Received: 7 July 2014, Revised: 6 December 2014, Accepted: 10 December 2014, Published online in Wiley Online Library: 25 February 2014

(wileyonlinelibrary.com) DOI: 10.1002/cem.2696

Hotelling trace criterion as a figure of merit for the optimization of chromatogram alignment

Edward J. Soares[a]*, Gopal R. Yalla[a], John B. O'Connor[a,b], Kevin A. Walsh[a] and Amber M. Hupp[b]

We present a methodology for optimization of chromatogram alignment using a class separability measure called the Hotelling trace criterion (HTC). This metric is a multi-class distance measure that accounts for within-class and between-class variation. We chose the correlation optimized warping algorithm as our alignment method and used the HTC to judge the effectiveness of the alignment based on algorithm parameters called segment length and max warp.

Biodiesel feedstock samples representing classes of soy, canola, tallow, waste grease, and hybrid were used in our experiments. Fatty acid methyl esters in each biodiesel were separated using gas chromatography-mass spectroscopy. The entire data set was baseline corrected, aligned, normalized, and mean-centered prior to principal components (PCs) analysis. The aligned, baseline corrected data sets were used to compute a figure of merit called warping effect, while the PC-transformed data sets were used to evaluate the HTC. The segment length and max warp parameters that maximized the warping effect and/or HTC were then determined. Scores plots of pairs of PCs, along with 95% confidence ellipses, were created and analyzed.

The results demonstrated that the parameters derived from maximizing the HTC more effectively aligned the data, as evidenced by better clustering of the biodiesels in the scores plots. This behavior was robust to the number of PCs used in the computation of the HTC. We conclude that the HTC is an objective measure of alignment quality that allows for optimal class separability and can be applied to optimize other methods of chromatogram alignment. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: Hotelling trace criterion; correlation optimized warping; principal components analysis; gas chromatography; biodiesel

1. INTRODUCTION

Complex chromatographic data can be challenging to analyze based on the sheer number of chemical components in a sample. Chemometric methods such as principal component analysis have been widely used to determine interesting trends from complex data sets [1–13]. Researchers have used several methods including extracting peak areas [1,4,5] as well as using the full, raw data set [8,14,15]. Extracting retention time peak areas can be straightforward. However, typically, a judgment of the number and type of chemical components must be made by the user. Using the raw data set does not require such a judgment, as every data point in the sample is investigated. However, this method requires more sophisticated data processing as every sample in the set must be aligned prior to any subsequent chemometric analysis [16,17]. With gas chromatography (GC), normal fluctuations occur in both peak height (because of variation in the manually injected volume) and retention time (because of slight differences in oven temperature, analyte interaction on column, etc.). Thus, without retention time alignment, the trends that are determined using chemometric methods of analysis could be skewed or meaningless.

Several authors have proposed alignment algorithms for GC measurements that operate on the entire chromatogram. Wang and Isenhour [18] presented a dynamic programming approach to time warp data derived from gas chromatography/Fourier transform infrared/mass spectroscopy experiments using a distance measure to produce an optimal match. This dynamic time warping (DTW) algorithm requires setting a window constraint and local constraint on the number of one-direction consecutive moves that can be made. Vest Nielsen *et al.* [14] introduced a correlation optimized warping (COW) method that uses piecewise stretching and compression of segments of the data, as well as the linear correlation between matching segments, to optimally align two chromatographic profiles. This algorithm requires the setting of two parameters: segment length, which is the fixed length used to divide up each chromatogram and maximum warp, which is the largest amount of stretching and/or compression a particular segment may undergo. Pierce *et al.*

* *Correspondence to: Edward J. Soares, Department of Mathematics and Computer Science, College of the Holy Cross, 1 College Street, Worcester, MA, 01610, USA.*
E-mail: esoares@holycross.edu

[a] *E. J. Soares, G. R. Yalla, J. B. O'Connor, K. A. Walsh*
Department of Mathematics and Computer Science, College of the Holy Cross, Worcester, MA 01610, USA

[b] *J. B. O'Connor, A. M. Hupp*
Department of Chemistry, College of the Holy Cross, Worcester, MA 01610, USA

[19] presented a variant of local retention time alignment called piece-wise alignment. Like COW, piece-wise alignment uses a segment length parameter to divide up the chromatogram and then shifts segments of the data to find the optimal correlation between the segments. However, it does not incorporate stretching and/or compression thus saving computation time. Several authors [20–22] have tested and compared the effectiveness of alignment algorithms including COW and DTW using experimental data, although none have incorporated objective measures of alignment quality assessment into their analysis.

Most algorithms that align the entire chromatogram require the selection of one or more input parameters and so a natural question arises regarding the selection of a "best" set of parameters that will produce an optimal alignment of the data. To perform such an optimization, one needs to define a performance metric that objectively quantifies the quality of the alignment. Pierce *et al.* [19] define a measure of alignment quality called *degree of class separation*, which is the ratio of the Euclidean distance between two principal component (PC) class means with the square root of their average variances. However, this quantity is only based on data from two of the classes, and it does not account for the linear correlation that may exist between the PCs for a particular class. Sinkov and Harynuk [23] use *cluster resolution* as their criterion for class separability, which is the maximum confidence limit for which confidence ellipses from two different classes do not overlap. For applications in which more than two classes are present, a value for cluster resolution is obtained from each possible pair of classes, and then, the product of these is computed. While this measure may account for separation between all of the classes, it does not measure their separability simultaneously. Skov *et al.* [15] define a measure called *warping effect*, which measures both the degree of similarity in the aligned data set and the amount of distortion the alignment has introduced. Because the data set may contain different types of samples, we may want to preserve these differences post-alignment. But, the warping effect measure may not serve to value difference preservation.

In this work, we investigate the use of the Hotelling trace criterion (HTC) [24] as a metric to determine the parameters that serve to optimally align GC data. The HTC is an omnibus measure of class separability [25] that incorporates both within-class and between-class variations and is the multi-class extension of the Mahalanobis distance [26]. Researchers have previously used the HTC as a quality metric for feature enhancement in image processing [27] and imaging system optimization [28].

We evaluated the suitability of the HTC using data from several biodiesels derived from various feedstocks (soybean oil, canola oil, waste cooking grease, and animal tallow) and analyzed using gas chromatography with mass spectrometry. The COW algorithm is employed as an alignment tool for the data; however, any method that requires input parameters may be employed. We compare the effectiveness of the HTC as a figure of merit to the warping effect metric of Skov *et al.* [15].

2. THEORY

2.1. Nomenclature and terminology

A measurement vector is used to represent a *sample chromatogram*. The *time* axis refers to the direction over which chemical components elute and along which warping and alignment occur. We use italics for scalars (i.e., a), lowercase bold for column vectors (i.e., $\mathbf{a}$), uppercase bold for matrices (i.e., $\mathbf{A}$), and superscript t to denote matrix/vector transpose. Data matrices are also denoted by uppercase bold (i.e., $\mathbf{X}$), where the row index n corresponds to sample chromatogram, and the column index m corresponds to retention time.

We assume a sample chromatogram has M elements (retention times), and that there are a total of N sample chromatograms in a data set. Furthermore, we assume that each sample chromatogram belongs to one of K distinct classes, where there are N_k sample chromatograms in the kth class, with $N_1 + N_2 + \cdots + N_K = N$. Thus, the quantity x_{knm} represents the measurement of peak height at retention time index m in the nth sample chromatogram that belongs to the kth class, while the vector $\mathbf{x}_{kn}$ is vector of measurements of the nth chromatogram from the kth class.

Raw sample chromatograms that undergo some kind of processing or correction will have the processing method denoted by a superscript in parentheses. Thus, a chromatogram that has processed with correction method Q will be denoted by $\mathbf{x}_{kn}^{(\mathrm{Q})}$.

2.2. Baseline correction

Before a sample chromatogram can be aligned and transformed using principal components, baseline correction (BC) should be performed, as baseline shifts can introduce artificial variability in peak height. In our study, the baseline shape of each chromatogram exhibited a non-linear increase as a function of retention time. We employed a variation of the baseline correction method of Eilers and Boelens [29] to correct for this curvature.

