Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 68(2) 2014 181-190

A Novel Metric Distance on Registered Curves with Application to a Fourier Transform-infrared Spectroscopy Analysis of Maize

Yishi Wang[1], Susan J. Simmons[1], Latasha L. Smith[2] and Ann E. Stapleton[2]
[1]*Department of Mathematics and Statistics, University of North Carolina Wilmington, Wilmington, NC*
[2]*Department of Biology and Marine Biology, University of North Carolina Wilmington, Wilmington, NC*

Received 11 October 2013; Revised 31 January 2014; Accepted 10 February 2014

SUMMARY

Registered curves containing a variety of information are becoming more and more frequent in natural sciences research. To date, most statistical analysis of such curves involves using only a portion of the information contained within these curves. In order to utilize information across the entire spectrum of the curve, we propose to consider shape and/or magnitude distance that measures the similarity of these non-smooth functional curves, and an object function to compare the effectiveness of different distances measures. Once a similarity/dissimilarity matrix is obtained, various statistical properties can be ascertained about the relationship between two or more curves. Herein, we develop an approach that can identify the most effective distance measure and apply it to an analysis of maize seed Fourier transform-infrared spectroscopy (FT-IR) spectral data. Dimension reduction techniques, such as multi-dimensional scaling (MDS), is then applied to represent the original curves in a lower dimensional space.

Keywords: Functional curve, Dimension reduction, Multidimensional scaling, Procrustes distance.

1. INTRODUCTION

The biological composition of maize and other crops provides essential information to breeders and researchers regarding inherent properties and attributes of different varieties. This information assists breeders and researchers in screening crops for advantageous seed composition traits, identifying potential crop contamination and understanding effects of environmental stress (Baye *et al.* 2006). Due to recent advancements in technology, this knowledge can be acquired rapidly and non-destructively by using FT-IR. This methodology provides accurate, valuable data without destroying the kernel or crop (Thygesen *et al.* 2003) and has been used quite frequently to obtain measurements of moisture, protein oil and starch (Stermer *et al.* 1977 and Cook *et al.* 2012). The National Institute of Standards and Technology (NIST) has developed standards for instrument design and calibration for infrared spectroscopy (http://www.nist.gov/pml/wmd/pubs/upload/5-57-09-HB44-FINAL.pdf).

The type of data produced by FT-IR and the analysis of this data is similar to other fields. For example, Sujatha *et al.* (2008) assessed use of protein profiles created by HPLC-LIF to identify cervical cancer. The authors use principal component analysis (PCA) to reduce the dimensionality of the curves and classify each profile into malignant or not malignant. Lund and Li (2009) defined a new distance measure that incorporates seasonal and autocorrelation in a time

Corresponding author: Yishi Wang
E-mail address: wangy@uncw.edu

series from climate related data from different stations. The information derived from the distance measure can be used in a cluster analysis to define climate zones. Flint-Garcia *et al*. (2009) compared maize inbred lines and landraces with HPLC analysis by first computing the area under the curves, identifying peaks and comparing this information across groups using analysis of variance. Metabolomic experiments produce data curves with similar properties as FT-IR. Sun and Weckwerth (2012) developed a toolbox to analyze this type of data that can perform analysis of variance, principal component analysis, independent component analysis, and clustering and correlation analysis.

Most analysis of registered curves involves one of two methodologies: either (1) the analysis revolves around peak identification or information about the peaks, which disregards the remaining wavelength information, or (2) the distance between the curves is measured and analyzed, which disregards the shape information contained within the curves. Therefore, we propose a new metric to compare these curves using the entire spectral of information including the distances between curves and the shape of the curves. This methodology identifies which distance measure is superior in terms of separating subjects from different groups and provides the best means to understand the differences inherent in the observations. In Section 2, a new shape distance is developed and illustrated on a sample data set. Section 3 develops the objective function which compares the new shape measure to a distance measure to determine which is most optimal. Two simulation examples are included in Section 4, and an application of this methodology to a maize data set is illustrated in Section 5. Conclusions and discussions are in Section 6.

2. METHODOLOGY

Measurements from an FT-IR experiment produce registered curves with wavelength frequency along the *x*-axis and intensity values along the *y*-axis. In other words, the FT-IR curves resemble a time series-type of curve. Commercial FT-IR instruments are calibrated to produce consistent wavelength accuracy on the *x*-axis, so the registration is handled at the point of data collection. In measuring similarities among these curves, we are interested in more than just identifying the number of peaks, frequency of peaks and the location of peaks within the curves, but rather the features of the entire output curves, and find out which distance measure is more effective in distinguishing the groups from each other.

In Fig. 1, we illustrate the FT-IR curves from eleven lines of maize with three replicates in each line, and replicates are shown in the same color. The *x*-axis is based on the fixed inputs (wavelength frequency). Therefore all the outputs are registered across the lines and replicates. In order to investigate the general pattern of the outputs, we begin by fitting cubic splines. However, several issues arise in this analysis. First, the fitted curves are relatively far from the original output. There are many points that are left out of the fitted curves, especially in the vicinity of the two bumps in the middle of the curves and the last bump. This result discourages the consideration of any parametric or nonparametric curve fitting technique that would smooth out information that could potentially be essential in distinguishing curves from one other. Second, the general patterns suggested by the splines are made of smooth sections as well as sections with sharp peaks. Thus techniques with the assumption of smooth curves would not be appropriate. Finally, both shape and magnitude of the curves could be the cause of the differences among FT-IR outputs, which is a question that has not been addressed before.

