Optimization of Chromatogram Alignment
Using a Class Separability Criterion

Gopal R. Yalla

Department of Mathematics and Computer Science
College of the Holy Cross
SUMMARY

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 statistical distance measure that accounts for within-class and between-class variation. We chose the correlation optimized warping (COW) 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 analysis (PCA). 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 the HTC were then determined. PC plots containing 95% confidence ellipses were created and analyzed.

The results demonstrated that the parameters derived from maximizing the HTC more effectively aligned the data than warping effect, 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.

KEYWORDS: Hotelling trace criterion, correlation optimized warping, principal components analysis, chromatogram alignment, biodiesel.
1 Introduction

1.1 Gas Chromatography and Mass Spectrometry

The gas chromatograph and mass spectrometer are two independent instruments which, when combined, create a powerful analytic technique for separating and identifying the components of complex mixtures. Even samples containing 300 compounds at amounts of $10^{-12}$ g can be separated using the relatively simple physical principles behind gas chromatography (GC) and mass spectrometry (MS).

The gas chromatograph, introduced in 1952, is the main instrument used for separating individual components of a mixture. A diagram of a chromatograph is shown below.

![Diagram of a typical chromatograph](image)

Figure 1: A schematic of a typical chromatograph [1].

GC involves the partitioning of a compound between two different phases—one mobile and one stationary. A sample is injected into the GC column
along with a carrier gas, which allows the sample to travel through the column. This denotes the *mobile phase*. Usually helium, hydrogen, or nitrogen gas is used as a carrier. The columns are packed with liquid-coated particles, which may be chosen from over one hundred different substances. As a mixture is carried along by the mobile phase, it travels past a fixed film of these particles, which denotes the *stationary phase*. The various compounds in a mixture partition to a different degree between these two phases depending on their respective solubility. Compounds are retarded to different extents, depending on their solubilities and degree of polarity, and will become separated physically. Higher solubility correlates positively with higher retention time in the stationary phase. The more selective the interactions, the greater the separation [2].

Eventually components emerge from the stationary bed one-by-one into the mass spectrometer, which is insensitive to the carrier gas and helps identify the amount and type of chemicals present in a sample. A diagram of a mass spectrometer typically used in GC is shown below.

![Diagram of a typical mass spectrometer used in combination with GC](image)

Figure 2: A schematic of a typical mass spectrometer used in combination with GC [3].

As components enter the mass spectrometer, their retention time in the GC column is recorded and they become ionized by a stream of electrons in order to break them down into their constituent fragments. Ions are then separated according to their mass by a grate-like device, which creates
definite patterns of the number of ions present at each mass level. This pattern, also known as a mass spectrum, is unique to a given compound and can therefore be used to identify compounds [2]. The mass spectrum is then scaled up by using basic amplification techniques. When the process is complete, the mass spectrometer produces a chromatogram, which displays chemical abundance versus retention time. A sample chromatogram is shown below.

![Sample Chromatogram](image)

Figure 3: An example of a sample chromatogram.

The chromatogram contains the analytical data for the components of a mixture. Qualitative information is contained on the $x$-axis, where retention time in the column can indicate chemical identity. The quantitative information, displayed along the $y$-axis, provides information regarding abundance of each molecular ion. The achievements of GC-MS instrumentation are striking and widely used in the field of analytic chemistry and chemometrics. As impressive as the GC-MS instrumentation is, much data preprocessing is required before any analysis can be performed on its results.
1.2 Introduction to Alignment

