Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Chromatography A, 1096 (2005) 101–110
JOURNAL OF CHROMATOGRAPHY A
www.elsevier.com/locate/chroma

Classification of gasoline data obtained by gas chromatography using a piecewise alignment algorithm combined with feature selection and principal component analysis

Karisa M. Pierce [a], Janiece L. Hope [a], Kevin J. Johnson [a,1], Bob W. Wright [b], Robert E. Synovec [a,*]

[a] *Department of Chemistry, Box 351700, University of Washington, Seattle, WA 98195, USA*
[b] *Pacific Northwest National Laboratory, Battelle Boulevard, P.O. Box 999, Richland, WA 99352, USA*

Available online 17 May 2005

Abstract

A fast and objective chemometric classification method is developed and applied to the analysis of gas chromatography (GC) data from five commercial gasoline samples. The gasoline samples serve as model mixtures, whereas the focus is on the development and demonstration of the classification method. The method is based on objective retention time alignment (referred to as piecewise alignment) coupled with analysis of variance (ANOVA) feature selection prior to classification by principal component analysis (PCA) using optimal parameters. The degree-of-class-separation is used as a metric to objectively optimize the alignment and feature selection parameters using a suitable training set thereby reducing user subjectivity, as well as to indicate the success of the PCA clustering and classification. The degree-of-class-separation is calculated using Euclidean distances between the PCA scores of a subset of the replicate runs from two of the five fuel types, i.e., the training set. The unaligned training set that was directly submitted to PCA had a low degree-of-class-separation (0.4), and the PCA scores plot for the raw training set combined with the raw test set failed to correctly cluster the five sample types. After submitting the training set to piecewise alignment, the degree-of-class-separation increased (1.2), but when the same alignment parameters were applied to the training set combined with the test set, the scores plot clustering still did not yield five distinct groups. Applying feature selection to the unaligned training set increased the degree-of-class-separation (4.8), but chemical variations were still obscured by retention time variation and when the same feature selection conditions were used for the training set combined with the test set, only one of the five fuels was clustered correctly. However, piecewise alignment coupled with feature selection yielded a reasonably optimal degree-of-class-separation for the training set (9.2), and when the same alignment and ANOVA parameters were applied to the training set combined with the test set, the PCA scores plot correctly classified the gasoline fingerprints into five distinct clusters.

Keywords: Alignment; Gas chromatography; Feature selection; Principal component analysis; ANOVA; Fuel; Chemometrics

1. Introduction

There is a need for efficient data processing methods for the large volumes of data produced by modern analytical instruments. For many applications, the analyst must compress such large volumes of data while retaining the essential information in order to classify samples, do long-term comparisons, or perform batch-to-batch reproducibility studies. Analysts dealing with large data sets of micro-arrays [1], spectroscopic profiles [2,3] and 2D images [4,5] use feature selection methods to reduce data sets down to features containing the essential information. Chemometric pattern recognition methods are then used to classify the reduced data set. However, application of feature selection and pattern recognition methodologies is limited in chromatography by uncontrollable retention time variations that obscure

* Corresponding author. Tel.: +1 206 685 2328; fax: +1 206 685 8665.
E-mail address: synovec@chem.washington.edu (R.E. Synovec).
[1] Present address: Chemical Sensing/Chemometrics Section, Code 6112, Chemical Dynamics and Diagnostics Branch, Naval Research Laboratory, Washington, DC 20375-5342, USA.

0021-9673/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2005.04.078

chemical variations in the data [6–14]. Retention time variations can be due to subtle, random, and often unavoidable variations in instrument parameters. Pressure, temperature and flow rate fluctuations may cause an analyte to elute at a different retention time in replicate runs [12]. Matrix effects and stationary phase decomposition may also cause retention time shifting. Ideally, one would want a comprehensive data analysis procedure that combines retention time alignment with feature selection and chemometric pattern recognition in order to classify large data sets of complex chromatograms with objectively optimized parameters.

Many retention time alignment algorithms have been reported. Some alignment algorithms operate by aligning specific features in the data [13,15]. However, there are many available alignment algorithms that, like piecewise alignment reported herein, do not require knowledge or identification of peaks. These algorithms contain some level of dynamic programming where iterated shifts are evaluated by a matching metric between the sample and target chromatogram. The matching metric indicates an optimal retention time correction for the sample. These algorithms fall under the categories of dynamic time warping [6,16,17], genetic algorithms [8], partial linear fit and minimization of residuals [9,10], correlation optimized warping (COW) [11], and local retention time alignment, also referred to as peakmatch alignment [12]. The piecewise alignment algorithm applied in this report is a form of local retention time alignment that is also related to the COW algorithm. COW operates by subdividing the data into local regions, or windows, which are iteratively stretched and compressed by interpolation so as to maximize the correlation between the sample and target chromatograms. COW seeks to find the global arrangement of stretches and shrinks that maximize correlation between the target and sample chromatograms [11,16]. Piecewise alignment is related to COW in that it operates by subdividing the data into windows, but then each window is iteratively shifted along the target chromatogram within a specified limit to find the maximum correlation and the best correction for each window. Thus, in comparison to COW, piecewise alignment does not apply the stretching and shrinking interpolation step just prior to calculating the correlation, thus saving computation time.

Herein, we introduce a data analysis procedure for reducing and classifying chromatographic data involving retention time alignment, feature selection and chemometric pattern recognition with objectively optimized parameters. The piecewise retention time alignment algorithm is demonstrated to quickly provide retention time corrections for a large GC data set of gasoline samples. It is shown that alignment combined with analysis of variance (ANOVA) feature selection and submission to principal component analysis (PCA) yields proper classification of unknown gasoline chromatograms. In order to make the method more robust, user-friendly and rapid to implement, it is demonstrated that the objective selection of alignment parameters and feature selection conditions can be achieved using a suitable training set containing a substantially smaller number of GC chromatograms and fewer sample classes than the entire data set (test set) of chromatograms to be evaluated.