The method uses asymmetric least squares smoothing to determine a baseline vector $\mathbf{b}'$ that minimizes

$$f(\mathbf{b}') = \|\mathbf{w}^t(\mathbf{b}' - \mathbf{x}_{kn})\|^2 + \lambda\|\mathbf{D}\mathbf{b}'\|^2 \tag{1}$$

where $\|\cdot\|$ is the Euclidean norm, $\mathbf{w}$ is a vector of weights, λ is a relaxation parameter, and $\mathbf{D}$ is a second-difference matrix (i.e., a tridiagonal matrix with value 2 on the main diagonal, value −1 on the first sub-diagonals above and below the main diagonal, and the rest of the elements zero). The first term of f ensures $\mathbf{b}'$ is a good fit to $\mathbf{x}_{kn}$, while the second term ensures that $\mathbf{b}'$ is smooth. The parameter λ controls the relative importance of these two properties: larger λ results in a smoother baseline and smaller λ results in a better fit to $\mathbf{x}_{kn}$. The weights $\mathbf{w}$ are used to prioritize fitting certain points in $\mathbf{x}_{kn}$ or to selectively ignore points in $\mathbf{x}_{kn}$.

Eilers and Boelens give an iterative algorithm for choosing suitable weights. For our purposes, a non-iterative approach sufficed, as follows. Intuitively, we should assign zero weight to points in $\mathbf{x}_{kn}$ near peaks in the chromatogram, because peaks have large deviations from the baseline. To identify peaks, let us assume

$$\mathbf{x}_{kn} = \mathbf{s} + \mathbf{b} + \boldsymbol{\epsilon} \tag{2}$$

where $\mathbf{s}$ is the non-random true peak height, $\mathbf{b}$ is the true non-random baseline to be estimated, and $\boldsymbol{\epsilon}$ denotes the random error. Furthermore, we assume that $\mathbf{s}$ is sparse (i.e., usually 0) with narrow, large deviation peaks of a fixed maximum width, $\mathbf{b}$ is smooth, and each component of $\boldsymbol{\epsilon}$ is normally distributed with a small standard deviation $\sigma_{\boldsymbol{\epsilon}}$.

Let $\mathbf{m}_i$ be the median vector of elements in $\mathbf{x}_{kn}$ over some appropriately-sized window of size T centered at time index i. Then, $\mathbf{m} \approx \mathbf{b}$, because the median is a robust measure of central tendency and $\mathbf{x}_{kn} \approx \mathbf{b} + \boldsymbol{\epsilon}$, except for some outliers due to peaks in $\mathbf{s}$. Furthermore, the median absolute deviation is a consistent estimator of σ_{ϵ}, with $\sigma_{\epsilon} \approx 1.4826 \times \text{median}(|\mathbf{x}_{kn} - \mathbf{m}|)$.

We consider each x_{kni} that lies outside an envelope defined by $m_i \pm 2\sigma_{\epsilon}$ as an outlier due to peaks in $\mathbf{s}$. So, we choose weight $w_i = 0$ if x_{kni} falls outside this envelope and choose weight $w_i = 1$ otherwise. An asymmetric least squares fitting using these weights is then performed to obtain $\mathbf{b}'$, an estimate for the true baseline $\mathbf{b}$. Subtracting the baseline yields a baseline-corrected chromatogram $\mathbf{x}_{kn}^{(\mathrm{BC})}$

$$\mathbf{x}_{kn}^{(\mathrm{BC})} = \mathbf{x}_{kn} - \mathbf{b}' \approx \mathbf{s} + \epsilon \qquad (3)$$

2.3. Correlation optimized warping

Prior to chemometric analysis, the full chromatograms need to be aligned, as small shifts in chromatographic profiles with respect to retention time can cause severe variations in chemometric analyses [16]. The COW algorithm [14] was chosen to align our data. This method is based on aligning a sample chromatogram to a target chromatogram (i.e., a reference sample) by piece-wise stretching and/or compression of segments of the data, in combination with linear interpolation and optimization with regard to the linear correlation coefficients between corresponding segments in the sample and target chromatograms. The details of the implementation of COW can be found in Vest Nielsen *et al.* [14]. Tomasi *et al.* [21] provide a conceptual example of the underlying algorithm. The COW algorithm uses two input parameters that specify the fixed size of each segment, and the maximum amount of warping each segment may undergo. We will refer to them as *segment length* and *max warp*, respectively. The quality of the alignment depends heavily on the selection of these parameters.

The choice of the reference sample is important, as it serves as the basis of alignment for all of the other samples. In their work, Skov *et al.* [15] discuss a number of approaches that can be taken to choose this target chromatogram. In particular, they refer to a quantity called the *similarity index* (*SI*) as a figure of merit for determining the best reference sample. To compute *SI* for the jth baseline-corrected chromatogram in the k_0th class, $\mathbf{x}_{k_0j}^{(\mathrm{BC})}$, we take the product of the absolute values of the sample correlation coefficients between this chromatogram and all of the other chromatograms in all of the classes

$$SI_j = \prod_{n=1,\, n\neq j}^{N} \left| r\left(\mathbf{x}_{k_0j}^{(\mathrm{BC})}, \mathbf{x}_{kn}^{(\mathrm{BC})}\right) \right| \qquad (4)$$

The chromatogram that possesses the highest *SI* is regarded as being the most similar to all of the other chromatograms. Thus, it is chosen as the target for alignment. One then applies the COW algorithm to each baseline-corrected chromatogram $\mathbf{x}_{kn}^{(\mathrm{BC})}$ using this reference sample to produce an aligned, baseline-corrected chromatogram $\mathbf{x}_{kn}^{(\mathrm{BC,\,COW})}$.

2.4. Data transformation

After baseline correction and alignment, each chromatogram $\mathbf{x}_{kn}^{(\mathrm{BC,\,COW})}$ should be normalized to account for variations in injection volume. To accomplish this, the intensity at each retention time was summed to define a total area under the nth chromatogram in the kth class

$$A_{kn} = \sum_{m=1}^{M} x_{knm}^{(\mathrm{BC,\,COW})} \qquad (5)$$

and the average total area of all of the chromatograms in the data set was also computed

$$\bar{A} = \frac{1}{N}\sum_{k=1}^{K}\sum_{n=1}^{N_k} A_{kn} \qquad (6)$$

Each component of a given chromatogram was subsequently divided by its total area A_{kn} so that each normalized chromatogram had a unit area. To return the data to the same order of magnitude before normalization, each chromatogram was scaled by the average total area previously computed. These steps can be accomplished via

$$\mathbf{x}_{kn}^{(\mathrm{BC,\,COW,\,NORM})} = \frac{\bar{A}}{A_{kn}} \cdot \mathbf{x}_{kn}^{(\mathrm{BC,\,COW})} \qquad (7)$$

Mean-centering (MC) of each chromatogram is often done prior to chemometric analysis in order to shift the relative location of the data to the origin. Centering the data preserves the relative inter-sample relationships and allows one to more easily consider relationships between samples [30]. After area normalization and scaling, the sample grand mean chromatogram is computed

$$\bar{\mathbf{x}}^{(\mathrm{BC,\,COW,\,NORM})} = \frac{1}{N}\sum_{k=1}^{K}\sum_{n=1}^{N_k} \mathbf{x}_{kn}^{(\mathrm{BC,\,COW,\,NORM})} \qquad (8)$$

and then subtracted from each sample chromatogram to compute a mean-centered, aligned, and baseline-corrected chromatogram $\mathbf{x}_{kn}^{(\mathrm{BC,\,COW,\,NORM,\,MC})}$

$$\mathbf{x}_{kn}^{(\mathrm{BC,\,COW,\,NORM,\,MC})} = \mathbf{x}_{kn}^{(\mathrm{BC,\,COW,\,NORM})} - \bar{\mathbf{x}}^{(\mathrm{BC,\,COW,\,NORM})} \qquad (9)$$

In order to more easily identify differences in the chromatographic profiles of the samples, the dimensionality of the chromatograms must be reduced while not eliminating important information contained in the data. The principal components transformation [26] is used for this purpose. It is a multivariate statistical technique that reorders the large numbers of possibly correlated measurements into a smaller set of uncorrelated features, called PCs. More importantly, these PCs still retain most of the variation in the original data set [30,31]. Ideally, only the important discriminating characteristics of the original data are retained within a small set of features, from which natural clusters of similar samples can be identified.