Complex chromatographic data can be challenging to analyze based on the sheer number of chemical components in a sample. When dealing with multiple samples especially, the entire data set must be preprocessed before chemometric techniques such as principal component analysis (PCA) can be used to determine interesting trends in the data. The need for this preprocessing comes from natural variations that occur between different runs on the chromatograph. Fluctuations in the chromatogram occur in both peak height—due to variation in the manually injected volume—and retention time—due to slight differences in oven temperature, analyte interaction on the column, starting time of the procedure, etc. Some researchers choose to extract retention time peak areas in order to combat fluctuations between samples. While this method is straightforward, a judgment of the number and type of chemical components must be made by the user, which may sacrifice interesting results during PCA. Another strategy is to use the entire raw data set, so there is no *a priori* judgment of what components should be studied. This method requires more sophisticated data processing, as every sample in the set must be aligned prior to any subsequent chemometric analysis. Without retention time alignment, the trends determined by 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 [4] 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 a local constraint on the number of one-direction consecutive moves that can be made. Vest Nielsen et al. [5] introduced a correlation optimized warping (COW) method, which uses piece-wise 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. [6] 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 [7, 8, 9] 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 the natural question arises regarding the selection of the “best” set of parameters that will produce optimal data alignment. Several performance metrics have been proposed to measure the quality of alignment. Pierce et al. [6] define a measure of alignment quality called degree of class separation, which is the ratio of the Euclidean distance between two 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 [10] 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. [11] 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. Since the data set may contain different types of samples, we may want to preserve these differences post-alignment and the warping effect measure may not serve to value this difference preservation.

In this work, we investigate the use of the Hotelling trace criterion (HTC) [12] as a metric to determine the parameters that serve to optimally align GC data. The HTC is an omnibus measure of class separability [13] that incorporates both within-class and between-class variation and is the multi-class extension of the Mahalanobis distance [14]. Researchers have previously used the HTC as a quality metric for feature enhancement in image processing [15] and imaging system optimization [16].

We evaluate 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 GC-MS. The COW algorithm is employed as an alignment tool for the data. We compare the effectiveness of
the HTC as a figure of merit to the warping effect metric of Skov et al. [11].

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. $a$), uppercase bold for matrices (i.e. $A$), and superscript $t$ to denote matrix/vector transpose. Data matrices are also denoted by uppercase bold (i.e. $X$), 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 $k$th 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 $n$th sample chromatogram that belongs to the $k$th class, while the vector $x_{kn}$ is vector of measurements of the $n$th chromatogram from the $k$th 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 $x_{kn}^{(Q)}$.

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. As shown in the Figure 4, the baseline shape of each chromatogram in our study exhibited a non-linear increase as a function of retention time. This increase may be caused from a variety of factors including a gradual increase in oven temperature over time or even slight bleeding of the column.

We employed a variation of the baseline correction method of Eilers and
Boelens [17] to correct for this curvature. The method uses asymmetric least squares (AsLS) smoothing to determine a baseline vector $\mathbf{b}'$ that minimizes

$$f(\mathbf{b}') = \| \mathbf{w}' (\mathbf{b}' - \mathbf{x}_{kn}) \|^2 + \lambda \| \mathbf{D}\mathbf{b}' \|^2,$$

where $\| \cdot \|$ is the Euclidean norm, $\mathbf{w}$ is a vector of weights, $\lambda$ 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 $\lambda$ controls the relative importance of these two properties: larger $\lambda$ results in a smoother baseline; smaller $\lambda$ 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, since peaks have large deviations from the baseline. To identify peaks, let us
assume
\[ x_{kn} = s + b + \epsilon, \]
where \( s \) is the non-random true peak height, \( b \) is the true non-random baseline to be estimated, and \( \epsilon \) denotes the random error. Furthermore, we assume that \( s \) is sparse (i.e. usually 0) with narrow, large deviation peaks of a fixed maximum width, \( b \) is smooth, and each component of \( \epsilon \) is normally distributed with a small standard deviation \( \sigma_\epsilon \).

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

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 \( s \). 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 \( b' \), an estimate for the true baseline \( b \). Subtracting the baseline yields a baseline-corrected chromatogram \( x_{kn}^{(BC)} \)
\[ x_{kn}^{(BC)} = x_{kn} - b' \approx s + \epsilon. \]

Applying this method to the same sample chromatogram shown in Figure 4, produces the baseline corrected chromatogram in Figure 5.

