

DR. JOHN V. GOODPASTER

05/08/2012

**MULTI-VARIATE STATISTICAL ANALYSIS OF UV-VISIBLE SPECTRA OBTAINED FROM
FIBER EVIDENCE COLLECTED IN STATE v. ECHOLS, BALDWIN AND MISSKELLEY**

May 8, 2012

To:

Christopher Bommarito
Forensic Science Consultants
1099 Grand River Avenue
Williamston, MI 48895
bommarito@forsci.com

Submitted By:

John V. Goodpaster, Ph.D.
Goodpaster Forensic Consulting
Zionsville, IN 36077
john.goodpaster@att.net

Contents

INTRODUCTION.....	1
RESULTS	3
COMPARISON 1	3
COMPARISON 2	4
COMPARISON 3	5
COMPARISON 4	6
COMPARISON 5	7
COMPARISON 6	8
CONCLUSIONS	9

SUMMARY

This report pertains to 95 UV-visible absorbance spectra that were acquired by Christopher Bommarito using a CRAIC QDI 2000 microspectrophotometer at Indiana University Purdue University Indianapolis (IUPUI) on April 10, 2012. These spectra were then exported into Microsoft Excel for statistical analysis. The software package used in this case was XLSTAT 2011 (AddinSoft, Paris, France). This program is an add-on for Microsoft Excel. Overall, the spectra corresponded to nine different known and questioned fiber samples. A total of six comparisons were made between the following known (K) and questioned (Q) fibers:

Known (K) Samples	Questioned (Q) Samples
E79 (green polyester)	E5
E79 (blue-green cotton)	E9, F2
E92 (red cotton)	E1, E3, F1

For each comparison, the spectra from the K and Q samples were background corrected and normalized. The spectra were then inspected visually to identify any systematic differences between the two samples. Then, Principal Components Analysis (PCA) was carried out to visualize the extent to which the spectra from the K and Q could be differentiated. Finally, a sub-set of significant principal components (PCs) was passed on to a Discriminant Analysis (DA) algorithm. DA was used to quantitatively test the extent to which the K and Q spectra could be reliably differentiated from one another. Ultimately, K/Q pairs were declared to be different IF they exhibited systematic differences in their spectra AND these differences could be visualized using PCA AND the differences could be confirmed by DA with a cross-validation accuracy of at least 95%.

INTRODUCTION

Chemometrics is a term used to describe a family of statistical methods that are applied to data generated by chemical instrumentation. Chemometric methods can reduce the complexity of a large data set (e.g., a large collection of replicate spectra obtained from multiple samples), thereby allowing users to quickly and easily compare groups of spectra. Chemometrics can also make predictions about unknown samples. Ultimately, chemometrics can be used to interpret the results of forensic analyses, especially those that involve trying to discern subtle differences between very similar samples.

The chemometric approach used in this case followed three distinct steps:

- 1) Data pre-processing was carried out in order to remove any unimportant differences between samples
- 2) "Unsupervised" chemometric methods were utilized to reduce the complexity of the data and to visualize its underlying structure (e.g., the extent of similarity/dissimilarity between two samples)
- 3) "Supervised" chemometric techniques were then utilized to test the extent to which samples can be differentiated by predicting class memberships

Preprocessing is the preparation of data before the application of chemometric algorithms. This step is important because it can remove noise and variation that might complicate data interpretation. In this case, background correction was used to remove varying background levels in the spectra. Background correction was accomplished by calculating the average absorbance value for a given spectrum and then subtracting that value from all absorbance values in the spectrum. Following background correction, the spectra were normalized. Normalization eliminates variations due to sample size, concentration, amount, and instrument response. In this case, normalization was accomplished by calculating the square root of the sum of squares of all absorbance values and then dividing the absorbance values by this quantity.

The next step of the analysis was to carry out Principal Component Analysis (PCA), which is a technique that transforms the original variables in a data set to a smaller number of significant principal components (PCs). PCA is considered to be "unsupervised" in that the data is not pre-classified or changed in any way; instead, the data is displayed so that the maximum amount of variation can be viewed in as few dimensions as possible. The information gained by PCA can be visually represented in a "scores plot". This plots the score of one PC against the score of another for each sample. Those objects that cluster closely together are similar to one another, while objects that far apart on a scores plot are dissimilar to one another.