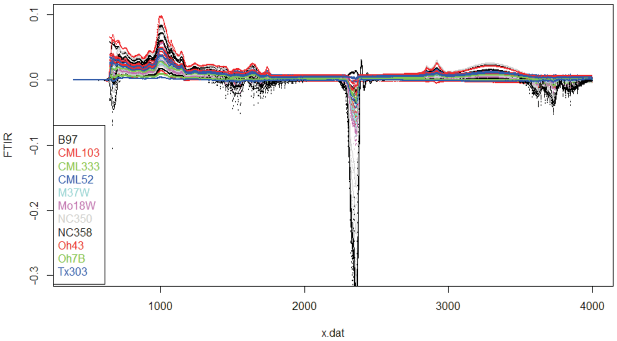

Fig. 1. Data from an FT-IR maize experiment with corresponding fitted curves. Actual data is displayed by points and fitted curves are displayed with colored lines.

Our proposed methodology takes into account information across the entire curve and allows this information to determine the differences among the curves. The methodology we propose is a hybrid measurement including shape and magnitude measures to describe the difference of registered outputs that are not necessarily smooth.

First, we will focus on defining differences in shapes of the registered curves. We will use the information from Fig. 2 as a motivating example in developing our measure for shape similarities/dissimilarities.

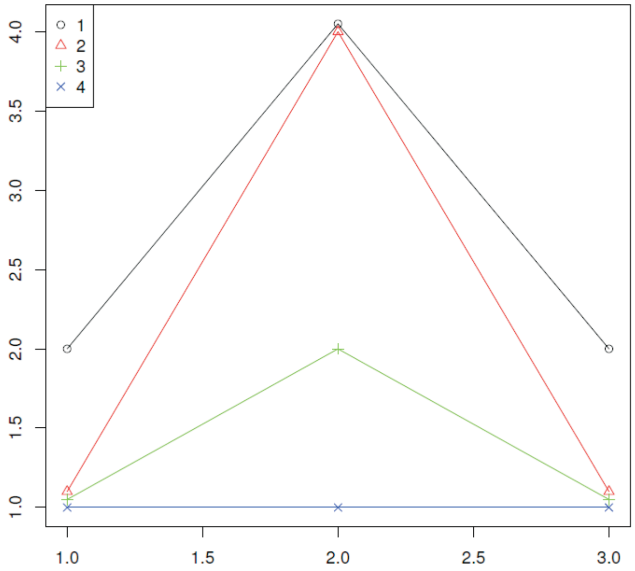

Fig. 2. Four simple shape curves illustrating points S_1, … S_4.

The example is comprised of four different sets of points. Since in this work we assume that all sets are registered, the *x*-values are set to be (1, 2, 3). The *y*-values of the four sets that we have in this example are: $S_1 = (2, 4, 2)^T$, $S_2 = (1 ,4, 1)^T$, $S_3 = (1, 2, 1)^T$ and $S_4 = (1, 1, 1)^T$.

Hausdorff distance, Fŕechet, and procrustes distance (Dryden and Mardia 1998) are ubiquitous measures of distances between the shape of objects and planar curves in 2-d and 3-d images. Unfortunately, due to the nature of the FT-IR curves, procrustes distance is not applicable in this situation. The Fŕechet distance between two curves can be thought of as "the minimum length of a leash required to connect a dog and its owner, constrained on two separate paths, as they walk without backtracking along their respective curves from one endpoint to the other". The Hausdorff distance is the greatest of all the distances from a point in one set to the closest point in the other set. Though both Fŕechet distance and Hausdorff distance are sensitive to the *x*-values, both distances are the same in the above examples with $x = (1, 2, 3)^T$. Table 1 displays the Hausdorff and Fŕechet distance matrix information for this example.

Table 1. Similarity matrix of curves from Figure 2 with Hausdorff distances

	S_1	S_2	S_3	S_4
S_1	0	1	2	3
S_2	1	0	2	3
S_3	2	2	0	1
S_4	3	3	1	0

It is interesting to observe from Table 1 that the distance between S_1 and S_3 is the same as the distance between S_2 and S_3, with a distance of 2, however $S_1 = 2^*S_3$. If we are interested in measuring the magnitude between the two curves, a distance of 2 makes sense, but if we are interested in estimating the shape distance between the two curves the distance should be zero. The drawback of the Fŕechet distance and Hausdorff distance is that they only consider the greatest discrepancy and ignore the patterns contained within the curves, in which case significant amounts of information may lost. Based on this example, we can conclude that the Fŕechet distance and Hausdorff distance measures are not appropriate for measuring the shape distance of registered curves, since the distances depend more on measuring magnitude than on measuring shape.

According to D.G. Kendall (1984), we use the following definition for shape.

Definition 1. Shape is all the geometrical information that remains when location, scale and rotational effects are filtered out from an object.

Since we consider registered curves with realized output at discrete points, each curve may be represented by a vector of the output values. The following distance was first proposed in Wang *et al*. (2013) without much discussion about its properties, and we adopt it here to measure the shape distance among curves.