This report is organized to describe the data reduction and classification method following the steps outlined in Fig. 1. Initially, the piecewise alignment parameters and ANOVA-based feature selection conditions are optimized using a suitable training set and the degree-of-class-separation metric. ANOVA-based feature selection is used to objectively select portions of the training set as a function of retention time. Then, the unknown test set is aligned using the optimal parameters. Next, the features that were selected in the training set are extracted from the test set. PCA is then applied for optimal classification of the test set. Classification is based on clustering in the scores plot. Finally, descriptions of classification results using piecewise alignment alone or feature selection alone are included to demonstrate the benefits of combining retention time alignment with ANOVA feature selection in order to allow PCA to focus on class variations. The reported classification method results are compared to linear discriminant analysis (LDA) at each stage of processing the data.

2. Theory

The following subsections describe the algorithms applied for the classification method reported herein.

2.1. Principal component analysis (PCA)

PCA is a data mining tool that is useful for providing unsupervised visual classification of multivariate data like GC data [18,19]. PCA converts each chromatographic vector into a single point in principal component space, essentially projecting the data onto a new set of orthogonal axes (principal components, i.e., PCs) that are sorted in order according to the amount of variance captured. If the captured variance is relevant to chemical variations and sample classification, similar sample scores should cluster together on a scores plot of PC 1 versus PC 2.

2.2. Degree-of-class-separation

For a classification method to be a truly objective process, the alignment parameters that are input by the user must be objectively optimized. Optimization can be achieved by determining the alignment parameters that yield the greatest degree-of-class-separation between two clusters of scores for two sample types in the training set on a PCA scores plot (PC 1 versus PC 2). The degree-of-class-separation metric provides a numeric measure of the quality of clustering within a PCA scores plot, as well as the classification by PCA. This can be used to evaluate the improvement in the PCA classification after data preprocessing or as a metric to optimize the input parameter values for a particular data set. For this work, the degree-of-class-separation on a scores plot was defined as

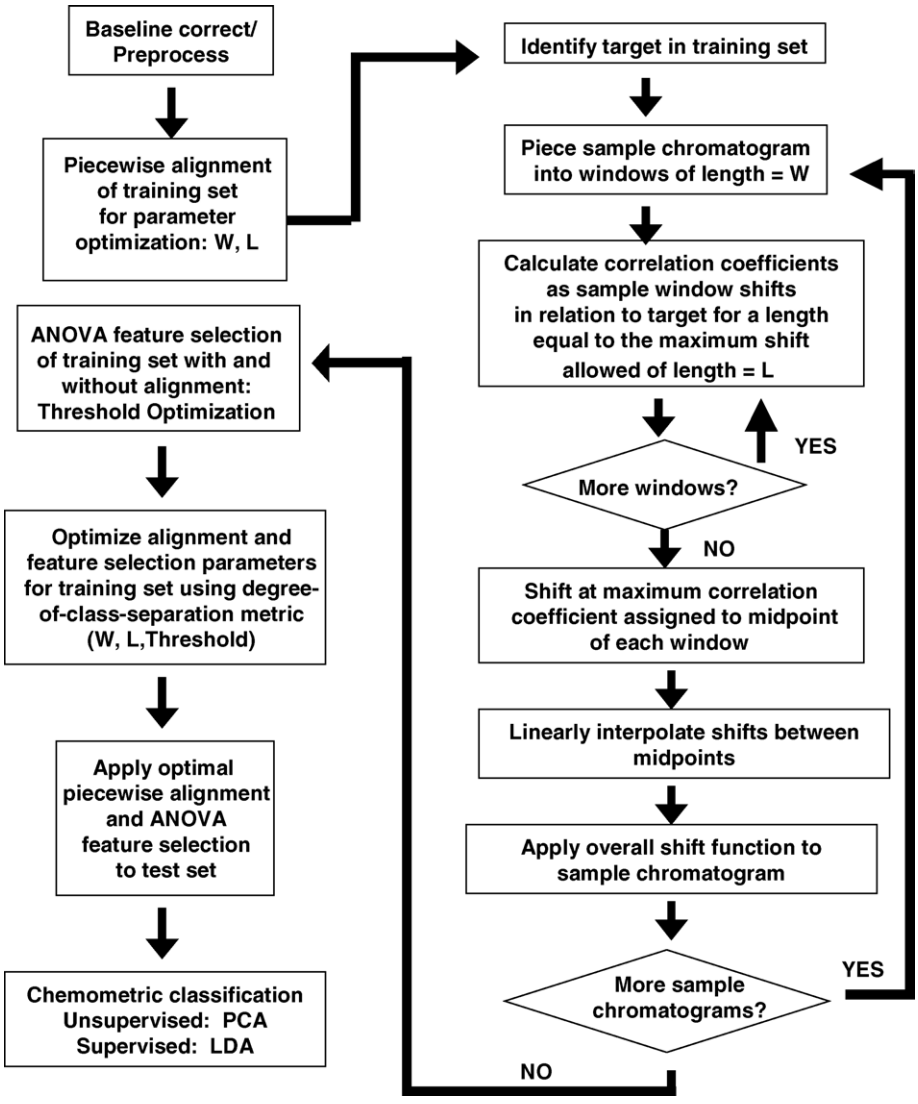

Fig. 1. Flowchart diagram of the piecewise alignment, ANOVA feature selection, and PCA classification process using a training set for optimization of the parameters, then application of those optimal parameters to a test set for classification.

the class-to-class variance divided by the sum of the within-class variance on a scores plot using two sample types in the training set. In other words, the degree-of-class-separation is the Euclidean distance between the centroids of two groups of sample replicates ($D_{A,B}$, where A and B are two different sample types) divided by the square root of the summed variances in the Euclidean distance of each sample replicate relative to the centroid of its group as in Eq. (1):

$$\text{Degree-of-class-separation} = \frac{D_{A,B}}{\sqrt{s_A^2 + s_B^2}} \quad (1)$$

The variance is defined as the square of the standard deviation (s) of the distance of each score in a group from the centroid of that group.

