International Journal of Chemistry; Vol. 10, No. 2; 2018
ISSN 1916-9698 E-ISSN 1916-9701
Published by Canadian Center of Science and Education

An Improved Comparison of Chemometric Analyses for the Identification of Acids and Bases With Colorimetric Sensor Arrays

Michael James Kangas[1], Christina L Wilson[1], Raychelle M Burks[2], Jordyn Atwater[1], Rachel M Lukowicz[1], Billy Garver[1], Miles Mayer[1], Shana Havenridge[1], Andrea E Holmes[1]

[1]Doane University, USA

[2]St. Edwards University, USA

Correspondence: Michael James Kangas, Doane University, USA.

Received: March 2, 2018 Accepted: April 25, 2018 Online Published: April 25, 2018

doi:10.5539/ijc.v10n2p36 URL: https://doi.org/10.5539/ijc.v10n2p36

Abstract

Colorimetric sensor arrays incorporating red, green, and blue (RGB) image analysis use value changes from multiple sensors for the identification and quantification of various analytes. RGB data can be easily obtained using image analysis software such as ImageJ. Subsequent chemometric analysis is becoming a key component of colorimetric array RGB data analysis, though literature contains mainly principal component analysis (PCA) and hierarchical cluster analysis (HCA). Seeking to expand the chemometric methods toolkit for array analysis, we explored the performance of nine chemometric methods were compared for the task of classifying 631 solutions (0.1 to 3 M) of acetic acid, malonic acid, lysine, and ammonia using an eight sensor colorimetric array. PCA and LDA (linear discriminant analysis) were effective for visualizing the dataset. For classification, linear discriminant analysis (LDA), (k nearest neighbors) KNN, (soft independent modelling by class analogy) SIMCA, recursive partitioning and regression trees (RPART), and hit quality index (HQI) were very effective with each method classifying compounds with over 90% correct assignments. Support vector machines (SVM) and partial least squares – discriminant analysis (PLS-DA) struggled with ~85 and 39% correct assignments, respectively. Additional mathematical treatments of the data set, such as incrementally increasing the exponents, did not improve the performance of LDA and KNN. The literature precedence indicates that the most common methods for analyzing colorimetric arrays are PCA, LDA, HCA, and KNN. To our knowledge, this is the first report of comparing and contrasting several more diverse chemometric methods to analyze the same colorimetric array data.

Keywords: chemometric analysis, colorimetric sensor array, hierarchical cluster analysis (HCA), hit quality index (HQI), k nearest neighbor analysis (KNN), linear discriminant analysis (LDA), soft independent modelling by class analogy (SIMCA), support vector machines (SVM)

1. Introduction

The examination of digital images in analytical chemistry has increased by more than 87% from 2005 to 2015, tracking with the increased availability of imaging devices (Capitán-Vallvey, López-Ruiz, Martínez-Olmos, Erena, & Palma, 2015). In particular, colorimetric tests and arrays have greatly benefited from the enhanced qualitative and quantitative analysis provided by that color space techniques (Askim, Mahmoudi, & Suslick, 2013). Colorimetric arrays are typically composed of 3-40 sensors that can interact with analytes and change color upon molecular interactions (Burks et al., 2010; Li, Jang, Askim, & Suslick, 2015; Salles, Meloni, de Aaujo, & Paixão, 2014). Various types of color changing sensors have been utilized in sensor arrays including pH indicators, metalloporphyrins, solvatochromic dyes, redox indicators, metal salts, ionic liquids, and nanoparticles (Askim et al., 2013; Galpothdeniya et al., 2015). Potential analyte – sensor interactions leading to colorimetric changes include Lewis acid/base interactions, hydrogen bonding, π-π interactions, and dipole-dipole interactions. Array sensor selection typically depends on an analyte's chief mode of interaction. For example, an acidic or basic analyte would warrant pH indicators as sensors, while the detection of a metal ion would point to complexometric sensors (Ariza-Avidad et al., 2014). The previously mentioned analyte – sensor interactions allow for a dynamic versatility and high applicability of colorimetric sensor arrays (Suslick, 2004). Effective arrays typically have the following criteria: high selectivity, high sensitivity, the ability to detect many analytes with the fewest numbers of sensors, and yield RGB data that can be analyzed via statistical analysis methods

for identification of unknowns. Furthermore, preferable sensors will also have the following qualities: solubility in commonly used solvents, are stable over time, both low cost and low toxicity, as well as demonstrating dramatic color changes (Burks et al., 2010).

Colorimetric sensor arrays have been shown to detect a diverse range of analytes including ions, acids and bases, metal nanoparticles, explosives, pesticides, warfare agents, drugs, various organic compounds, complex mixtures (including coffee, beer, and soft drinks), and even biological molecules (steroids and proteins) (Askim et al., 2013; Bang, Lim, Park, & Suslick, 2008; Batres et al., 2014; Capitán-Vallvey et al., 2015; Chulvi et al., 2012; Johnke, Batres, Wilson, Holmes, & Sikich, 2013; Kitamura, Shabbir, & Anslyn, 2009; Lim, Feng, Kemling, Musto, & Suslick, 2009; Mahmoudi, Lohse, Murphy, & Suslick, 2016; Soga, Jimbo, Suzuki, & Citterio, 2013; Kangas, 2017). Versatile, colorimetric arrays have been used to detect analytes in solid, liquid, and gas phases (Feng, Musto, Kemling, Lim, & Suslick, 2010). More and more, analyte induced sensor color changes are analyzed by computational methods, rather than the traditional user vision color acuity. The patterns of color changes in colorimetric arrays, when analyzed with chemometric methods including Euclidean distance, binary codes, principal component analysis (PCA), hierarchical cluster analysis (HCA), linear discriminant analysis (LDA), and matrix discriminant analysis (MDA), can be used for the identification and quantification of different compounds (Askim et al., 2013; Burks et al., 2010; Capitán-Vallvey et al., 2015; Zhang, Askim, Zhong, Orlean, & Suslick, 2014). Most relevant studies only focus on one - or very few - chemometric methods, which limits comparisons between techniques for data analysis of colorimetric sensor array output. The goal of our research was to explore several multivariate techniques, including less often reported on methods, to expand the arsenal of chemometric techniques used in tandem with colorimetric detection.

In our previous study, we investigated the performance of HCA, LDA, k-nearest neighbors (KNN), and hit quality index (HQI) in classifying samples of water, HCl (0.5 - 10 M), and NaOH (0.5 - 10 M) using an eight sensor colorimetric sensor array (Kangas, 2018). In this study, samples were classified based on the analyte and concentration allowing for some quantitation. For the classification of analyte concentration, LDA slightly outperformed HQI and KNN, with 96%, 94%, and 90% accuracies, respectively. The work described herein compares the categorizing accuracy of nine different statistical analysis methods using one dataset comprised of 631 formulations of 0.1 to 3 M of acetic acid ($pK_a = 4.76$), malonic acid ($pK_{a1} = 2.83$; $pK_{a2} = 5.69$), lysine ($pK_{a1} = 2.18$; $pK_{a2} = 8.95$), and ammonia ($pK_a = 9.25$) solutions with an eight sensor colorimetric array, with accompanying image collection and image analysis (Figure 1).

The chemometric methods selected for this study include PCA, HCA, KNN, LDA, and HQI which were used in the previous study, as well as, soft independent modelling by class analogy (SIMCA), recursive partitioning (RPART), partial least squares – discriminant analysis (PLS-DA), and support vector machines (SVM). The selected chemometric methods were chosen because they are common chemometric methods, range in complexity from simple non-parametric methods to very sophisticated methods.

These test compounds were selected for three main reasons: (1) they are inexpensive and water soluble; (2) readily available pH indicators could be used as sensors; (2) they contain functional groups often targeted by color tests such as such as a(n) carboxylic acid or dicarboxylic acid, amine (primary or secondary), alcohol (primary or secondary), aldehyde, ketone, ester, and many others (Gilbert & Martin, 2010; Khan, Kennedy, & Christian, 2012; Kovar & Laudszun, 1989; United Nations International Drug Crime Programme, 1994); (3)acids and bases encompass analytes of interest to pharmaceutical industry, forensic science, and environmental fields (Charifson & Walters, 2014).

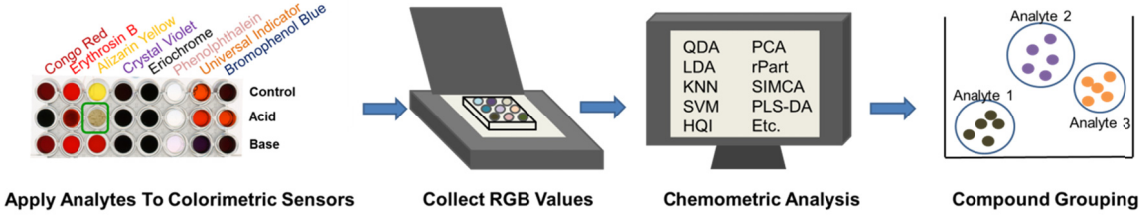

Figure 1. Experiment schematic overview for classifying selected acids and bases using the sensor array and chemometric analysis presented in the work

2. Experimental

All reagents were purchased from various chemical supply companies at technical grade or better and were used as received without further purification. A solution of universal indicator was prepared as previously described (Kangas, 2018). One percent weight-by-weight solutions of Congo red (CR), erythrosin B (EB), alizarin yellow R (AY), crystal violet (CV) , eriochrome black T (ER), phenolphthalein (PH), universal indicator (UV), and bromophenol blue (BB)

were prepared by dissolving each into aliquots of a solvent mixture consisting of acetate buffer (0.1 M, pH 5), ethylene glycol, triethylene glycol monobutyl ether, and glycerol in a ratio of 14:1.6:1:3.2. The sensor solutions sonicated for 1 hour in a bath sonicator (30 ºC), followed by 5 minutes mixing with a probe sonicator, and then vacuum filtered twice through Whatman #1 filter paper. Acetic acid (0.1 – 3 M) and ammonia (0.1 – 3 M) solutions were prepared by diluting concentrated reagent solutions with milli-Q water (18 MΩ-cm). Solutions of malonic acid (0.1 – 2 M) and lysine (HCl salt, 0.1-2 M) were prepared by dissolving appropriate amount of the analytes in milli-Q water. The sensor array was laid out in a 96-well plate as shown in Figure 2 by dispensing 100 µL of each sensor in designated rows. The same volume of an analyte or control were added to the 12 columns of the well plate. To explore reproducibility, each plate contained 4 replicates of a water control and 8 replicates of each analyte.