Definition 2. Let $U = (u_1, ..., u_n)^T$ and $V = (v_1, ..., v_n)^T$ be the registered output values from two subjects. When $||U||_2||V||_2 > 0$, the shape distance between U and V is

$$D_n(U,V)\pi := \sqrt{\frac{(U,V)^2}{\|U\|_2^2\|V\|_2^2}} \qquad (1)$$

where n is the number of output, $< \cdot, \cdot >$ *is the inner product of two vectors, and* $|| \cdot ||_2$ *is the* L_2 *norm.*

By applying the distance to the aforementioned example, the results are contained in Table 2. The

Table 2. Similarity matrix of curves from Fig. 2 with the proposed shape distances from (1)

	S_1	S_2	S_3	S_4
S_1	0	0.27	0	0.33
S_2	0.27	0	0.27	0.58
S_3	0	0.27	0	0.33
S_4	0.33	0.58	0.33	0

distances in Table 2 make more sense when we are interested in estimating the differences due to the shape of the curves. The distance between S_1 and S_3 is now zero, which it should be since these two curves have an identical shape. The largest distance in shape occurs between S_2 (the curve with the highest peak) and S_4 (a flat curve).

Following the example of Fig. 2, the rationale of proposing such a distance in (1) is to compare the shapes of the different curves. If the ratio is a constant, it suggests that the two curves share the "same" shape. Using the previous notation, let $U = (u_1, ..., u_n)^T$ and $V = (v_1, ..., v_n)^T$ be the registered output values. When $||U||_2 \cdot ||V||_2 > 0$, the idea is to minimize the following object function:

$$\tilde{L}(U,V) = \min_r \left\| \frac{U}{\|U\|_2} - r\frac{V}{\|V\|_2} \right\|. \qquad (2)$$

This expression turns out to be a special case of the general procrustes distance proposed in Dryden and Mardia (1998), where the *U's* and *V's* are vectors of complex numbers. Since in this work we only focus on real functions, the advantage of (2) is that only the y-values of the points on the curves are involved, not the x-values, which could potentially save computational time. Besides, all x-values from FT-IR curves are registered, and it makes little sense to re-register them again.

When $||U||_2 \cdot ||V||_2 > 0$, we define $X = (x_1, \ldots, x_n)^T$ and $Y = (y_1, \ldots, y_n)^T$, such that $X = \frac{U}{\|U\|_2}$ and $Y = \frac{V}{\|V\|_2}$. Then (2) becomes:

$$\tilde{L}(X,Y) = \min_r \|X - rY\|_2 . \qquad (3)$$

We might rewrite the second half on the right-hand side of equation (3) as

$$L(r, X, Y) = \|X - rY\|_2 = \sqrt{\Sigma_{i=1}^n (x_i - ry_i)^2}. \qquad (4)$$

The first derivative of $L(r, X, Y)$ with respect to r is

$$\frac{\partial L^2(r, X, Y)}{\partial r} = -2\Sigma_{i=1}^n (x_i - ry_i) y_r \qquad (5)$$

Since $L(r, X, Y)$ is a convex function for r, the minimum exists for given X and Y. By setting equation (5) equal to zero, the solution for r is:

$$\hat{r} = \frac{(X,Y)}{\|Y\|_2^2} = (X,Y). \qquad (6)$$

Hence, the minimizer of (2) is

$$\hat{r} = \frac{(U,V)}{\|U\|_2 \|V\|_2}. \qquad (7)$$

Notice that when $U = V \neq 0$, $r = 1$ and when $U = -V \neq 0$, $r = -1$. The value of r is actually the cosine of the angle between the two vectors U and V. By substituting the solution in (7) back to equation (2), we have

$$\tilde{L}(U,V) = \left\| \frac{U}{\|U\|_2} - \frac{\langle U,V \rangle}{\|U\|_2 \|V\|_2 \|V\|_2} \right\|_2$$

$$= \sqrt{1 - \frac{\langle U,V \rangle^2}{\|U\|_2^2 \|V\|_2^2}}, \qquad (8)$$

and therefore the definition of distance in (1) follows.

It is obvious that the distance defined in (1) is not a metric distance, since $D_n(U, aU) = 0$ for any nonzero scalar a. The following discussion shows that the distance is actually a pseudo-metric distance.

Let c_1, c_2 and c_3 be continuous curves on closed domains, let $X = (x_1, \ldots, x_n)^T$, $Y = (y_1, ..., y_n)^T$ and $Z = (z_1, ..., z_n)^T$ be their corresponding registered heights, with $\| X \|_2 \cdot ||Y||_2 \cdot ||Z||_2 > 0$. It is obvious that $D_n(c_1, c_1) = 0$, and $D_n(c_1, c_2) = D_n(c_2, c_1) \geq 0$. If $D_n(c_1, c_2) = 0$, it indicates that $\Sigma_{i=1}^n |x_i y_i| = ||x||_2||y||_2$. By Cauchy-Schwartz inequality, we have $X = rY$, where r is a scalar. As to the triangle inequality, we refer to the following theorem.

Theorem 3. With notations in the previous paragraph,

$$D_n(c_1, c_2) \leq D_n(c_1, c_3) + D_n(c_2, c_3).$$

Proof: Since X, Y and Z can be seen as vectors in R^n, let α be the angle between X and Y, β be the angle between X and Z, and θ be the angle between Y and Z. Without loss of generality, we assume that α and $\beta \le 90^\circ$, since we can multiply the vector by negative one, and we only consider the squared cosine values in the definition of $D_n(\cdot, \cdot)$. Notice that when α and $\beta \le 90°$, $\theta \le 180°$.

