.
[1]Values in parentheses are determination of rank by augmentation loading vector number rounded to the nearest whole number.

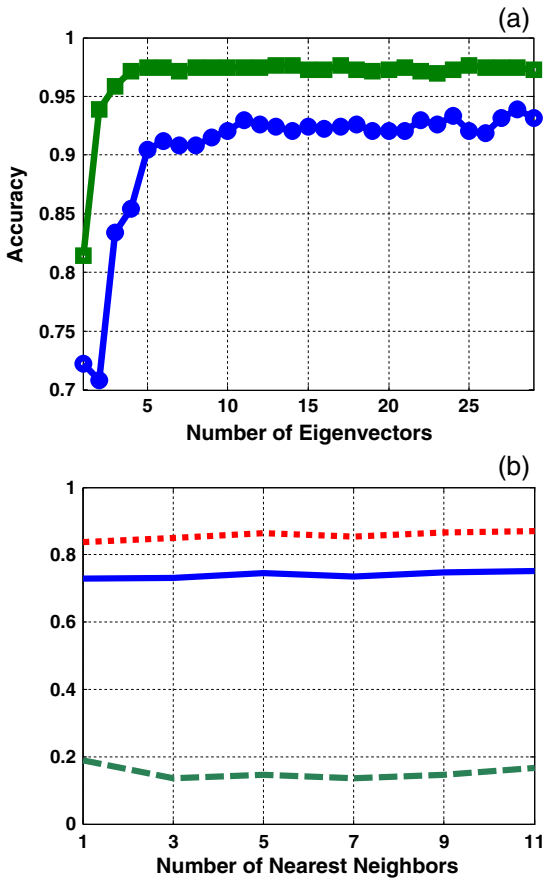

Figure 10. (a) Overall accuracy values for all extra virgin olive oil adulterants using classification methods (green, squares) orthogonal projection analysis and (blue, circles) Mahalanobis distance. (b) Overall *k*-nearest neighbors accuracy (blue), sensitivity (dash green), and specificity (dot red).

The overall OPA and MD accuracies for each data set are also plotted as a function of the number of loading vectors. Because no method is used to identify the optimal number of neighbors for KNN, the overall accuracy, SE, and SP are plotted for each data set. The focus of the paper is not to evaluate the best way to determine the number of eigenvectors, number of nearest neighbors, or the best way to perform classification. The primary goal is to assess the ability of OPA to act as a simple stand-alone classification tool given the same circumstance for all data sets.

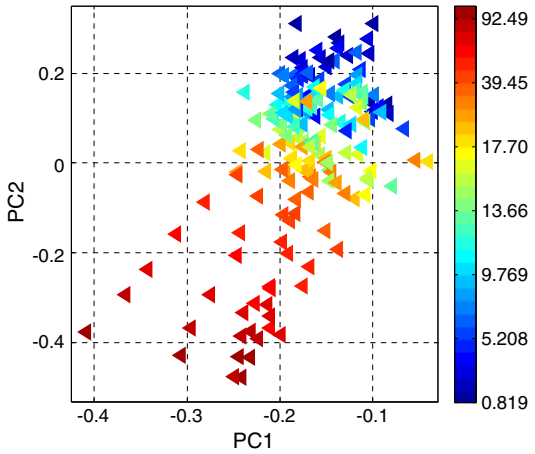

Figure 11. Extra virgin olive oil score plot as in Figure 9 except that points are now color coded to the respective adulterant concentration as indicated in the Figure. PC, principal components.

4.1. Plastic

Shown in Figure 2 are the score and scree plots from principal component analysis. The score plot in Figure 2a reveals that the plastic types do not uniquely cluster out within the first two principal components (PCs). The scree plot in Figure 2b for each library set of spectra identifies over 90% of the variation being captured with the first two PCs, and the first PC provides most of it. The scree plot is shown for each library set because the number of loading vectors for OPA and MD are library set specific. In [21], score plots show unique clusters only after preprocessing the data with second derivatives. Simple classification tools could then be used including KNN. As Figure 2 reveals, using the raw data should make the classification more difficult.