The centroids of groups A and B are located at $(\bar{X}_A, \bar{Y}_A)$ and $(\bar{X}_B, \bar{Y}_B)$ on a plot of PC 1 versus PC 2 where $\bar{X}_A$ is the mean of the scores for the samples in A on PC 1 and $\bar{Y}_A$ is the mean of the scores for the samples in A on PC 2, determined using Eqs. (2) and (3):

$$\bar{X}_A = \frac{\sum_{j=1}^{n} x_{A,j,PC\,1}}{n} \quad (2)$$

$$\bar{Y}_A = \frac{\sum_{j=1}^{n} y_{A,j,PC\,2}}{n} \quad (3)$$

The Euclidean distance between these centroids, $D_{A,B}$, is calculated as in Eq. (4).

$$D_{A,B} = \sqrt{(|\bar{X}_A - \bar{X}_B|^2 + |\bar{Y}_A - \bar{Y}_B|^2)} \quad (4)$$

2.3. Analysis of variance (ANOVA) feature selection

ANOVA-based feature selection was explained in detail in a previous publication [20] and a brief introduction is included here. Feature selection discards chromatographic

signals that are not useful for classification, while primarily retaining signals that have chemical information correlating with sample groups [18,20–22]. Features that have a large Fisher ratio are retained, where the Fisher ratio is the class-to-class variance in the chromatographic signal divided by the summed within-class variances in the chromatographic signal for a training set of known sample types [18,20]. Note that the Fisher ratio calculation uses the actual chromatographic signal to indicate how much classification information is in each chromatographic peak while the degree-of-class-separation uses the scores on a PCA plot to indicate how successful classification is based on clustering. The ANOVA program first calculates a Fisher ratio for each point along the retention time axis for the training set. Then, the features that have Fisher ratios above a defined threshold can be extracted from the test set data. A pattern recognition method is applied (e.g., PCA) to this reduced data set for classification. ANOVA feature selection is a partially supervised classification method where the training set can be composed of fewer classes than are in the test set. Two sample types are used for the training set in this report.

2.4. Linear discriminant analysis

Linear discriminant analysis (LDA) is a traditional statistical approach for supervised classification and pattern recognition [18,19,22,23]. LDA requires that representative samples from all of the classes in the test set be present and identified in the training set. The linear discrimination function then fits a multivariate normal density to each group in the training set, with a pooled estimate of covariance to determine class membership for individual chromatograms alternately treated as unknowns [19]. The algorithm classifies the samples by type and yields a percentile misclassification rate for each sample present in the test set. The improvements in LDA misclassification rates at each step of the classification method are used to show that both alignment and feature selection are beneficial for successful classification by LDA. Since LDA is a common pattern recognition method, it is used to validate the improvements in PCA clustering at each step.

2.5. Piecewise alignment algorithm

The piecewise alignment algorithm performs retention time alignment for a target chromatogram and sample chromatograms from various classes. Piecewise alignment is schematically depicted in Fig. 1. Piecewise alignment begins by choosing or generating the target chromatogram. In this case, the target was a chromatogram randomly chosen from the training set. In the next step, the sample and target chromatograms are divided into windows of a user-specified length (window length = W). Every window in the sample chromatogram contains multiple chromatographic peaks, and it is assumed that the shifting in these windows is a scalar offset rather than a more complicated stretch or shrink function. Each window in the sample chromatogram is iteratively shifted point-by-point by the algorithm, within a specified limit of the maximum shift allowed (limit = L), along the retention time axis, where a point is defined as the actual data points collected during data acquisition. The Pearson correlation coefficient between the sample and target is calculated at each shift [20,21]. Solely for the purpose of determining the correlation coefficient at each shift, the algorithm applies a temporary Wallis filter to both the sample and target chromatograms in order to minimize the effect of varying peak heights [11,12]. As the Wallis-filtered alignment window of length W is shifted, point-by-point, along the retention time axis, a list of correlation coefficients is generated. The shift that gives the maximum correlation coefficient is used to correct that window of the sample chromatogram. The desired retention time corrections are assigned to the center point of the windows. The shifts to be applied in regions between window centers are calculated by linear interpolation.

3. Experimental

Unleaded gasoline samples with an octane rating of 89 were arbitrarily obtained from the pump at five local gasoline stations (Seattle, WA, USA): Type A (A), Type C (C), Type M (M), Type S (S), and Type T (T). These five fuels were analyzed with an Agilent 6890 gas chromatograph equipped with an electronic pressure controller and a flame ionization detector (FID). The separation column was fused-silica capillary, 10 m long, with a 100 μm diameter, and a 0.4 μm DB-5 stationary phase. The inlet temperature was 275 °C and a 300:1 split ratio was used with a temperature program set initially at 30 °C for 2 min, then ramped at 25 °C/min to 200 °C. Each sample was run in replicate for five consecutive days to yield a data set of 210 chromatograms (40 A, 40 C, 45 M, 45 S, and 40 T). FID readings were acquired at a rate of 20 Hz. The chromatograms were imported from Chemstation (Agilent Technologies, Palo Alto, CA, USA) into Matlab 6.1 (The Mathworks, Natick, MA, USA) where the alignment and chemometric analyses were performed on a Pentium 4 Intel 2.8 GHz processor with 1 GB of RAM and Microsoft 2000 Operating System. Each chromatogram was loaded into a Matlab workspace as a vector composed of the FID signal gathered over the duration of the GC run. The chromatograms for the training set and the test set were appended into a matrix where each row was a chromatogram.

The chromatograms were individually baseline corrected by subtracting the best-fit line through the first and last 2 s of the chromatogram (regions of baseline noise only) from the entire length of the chromatogram. The chromatograms were individually normalized to account for injection volume deviations by dividing each data point in the chromatogram by the sum of the absolute value of all the data points in the chromatogram. This baseline corrected and normalized data is referred to as unaligned data. The training set was composed of 25 Type M and 25 Type S replicates that were run over the

course of 5 days (5 of each per day) while gathering the entire set of A, C, M, S, and T replicate chromatograms. It is important to have training set chromatograms from each of the days that test set data are collected in order to obtain optimal retention time alignment parameters. The test set was composed of the remaining 40 A, 40 C, 20 M, 20 S, and 40 T replicates. Prior to PCA, the data were mean-centered. The LDA algorithm was run 1000 times and averaged to yield a misclassification rate for each sample at each of the data processing stages. The ANOVA feature selection program was written in house [20]. The PCA and LDA algorithms were from Eigenvector's PLS Toolbox (Eigenvector Research, Inc., Manson, WA, USA). The Matlab implementation for COW was downloaded from www.models.kvl.dk/source.

4. Results and discussion

4.1. Data set characteristics

A typical chromatogram of one of the fuels is shown in Fig. 2A. Fuels are mixtures of many chemical components that yield complex chromatograms. The fuels were very similar in both their number and type of chemical components, however, many compounds varied in amount between samples and a smaller number of compounds were absent in some samples. Retention time alignment, as discussed in Section 2, retains this chemical selectivity. Chemometric data reduction (ANOVA feature selection) and pattern recognition methods (PCA) are a natural choice for analysis of such complex, inter-related data.