First we are going to show that

$$\sin \alpha \le \sin \beta + \sin \theta. \tag{9}$$

When $\alpha \le \beta$, we have $\sin \alpha \le \sin \beta$, equation (9) follows. When $\beta = 0$, $\alpha = \theta$, the equation also follows. When $\alpha > \beta > 0$, we have

$$\sin(\alpha + \beta) > \sin(\alpha - \beta). \tag{10}$$

Due to the geometry fact that $\theta \le \alpha + \beta < 180°$, and $\theta \ge \alpha - \beta$,

$$\sin(\theta) \ge \sin(\alpha - \beta). \tag{11}$$

With equation (11), equation (9) would follow if we can show that

$$\sin\alpha \le \sin\beta + \sin(\alpha - \beta). \tag{12}$$

By using geometry, the following equations follow:

$$\tan(\alpha/2) > \tan(\beta/2); \tag{13}$$

$$\frac{2\sin^2(\alpha/2)}{2\sin^2(\alpha/2)\cos(\alpha/2)} > \frac{2\sin^2(\beta/2)}{2\sin^2(\beta/2)\cos(\beta/2)}; \tag{14}$$

$$\frac{1-\cos\alpha}{\sin\alpha} > \frac{1-\cos\beta}{\sin\beta}; \tag{15}$$

$$\sin\alpha(1-\cos\beta) \le \sin\beta(1-\cos\alpha); \tag{16}$$

$$\sin\alpha \le \sin\beta + \sin\alpha\cos\beta - \cos\alpha\sin\beta. \tag{17}$$

Thus, equation (12) follows, and therefore equation (9) is proved.

Since α is the angle between vector X and Y, we have $\cos\alpha = \frac{(X, Y)}{\|X\|_2 \|Y\|_2}$. Therefore, based on equation (9), we have

$$\sqrt{1-\cos^2\alpha} \le \sqrt{1-\cos^2\beta} + \sqrt{1-\cos^2\theta}; \tag{18}$$

$$\sqrt{1-\frac{\langle X, Y\rangle^2}{\|X\|_2^2\|Y\|_2^2}} \le \sqrt{1-\frac{\langle X, Z\rangle^2}{\|X\|_2^2\|Z\|_2^2}} + \sqrt{1-\frac{\langle Z, Y\rangle^2}{\|Z\|_2^2\|Y\|_2^2}}. \tag{19}$$

The theorem follows.

In addition to measuring the shape distance, the magnitude distance between curves is also an important feature to capture. Using the metric distance of the L_2 norm for measuring magnitude distances, we propose the following hybrid distance measure between curves:

Definition 4. Let $U = (u_1,\ldots,u_n)^T$ and $V = (v_1, \ldots, v_n)^T$ be the registered output values from two subjects. When $\|U\|_2 \cdot \|V\|_2 > 0$, the hybrid distance between U and V is

$$H_n(U, V) = \lambda D_n(U, V) + (1 - \lambda)\|U - V\|_2, \tag{20}$$

where $0 \le \lambda \le 1$.

It is obvious that $H_n(U, V)$ is a metric distance for $0 \le \lambda \le 1$. When $0 \le \lambda \le 1$, the value of λ can be regarded as a weighting variable for the two distance measures. For $\lambda = 0.50$, equal weight is placed on both the shape distance and the magnitude distance. For $\lambda > 0.50$, more emphasis is placed on the shape distance; and when $\lambda < 0.50$, more emphasis is placed on the magnitude distance.

Once the distance matrix is estimated between registered curves, these distances can be used in various ways to provide a variety of information about the curves. For example, when there are N curves, we end up with $N(N - 1)/2$ distances. With the $N \times N$ distance matrix, we can use dimension reduction techniques such as Multidimensional Scaling (MDS) to represent the distance relationship in a R^d space. For large N and small d, this approach is essentially a new way to project functional curves in lower dimensional space. Hence, analysis of the curves may be reduced to the analysis of their corresponding lower dimensional representations. In summary, the algorithm that we propose, for registered curves is as following: step1: keep all the y-values; step2: find the hybrid distance; step3: use dimension reduction technique to find the lower dimensional representation.

3. CHOICE OF THE PARAMETER λ

The next question is then the choice of the parameter λ. Let M be the total number of types of subjects and K be the number of replicates for each type. We define x_{ij} as the observed vector for the jth replicate from the ith type of subject, for $i = 1, \ldots, M$ and $j = 1, \ldots, K$. We further assume that replicates within the same type of subject are not all identical. Following the definition in (20), we define

$$W_i = \Sigma_{j=1}^{K} \Sigma_{l=1,\, l \neq j}^{K} H_n(X_{ij}, X_{il}), \tag{21}$$

which represents the summation of all distances between the replicates for the i^{th} type of subject. Furthermore, we define

$$B_{ij} = \Sigma_{h=1,\, h \neq i}^{M} \Sigma_{l=1}^{K} H_n(X_{ij}, X_{lh}), \tag{22}$$

which represents the summation of all distances between the ij^{th} vector and every other vector that is not from the i^{th} type of subject. Thus $B_{i+} = \Sigma_{j=1}^{K} B_{ij}$ is the summation of all distances between the vectors from the ith type and all other vectors that are not. Then the objective function for searching the optimal λ is defined as

$$Q(\lambda) = \frac{\Sigma_{i=1}^{M} B_{i+}}{\Sigma_{i=1}^{M} W_i}. \tag{23}$$

The construction of the object function in (23) mimics the test statistics in an analysis of variance setting; in other words, the difference between groups over the differences within groups. If λ_0 = argmax$_\lambda Q(\lambda)$, we can claim that the choice of λ_0 in (20) makes the groups more distinguishable than any other possible value of λ.

