Current Development in Theory and Applications of Wavelets
Volume 1, Issue 2, 2007, Pages 165-182
This paper is available online at http://www.pphmj.com

REPEATED MEASURES EXPERIMENTAL DESIGN IN NEAR INFRARED SPECTROSCOPY USING ADAPTIVE WAVELETS

DAVID DONALD[1*], DANNY COOMANS[1], YVETTE EVERINGHAM[1], DANIEL COZZOLINO[2] and MARK GISHEN[2]

[1]BSES Limited, 713178 Bruce Highway
Meringa, Queensland, 4865, Australia
e-mail: DDonald@BSES.org.au

[2]The Australian Wine Research Institute
PO Box 197, Glen Osmond 5064, Australia
and
[2]Cooperative Research Centre for Viticulture
PO Box 154, Glen Osmond SA 5064, Australia

[*]School of Mathematical and Physical Sciences
James Cook University
Townsville, Queensland 4811, Australia
e-mail: David.Donald@jcu.edu.au

Abstract

We introduce a novel method to analyse near infrared (NIR) spectra collected in a repeated measures experiment. Using an adaptive discrete wavelet transform, our method initially extracts features from the spectra that correlate with the design of the experiment. Then the extracted features are then mapped onto a five-dimensional hyperplane

2000 Mathematics Subject Classification: 62P30, 42-02.

Keywords and phrases: adaptive wavelet, repeated measures, multivariate MANOVA, near infrared, wine grape, penalized discriminate mapping.

Received September 20, 2006

using penalised discriminate mapping (PDM) to form PDM scores. The PDM scores are analysed using a multivariate mixed model (MMM) to determine if the experimental design affects the NIR spectra. Illustration of the method is given by a case study from the viticulture industry, where NIR reflectance measurements (400, 402, ..., 2500nm) were taken from red grape homogenates sampled from a nested repeated measures experimental design consisting of the following factors: various growing regions, vineyards, grape varieties and storage durations. Analyses of the viticulture example using our proposed method identified all main effects and two-way interactions between regions, grape variety and, most importantly, storage duration to all be significant ($P < 0.001$). By visualization and univariate analysis of the PDM scores, we identified regions in the NIR spectrum associated with the storage duration, variety and interaction effects.

1. Introduction

Repeated measures experimental designs using a large number of system variables, such as hundreds of wavelengths from the near infrared (NIR) spectrum, with relatively small sample sizes poses a number of technical difficulties; particularly: degradation of power for statistical analysis [3, 21], and numerical instability [18]. The purpose of this paper is to provide a practical means, circumventing the pre-mentioned difficulties, to analyse repeated measures experimental designs where the sample to variable size ratio is small for NIR spectra.

Complications involved with the analysis of repeated measures or more generally in multivariate mixed models (MMM), with NIR spectra are: firstly the number of variables (sampled wavelengths) excessively exceeds the number of samples, and secondly the variables are highly correlated. The former concern results in a lack of degrees of freedom while the latter results in numerical instability of the MMM. Both of these complications result in a degradation of the power of statistical analysis [12, 18, 21]. To overcome these issues of high variable correlation and low variable to sample ratios, we need to employ a method of feature extraction and/or data compression.

Principal component analysis (PCA) and parallel factor analysis (PARFAC) [2] are the most commonly used methods in chemometrics for feature extraction and data compression. PCA or similar bi-linear models,

such as multivariate curve resolution [19], extract features which describe global variability and compress the spectral data in relatively few latent factors. While PARAFAC and similar tri-linear models are an improvement to the bi-linear cases since the data is known to consist of a time varying component [2]. However, both PCA and PARAFAC do not utilize the group structure (experimental design) which is known a priori in compressing the data. Furthermore, in the case of the PARAFAC model; it is assumed that the time varying experimental unit is the NIR sample [8, 9], which is not necessarily the case as illustrated within the given case study provided.

The goal for the feature extraction process is then to extract information from the spectra that expresses the variation in the spectra resulting from the experimental design. To achieve this, we propose using the wavelet transforms in conjunction with a penalised discriminate mapping (PDM) [13]. Here the wavelets used are trained to extract features and dimension reduction resulting in a high group separation as defined by a Fishers discriminant ratio as well as performing dimension reduction. Further dimension reduction is achieved with the use of the penalised discriminate mapping.

The rational for integrating the wavelet transform (WT) with PDM are based on the philosophy that information contained within the NIR spectrum exists in two general frameworks (i) juxtapositional information within localized wavelengths and (ii) interactions between juxtapositional groupings. The wavelet transform is used to extract juxtapositional information while PDM is used to relate the information between juxtapositional groupings succinctly. The choice of the wavelet transform is of some importance as different wavelets can lead to heteromorphic features [22].