Let $\mathbf{S}$ represents the sample covariance matrix of the entire set of processed sample chromatograms with eigen decomposition [26,32,33] given by

$$\mathbf{S} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^t \qquad (10)$$

where $\mathbf{U}$ is the orthogonal matrix whose columns are the eigenvectors (loadings) of $\mathbf{S}$, and $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues that represent the variances related to each PC variable. Then, $\mathbf{y}_{kn}$, the vector of PCs for sample chromatogram $\mathbf{x}_{kn}^{(\mathrm{BC,\,COW,\,NORM,\,MC})}$ is computed via the matrix transformation

$$\mathbf{y}_{kn} = \mathbf{U}^t \mathbf{x}_{kn}^{(\mathrm{BC,\,COW,\,NORM,\,MC})} \tag{11}$$

Each PC is a linear combination of the original measurements. Furthermore, only the first few elements of $\mathbf{y}_{kn}$ will likely contain useful information for the purpose of discrimination between sample classes.

2.5. Optimization criteria

As previously stated, the COW algorithm requires two user-defined input parameters: segment length and max warp. The selection of these parameters affects how well the alignment is performed. In order to identify the best parameter values, we have to decide on a figure of merit for judging the effectiveness (or quality) of the alignment.

2.5.1. *Warping effect*

Some authors have conjectured that making the entire set of chromatograms as similar as possible while retaining peak shape and area should be the goal of alignment. Skov *et al.* [15] have defined a figure of merit to quantify this similarity called *warping effect*, which is the sum of two quantities: *simplicity* and *peak factor*. Simplicity is related to the rank of the data matrix for the aligned, baseline-corrected chromatograms. A data matrix with rank 1 means that there is only one linearly independent sample chromatogram, and that all of the other chromatograms are scalar multiples of the first. Thus, higher values for simplicity means that the chromatograms are more similar, thus reflecting that they are better aligned.

If **X** is the data matrix for the aligned, baseline-corrected chromatogram profiles, then simplicity is defined to be [15]

$$\text{simplicity} = \sum_{r=1}^{R} \left(\mathrm{SVD}\left(\mathbf{X} \Big/ \sqrt{\sum_{k=1}^{K} \sum_{n=1}^{N_k} \sum_{m=1}^{M} x_{knm}^2} \right) \right)^4 \tag{12}$$

where r is the singular value index and division by the total sum of the elements in **X** scales the singular values so that they sum to 1. Values of simplicity closer to 1 indicate that the chromatograms are better aligned, while values closer to 0 correspond to deviations from ideal alignment.

The second quantity, peak factor, is intended to measure how much the shape and peak area of chromatograms have been changed by the warping. If we define

$$c_{kn} = \left| \frac{\| \mathbf{x}_{kn}^{(\mathrm{BC,\,COW})} \| - \| \mathbf{x}_{kn}^{(\mathrm{BC})} \|}{\| \mathbf{x}_{kn}^{(\mathrm{BC})} \|} \right. \tag{13}$$

as the relative error between a baseline-corrected chromatogram before alignment and after alignment, then peak factor can be computed as [15]

$$\text{peak factor} = \frac{1}{N} \sum_{k=1}^{K} \sum_{n=1}^{N_k} \left(1 - \min\left(c_{kn}, 1 \right)^2 \right) \tag{14}$$

When alignment distorts a sample, c_{kn} will be large, and so its contribution to peak factor will be zero. However, when the sample stays relatively unchanged, c_{kn} will be small and thus will contribute a $1/N$ to the sum. Thus, better alignment corresponds to larger values for simplicity and peak factor and consequently for warping effect.

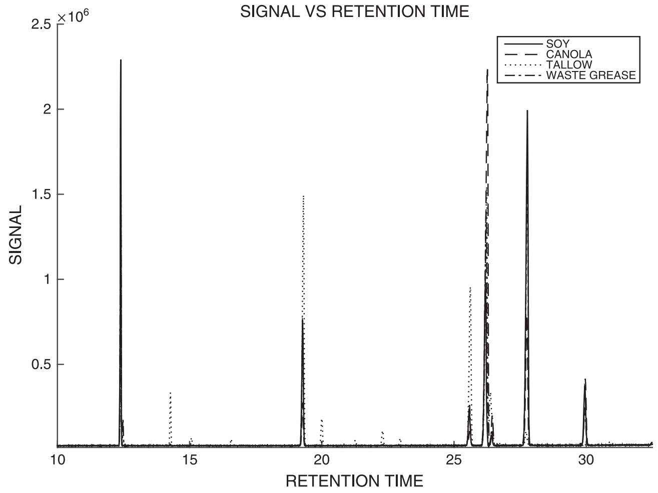

Figure 1. Representative total ion chromatograms from each fuel class showing separation of FAME components for m/z = 20 to 400 for biodiesel fuels produced from different feedstock types.

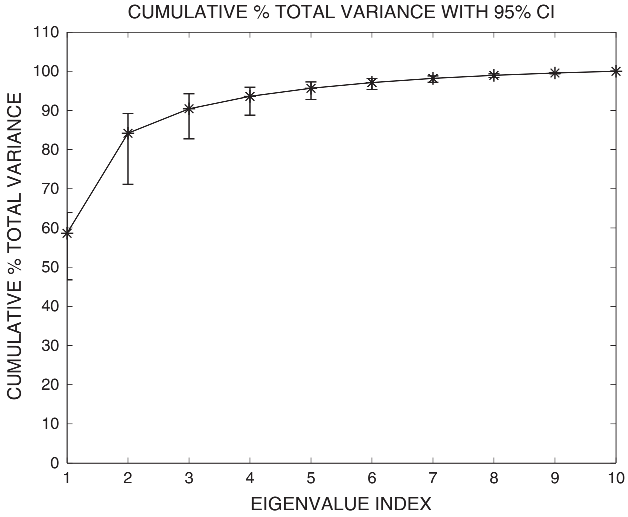

Figure 2. A plot of median cumulative percent total variation versus eigenvalue index, along with 95% confidence bounds.

2.5.2. *Hotelling trace criterion*

For a set of chromatograms that contains samples from different classes, we would like to identify those differences. Therefore, it is desirable to remove variation in peak location along the time axis but retains variation in peak height. Simplicity is a global measure of similarity between all of the samples that does not quantify class separability, and so maximizing it might not serve to identify the segment length and max warp parameters that best retain these differences. Ideally, we would like to use a measure that reflects our ability to discriminate between the different classes of biodiesels.

When there are two multivariate populations present, the sample Mahalanobis distance [26] gives a numerical measure of the distance between the distributions. However, we have $K > 2$ multivariate populations to consider. In this scenario, the *HTC*

[24,27,28], which is the multi-class extension of the Mahalanobis distance, provides us with this same numerical measure of distribution separability. Like the Mahalanobis distance, the HTC incorporates both within-class and between-class variations in the data set. We evaluated the HTC based on the PCs of our transformed sample chromatograms.

Let $\mathbf{z}_{kn} = (y_{kn1}, y_{kn2}, \cdots, y_{knL})^t$ denotes the $L \times 1$ vector corresponding to the first L PCs of $\mathbf{y}_{kn}$, where $\mathbf{y}_{kn}$ is the nth PC vector belonging to the kth class. The sample mean vector $\bar{\mathbf{z}}_k$ and sample covariance matrix $\mathbf{S}_k$ for the kth class are given respectively by

$$\bar{\mathbf{z}}_k = \frac{1}{N_k} \sum_{n=1}^{N_k} \mathbf{z}_{kn} \tag{15}$$

and

$$\mathbf{S}_k = \frac{1}{N_k - 1} \sum_{n=1}^{N_k} \left(\mathbf{z}_{kn} - \bar{\mathbf{z}}_k\right)\left(\mathbf{z}_{kn} - \bar{\mathbf{z}}_k\right)^t \tag{16}$$

Furthermore, we define the grand mean vector of all of the classes as

$$\bar{\bar{\mathbf{z}}} = \sum_{k=1}^{K} P_k \bar{\mathbf{z}}_k \tag{17}$$

where $P_k = N_k/N$ is the probability of occurrence for class k. Using these quantities, we define the within-class scatter matrix $\mathbf{S}_{wc}$ as

$$\mathbf{S}_{wc} = \sum_{k=1}^{K} P_k \mathbf{S}_k \tag{18}$$

and the between-class scatter matrix $\mathbf{S}_{bc}$ as

$$\mathbf{S}_{bc} = \sum_{k=1}^{K} P_k \left(\bar{\mathbf{z}}_k - \bar{\bar{\mathbf{z}}}\right)\left(\bar{\mathbf{z}}_k - \bar{\bar{\mathbf{z}}}\right)^t \tag{19}$$

The matrix $\mathbf{S}_{wc}$ quantifies the average multi-dimensional dispersion within each class about the class mean, while $\mathbf{S}_{bc}$ quantifies the average multi-dimensional dispersion between each class mean and the grand mean. The HTC is then defined to be

$$J = \mathrm{tr}\left(\mathbf{S}_{wc}^{-1}\mathbf{S}_{bc}\right) \tag{20}$$

where tr(·) denotes the trace of the matrix argument. Large values of J correspond to better class separability. Smaller within-class variation increases the value of J, as does larger between-class variation.