Noticing that the function $Q(\cdot)$ is differentiable, finding its derivative becomes easier after the following steps:

$$Q(\lambda) = \frac{\Sigma_{i=1}^{M} B_{i+}}{\Sigma_{i=1}^{M} W_i} = \frac{2\,\Sigma_{i=1}^{M} \Sigma_{h=i+1}^{M} \Sigma_{j=1}^{K} \Sigma_{l=1+1}^{K} H_n(X_{ij}, X_{hl})}{2\,\Sigma_{i=1}^{M} \Sigma_{j=1}^{K-1} \Sigma_{l=1+1}^{K} H_n(X_{ij}, X_{il})}$$

$$= \frac{\Sigma_{i,h,j,k} \| X_{ij} - X_{hl} \|_2 + \lambda \Sigma_{i,h,j,k} (D_n(X_{ij}, X_{hl}) - \| X_{ij} - X_{hl} \|_2)}{\Sigma_{i,j,l} \| X_{ij} - X_{hl} \|_2 + \lambda \Sigma_{i,j,l} (D_n(X_{ij}, X_{hl}) - \| X_{ij} - X_{hl} \|_2)}$$

$$= \frac{a + \lambda b}{c + \lambda d}. \tag{24}$$

where

$$a = \Sigma_{i,h,j,k} \| X_{ij} - X_{hl} \|_2$$

$$b = \Sigma_{i,h,j,k} (D_n(X_{ij}, X_{hl}) - \| X_{ij} - X_{hl} \|_2)$$

$$c = \sum_{i,j,l} \left\| X_{ij} - X_{hl} \right\|_2$$

$$d = \Sigma_{i,j,l} (D_n(X_{ij}, X_{hl}) - \| X_{ij} - X_{hl} \|_2).$$

Following from (24),

$$\frac{\partial Q}{\partial \lambda} = \frac{bc - ad}{(c + \lambda d)^2}. \tag{25}$$

When there are random variations within any type of subjects, $c + \lambda d \neq 0$, the derivative in (26) is valid. In this situation, we may establish the rule of thumb to determine which distance measure is more appropriate for finding the differences among groups.

- When $bc > ad$, $\frac{\partial Q}{\partial \lambda} > 0$, which means that Q is monotone increasing. Therefore $\lambda = 1$ is the optimal parameter, indicating that the shape information in the curves differentiates between the groups better than the L_2 norm does.
- When $bc < ad$, $\frac{\partial Q}{\partial \lambda} < 0$, which means that Q is monotone decreasing. Therefore, $\lambda = 0$ is the optimal parameter indicating that the L_2 norm differentiates between the groups better than the shape information does.
- When $bc = ad$, then the two distances are equivalent in terms of differentiating between the two curves.

4. SIMULATION EXAMPLES

4.1 First Simulation

In the first simulated data set, there are four different groups of curves with dimension 20. Each group has 15 replications. Each of the curves from the four groups can be expressed as a vector $X_{ij} = \alpha_i + \varepsilon_j$, for $i = 1, \ldots, 4$ and $j = 1, \ldots, 15$. The term ε_j creates the randomization within each group and it follows multivariate normal distribution with mean vector $0_{20\times 1}$ and a variance matrix Σ, which is a diagonal matrix with $\sigma_1 = \sigma_{20} = 0$ and $\sigma_2 = \ldots. = \sigma_{19} = 0.2$. The major differences among the four groups are on the α'_i's.