4.2. PCA and LDA applied to unaligned raw data

When PCA was applied to the unaligned data set (training set combined with test set), in the scores plot in Fig. 2B, PC 1 captured 35% of the variance and PC 2 captured 27% of the variance. PCA alone was not able to provide accurate clustering of the chromatograms by fuel type. The Type A replicates (A) are separated from the other types of fuels, but the remaining four fuels are clustered together. Thus, unsupervised PCA classification using unaligned raw data fails. On the other hand, fully supervised LDA applied to the first 10 PCs of the unaligned data resulted in perfect classification for A, S, and T, but M had a 10% misclassification rate and C had a 7.5% misclassification rate. However, LDA requires standards of all sample types present in the test set, which may not be feasible in most applications of interest when not all sample sources are known beforehand. PCA is an unsupervised pattern recognition tool, but PCA alone failed to capture the class variations. This prompts the development of a partially supervised method of classification that can use knowledge of two samples to improve PCA clustering for all five of the unknown samples, i.e., implementing retention time alignment and ANOVA feature selection.

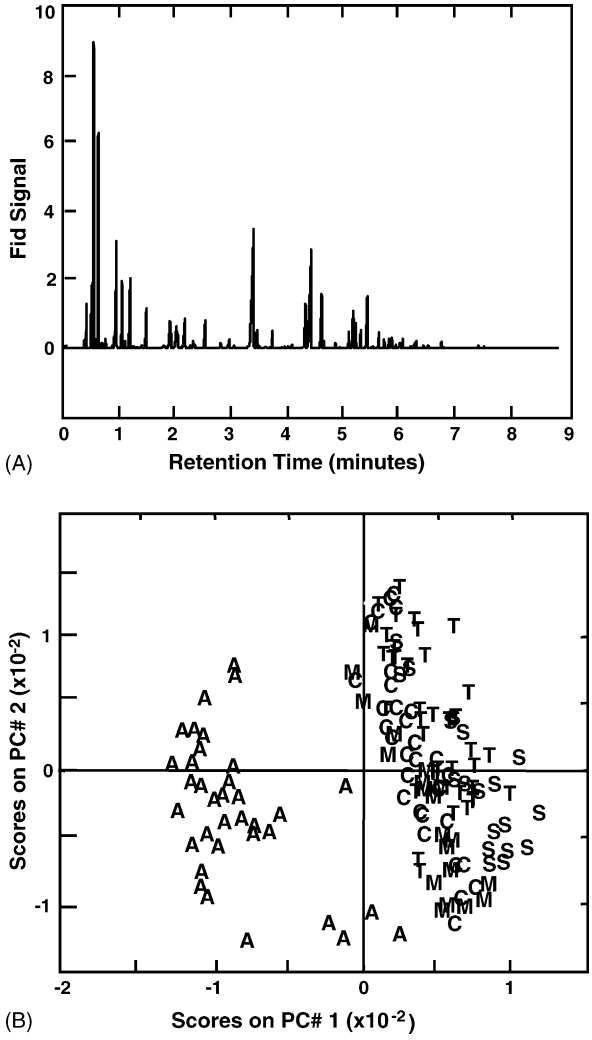

Fig. 2. (A) A typical gas chromatogram of a gasoline sample. (B) Scores plot from PCA of unaligned chromatograms (all chromatograms). Retention time variation and other sources of variation that were not related to class differences led to poor clustering among replicate chromatograms. Gasoline samples: A = Type A, C = Type C, M = Type M, S = Type S, T = Type T.

4.3. Piecewise alignment demonstration

Visual inspection of the training set data (25 M and 25 S collected 5 each per day over 5 days) revealed that retention time shifting was present in the raw chromatograms. In an effort to rid the data set of retention time variation, the raw data set was subjected to piecewise alignment. An overlay of a section of the chromatograms is shown in Fig. 3A (before alignment) and in Fig. 3B (after piecewise alignment with $W = 10$ s, $L = 1.5$ s). Run-to-run retention time shifting is apparent in Fig. 3A, but after piecewise alignment was applied to the training set the retention time shifting was corrected as seen in Fig. 3B. The standard deviations (s) of the locations of the peaks shown in Fig. 3A and B were evaluated to quantify the improvement gained from piecewise alignment. These standard deviations were significantly reduced for the five peaks indicated. Overall, the retention time pre-

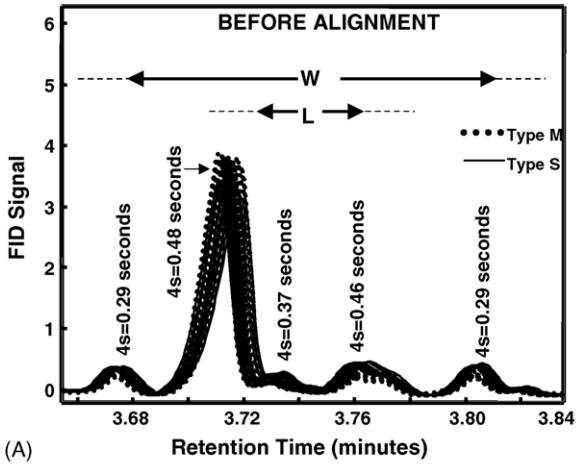

Fig. 3. Selected region of training set data (A) before piecewise alignment and (B) after piecewise alignment ($W = 10$ s and $L = 1.5$ s). An improvement in retention time precision provided by piecewise alignment is noted by comparing the run-to-run peak retention time precision before and after alignment. The improvement is quantified by comparing four times the standard deviation (s) of peak locations.

cision from run-to-run was improved along the entire length of the chromatographic axis.

