J Forensic Sci, September 2008, Vol. 53, No. 5
doi: 10.1111/j.1556-4029.2008.00824.x
Available online at: www.blackwell-synergy.com

Nicholas D. K. Petraco,[1,2] *Ph.D.; Mark Gil,*[3] *M.S.; Peter A. Pizzola,*[4] *Ph.D.; and T. A. Kubic,*[1,2] *Ph.D.*

Statistical Discrimination of Liquid Gasoline Samples from Casework

ABSTRACT: The intention of this study was to differentiate liquid gasoline samples from casework by utilizing multivariate pattern recognition procedures on data from gas chromatography-mass spectrometry. A supervised learning approach was undertaken to achieve this goal employing the methods of principal component analysis (PCA), canonical variate analysis (CVA), orthogonal canonical variate analysis (OCVA), and linear discriminant analysis. The study revealed that the variability in the sample population was sufficient enough to distinguish all the samples from one another knowing their groups a priori. CVA was able to differentiate all samples in the population using only three dimensions, while OCVA required four dimensions. PCA required 10 dimensions of data in order to predict the correct groupings. These results were all cross-validated using the "jackknife" method to confirm the classification functions and compute estimates of error rates. The results of this initial study have helped to develop procedures for the application of multivariate analysis to fire debris casework.

KEYWORDS: forensic science, gasoline, chromatography, multivariate, pattern recognition, classification

In fire investigations, gasoline is the most commonly encountered ignitable liquid. Ignitable liquids are petroleum based or related products that have certain flammable or combustible properties. Ignitable liquids are commonly used as accelerants in arsons to initiate or promote the spread of the fire.

If an individual sample can be discriminated from the larger group, this can be of forensic interest. In fire debris analysis casework, liquid gasoline samples recovered during a fire investigation have an unknown history and are subjected to various real world conditions. These conditions ultimately introduce some variation to the liquid gasoline sample as compared to gasoline from a service station. Liquid gasoline samples recovered during fire investigations were examined in order to determine the magnitude of variability present in these types of samples. The goal of this study was to determine whether liquid gasoline samples from casework could be discriminated using statistical methods. Because of the sample size and the number of components present in gasoline, traditional methods of peak comparisons of the analytical data are difficult and time consuming.

Gasoline is a volatile flammable liquid hydrocarbon mixture of over 400 compounds (1). Depending on the production date of the gasoline, these compounds may also include certain oxygenates such as methyl-*tert*-butyl-ether (MTBE) or ethanol. Gasoline is produced in oil refineries from material that is separated from crude oil via cracking and distillation and then subjected to various processes that refine the product. The resulting reformulated gasoline base fuel is a carefully blended and formulated product that must meet specific guidelines and regulations as to its physical properties and engine performance characteristics (2). It is generally distributed through pipelines to storage terminals as reformulated gasoline base fuel to which an additive package is later blended in by the individual gasoline companies. The additives in these packages are added to the gasoline base fuel to either enhance fuel performance or to correct deficiencies arising from its production (3).

The basis of the published research in the comparison of gasoline samples has been the variability of the refining operations and blending processes. Mann conducted research to determine the variability in gasoline samples utilizing heated headspace concentration using a 60 m column in GC-FID (4). He noted significant variation in the alkylate region between C4 and C8. The comparison methodology was based on a qualitative and quantitative examination (4). The qualitative examination was conducted first to determine whether to include or exclude a sample based on visual examination. If the samples appeared to be similar, the quantitative examination was conducted using a sequential peak normalization procedure. The research demonstrated the potential of gasoline comparisons to eliminate or associate gasoline samples in forensic casework. The limitations of the study include a comparison threshold of up to 75% weathered sample and samples that had undergone significant degradation. At best, the methodology could only determine limited positive associations (5).

Recently, multivariate analysis techniques have been applied in an attempt to differentiate liquid gasoline samples under controlled conditions (6–12). Multivariate analysis is the simultaneous statistical consideration of relationships among many measured properties of a given system (13). It is typically applied to high order systems, or systems that have more than one dependent variable, where the variability cannot be observed by visual means or analyzed by univariate measures. This type of analysis allows one to focus on the variables that best summarize a data set by using as few variables as possible. The summary of the data is based on how the measured variables, within a single observation of a sample, vary with each other. This snapshot of variability is useful in testing hypotheses about the data quickly and efficiently. Inferences about the data set can then be made based on observed patterns or trends (14,15).

[1]Department of Science, John Jay College of Criminal Justice, City University of New York, 899 10th Avenue, New York, NY 10019.
[2]Faculty of Chemistry, Graduate Center, City University of New York, 365 5th Avenue, New York, NY 10016.
[3]New York City Police Department Crime Laboratory, Fire Debris Analysis Unit, 150-14 Jamaica Avenue, Jamaica, NY 11432.
[4]New York City Police Department Crime Laboratory, 150-14 Jamaica Avenue, Jamaica, NY 11432.

Received 18 July 2007; and in revised form 11 Jan. 2008; accepted 28 Jan. 2008.

Sandercock and Du Pasquier (9) conducted research to determine the variability of polar compounds and polycyclic aromatic compounds present in gasoline samples utilizing solid phase microextraction and gas chromatography-mass spectrometry (GC-MS). The researchers analyzed the data using two-dimensional (2D) and three-dimensional (3D) principal component analysis (PCA) and linear discriminant analysis (LDA) with cross-validation. The results of this multivariate analysis highlighted the variation in the polycyclic aromatic hydrocarbon content across the sample set. The reported overall correct classification rate was 96–100% depending on the location of the collected samples.

Also recently, a covariance mapping technique has been applied to aid in computer-assisted pattern recognition procedures for GC-MS data (11,12). This technique was used for both ignitable liquid samples and samples collected from the passive headspace procedure. Using the entire chromatogram, including the light volatile components, the authors were able to distinguish between 10 different samples of liquid gasoline.

GC-MS is utilized by the majority of forensic laboratories conducting fire debris analysis. Typically, the laboratory's procedures refer to or reflect the guidelines set forth by ASTM-E 1618-06: Standard Test Method for Ignitable Liquid Residues from Fire Debris Samples by GC-MS (16). The comparison methodology in this research, therefore, was designed to incorporate the latter instrumentation.

This study was undertaken to examine the variability of gasoline components in twenty retained liquid gasoline samples from fire investigations in the New York area via gas chromatography coupled with a mass spectrometer (GC-MS). The objective of multivariate analysis in this research was to reduce the dimensions of the generated GC-MS data and to determine if any correlations existed between the variables and or between the samples drawn from the general population. A selection of the common components present in gasoline was utilized to test the discrimination potential of the multivariate methods. The multivariate methods employed in this study were PCA, canonical variate analysis (CVA), and orthogonal canonical variate analysis (OCVA) coupled with LDA for numerical discrimination.