Thus, we seek to use the HTC as our optimization metric, in order to identify the values of segment length and max warp needed for COW alignment that will maximize the separability of the different classes of biodiesels. It is important for the reader to note that our computation of the HTC is dependent on the number of PCs (L) that we include in $\mathbf{z}_{kn}$. In fact, as L increases, the value of the HTC will also increase.

3. EXPERIMENTAL METHODS

3.1. Chemicals

Biodiesel fuel samples were obtained from various manufacturers throughout the USA (Minnesota Soybean Processors (soybean biodiesel, Minn Soy 2010, 2011), Western Dubuque Biodiesel (soybean biodiesel, Iowa Soy 2010), Iowa Renewable Energy (soybean biodiesel, canola biodiesel, tallow biodiesel, IRE Soy, canola, Tallow 2012), National Institute of Standards and Technology (Standard Reference Material (SRM) 2772, soy SRM, soybean biodiesel from Ag Processing Inc. and SRM 2773, animal SRM, tallow/soybean biodiesel mixture from Smithfield BioEnergy LLC), ADM Company (canola biodiesel, ADM Canola 2010, 2011), TMT Biofuels (waste grease biodiesel, Waste Grease 2010, 2011), Texas Green Manufacturing (beef tallow biodiesel, Texas Tallow 2010, 2012), and Keystone Biofuels (unknown biodiesel, Keystone 2010)) and were stored in their original shipping container at 4 °C. Prior to dilution, each biodiesel was gradually warmed to room temperature and was inverted multiple times to ensure homogeneity. An amount of 1 mL of each biodiesel sample was diluted to 100 mL total volume with methylene chloride (BDH chemicals distributed by VWR, West Chester, PA, USA), and 1 mL of 0.30 M tridecanoic acid methyl ester (Fluka) was added to a 50-mL volumetric flask and was diluted to volume with the 100:1 biodiesel. Tridecanoic acid methyl ester (C13) was chosen as an internal standard as it was not present in any of the biodiesel samples originally. All diluted biodiesel solutions were stored in amber bottles at 4 °C and were gradually warmed to room temperature prior to analysis.

3.2. Instrumentation

Separations were performed using an Agilent 6890 gas chromatograph coupled with an Agilent 5937 mass spectrometer (Agilent Technologies, Santa Clara, CA, USA) and have been presented in detail previously [34]. The gas chromatography with mass spectrometry was equipped with a polyethylene glycol fused-silica capillary column of dimensions 30 m × 0.25 mm × 0.25 μm (ZB-WAXplus, Phenomenex). The oven temperature was optimized to ensure baseline resolution of all Fatty acid methyl esters (FAME) in a 37-component FAME standard (Supelco) and was as follows: 60 °C (hold 2 min) to 150 °C at 13 °C/min to 230 °C at 2 °C/min. High purity helium was used

Table I. Segment length and max warp at the maximum Hotelling trace criterion (HTC) value as a function of the number of principal components (PCs) used in the calculation of the HTC. The value of warping effect for these segment length-max warp combinations was also included for comparison purposes

Number of PCs (L)	Segment length	Max warp	Max HTC	Warp effect
1	64	3	143.5	1.65
2	55	8	244.2	1.69
3	70	6	298.6	1.69

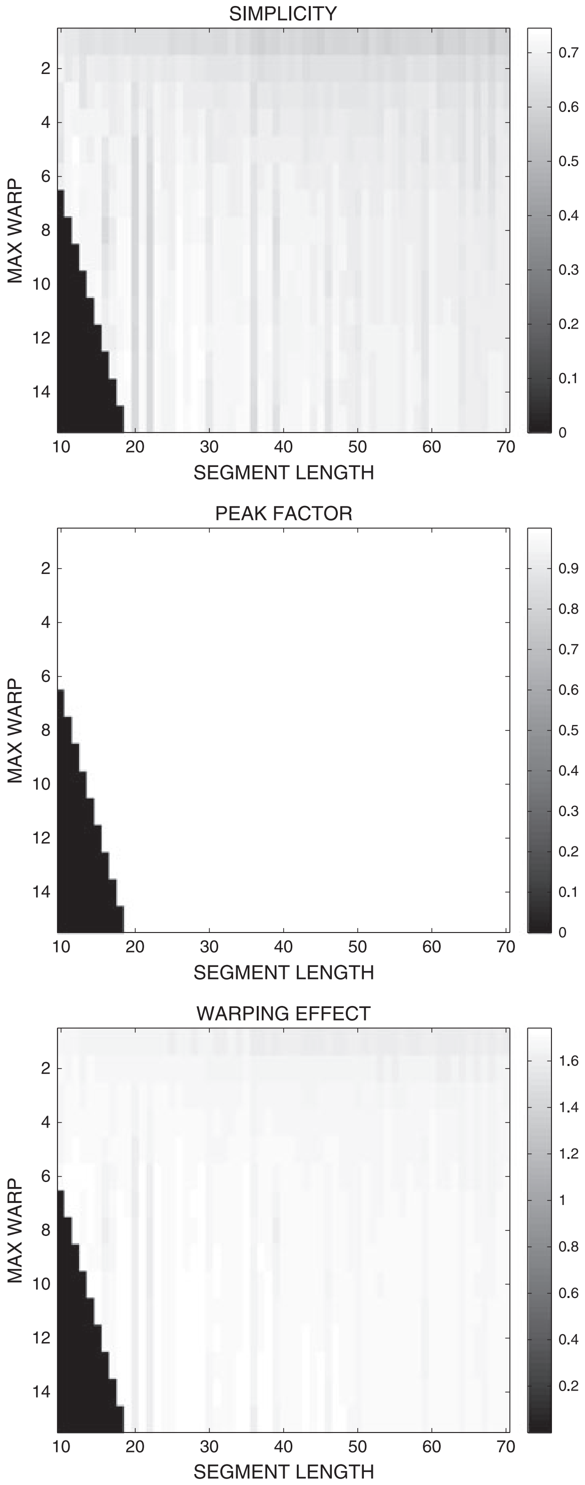

Figure 3. Two-dimensional (2D) density plots of simplicity, peak factor, and warping effect. Maximum values for both simplicity and warping effect occurred for segment length-max warp of (26,15).

as a carrier gas at a nominal flow rate of 1.5 mL/min. Each sample was injected via syringe (1 μL injected from 10 μL syringe, Hamilton Company) with a split ratio of 50:1. The inlet and transfer line temperatures were held at 250 °C and 280 °C, respectively. An electron-impact ionization source was utilized with a quadrupole mass analyzer operated in full-scan mode (m/z 20,300) with a sampling rate of 4.94 scans/s. The mass spectrometer source and quadrupole were held at 230 °C and 150 °C, respectively. FAME identification was performed using the mass spectra library (National Institute of Standards and Technology mass spectral search program version 2.0a, Gaithersburg, MD, USA) as well as retention time comparison to the FAME standard. Representative total ion chromatograms from each fuel class showing separation of FAME components for m/z = 20 to 400 for biodiesel fuels produced from different feedstock types are shown in Figure 1.

