Model-Based Discriminant Analysis of Near-Infrared Spectroscopic Data in Food Authenticity.

Deirdre Toher*†, Gerard Downey* and Thomas Brendan Murphy†

* Teagasc, The National Food Centre, Ashtown
† Department of Statistics, Trinity College Dublin.

Motivation

After the food scares of recent years, food authenticity studies have become an increasingly important tool for boosting consumer confidence. Honey is a highly variable natural substance that is relatively expensive to produce, therefore is more likely to be the subject of intentional adulteration. As honey is amongst the more difficult of foodstuffs to authenticate, it should be possible to generalize techniques developed specifically for honey for use on other foodstuffs.

Near-infrared spectroscopy is an inexpensive non-destructive analytical technique, thus is ideally suited for food authenticity studies. The technique is useful in preliminary studies where the aim is to develop methods to reliably classify samples into "pure" and "needs further testing, but probably adulterated".

As can be seen from figure 1a, authentic honey, even when from the same geographic region, is extremely variable – the wider the line, the more variable the honey samples are at that wavelength. Indeed the region of the near-infrared spectrum with the greatest variability happens to represent the natural sugars in honey. Honey is often adulterated by these sugars, or by compounds which echo the spectral composition of these sugars.

Thus visually examining the spectra alone is unable to provide satisfactory classification results. Therefore, statistical procedures are required in order to ensure that the analysis of the data is performed in a consistent manner.

25 honey samples from throughout Ireland were adulterated with adulterant solutions consisting of fructose and glucose in the following ratios: 0.7:1, 1.2:1 and 2.3:1 weight/weight (w/w), each of the following three levels 7,14 and 21% w/w, producing 225 adulterated honeys. A further 50 unadulterated honeys were added to the sample set, as outlined in [1].