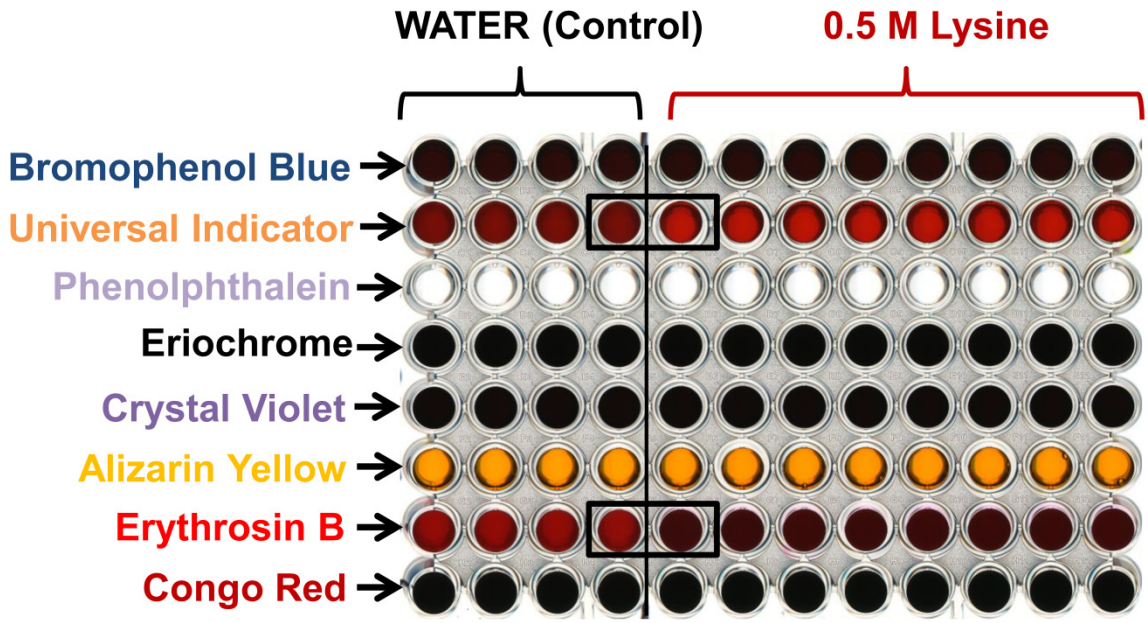

Figure 2. Sensor array housed in a 96-well plate, with each sensor was placed in a designated row. For each array, the first four columns are controls (water), and the final eight columns are analyte (shown above: 0.5 M lysine). The black boxes highlight color differences between the control (water) and the analyte (0.5 M lysine).

All array images, including Figure 2, were collected as 24-bit color images using an Epson Perfection V700 desktop scanner in transparency mode. To eliminate interferences from stray light, the scanner was draped in black cloth. The images were analyzed with ImageJ (Schneider, Rasband, & Eliceiri, 2012), and the extraction of mean RGB values for each well was automated with a macro (Lyon et al., 2012; Soldat, Barak, & Lepore, 2009). No attempts were made to correct for image-to-image variation by subtracting a control row, as our previous work showed such a correction to be unnecessary (Kangas 2018). The RGB dataset is provided in the supplemental information as SI.4 to facilitate further chemometric studies.

All statistical analysis was performed using the statistical programming language R. PCA was performed using the function prcomp. The data was mean-centered, but was not scaled to unit variance because all of the data was on a consistent scale of 0 to 255 RGB units. Loading plots are included in the supplemental information as SI.2. Score plots of the resulting data were constructed using the function pca2d from the library pca3d (Weiner, 2017). Hierarchical clustering was conducted in agglomerative mode using Ward's method based on Euclidean distances using the hclust function from the stats library (R Core Team, 2017). LDA was performed using the MASS library (Venables & Ripley, 2002) and the classification ability of the LDA model was tested by using all but one cross validation. The concentrations of each analyte were treated as a separate class, and for classification the prior probabilities were equal for all classes. KNN based on Euclidean distances was performed for k=1, 3, 5, 7, and 9 using the class package (Venables & Ripley, 2002). HQI values were calculated using a custom R program given in the supplemental information (Supplementary Figure, SI.1. Classification was performed using all but one cross-validation, and classifications were assigned based on the library sample with the highest HQI value. PLS-DA was performed using the plsDA function from the DiscriMiner library with leave one out cross validation (Sanchez, 2013). For RPART, SVM, and SIMCA, the data was randomly split into a training set containing 75% of the data and a test set comprised of the remaining 25%, and the analysis was repeated in triplicate with different train/test sets. RPART was conducted using the

rpart package (Therneau, Atkinson, & Ripley, 2017). SVM was performed using the rrcovHD package (Todorov, 2016). SIMCA was performed using the CSIMCA function from the rrcovHD package.

3. Results and Discussion

In the present study, the selected sensors were chosen for many of the reasons listed in the Introduction, plus our previous work (Kangas, 2018) showed these sensors to be well-suited for the qualitative and quantitative classification of NaOH and HCl solutions. As shown in Figure 2, some color differences between the control and analyte test wells are easily visible to the naked eye. However, some sensor changes more subtle and are not immediately visible to a user. Thereby, with chemometric analysis the investigator can more easily detect and use subtle color changes identified by image analysis for the identification and quantification of analytes.

3.1 Principal Component Analysis

PCA is one of the more commonly used chemometric analysis methods for large data sets collected from colorimetric tests or sensor arrays (Capitán-Vallvey et al., 2015). PCA is an algorithm that uses an orthogonal transformation a set of observable - and possibly related - variables to change them into a set of linearly uncorrelated variables known as principal components (Graham, 1993). Usually, for use with colorimetric sensors, this means that the principal components are statistically weighted combinations of the R,G and B values from all the sensors. Although calculated in multiple components, only a few components are typically needed to visualize and analyze trends in the data set. In this study, for each array, 24 variables were available in the original data set (i.e. 8 sensors x 3 channels = 24 variables). However, only 4 principal components were required to assess 95% of the variance in the data set. The number of components needed to describe the variance in the data set was consistent with observations from other colorimetric sensor array studies (Li et al., 2015; Salinas et al., 2014).

When analyzing with PCA, variables that are strongly correlated typically remain closely related when converted to the new components. Similarly, data points that are clustered in the original data set can usually be found together in the principal component space, thereby allowing the visualization of similar data when multiple component are plotted together (Graham, 1993). In our work here, the biplots of PC1 and PC2 (Figure 3) show that water, each acid, and each base create distinct clusters in the plot. These grouping indicate the ease with which each analyte of study can be identified, whereas the biplot of all analytes (Figure 3) struggles to distinguish malonic from acetic acid in the same space. However, in the biplot of PC2 and PC3 (Supplementary Figure SI.2), the distinction of acetic and malonic acid is much more apparent while the distinction of water, ammonia and lysine are less so. The individual concentrations of the analytes were also observed via plotting in the PC1 vs PC2 biplot. Again, the acids did not perform well with very little selective grouping to mark the concentration changes while ammonia showed very distinct clusters of grouping by concentration below 2 M. It was also observed that the low concentrations of lysine, 0.1-1 M, were very difficult to distinguish from one another while there was a distinct separation between those and the 2 M samples. However, some distinction arises in all groups by concentration when viewed in the PC2 vs PC3 biplot suggesting that this space is a better space to observe the clustering and concentration differences found by the PCA analysis (Supplementary Figure SI.2).