Methodology

Fifteen peaks were chosen in this study that represented the common components present in gasoline. The peaks and their identity are given in Table 1. Gasoline contains abundant aromatics in a specific pattern such as ethylbenzene, *m*- & *p*- xylenes, *o*-xylene, and the tri-substituted methyl benzenes, specifically 1,2,4 trimethylbenzene (16). Keto and Wineman identified target compounds present in gasoline including the isomers of trimethylbenzene and tetramethylbenzene (17). The latter literature sources were used as a guideline to select the compounds of interest. The volatility of the compounds chosen was also taken into consideration and was expected to remain consistent up to roughly 75% evaporated (weathered) by volume based on prior research (4). Twenty liquid gasoline samples from casework were chosen for this initial study. The first eight samples were analyzed using seven replicate analyses per sample. For the remaining 12 samples, three replicate analyses per sample were performed in order to test the discrimination power of the multivariate procedures using different numbers of replicates.

The liquid gasoline casework samples were stored in amber colored teflon lined screw top vials and kept in a storage cabinet to limit light exposure. All samples were verified as containing gasoline prior to statistical analysis according to the criteria outlined in ASTM-E 1618-06 (16).

TABLE 1—*The* m/z *values of the fifteen peaks integrated in each chromatogram of gasoline plus a deuterated internal standard.*

Ret. Time (±0.05 min)	Chemical Compound/Class
11.54*	Ethylbenzene, d10*
11.70	Ethylbenzene
11.98	*m*- & *p*-xylene
12.68	*o*-xylene
14.62	Propylbenzene
14.85	*m*-ethyltoluene
14.91	C3 alkyl benzene, unidentified
15.08	C3 alkyl benzene, unidentified
15.36	*o*-ethyltoluene
15.81	1,2,4 trimethylbenzene
16.59	1,2,3 trimethylbenzene
17.40	C4 alkyl benzene, unidentified
17.48	C4 alkyl benzene, unidentified
17.59	C4 alkyl benzene, unidentified
17.66	C4 alkyl benzene, unidentified
18.38	C4 alkyl benzene, unidentified

*Deuterated internal standard.

All the samples in this study were analyzed directly on the GC-MS using 1:50 dilutions of the gasoline samples in an internal standard solution mixture. The internal standard solution mixture consisted of deuterated aromatics, specifically benzene-d_6 40% (w/w), ethylbenzene-d_{10} 40% (w/w), and naphthalene-d_8 20% (w/w) and prepared at a dilution of (1:500) in carbon disulfide (CS_2). A Hewlett Packard 6890 series GC with a 5973 series MS detector was used to analyze all the samples with the following conditions: column, HP-1 Methyl Siloxane 60 m length × 250 μm bore × 0.25 μm film thickness; flow rate, 1.6 mL/min; carrier gas: Helium; injection volume: 1 μL; split ratio, 20:1; temperature programming: 35°C 4 min, 6°C/min to 260°C, hold 12 min; total run time: 53.50 min.

The data were preprocessed before being introduced into the multivariate procedures. The peak areas for the 15 peaks in the 92 chromatograms were determined by using the RTE integrator within the HP Chemstation software.

The area of each peak was scaled relative to the area of the ethylbenzene-d_{10} peak with a retention time of 11.54 minutes in all the chromatograms. The scaled areas of the peaks were assembled into a 15 column by 92 row data matrix for use with PCA, CVA, OCVA, and LDA.

Statistical Methods

The normalized integration results from the gas chromatographic analysis were first arranged into an $n \times p$ data matrix ($\mathbf{X}$) for analysis:

$$\mathbf{X} = \begin{bmatrix} X_{11} & .. & X_{1j} & .. & X_{1p} \\ : & & : & & : \\ X_{i1} & .. & X_{ij} & .. & X_{ip} \\ : & & : & & : \\ X_{n1} & .. & X_{nj} & .. & X_{np} \end{bmatrix}$$

where $n = 92$ is the number of chromatograms and $p = 15$ is the number of peaks in the chromatogram. Each X_{ij} represents an area under peak j in chromatogram i. The symbol $\mathbf{X}_i$ designates row i of $\mathbf{X}$ and is a vector of data representing chromatogram i. The average of all row vectors in $\mathbf{X}$ is the average vector $\overline{\mathbf{X}}$. Multivariate statistical methods were used to transform the data set ($\mathbf{X}$) into a new data set ($\mathbf{Z}$) that contains the optimal variables from the original data set that accounts for a majority of the variation. These

optimal variables retain most of the underlying structure of the original data set that may be used to discriminate between different samples of gasoline. In this study, $k = 20$ different samples of gasoline were analyzed. The raw integration data were processed using computer programs written using the Mathematica version 5.1 computer algebra system. The Mathematica notebooks developed for this study are available upon request from the authors. The multivariate analysis of data set (**X**) undertaken in this study were PCA, CVA, OCVA, and LDA.

PCA is a multivariate procedure that is used to reduce the dimensionality of a data set (**X**) to a new data set ($\mathbf{Z}_{PC}$) of "derived variables" (13,18). The derived variables are linear combinations of the original variables

$$Z_{ij} = \sum_{l=1}^{p} a_{il} X_{il}$$

or in matrix form

$$\mathbf{Z}_{PC} = \mathbf{X}\mathbf{A}_{PC}^{T}$$

where the superscript T is the transpose of $\mathbf{A}_{PC}$. The subscript PC will now be dropped for typographical convenience. This transformation simply rotates the coordinate axes in feature space and the above equation is a transformation of the data (**X**) into the basis of principal components. The entire set of derived variables is equivalent to the original data (**X**). The new data set (**Z**), however, orders the variables (columns) according to the amount of variance of the data set they contain, from highest to lowest. If the first few variables in **Z** contain a majority of the variance, then the remaining variables can be deleted with a minimum loss of information contained in the data. The dimensionality of the data set is then effectively reduced to include only those variables that adequately represent the data. As a note of caution, while the information contained in the low variance variables of **Z** that are removed may not be important to the overall structure of the data, they may contain the derived features needed to discriminate between different samples (18). This issue is discussed further in the Results and Discussion section below.