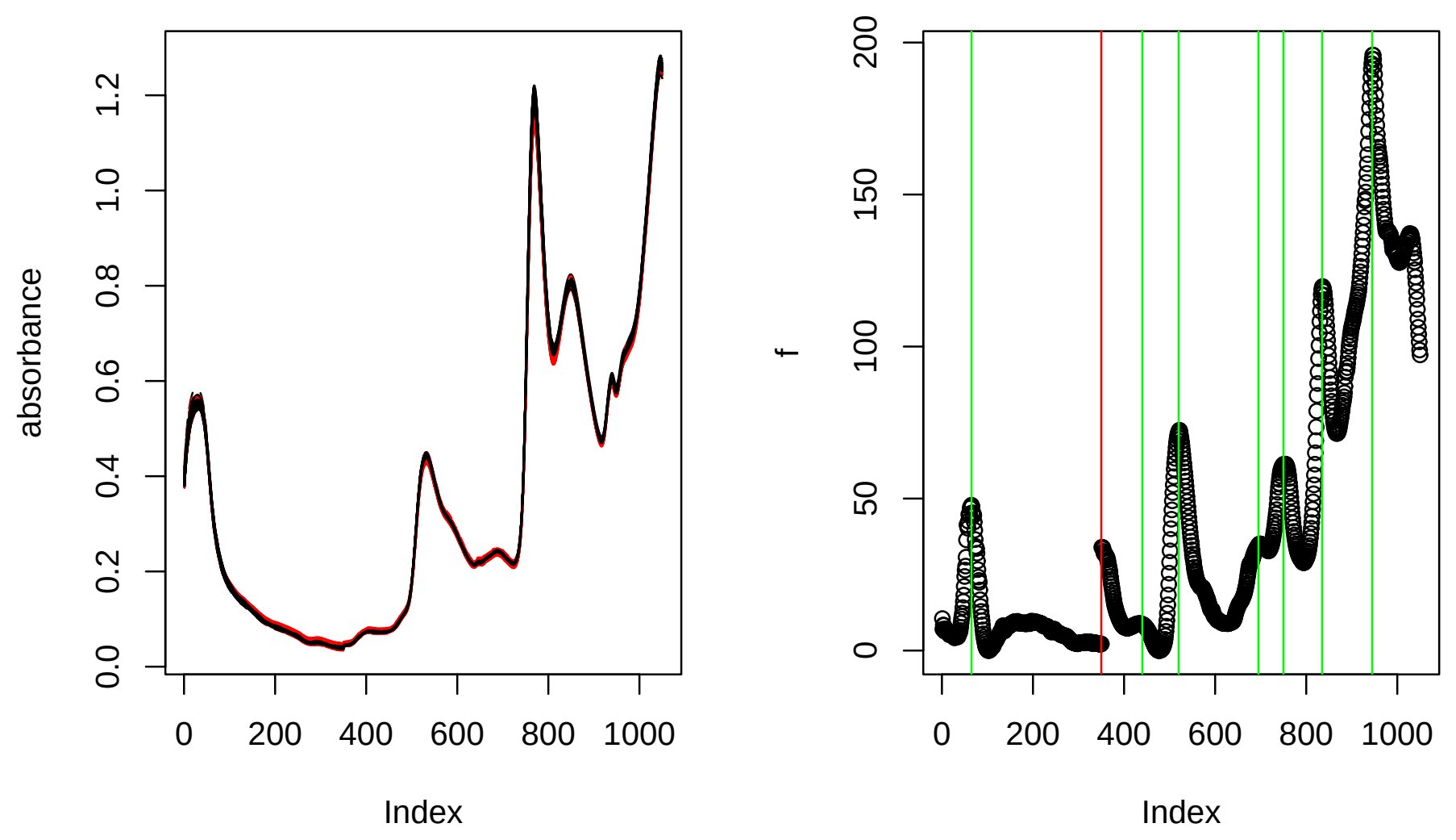

FIGURE 1a *Near-Infrared Spectra of Authentic Honey Samples*
FIGURE 1b *F-statistic for the spectral range 400-2500nm*

Data Reduction Methods

The near-infrared data span from 400-2500 nm, where measurements are taken every 2 nm. Adjacent values are closely correlated. Therefore dimension reduction is the first issue to be confronted. Two techniques for dimension reduction are examined here – using the peaks of the F-statistic for the spectra and using Wavelet Analysis.

F-statistic

The basis of the F-statistic is to examine the between-group variability. A simple method to dramatically reduce the dimensionality of the data is to use the peaks of the graph to choose the wavelengths with the most variability between authentic and adulterated honey. However, there is a discontinuity at 1100 nm, where the sensors used are changed, at the boundary of the visible and infrared frequencies; this is not visible within the spectra of figure 1a.

As this discontinuity is likely to be one caused by the process, rather than by the honey, this wavelength is excluded for the purposes of further evaluation. The major peaks are highlighted in figure 1b in green, with the discontinuity marked in red. Using major peaks, it is possible to select less than 10 wavelengths on which to perform further analysis.

Wavelet Analysis

Wavelet analysis is used to decompose a spectrum into a series of wavelet coefficients. The coefficients can be used to reconstruct the original spectrum, so no data reduction occurs. However, on examining the coefficients produced by the wavelet analysis, it is evident that many are zero or close to zero.

Thresholding is used to select the coefficients that contain important information on the structure of the spectrum. Many thresholding techniques have been proposed and the choice of methods is a subjective one.

The Daubechies' wavelet is a consistently reliable type to use and is the default within `wavethresh` [2]. Efficient wavelet analysis methods require that the dimension of the data must be 2^m, where m is an integer. This is not a restriction in this application because we have 1050 measurements and we use $2^{10} = 1024$ of these.