In most NIR WT applications to date, the wavelet used is selected from one of eight standard types of wavelets [16] mainly as a matter of convenience. However, it is possible to develop wavelets specifically for a particular application. These application specific wavelets iteratively adapt themselves towards a user defined criteria and are generally termed adaptive wavelets [10]. It has been demonstrated in supervised settings that adaptive wavelets–ones characteristic to the modelling

process, result in higher classification rates [4, 5, 14] and more accurate regression models [6].

Initially we describe the methods used: repeated measures analysis, adaptive wavelet transforms, and penalised discriminant analysis. Then we describe the method using of the adaptive wavelet transform in conjunction with a penalised discriminant mapping to analyse a repeated measures experimental design. Finally, we illustrate the method using a case study borrowed from the viticulture industry.

2. Theory

2.1. Analysis of repeated multivariate measures

Measurements on a set of p variables made at several occasions on the same experimental unit leads to repeated multivariate measures or longitudinal data. Analysis of these data needs special care since measurements made on the same units are likely to be correlated in time. A typical set of repeated measures is usually taken on $n = n_1 + \cdots + n_g$, individuals in g groups over t time points. The problems of interest are to test for the (i) *time effect* (ii) *group effect*, and (iii) *the effect of interaction between time and group*. Several approaches to analyse these data exists in the literature. A brief review follows.

Let $\mathbf{y}_{ijk}$, where $k = 1, ..., t$; $j = 1, ..., n_i$; $i = 1, ..., g$; be a $p \times 1$ vector of measurements on the jth individual measurement in the ith group on the kth time point and $\mathbf{y}_{ij} = (\mathbf{y}'_{ij1}, ..., \mathbf{y}'_{ijt})'$. Then $\mathbf{y}_{ij}$ is $pt \times 1$ random observation vector corresponding to the jth individual in the ith group. Let $\text{cov}(\mathbf{y}_{ij}) = \mathbf{\Omega}$, for $j = 1, ..., n_i$; $i = 1, ..., g$, where $\mathbf{\Omega}$ is a positive definite matrix. Using a multivariate linear model of the form $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$, where the $n \times pt$ matrix $\mathbf{Y}$ is the observation matrix by taking each $\mathbf{y}'_{ij}$ in a row, $\mathbf{X}$ is the $n \times k$ design matrix, $\mathbf{B}$ is the $k \times pt$ matrix of unknown parameters and, assuming rows of $\mathbf{E}$ independently follow multivariate normal distribution with a zero mean vector and covariance matrix $\mathbf{\Omega}$, any linear hypothesis about the effect of time, groups or interactions can be formulated in the form of a general linear

hypothesis H_0 : **LBM** = **0**, for known full rank matrices **L** and **M**. This approach for analysing the repeated multivariate measures, known as doubly multivariate measures (DMM) analysis, is commonly adopted in practice [12].

An alternative approach taken to analyse the repeated multivariate measures data is the multivariate mixed model (MMM) analysis, the path taken in this paper. Consider a mixed effects MANOVA model with the effects of the subjects (experimental units) within a group being random. Then the MANOVA table can be given as in Table 1.

Table 1. MANOVA table for mixed effects model

Source	D.O.F	SS&CP	Distribution under H_0
Between Groups			
Groups	$g-1$	$\mathbf{Q}_1$	$W_p(g-1, \mathbf{\Omega})$
Individuals	$n-g$	$\mathbf{Q}_2$	$W_p(n-g, \mathbf{\Omega})$
Within Groups			
Time	$t-1$	$\mathbf{Q}_3$	$W_p(t-1, \mathbf{\Omega})$
Time*Groups	$(g-1)(t-1)$	$\mathbf{Q}_4$	$W_p((t-1)(g-1), \mathbf{\Omega})$
Error	$(t-1)(n-g)$	$\mathbf{Q}_5$	$W_p((t-1)(n-g), \mathbf{\Omega})$
Total	$nt-1$	$\mathbf{Y}\left(\mathbf{I}_{nt} - \frac{1}{nt}\mathbf{J}_{nt}\right)\mathbf{Y}'$	