The matrix **A** contains the p principal components as rows and was computed by diagonalizing the $p \times p$ covariance matrix (**S**) of **X**

$$\mathbf{S}\mathbf{A}^{T} = \mathbf{A}^{T}\mathbf{\Lambda}$$

Standard eigenvector and eigenvalue routines were used to determine the PCs **A** (eigenvectors of **S**) and their variances $\mathbf{\Lambda}$ (eigenvalues of **S**) (19). The PCs were all normalized to unity. The maximum likelihood covariance matrix, **S** used in this study was computed as

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{X}_i - \overline{\mathbf{X}}) \otimes (\mathbf{X}_j - \overline{\mathbf{X}})^{T}$$

where $\otimes$ is the Kronecker (direct) product of vectors (13).

The ratio of eigenvalues

$$\lambda_i \Big/ \sum_{j=1}^{p} \lambda_j \; \mathbf{\Lambda} = diag(\lambda_1, \lambda_2, \ldots, \lambda_p)$$

gives the proportion of variance explained by the ith principal component and is useful in selecting the number of principal components required to adequately represent the data. No prior grouping of sample chromatograms was assumed in the PCA computations.

CVA (also called Fisher linear discriminant analysis) seeks to characterize the ratio of between group variance (**B**) to within group variance (**W**) (13,20). Unlike PCA, CVA requires that at least some of the sample groups are known a priori in order to characterize the variations in the data set. These a priori labeled samples act as a training set in order to compute the canonical variates (CVs). Geometrically, the CVs define axes onto which the data are projected that best separate the samples into discrete clusters (13,20). In p-dimensional space, p CVs can be computed. However, CVA can be used to reduce the dimensionality of the data by retaining only the first few CVs. Additionally, like PCA, CVA can be formulated as an eigenvector-eigenvalue problem with the magnitude of the eigenvalues providing a guide as to the number of CVs to be retained. The CVs, $\mathbf{A}_{CV}$, and their eigenvalues $\mathbf{\Lambda}_{CV}$ were computed by diagonalizing the matrix $\mathbf{W}^{-1}\mathbf{B}$ with

$$\mathbf{B} = \sum_{i=1}^{k} n_i (\overline{\mathbf{X}}_i - \overline{\mathbf{X}}) \otimes (\overline{\mathbf{X}}_i - \overline{\mathbf{X}})^{T}$$

and

$$\mathbf{W} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\mathbf{X}_{ij} - \overline{\mathbf{X}}_i) \otimes (\mathbf{X}_{ij} - \overline{\mathbf{X}}_i)^{T}$$

The subscript CV will now be dropped for typographical convenience. A standard inversion method was used to invert nonsingular **W** (19). $\mathbf{X}_{ij}$ represents the jth chromatogram in the ith sample and $\overline{\mathbf{X}}_i$ is the average of all the chromatograms in the ith sample. Note that there are n_i replicate chromatograms in sample i. Because the eigenproblem for CVA

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{A}^{T} = \mathbf{A}^{T}\mathbf{\Lambda}$$

is not symmetric, its eigenvectors are not guaranteed to be orthogonal. Thus unlike in PCA the CVs are not necessarily at right angles to each other (13). However, the eigenvectors in **A** are normalized to unity. The data is then transformed to the basis of (retained) CVs as

$$\mathbf{Z} = \mathbf{X}\mathbf{A}^{T}$$

OCVA seeks to find axes in p-dimensional feature space (defined by the data set) which maximize the ratio of between-group to within-group variance as in CVA (21). OCVA, however, has the additional restriction that the axes it finds are orthogonal to each other as in PCA. Because of the added constraint the OCVA procedure cannot be formulated as an eigenproblem and was a bit more difficult to program. As prescribed by Krzanowski (21), we first form the equation of weighted variables, V for the ith OCV

$$V = \mathbf{e}_i^{T}\mathbf{B}\mathbf{e}_i / \mathbf{e}_i^{T}\mathbf{W}\mathbf{e}_i$$

where $\mathbf{e}_i$

$$\mathbf{e}_i^{T} = (y_1, y_2, \ldots, y_p)_i$$

is a vector of variables to be determined. Next the Lagrangian

$$L(\mathbf{e}_i, \lambda) = V - \sum_{j=1}^{i} \lambda_j (\mathbf{e}_j^{T}\mathbf{e}_j - 1)$$

is formed and maximized to determine the values of the variables in $\mathbf{e}_i$ and Lagrange multipliers λ_i. A standard maximization routine is used to perform this optimization task (19). The resulting OCVs, $\mathbf{e}_i$ are collected (as rows) into the matrix $\mathbf{A}_{\mathrm{OCV}}$. The subscript OCV will now be dropped for typographical convenience. The above procedure was repeated for the desired number of OCVs. A total of p OCVs can be computed; however, like the latter multivariate methods described above, the goal is to reduce the dimensionality of the data so that it can be visualized in Cartesian plots. Like PCA and CVA the OCVs can be ordered in increasing importance. The Lagrange multipliers serve this task in OCVA; however, they no longer represent variances as in PCA (21). Typically, only the top 2, 3, or 4 OCVs are retained. The data in $\mathbf{X}$ can then be transformed to the basis of (retained) OCVs as

$$\mathbf{Z} = \mathbf{X}\mathbf{A}^{\mathrm{T}}$$

The samples in the data contained in the original data matrix $\mathbf{X}$ or any of the derived data matrices $\mathbf{Z}$ can be numerically discriminated between using LDA (also called classification analysis) (13). This decision model uses a distance function to find the mean vector $\overline{\mathbf{Y}}_i$ (average chromatogram of sample i contained in matrix $\mathbf{X}$ or any of the derived data matrices $\mathbf{Z}$) that is closest to the "test" chromatogram $\mathbf{Y}_j$ (original or derived). The distance metric used is of the Mahalanobis type

$$D^2(\mathbf{Y}_j) = (\mathbf{Y}_j - \overline{\mathbf{Y}}_i)^{\mathrm{T}}\mathbf{S}_{\mathrm{pl}}^{-1}(\mathbf{Y}_j - \overline{\mathbf{Y}}_i)$$

which employs a pooled covariance matrix for all the samples

$$\mathbf{S}_{\mathrm{pl}} = \frac{1}{n-k}\sum_{i=1}^{k}(n_i - 1)\mathbf{S}_i$$

$\mathbf{S}_i$ is the covariance matrix for sample i. No singular $\mathbf{S}_{\mathrm{pl}}$ were encountered in this study. The actual discriminant function constructed for sample i is given as

$$L_i(\mathbf{Y}_j) = \overline{\mathbf{Y}}_i^{\mathrm{T}}\,\mathbf{S}_{\mathrm{pl}}^{-1}\,\mathbf{Y}_j - \frac{1}{2}\,\overline{\mathbf{Y}}_i^{\mathrm{T}}\,\mathbf{S}_{\mathrm{pl}}^{-1}\,\overline{\mathbf{Y}}_i$$

Thus LDA is essentially a method to "train" a set of linear functions, L_i to be able to recognize which data group a particular pattern came from (i.e., a "supervised" machine learning technique). For this study, the data groups are the different samples of gasoline and the patterns are the areas of the chosen fifteen peaks in the gas chromatograms. A total of k = 20 discriminant functions were constructed, one for each sample of gasoline. Chromatogram j is then assigned to sample i according to the decision rule

$$\arg\max_i L_i(\mathbf{Y}_j)$$

i.e., assign $\mathbf{Y}_j$ to sample i whose discriminant function yields the largest numerical value (13). In other words, this decision rule means: "the chromatogram $\mathbf{Y}_j$ is most similar to the set of chromatograms from gasoline sample i."