2.3 Correlation Optimized Warping (COW)

Full chromatograms need to be aligned prior to chemometric analysis, as even small shifts in chromatographic profiles with respect to retention time can cause severe variations in chemometric analyses [18]. The COW algorithm was introduced by Nielsen et al. [5] as a method to correct for these small shifts in discrete data signals, and was used to align our data. COW uses dynamic programming to piecewise align a sample chromatogram towards a reference chromatogram by stretching or compressing segments of the sample using linear interpolation [11].

The choice of a reference sample is important, as it serves as the basis of alignment for all other samples. Skov et al. [11] 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) in order to determine the optimal chromatogram to be used as a reference. The SI is based on the product of the correlation coefficients between all individual chromatograms. This figure of merit is intuitive—the chromatogram with the highest correlation with all other chromatograms will be chosen as a reference. To compute SI for the $j$th baseline-corrected chromatogram in the $k_0$th class, $x^{(BC)}_{k_0j}$, 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,

$$SI_j = \prod_{n=1}^{N} |r(x^{(BC)}_{k_0j}, x^{(BC)}_{kn})| .$$

Once the target chromatogram has been selected, we must establish how to align other sample chromatograms to this target. In order to simplify the following explanation, let’s abandon our previous notation for a moment, and consider two chromatographic profiles, each of length $L$, to be aligned—the target ($T$) and another profile ($P$)—which leads to an aligned profile ($P'$).
The number of sections, $S$, to be warped is simply,

$$S = \frac{L}{m}$$

where $m$ is a given section length. Each section may be warped to smaller or greater length by linear interpolation. As described in [5], a section having starting point $x_s$ and end point $x_e$ in $P$ is warped to a starting position $x'_s$ and end position $x'_e$ in $P'$ by calculating points,

$$p_j = \frac{j}{x'_e - x'_s}(x_e - x_s) + x_s; \quad j = 0, \ldots, x'_e - x'_s$$

and then calculating the value of $P'(x'_s + j)$ by linear interpolation between points in $P$ adjacent to $p_j$. Each section can only be warped by a number between 0 and $t$ where $t$ is a given max warp value that is uniform for each section. In order to determine the optimal $P'$, each warping value must be tried on each segment; however, the warping of one segment directly effects the warping of its neighboring segments. Therefore, we need an efficient way to test each warping value on each segment.

Dynamic programming solves combinatorial optimization problems such as this. Essentially, this style of programming is a way to test all possibilities without testing any individual more than once. The COW algorithm uses a variant of dynamic programming, known as backward dynamic programming. The optimal combination of warping—the one that gives the largest value of summed correlation coefficients—defines the optimal set of node positions, $x_i$, after warping, so we have

$$x_0 = 0 < x_1 < \cdots < x_{S-1} < x_S = L$$

$$u_i \in [-t, t]; \quad i = 0, \ldots, S - 1$$

$$x_{i+1} = x_i + m + u_i; \quad i = 0, \ldots, S - 1$$

where $u_i$ is the warping value for the $i$th segment. The algorithm is based on the use of two matrices, $F$ and $U$ of size $(S + 1) \times (L + 1)$, that contain correlation values for each node placement and the corresponding warping values, respectively. To clarify, the entry $(i, j)$ in the matrix $F$ contains the highest correlation value if a given node $x_i$ were to be placed at position $j$, \[11\]
and the matrix $U$ stores the optimal warping value for node $x_i$ to end up at position $j$.

Consider the following example. Suppose $L = 40$, $m = 10$, and $t = 5$. The structure of $F$, and $U$ are shown in Figure 6. Each matrix has a row for each node, and a column for each possible node position. For instance, consider entry $(2, 29)$ in the matrix. Node 2 may be placed at position 29 in two ways: if node 3 was at position 35 ($u_3 = -5$) in which case $u_2 = -4$, or if node 3 was placed at position 34 ($u_3 = -4$), then $u_2 = -5$. Whatever combination leads to the highest correlation value between segments 2 through 4 is chosen, and $u_2$ is stored in $U_{2,29}$, while the correlation value is stored in $F_{2,29}$. The nice quality of dynamic programming is that examining positions for node 1, will lead to positions that were examined for node 2, so the optimal combination of node placings for the rest of the nodes has already been determined.