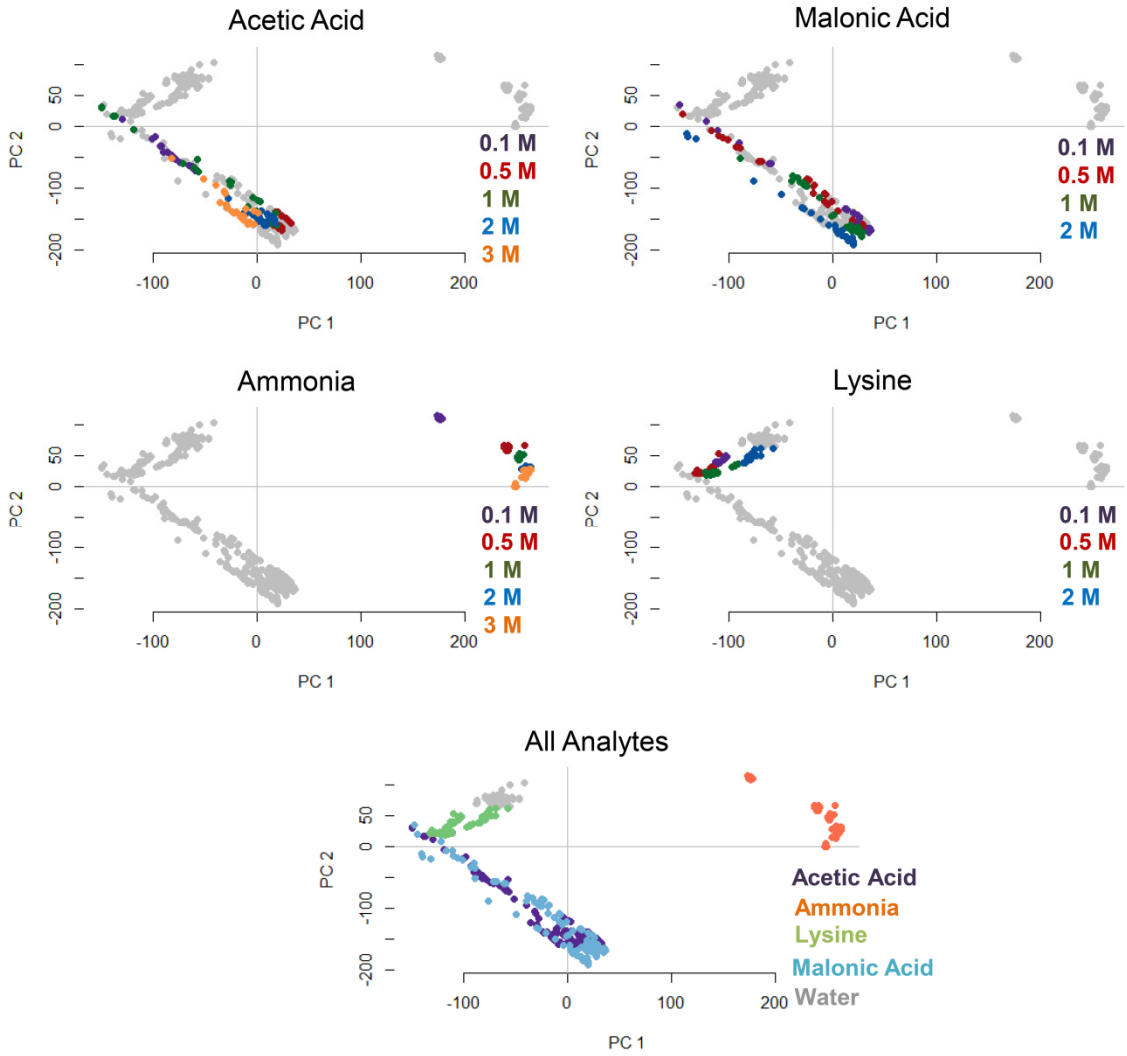

Figure 3. Biplots of PCA results of acetic acid, ammonia, lysine, malonic acid, and water by concentration and all analytes together showing analyte and concentration grouping as viewed by the first and second components. PC1 and PC2 are the first two components from PCA, respectively

By observing the principle component values for each channel, PCA analysis can be used to also determine the sensors with the most influence within each component. For instance, as observed in Supplemental figure SI.3, PC1 appears to be dominated by the red channels of CR, EB, AY, and UV and the green channels of AY and PH. In PC2, the red channels of EB and AY are strongly influential while the red channels of CR, UV and BB and green channels of AY and PH are only mildly so. Finally, PC3 is strongly dominated by the red channel of UV and BB and the blue channel of AY while only mildly influenced by the green channel of AY. This may explain why the acid samples are not well distinguished in the PC1 and PC2 biplot as both channels are strongly dominated by similar red channels, especially EB and AY, while the PC3 is strongly influenced by other dyes in different channels. In addition to visualizing the data with the scores plot, PCA can also be used to determine which sensors are responsible for the analyte discrimination by analyzing the loading plots (Supplementary Figure, SI.3). This information could be used for sensor selection and array optimization as the loading plots reveal which sensors make the biggest contributions toward analyte detection as evidenced by the highest loadings on the y-axis.

Overall, the plots show that PC1 and PC2 space provide excellent separation of ammonia, lysine, and water, but not acetic and malonic acid. This is likely because PC1 and PC2 are strongly dominated by similar principle components. To achieve separation of the acids and their concentrations, one must look at PC2 and PC3 space which provides better separation but loses the separation of the bases, which like the acids, seems to coincide with sensor pKa values and the complementary color change. This indicates that simply relying on two dimensions of PCA is not best for identifying the acids and bases tested in this study, and at least three components should be used. Furthermore, by viewing the

single component variables (Supplementary Figure, SI.2), the array of dyes can be improved by noting those sensors which strongly influence each components (i.e. CR, EB, AY, UV, BB, PH) and those that have very little influence in the first three components, (i.e. CV and ER). It may be advisable to use other components to see if those sensors have any influence elsewhere in the space or exchange non-influential sensors for others which may provide more information.

3.2 Hierarchical Cluster Analysis

Similar to PCA, HCA is an unsupervised, no bias multivariate clustering analysis that is commonly used to analyze colorimetric arrays (Bueno, Meloni, Reddy, & Paixão, 2015; Capitán-Vallvey et al., 2015). When performing HCA, the analysis receives no information on the classes of the samples except for the 24 variable values per sample. Therefore, the grouping of samples is determined by closeness, typically a Euclidean distance, in that 24 dimension space. This clustering is an iterative process resulting in a tree of relational closeness where well-related samples are near to one another on the tree while samples with less relation are further away. Compared to other clustering algorithms, HCA has two main advantages: (1) it provides a quantitative metric for the similarity of groups and (2) it defines clusters in all size scales, ranging from individual samples up to a single group that contains all samples (Graham, 1993). Figure 4 shows the clustering results of the analyte group means, with all variables averaged within a particular sample label, from the colorimetric data to provide a more visually appealing example of the clustering capabilities. When viewing the HCA plot, each of the T junctions are flexible when interpreting similarity within the plot. What this means is that groups within the same branch, but not necessarily samples within the same region of the graph, are considered similar. For example, all of the ammonia samples fall within the same branch noting their similarity. However, 2 M ammonia is equally similar to 0.5 M ammonia as it is to 1 M ammonia due to the flexibility of the branch junctions. Furthermore, in the branch containing lysine and water, 1 M lysine and 2 M lysine are classified equally similar to water although 0.5 M lysine is less similar to water than 1 M lysine. Ammonia and all the rest of the data form the largest groupings showing that ammonia is the least similar to all of the rest of the samples. Within the next tier, lysine-water and the acids form the next families demonstrating how quickly HCA is able to discern the selected acids from the bases. While lysine and ammonia are easily distinguished, HCA struggles with the individual acids especially in the low concentrations, such as in the case of 0.1 M malonic and 1 M acetic acid.

Overall, HCA provided useful analysis for understanding similarities and differences in the data sets and was able to distinguish ammonia and lysine as having distinct signal from the acids. However, given that the branches of the dendromer are freely rotating, HCA was unable to distinguish the variables of lysine as being uniquely basic or distinct from water. Furthermore, HCA does not indicate why the groups were clustered as they were or how each individual variable contributes to the classification, which would – as in the case of PCA – be helpful sensor array optimization data. Although HCA is a powerful classification tool that provides sufficient analysis of the data set, we judge it inferior to PCA in the ability to draw conclusions on the sufficiency of the sensor array as HCA provides little insight into how the classification may be improved.

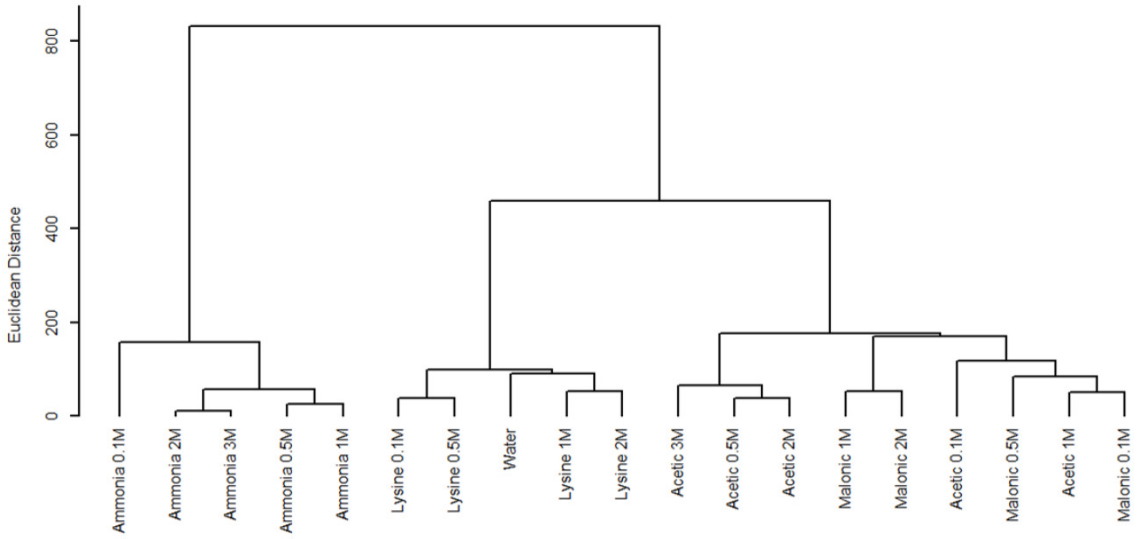

Figure 4. Dendrogram from HCA of 0.1 – 3 M acetic acid, 0.1-3 M ammonia, 0.1-2 M lysine, 0.1-2 M malonic acid and water. Distinct clusters formed for water, ammonia, and lysine

3.3 Linear Discriminant Analysis

Like PCA, LDA is also a method that generates new variables called discriminants which consist of linear combinations of the original variables and has been applied in colorimetric sensor array analysis (Minami et al., 2013; Zhang et al., 2014). Unlike PCA, LDA considers and exploits the differences in the group means and often outperforms PCA in the separation of groups (Askim et al., 2013). The main disadvantage of LDA is that it requires a data set larger than that required for PCA (Wold, Johansson, Jellum, Bjørnson, & Nesbakken, 1981). In addition, the size and composition of the classes within a dataset will affect the discriminant and result of LDA, thus influencing LDA's ability to correctly identify analytes and their concentrations. Figure 5 shows a panel of plots from the first and second discriminants of LDA analysis which was input as groups by concentration. In the plot of analytes it is observed that LDA appears to cluster acetic acid, malonic acid and ammonia as distinct groups for analysis and identification of analyte while lysine is hardly distinguishable from water. Malonic acid was easily separated into of concentration groups, while acetic acid and ammonia solutions were well-resolved at lower concentrations but distinguishing between higher concentrations - especially 2 and 3 M acetic acid and 1, 2, and 3 M ammonia – was a challenge. For this particular data set it appears that LDA is superior to PCA at a two dimensional separation of acetic and malonic acid with regard to class and concentration, but is more comparable to HCA in the ability to distinguish lysine from water.