The ability of discriminant functions to accurately predict the sample identity of a pattern which they have not been trained with, is called classification error analysis (22). This is a very important topic whenever statistical pattern recognition techniques are applied to evidence in forensic science. The reason is because discriminant functions are necessarily trained on a finite (probably small) set of data. The functions, however, will be used to classify or identify a piece of evidence (data) that they have not been trained with. Thus, rigorously derived accurate estimates for error rates of computed sets of discriminant functions are critical in forensic science applications. For this study, we estimate the error rates of the k discriminant functions in three different ways. We actually compute estimates of the "correct classification rate" which is one minus the error rate and is reported as a percentage.

The first estimate used is the "apparent" correct classification rate computed by determining the number of chromatograms assigned to their correct sample (by the discriminant functions) divided by the total number of chromatograms. This performance estimate is known to be biased and tends to yield an overly optimistic correct classification rate (13).

The second estimate is the overall "hold-one-out" correct classification rate (14,20). This is computed by first recalculating the linear discriminant functions omitting a single chromatogram from the data set. Thus, the recalculated discriminant functions are not trained to identify the held out chromatogram. This omitted chromatogram is then classified with the recalculated linear discriminant functions and the process is repeated sequentially for each chromatogram in the data set. The number of correctly classified "held-out" chromatograms is divided by the total number of chromatograms in the entire data set (92 in this study) to yield the overall hold-one-out correct classification rate.

Finally the "jackknife" correct classification rate is computed. "Jackknifing" a data set is the process of replicating a data set composed of n observation vectors, n-times. Each replicate data set, however, contains all but one of the original data vectors (23). Thus the jackknife method employs the hold-one-out process. The n "jackknifed" data sets are then used to recalculate a statistic on that data set in the absence of the deleted data vector, producing a set of estimates of the statistic. The set can then be used to produce an average and standard deviation for the statistic (23). The jackknife correct classification rate is mathematically the least biased estimation of the discrimination functions' classification performance (13).

Here we compute the jackknife correct classification rate by first computing all the samples' hold-one-out correct classification rates and recording them in a "jackknife cross-validation table." Next the average and standard deviation of the samples' hold-one-out correct classification rates is found yielding the jackknife correct classification rate (14,20,23). The difference between the jackknife correct classification rates for samples containing three replicates and seven replicates was analyzed using the Student's t-test (24).

Results and Discussion

PCA

PCA is typically used to reduce the number of dimensions in multivariate data. The dimensions that were excluded were those describing the least amount of variance. The method minimizes the dimensionality of the data by discounting variables with minimal contributions to the overall spread of the data. As a result, the more highly correlated the data variables are, the fewer high variance PCs there will be, and thus, the lower the dimension of the subspace in which most of the structure of the data resides (18). If the number of dimensions can be reduced to two or three, then the data can be graphically visualized in Cartesian plots. The scatter of data points in these graphical plots may cluster into groups; however, it is important to note that PCA is not itself a clustering technique. PCA in this application simply projects the compressed data into a subspace where clusters of data, if present in that subspace, can be visualized.

TABLE 2—*The fifteen principal components and corresponding eigenvalues ordered by their fractional variance.*

Principal Component Number	Eigenvalue	Fractional Variance (%)	Cumulative Variance (%)
1	0.2102970	88.9560	88.956
2	0.0198672	8.4039	97.360
3	0.0031579	1.3358	98.696
4	0.0018320	0.7749	99.471
5	0.0005960	0.2521	99.723
6	0.0003730	0.1578	99.881
7	0.0001080	0.0457	99.926
8	0.0000743	0.0314	99.958
9	0.0000396	0.0168	99.974
10	0.0000306	0.0129	99.987
11	0.0000103	0.0044	99.992
12	0.0000078	0.0033	99.995
13	0.0000063	0.0027	99.998
14	0.0000032	0.0014	99.999
15	0.0000024	0.0010	100.000

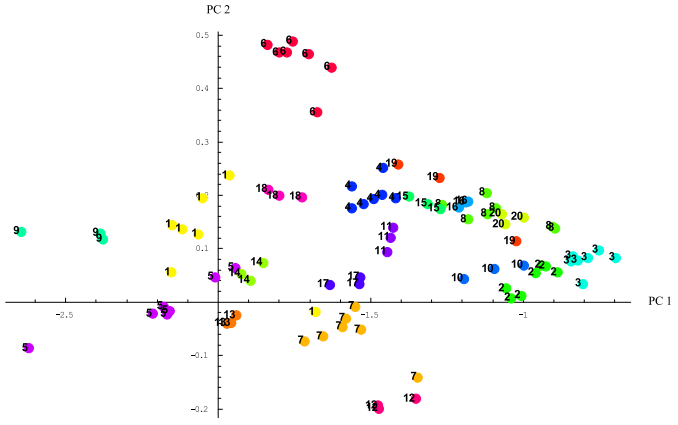

FIG. 1—*Chromatogram data projected into the space of the first two principal components (2D PCA plot).*

The variance associated with the 15 principal components is given in Table 2. Here, one can see that principal components 1 through 6 accounts for 99.9% of the variance structure in the data. Figure 1 shows the samples plotted against the first two principal components (97.4% of variance accounted for) and Fig. 2 shows the samples plotted against the first three principal components (98.7% of variance accounted for). Finally, while the jackknife correct classification rates between the 2D and 3D PCA data for the samples with seven replicates are not equivalent at the 95% level, they are equivalent at the 95% level for the samples with three replicates. Overall, some clustering of points was noted but well-formed (low intra-cluster spread) well-separated (high inter-cluster spread) was not readily apparent.