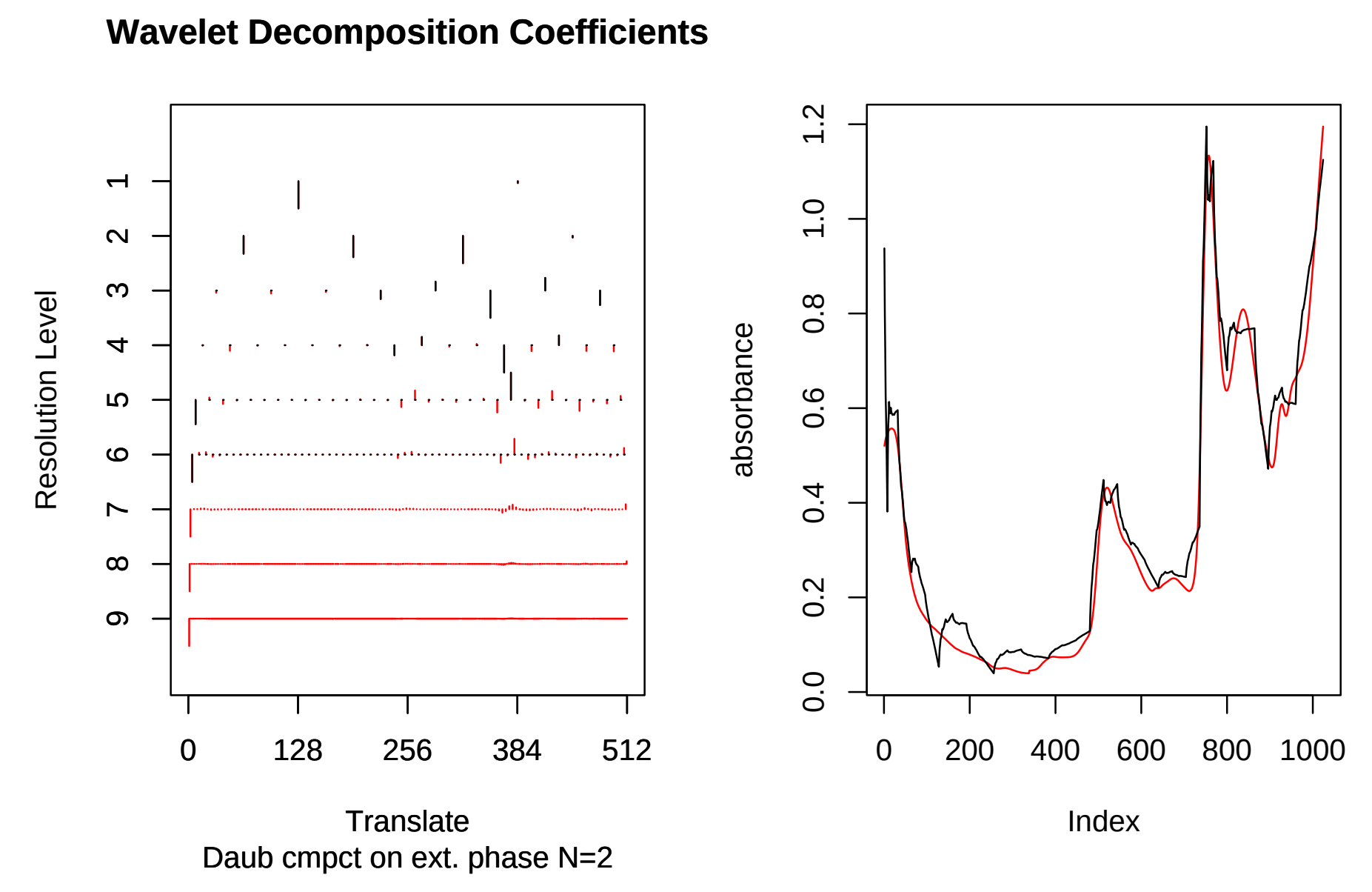

FIGURE 2a *Wavelets Decomposition – thresholded & non-thresholded*
FIGURE 2b *Actual and Reconstructed Thresholded Wavelets*

Figure 2a shows the structure of the wavelet analysis for one sample of pure honey, both in its full and thresholded form, while figure 2b shows the reconstructed spectrum of the same sample after thresholding, in comparison to the actual spectrum for the sample.

Classification Techniques

The classification techniques used on this data set are based on Gaussian mixture models; each group is modelled using a Gaussian distribution. The covariance structure of the Gaussian models are structured in a parsimonious manner using constraints. This approach offers the ability to model groups that have distinct volume, shape and orientation properties.

Fraley and Raftery's paper [3] describes a methodological approach towards cluster analysis, with specific mention towards model-based Discriminant Analysis. Their **`mclust`** [4] package was used to perform the model-based Discriminant Analysis.