Shown in Figure 3a are the accuracy plots for OPA and MD as the number of loading vectors vary. From Figure 3a, it is observed that MD is never able to achieve 100% accuracy,

Table 4. Accuracy, sensitivity, and specificity values for the extra virgin olive oil data set

Library oil[1]	Accuracy (%)		Sensitivity (%)		Specificity (%)	
	OPA	MD	OPA	MD	OPA	MD
Corn (8)	98	89	93	68	99	94
Olive-pomace (4)	100	92	100	77	100	95
Rapeseed (6)	93	88	81	65	96	93
Soybean (6)	100	93	100	80	100	96
Sunflower (6)	97	87	90	61	98	92
Walnut (4)	99	84	97	51	99	90

Values are broken down by each library set (adulterant oil). OPA, orthogonal projection analysis; MD, Mahalanobis distance. [1]Values in parentheses are determination of rank by augmentation loading vector number rounded to the nearest whole number.

Table 5. Minimum adulterant concentration correctly classified for the extra virgin olive oil data set

Library oil[1]	Minimum adulterant concentration (%)	
	OPA	MD
Corn (8)	1.74	14.91
Olive-pomace (4)	0.85	14.53
Rapeseed (6)	15.50	20.64
Soybean (6)	1.05	17.06
Sunflower (6)	4.03	18.62
Walnut (4)	0.82	21.23

Values are broken down by each library set (adulterant oil). OPA, orthogonal projection analysis; MD, Mahalanobis distance. [1]Values in parentheses are determination of rank by augmentation eigenvector number rounded to the nearest whole number.

whereas it is possible with OPA. The ROC plot in Figure 4 also shows that MD does not classify as well as OPA. For example, 11 and 12 loading vectors are used (12 is the maximum available) with MD, whereas OPA uses two and three loading vectors to obtain the similar TP and FP rates. As a reminder, Figure 4 deviates from a traditional ROC plot that monotonically increases as the FP rate increases. Hence, only points are shown using the number of loading vectors for the ROC threshold values.

Tabulated in Table 1 are the accuracies, sensitivities, and specificities broken down by each library set (plastic type) at the number of loading vectors determined best by the DRAUG. As expected from Figures 3a and 4, OPA outperforms MD. For MD, Figures 3 and 4 indicate that all the loading vectors should be used to obtain the best accuracy. However, the MD accuracy with all the loading vectors is still not as good as OPA.

The overall accuracy, SE, and SP plot for KNN as the number of neighbors vary is shown in Figure 3b. Figure 3b reveals that regardless of the number of neighbors, the accuracy is worse than OPA. In [21], spectra need to be preprocessed by second derivative in order for KNN to correctly classify all plastic samples. Because no preprocessing was performed on the plastic Raman spectra in this study, the simple OPA approach appears to be more useful.

Lastly, in previous work with this data set, a traditional library search with cos θ was used where each plastic sample spectrum was matched to one representative plastic-type spectrum from each library set. The results were deficient. A similar approach was used without second derivative preprocessing in this study. Specifically, rather than using a representative spectrum from each library set, individual cos θ values were instead obtained in the LOOCV process between the test sample spectrum and the spectra making up each library class. The classwise mean cos θ values were compared, and the library set with mean cos θ value closest to 1 was identified as the test sample plastic type. With this approach, the results were even worse (not shown), and hence, traditional library searching is not applicable.

4.2. Archeological

Score and scree plots shown in Figure 5 indicate that the third and fourth classes are well separated from each other and the first and second classes. However, the first and second classes slightly overlap. The first two PCs characterize over 90% of the concentration information with most of that coming with the first PC. Displayed in Figure 6a is the accuracy plot for OPA and MD. The accuracy plot reveals a small difference from the plastic accuracy plots in that OPA degrades as more loading vectors beyond four are used while MD continues to improve. Listed in Table 2 are the results based on the DRAUG-determined number of loading vectors. Because DRAUG identifies a small number of loading vectors as best for each library class, OPA outperforms MD. If a different loading vector selection approach were used [24,25], it may be that a greater number would be identified and MD would now perform slightly better. Again, the focus of this paper is not to compare methods for determining the number of loading vectors. Trends in the ROC plots follow that of the accuracy plots. Specifically, the TP and FP rates for OPA degrade after four loading vectors are included.

The overall accuracy, SE, and SP plots for KNN as the numbers of neighbors vary are shown in Figure 6b. The Figure indicates that class identification by KNN is not significantly affected by the number of neighbors until larger numbers are used. This probably stems from the good class separation shown in Figure 5a. All three approaches work equally well with this data set.

4.3. Gasoil

The score plot in Figure 7a shows that class clustering is poor. As with the previous two data sets, the scree plot in Figure 7b specifies that over 90% of the spectral variance is described by the first two PCs for each class. The accuracy plots in Figure 8 and tabulated values in Table 3 demonstrate that OPA again outperforms MD. Similar to the plastic data, the accuracy plots in Figure 8a show that MD does not do as well as OPA regardless of the number of loading vectors. Unlike the plastic and archeological data, 100% accuracies by OPA are not obtained for the classes. Figure 8b reveals that KNN does not classify as well as OPA or MD regardless of the number of neighbors.