Here $\mathbf{\Omega}$ is the variance covariance matrix of **Y**. The matrix quadratic forms $\mathbf{Q}_1, \ldots, \mathbf{Q}_5$ are

$$\mathbf{Q}_1 = t\sum_{i=1}^{g} n_i(\bar{\mathbf{y}}_{i..} - \bar{\mathbf{y}}_{...})(\bar{\mathbf{y}}_{i..} - \bar{\mathbf{y}}_{...})' = \mathbf{Y}'\mathbf{A}_1\mathbf{Y}$$

$$\mathbf{Q}_2 = t\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\bar{\mathbf{y}}_{ij.} - \bar{\mathbf{y}}_{i..})(\bar{\mathbf{y}}_{ij.} - \bar{\mathbf{y}}_{i..})' = \mathbf{Y}'\mathbf{A}_2\mathbf{Y}$$

$$\mathbf{Q}_3 = n\sum_{k=1}^{t} t(\bar{\mathbf{y}}_{..k} - \bar{\mathbf{y}}_{...})(\bar{\mathbf{y}}_{..k} - \bar{\mathbf{y}}_{...})' = \mathbf{Y}'\mathbf{A}_3\mathbf{Y}$$

$$\mathbf{Q}_4 = \sum_{i=1}^{g} n_i \sum_{k=1}^{t} (\overline{\mathbf{y}}_{i.k} - \overline{\mathbf{y}}_{i..} - \overline{\mathbf{y}}_{..k} + \overline{\mathbf{y}}_{...})(\overline{\mathbf{y}}_{i.k} - \overline{\mathbf{y}}_{i..} - \overline{\mathbf{y}}_{..k} + \overline{\mathbf{y}}_{...})' = \mathbf{Y}'\mathbf{A}_4\mathbf{Y}$$

$$\mathbf{Q}_5 = \sum_{i=1}^{g} \sum_{j=1}^{n_i} \sum_{k=1}^{t} (\overline{\mathbf{y}}_{ijk} - \overline{\mathbf{y}}_{ij.} - \overline{\mathbf{y}}_{i.k} + \overline{\mathbf{y}}_{i..})(\overline{\mathbf{y}}_{ijk} - \overline{\mathbf{y}}_{ij.} - \overline{\mathbf{y}}_{i.k} + \overline{\mathbf{y}}_{i..})' = \mathbf{Y}'\mathbf{A}_5\mathbf{Y}$$

with the appropriate choice of symmetric matrices $\mathbf{A}_1$, ..., $\mathbf{A}_5$ of order $nt \times nt$ and with the usual notations for the sample averages. The matrices $\mathbf{A}_1$, ..., $\mathbf{A}_5$ are derived from the known design matrix $\mathbf{X}$ [7]. The matrix quadratic forms $\mathbf{Q}_1$, ..., $\mathbf{Q}_5$ are independent of each other and under the appropriate null hypothesis each has a scale multiple of a Wishart distribution, $W_p(df, \mathbf{\Omega})$, with a shape parameter $\mathbf{\Omega}$ and an appropriate degrees of freedom [15]. For both DMM and MMM, the covariance matrix $\mathbf{\Omega}$ can be estimated using a variety of methods.

The simplest method of estimating $\mathbf{\Omega}$ is the general covariance structure, whose only constrains are positive definiteness and symmetry. This method can reliably be applied when n is large, which is often not the case. Typically, $\mathbf{\Omega}$ is parameterized into simpler structures. The most common of these are the (i) multivariate compound symmetric structure, (ii) auto-regressive and exponential multivariate compound symmetric structure, and (iii) general Kronecker product [17, 20].

Of these three types of parameterizations of $\mathbf{\Omega}$, we discuss the Kronecker product since it has the greatest flexibility with respect to the time sampling component (the other two models generally require regularly sampled time intervals).

The structure of the Kronecker product for $\mathbf{\Omega}$ is given as:

$$\mathbf{\Omega} = \mathbf{V} \otimes \mathbf{\Sigma}, \tag{1}$$

where $\mathbf{V}$ and $\mathbf{\Sigma}$ are $t \times t$ and $p \times p$ are positive definite, symmetric matrices, respectively. This parameterization separates $\mathbf{\Omega}$ into a time and variables partition. This structure has several advantages over the general covariance structure. First, it is well known that the correlation structure of the repeated measures usually has a simpler structure as

opposed to the general structure [17]. Secondly, the number of unknown parameters of the covariance matrix for the Kronecker product is much less, $[t(t+1)/2 + p(p+1)/2]$ as compared $[pt\,(pt+1)/2]$ for the general case.

Given these advantages, the use of the Kronecker product structure would seem to be the logical choice for DMM and MMM. However, before using this separated structure, the likelihood ratio test for separability [17] is used to verify the validity of equation (1).