2D PCA

Figure 1 shows the projection of the data into the space of the first two principal components. While the first two principal components account for roughly 97.4% of the variance in the data, the clustering of some of the data points into well-defined groups containing only data from the same sample (especially in the first quadrant), is difficult to discern. 2D PCA performed reasonably well in confirming a cluster structure in the data considering how closely the samples were projected into the subspace.

Table 3 shows the jackknife cross-validation table for classification of each 2D principal component reduced gas chromatogram using LDA. The overall hold-one-out correct classification rate was 80% (82% apparent correct classification rate). The jackknife correct classification rate was 83 ± 24%. For the samples with seven replicates the jackknife correct classification rate was 75 ± 26%. For the samples with three replicates the jackknife correct classification rate was 89 ± 22%. These averages are statistically different at the 95% level of significance using the Student's *t*-distribution. However, such large standard deviations are not surprising considering the small number of replicate gas chromatograms generated for each sample.

As can be seen in Fig. 1, samples **4**, **8**, **15**, **16**, **19**, and **20** were the most intermingled. Samples **8** and **19** had the worst hold-one-out correct classification rates (29% and 33% respectively, cf. Table 3) where sample **8** is difficult to distinguish from samples **15**, **16**, and **20** while replicates from sample **19** are close to the tightly clustered samples of **4** and **20**.

3D PCA

Table 4 shows the jackknife cross-validation classification for the data projected into the space of the first three PCs. Clearly, the addition of the extra dimension provided by PC_3 (1.3% of the variance) markedly increases the correct classification rates, now 88% overall hold-one-out correct classification (93% apparent correct classification rate) up from 80% in 2D PCA. Note, however, how tightly packed the data are about PC_3 in Fig. 2. In fact, a 100% jackknife correct classification rate was not achieved until the PCA data reached 10 dimensions. Thus, for this study, we observed that complete discrimination between the samples is very subtle.

The jackknife correct classification rate was 88 ± 21% while with the seven replicate samples it was 87 ± 14%, and with the three replicate samples it was 89 ± 30%. These averages are statistically distinct at the 95% level of confidence. We can interpret this to mean that in order to obtain slightly more reliable classification rates, not only should one use the same number of replicates in each sample, but one should use as many replicates in the samples as possible.

Higher Dimensional PCA

Finally, we note that PCA required 10 dimensions order to yield a 100% jackknife correct classification rate. We pay particular note to this result because PCA is often used as a data preprocessing technique before some sort of discrimination procedure is performed. Dimensional reduction for our data set actually hindered discrimination in that we needed such a high dimensional feature space in order to obtain a perfect classification rate. In fact, we observed that as we added dimensions to the derived data set (i.e., retained more PCs) the jackknife correct classification rate was slow to converge to 100%. Here we note that using PCA first to reduce the dimension of data sets before processing with a discrimination algorithm may inadvertently remove necessary discriminating power. We recommend that, if possible, one should compare the jackknife correct classification rate for a discrimination method by *both* preprocessing and not preprocessing the raw data set with PCA. This will help to indicate if higher dimensions of the data, while low in variance, will aid in discrimination between different samples. Below we find that this is in fact the case with the data set in this study. CVA and OCVA have much better discrimination power at lower dimension than PCA when used on the raw data.

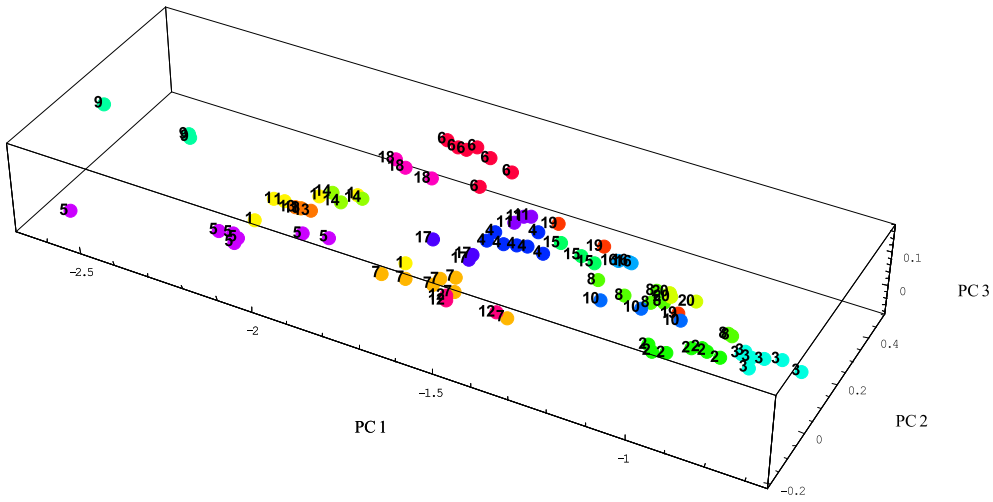

FIG. 2—*Chromatogram data projected into the space of the first three principal components (3D PCA plot).*

TABLE 3—*LDA jackknife cross-validation table for 2D PCA.*

Sample ID	Number of Replicates in Sample	Number of Misidentified Replicates in Sample	Incorrectly Predicted Sample IDs of Misidentified Replicates	Sample "Hold-One-Out" Correct Classification Rates (%)
1	7	3	**5**, **7**, **18**	57
2	7	1	**10**	86
3	7	0	None	100
4	7	0	None	100
5	7	3	**9**, **14** × 2	57
6	7	0	None	100
7	7	2	**12, 17**	71
8	7	5	**15**, **16** × 2, **20** × 2	29
9	3	0	None	100
10	3	1	**2**	67
11	3	0	None	100
12	3	0	None	100
13	3	0	None	100
14	3	0	None	100
15	3	0	None	100
16	3	0	None	100
17	3	0	None	100
18	3	0	None	100
19	3	2	**4**, **20**	33
20	3	1	**8**	67

Gasoline sample numbers shown in boldface. The jackknife correct classification rate was 83 ± 24%.

Thus preprocessing our raw data with PCA in order to reduce its dimension, before processing it with CVA and OCVA decreased the jackknife correct classification rates for these methods.

CVA

CVA, like PCA, is used to reduce the number of dimensions in multivariate data by utilizing linear combinations of functions. CVA's strength as a method is its ability to discriminate between groups in data. The scatter of data points in these graphical plots try to maximize the difference between groups by exploiting the inter- and intra-group variance in determining clustering. It is important to note that CVA is a clustering technique given that the groups are known a priori in a training sense. However, CVA has an additional property that it searches for planes on which to project the data which optimally separate it into groups with minimal information loss. CVA has the weaker property as compared to PCA or OCVA of not utilizing orthogonal Cartesian planes. As with PCA, if the number of dimensions can be reduced to two or three, then the data can be graphically visualized in Cartesian plots.