Next, a number of principal components were selected to represent the data set in the final step, which was linear discriminant analysis (DA). If too many principal components are used at this stage, the "noise" from extra principal components may interfere with the formation and verification of classes. Several methods can be used to choose the correct number of PCs. In

this case, enough PCs were retained so that the total amount of variance represented was at least 95%.

With this data in hand, DA was used to create a new set of axes that placed spectra from a given sample (Q or K) as close together as possible, and placed spectra from the opposing sample as far away as possible. DA is a form of "supervised" pattern recognition, as it requires knowledge of group memberships for each sample. Ultimately, the validity of describing the spectra as falling into two distinct groups can be tested. In this method, a sample is removed from the data set temporarily. The classification model is then built from the remaining samples, and then used to predict the classification of the deleted sample. This "leave one out" process continues through all of the samples, treating each sample as an unknown to be classified using the remaining samples. The result of this validation is expressed as a percentage of the observations that were correctly classified into their original class membership. Only K/Q pairs that had an accuracy rate of 95% or higher were considered to be reliably differentiated.

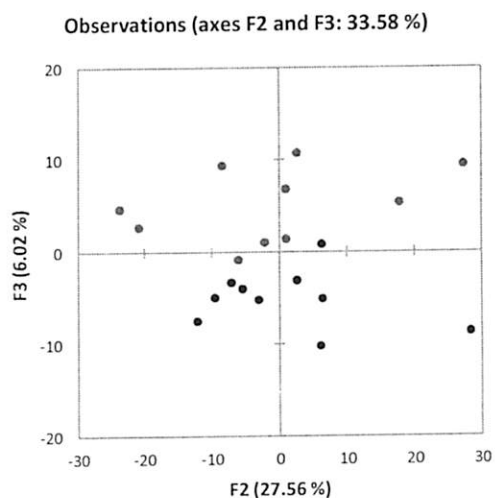
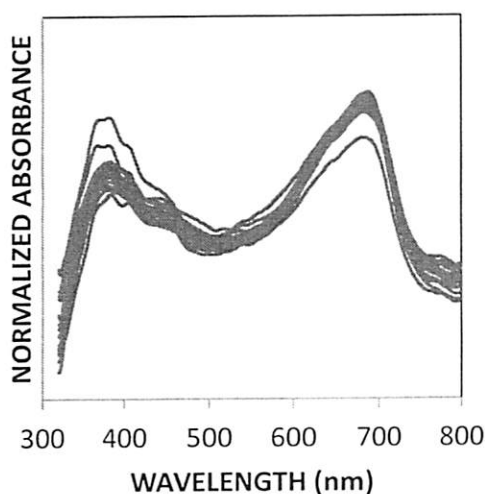
RESULTS

COMPARISON 1

Known = E79 green polyester

Unknown = E5 (single fiber)

The background subtracted and normalized spectra from the known (in blue) and unknown (in red) are shown below plotted on the same set of axes as well as in a PCA scores plot (F2 vs. F3):



