rich2/ccj-cjc/ccj-cjc/ccj99914/ccj1054d14z | xppws | S=3 | 6/5/18 | 8:18 | 4/Color Fig: F1-F8 | Art: **cjc-2017-0730** | Input-1st disk, 2nd ??
Pagination not final (cite DOI) / Pagination provisoire (citer le DOI)

NRC Research Press

ARTICLE

Implications of measurement error structure on the visualization of multivariate chemical data: hazards and alternatives

Peter D. Wentzell, Chelsi C. Wicks, Jez W.B. Braga, Liz F. Soares, Tereza C.M. Pastore, Vera T.R. Coradin, and Fabrice Davrieux

Abstract: The analysis of multivariate chemical data is commonplace in fields ranging from metabolomics to forensic classification. Many of these studies rely on exploratory visualization methods that represent the multidimensional data in spaces of lower dimensionality, such as hierarchical cluster analysis or principal components analysis. However, such methods rely on assumptions of independent measurement errors with uniform variance and can fail to reveal important information when these assumptions are violated, as they often are for chemical data. This work demonstrates how two alternative methods, maximum likelihood principal components analysis (MLPCA) and projection pursuit analysis (PPA), can reveal chemical information hidden from more traditional techniques. Experimental data to compare different methods consists of near-infrared reflectance spectra from 108 samples of wood that are derived from four different species of Brazilian trees. The measurement error characteristics of the spectra are examined and it is shown that, by incorporating measurement error information into the data analysis (through MLPCA) or using alternative projection criteria (i.e., PPA), samples can be separated by species. These techniques are proposed as powerful tools for multivariate data analysis in chemistry.

Key words: chemometrics, exploratory data analysis, near-infrared spectroscopy, measurement errors, projection pursuit.

Résumé : L'analyse de données chimiques à plusieurs variables est couramment employée dans divers domaines allant de la métabolomique à la classification criminalistique. Bon nombre de ces études reposent sur des méthodes exploratoires de représentation des données multidimensionnelles dans des espaces de faible dimensionnalité, comme la classification hiérarchique ou l'analyse en composantes principales. Cependant, de telles méthodes ne sont fiables que si l'on admet que les erreurs de mesure sont indépendantes et que la variance est uniforme. Elles peuvent donc faillir à mettre en lumière d'importantes informations si ces suppositions devaient être erronées, ce qui est souvent le cas des données en chimie. Ces travaux démontrent comment deux autres méthodes, l'analyse en composantes principales à probabilité maximale (ACPPM) et la poursuite de projection (PP), permettent révéler des informations chimiques omises par des techniques plus traditionnelles. Afin de comparer expérimentalement les différentes méthodes, nous avons recueilli les spectres de réflectance dans le proche infrarouge de 108 échantillons de bois provenant de quatre espèces d'arbres brésiliens. Nous avons examiné les caractéristiques des erreurs de mesure des spectres et nous avons observé que, en incorporant ces informations d'erreur de mesure dans l'analyse des données (par ACPPM) ou en utilisant d'autres critères de projection (c.-à-d. la PP), nous pouvons distinguer les échantillons selon l'espèce d'arbre. Ces techniques peuvent constituer de puissant outils pour l'analyse multidimensionnelle de données en chimie. [Traduit par la Rédaction]

Mots-clés : chimiométrie, analyse exploratoire des données, spectroscopie proche infrarouge, erreurs de mesure, poursuite de projection.

Introduction

Modern chemical measurements are often multivariate in nature, taking the form of vectors (e.g., mass spectra, NMR spectra, chromatograms, lists of protein abundances), matrices (e.g., two-dimensional NMR spectra, LC–MS data, hyperspectral images), or higher order tensors. To understand the complex relationships among different sets of measurements (e.g., samples), simplification is often sought through visualization of the high-dimensional data in low-dimensional spaces, sometimes referred to as exploratory data analysis. Two methods that are widely used for this purpose are hierarchical cluster analysis (HCA)[1–3] and principal components analysis (PCA).[1–5] HCA is a nonlinear mapping technique that renders the information about the distance among objects (samples) in a high-dimensional space into a two-dimensional representation known as a dendrogram. Often these are used in conjunction with so-called heat maps to display the characteristics of variables, such as the expression levels of genes or proteins. PCA is a linear projection technique that projects the multivariate data into a two- or three-dimensional space while preserving information about the relationships among objects. These projections, commonly known as scores plots, are often used to determine

Received 30 November 2017. Accepted 6 May 2018.