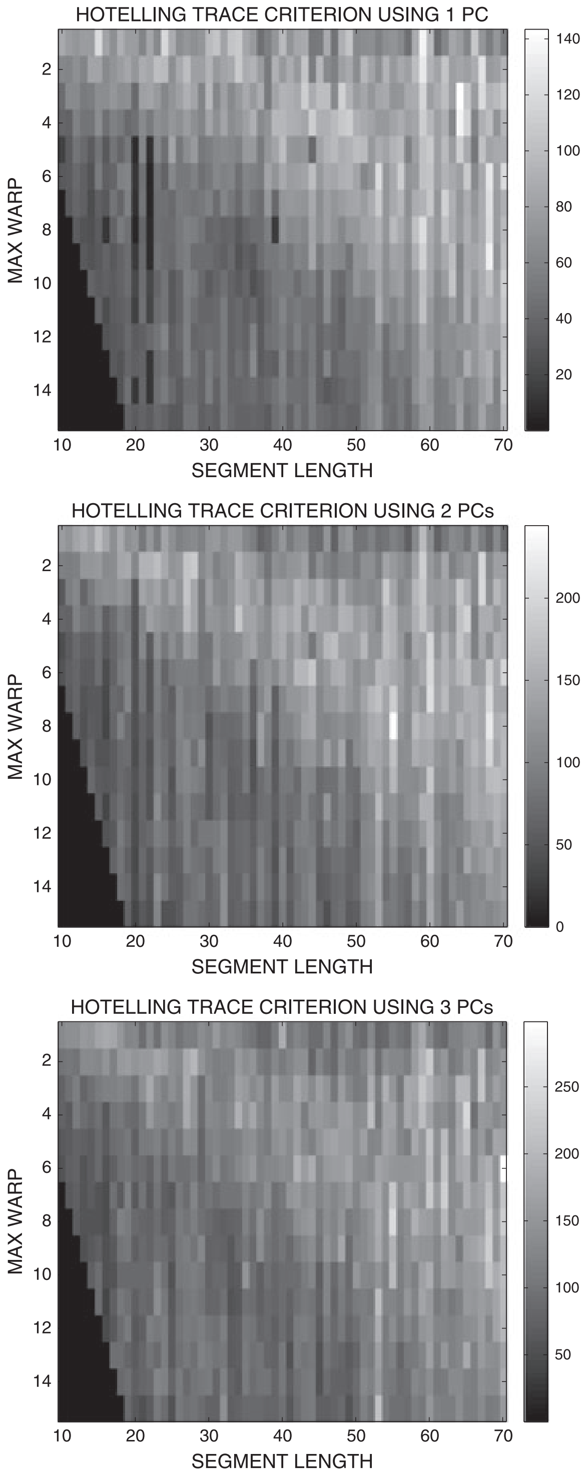

Figure 4. Two-dimensional (2D) density plots of Hotelling trace criterion (HTC). Maximum values occurred for segment length-max warp combinations of (64,3) using one principal component (PC), (top), (55,8) using two PCs (middle), and (70,6) using three PCs (bottom).

3.3. Data processing

Total ion chromatograms were extracted from Chemstation using a macro developed by Infometrix (Bothell, WA, USA). All chromatograms were first baseline corrected using a python implementation of the method previously described with window size $T = 1000$, a maximum peak detection width of 20, and relaxation parameter $\lambda = 10^7$. In addition, portions of the chromatogram that did not contain chemical information (0–5 min and 40.37–48.92 min) were removed prior to further chemometric analysis, resulting in 10,500 sample values in each chromatogram.

Next, the chromatograms were aligned using a MATLAB implementation of the COW algorithm (http://www.models.life.ku.dk/algorithms) under the same combinations of segment length-max warp as seen in Skov *et al.* [15]. Segment lengths ranged from 10 through 70. For segment lengths between 10 and 19 (inclusive), max warp was equal to segment length minus 4. For segment lengths greater than or equal to 20, max warp was fixed at 15. This produced 870 total aligned, baseline-corrected data sets. The reference sample, a waste grease, was determined as the sample chromatogram that produced the largest *SI*.

The 870 aligned, baseline-corrected data sets were then normalized and scaled as previously described using MATLAB scripts written in-house. The PC transform was then computed for each data set and was applied to each chromatogram to generate the corresponding PC scores using MATLAB's statistics toolbox. Only PC information regarding the 10 largest eigenvalues was retained.

After all of the data had been fully processed, the figures of merit were tabulated. For each of the 870 aligned, baseline-corrected data sets, the value of warping effect was computed. Because each data file correspondeds to COW processing with a particular combination of segment length-max warp, we arranged the values of warping effect into a two-dimensional (2D) density plot, with segment length along the horizontal axis and max warp along the vertical axis. Furthermore, for each of the 870 PC-transformed data files, the HTC was computed as a function of the number of PCs *L*. There were five classes of biodiesels: soy (six different samples), canola (three different samples), tallow (three different samples), waste grease (two different samples), and hybrid (one sample – 15% soy and 85% tallow) with each sample measured in three different runs. The HTC values also corresponded to COW processing with a particular segment length-max warp, so the HTC values were similarly arranged into 2D density plots. MATLAB scripts to perform these computations were written in-house and are available from the authors upon request.

The maximum warping effect was found to be 1.74, obtained using a segment length-max warp pair of (26,15). The reader should also note that processing with this parameter combination produced the following values for the HTC: 31.3 (L = 1), 55.9 (L = 2), and 104.6 (L = 3). We chose L = 3 as the maximum number of PCs to include in the calculation of HTC, as over 90% of the cumulative percent total variation is accounted for when L = 3, as can be seen in Figure 2. We also found the maximum HTC value as a function of L and determined the combinations of segment

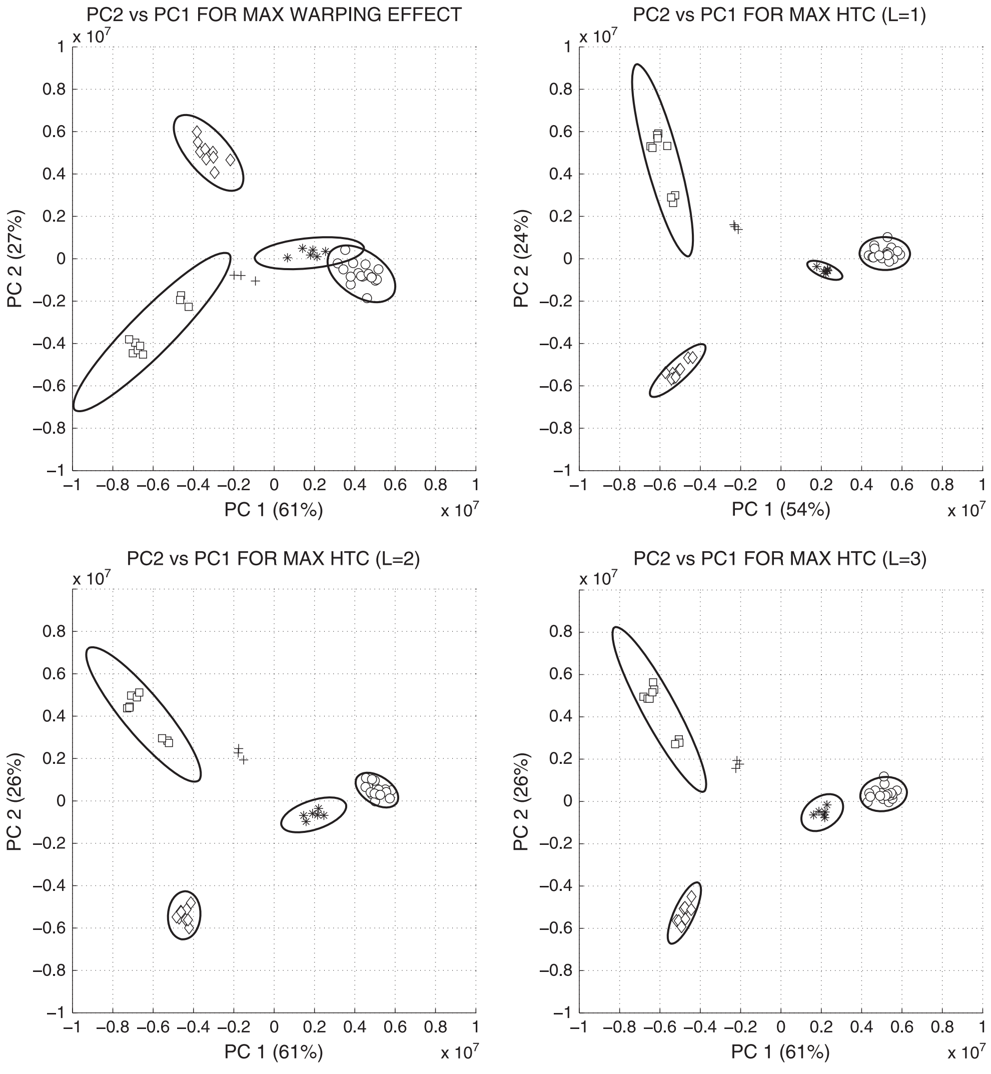

Figure 5. Scatter plots of principal component (PC)2 versus PC1 for combinations of segment length-max warp (26,15) (top left), (64,3) (top right), (55,8) (bottom left), and (70,6) (bottom right). Classes displayed are as follows: soy (○), canola (◇), tallow (□), waste grease (*), and hybrid (+).

length-max warp that produced them. It should be noted that these combinations changed with *L*. These results can be seen in Table I.