4.4. Extra virgin olive oil

Clustering of the adulterants is poor as demonstrated by the score plot in Figure 9a. Similar to the previous three data sets, the scree plot in Figure 9b shows that over 90% of the spectral variance is described by the first two PCs. Accuracy plots in Figure 10a and tabulated values in Table 4 pattern those of the gasoil data demonstrating that OPA again outperforms MD. For OPA, the poorer performance with rapeseed oil is due to the lower SE value. Similar observations can be made for the low accuracy values with MD. The accuracy plots in Figure 10a show that MD does not do as well as OPA regardless of the number of loading vectors. The KNN results plotted in Figure 10b reveal that KNN does not perform as well as OPA and MD regardless of the number neighbors used.

The EVOO data is different from the other three data sets in that adulterant concentration values are available. Plotted in Figure 11 is the same score plot in Figure 9a except the points are color coded to respective adulterant concentrations. Listed in Table 5 are the minimum adulterant concentrations that could be correctly classified. The concentrations range from 0.85% to 15.50% for OPA and from 14.53% to 21.24% for MD. The adulterant with the worse accuracy for both methods is rapeseed oil. Samples incorrectly classified are those in the lower-concentration range. Otherwise, adulterants can be accurately classified at the lower concentrations with OPA, a desired ability in the EVOO adulteration problem.

5. CONCLUSION

This paper showed that the OPA methods of TFA and NAS are essentially the same with TFA being residual based and NAS is angle based. Results from this study on a variety of data sets demonstrated that the OPA in TFA or NAS format generally outperforms MD and KNN, conventional approaches to classification problems. When score plots do not mark clear clusters, the TFA and NAS measures always performed better than MD and KNN. The OPA approaches, MD, and KNN require selection of tuning parameters. From the results, it appears that DRAUG performs well at this task for OPA and MD. The focus of the paper is not on evaluating the abundance of methods to identify the optimal tuning parameters. Instead, accuracy trends were also plotted across the tuning parameters. From these plots, it was ascertained that even if the optimal set of tuning parameters were determined, the OPA process performed best overall.

The OPA process can be generalized to *N*th-order data [18,32]. Thus, the TFA and NAS are also generalizable to *N*th-order data.

The focus in this paper is on a vector for the test sample being projected orthogonally (or into) a library set (a matrix).

Lastly, none of the data was preprocessed in any way. From previous work with the plastic data set, it was necessary to preprocess the Raman spectra with second derivatives to correctly classify plastic samples. With OPA, 100% correct identification was possible without preprocessing. Additionally, in previous work with the plastic data set, a traditional library searching approach used cos θ to match each sample spectrum to a representative library spectrum, and the results were not acceptable. Similar poor results were obtained in this study with a traditional library search using raw spectra indicating the advantage of using OPA with a set of library spectra for a chemical substance.

Acknowledgements

This material is based upon work supported by the National Science Foundation under grant no. CHE-0715149 (cofunded by the MPS Chemistry and DMS Statistical Divisions and the NSPA Program) and by the University Research Committee under grant no. SU11-7U at Idaho State University, Pocatello, Idaho, and is gratefully acknowledged by the authors. The authors are thankful to the authors of [23] for providing the data.