The optimal sequence of warpings may then be reconstructed by back-tracking through the matrix $U$. The pseudo-code for the COW algorithm is listed in Appendix A of Vest Nielsen, *et al.* [5]

### 2.4 Data Transformation

After baseline correction and alignment, each chromatogram $x_{kn}^{(BC,COW)}$ 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 $n$th chromatogram in the $k$th class

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

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

\[ \bar{A} = \frac{1}{N} \sum_{k=1}^{K} \sum_{n=1}^{N_k} A_{kn}. \]  

(5)

Each component of a given chromatogram was subsequently divided by its total area \( A_{kn} \) so that each normalized chromatogram had 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

\[ x_{kn}^{(\text{BC,COW,NORM})} = \frac{\bar{A}}{A_{kn}} \cdot x_{kn}^{(\text{BC,COW})}. \]  

(6)

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 [19]. After area normalization and scaling, the sample mean chromatogram is computed

\[ \bar{x}^{(\text{BC,COW,NORM})} = \frac{1}{N} \sum_{k=1}^{K} \sum_{n=1}^{N_k} x_{kn}^{(\text{BC,COW,NORM})}, \]  

(7)

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

\[ x_{kn}^{(\text{BC,COW,NORM,MC})} = x_{kn}^{(\text{BC,COW,NORM})} - \bar{x}^{(\text{BC,COW,NORM})}. \]  

(8)

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 [14] 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 principal components (PCs). More importantly, these PCs still retain most of the variation in the original data set [19, 20]. 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 \( S \) represent the sample covariance matrix of the entire set of processed sample chromatograms with eigen-decomposition \([14, 21, 22]\) given by

\[
S = U\Lambda U^t, \tag{9}
\]

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

\[
y_{kn} = U^t x_{kn}^{(BC,COW,NORM,MC)}. \tag{10}
\]

Each PC is a linear combination of the original measurements. Furthermore, only the first few elements of \( y_{kn} \) 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. \([11]\) have defined a figure of merit to quantify this similarity called \textit{warping effect}, which is the sum of two quantities: \textit{simplicity} and \textit{peak factor}. Simplicity is related to the rank of the data matrix for the aligned, baseline-corrected chromatograms. A data matrix with rank one 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 \([11]\)
simplicity = \sum_{r=1}^{R} \left( \text{SVD} \left( \frac{X}{\sqrt{\sum_{k=1}^{K} \sum_{n=1}^{N_k} \sum_{m=1}^{M} x_{knm}^2}} \right) \right)^4 . \quad (11)

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

\[ c_{kn} = \left| \frac{\| x_{kn}^{(BC,COW)} \| - \| x_{kn}^{(BC)} \|}{\| x_{kn}^{(BC)} \|} \right| \quad (12) \]

as the relative error between a baseline-corrected chromatogram before alignment and after alignment, then peak factor can be computed as [11]

\[ \text{peak factor} = \frac{1}{N} \sum_{k=1}^{K} \sum_{n=1}^{N_k} (1 - \min(c_{kn}, 1)^2) . \quad (13) \]

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 contribute a \( 1/N \) to the sum. Thus, better alignment corresponds to larger values for simplicity and peak factor, and consequently for warping effect.

### 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 locations along the time axis, but retain variation in peak height. Simplicity is a global measure of similarity between all of the samples which does not quantify class separability, and so maximizing it might not serve to identify the segment length and max warp parameters that best retrain these differences. Ideally, we would like to use a measure that reflects our ability to discriminate between the different classes of biodiesels. Therefore, we will look to identify an appropriate statistical metric that measures
the separability between groups of multivariate measurements.

The Hotelling trace criterion (HTC) is a statistical class separability metric, which incorporates both within-class and between-class variation in a data set. We will use it to optimize the COW parameters, segment length and max warp, that will result in the retention of difference between classes of chromatograms.