2.2. Penalised discriminant mapping (PDM)

Penalised discriminant mapping [24] is an extension of Fisher's linear discriminant analysis (LDA) which aims to find linear combinations of the variables, $\boldsymbol{b}$, that best separate the g different groups within the dataset such that the between group variability is maximised as much as possible relative to the within group variability. LDA assumes that the data, $\mathbf{X}$, are drawn from g groups with proportions $\pi_1, ..., \pi_G$ that have K dimensional mean vectors, $\overline{\mathbf{x}}_j$, $j = 1, ..., g$ and a common within group covariance, $\boldsymbol{\Omega}_w$. Specifically, LDA finds $\boldsymbol{b} \in \Re^K$ with $\boldsymbol{b}^T\boldsymbol{\Omega}_w\boldsymbol{b} = 1$ such that $f = \sum_{j=1}^{G} \pi_j(\boldsymbol{b}^T\overline{\mathbf{x}}_j - \boldsymbol{b}^T\overline{\mathbf{x}})^2$ is maximised. Here $\overline{\mathbf{x}} = \sum_j \pi_j\overline{\mathbf{x}}_j$ is the overall population mean vector. Maximising f is identical to maximising the ratio $g = \boldsymbol{b}^T\boldsymbol{\Omega}_B\boldsymbol{b}/\boldsymbol{b}^T\boldsymbol{\Omega}_w\boldsymbol{b}$ under the constraint $\boldsymbol{b}^T\boldsymbol{\Omega}_w\boldsymbol{b} = 1$. Differentiation leads to the eigensystem: $\boldsymbol{\Omega}_w^{-1}\boldsymbol{\Omega}_B\boldsymbol{b} = g\boldsymbol{b}$. In this way we can see that the eigenvectors of $\boldsymbol{\Omega}_w^{-1}\boldsymbol{\Omega}_B$ lead to the discriminant space, where $\boldsymbol{\Omega}_B$ is the between group covariance matrix.

In many NIR spectra situations, $\boldsymbol{\Omega}_w$ is near singular due to the high correlations between adjacent wavelengths (variables), thus the eigenvalues of $\boldsymbol{\Omega}_w^{-1}\boldsymbol{\Omega}_B$ cannot easily be computed. To overcome this near singularity, $\boldsymbol{\Omega}_w$ is replaced with $\boldsymbol{\Omega}'_w = \boldsymbol{\Omega}_w + \mathbf{K}$, where $\mathbf{K}$ is a K by K matrix such that $\boldsymbol{b}^T\mathbf{K}\boldsymbol{b}$ is large for undesirable $\boldsymbol{b}$. This $\mathbf{K}$ is the central idea in PDM, where $\mathbf{K}$ penalises the $\boldsymbol{b}$'s. We refer the reader to [24] for a detailed description of $\mathbf{K}$.

2.3. Discrete wavelet transform (DWT)

The discrete wavelet transform (DWT) [23] like the Fourier transform, can be used to reformulate a spectrum into an alternative "feature space", by mapping the spectrum onto an analysing function. In Fourier analysis, the analysing functions are the set of sine functions (spectra are mapped onto "frequency space"), where as for the DWT wavelets are the analysing functions (spectra are mapped onto a "wavelet space"). The DWT is given by:

$$x(t) = \sum_{j=1}^{l} \sum_{k=0}^{2^l} c_{j,k} \psi_{j,k}, \tag{2}$$

where $\psi_{0,0}$ is the father wavelet, from which all the other wavelets $\psi_{j,k}$ are derived, $x(t)$ is the spectrum and $c_{j,k}$ is the wavelet coefficient calculated by the inner product between $x(t)$ and $\psi_{j,k}$.

$$c_{j,k} = \langle x(t) | \psi_{j,k} \rangle. \tag{3}$$

Unlike Fourier analysis, there are many types of analysis functions (wavelets) that can be used for the DWT-each resulting in different wavelet coefficients (mapped features), where typical (standard) wavelets used are Daubechies, Symlets or Coiflets. Since we do not know which wavelets will result in the best feature extraction a priori for classification, this paper will use Pollen's adaptive wavelets [11, 23] to extract features.

An advantage of the Pollen adaptive wavelets, is that the wavelet can be parameterized into $q+1$ normalized vectors $\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_q$ and $\mathbf{v}$; where $q \in Z^+$ is a smoothness parameter for the resulting wavelet. This means that we can assess the "fitness" of the wavelet as a function of the normalized vectors, which can then be iteratively updated to achieve a high "fitness". In this study, we define the fitness as the ability to discriminate between the various homogenizers, varieties and storage combinations. To achieve this, we introduce a fitness function based on the wavelet coefficients from the DWT and the experimental design.

The fitness function is defined as:

$$f(u_1, ..., u_1, \nu) = \frac{|\mathbf{\Omega}_B|}{|\mathbf{\Omega}_B| + |\mathbf{\Omega}_w|}. \quad (4)$$

The Pollen adaptive wavelets can be summarized in the following steps:

(1) Define the integer values for m and q.

(2) Initialize the normalized vectors $\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_q$ and $\mathbf{v}$.

(3) Perform the DWT and evaluate the performance of the wavelet with equation (4).