4.4. ANOVA feature selection demonstration after piecewise alignment

Coupling alignment with ANOVA feature selection should enhance the multivariate classification by reducing the data set and allowing the pattern recognition tool to focus on chemical variations rather than other sources of variation [1,13]. When ANOVA feature selection is applied to a data set, replicates from as few as two samples need to be identified. The training set of M and S replicates used in the previous section was chosen for the ANOVA training set, in order to be consistent for the subsequent alignment parameter optimization. Samples from other classes could have been chosen with similar results, but are not reported herein for brevity. Fig. 4A contains the training set Fisher ratios calculated by ANOVA

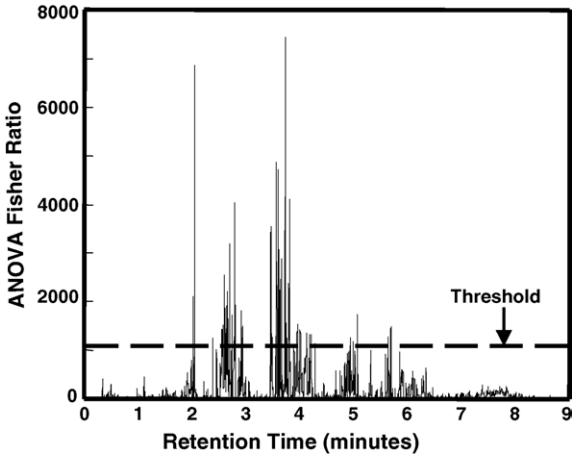

Fig. 4. Analysis of variance (ANOVA) feature selection. Fisher ratios calculated by ANOVA after piecewise alignment ($W = 10$ s and $L = 1.5$ s) for the training set (25 M and 25 S). The dashed line is an example of choosing a threshold of 1000 that retains 3% of the data.

at each data point along the retention time axis after piecewise alignment was applied ($W = 10$ s, $L = 1.5$ s). As the threshold for the Fisher ratio increases the number of features retained by feature selection decreases. At a Fisher ratio threshold of zero, 100% of the data points in the chromatograms are retained. Illustrated in Fig. 4A is the example of a Fisher ratio threshold of 1000, marked by a dashed line, where 3% of the data is retained. ANOVA feature selection is shown to be useful as a partially supervised data reduction tool whereby indexing certain features based on the two-class training set will reduce the five-class test set to features containing classification information. Ideally, the selected features will be useful to distinguish the other sample classes that are in the test set. We shall see in subsequent sections herein that this is indeed the case.

4.5. Parameter optimization

The optimization of the piecewise alignment and feature selection parameters by analysis of the degree-of-class-separation information is shown in Fig. 5.

Objective selection of alignment parameters and feature selection conditions was achieved using the training set. The training set contained a smaller number of GC chromatograms (50), i.e., 25 M and 25 S collected 5 each per day over 5 days, than the test set of chromatograms (160). As illustrated in Sections 4.3 and 4.4, optimization of W, L, and the ANOVA threshold was performed for classification by PCA using the selected training set. Then, the optimal W and L are applied to the test set for objective alignment, and the retention time indices of the features retained in the training set are used to extract features from the test set. Finally, PCA is applied for optimal classification of the test set (or, if desired, the test set combined with the training set). The quantitative metric used for optimization of the alignment and feature selection parameters using the training set