REFERENCES

1. Næs T, Isaksson T, Fern T, Davies T. *A User Friendly Guide to Multivariate Calibration and Classification*. NIR Publications: Chichester, UK, 2002.
2. Hastie TJ, Tibshirani RJ, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd edn). Springer-Verlag: New York, 2009.
3. Brereton RG. *Chemometrics for Pattern Recognition*. Wiley: Chichester, UK, 2009.
4. Lavine BK, Rayens WS. Classification: basic concepts. In *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis Vol. 3*, Brown SD, Tauler R, Walczak B (eds.). Elsevier: Amsterdam, 2009; 507–515.
5. McCue M, Malinowski ER. Target factor analysis of ultraviolet spectra of unresolved liquid chromatographic fractions. *Appl. Spectrosc.* 1983; **37**: 463–469.
6. Malinowski ER. *Factor Analysis in Chemistry* (3rd edn). Wiley: New York, 2002.
7. Morgan DR. Spectral absorption pattern detection and estimation. I. analytical techniques. *Appl. Spectrosc.* 1977; **51**: 404–415.
8. Lorber A. Error propagation and figures of merit for quantification by solving matrix equations. *Anal. Chem.* 1986; **58**: 1167–1172.
9. Zeaiter M, Rutledge D. Preprocessing methods. In *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis Vol. 3*, Brown SD, Tauler R, Walczak B (eds.). Elsevier: Amsterdam, 2009; 121–231.
10. Sánchez FC, Toft J, van den Bogart B, Massart DL. Orthogoanl projection approach applied to peak purity assessment. *Anal. Chem.* 1996; **68**: 79–85.
11. Xie YL, Kalivas JH. Use of matrix orhtogoanl projection peak purity assessment. *Anal. Lett.* 1997; **30**: 395–416.
12. Zhang P, Littlejohn D. Mathematical prediction and correction of interferences for optimization of line selection in inductively coupled plasma optical emission spectrometry. *Spectrochim. Acta* 1993; **48B**: 1517–1555.
13. Ruyken MMA, Visser JA, Smilde AK. On-line detection and identification of interferences in multivariate predictions of organic gases using FT-IR spectroscopy. *Anal. Chem.* 1995; **67**: 2170–2179.
14. Skibsted ETS, Boelens HFM, Westerhuis JA, Smilde AK, Broad NW, Rees DR, Witte DT. Net analyte signal based statistical quality control. *Anal. Chem.* 2005; **77**: 7103–7114.
15. Wold S, SØstrØm M. SIMCA. A method for analyzing chemical data in terms of similarity and analogy. In *Chemometrics: Theory and Publications*, Kowalski BR (ed.). American Chemical Society: Washington DC, USA, 1977; 243–282.
16. Krzanowski WJ. Between-group comparison of principal components. *J. Am. Stat. Assoc.* 1979; **74**: 703–707.
17. Carlosens A, Anrade JM, Kubista M, Prada D. Procrustes rotation as a way to compare different sampling seasons in soils. *Anal. Chem.* 1995; **67**: 2373–2378.
18. Anderson CE, Nieves RG, Kalivas JH. Orthogonality considerations for library searching Nth-order data. *Chemometr. Intell. Lab. Syst.* 1998; **41**: 115–125.
19. Standard Practices for Infrared, Multivariate, Quantitative Analysis, E 1655–94, American Society for Testing and Materials: Philadelphia, 1995.
20. Kowalski BR, Schatzki TF, Stross FH. Classification of archaeological artifacts by applying pattern recognition to trace element data. *Anal. Chem.* 1972; **44**: 2176–2180.
21. Allen V, Kalivas JH, Rodriguez RG. Post-consumer plastic identification using Raman spectroscopy. *Appl. Spectrosc.* 1999; **53**: 672–681.
22. Wentzell P, Andrews D, Walsh J, Cooley J, Spencer P. Estimation of hydrocarbon types in light gas oils and diesel fuels by ultraviolet absorption spectroscopy and multivariate calibration. *Can. J. Chem.* 1999; **77**: 391–400.
23. Poulli KI, Mousdis GA, Georgiou CA. Rapid synchronous fluorescence method for virgin olive oil adulteration assessment. *Food Chem.* 2007; **105**: 369–375.
24. Wasim M, Brereton RG. Determination of the number of significant components in liquid chromatography nuclear magnetic resonance spectroscopy. *Chemometr. Intell. Lab. Syst.* 2004; **72**: 133–151.
25. Meloun M, Čapek J, Mikšik P, Brereton RG. Critical comparison of methods predicting the number of components in spectroscopic data. *Anal. Chim. Acta* 2000; **423**: 51–68.
26. Malinowski ER. Determination of rank by augmentation. *J. Chemom.* 2011; **25**: 323–328.
27. Trullols E, Ruisánchez I, Rius FX. Validation of qualitative analytical methods. *Trends Anal. Chem.* 2004; **23**: 137–145.
28. Myatt GJ, Johnson WP. *Making Sense of Data II: A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications*. Wiley: Hoboken, New Jersey, 2009.
29. Rizzi A, Fioni A. Virtual screening using PLS discriminant analysis and ROC curve approach: an application study on PDE4 inhibitors. *J. Chem. Inf. Model.* 2008; **48**: 1686–1692.
30. Brown CD, Davis HT. Receiver operating characteristics curves and related decision measures: a tutorial. *Chemometr. Intell. Lab. Syst.* 2006; **80**: 24–38.
31. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 2000; **16**: 412–424.
32. Messick NJ, Kalivas JH, Lang PM. Selectivity and related measures for nth-order data. *Anal. Chem.* 1996; **68**: 1572–1579.