(4) Iteratively update $\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_q$ and $\mathbf{v}$ until a convergence criterion is met.

In this study, $\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_q$ and $\mathbf{v}$ are initially assigned elements from the uniform distribution, which in previous supervised studies [4, 14] as shown to converge (tolerance of $f(u_1, ..., u_1, \nu) \leq 10^{-6}$) for fitness functions similar to that used in this study (equation (4)). For a comprehensive account of the theory of the Pollen Factorization, the reader is referred to [11].

3. Experimental Design

3.1. Data

The data comprises of the reflectance NIR spectra of 234 red grape homogenates; where the grapes were sampled using a nested repeated measures design (Figure 1). Red grapes of three varieties (A, B, C) were grown in two regions (R1 and R2) with varieties A and B replicated (in different vineyards) twice in region R1 and thrice in R2, and variety C replicated once in each region. Further more, each variety/region/ vineyard combination was replicate four times. The grapes where frozen for 0, 1, 3, 6 and 12 months before being homogenized then measured using a FOSS NIRSystems6500 instrument (400-2500nm). The spectra were truncated to 400-2448nm and then normalized via the SNV transform [1]. Figure 2 shows the spectra in the visible and NIR range of red grape homogenates.

$$\frac{\begin{pmatrix} Region(F) \\ | \\ Vineyard(R)^{*} \end{pmatrix} \times Time(F) \times Variety(F)}{| \\ Replication(R)}$$

Figure 1. Experimental design. Fixed effects and random effects are indicated in parenthesis as (*F*) and (*R*) respectively. Nested factors are indicated by | and are nested within the factor directly about. The astrix denotes the level the experimental unit that is repeatedly measured over time.

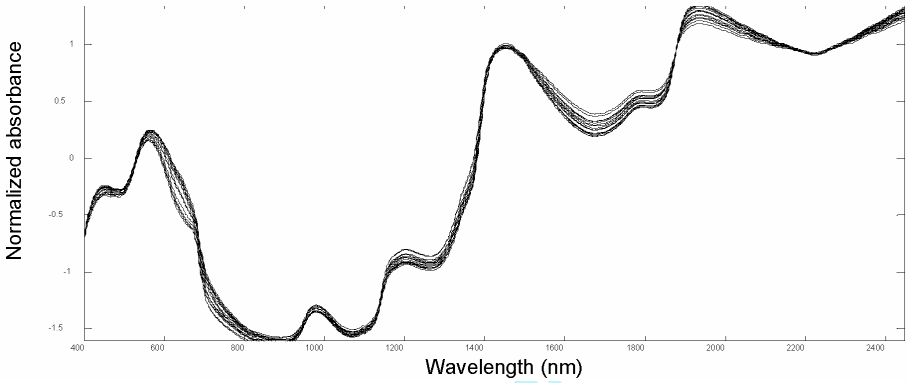


Figure 2. Sample spectra in the visible and NIR range of red grape homogenates.

3.2. Analysis

Analysis of the NIR spectra was performed via a three step process:

1. Extraction of NIR features via the adaptive wavelet (AWPT) transform.

2. Projection of the NIR wavelet features (from step 1) onto a five dimensional space using PDM. Five dimensions were chosen as over 99% of the between group variation was represented within the first five dimensions.

3. Analysis of the PDM scores (from step 2) via a multivariate mixed model (repeated multivariate measures) using the experimental design shown in Figure 1.

4. Results and Discussion

Using a multivariate mixed model (MMM) on the spectral scores from the adaptive discrete wavelet transform/penalised discriminate analysis (ADWT/PDM) (mapping), the region, variety and time factors significantly affect the spectra while the vineyard factor did not. Consequently, the MMM was re-analysed without the vineyard factor (Table 2). In the revised ADWT/PDM MMM, interaction between variety and region factors significantly affect the between subjects while the interaction between time, variety and region significantly influences the within subject effects.

To determine if the ADWT contributes to the ADWT/PDM MMM analysis, a PDM MMM was performed as a comparison. As with the ADWT/PDM MMM, the vineyard factor did not affect the spectra and was removed from further analysis (Table 3). A comparison between the ADWT/PDM and PDM MMMs' can be made by comparing Tables 2 and 3. For the PDM MMM, only variety is significant as a between subjects. While the time∗variety interaction affect is the only significant within subject factor. Notably, for the PDM MMM, the region factor does not contribute a significant role where it does in the ADWT/PDM MMM. Consequently, in this case study, ADWT/PDM method identified more of the underlying experimental design factors than PDM alone.