Two-dimensional scores plots of combinations of the first, second, and third PCs of the data sets corresponding to segment length-max warp combinations of (26,15), (64,3), (55,8), and (70,6), were then created. Confidence ellipses were also determined using a method similar to that described in [35] and implemented by Schwarz [36]. These were included on the scores plots.

4. RESULTS AND DISCUSSION

The results of our investigation into the optimization of the COW algorithm parameters can be seen in Figures 3–7 and Tables II–III. Figure 3 displays 2D density plots of simplicity, peak factor, and warping effect, as functions of segment length-max warp. The analogous 2D density plots of the HTC, using one, two, or three PCs in its computation are given in Figure 4. 2D scores plots of pair-combinations of the PCs, along with corresponding 95% confidence ellipses are displayed in Figure 5 (PC2 vs PC1), in Figure 6 (PC3 vs PC1), and in Figure 7 (PC3 vs PC2), for the four groupings of segment length-max warp discussed in the Experimental Methods section. Table II lists the Euclidean distances between each pair of class means, while Table III lists the ratios of the standard deviations along the principal axes of each class, where the numerator is the class standard deviation of the data derived from maximizing the HTC, while the denominator is the class standard deviation of the data derived from maximizing the warping effect.

Considering Figure 3, the density plot for peak factor (middle) is fairly uniform. In fact, peak factor values ranged from 0.9934 to 1.0. Because of this narrow range, the density plot for warping effect was approximately the same as the density plot for simplicity plus a constant factor. We note that maximum values of both simplicity and warping effect occurred at segment length-max warp combination (26,15). However, the range of values of both of these figures of merit is small. Therefore, based only on these density plots, it is difficult to determine if the segment length-max warp parameters corresponding to the maximum will result in any meaningful differences in discriminability between the classes.

This limited range in values is not seen for the HTC figure of merit. In Figure 4, the corresponding 2D density plots for

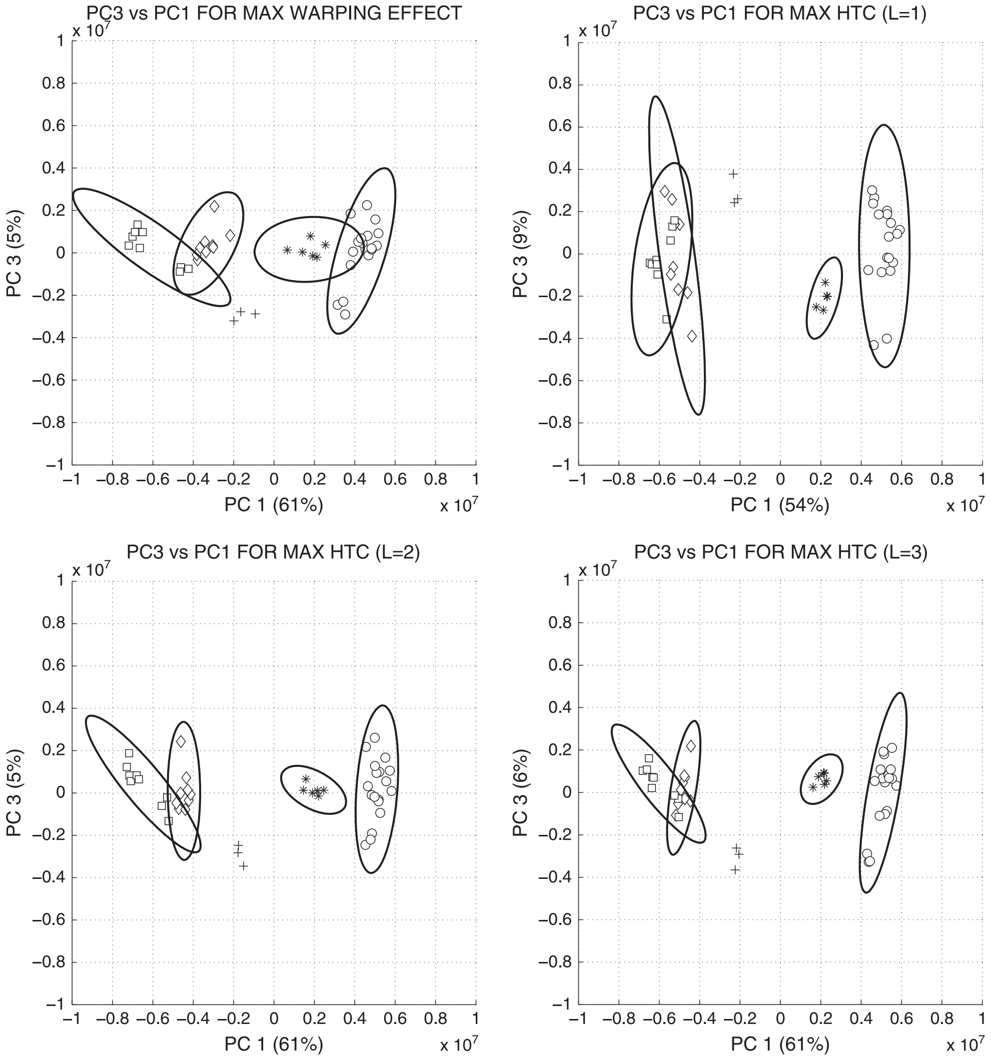

Figure 6. Scatter plots of principal component (PC)3 versus PC1 for combinations of segment length-max warp (26,15) (top left), (64,3) (top right), (55,8) (bottom left), and (70,6) (bottom right). Classes displayed are as follows: soy (○), canola (◇), tallow (□), waste grease (*), and hybrid (+).

the HTC, as a function of the segment length-max warp combination, are given. The broader range in values can be seen visually and by examining the color bars in each plot. As previously noted, maximum values occurred for segment length-max warp combinations of (64,3) using one PC, (55,8) using two PCs, and (70,6) using three PCs. Two important observations can be noted. First, the HTC values exhibit greater variation in magnitude as compared with the measures of simplicity and warping effect. Thus, there should be substantive differences in class separability when using different combinations of segment length-max warp. Second, the magnitude of the HTC increases as the number of PCs used in the calculation increases. Thus, the user must decide to either use a specific number of PCs in the calculation of the HTC or to evaluate the results for a variety of numbers of PCs.

We now turn our discussion to the PC scores plots. Recalling Figure 2, the first three PCs account for approximately 90% of the total variation, on average. Thus, scores plots of pair-combinations of the first three PCs should indicate optimal clustering of the different types of biodiesels.

Examining Figure 5 (top left), when the data are aligned using a segment length-max warp combination of (26,15), parameters found to maximize the warping effect; canola and tallow classes are well separated. However, the soy and waste grease classes overlap. Moreover, the samples from the hybrid class lie outside of the 95% confidence ellipses of the other classes but are spatially close to the tallow class. This makes sense because the hybrid samples contain 85% tallow. Selecting alignment parameters that maximize the HTC (top right and bottom left and right) figure of merit leads to stronger separation for all classes with no overlapping. Again, the hybrid samples remain spatially close to the tallow class.

The reader will note that confidence ellipses were not calculated for the hybrid class. This is because of the fact that there were only three observations in this class. This sample size was not sufficient to derive an accurate estimate for the covariance matrix of that class [37]. The eigenvectors derived from the diagonalization of this covariance matrix are used to determine a confidence ellipse.