This allows for the possibility of the following models:

TABLE 1: Parametrizatons of the covariance matrix Σ_k

Model ID	**Decomposition**	**Distribution**
EII	$\Sigma_g = \lambda I$	Spherical
VII	$\Sigma_g = \lambda_g I$	Spherical
EEI	$\Sigma_g = \lambda D D^T$	Diagonal
VEI	$\Sigma_g = \lambda_g D D^T$	Diagonal
EVI	$\Sigma_g = \lambda D_g D_g^T$	Diagonal
VVI	$\Sigma_g = \lambda_g D_g D_g^T$	Diagonal
EEE	$\Sigma_g = \lambda D A D^T$	Ellipsoidal
EEV	$\Sigma_g = \lambda D_g A D_g^T$	Ellipsoidal
VEV	$\Sigma_g = \lambda_g D_g A D_g^T$	Ellipsoidal
VVV	$\Sigma_g = \lambda_g D_g A_g D_g^T$	Ellipsoidal

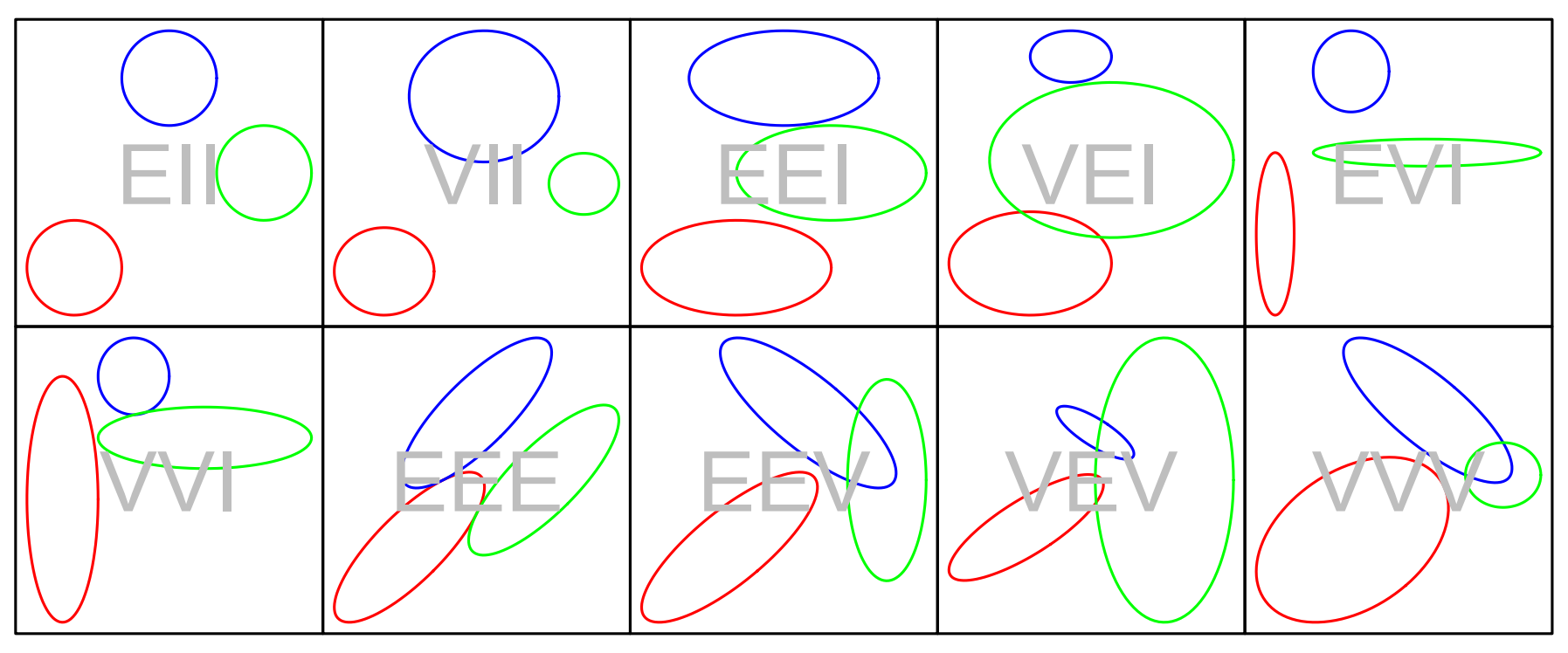

FIGURE 3: *General Shapes of Models*

Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) correspond to the EEE and VVV models respectively. The letters of the Model ID represent the volume, shape and orientation of the groups.