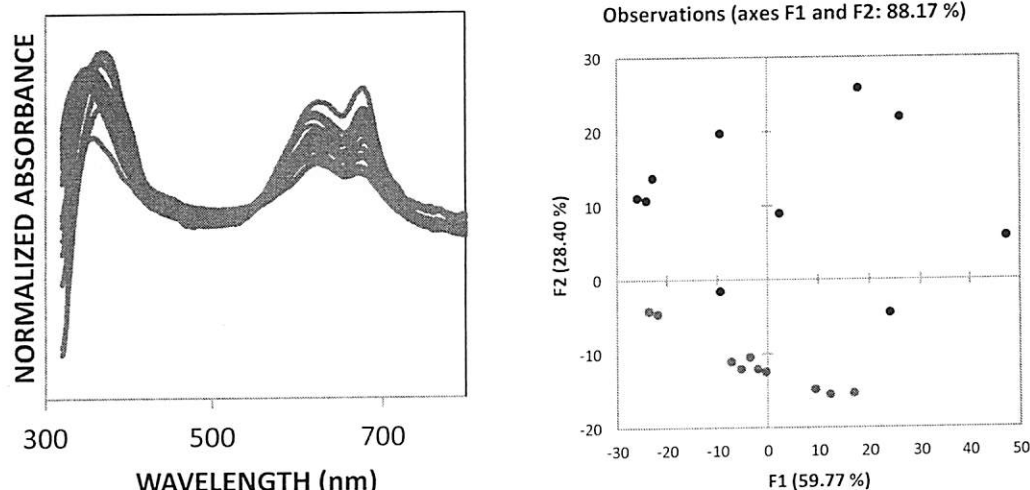
Visual inspection of the spectra shows that they have the same general shape and their wavelengths of maximum and minimum absorbance are similar. However, there are also several small peaks that appear between 350 – 500 nm in the red spectra (E5) but do not appear in the blue spectra (E79). Principal Component Analysis (PCA) indicated that although the spectra from these two samples were similar, they could be distinguished as two separate groups (as shown above). The first four principal components were retained and they represented 97.1% of the total variability in the data. These four PCs were then used to construct a model in discriminant analysis and the two groups were clearly distinguished. This model was cross-validated with 95% accuracy. Therefore, the two groups of spectra can be considered to be different and reliably differentiated.

COMPARISON 2

Known = E79 blue-green cotton

Unknown = E9 (single fiber)

The background subtracted and normalized spectra from the known (in blue) and unknown (in red) are shown below plotted on the same set of axes as well as in a PCA scores plot (F1 vs. F2):



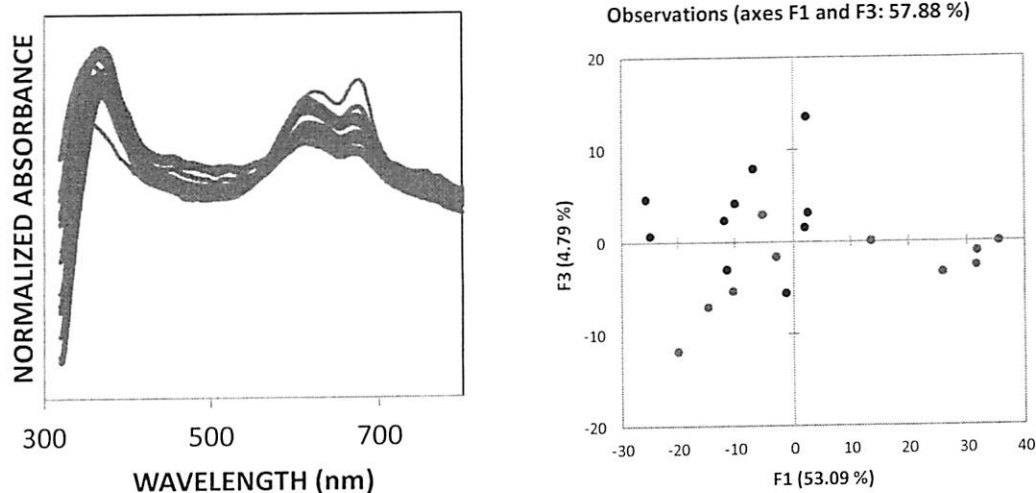
Visual inspection of the spectra shows that they have the same general shape and their wavelengths of maximum and minimum absorbance are similar. However, there is also a systematic difference between the known and unknown as seen in the peak height ratios between 550 – 750 nm. Principal Component Analysis (PCA) indicated that although the spectra from these two samples were similar, they could be distinguished as two separate groups (as shown above). The first four principal components were retained and they represented 96.7% of the total variability in the data. These four PCs were then used to construct a classification model in discriminant analysis and the two groups were clearly distinguished. This model was cross-validated with 95% accuracy. Therefore, the two groups of spectra can be considered to be different and reliably differentiated.