Regions within the spectrum, identifiable with the significant effects found in the ADWT/PDM MMM, are highlighted using: (i) scatter plots, (ii) univariate mixed models and, (iii) inversion of the PDM axes via the inverse discrete wavelet transform. Scatter-plots of the ADWT/PDM scores reveal clustering effects consistent with the experimental design (Figure 3) where: PDM axis 1 (PDA1) is largely dominated by variety and variety∗region affects (Figure 3a), PDA2 is predominated by time effects – with longer storage times resulting in lower PDA2 scores (Figure 3b) and, PDA3 scores are substantially influenced by region and region∗variety effects (Figure 3c).

Univariate mixed models on each of the PDM axes, highlights the relative magnitude of the factor affects of the various factors form the experimental design – summarized in Table 4. Dominant effects for each

PDM axis, highlighted by the univariate mixed models, are similar to those identified by visual inspection of the PDM scatter-plots. So in correlating experimental affects with PDM axes, both scatter-plots and univariate mixed models can be used.

The correlation of PDM axes with experimental affects makes it possible to identify regions in the spectrum related to the experimental design. Regions in the spectrum related to the PDM axes are found by transforming the PDM factor loadings using the inverse discrete wavelet transform (Figure 4). Hence, experimental affects can be identified with regions within the spectrum. For instance, the regions 295-975nm and 1275-1325nm (Figure 4a) can be attributed to variety and variety∗region affects, since these factors predominate PDA1.

Table 2. MANOVA table for final NIR adaptive wavelet PDM linear mixed model

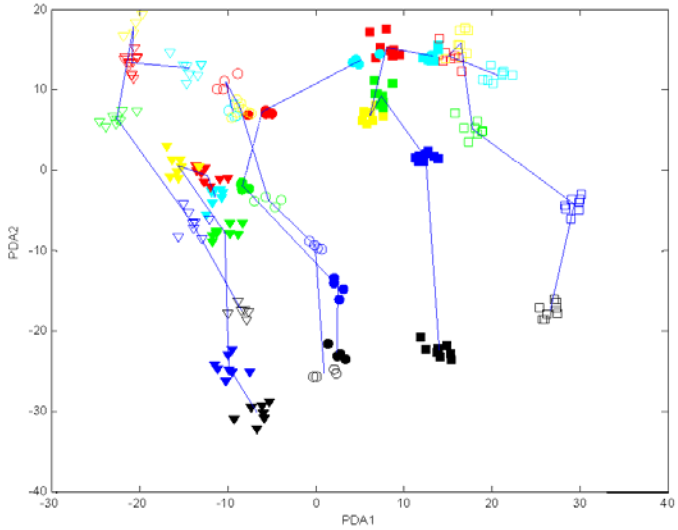
Effect	Wilks' Lambda Value	F	Hypothesis df	Error df	Sig.
Between Subjects					
Intercept	.056	77.554	5.000	23.000	.000
Variety	.000	1272.384	10.000	46.000	.000
Region	.003	1571.905	5.000	23.000	.000
Variety* Region	.002	2515.477	5.000	23.000	.000
Within Subjects					
Time	.000	4621.284	25.000	3.000	.000
time* Variety	.000	52.168	50.000	6.000	.000
time* Region	.006	21.554	25.000	3.000	.014
time* Variety* Region	.002	58.362	25.000	3.000	.003

Table 3. MANOVA table for NIR PDM linear mixed model

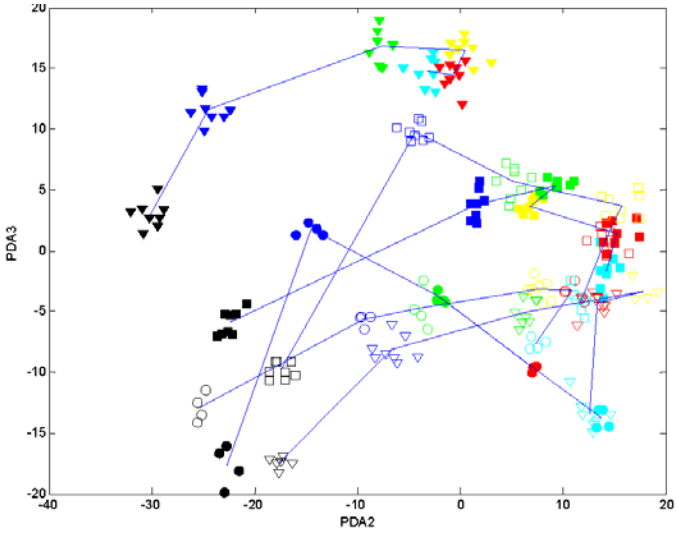
Effect	Wilks' Lambda Value	F	Hypothesis df	Error df	Sig.
Between Subjects					
Intercept	.041	143.724	5.000	23.000	.000
Variety	.000	1020.277	10.000	46.000	.000
Region	.675	1.345	10.000	23.000	.227
Variety* Region	.793	0.761	10.000	23.000	.665
Within Subjects					
Time	.000	2216.599	25.000	3.000	.000
time* Variety	.001	11.473	50.000	6.000	.000
time* Region	.078	1.134	50.000	3.000	.384
time* Variety* Region	.058	1.383	50.000	3.000	.206