Let \( \mathbf{x}_{kn} \) denote the \( p \times 1 \) vector corresponding to the \( n \)th chromatogram of length \( p \) in the \( k \)th class. The sample mean vector \( \bar{\mathbf{x}}_k \) and sample covariance matrix \( \mathbf{S}_k \) for the \( k \)th class are given respectively by

\[
\bar{\mathbf{x}}_k = \frac{1}{N_k} \sum_{n=1}^{N_k} \mathbf{x}_{kn},
\]

and

\[
\mathbf{S}_k = \frac{1}{N_k - 1} \sum_{n=1}^{N_k} (\mathbf{x}_{kn} - \bar{\mathbf{x}}_k)(\mathbf{x}_{kn} - \bar{\mathbf{x}}_k)^t.
\]

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

\[
\bar{\mathbf{x}} = \sum_{k=1}^{K} P_k \bar{\mathbf{x}}_k,
\]

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

\[
\mathbf{S}_{wc} = \sum_{k=1}^{K} P_k \mathbf{S}_k,
\]

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

\[
\mathbf{S}_{bc} = \sum_{k=1}^{K} P_k (\bar{\mathbf{x}}_k - \bar{\mathbf{x}})(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})^t.
\]

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
where \( \text{tr}(\cdot) \) 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.

### 2.5.3 Mahalanobis Distance

When only two classes are present \((k = 2)\), the HTC reduces to the Mahalanobis distance. Using the previous notation, the sample means are given by

\[
\begin{align*}
\bar{x}_1 &= \frac{1}{N_1} \sum_{n=1}^{N_1} x_{1n} \\
\bar{x}_2 &= \frac{1}{N_2} \sum_{n=1}^{N_2} x_{2n}
\end{align*}
\]  
(20)  

and the sample covariance matrices are

\[
\begin{align*}
S_1 &= \frac{1}{N_1 - 1} \sum_{n=1}^{N_1} (x_{1n} - \bar{x}_1)(x_{1n} - \bar{x}_1)^t \\
S_2 &= \frac{1}{N_2 - 1} \sum_{n=1}^{N_2} (x_{2n} - \bar{x}_2)(x_{2n} - \bar{x}_2)^t.
\end{align*}
\]  
(22)  

Assuming equal probability of the occurrence of each class \((P_1 = P_2 = \frac{1}{2})\), we define the grand mean vector of the two classes to be

\[
\bar{\mathbf{x}} = \frac{1}{2} \bar{x}_1 + \frac{1}{2} \bar{x}_2.
\]  
(24)
\[ S_{wc} = \frac{1}{2} S_1 + \frac{1}{2} S_2 = S , \]  

and the between-class scatter matrix is

\[
S_{bc} = \frac{1}{2} (\bar{x}_1 - \bar{x})(\bar{x}_1 - \bar{x})' + \frac{1}{2} (\bar{x}_2 - \bar{x})(\bar{x}_2 - \bar{x})'
= (\bar{x}_1 - \bar{x}_2)(\bar{x}_1 - \bar{x}_2)' .
\]  

Substituting these quantities into the formula for HTC gives

\[
J = tr (S_{wc} S_{bc})
= tr \left( S^{-1}(\bar{x}_1 - \bar{x}_2)(\bar{x}_1 - \bar{x}_2)' \right)
= tr \left( (\bar{x}_1 - \bar{x}_2)' S^{-1}(\bar{x}_1 - \bar{x}_2) \right)
= (\bar{x}_1 - \bar{x}_2)' S^{-1}(\bar{x}_1 - \bar{x}_2) ,
\]

which is the definition of the sample Mahalanobis distance.