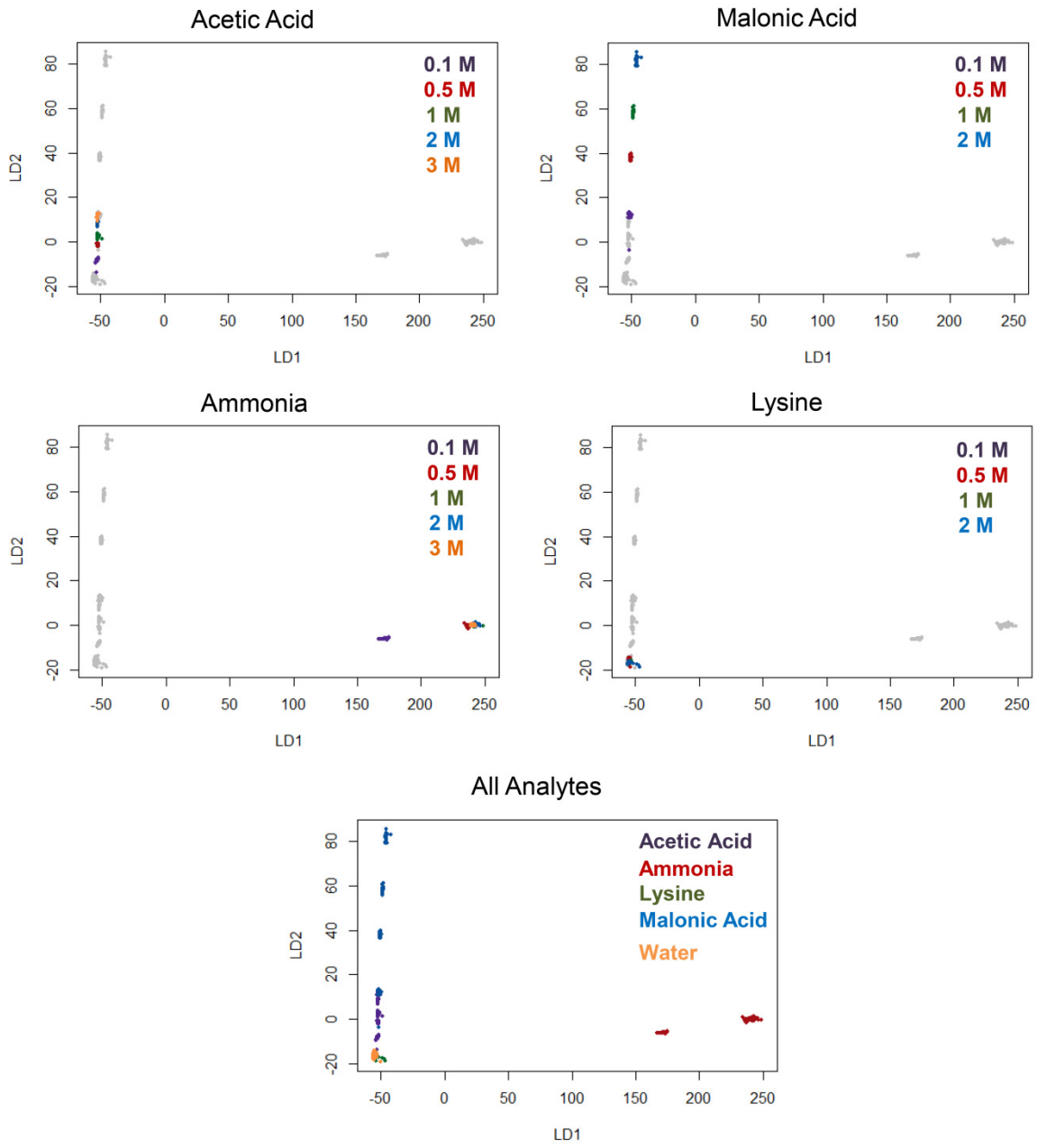

Figure 5. LDA for 0.1 -3 M acetic acid, and 0.1-3 M ammonia, 0.1-2 M lysine, 01-2 M malonic acid and water. LD1 and LD2 are the first and second discriminants from LDA, respectively

An advantage of LDA over PCA is LDA's ability to report quantitative means of classifying unknown compounds. Table 1 shows the classification of the LDA analysis in which groups were input by concentration. LDA was able to correctly classify samples by analyte identity and concentration in 626 out of 631 samples (99.2%). When misclassifications occurred, the analysis was often still able to classify an analyte as an acid or a base. For example, one 3 M ammonia sample was misclassified as 1 M ammonia and one 0.1 malonic acid sample was misclassified as 0.5 M acetic acid. As mentioned previously, LDA analysis struggled most with lysine samples, misclassifying one sample of 0.1 M acetic acid as 0.1 M lysine and one sample of water as 2 M lysine. One sample of 0.5 M ammonia could not be classified, as replicate trials resulted in different classifications. In addition, subsequent analysis of the posterior probabilities indicated a modeling error in the analysis. This result was likely the result of unusually low RGB values for phenolphthalein other possibilities include a shadow or air bubble in the image of the well. LDA was also used to qualitatively classify analytes correctly 622 out of 631 (98.6%). The previously observed trends were also true when considering only analyte classification with one sample of acetic acid misclassified as lysine (struggles with lysine) and seven samples of malonic acid classified as acetic acid (still classifies acids as acids). Overall, LDA was observed to

classify the data 99.2% correctly when considering analyte identify and concentration, while data classification was 98.6% correct when only analyte identity was considered. Although PCA has an advantage of better grouping lysine samples, LDA clearly has the advantage of quantitative classification results which may be more useful in certain reporting schemes.

Table 1. Summary of LDA sample classification (grouped by [Analyte])

Analyte	Total	Correct	Incorrect	Misclassified
0.1 M acetic	24	23	1	0.1 M lysine
0.5 M acetic	24	24	0	-
1 M acetic	24	24	0	-
2 M acetic	24	24	0	-
3 M acetic	24	24	0	-
0.1 M ammonia	24	24	0	-
0.5 M ammonia	24	23	1	Modeling Error
1 M ammonia	24	24	0	-
2 M ammonia	24	24	0	-
3 M ammonia	24	23	1	1 M ammonia
0.1 M lysine	24	24	0	-
0.5 M lysine	24	24	0	-
1 M lysine	24	24	0	-
2 M lysine	24	24	0	-
0.1 M malonic	24	23	1	0.5 M acetic
0.5 M malonic	24	24	0	-
1 M malonic	24	24	0	-
2 M malonic	24	24	0	-
Water	199	198	1	2 M lysine
Overall	631	626	5	

Table 2. Summary of LDA sample classification (grouped by analyte)

Analyte	Total	Correct	Incorrect	Misclassified
acetic acid	120	199	1	lysine
ammonia	120	119	1	Modeling Error
lysine	96	96	0	-
malonic acid	96	89	7	7*acetic acid
water	199	199	0	-
Overall	631	622	5	

3.4 K Nearest Neighbor

KNN is another chemometric method which classifies unknown samples by comparing them to a library of known samples. Previous applications of KNN include pH determination using a sensor array (Capel-Cuevas, Cuéllar, Orbe-Payá, Pegalajar, & Capitán-Vallvey, 2010) and melting point estimation of organic compounds (Nigsch et al., 2006). With KNN, the classification of the unknown is determined by measuring the distance to the most similar known samples. The unknown is then identified by association with the predetermined "K" number of nearest neighbors, with the nearest neighbor defined as the known samples with the shortest distance to the unknown (Ma, Yang, & Cheng, 2014). For example, if K = 1 the unknown is identified as the classification of the one closest known sample whereas if K = 4 then the unknown is classified with the identity of the four closest neighbors. KNN classification treats the data sets as points in dimensional space equal to the number of variables (Balabin, Safieva, & Lomakina, 2010). In this study with 24 values per array, there are 24 dimensions, this makes the calculation of the distance between points and the

implementation of KNN relatively straight-forward. Furthermore, KNN often performs as well as or better than more complicated classifiers, such as SVM (Ma et al., 2014).

For our work, K was varied and the Euclidean distance was calculated for KNN analysis. Increasing the k value has previously been shown to influence method performance and accuracy, but the relationship between k and performance can vary (Ma et al., 2014). In our study, increasing K = 1 to K = 9 resulted in lower classification accuracy (Table 3). Overall, KNN was used to correctly identify 98.1% of samples for K = 1, decreasing to 93.0% for K=9. This indicates that K = 1 was the optimal parameter. Table 4 contains K = 1 classification data. Our results indicate that KNN is a very robust method for sample classification for this sample size since the least accurate run with the largest number of neighbors (K = 9) is able to classify samples correctly 93.0% of the time. Some of the mislabeled samples involved the correct analyte, but the wrong concentration, such as 0.1 M lysine was identified as 0.5 M lysine or 0.5 M acetic acid was identified as 1 M. Six misclassifications occurred in acetic acid.

Comparing the two classification methods, KNN (K = 1) is comparable to LDA with 98.1% accuracy and 98.6% accuracy, respectively. While LDA utilizes input variables with optimal weights to separate the group means and KNN gives equal significance to all of the variables, both methods still result in similar results. Alternative KNN algorithms do apply various transformations to the dataset to optimize the accuracy of KNN. However, these methods were not pursued in this study (Nigsch et al., 2006).