Table 4. Main experimental design effects associated by PDM axis. Mean squared error (MSE) values shown in brackets

AWPT PDM axis	Associated experimental effects
PDA1	Variety(18400), Variety∗Region(2656)
PDA2	Time(4688), Region(2248), Variety(1619), Variety∗Region(2222)
PDA3	Region(4426), Variety∗Region(4380)



(a)



(b)

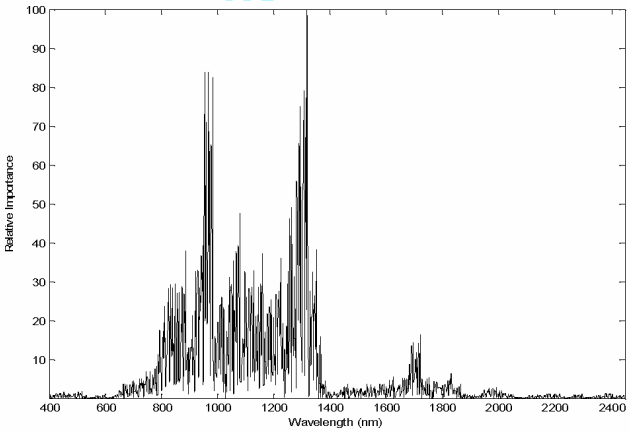
Figure 3. AWPT PDM scores for the (a) discriminant functions PDA1 and PDA2 and (b) discriminant functions PDA2 and PDA3. Legend: Time; Fresh = cyan, Overnight = red, 1 month = yellow, 3 months = green, 6 months = blue, 12 months = black, Region: Solid fill = R1, Empty fill = R2; Variety: A = o, B = ▼, C = ■. Blue lines represent the sample mean change in time.

5. Conclusion

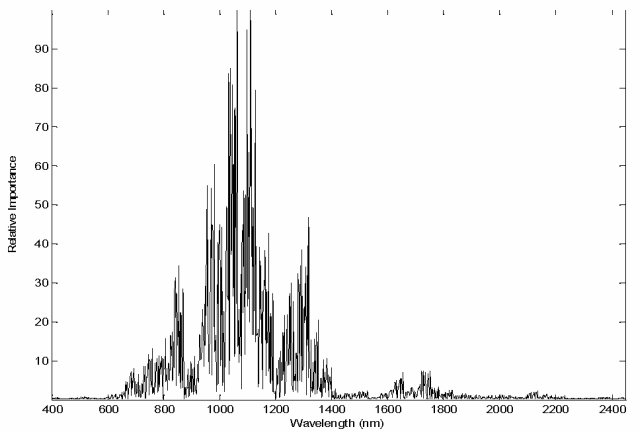
In this paper, we have illustrated firstly how a repeated measures design can be implemented using near infra-red spectra and secondly, with pre-processing via the adaptive wavelet transform, the analysis of the repeated measures design is more sensitive to the between subjects and within effects.

In this study of red grape homogenates, by using the ADWT/PDM MMM, we were able to identify the main effects and interactions resulting from (a) the flow of time (b) the variety of the grape and (c) the region from which it grew. However, in the analysis where the AWT is absent, the region effect was found to be insignificant. This demonstrates the effectiveness of the AWT ability to elucidate information form the spectral pertinent to repeated measures experimental design. Additional to this, the particular wavelengths attributed to the repeated multivariate measures with the AWT were identified (600 to 1400nm).

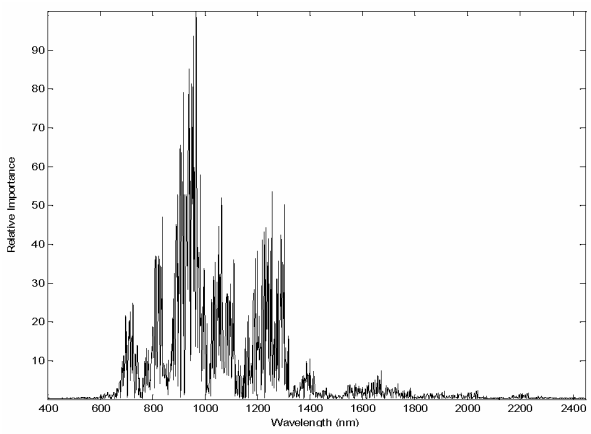
By using repeated multivariate measures, we can identify if and how the spectra are being affected by the flow of time or by other measurable factors such as variety effect. Furthermore, the information contained within the spectra can be better utilized through the application of adaptive wavelets, generated specifically to extract the information from the spectra pertinent to the multivariate analysis of the experimental design.