$\alpha_1 = (0, 0.1, 1, 0.5, 0.7, 0.3, 0.1, 0.5, 0.2, 0, \ldots.0)^T$,

$\alpha_2 = (0, 0.1, 1, 0.2, 0.4, 0.6.\ 0.1, 0.5, 0.2, 0, \ldots.0)^T$,

$\alpha_3 = 2\alpha_1$,

$\alpha_4 = 2\alpha_2$

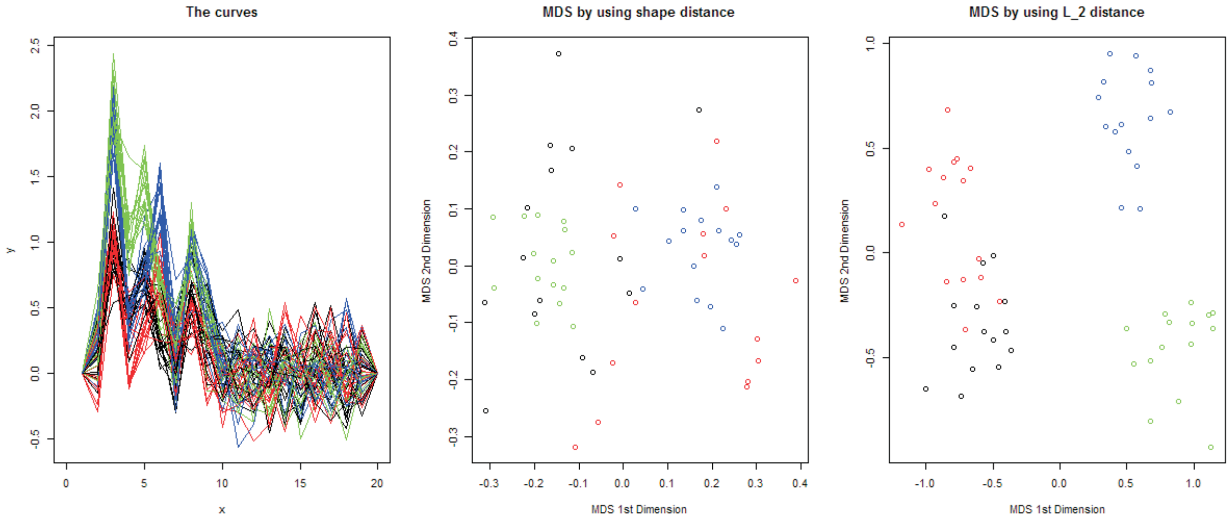

Fig. 3. Simulated results for example 1. Left plot is the actual data; middle plot is a 2D MDS plot for shape distance; right plot is a 2D MDS plot for L_2 distance.

Computation by using R reveals that $bc - ad < 0$. According to the derivations in the previous section, we expect that L_2 norm to outperform the shape distance. In Fig. 3, the left plot shows the overall pattern of the four groups of curves with colors black, red, green, and blue corresponding to the groups from one to four. The middle plot shows the first two dimensions after performing multidimensional scaling (MDS) on the distance matrix by using the shape distances on these curves, and the right plot shows the first two dimensions after using MDS on the L_2 distances on these curves. It is obvious from these plots that the L_2 norm is better at differentiating between the different groups.

4.2 Second Simulation

In the second simulation data set, there are again four different groups of curves with dimension 20. The design is similar to the first simulated data set; however, the major difference is that the fixed term is not constant within groups. Instead, it $\alpha_{ij} = (j \bmod 5)v_i$ for $i = 1, \ldots, 4$ and $j = 1, \ldots, 15$. Notice that (j mode 5) returns 1 for the first 5 replicates, 2 for the second 5 replicates and so forth. This term serves as an amplifier

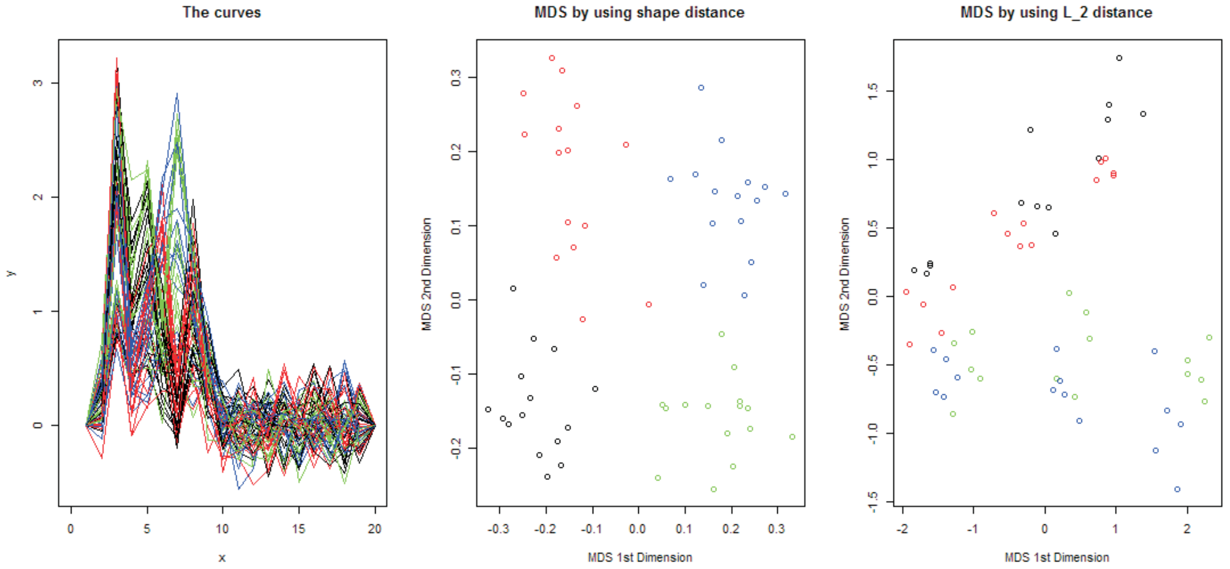

Fig. 4. Simulated results for example 1. Left plot is the actual data; middle plot is a 2D MDS plot for shape distance; right plot is a 2D MDS plot for L_2 distance.

which provides shape variation within each group. The v_i's are defined as follows:

$v_1 = (0, 0.1, 1, 0.5, 0.7, 0.3, 0.1, 0.5, 0.2, 0, \ldots .0)^T$,

$v_2 = (0, 0.1, 1, 0.2, 0.4, 0.6. 0.1, 0.5, 0.2, 0, \ldots .0)^T$,

$v_3 = (0, 0.1, 1, 0.5, 0.7, 0.3, 0.1, 0.5, 0.2, 0, \ldots .0)^T$,

$v_4 = (0, 0.1, 1, 0.2, 0.4, 0.6. 0.1, 0.5, 0.2, 0, \ldots .0)^T$.