COMPARISON 3

Known = E79 blue-green cotton

Unknown = F2 (multiple fibers)

The background subtracted and normalized spectra from the known (in blue) and unknown (in red) are shown below plotted on the same set of axes as well as in a PCA scores plot (F1 vs. F3):



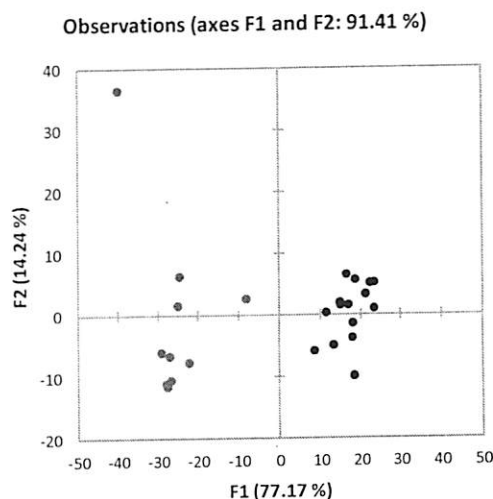
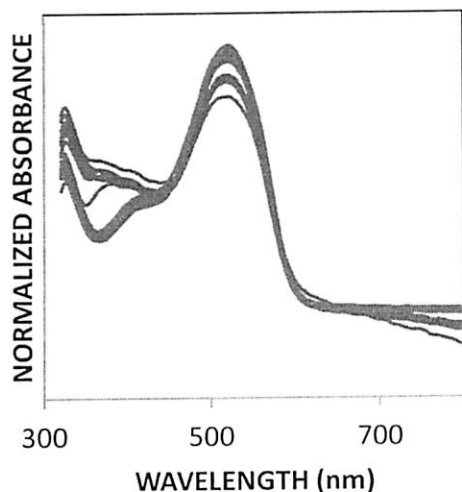
Visual inspection of the spectra shows that the fibers have the same general shape and their wavelengths of maximum and minimum absorbance are similar. Principal Component Analysis (PCA) indicated that the spectra from these two samples were very similar and they could not be displayed as two separate groups. The first three principal components were retained and they represented 95.6% of the total variability in the data. These three PCs were then used to construct a model in discriminant analysis. However, the model was cross-validated with only 80% accuracy. This result indicates that these two samples cannot be reliably distinguished from one another.

COMPARISON 4

Known = E92 red cotton

Unknown = E1 (single fiber)

The background subtracted and normalized spectra from the known (in blue) and unknown (in red) are shown below plotted on the same set of axes as well as in a PCA scores plot (F1 vs. F2):



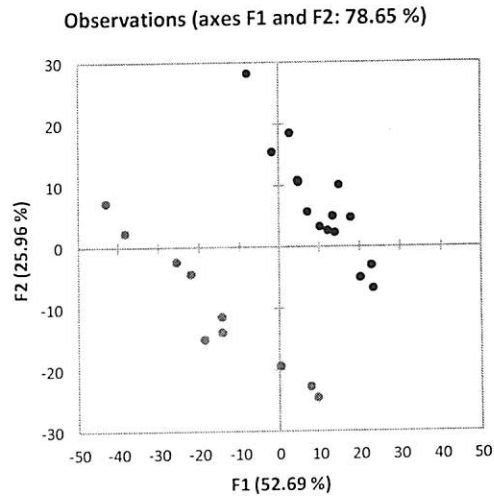
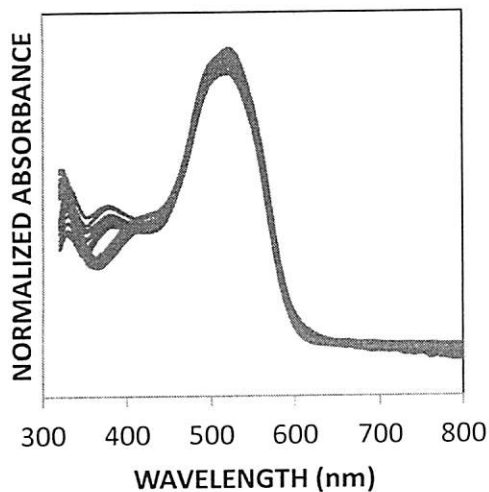
Visual inspection of the spectra shows that they exhibit differences throughout the region between 320 nm and 800 nm. Principal Component Analysis (PCA) indicated that the spectra from these two samples can be clearly distinguished as two separate groups. The first three principal components were retained and they represented 97.4% of the total variability in the data. These three PCs were then used to construct a model in discriminant analysis. The model was cross-validated with 100% accuracy. Therefore, the two groups of spectra can be considered to be different and reliably differentiated.