Model Selection and Verification

1000 simulations were performed, each one taking a random split of the data to form a training set of 200 observations. 50 of the observations in the training set were pure samples and 150 were adulterated samples. The remaining 100 observations then formed a test set for the simulation. Over the 1000 simulations the mean classification error for both the training and the test data were calculated. The models were selected on the basis of their respective performance on the training data.

Results

The top three performing models were **"EEV"**, **"VEV"** and **"EEE"**. These models performed significantly better than the other models on the training data and were also the three best performing models on the test data. The range of the error rates for each of these models is given in Table 2. Example classification tables for the test sets of the three best models are also shown.

The initial study [1] achieved a misclassification rate of just under 3% using partial least squares regression on the 2nd derivatives of the entire spectra.

TABLE 2: Range of Error Rates

	Training	**Test**
EEV		
F-statistic	1.5 – 6.0%	1.0 – 11.0%
Wavelet Analysis	0.0 – 3.5%	1.0 – 12.0%
VEV		
F-statistic	1.5 – 6.5%	1.0 – 11.0%
Wavelet Analysis	0.0 – 4.0%	1.0 – 12.0%
EEE		
F-statistic	1.5 – 7.0%	1.0 – 11.0%
Wavelet Analysis	0.5 – 5.5%	0.0 – 9.0%

EEV

F-statistic

		Predicted	
		Pure	Adult.
Actual	Pure	22	**3**
	Adult.	**1**	74

Mean training set error: 3.9%
Mean test set error: 5.1%

Wavelet Analysis

		Predicted	
		Pure	Adult.
Actual	Pure	20	**5**
	Adult.	**0**	75

Mean training set error: 1.7%
Mean test set error: 6.3%

VEV

F-statistic

		Predicted	
		Pure	Adult.
Actual	Pure	21	**4**
	Adult.	**1**	74

Mean training set error: 3.6%
Mean test set error: 5.3%

Wavelet Analysis

		Predicted	
		Pure	Adult.
Actual	Pure	20	**5**
	Adult.	**1**	74

Mean training set error: 1.7%
Mean test set error: 6.3%

EEE or Linear Discriminant Analysis

F-statistic

		Predicted	
		Pure	Adult.
Actual	Pure	23	**2**
	Adult.	**2**	73

Mean training set error: 4.2%
Mean test set error: 5.4%

Wavelet Analysis

		Predicted	
		Pure	Adult.
Actual	Pure	24	**1**
	Adult.	**2**	73

Mean training set error: 2.4%
Mean test set error: 4.1%

Conclusions and Further Work

The dimension reduction techniques demonstrated above yield quite promising results, given that only one approach towards classification was examined. Using the F-statistic proved to be a surprising effective technique, for one that can be explained to those with limited knowledge of statistics.

The reliability of these methods on detecting other forms of adulteration must also be examined – other methods of adulteration that prove problematic to detect include adulteration with beet/cane inverts, and adulteration with high fructose corn syrup. The ability to quantify the level of adulteration also requires further exploration. These techniques are being developed towards a commerical and regulatory standard, and so must be proven to perform not only on Irish honey but on samples from throughout the world.

The potential for using updating algorithms should also be explored.

Acknowledgements

This work was funded by the Teagasc, under the Walsh Fellowship Scheme.

References

[1] G. Downey, V. Fouratier and J.D. Kelly, Detection of honey adulteration by addition of fructose and glucose using near infrared transflectance spectroscopy, *J. Near Infrared Spectrosc.* **11**, 447–456 (2003)

[2] G. Nason, A. Kovac (1997) and M. Maechler (1999), wavethresh: Software to perform wavelet statistics and transforms, R package version **2.2-8**, (2004)

[3] C. Fraley, A.E. Raftery, Model-based clustering, discriminant analysis, and density estimation, *J. Amer. Stat. Assoc.* **97**, 611–631, (2002)
available online as Technical Report no. 380, October 2000, www.stat.washington.edu/fraley/mclust/rep.shtml

[4] C. Fraley, A.E. Raftery and R. Wehrens (R-port), mclust: Model-based cluster analysis