Computations via *R* reveals that $bc - ad > 0$. According the previous section, we expect that the shape distance will outperform the L_2 norm. In Fig. 4, the left plot shows the overall pattern of the four groups of curves with colors black, red, green, and blue corresponding to the groups from one to four. The middle plot shows the first two dimensions after performing MDS on the distance matrix by using the shape distances on these curves, and the right plot shows the first two dimensions after using MDS on the L_2 distances on these curves. It is obvious from these plots that the shape distance is better at differentiating between the different groups.

5. EXAMPLE

5.1 Grain Data Set

Eleven different inbred genotypes of maize (listed in Table 3) were grown in a single field environment at the Central Crops Research Station in Clayton, NC in 2008. The eleven genotypes are inbreds that have full single-nucleotide polymorphism sequence information (Chia *et al.* 2012) (see also http://www.panzea.org/) and were parents of a widely used nested association mapping population (Bukler *et al*. 2009) The plants were hand pollinated and the grain harvested and dried. For each genotype, three different plant cobs were randomly chosen for shelling, to create three replicate samples of whole grain from each genotype. Three kernels from each cob were ground using a ball mill (SPEX SamplePrep 5100 mixer mill, SpexSamplePrep Co, Metuchen, NJ) and the milled sample was stored dry in paper envelopes until measurement of 0.1g of sample in the collection window of the FT-IR instrument.

Table 3. Maize Genotypes by groups classified according to Liu *et al*. (2003)

Number	Genotype	Genetic relatedness group
1	B97	NSS
2	CML103	TS
3	CML333	TS
4	CML52	TS
5	M37W	Mix
6	Mo18W	Mix
7	NC350	TS
8	NC358	TS
9	Oh43	NSS
10	Oh7B	NSS
11	Tx303	Mix

Grain quality was examined essentially as described in Jiang *et al*. (2007) and Kuhnen *et al.* (2010), except that the instrument used was a Thermo Scientific Nicolet 6700 with the SmartPerformer ATR attachment. Ten scans with 4 scan/cm were run for each sample, from wavenumber of 4000 to 450cm-1 and the average of the ten scans was retained for curve analysis.

5.2 Analysis

To begin the analysis, we examine the distances due to shape alone, *i.e*. set $\lambda = 1$. Fig. 5 is an MDS plot with $d = 2$ that displays the distances between these two curves. Due to too much noise in the shape distance within each genotype, it is difficult to discern differences between the genotypes due to shape.

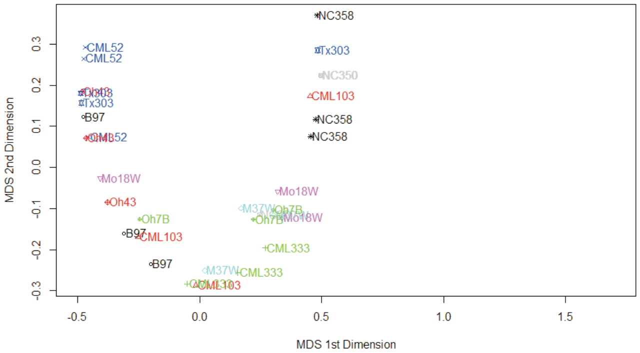

Fig. 5. 2D MDS representation of FT-IR curves by using the shape distance in equation (1).

Now we examine the distances due to just magnitude alone, or in other words, set $\lambda = 0$. Fig. 6 illustrates the 2 dimension MDS plot for magnitude distances alone. Looking at Fig. 6, it is obvious that the curves for genotypes NC350 and NC358 are well separated from the rest of the curves and are in close proximity to each other. It is interesting to note the close proximity of these two genotypes, since these two genotypes share a common origin (developed by the same individual, same time and same place). From this analysis, the FT-IR curves appear to capture this information.

Using the methodology described in Section 3, we determine that $bc < ad$, or in other words the optimal value of λ should be 0. When $\lambda = 0$, the L_2 distance should outperform the shape distance. Looking at Figs. 5 and 6, we can see that the L_2 distance helps discern differences between the groups better than the shape distance.

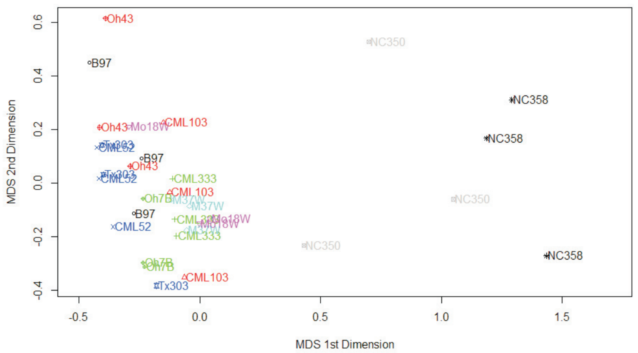

Fig. 6. 2D MDS representation of FT-IR curves by using the L_2 distance

6. CONCLUSION AND DISCUSSION

With continued technological advancements, more and more data will be obtained in a high dimensional structure, such as data generated via FT-IR. The need for identifying relationships and understanding the information contained in such data is a growing area of research. In the case of high dimensional data, it is important that all relevant information is captured in its analysis. Thus, in the analysis of FT-IR data, we proposed using shape, magnitude and their hybrid distances to measure the differences among the data. This information ensures that we are capturing information along the entire curve, in order to better observe and understand relationships between the entities in which the curve represents.