2.5.4 The t-statistic

In the two class case where each sample contains only one retention time measurement, the Mahalanobis distance further reduces to the square of a t-statistic. Let \( x_{kn} \) be the \( n \)th sample measurement from the \( k \)th class. For simplicity assume \( N_1 = N_2 = N \), and similarly that the two populations have a common but unknown variance \( \sigma_1^2 = \sigma_2^2 = \sigma^2 \). As in the previous section, we define the means of the two samples to be

\[
\bar{x}_1 = \frac{1}{N} \sum_{n=1}^{N} x_{1n} \]  
\[
\bar{x}_2 = \frac{1}{N} \sum_{n=1}^{N} x_{2n} .
\]

The pooled sample variancne \( S_p^2 \), which estimates the common variance \( \sigma^2 \), is
defined as

\[
S_p^2 = \frac{(N - 1) \sum_{n=1}^{N} (x_{1n} - \bar{x}_1)^2 + (N - 1) \sum_{n=1}^{N} (x_{2n} - \bar{x}_2)^2}{2N - 2}
= \frac{(S_1^2 + S_2^2)}{2},
\]

(30)

where \( S_k^2 \) is the sample variance for the \( k \)th class, \( k = 1, 2 \). Using the definition for Mahalanobis distance, we have

\[
J = (\bar{x}_1 - \bar{x}_2)(S_p^2)^{-1}(\bar{x}_1 - \bar{x}_2)
= (\bar{x}_1 - \bar{x}_2)^2 \frac{S_p^2}{S_p^2}
= t^2 \left( \frac{2}{N} \right),
\]

(31)

where \( t \) is the \( t \)-statistic under equality of variances with \( 2N - 2 \) degrees of freedom.

### 2.6 Evaluation of HTC Using Principal Components

In Section 2.4 we discussed how one would transform each chromatogram into its PC scores (Equation (10)). We will now discuss the evaluation of the HTC using the PC scores in place of the original chromatogram measurements.

Recall that the PC-transformed chromatograms are denoted by the \( p \times 1 \) vectors \( y_{kn} \), where \( k \) is the class index and \( n \) is the sample chromatogram index. Let \( z_{kn} = (y_{k1n}, y_{kn2}, \ldots, y_{kLn})^t \) denote the \( L \times 1 \) vector corresponding to the first \( L \) PCs of \( y_{kn} \). The sample mean vector \( \bar{z}_k \) and sample covariance matrix \( S_k \) for the \( k \)th class are given respectively by

\[
\bar{z}_k = \frac{1}{N_k} \sum_{n=1}^{N_k} z_{kn},
\]

(32)

and
\[ S_k = \frac{1}{N_k - 1} \sum_{n=1}^{N_k} (z_{kn} - \bar{z}_k)(z_{kn} - \bar{z}_k)^t. \] (33)

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

\[ \bar{\bar{z}} = \sum_{k=1}^{K} P_k \bar{z}_k, \] (34)

where \( P_k = N_k/N \) is the probability of occurrence for class \( k \). Using these quantities, we compute the within-class scatter matrix \( S_{wc} \) as

\[ S_{wc} = \sum_{k=1}^{K} P_k S_k, \] (35)

and the between-class scatter matrix \( S_{bc} \) as

\[ S_{bc} = \sum_{k=1}^{K} P_k (\bar{z}_k - \bar{\bar{z}})(\bar{z}_k - \bar{\bar{z}})^t, \] (36)

As previously stated, the matrix \( S_{wc} \) quantifies the average multi-dimensional dispersion within each class about the class mean, while \( S_{bc} \) quantifies the average multi-dimensional dispersion between each class mean and the grand mean. The HTC is then computed as

\[ J = \text{tr}(S_{wc}^{-1}S_{bc}). \] (37)

Thus, we seek to use this implementation of the HTC as our optimization metric. It is important for the reader to note that this computation of the HTC is dependent on the number of PCs \( L \) that we include in \( z_{kn} \). 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 United States (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), NIST (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 stored in their original shipping container at 4 °C. Prior to dilution, each biodiesel was gradually warmed to room temperature and 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) and 1 mL of 0.30 M tridecanoic acid methyl ester (Fluka) was added to a 50 mL volumetric flask and 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 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) and have been presented in detail previously [23]. The GCMS 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 FAMEs in a 37 component FAME standard (Supelco) and is 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 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 (NIST mass spectral search program version 2.0a, Gaithersburg, MD) as well as retention time
comparison to the FAME standard. A representative chromatogram showing separation of FAME components is shown in Figure 7.