The eigenvalues associated with the 15 CVs are given in Table 5. Plots of the data in both 2D and 3D CV space revealed well-formed low intra-cluster spread and well-separated high inter-cluster spread that could be easily identified in two or three dimensions. In general, clustering of points was noted in 2D; however, only by adding the third dimension to the plot did the latter characteristics become readily apparent.

2D CVA

Figure 3 shows the projection of the data into the space of the first two CVs. In general, the majority of the data points separated into clear packed clusters, while a few of the groups showed some overlapping.

TABLE 4—*LDA jackknife cross-validation table for 3D PCA.*

Sample ID	Number of Replicates in Sample	Number of Misidentified Replicates in Sample	Incorrectly Predicted Sample IDs of Misidentified Replicates	Sample "Hold-One-Out" Correct Classification Rates (%)
1	7	2	**5, 7**	71
2	7	0	None	100
3	7	0	None	100
4	7	0	None	100
5	7	2	**1** × 2	71
6	7	0	None	100
7	7	1	**17**	86
8	7	2	**3** × 2	71
9	3	0	None	100
10	3	0	None	100
11	3	0	None	100
12	3	0	None	100
13	3	0	None	100
14	3	0	None	100
15	3	1	**19**	66
16	3	0	None	100
17	3	0	None	100
18	3	0	None	100
19	3	3	**15** × 2, **20**	0
20	3	0	None	100

Gasoline sample numbers shown in boldface. The jackknife correct classification rate was 88 ± 21%.

TABLE 5—*The fifteen eigenvalues associated with the fifteen CVs and ordered by their magnitude.*

Canonical Variate Number	Eigenvalue
1	278.3
2	201.1
3	109.5
4	79.1
5	54.7
6	35.0
7	27.6
8	13.2
9	10.0
10	3.2
11	1.6
12	1.1
13	0.7
14	0.3
15	0.1

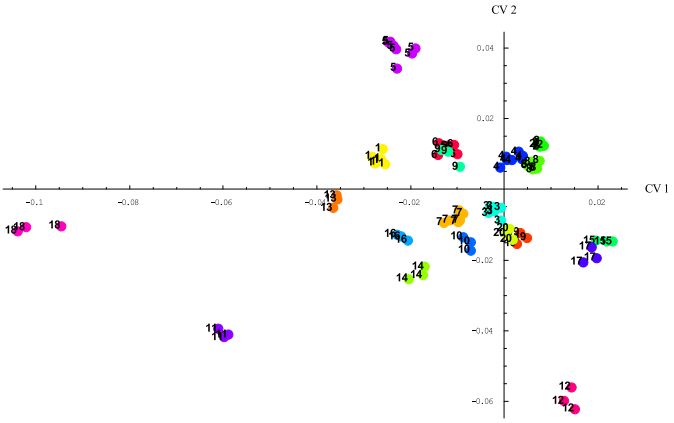

FIG. 3—*Chromatogram data projected into the space of the first two CVs (2D CVA plot).*

Table 6 shows the jackknife cross-validation table for classification of each 2D CV reduced gas chromatogram using LDA. The overall hold-one-out correct classification rate was 93% (97% apparent correct classification rate). The jackknife correct classification rate was 92 ± 18%. For the samples with seven replicates, the jackknife correct classification rate was 96 ± 10%. For the samples with three replicates, the jackknife correct classification rate was 89 ± 22%. These averages are statistically distinct at the 95% level of confidence.

From Fig. 3, one can notice that groups 9 and 6 were quite intermingled. Table 6 shows that the LDA classification algorithm confused these groups with one another which is confirmed by these samples' individual hold-one-out correct classification rates of 71% and 33% respectively (cf. Table 6). There was also slight overlapping noted between groups 15 and 17 as well as groups 19 and 20 as can be seen in Fig. 3 and Table 6.

3D CVA

The jackknife classification rate (and thus overall hold-one-out and apparent correct classification rates) for the data projected into the subspace of the first three CVs was 100% (thus no jackknife cross-validation table is given). Figure 4 shows the plot of this data in the space of the first three CV dimensions. While it may appear that there is some overlap between groups **12**, **6**, **13**, **19** and groups **10**, **7** and groups **9**, **5**, by viewing the data down each of the coordinate axes this issue can be resolved. These groups of gasoline actually form distinct clusters in 3D CV space. Alternative view points of Fig. 4 illustrating the fact that all twenty groups of gasoline form distinct clusters are available from the authors upon request.

For completeness we note here the angles between the (non-orthogonal) CV axes are 92° (CV_1, CV_2), 71° (CV_1, CV_3), and 69° (CV_2, CV_3). We computed the angle θ between each pair of normalized CV axes using the scalar (dot) product:

$$\theta = \arccos[CV_i \bullet CV_j]$$

Note that they are all displayed as 90° in Fig. 4. This practice is standard in the statistical literature because the distortion is generally assumed to be small (13).

TABLE 6—*LDA jackknife cross-validation table for 2D CVA and OCVA*.*

Sample ID	Number of Replicates in Sample	Number of Misidentified Replicates in Sample	Incorrectly Predicted Sample IDs of Misidentified Replicates	Sample "Hold-One-Out" Correct Classification Rates (%)
1	7	0	None	100
2	7	0	None	100
3	7	0	None	100
4	7	0	None	100
5	7	0	None	100
6	7	2	**9** × 2	71
7	7	0	None	100
8	7	0	None	100
9	3	2	**6** × 2	33
10	3	0	None	100
11	3	0	None	100
12	3	0	None	100
13	3	0	None	100
14	3	0	None	100
15	3	0	None	100
16	3	0	None	100
17	3	1	**15**	66
18	3	0	None	100
19	3	0	None	100
20	3	1	**19**	66

Gasoline sample numbers shown in boldface. The jackknife correct classification rate for both CVA and OCVA was 92 ± 21%.
*The jackknife cross-validation tables for CVA and OCVA were identical and thus only one table is given.