Table 3. Effect of K on K nearest neighbor (KNN) accuracy

K	Total	Correct	Incorrect
1	631	619	12
3	631	611	20
5	631	598	33
7	631	592	39
9	631	587	44

Table 4. Summary of k nearest neighbor (KNN) sample classifications. (K = 1)

Analyte	Total	Correct	Incorrect	Misclassified
0.1 M acetic acid	24	23	1	0.5 M lysine
0.5 M acetic acid	24	23	1	1 M acetic acid
1 M acetic acid	24	21	3	2* 0.5 M acetic acid, 0.1 M malonic acid
2 M acetic acid	24	23	1	3 M acetic acid
3 M acetic acid	24	23	1	1 M acetic acid
0.1 M ammonia	24	24	0	-
0.5 M ammonia	24	23	1	3 M ammonia
1 M ammonia	24	24	0	-
2 M ammonia	24	24	0	-
3 M ammonia	24	24	0	-
0.1 M lysine	24	23	1	0.5 M lysine
0.5 M lysine	24	24	0	-
1 M lysine	24	24	0	-
2 M lysine	24	23	0	-
0.1 M malonic acid	24	24	1	1 M acetic acid
0.5 M malonic acid	24	23	0	-
1 M malonic acid	24	22	2	0.5 M malonic acid, 2 M malonic acid
2 M malonic acid	24	24	0	-
Water	199	194	5	2 M lysine, 3 * 3 M malonic acid
Overall	631	619	12	

3.5 Hit Quality Index

HQI is commonly used as a spectral comparison method when working with an unknown FTIR or Raman spectra and a database of known spectra (Gryniewicz-Ruzicka, Rodriguez, Arzhantsev, Buhse, & Kauffman, 2012; Lee, Lee, & Chung, 2013, p. 201). HQI treats the unknown spectra as vectors in 24 dimensional space by calculating the dot product according to the following equation:

$$HQI = \frac{(x \cdot y) * (x \cdot y)}{(x \cdot x) * (y \cdot y)} \quad (1)$$

The terms x and y are the unknown and one of the many known spectra in the database, respectively. Classification is then assigned by assessing the closeness of fit which is determined by the result of the dot product being close to 1. The use of HQI in comparative colorimetric studies have shown similar accuracies to other chemometric methods we compare (Gryniewicz-Ruzicka et al., 2012; Lee et al., 2013, p. 201).

In the present study, HQI showed an overall 98% accuracy for analyte identity and concentration when classifying acetic acid, malonic acid, lysine and ammonia (see Table 5 for results). Common misclassifications include concentrations within the same analyte (6 samples), such as 1 M acetic acid misclassified as 0.5 M acetic acid and 0.5 M lysine misclassified as 0.1 M lysine. Two samples were classified outside their analyte; 0.5 M lysine misclassified as 0.1 M acetic acid and 0.1 M malonic acid misclassified as 1 M acetic acid. Similar to KNN, HQI underperformed both LDA and PCA, especially with respect to classifying by concentration within an analyte. These results are similar to our previous work with NaOH and HCl which also showed a high level of accuracy for analyte classification efficacy while suffering in the concentration identification within analytes, especially high concentration HCl (Kangas, 2018). HQI is a straightforward, mathematical analysis method which could be useful as an alternative or additional chemometric analysis of colorimetric arrays. However, it suffers from inaccuracies which make it inferior to LDA and PCA-especially in the acetic acid concentrations tested in the present study (Table 5).

Table 5. Summary of hit quality index (HQI) sample classifications

Analyte	Total	Correct	Incorrect	Misclassified
0.1 M acetic acid	24	23	1	0.5 M lysine
0.5 M acetic acid	24	23	1	1 M acetic acid
1 M acetic acid	24	21	3	0.5 M acetic acid, 0.1 malonic acid
2 M acetic acid	24	23	1	3 M acetic acid
3 M acetic acid	24	23	1	1 M acetic acid
0.1 M ammonia	24	24	0	-
0.5 M ammonia	24	24	0	-
1 M ammonia	24	24	0	-
2 M ammonia	24	24	0	-
3 M ammonia	24	24	0	-
0.1 M lysine	24	23	1	0.5 M lysine
0.5 M lysine	24	24	0	-
1 M lysine	24	24	0	-
2 M lysine	24	24	0	-
0.1 M malonic acid	24	23	1	1 M acetic acid
0.5 M malonic acid	24	24	0	-
1 M malonic acid	24	22	2	0.5 malonic acid, 2 M acetic acid
2 M malonic acid	24	24	0	-
Water	199	199	0	-
Overall	631	620	11	

3.6 Partial Least Squares Discriminant Analysis

PLS-DA is an extension of the PLS methodology used for classifying samples. Briefly, PLS or PLS-DA are performed by generating components similar to those in PCA, but the components are selected to correlate with y-values or classes, respectively (Brereton & Lloyd, 2014). Our PLS-DA results for classifying samples into groups based on the analyte

and concentration are given in Table 6. Unlike the other classification methods used in this study, PLS-DA was only able to correctly classify water and the most concentrated ammonia samples (3 M), resulting in an overall accuracy of 39%. The poor performance may be ascribed to PLS-DA using a one-versus-all approach for multiple class problems (Brereton & Lloyd, 2014). With the present data set, the samples in target class may be very similar to those in the rest of the dataset. Since water was the majority of our observations, the class weights may have been affected. For example, when classifying 0.5 M acetic acid, both 0.1 and 1 M acetic acid would be in the other class. PLS-DA also can have problems setting the boundaries when there are groups of unequal sizes (Brereton & Lloyd, 2014), which would also be present with the one-versus-all groupings. PLS-DA analysis with groups based on the analytes rather than both the analytes and concentration resulted in a much higher accuracy with 610 of 631 correct, data not shown. In this case, the one-versus-all groupings should be more distinct and the classes should be closer in size, and there are fewer classes to assign observations to.

Table 6. Summary of partial least squares discriminant analysis (PLS-DA) sample classifications

Analyte	Total	Correct	Incorrect	Misclassified
0.1 M acetic acid	24	0	24	24 * water
0.5 M acetic acid	24	0	24	24 * water
1 M acetic acid	24	0	24	24 * water
2 M acetic acid	24	0	24	24 * water
3 M acetic acid	24	0	24	24 * water
0.1 M ammonia	24	0	24	24 * water
0.5 M ammonia	24	0	24	24 * 3 M ammonia
1 M ammonia	24	0	24	24 * 3 M ammonia
2 M ammonia	24	0	24	24 * 3 M ammonia
3 M ammonia	24	24	0	-
0.1 M lysine	24	0	24	24 * water
0.5 M lysine	24	0	24	24 * water
1 M lysine	24	0	24	24 * water
2 M lysine	24	0	24	24 * water
0.1 M malonic acid	24	0	24	17 * 3 M malonic acid, 7 * water
0.5 M malonic acid	24	0	24	24 * 3 M malonic acid
1 M malonic acid	24	0	24	24 * 3 M malonic acid
2 M malonic acid	24	0	24	-
Water	199	199	0	-
Overall	631	247	384	

3.7 Recursive Partitioning

RPART is fast and simple to implement (Miller, 2001) classification method that generates a decision tree to classify samples. At each branch, the value of a single variable is tested with a rule. For example, a classifier for patients may use rules like is the patient age >18, while a classifier for colorimetric data may test the intensity of a specific sensor. Examples of the usage of RPART in the chemical literature include calculating phase diagrams for surfactants (Bell, 2016) and identifying new pharmaceutical and antibiotic compounds (Rusinko, Farmen, Lambert, Brown, & Young, 1999; Wang et al., 2014). An advantage of RPART is that only the most important variables from the dataset are used in the rules, and variables that have a low impact on classification are ignored.

Figure 6 shows the decision tree used for classifying the samples. As shown in Figure 6, the first rule is based on the blue intensity of AY. The numbers below the groups indicate the confidence in that classification with the training set. The 0.1 M lysine sample was the only one with a confidence less than 1, and is consistent with the classification results for the test set, where there were acetic acid samples classified as 0.1 M lysine and lysine samples classified as acetic acid. In addition, the decision tree shows that phenolphthalein and eriochrome black T were not used in any rules, while bromophenol blue and alizarin yellow were the most utilized sensors.

For RPART, the data was randomly split with R into a training set which acts as the database with 473 samples and a testing set which acts as the unknowns with 158 samples. To test the reproducibility, the analysis was repeated in triplicate with a new training and testing set of data randomly chosen for each trial. The classification accuracies for the three trials were 95.6%, 97.4%, and 96.2% for an average accuracy of 96.4%, and the classification results for trial 1 are summarized in Table 7. The decision tree for trial 1 is shown in Figure 6.

As shown in Table 7, there were 7 incorrect classifications including three acetic acid samples that were classified as the correct analyte but the wrong concentration. The remaining four samples were classified as the wrong analyte. These include two lysine samples (0.1 and 1 M) that were classified as 0.1 M acetic acid and two acetic acid samples (0.1 M) that were classified as lysine (0.1 M). These results are consistent with the results for KNN and LDA (98% and 98.6%, respectively), which also showed some confusion between dilute acetic acid and lysine. This is likely due to the similarity in pH between the two compounds.