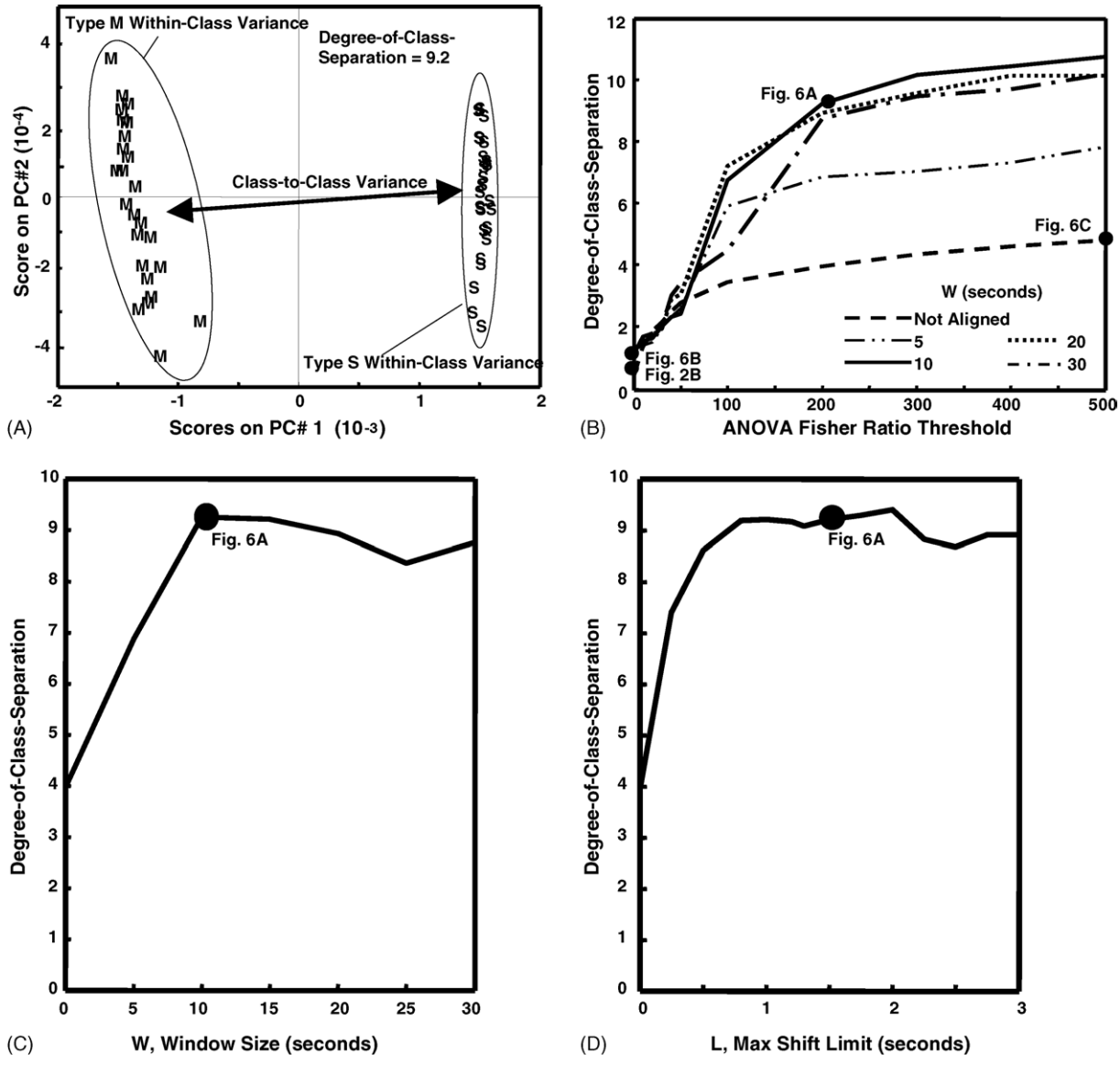

Fig. 5. Parameter optimization based on degree-of-class-separation. (A) PCA scores plot illustrating the degree-of-class-separation for the M and S training set, for $W = 10$ s, $L = 1.5$ s, and threshold = 200. Degree-of-class-separation is defined as the class-to-class variance divided by the sum of the within-class variance for the PCA scores of replicates of two classes in a training set. (B) Degree-of-class-separation between training set scores as a function of window size and threshold. A reasonably optimal degree-of-class-separation was 9.2, achieved using the alignment parameters of $W = 10$ s, $L = 1.5$ s, and ANOVA threshold = 200. (C) Degree-of-class-separation as a function of W for a constant threshold of 200 for piecewise alignment ($L = 1.5$ s). (D) Degree-of-class-separation as a function of L with a constant threshold = 200 and constant $W = 10$ s. Degree-of-class-separation is independent of values of L greater than 1.0 s (the maximum shift present in the raw data).

is the degree-of-class-separation. As discussed in Section 2, an increase in the degree-of-class-separation quantifies an improvement in PCA scores clustering as a function of W, L, and ANOVA threshold because it measures the distance between two clusters of scores as well as the tightness of each cluster. An illustration of the degree-of-class-separation for the training set is shown in the PCA scores plot in Fig. 5A, with the training set first submitted to piecewise alignment ($W = 10$ s, $L = 1.5$ s), then feature selection (threshold = 200) prior to PCA. In this case, applying Eqs. (1)–(4) the degree-of-class-separation between the Type M and Type S training set members was 9.2. The alignment parameters and feature selection indices that yield the maximum degree-of-class-separation are then applied to the test set, and presumed to provide an acceptable PCA scores plot for classification.

The training set was subjected to piecewise alignment and ANOVA feature selection for a range of W values (with $L = 1.5$ s) as well as for a range of ANOVA thresholds. The degree-of-class-separation between the training set members was determined for each of the resulting scores plots with the results shown in Fig. 5B. The other pair combinations of fuels were used as the training set and trends similar to those in Fig. 5B were found for degree-of-class-separation as a function of W and threshold, though not shown here for brevity.

According to Fig. 5B, piecewise alignment coupled with feature selection, where $W = 10$ s and threshold = 200, yielded a reasonably optimal degree-of-class-separation (9.2) for the training set (see Fig. 5A for PCA scores plot), while still using a reasonably low threshold. The reasonably low threshold was chosen in order to more fully utilize the data for the all the other sample types in the test set. Note that at a threshold of 200, 16% of the piecewise aligned data are retained for submission to PCA. For the fuels studied in this work, the unaligned data, with no feature selection (Fig. 2B) yielded the lowest degree-of-class-separation (0.4) between the M and S training set classes on a scores plot. The degree-of-class-separation for data submitted to piecewise alignment alone ranged between 0.4 and 1.2 for $0\,s \leq W \leq 30\,s$. Conversely, ANOVA feature selection applied directly to the unaligned training set with a threshold of 500 does increase the degree-of-class-separation (4.8). In Fig. 5B, the leveling off in degree-of-class-separation that occurs for each window size as the threshold increases beyond ∼200 is expected due to the diminishing number of peaks that are retained after applying ANOVA to the training set. The steep rise in the degree-of-class-separation occurring at the beginning of each window size trace is due to the removal of noise and features with low Fisher ratios. The analysis of degree-of-class-separation indicates that piecewise alignment needs to be coupled with ANOVA feature selection in order to allow PCA to focus on class variations for the best achievable clustering on a PCA scores plot.

For each constant threshold in Fig. 5B the degree-of-class-separation increases with increasing window size up to a value of $W = 10$ s. However, for values of W greater than 10 s the degree-of-class-separation, as well as the clustering in the scores plots, slightly declines. Another view of degree-of-class-separation changing as a function of W for threshold = 200 is shown in Fig. 5C. The slight decrease in degree-of-class-separation that occurs for $W > 10$ s results from the fact that the retention time shifting in a window length greater than 10 s can no longer be modeled by simple scalar shifting. Window sizes ranging from 0 s (not aligned) up to 30 s were explored and were found to comply with the described trends. Although the degree-of-class-separation slightly decreases beyond the optimal value of W ($W = 10$ s), it is consistently higher than the degree-of-class-separation for data that underwent no alignment or feature selection alone when the window sizes are large ($5\,s \leq W \leq 30\,s$).

As discussed in Section 2, the piecewise alignment algorithm has two input parameters: W and L. Intuitively, the value for L should be equal to or greater than the greatest amount of retention time shifting seen in the unaligned data. Therefore, it can be concluded that as long as L is larger than the greatest amount of shifting in the data, the value of L should have no effect on the performance of piecewise alignment. The observation that alignment performance is only affected when L is too small (not when L is too big) suggests that the subjectivity of an analyst choosing the optimal L parameter is lessened. The analyst simply looks at the raw data, determines how severe the shifting is, and selects an L that is larger than that shifting, without worrying about crossing an upper limit of L that might begin to affect the alignment. In these GC data the retention time shifting was less than 1.5 s so all of the work presented was performed with $L = 1.5$ s. To validate this choice of L, the degree-of-class-separation as a function of L was analyzed via the plot in Fig. 5D for $W = 10$ s and threshold = 200. From this plot one can see that limiting the value of L to 1.0 s or longer yields a consistently large degree-of-class-separation, so L equal to 1.5 s is a suitably optimal value for L.