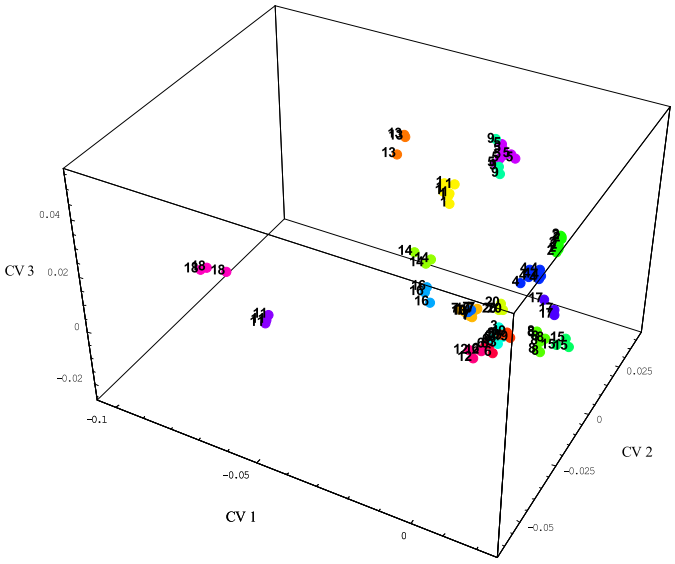

FIG. 4—*Chromatogram data projected into the space of the first three CVs (3D CVA plot).*

TABLE 7—*The fifteen Lagrange multipliers associated with the fifteen orthogonal CVs and ordered by their magnitude.*

Orthogonal Canonical Variate Number	Lagrange Multiplier
1	278.3
2	201.2
3	145.4
4	132.3
5	95.2
6	86.4
7	73.1
8	65.5
9	59.2
10	47.5
11	31.2
12	30.6
13	20.7
14	20.6
15	17.1

OCVA

OCVA, like CVA, is used to reduce the number of dimensions in multivariate data by utilizing combinations of linear functions that search for planes that optimally separate the data into groups. The method projects the data into planes that maximize the differences in groups by exploiting the inter and intra-group variance. OCVA, however, has the stronger property in that the OCV projection planes are orthogonal. The projection planes are not necessarily orthogonal in CVA. It is important to note that OCVA is a clustering technique, like CVA, given that the groups are known a priori.

OCVA can also be used to reduce the number of dimensions in multivariate data. The method does this by generating a set of derived variables which maximize the ratio of between group to within group variance. Derived variables corresponding to larger between group to within group ratios have larger Lagrange multipliers. Thus by retaining only those derived variables with the largest Lagrange multipliers we obtain a data set reflective of the widest possible inter-group separation and smallest intra-group separation. If only two or three OCVs are kept then the new data sets can be plotted and the clusters (if any) can be visualized.

The Lagrange multipliers associated with the fifteen orthogonal CVs are given in Table 7. Like CVA, the plots of the data in 2D and 3D OCV space revealed well formed low intra-cluster spread and well separated high inter-cluster spread that could be easily identified in two or three dimensions.

2D OCVA

In the space of the first two OCVs (cf. Fig. 5) the scatter of data points appears to be nearly identical to that of the 2D scatter of

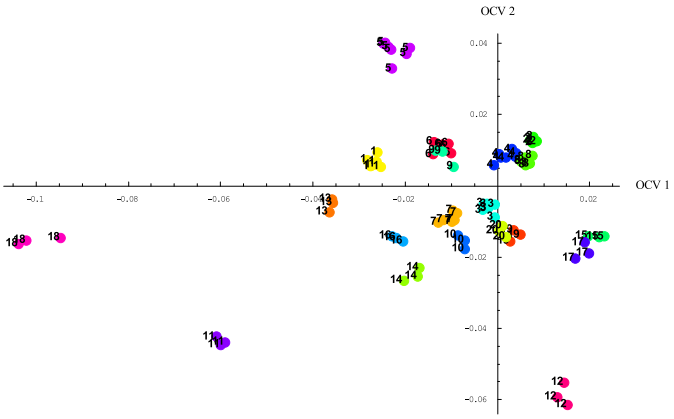

FIG. 5—*Chromatogram data projected into the space of the first two orthogonal CVs (2D OCVA plot).*

data points of CVA. The distortion of the 2D CVA plot as compared to the OCVA was not severe since the angles of the first two CVs were only a few degrees from being a right angle (92°). The axes of OCVA are always at 90° to one another. Thus it is not surprising that the jackknife cross-validation classification analysis for 2D OCVA, was identical to 2D CVA in this study (cf. Table 6).

3D OCVA

The graphical results for 3D OCVA are shown in Fig. 6. The spatial distribution of the data points differs to some degree from that obtained by 3D CVA. This makes sense in light of the fact that CVA seeks projections of the data into a subspace which best displays the separation of groups in a data set while OCVA has the added property of keeping the coordinate system of the data orthogonal. Therefore, it can be expected that there is some distortion between Figs. 4 and 6.

By viewing Fig. 6 along each of its coordinate axes it becomes apparent that some of the samples of gasoline are quite close and

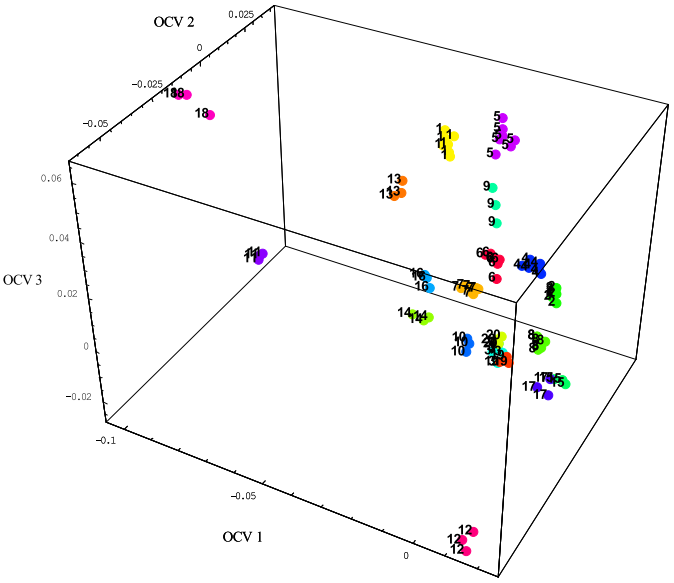

FIG. 6—*Chromatogram data projected into the space of the first three orthogonal CVs (3D OCVA plot).*

not entirely linearly separable in 3D OCV space. Alternative viewpoints of Fig. 6 are available from the authors upon request.

The jackknife correct classification rate, however, is nearly perfect (and thus no jackknife cross-validation table is given), with the exception being a replicate of group **17** is misclassified as group **15**. A 100% jackknife correct classification rate for the samples in 4D OCV space was found.

Conclusion

The intention of this study was to differentiate casework liquid gasoline samples by utilizing multivariate procedures from data generated by the GC-MS. A supervised learning approach was undertaken to achieve this goal. In other words, the procedures were tested knowing a priori the correct group assignments for the gasolines.