(a)



(b)



(c)

Figure 4. Relative wavelength importance for (a) PDA1 (b) PDA2 and (c) PDA3.

References

[1] R. Barnes, M. Dhanoa and S. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Applied Spectroscopy 43 (1989), 772-777.

[2] R. Bro, PARAFAC, tutorial and applications, Chemometrics and Intelligent Laboratory Systems 38(2) (1997), 149-171.

[3] J. D. Dawson, Sample size calculations based on slopes and other summary statistics, Biometrics 54(1) (1998), 323-330.

[4] D. Donald, D. Coomans, Y. Everingham, D. Cozzolino, M. Gishen and T. Hancock, Adaptive wavelet modelling of a nested 3 factor experimental design in NIR chemometrics, Chemometrics and Intelligent Laboratory Systems (in Press).

[5] D. Donald, Y. Everingham and D. Coomans, Integrated wavelet principal component mapping for unsupervised clustering on near infra-red spectra, Chemometrics and Intelligent Laboratory Systems 77(1-2) (2005), 32-42.

[6] R. K. H. Galvao, G. E. Jose, H. A. Dantas, M. C. U. Araujo, E. C. da Silva, H. M. Paiva, T. C. B. Saldanha and E. S. O. N. de Souza, Optimal wavelet filter construction using *X* and *Y* data, Chemometrics and Intelligent Laboratory Systems 70(1-2) (2004), 1-10.

[7] S. Geisser and S. Greenhouse, An extension of Box's results on the use of the *F* distribution in multivariate analysis, Ann. Math. Statist. 29 (1958), 885-891.

[8] P. Geladi and P. Aberg, Three-way modelling of a batch organic synthesis process monitored by near infrared spectroscopy, J. Near Infrared Spectroscopy 9(1) (2001), 1-9.

[9] P. Geladi, H. Bergner and L. Ringqvist, From experimental design to images to particle size histograms to multiway analysis, An example of peat dewatering, J. Chemometrics 14(3) (2000), 197-211.

[10] J. Kaustsky and R. Turcajova, Pollen product factorization and construction of higher multiplicity wavelets, Numer. Algorithms 8 (1994), 27-54.

[11] J. Kautsky and R. Turcajova, Pollen product factorization and construction of higher multiplicity wavelets, Linear Algebra Appl. 222 (1994), 241-260.

[12] P. R. Krishnaiah, Multivariate Measurements of Repeated Measurements, Handbook of Statistics, Vol. 1, New York, 1980.

[13] Y. Mallet, D. Coomans and O. deVel, Recent developments in discriminant analysis on high dimensional spectral data, Chemometrics and Intelligent Laboratory Systems 35(2) (1996), 157-173.

[14] Y. Mallet, D. Coomans, J. Kautsky and O. DeVel, Classification using adaptive wavelets for feature extraction, IEEE Transactions on Pattern Analysis and Machine Intelligence 19(10) (1997), 1058-1066.

[15] T. Mathew, MANOVA in the multivariate components of variance model, J. Multivariate Anal. 29(1) (1989), 30-38.

[16] Maths Works, MATLAB Wavlet Toolbox, 1998.

[17] D. N. Naik and S. S. Rao, Analysis of multivariate repeated measures data with a Kronecker product structured covariance matrix, J. Appl. Stat. 28(1) (2001), 91-105.

[18] A. C. Rencher, The contribution of individual variables to Hotelling's T^2, Wilks' λ, and R^2, Biometrics 49(2) (1993), 479-489.

[19] R. Tauler, Multivariate curve resolution applied to second order data, Chemometrics and Intelligent Laboratory Systems 30(1) (1995), 133-146.

[20] D. Thomas, Univariate repeated measures techniques applied to multivariate data, Psychometrica 48 (1983), 451-464.

[21] E. F. Vonesh and M. A. Schork, Sample sizes in the multivariate-analysis of repeated measurements, Biometrics 42(3) (1986), 601-610.

[22] B. Walczak, ed., Wavelets in Chemistry, Elsevier, Amsterdam, 2000, pp. 177-202.

[23] B. Walczak, ed., Wavelets in Chemistry, Elsevier, Amsterdam, 2000.

[24] B. Yu, I. M. Ostland, P. Gong and R. L. Pu, Penalised discriminant analysis of in situ hyperspectral data for conifer species recognition, IEEE Transactions on Geoscience and Remote Sensing 37(5) (1999), 2569-2577.