4.6. PCA and LDA after optimal piecewise alignment and ANOVA feature selection

Fig. 6A contains the scores plot for the test set combined with the training set after submission to piecewise alignment and ANOVA feature selection using the optimal parameters ($W = 10$ s, $L = 1.5$ s, threshold = 200). PC 1 captured 66% of the variance in the data and PC 2 captured 21% of the variance in the data. Interestingly, the degree-of-class-separation for M and S training set members in the scores plot of the optimally aligned and feature selected data set (combination of test set and training set) was 22.6 (Fig. 6A) while being only 9.2 for the training set submitted alone to optimal alignment and feature selection (Fig. 5B). This is due to the fact that as other classes of unknown chromatograms are added to the data set, PC 1 and PC 2 capture the more prominent chemical variations between the classes. At the same time, the within-class variations for the Type S and Type M training set members are buried in higher order PCs, thus diminishing the within-class variation for Type S and Type M scores and consequently increasing the degree-of-class-separation for the training set members. A degree-of-class separation of 22.6 for M and S training set members in Fig. 6A is a significant improvement over the degree-of-class-separation for M and S training set members in the combined training and test set in Fig. 2B that was not submitted for alignment (0.9). The fuels are correctly and clearly grouped into five clusters corresponding to the five classes after optimal alignment and feature selection in Fig. 6A. Piecewise alignment combined with feature selection also improved the supervised LDA pattern recognition results compared to LDA for unaligned data. Piecewise alignment with feature selection yielded 100% accurate LDA classification rates for all five classes. It is important to point out that LDA is a completely supervised pattern recognition tool that requires a training set containing replicates of all five classes while the method outlined herein is a partially supervised classification method trained on replicates of only two classes.

4.7. PCA and LDA after piecewise alignment only

PCA was applied to the combined test and training data set after it was submitted to optimized piecewise alignment alone and the resulting scores plot is shown in Fig. 6B. PC 1

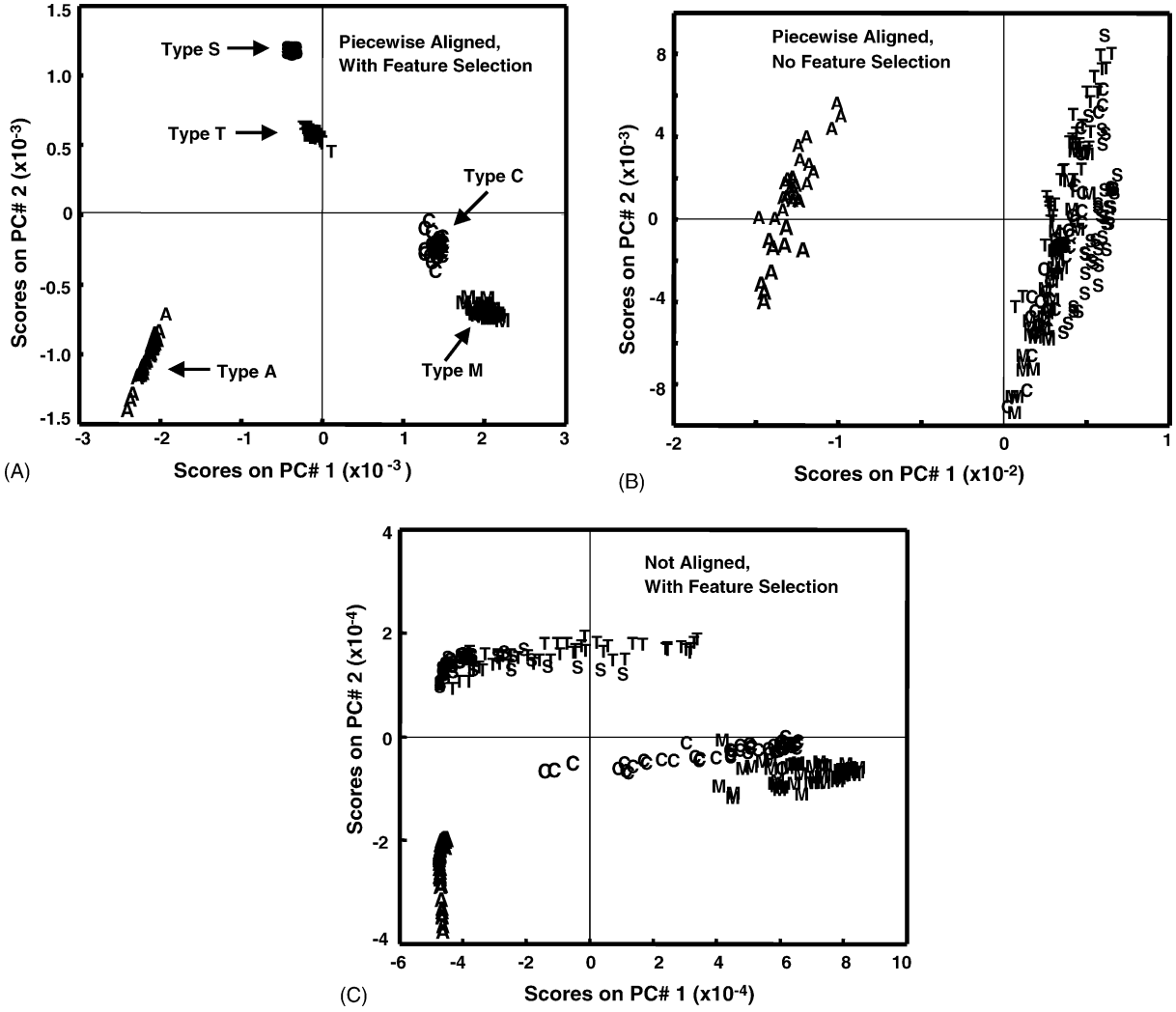

Fig. 6. Benefit of piecewise alignment coupled with ANOVA feature selection for PCA. (A) Scores plot after optimal piecewise alignment and feature selection (W = 10 s, L = 1.5 s, threshold = 200). Every sample is correctly clustered into specific fuel type and each cluster is clearly separate from every other cluster. (B) Scores plot of test set after only piecewise alignment (W = 10 s, L = 1.5 s). The A and S scores cluster more tightly compared to the unaligned data in Fig. 2B indicating that piecewise alignment corrected retention time variation. (C) Scores plot of unaligned test set after only ANOVA feature selection was applied (threshold = 500). The S and T scores cluster together and the C and M scores cluster together. Alignment coupled with feature selection is required for successful classification by PCA. Gasoline samples: A = Type A, C = Type C, M = Type M, S = Type S, T = Type T.

captured 62% of the variance in the aligned data set while PC 2 captured 18% of the variance in the aligned data set. The degree-of-class separation for M and S for the training set alone under these conditions was 1.2 (Fig. 5B). The degree-of-class-separation for the M and S training set members in the combined training set and test set under these conditions was only 1.6 (Fig. 6B). Only the A replicates were loosely clustered in the unaligned data, but in this case, where the optimal retention time alignment was applied to the test set data, the A and S replicates are both more tightly clustered. This indicates that superfluous retention time variation was removed from the data by correcting retention time shifting. However, the variations captured in the first two PCs did not fully differentiate the samples. The A and S replicates were more tightly clustered compared to unaligned data, but the T, C, and M replicates were clustered together in one group. Thus, PCA with piecewise alignment was an improvement over no alignment, but was not fully sufficient to correctly classify the test set data. Piecewise alignment also improved the supervised LDA pattern recognition results over the LDA of unaligned test set data. With LDA, alignment reduced the misclassification rates for both M (2.5%) and C (5%), while maintaining perfect classification rates for the other classes. Thus, in order to further improve the classification, not only alignment, but also feature selection, was required.