Table 7. Summary of RPART sample classifications

Analyte	**Total**	**Correct**	**Incorrect**	**Misclassified**
0.1 M acetic acid	9	7	2	2 * 0.1 M Lysine
0.5 M acetic acid	5	5	0	-
1 M acetic acid	6	5	1	2 M acetic acid
2 M acetic acid	9	8	1	1 M acetic acid
3 M acetic acid	7	7	0	-
0.1 M ammonia	8	8	0	-
0.5 M ammonia	3	3	0	-
1 M ammonia	6	6	0	-
2 M ammonia	5	5	0	-
3 M ammonia	9	8	1	2 M ammonia
0.1 M lysine	5	4	1	0.1 M acetic acid
0.5 M lysine	5	5	0	-
1 M lysine	6	5	1	0.1 M acetic acid
2 M lysine	5	5	0	-
0.1 M malonic acid	7	7	0	-
0.5 M malonic acid	7	7	0	-
1 M malonic acid	10	10	0	-
2 M malonic acid	5	5	0	-
Water	41	41	0	-
Overall 158 151 7				

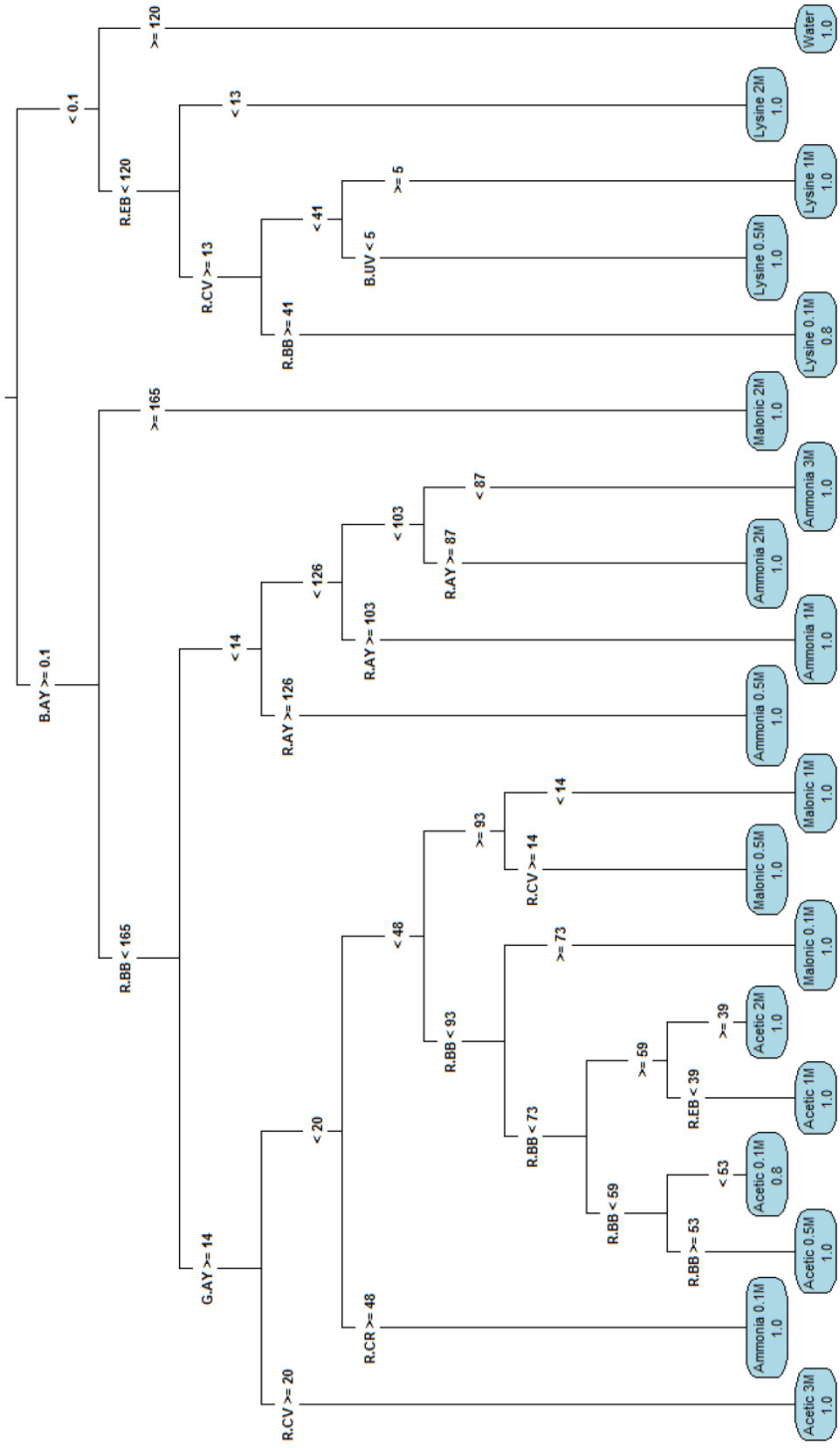

Figure 6. Decision tree generated with trial one of RPART

Each branch provides a freely rotating, more selective classification of the samples. The blue boxes at the bottom show the classes and the confidence in that assignment. Variables used in the rules are indicated as R, G, or B for the color channel and an abbreviation for the sensor. Sensors are defined as Congo red (CR), erythrosin B (EB), alizarin yellow R (AY), crystal violet (CV), eriochrome black T (ER), phenolphthalein (PH), universal indicator (UV), and bromophenol blue (BB).

3.8 Support Vector Machines

Support vector machines were utilized to investigate whether more novel and less reported learning schemes could

improve the classification results of our data set. SVMs establish a space complete with hyperplanes that are based on a training set of data with known classifications. This gives maximum distance between groups clearly dividing the possible combinations of data. When an unknown sample is added to the space, the sample is subsequently classified based on its closeness to a particular hyperplane receiving the identity of the training samples which made up that hyperplane. For colorimetric sensor arrays, SVMs has been applied for the detection, prediction, and classification of various explosives (Askim, Li, LaGasse, Rankin, & Suslick, 2016).

Table 8 demonstrates one trial of SVM identification and classification with the presented dataset. Out of 158 samples, 140 were identified correctly in analyte identity and concentration. Most misclassifications occurred within the acetic acid samples. However, it was only the concentrations of the acetic acid samples there were misidentified but not the analyte itself - such as 1 M of acetic acid was misclassified as 0.5 and 2 M acetic acid. Only two concentrations were misclassified in the ammonia dataset: 1 M of ammonia was predicted to be 0.5 M of ammonia. In the end, the accuracy of SVMs was 89%, which makes the performance of SVMs comparable to other learning schemes such KNN (98%), LDA (98.6%), etc.

Table 8. Summary of Support Vector Machines (SVM) sample classifications

Analyte	Total	Correct	Incorrect	Misclassified
0.1 M acetic acid	4	4	0	-
0.5 M acetic acid	7	3	4	4 *2 M acetic
1 M acetic acid	5	2	3	0.5 M acetic acid, 2 * 2 M acetic acid
2 M acetic acid	4	4	0	-
3 M acetic acid	12	3	9	9 * 2 M acetic acid
0.1 M ammonia	7	7	0	-
0.5 M ammonia	6	6	0	-
1 M ammonia	7	5	2	2 * 0.5 M ammonia
2 M ammonia	6	6	0	-
3 M ammonia	4	4	0	-
0.1 M lysine	4	4	0	-
0.5 M lysine	10	10	0	-
1 M lysine	4	4	0	-
2 M lysine	12	12	0	-
0.1 M malonic acid	5	5	0	-
0.5 M malonic acid	4	4	0	-
1 M malonic acid	7	7	0	-
2 M malonic acid	10	10	0	-
Water	40	40	0	-
Overall 158 140 18				

3.9 Soft Independent Modelling by Class Analogy

Soft and hard chemometric methods have been developed to analyze data obtained from chemical systems (Kakhki & Abedi, 2012). SIMCA is usually defined as soft method, meaning that samples can be classified in one group, multiple groups, or no groups. SIMCA is also an independent modelling method, which means that a sample can be categorized into to more than one group. Unlike many other classification algorithms, in SIMCA, a sample could be assigned to multiple groups in the case of orthogonal or hierarchical groups. For each class in the training set, PCA performed and a model describing the group is generated. Afterwards each unknown sample is projected into all of the models for the groups, and the unknown can be assigned to a group based on the similarity with the group (Gemperline, 2006). In addition, outliers can be rejected from all classes. Other advantages of SIMCA include the ability to work well with data sets with small numbers of variables and large numbers of variables (Esbensen, Guyot, Westad, & Houmøller, 2010). There are many examples of SIMCA in in the chemical literature including classifying samples of spray paint using FTIR spectra (Muehlethaler, Massonnet, & Esseiva, 2014), identifying contaminated pharmaceuticals with Raman spectroscopy (Gryniewicz-Ruzicka et al., 2012), identifying potential pharmaceutical compounds (Tominaga, 1999),

and classifying healthy brain tissue samples from GC-MS chromatograms (Wold et al., 1981).

Similar to RPART, SVM, and PLS-DA the data set was randomly split into a training set and a test set using R. To check the reproducibility, the analysis was performed in triplicate, with new training and test sets each time, providing accuracies of 97%, 89%, and 92% with an average of 92%. These results are comparable to the results from KNN (98%), RPART (96.4%), HQI (98%), and LDA (98.6%). A summary of the classification results for SIMCA trial 1 are given below in Table 9. Of the five misclassifications, four were the correct analyte but the wrong concentration. The final misclassification was a sample of 0.5 M lysine which was classified as 0.1 M acetic acid. Misclassifications between lysine and dilute acetic acid were also observed with RPART.