This study revealed that the variability in the sample population was sufficient enough to distinguish all the samples from one another knowing their groups a priori using PCA, CVA, and OCVA. It was observed that CVA was able to differentiate all samples in the population using three dimensions while OCVA required four dimensions. These results were cross-validated using the "jackknife" method to confirm the classification functions. By plotting the CVA and OCVA data in two and three dimensions, clearly defined and easily interpretable clusters were evident in the sample population.

PCA required at least 10 dimensions of data in order to predict the correct groupings. It was observed that by plotting the PCA data in two and three dimensions that the samples did not cluster into well-defined groups when compared with the results obtained from CVA and OCVA.

Preliminary studies conducted on weathered gasoline samples showed that group predictions using CVA and OCVA were applicable to about 75% to 80% weathered by volume. A more formal and detailed study will be conducted with these weathered samples. The outcome of this initial study served to develop the multivariate procedures and methods to be adaptable to planned future studies. It is hoped that future studies in this area will be developed into practical procedures that possess the scientific rigor required of a technique applicable to fire debris casework.

Finally, the authors feel that statistical methods of pattern recognition must be applied to as many fields of empirical science as possible. This is especially true in the field of forensic science, where more and more of the findings of traditional methods used in trace evidence, firearms, and toolmark analysis are being thrown out of court because many statistical studies have not been carried out. When using statistical methods to make numerical discriminations between data, it is critical to lay out all of the details of these methods if they are not in common use in the particular scientific community the study is directed at. Thus we presented a detailed discussion of the statistical methods used for this study. The discussion should not be a barrier to the common application of these methods in the field for a number of reasons. They have been in the scientific literature for ten to almost 100 years (depending on the method), industry has been widely using them for the last 40 years (i.e., since the availability of computers), and three of the methods used in this study (PCA, CVA, and LDA) are implemented in widely available easy to use statistical analysis software such as SPSS, Minitab, and SAS (25–27). Lastly, the authors would be happy to share the data set and Mathematica software written by them for this study, upon request.

Acknowledgments

The authors are grateful to Professor W. J. Krzanowski for guidance in implementing OCVA and kindly providing a reprint of his paper. We also thank Ms. Lauren Gunderson, Ms. Dayhana Olivo and Ms. Helen Chan for kindly reading and commenting on our manuscript.

References

1. Speight JG. Handbook of petroleum product analysis. 1st edn. New York: Wiley, 2002.
2. Lucas AG. Modern petroleum technology: downstream. 6th edn. New York: Wiley, 2000.
3. Gibbs LM. Understanding gasoline additives. Automotive Eng 1990;98(1):43–8.
4. Mann DC. Comparison of automotive gasolines using capillary gas chromatography I: comparison methodology. J Forensic Sci 1987;32(3): 606–15.
5. Mann DC. Comparison of automotive gasolines using capillary gas chromatography II: limitations of automotive gasoline comparisons in casework. J Forensic Sci 1987;32(3):616–28.
6. Sandercock PML, Du Pasquier E. Chemical fingerprinting of gasoline 3. Comparison of unevaporated automotive gasoline samples from Australia and New Zealand. Forensic Sci Int 2004;140(1):71–7.
7. Tan B, Hardy JK, Snavely RE. Accelerant classification by gas chromatography/mass spectrometry and multivariate pattern recognition. Anal Chim Acta 2000;422(10):37–46.
8. Doble P, Sandercock M, Du Pasquier E, Petocz P, Rouxa C, Dawson M. Classification of premium and regular gasoline by gas chromatography/mass spectrometry, principal component analysis and artificial neural networks. Forensic Sci Int 2003;132(1):26–39.
9. Sandercock PML, Du Pasquier E. Chemical fingerprinting of unevaporated automotive gasoline samples. Forensic Sci Int 2003;134(1): 1–10.
10. Sandercock PML, Du Pasquier E. Chemical fingerprinting of gasoline 2. Comparison of unevaporated and evaporated automotive gasoline samples. Forensic Sci Int 2004;140(1):43–59.
11. Sigman ME, Williams MR. Covariance mapping in the analysis of ignitable liquids by gas chromatography/mass spectrometry. Anal Chem 2006;78(5):1713–8.
12. Sigman ME, Williams MR, Ivy RG. Individualization of gasoline samples by covariance mapping and gas chromatography/mass spectrometry. Anal Chem 2007;79(9):3462–8.
13. Rencher AC. Methods of multivariate analysis, 2nd edn. Hoboken: Wiley, 2002.
14. Duda RO, Hart PE, Stork DG. Pattern classification, 2nd edn. New York: Wiley, 2001.
15. Theodoridis S, Koutroumbas K. Pattern recognition, 3rd edn. San Diego: Academic Press, 2006.
16. ASTM-E 1618-06. Standard test method for ignitable liquid residues in extracts from fire debris samples by gas chromatography-mass spectrometry. West Conshohocken, PA: American Society of Testing and Materials, 2006;14.02.
17. Keto RO, Wineman P. Detection of petroleum-based accelerants in fire debris by target compound gas chromatography/mass spectrometry. Anal Chem 1991;63:1964–71.
18. Jolliffe IT. Principal component analysis, 2nd edn. New York: Springer, 2004.
19. Wolfram Research, Inc. Mathematica [computer program]. 5.1. Champaign (IL): Wolfram Research, Inc. 2005.
20. Mardia KV, Kent JT, Bibby JM. Multivariate analysis, 1st edn. Amsterdam: Academic Press, 1980.
21. Krzanowski WJ. Orthogonal canonical variates for discrimination and classification. J Chemometrics 1994;9(6):509–20.
22. Shawe-Taylor J, Cristianini N. Kernel Methods for Pattern Analysis. London: Cambridge University Press, 2004.
23. Lanyon SM. Jackknifing and bootstrapping: important ''new'' statistical techniques for ornithologists. Auk 1987;104:144–6.
24. Schaeffer RL, McClave JT. Probability and statistics for engineers, 2nd edn. Boston: Duxbury, 1986.
25. SPSS, Inc. SPSS [computer program]. Chicago (IL): SPSS, Inc., 2007.
26. Minitab, Inc. Minitab [computer program]. State College (PA): Minitab, Inc., 2007.
27. SAS, Inc. SAS [computer program]. Raleigh, (NC): SAS, Inc., 2007.

Additional information and reprint requests:
Nicholas D. K. Petraco, Ph.D.
Department of Science
John Jay College of Criminal Justice
899 10th Ave.
New York, NY 10019
E-mail: npetraco@jjay.cuny.edu