4.8. PCA after ANOVA feature selection of unaligned data

Feature selection was applied directly to the unaligned data (training set combined with test set) with a threshold of 500. Fig. 6C contains the corresponding scores plot. The

Fisher ratio threshold of 500 was chosen because it yielded the maximum degree-of-class-separation (4.8) achievable by applying feature selection alone to the training set according to Fig. 5B. At this threshold, PC 1 captured 90% of the variance in the data and PC 2 captured 9% of the variance. The S and T scores clustered together as did the C and M scores, despite the attempt to reduce the test set. The fact that feature selection alone fails to classify the samples, but alignment coupled with feature selection succeeds at classifying the samples, indicates that under the conditions for Fig. 6C, PCA is likely modeling variations due to retention time shifting rather than chemical variation.

4.9. Comparison to other alignment methods

A positive characteristic of piecewise alignment is that it is written to be a fast algorithm that can quickly correct the retention time shifting in large data sets of complex chromatograms. The piecewise alignment algorithm required 1 min to align 210 chromatograms in Matlab. The same set of 210 chromatograms aligned by the Matlab implementation of COW (www.models.kvl.dk/source) required 8 min (for COW the parameters were window = 200 data points and slack = 1 data point). Piecewise alignment works at a suitable pace that makes it a good choice for inclusion in the optimized data reduction and classification method introduced in this report. Also, a feature based alignment and genetic feature selection algorithm recently reported required 3–4 h for a comparably sized data set [13]. In the work reported herein, the total time required for parameter optimization, piecewise alignment, data reduction and classification by clustering was 30 min.

5. Conclusion

A novel data reduction and classification method using optimized parameters was shown to objectively and successfully classify a large data set of GC fingerprints by producing a scores plot with accurate clustering. The piecewise alignment algorithm quickly and accurately corrected retention time variations for the complex chromatograms in order to reduce the complexity of the data set and restore bilinearity prior to PCA. ANOVA feature selection of a training set composed of two classes proved to be a good tool for reducing the data set to features useful for classifying all five classes, which is important in situations where the analyst does not know the source of every sample in the data set. The piecewise alignment algorithm parameters were objectively optimized by analyzing the degree-of-class-separation for a subset of the data (training set) that included only two of the five classes. Finally, accurate clustering of all samples (training set plus the test set) was achieved using piecewise alignment coupled with ANOVA feature selection prior to PCA. Improvements due to alignment and feature selection were noted with PCA clustering as well as with LDA.

Acknowledgments

This work was supported by the Internal Revenue Service through an Interagency Agreement with the U.S. Department of Energy. The Pacific Northwest National Laboratory is operated by Battelle Memorial Institute for the U.S. Department of Energy under contract DE-AC05-76RLO 1830. The views, opinions, and/or findings contained in this report are those of the authors and should not be construed as the official International Revenue Service position, policy, or decision unless designated by other documentation.

References

[1] G. Stephanopoulos, D. Hwang, W.A. Schmitt, J. Misra, Bioinformatics 18 (2002) 1054.
[2] U.G. Indahl, Chemom. Intell. Lab. Syst. 49 (1999) 19.
[3] W. Wu, S.C. Rutan, A. Baldovin, D.-L. Massart, Anal. Chim. Acta 335 (1996) 11.
[4] D.L. Swets, J.J. Weng, IEEE Trans. Pattern Anal. Machine Intell. 18 (1996) 831.
[5] C. Kan, Pattern Recognit. 35 (2002) 143.
[6] C.P. Wang, T.L. Isenhour, Anal. Chem. 59 (1987) 649.
[7] P.H.C. Eilers, Anal. Chem. 76 (2004) 404.
[8] J. Forshed, I. Schuppe-Koistinen, S.P. Jacobsson, Anal. Chim. Acta 487 (2003) 189.
[9] J.T.W.E. Vogels, A.C. Tas, J.V.D. Greef, Chemom. Intell. Lab. Syst. 21 (1993) 249.
[10] K.S. Booksh, C.M. Stellman, W.C. Bell, M.L. Myrick, Appl. Spectrosc. 50 (1996) 139.
[11] N.-P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, J. Chromatogr. A 805 (1998) 17.
[12] K.J. Johnson, B.W. Wright, K.H. Jarman, R.E. Synovec, J. Chromatogr. A 996 (2003) 141.
[13] B.K. Lavine, Anal. Chim. Acta 437 (2001) 233.
[14] G. Malmquist, R. Danielsson, J. Chromatogr. A 687 (1994) 71.
[15] R.J.O. Torgrip, M. Aberg, B. Karlberg, S.P. Jacobsson, J. Chemom. 17 (2003) 573.
[16] G. Tomasi, F. van den Berg, C. Andersson, J. Chemom. 18 (2004) 231.
[17] E. Reiner, Biomed. Mass Spectrom. 6 (1979) 491.
[18] D.L. Massart, Chemometrics: A Textbook, Elsevier Sciences Ltd, New York, 1988.
[19] R.G. Brereton, Chemometrics: Data Analysis for the Laboratory and Chemical Plant, Wiley, New York, 2003.
[20] K.J. Johnson, R.E. Synovec, Chemom. Intell. Lab. Syst. 60 (2002) 225.
[21] R.O. Duda, P.E. Hart, Pattern Classifications and Scene Analysis, Wiley, New York, 1973.
[22] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis, Academic Press, New York, 1979.
[23] Y. Tominaga, Chemom. Intell. Lab. Syst. 49 (1999) 105.