Considering the scores plots of PC3 versus PC1 in Figure 6, all of the combinations of segment length-max warp result in an overlap of the canola and tallow classes. However, only those combinations that correspond to a maximized HTC kept the soy and waste grease classes separated. None of the combinations

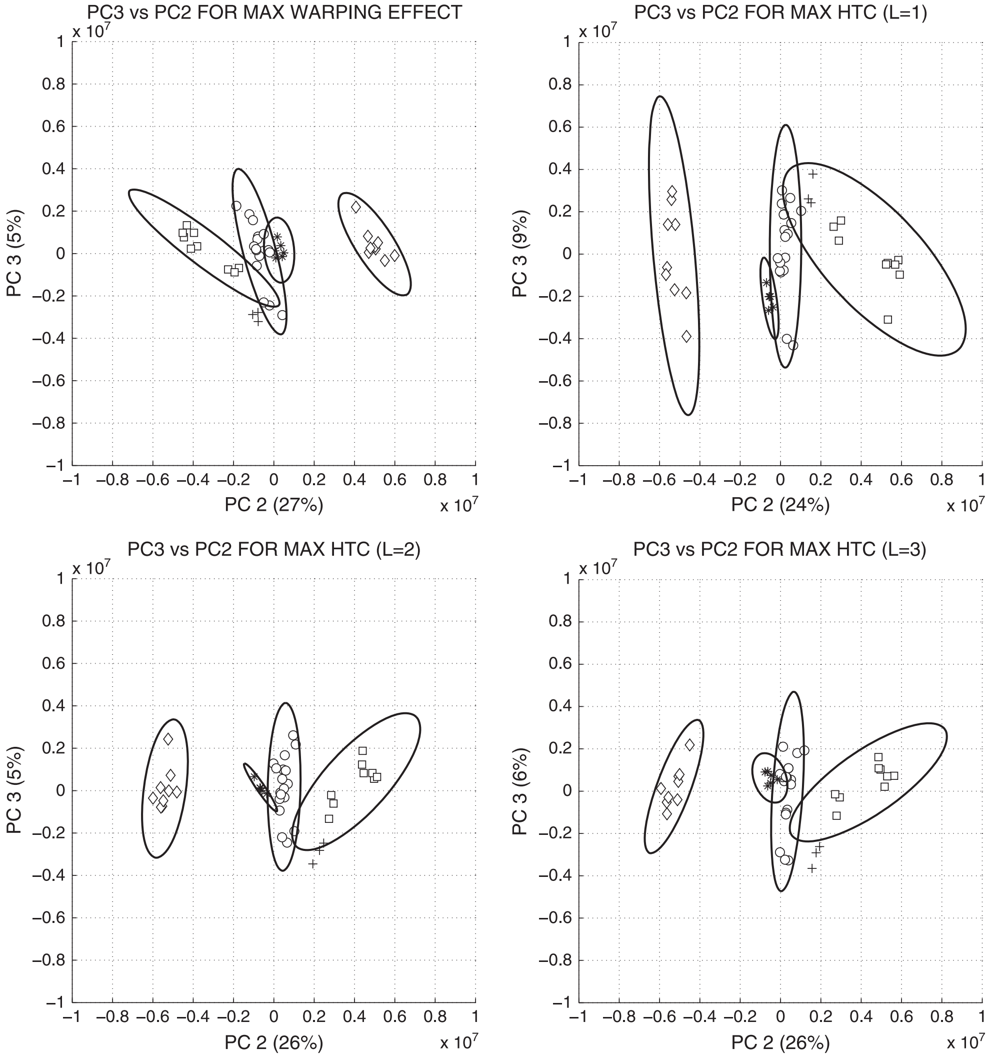

Figure 7. Scatter plots of principal component (PC)3 versus PC2 for combinations of segment length-max warp (26,15) (top left), (64,3) and (55,8) (top right), (55,8) (bottom left), and (70,6) (bottom right). Classes displayed are as follows: soy (○), canola (◇), tallow (□), waste grease (*), and hybrid (+).

Table II. Euclidean distances between pairs of class means for data in scores plots comparing principal component (PC)2 versus PC1. All values should be scaled by 10^6

Segment length/max warp (26,15)

Class	Soy	Canola	Tallow	Waste grease
Soy	0	–	–	–
Canola	9.49	0	–	–
Tallow	10.74	8.91	0	–
Waste grease	2.76	6.88	8.64	0

Segment length/max warp (64,3)

Class	Soy	Canola	Tallow	Waste grease
Soy	0	–	–	–
Canola	11.66	0	–	–
Tallow	11.87	9.95	0	–
Waste grease	3.08	8.69	9.58	0

Segment length/max warp (55,8)

Class	Soy	Canola	Tallow	Waste grease
Soy	0	–	–	–
Canola	11.24	0	–	–
Tallow	12.11	9.69	0	–
Waste grease	3.35	7.98	9.68	0

Segment length/max warp (70,6)

Class	Soy	Canola	Tallow	Waste grease
Soy	0	–	–	–
Canola	11.40	0	–	–
Tallow	11.80	9.71	0	–
Waste grease	3.16	8.33	9.44	0

Table III. Ratios of standard deviations between corresponding principal axes for data derived from maximizing Hotelling trace criterion (HTC) versus data derived from maximizing warping effect

Ratios for segment length/max warp (64,3) to (26,15)

Class	first major axis	second major axis
Soy	1.36	1.02
Canola	2.46	0.97
Tallow	1.03	2.15
Waste grease	0.74	0.47

Ratios for segment length/max warp (55,8) to (26,15)

Class	first major axis	second major axis
Soy	0.94	0.92
Canola	1.06	0.80
Tallow	0.86	1.30
Waste grease	0.68	0.68

Ratios for segment length/max warp (70,6) to (26,15)

Class	first major axis	second major axis
Soy	1.14	0.71
Canola	1.10	0.69
Tallow	0.86	1.30
Waste grease	0.49	0.69

obscure the hybrid class; however, it appears less isolated from the other classes in the plot for the (26,15) combination.

For completeness, we also wanted to determine how well the second and third PCs together separate the classes. This can be seen in Figure 7. Examining this plot, all combinations separate the canola class well. However, none of the combinations allow for easy discrimination between the soy, tallow, waste grease, and hybrid classes. PC3 only accounts for about 5% of the total variation in the data, while PC2 accounts for around 25% of the total variability. We conclude that these two PCs alone do not account for enough of the variation in the data to separate the classes.

At this point, it is clear that comparison of the first two PCs best allows for discrimination between the classes. Between-class variability seems larger for the combinations where the HTC is maximized, as opposed to the combination where the warping effect is maximized. Also, within-class variability seems to be reduced, at least for some of the classes.

To quantify these observations, we computed the Euclidean distances between each pair of class means, for each combination of segment length-max warp that we analyzed. We also computed the ratios of the standard deviations between the classes where the HTC was maximized versus those where the warping effect was maximum. This was accomplished by finding the eigen decomposition of the covariance matrix for each class separately and by using the square root of each eigenvalue to measure the length of each principal axis. The ratios of the standard deviations of the corresponding principal axes were then tabulated. A ratio of 1.0 would mean that the two methods produced the same amount of within-class variability in that class for that principal direction of the distribution, while a ratio less than one means that the data derived from maximizing the HTC has less within-class variation in that class for that principal direction of the distribution. The reader should note that the orientation of the axes are not incorporated into this quantity. The results can be seen in Tables II and III.

Examining Table II, the Euclidean distance between each pair of class means is greater for the data produced from the segment length-max warp combinations derived by maximizing the HTC, as compared to that combination derived by maximizing the warping effect. This is expected because of the fact that the HTC does incorporate between-class variation into its estimate of class separability. Also, this result is consistent regardless of the number of PCs that are used in the calculation of the HTC.

Considering Table III, there is no discernible pattern to whether one method consistently reduces within-class variability over another method. For some classes, within-class variability is smaller using the segment length-max warp derived from maximizing the HTC, while for others, it is smaller using the combination derived from maximizing warping effect. However, it is worth noting that when using two PCs to compute the HTC, within-class variability is reduced in all of the classes with respect to both principal axes, except for canola along its first major axis and tallow along its second major axis.

We remind the reader that within-class variation is not minimized, and between-class variation is not maximized *simultaneously* when the HTC is maximized. The HTC is a summary measure that incorporates estimates of both within-class and between-class variations. Thus, we would not expect within-class variation to be systematically smaller when the HTC is at a maximum.

5. CONCLUSIONS

We have presented a method for optimization of chromatogram alignment using a class separability criterion. The optimal segment length and max warp for the COW algorithm were found by evaluating a figure of merit called the HTC. In addition, we compared our results with those derived from maximizing the warping effect figure of merit of Skov *et al.* [15]. These metrics were tested on data derived from biodiesel feedstock samples representing classes of soy, canola, tallow, waste grease, and hybrid.