Table 9. Summary of SIMCA sample classifications

Analyte	Total	Correct	Incorrect	Misclassified
0.1 M acetic acid	9	9	0	-
0.5 M acetic acid	6	6	0	-
1 M acetic acid	6	5	1	0.5 M acetic acid
2 M acetic acid	3	2	1	3 M acetic acid
3 M acetic acid	7	6	1	1 M acetic acid
0.1 M ammonia	6	6	0	-
0.5 M ammonia	7	7	0	-
1 M ammonia	4	4	0	-
2 M ammonia	7	7	0	-
3 M ammonia	9	9	0	-
0.1 M lysine	5	5	0	-
0.5 M lysine	7	6	1	0.1 M acetic acid
1 M lysine	5	5	0	-
2 M lysine	8	8	0	-
0.1 M malonic acid	8	8	0	-
0.5 M malonic acid	2	2	0	-
1 M malonic acid	6	5	1	0.1 M malonic acid
2 M malonic acid	6	6	0	-
Water	47	47	0	-
Overall	158	153	5	

4. Conclusions and Future Outlook

Colorimetric sensor arrays are rapidly becoming a common tool for the identification and quantification of analytes. The multidimensional nature of colorimetric data is well-served by the use of chemometric methods. While HCA and PCA are popular chemometric methods, we sought to explore the use of other algorithms to compare and contrast their usefulness in qualitative and quantitative analysis. In this work, an eight sensor colorimetric array was used to compare the performance of PCA, HCA, LDA, KNN, HQI, PLS-DA, RPART, SVM, and SIMCA for efficacy in identification and quantification of acetic acid, malonic acid, ammonia, and lysine. PCA, HCA, and LDA were used to qualitatively visualize the data and relationships between the analytes. In PCA, PC1 and PC2 provide excellent separation of ammonia, lysine, and water - but not acetic and malonic acid. These analytes were separated much better with PC2 and PC3, indicating that greater than bidimensional PCA components should be evaluated to obtain optimal clustering of analytes. HCA was unable to distinguish the variables of lysine as being uniquely basic or distinct from water, making this method not as effective for classification as PCA for our selected analytes. The two dimensional separation of acetic and malonic acid with regard to class and concentration was achieved with LDA, making this method for our data set superior to PCA. However, the lysine separation from water was similar in performance to HCA. Therefore, for the present data set and presented methods, the effectiveness in regards to visualization and classification can be arranged as LDA > PCA (if only PC 1 and 2 are used) > HCA.

LDA, KNN, HQI, PLS-DA, RPART, SVM, and SIMCA were used to quantitatively classify the samples. LDA is unique

in that it can achieve visualization of the data as well as report quantitative means of classifying unknown compounds with high accuracy (>99% in this data set). KNN is advantageous because it is relatively simple to execute, performing similar to HQI (98% accuracy when k = 1) and better than PLS-DA, RPART, SVM, and SIMCA. PLS-DA was the lease discriminating chemometric method for this data set as it was only able to correctly classify water and the most concentrated ammonia samples (3 M), resulting in an overall accuracy of 39%. RPART results were consistent with KNN and LDA showing misclassifications between dilute acetic acid and lysine. In comparison to all methods except for PLS-DA (39%), SVM under performed (85% correct classification). Therefore, the effectiveness of the quantitative methods for this dataset for an analyte concentration range from 0.1M to 3.0M can be ranked as LDA > HQI > KNN > SIMCA > RPART > SVM >> PLS-DA. This coincides with our previously published ranking of LDA> HQI > KNN for classifications and quantification of HCl and NaOH. (Kangas paper) Therefore, it appears that these classification methods follow a general trend for inorganic and organic acids and bases. If other analytes were to be analyzed, it is recommended that all these chemometric methods are examined for effectiveness as analytes and sensors can have completely different mechanistic interactions that lead to different types of color changes, different RGB values, and data sets. However, based on the fact that PLS-DA was much more inferior to the other methods with only 39% accuracy, it may also not perform well for other datasets with a high number of analytes. Also, depending on the number of samples, LDA may not work because it requires a large data set, while KNN, HQI and SIMCA can accommodate smaller data sets. Finally, KNN can be employed if an easy algorithm and quick results are desired if a slightly lower accuracy is acceptable.

Table 10. Summary of quantitative chemometric analysis of data by method. LOO = Leave one out, all but one. T/T = Train and Test

Method	Validation	Trial%	(Avg) Classification %
LDA	LOO	N/A	99.2
KNN	LOO	N/A	98.1
HQI	LOO	N/A	98.3
SVM	T/T	88.6	85.0
		82.3	
		84.2	
SIMCA	T/T	96.8	92.4
		88.6	
		91.8	
RPART	T/T	95.6	96.4
		97.5	
		96.2	
PLS-DA	LOO	N/A	39.1

Herein nine chemometric methods were applied to the data set. The data set is provided in the supplemental information (SI.4) to the readers for analysis with the many other methods available for further processing and comparison. The methods that were reported here offer a suitable balance that was reached between data set requirements, analysis time, and robustness of response for our chemical classification application.

Acknowledgements

This publication was made possible by

1. US Army W911SR-15-C-0027 SBIR Phase I -Chemical Biological Radiological Nuclear and Explosives (CBRNE) Reconnaissance Sampling Kit (A15-048).
2. SBIR Phase II - Chemical Biological Radiological Nuclear and Explosives (CBRNE) Reconnaissance Sampling Kit (W911SR-16-C-0051)
3. National Institute for General Medical Science (NIGMS) (5P20GM103427), a component of the National Institutes of Health (NIH). (RL INBRE Scholar, DRPP)
4. Camille and Henry Dreyfus Foundation (AH-2015 Dreyfus Teacher Scholar Award).

References

Ariza-Avidad, M., Salinas-Castillo, A., Cuéllar, M. P., Agudo-Acemel, M., Pegalajar, M. C., & Capitán-Vallvey, L. F. (2014). Printed Disposable Colorimetric Array for Metal Ion Discrimination. *Analytical Chemistry*, *86*(17), 8634–8641. https://doi.org/10.1021/ac501670f

Askim, J. R., Li, Z., LaGasse, M. K., Rankin, J. M., & Suslick, K. S. (2016). An optoelectronic nose for identification of explosives. *Chemical Science*, *7*(1), 199–206. https://doi.org/10.1039/C5SC02632F

Askim, J. R., Mahmoudi, M., & Suslick, K. S. (2013). Optical sensor arrays for chemical sensing: the optoelectronic nose. *Chemical Society Reviews*, *42*(22), 8649–8682. https://doi.org/10.1039/C3CS60179J

Balabin, R. M., Safieva, R. Z., & Lomakina, E. I. (2010). Gasoline classification using near infrared (NIR) spectroscopy data: Comparison of multivariate techniques. *Analytica Chimica Acta*, *671*(1–2), 27–35. https://doi.org/10.1016/j.aca.2010.05.013

Bang, J. H., Lim, S. H., Park, E., & Suslick, K. S. (2008). Chemically Responsive Nanoporous Pigments: Colorimetric Sensor Arrays and the Identification of Aliphatic Amines. *Langmuir*, *24*(22), 13168–13172. https://doi.org/10.1021/la802029m

Batres, G., Jones, T., Johnke, H., Wilson, M., Holmes, A. E., & Sikich, S. (2014). Reactive Arrays of Colorimetric Sensors for Metabolite and Steroid Identification. *Journal of Sensor Technology*, *4*(1), 1–6. https://doi.org/10.4236/jst.2014.41001

Bell, G. (2016). Non-ionic surfactant phase diagram prediction by recursive partitioning. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2072). https://doi.org/10.1098/rsta.2015.0137

Brereton, R. G., & Lloyd, G. R. (2014). Partial least squares discriminant analysis: taking the magic away. *Journal of Chemometrics*, *28*(4), 213–225. https://doi.org/10.1002/cem.2609

Bueno, L., Meloni, G. N., Reddy, S., & Paixão, T. R. L. C. (2015). Use of plastic-based analytical device, smartphone and chemometric tools to discriminate amines. *RSC Advances*, *5*(26), 20148–20154. https://doi.org/10.1039/c5ra01822f

Burks, R. M., Pacquette, S. E., Guericke, M. A., Wilson, M. V., Symonsbergen, D. J., Lucas, K. A., & Holmes, A. E. (2010). DETECHIP: A Sensor for Drugs of Abuse. *Journal of Forensic Sciences*, *55*(3), 723–727. https://doi.org/10.1111/j.1556-4029.2010.01323.x

Capel-Cuevas, S., Cuéllar, M. P., Orbe-Payá, I. de, Pegalajar, M. C., & Capitán-Vallvey, L. F. (2010). Full-range optical pH sensor based on imaging techniques. *Analytica Chimica Acta*, *681*(1–2), 71–81. https://doi.org/10.1016/j.aca.2010.09.033

Capitán-Vallvey, L. F., López-Ruiz, N., Martínez-Olmos, A., Erena, M. M., & Palma, A. J. (2015). Recent developments in computer vision-based analytical chemistry: A tutorial review. *Analytica Chimica Acta*, *899*, 23–56. https://doi.org/10.1016/j.aca.2015.10.009

Charifson, P. S., & Walters, W. P. (2014). Acidic and Basic Drugs in Medicinal Chemistry: A Perspective. *Journal of Medicinal Chemistry*, *57*(23), 9701–9717. https://doi.org/10.1021/jm501000a

Chulvi, K., Gaviña, P., Costero, A. M., Gil, S., Parra, M., Gotor, R., … Vivancos, J. L. (2012). Discrimination of nerve gases mimics and other organophosphorous derivatives in gas phase using a colorimetric probe array. *Chemical Communications*, *48*(81), 10105–10107. https://doi.org/10.1039/C2CC34662A

Esbensen, K. H., Guyot, D., Westad, F., & Houmøller, L. P. (2010). *Multivariate Data Analysis - in practice* (Vol. 5). CAMO Sosftware.

Feng, L., Musto, C. J., Kemling, J. W., Lim, S. H., & Suslick, K. S. (2010). A colorimetric sensor array for identification of toxic gases below permissible exposure limits. *Chemical Communications*, *46*(12), 2037–2039. https://doi.org/10.1039/B926848K

Galpothdeniya, W. I. S., Regmi, B. P., McCarter, K. S., de Rooy, S. L., Siraj, N., & Warner, I. M. (2015). Virtual Colorimetric Sensor Array: Single Ionic Liquid for Solvent Discrimination. *Analytical Chemistry*, *87*(8), 4464–4471. https://doi.org/10.1021/acs.analchem.5b00714

Gemperline, P. (Ed.). (2006). *Practical Guide to Chemometrics* (2nd ed.). Boca Raton: CRC Press.