The analysis of FT-IR spectral data for maize captured an interesting characteristic of the genetic data. Two genotypes, NC350 and NC358 were very close to each other, but separated from the other genotypes. These two genotypes share a common origin, which could account for the relatedness of these two genotypes. The remaining homozygous inbreds in this study are likely to have a mixture of positive and negative genetic effect alleles that could explain the relatively small and insignificant differences in their FT-IR curves. Future studies involving selection and mapping population testing could be used to separate these allele effects to allow FT-IR phenotype analysis of progeny using our new method.

Through this research, we discovered that λ is important in determining the correct distance measure to use to differentiate between different curves. However, the important values of λ are only 0 and 1 due to the monotonicity of the objective function. The simulated data sets, as well as the derivations indicate that it is important to make the correct chose for λ to correctly analyze data from these types of experiments.

ACKNOWLEDGEMENTS

The authors would like to thank the reviewer for the helpful comments that significantly improves the presentation of the paper.

REFERENCES

Baye, T.M., Pearson, T.C. and Settles, A.M. (2006). Development of a calibration to predict maize seed composition using single kernel near infrared spectroscopy. *J. Cereal Sci.*, **43(2)**, 236-243.

Buckler, E.S., Holland, J.B., Bradbury, P.J., Acharya, C.B., Brown, C.B., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia, A. and Glaubitz, J.C. *et al*. (2009). The genetic architecture of maize flowering time. *Science*, **325(5941)**, 714-718.

Chia, J.-M., Song, C., Bradbury, P.J., Costich, D., de Leon, N., Doebley, J., Elshire, R.J., Gaut, B., Geller, L. and Glaubitz, J.C. *et al.* (2012). Maize hapmap2 identifies extant variation from a genome in flux. *Nature Genet.*, **44(7)**, 803-807.

Cook, J.P., McMullen, M.D., Holland, J.B., Tian, F., Bradbury, P., Ross-Ibarra, J., Buckler, E.S. and Flint-Garcia, S.A. (2012). Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. *Plant Physiology*, **158(2)**, 824-834.

Dryden, I. and Mardia, K. (1998). *Statistical Shape Analysis*. John Wiley and Sons.

Flint-Garcia, S.A., Bodnar, A.L. and Scott, M.P. (2009). Wide variability in kernel composition, seed characteristics, and zein profiles among diverse maize inbreds, landraces, and teosinte. *Theo. Appl. Genet.*, **119(6)**, 1129-1142.

Jiang, H., Zhu, Y., Wei, L., Dai, J., Song, T., Yan, Y. and Chen, S. (2007). Analysis of protein, starch and oil content of single intact kernels by near infrared reflectance spectroscopy (nirs) in maize (zea mays l.). *Plant Breeding*, **126(5)**, 492-497.

Kendall, D. (1984). Shape manifolds, procrustean metrics, and complex projective spaces. *Bull. London Math. Soc.*, **16(2)**, 81-121.

Kuhnen, S., Ogliari, J.B., Dias, P.F., Boffo, E.F., Correia, I., Ferreira, A.G., Delgadillo, I. and Maraschin, M. (2010). Atr-ftir spec- troscopy and chemometric analysis applied to discrimination of landrace maize flours produced in southern brazil. *Int. J. Food Sci. Tech.*, **45(8)**, 1673-1681.

Liu, K., Goodman, M., Muse, S., Smith, J., Buckler, E. and Doebley, J. (2003). Genetic structure and diversity among maize inbred lines as inferred from dna microsatellites. *Genetics*, **165,** 2117-2128.

Lund, R. and Li., B. (2009). Revisiting climate region definitions via clustering. *J. Climate*, **22(7)**, 1787-1800.

Stermer, R., Pomeranz, Y. and McGinty. R. (1977). Infrared reflectance spectroscopy for estimation of moisture of whole grain. *Cereal Chemistry*, **54(2),** 345-351.

Sujatha, L., Rai, P., Kumar, K., Mahato, V., Kartha and C. Santhosh. (2008). Serum protein profile study of normal and cervical cancer subjects by high performance liquid chromatography with laser-induced fluorescence. *J. Biomed. Optics*, **13(5)**, 54-62.

Sun, X. and Weckwerth, W. (2012). Covain: a toolbox for uni-and multivariate statistics, time-series and correlation network analysis and inverse estimation of the differential jacobian from metabolomics covariance data. *Metabolomics*, **8(1),** 81-93.

Thygesen, L.G., Løkke, M.M., Micklander, E. and Engelsen, S.B. (2003). Vibrational microspectroscopy of food. raman vs. ft-ir. *Trends Food Sci. Tech.*, **14(1-2)**, 50-57.

Wang, Y., Chen, C., Albert, M., Chang, Y. and Ricanek, K. (2013). Eyebrow shape analysis by using a modified functional curve procrustes distance. *accepted by IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS).*