The results demonstrated that the combination of segment length and max warp derived from maximizing the HTC produced scores plots in which different classes of biodiesels were optimally separated, while the parameters derived from maximizing warping effect did not separate the classes as well. This behavior was robust to the number of PCs used in the computation of the HTC. Thus, the HTC can be used to find the optimal parameter values for the COW algorithm.

One limitation in using the HTC is that the classes to which each biodiesel belongs must be known. Thus, the HTC is appropriate to use to optimize a particular known multi-class data set or to aid in the construction of an optimal linear discriminant [25,38] for classification of unknown biodiesels, as long as the unknown samples share similar chemical properties with the known training set. We conclude that the HTC is an objective measure of the quality of chromatogram alignment that allows for optimal class separability and which can be applied to optimize other methods of chromatogram alignment.

Acknowledgements

The authors are greatly appreciative of financial support provided by the University Syringe Program Grant from Hamilton Company (AMH), Robert L. Ardizzone Fund for Junior Faculty Excellence (AMH), and College of the Holy Cross. In addition, the authors thank the National Institute of Standards and Technology (NIST, Gaithersburg, MD, USA) for supplying the FAME standards and Minnesota Soybean Processors, Western Dubuque Biodiesel, ADM Company, Keystone Biofuels, TMT Biofuels, Texas Green Manufacturing, and Iowa Renewable Energy for supplying biodiesel samples.

REFERENCES

1. Doble P, Sandercock M, Du Pasquier E, Petocz P, Roux 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**: 26–39.
2. Eide I, Zahlsen K. A novel method of chemical fingerprinting of oil and petroleum products based on electrospray mass spectrometry and chemometrics. *Energ. Fuel.* 2005; **19**: 964–967.
3. Eide I, Zahlsen K. Standardizing the novel method for chemical fingerprinting of oil and petroleum products based on positive electrospray mass spectrometry and chemometrics. *Energ. Fuel.* 2006; **20**: 265–270.
4. Sandercock PML, Du Pasquier E. Chemical fingerprinting of unevaporated automotive gasoline samples. *Forensic Sci. Int.* 2003; **134**: 1–10.
5. Sandercock PML, Du Pasquier E. Chemical fingerprinting of gasoline part 3. Comparison of unevaporated automotive gasoline samples from Australia and New Zealand. *Forensic Sci. Int.* 2004; **140**: 71–77.
6. Gaines RB, Hall GJ, Frysinger GS, Gronlund WR, Juaire JL. Chemometric determination of target compounds used to fingerprint unweathered diesel fuels. *Environ. Forensics* 2006; **7**: 77–87.
7. Johnson KJ, Rose-Pehrsson SL, Morris RE. Characterization of fuel blends by GC-MS and multi-way chemometric tools. *Pet. Sci. Technol.* 2006; **24**: 1175–1186.
8. Hupp AM, Marshall LJ, Campbell DI, Waddell Smith R, McGuffin VL. Chemometric analysis of diesel fuel for forensic and environmental applications. *Anal. Chim. Acta* 2008; **606**: 159–171.
9. Marshall LJ, McIlroy JW, McGuffin VL, Waddell Smith R. Association and discrimination of diesel fuels using chemometric procedures. *Anal. Bioanal. Chem.* 2009; **394**: 2049–2059.
10. Flood MF, Goding JC, OConnor JB, Ragon DY, Hupp AM. Analysis of biodiesel feedstock using GCMS and unsupervised chemometric methods. *J. Am. Oil Chem. Soc.* 2014; **91**: 1443–1452.
11. Jalali-Heravi M, Moazeni RS, Sereshti H. Analysis of Iranian rosemary essential oil: application of gas chromatography-mass spectrometry combined with chemometrics. *J. Chromatogr. A* 2011; **1218**: 2569–2576.
12. Pizarro C, Rodriquez-Tecedor S, Perez-del-Notario N, Gonzalez-Saiz JM. Recognition of volatile compounds as markers in geographical discrimination of Spanish extra virgin olive oils by chemometric analysis of non-specific chromatography volatile profiles. *J. Chromatogr. A* 2011; **1218**: 518–523.
13. Tistaert C, Dejaegher B, Vander Heyden Y. Chromatographic separation techniques and data handling methods for herbal fingerprints: a review. *Anal. Chim. Acta* 2011; **690**: 148–161.
14. Vest Nielsen NP, Carstensen JM, Smedsgaard J. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *J. Chromatogr. A* 1998; **805**: 17–35.
15. Skov T, van den Berg F, Bro R. Automated alignment of chromatographic data. *J. Chemometrics* 2006; **20**: 484–497.
16. Malmquist G, Danielsson R. Alignment of chromatographic profiles for principal component analysis: a prerequisite for fingerprinting methods. *J. Chromatogr. A* 1994; **687**: 71–88.
17. Johnson KJ, Wright BW, Jarman KH, Synovec RE. High-speed peak matching algorithm for retention time alignment of gas chromatgraphic data for chemometric analysis. *J. Chromatogr. A* 2003; **996**: 141–155.
18. Wang CP, Isenhour TL. Time-warping algorithm applied to chromatographic peak matching gas-chromatography Fourier-transform infrared mass-spectrometry. *Anal. Chem.* 1987; **59**: 649–654.
19. Pierce KM, Hope JL, Johnson KJ, Wright BW, Synovec RE. Classification of gasoline data obtained by gas chromatography using a piecewise alignment algorithm combined with feature selection and principal component analysis. *J. Chromatogr. A* 2005; **1096**: 101–110.
20. Pravdva V, Walczak B, Massart DL. A comparison of two algorithms for warping of analytical signals. *Anal. Chim. Acta* 2002; **456**: 77–92.
21. Tomasi G, van den Berg F, Andersson C. Correlation optimized warping and dynamic time warping as pre-processing methods for chromatographic data. *J. Chemometrics* 2004; **18**: 231–241.
22. van Nederkassel AM, Daszykowski M, Eilers PHC, Vander Heyden Y. A comparison of three algorithms for chromatograms alignment. *J. Chromatog. A* 2006; **1118**: 199–210.
23. Sinkov NA, Harynuk JJ. Cluster resolution: a metric for automated, objective and optimized feature selection in chemometric modeling. *Talanta* 2011; **83**: 1079–1087.
24. Hotelling H. The generalization of student's ratio. *Ann. Math. Stat* 1931; **2**: 360–378.
25. Fukunaga K. *Introduction to Statistical Pattern Recognition* (2nd edn). Academic Press: Boston, MA, 1990.

26. Mardia KV, Kent JT, Bibby JM. *Multivariate Analysis*. Academic Press: New York, NY, 1979.
27. Fiete RD, Barrett HH. Using the Hotelling trace criterion for feature enhancement in image processing. *Opt. Lett.* 1987; **12**(9): 643–645.
28. Smith WE, Barrett HH. Hotelling trace criterion as a figure of merit for the optimization of imaging systems. *JOSA A* 1986; **3**(5): 717–725.
29. Eilers PHC, Boelens HFM. Baseline correction with asymmetrical least squares smoothing. http://www.science.uva.nl/~hboelens [19 June 2014].
30. Gemperline PJ. *Practical Guide to Chemometrics* (2nd edn). CRC Press: Boca Raton, FL, 2006.
31. Massart DL, Vandeginste BGM, Deming SN, Michotte Y, Kaufman L. *Chemometrics: A Textbook*. Elsevier: New York, NY, 1988.
32. Johnson RA, Wichern DW. *Applied Multivariate Statistical Analysis*. Prentice-Hall: Englewood Cliffs, New Jersey, 1982.
33. Lay DC. *Linear Algebra and Its Applications* (4th edn). Addison Wesley: Boston, MA, 2003.
34. Goding JC, Ragon DY, OConnor JB, Boehm SJ, Hupp AM. Comparison of GC stationary phases for the separation of fatty acid methyl esters in biodiesel fuels. *Anal. Bioanal. Chem.* 2013; **405**: 6087–6094.
35. Jackson JE. *A User's Guide to Principal Components*. John Wiley and Sons: New York, NY, 1991.
36. Schwarz D. Confellipse2.m. http://www.mathworks.com/matlabcentral/answers/ [19 June 2014].
37. Eaton ML, Perlman MD. The non-singularity of generalized sample covariance matrices. *Ann. Stats.* 1973; **1**: 710–717.
38. Barrett HH, Gooley T, Girodias K, Rolland J, White T, Jao J. Linear discriminants and image quality. *Image Vision Comput.* 1992; **10**(6): 451–460.