COMPARISON 5

Known = E92 red cotton

Unknown = E3 (two fibers)

The background subtracted and normalized spectra from the known (in blue) and unknown (in red) are shown below plotted on the same set of axes as well as in a PCA scores plot (F1 vs. F2):



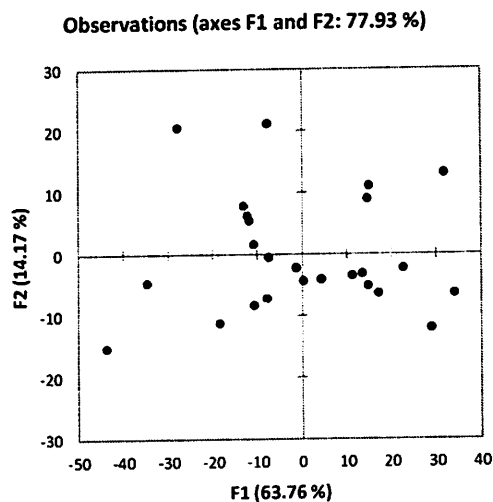
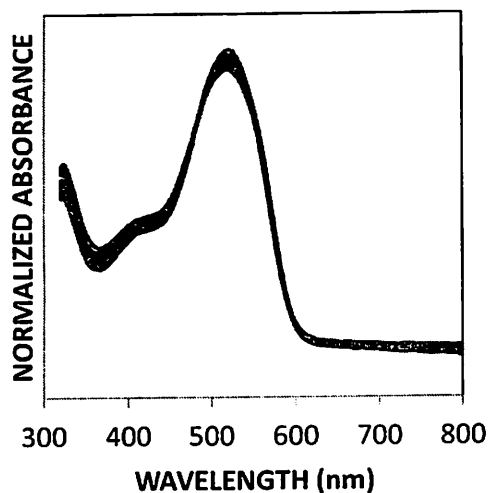
Visual inspection of the spectra shows that they have the same general shape but there are significant differences between 320 and 420 nm. Principal Component Analysis (PCA) clearly indicated that the spectra from these two samples may be distinguished as two separate groups. The first four principal components were retained and they represented 97.1% of the total variability in the data. These four PCs were then used to construct a model in discriminant analysis. The model was cross-validated with 100% accuracy. Therefore, the two groups of spectra can be considered to be different and reliably differentiated.

COMPARISON 6

Known = E92 red cotton

Unknown = F1 (single fiber)

The background subtracted and normalized spectra from the known (in blue) and unknown (in red) are shown below plotted on the same set of axes as well as in a PCA scores plot (F1 vs. F2):



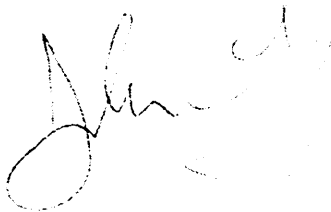
Visual inspection of the spectra shows that they are extremely similar. Principal Component Analysis (PCA) indicated that the spectra from these two samples were very similar and did not form two separate groups. The first four principal components were retained and they represented 96.4% of the total variability in the data. These four PCs were then used to construct a model in discriminant analysis. However, the model was cross-validated with only 84% accuracy. This result indicates that these two samples cannot be reliably distinguished from one another.

CONCLUSIONS

A statistical analysis of the spectra that were generated in this case has resulted in a number of critical findings. These are summarized in the table below where the term "different" means that the spectra were visibly different, the projection of the data using PCA revealed two clear groups of spectra and discriminant analysis could accurately ($\geq 95\%$) discriminate between the two groups. The term "not different" means that the spectra from the two groups are very similar, the projection of the data using PCA revealed that the spectra from the two groups were co-mingled, and discriminant analysis was not able to accurately ($< 95\%$) discriminate between the two groups.

	E1	E3	E5	E9	F1	F2
E79 (green polyester)			different			
E79 (blue-green cotton)				different		not different
E92 (red cotton)	different	different			not different	

Signed:



John V. Goodpaster, Ph.D.
Goodpaster Forensic Consulting, LLC