![MOLECULAR ABUNDANCE VS RETENTION TIME](image)

**Figure 7:** Sample chromatograms for biodiesel fuels produced from different feedstocks.

### 3.3 Data Processing

Total ion chromatograms were extracted from Chemstation using a macro developed by Infometrix (Bothell, WA). All chromatograms were baseline corrected using the method previously described (and implemented in python) 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 minutes and 40.37 - 48.92 minutes) were removed prior to further chemometric analysis.

Next, the chromatograms were aligned using the COW algorithm under the same combinations of segment length-max warp as seen in Skov et al. [11]. 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. The PC transform was then computed for each data set and applied to each chromatogram to generate the corresponding PC scores. 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. Since each data file corresponded to COW processing with a particular combination of segment length-max warp, we arranged the values of warping effect into a 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 (6 different samples), canola (3 different samples), tallow (3 different samples), waste grease (2 different samples), and hybrid (1 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.

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 8. We also found the maximum HTC value as a function of $L$ and determined the combinations of segment length-max warp that produced them. It should be noted that these combinations changed with $L$. These results can be seen in Table 1.

Two-dimensional (2D) 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
Figure 8: A plot of median cumulative percent total variation versus eigenvalue index, along with 95% confidence bounds.

in [24] and implemented by Schwarz [25]. These were included on the scores plots.

4 Results

The results of our investigation into the optimization of the COW algorithm parameters can be seen in Figures 9-13 and Tables 2-3. Figure 9 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 10. Two-dimensional scores plots of pair-combinations of the PCs, along with corresponding 95% confidence ellipses are displayed in Figure 11 (PC2 vs PC1), Figure 12 (PC3 vs PC1), and Figure 13 (PC3 vs PC2), for the four groupings of segment length-max warp discussed in the Experimental Methods section. Table 2 lists the Euclidean distances between each pair of class means, while Table 3 lists the ratios of the standard deviations along the principal axes of each class, where the numerator is the class standard devia-
Table 1: Segment length and max warp at the maximum HTC value as a function of the number of 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.

<table>
<thead>
<tr>
<th>No. of PCs ((L))</th>
<th>segment length</th>
<th>max warp</th>
<th>max HTC</th>
<th>warp effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>64</td>
<td>3</td>
<td>143.5</td>
<td>1.65</td>
</tr>
<tr>
<td>2</td>
<td>55</td>
<td>8</td>
<td>244.2</td>
<td>1.69</td>
</tr>
<tr>
<td>3</td>
<td>70</td>
<td>6</td>
<td>298.6</td>
<td>1.69</td>
</tr>
</tbody>
</table>

Considering Figure 9, 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 10, the corresponding 2D density plots for 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 1 PC, (55,8) using 2 PCs, and (70,6) using 3 PCs. Two important observations can be noted. The first is that the HTC values exhibit greater variation in magnitude as compared to 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.
Figure 9: 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).
Figure 10: 2D density plots of HTC. Maximum values occurred for segment length-max warp combinations of (64,3) using 1 PC (top), (55,8) using 2 PCs (middle), and (70,6) using 3 PCs (bottom).
We now turn our discussion to the PC scores plots. Recalling Figure 8, 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 11 (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, 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.

Confidence ellipses were not calculated for the hybrid class. This is due to 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 [26]. The eigenvectors derived from the diagonalization of this covariance matrix are used to determine a confidence ellipse.

Considering the scores plots of PC3 vs PC1 in Figure 12, 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 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 13. 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.
Figure 11: Scatter plots of PC2 vs 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 (+).
Figure 12: Scatter plots of PC3 vs 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 (+).
Figure 13: Scatter plots of PC3 vs PC2 for combinations of segment length-max warp (26,15) (top left), (64,3)(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 (+).
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 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 orientation of the axes are not incorporated into this quantity. The results can be seen in Tables 2 and 3.