Gilbert, J. C., & Martin, S. F. (2010). *Experimental Organic Chemistry: A Miniscale and Microscale Approach* (5th ed.). Boston: Cengage Learning.

Graham, R. C. (1993). *Data Analysis for the Chemical Sciences: A Guide to Statistical Techniques*. New York: VCH Publishers, Inc.

Gryniewicz-Ruzicka, C. M., Rodriguez, J. D., Arzhantsev, S., Buhse, L. F., & Kauffman, J. F. (2012). Libraries, classifiers, and quantifiers: A comparison of chemometric methods for the analysis of Raman spectra of contaminated pharmaceutical materials. *Journal of Pharmaceutical and Biomedical Analysis*, *61*, 191–198. https://doi.org/10.1016/j.jpba.2011.12.002

Johnke, H., Batres, G., Wilson, M., Holmes, A. E., & Sikich, S. (2013). Detecting Concentration of Analytes with DETECHIP: A Molecular Sensing Array. *Journal of Sensor Technology*, *3*(3), 94–99. https://doi.org/10.4236/jst.2013.33015

Kakhki, J. F., & Abedi, M. R. (2012). Application of soft and hard modeling methods to resolve the three competitive complex formation of 13 lanthanide-Arsenazo III complexes. *International Journal of Industrial Chemistry*, *3*(1), 9. https://doi.org/10.1186/2228-5547-3-9

Kangas, M. J., Burks, R. M., Atwater, J., Lukowicz, R. M., Garver, B., & Holmes, A. E. (2018). Comparative chemometric analysis for classification of acids and bases via a colorimetric sensor array. *Journal of Chemometrics*, in press. https://doi.org/10.1002/cem.2961

Kangas, M. J., Burks, R. M., Atwater, J., Lukowicz, R. M., Williams, P., & Holmes, A. E. (2017). Colorimetric Sensor Arrays for the Detection and Identification of Chemical Weapons and Explosives. *Critical Reviews in Analytical Chemistry*, *47*(2), 138–153. https://doi.org/10.1080/10408347.2016.1233805

Khan, J. I., Kennedy, T. J., & Christian, D. R., Jr. (2012). *Basic Principles of Forensic Chemistry*. New York: Humana Press.

Kitamura, M., Shabbir, S. H., & Anslyn, E. V. (2009). Guidelines for Pattern Recognition Using Differential Receptors and Indicator Displacement Assays. *The Journal of Organic Chemistry*, *74*(12), 4479–4489. https://doi.org/10.1021/jo900433j

Kovar, K.-A., & Laudszun, M. (1989). *Chemistry and Reaction Mechanisms of Rapid Tests for Drugs of Abuse and Precursors Chemicals* (Scientific and Technical Notes No. 6) (p. 19). New York: United Nations Office on Drugs and Crime. Retrieved from https://www.unodc.org/pdf/scientific/SCITEC6.pdf

Lee, S., Lee, H., & Chung, H. (2013). New discrimination method combining hit quality index based spectral matching and voting. *Analytica Chimica Acta*, *758*, 58–65. https://doi.org/10.1016/j.aca.2012.10.058

Li, Z., Jang, M., Askim, J. R., & Suslick, K. S. (2015). Identification of accelerants, fuels and post-combustion residues using a colorimetric sensor array. *Analyst*, *140*(17), 5929–5935. https://doi.org/10.1039/C5AN00806A

Lim, S. H., Feng, L., Kemling, J. W., Musto, C. J., & Suslick, K. S. (2009). An optoelectronic nose for the detection of toxic gases. *Nature Chemistry*, *1*(7), 562–567. https://doi.org/10.1038/nchem.360

Lyon, M., Wilson, M. V., Rouhier, K. A., Symonsbergen, D. J., Bastola, K., Thapa, I., … Jackson, A. (2012). Digital Image Analysis for DETECHIP® Code Determination. *Signal & Image Processing : An International Journal (SIPIJ)*, *3*(4), 51–63. https://doi.org/10.5121/sipij.2012.3405

Ma, C.-M., Yang, W.-S., & Cheng, B.-W. (2014). How the Parameters of K-nearest Neighbor Algorithm Impact on the Best Classification Accuracy: In Case of Parkinson Dataset. *Journal of Applied Sciences*, *14*(2), 171–176. https://doi.org/10.3923/jas.2014.171.176

Mahmoudi, M., Lohse, S. E., Murphy, C. J., & Suslick, K. S. (2016). Identification of Nanoparticles with a Colorimetric Sensor Array. *ACS Sensors*, *1*(1), 17–21. https://doi.org/10.1021/acssensors.5b00014

Miller, D. W. (2001). Results of a New Classification Algorithm Combining K Nearest Neighbors and Recursive Partitioning. *Journal of Chemical Information and Computer Sciences*, *41*(1), 168–175. https://doi.org/10.1021/ci0003348

Minami, T., Esipenko, N. A., Akdeniz, A., Zhang, B., Isaacs, L., & Anzenbacher, Jr., P. (2013). Multianalyte Sensing of Addictive Over-the-Counter (OTC) Drugs. *Journal of the American Chemical Society*, *135*(40), 15238–15243. https://doi.org/10.1021/ja407722a

Muehlethaler, C., Massonnet, G., & Esseiva, P. (2014). Discrimination and classification of FTIR spectra of red, blue and green spray paints using a multivariate statistical approach. *Forensic Science International*, *244*, 170–178. https://doi.org/10.1016/j.forsciint.2014.08.038

Nigsch, F., Bender, A., van Buuren, B., Tissen, J., Nigsch, E., & Mitchell, J. B. O. (2006). Melting Point Prediction Employing k-Nearest Neighbor Algorithms and Genetic Parameter Optimization. *Journal of Chemical Information and Modeling*, *46*(6), 2412–2422. https://doi.org/10.1021/ci060149f

R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rusinko, A., Farmen, M. W., Lambert, C. G., Brown, P. L., & Young, S. S. (1999). Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning. *Journal of Chemical Information and*

Computer Sciences, *39*(6), 1017–1026. https://doi.org/10.1021/ci9903049

Salinas, Y., Ros-Lis, J. V., Vivancos, J.-L., Martínez-Máñez, R., Aucejo, S., Herranz, N., … Garcia, E. (2014). A chromogenic sensor array for boiled marinated turkey freshness monitoring. *Sensors and Actuators B: Chemical*, *190*, 326–333. https://doi.org/10.1016/j.snb.2013.08.075

Salles, M. O., Meloni, G. N., de Aaujo, W. R., & Paixão, T. R. L. C. (2014). Explosive colorimetric discrimination using a smartphone, paper device and chemometrical approach. *Analytical Methods*, *6*(7), 2047–2052. https://doi.org/10.1039/c3ay41727a

Sanchez, G. (2013). *DiscriMiner: Tools of the Trade for Discriminant Analysis. R package version 0.1-29.* Retrieved from https://CRAN.R-project.org/package=DiscriMiner

Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*, *9*(7), 671–675. https://doi.org/10.1038/nmeth.2089

Soga, T., Jimbo, Y., Suzuki, K., & Citterio, D. (2013). Inkjet-Printed Paper-Based Colorimetric Sensor Array for the Discrimination of Volatile Primary Amines. *Analytical Chemistry*, *85*(19), 8973–8978. https://doi.org/10.1021/ac402070z

Soldat, D. J., Barak, P., & Lepore, B. J. (2009). Microscale Colorimetric Analysis Using a Desktop Scanner and Automated Digital Image Analysis Douglas. *Journal of Chemical Education*, *86*(5), 617–620. https://doi.org/10.1021/ed086p617

Suslick, K. S. (2004). An Optoelectronic Nose:"Seeing" Smells by Means of Colorimetric Sensor Arrays. *MRS Bulletin*, *29*(10), 720–725. https://doi.org/10.1557/mrs2004.209

Therneau, T., Atkinson, B., & Ripley, B. (2017). *rpart: Recursive Partitioning and Regression Trees. R package version 4.1-11.* Retrieved from https://CRAN.R-project.org/package=rpart

Todorov, V. (2016). *rrcovHD: Robust Multivariate Methods for High Dimensional Data. R package version 0.2-5.* Retrieved from https://CRAN.R-project.org/package=rrcovHD

Tominaga, Y. (1999). Comparative study of class data analysis with PCA-LDA, SIMCA, PLS, ANNs, and k-NN. *Chemometrics and Intelligent Laboratory Systems*, *49*(1), 105–115. https://doi.org/10.1016/S0169-7439(99)00034-9

United Nations International Drug Crime Programme. (1994). *Rapid Testing Methods of Drugs of Abuse: Manual for use by Narcotics Laboratory Personnel* (p. 116). Vienna. Retrieved from https://www.unodc.org/pdf/publications/st-nar-13-rev1.pdf

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth). New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

Wang, L., Le, X., Li, L., Ju, Y., Lin, Z., Gu, Q., & Xu, J. (2014). Discovering New Agents Active against Methicillin-Resistant Staphylococcus aureus with Ligand-Based Approaches. *Journal of Chemical Information and Modeling*, *54*(11), 3186–3197. https://doi.org/10.1021/ci500253q

Weiner, J. (2017). *pca3d: Three Dimensional PCA Plots. R package version 0.10.* Retrieved from http://CRAN.R-project.org/package=pca3d

Wold, S., Johansson, E., Jellum, E., Bjørnson, I., & Nesbakken, R. (1981). Application of simca multivariate data analysis to the classification of gas chromatographic profiles of human brain tissues. *Analytica Chimica Acta*, *133*(3), 251–259. https://doi.org/10.1016/S0003-2670(01)83199-8

Zhang, Y., Askim, J. R., Zhong, W., Orlean, P., & Suslick, K. S. (2014). Identification of pathogenic fungi with an optoelectronic nose. *Analyst*, *139*(8), 1922–1928. https://doi.org/10.1039/C3AN02112B