Examining Table 2, 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 due to 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 3, 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 2 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 variation. Thus, we would not expect within-class variation to be systematically smaller when the HTC is at a maximum.
Table 2: Euclidean distances between pairs of class means for data in scores plots comparing PC2 versus PC1. All values should be scaled by $10^6$.

### Segment Length/Max Warp (26,15)

<table>
<thead>
<tr>
<th>Class</th>
<th>Soy</th>
<th>Canola</th>
<th>Tallow</th>
<th>Waste Grease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soy</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Canola</td>
<td>9.49</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tallow</td>
<td>10.74</td>
<td>8.91</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>Waste Grease</td>
<td>2.76</td>
<td>6.88</td>
<td>8.64</td>
<td>0</td>
</tr>
</tbody>
</table>

### Segment Length/Max Warp (64,3)

<table>
<thead>
<tr>
<th>Class</th>
<th>Soy</th>
<th>Canola</th>
<th>Tallow</th>
<th>Waste Grease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soy</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Canola</td>
<td>11.66</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tallow</td>
<td>11.87</td>
<td>9.95</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>Waste Grease</td>
<td>3.08</td>
<td>8.69</td>
<td>9.58</td>
<td>0</td>
</tr>
</tbody>
</table>

### Segment Length/Max Warp (55,8)

<table>
<thead>
<tr>
<th>Class</th>
<th>Soy</th>
<th>Canola</th>
<th>Tallow</th>
<th>Waste Grease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soy</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Canola</td>
<td>11.24</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tallow</td>
<td>12.11</td>
<td>9.69</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>Waste Grease</td>
<td>3.35</td>
<td>7.98</td>
<td>9.68</td>
<td>0</td>
</tr>
</tbody>
</table>

### Segment Length/Max Warp (70,6)

<table>
<thead>
<tr>
<th>Class</th>
<th>Soy</th>
<th>Canola</th>
<th>Tallow</th>
<th>Waste Grease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soy</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Canola</td>
<td>11.40</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tallow</td>
<td>11.80</td>
<td>9.71</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>Waste Grease</td>
<td>3.16</td>
<td>8.33</td>
<td>9.44</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 3: Ratios of standard deviations between corresponding principal axes for data derived from maximizing HTC versus data derived from maximizing warping effect.

<table>
<thead>
<tr>
<th>Class</th>
<th>1st Major Axis</th>
<th>2nd Major Axis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soy</td>
<td>1.36</td>
<td>1.02</td>
</tr>
<tr>
<td>Canola</td>
<td>2.46</td>
<td>0.97</td>
</tr>
<tr>
<td>Tallow</td>
<td>1.03</td>
<td>2.15</td>
</tr>
<tr>
<td>Waste Grease</td>
<td>0.74</td>
<td>0.47</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
<th>1st Major Axis</th>
<th>2nd Major Axis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soy</td>
<td>0.94</td>
<td>0.92</td>
</tr>
<tr>
<td>Canola</td>
<td>1.06</td>
<td>0.80</td>
</tr>
<tr>
<td>Tallow</td>
<td>0.86</td>
<td>1.30</td>
</tr>
<tr>
<td>Waste Grease</td>
<td>0.68</td>
<td>0.68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
<th>1st Major Axis</th>
<th>2nd Major Axis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soy</td>
<td>1.14</td>
<td>0.71</td>
</tr>
<tr>
<td>Canola</td>
<td>1.10</td>
<td>0.69</td>
</tr>
<tr>
<td>Tallow</td>
<td>0.86</td>
<td>1.30</td>
</tr>
<tr>
<td>Waste Grease</td>
<td>0.49</td>
<td>0.69</td>
</tr>
</tbody>
</table>
5 Conclusion

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 Hotelling trace criterion (HTC). In addition, we compared our results with those derived from maximizing the warping effect figure of merit of Skov et al. [11]. 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 [13, 27] 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.

References


38