P.D. Wentzell and C.C. Wicks. Trace Analysis Research Centre, Department of Chemistry, Dalhousie University, P.O. Box 15000, Halifax, NS B3H 4R2, Canada.
J.W.B. Braga and L.F. Soares. Chemistry Institute, University of Brasilia, Brasília, 72910-000, Brasilia, DF, Brasil.
T.C.M. Pastore and V.T.R. Coradin. Forest Products Laboratory, Brazilian Forest Service, 70818-970, Brasilia, DF, Brasil.
F. Davrieux. French Agricultural Research Center for International Development, CIRAD-UMR Qualisud, F-34398, Montpellier Cedex 5, France.
Corresponding author: Peter D. Wentzell (email: peter.wentzell@dal.ca).

This paper is part of a Special Issue to celebrate the 200th anniversary of Dalhousie University in Halifax, Nova Scotia, Canada.

which samples group together and can therefore be considered to represent a cluster or class.

Both HCA and PCA are extensively used in chemical applications that include proteomics,[6,7] metabolomics,[8–10] food science,[11,12] forensics,[13–15] medical diagnostics,[16,17] and threat detection.[18] An important goal of both techniques in these and other applications is to either identify or confirm groupings of samples that are consistent with external classifications that are based on other factors, such as disease state (medicine), geographic origin (food analysis), provenance (forensics), and biological species (chemotaxonomy). The widespread use of these tools is based, in part, on the fact that they are unsupervised methods, which means that the visualization uses no prior knowledge of the class structure. This is in contrast to supervised methods, such as partial least squares discriminant analysis (PLSDA),[19,20] that actively employ class information to build a model and therefore require careful validation to avoid overfitting. Because no class information is employed in HCA and PCA, they have gained acceptance as suitable methods for hypothesis confirmation where the key question is whether the data contain sufficient information to distinguish different groups of samples, especially when the number of samples is small and the number of variables is large. This is often a critical question in research and can determine whether a line of inquiry continues or is abandoned. This accounts for the pervasive application of these methods across all areas of chemistry.

Although HCA and PCA are powerful and useful techniques, they can be subject to serious limitations when applied to problems where the data do not meet certain criteria. HCA is based on the calculation of Euclidean distances among objects in higher dimensions, whereas PCA creates a subspace that maximizes the amount of variance retained in the data. Both of these methods are sensitive to the scale of the data, which means that variables that have a larger range will be weighted more heavily in mapping the high-dimensional data to lower dimensions, even if the information content is greater for variables with a smaller range. For example, a small mass spectral or NMR peak that contains important information for the separation of classes may be eclipsed by larger peaks with a variability that does not correlate with class separation, resulting in a projection that does not reveal the critical relationships. In some cases, this problem may be mitigated by appropriate pretreatment of the data (e.g., variable scaling, log transformation); however, this may give rise to other problems.[2,21,22] For example, scaling of variables that are predominately associated with noise (e.g., baseline regions) increases their influence in the mapping process even though they have no relevance in classification. This problem is further exacerbated by complex measurement noise structures that may include non-uniform error variance among variables (referred to as heteroscedastic noise) or correlated errors.[23]

The principal weaknesses of HCA and PCA for unsupervised clustering with multivariate chemical data are (*i*) lack of a criterion to distinguish meaningful chemical variance in a data set from the noise variance, and (*ii*) a reliance on variance and distance metrics to develop interesting and useful projections of the data. In this paper, two alternative approaches are presented to address these shortcomings. The first approach is the use of maximum likelihood principal components analysis (MLPCA), which directly incorporates prior information about the measurement error variance into the decomposition process, thereby more effectively distinguishing the chemical variance from the noise variance.[23–25] The second approach employs a new implementation of an old idea known as projection pursuit analysis (PPA), which is not based on variance or distance metrics.[26–31] To demonstrate these methods, near-infrared (NIR) reflectance spectra, which exhibit a heteroscedastic and correlated noise structure, are employed to show how the new approaches provide superior clustering information.

Background

PCA and HCA

Because PCA and HCA are widely used techniques, only a brief description will be provided here to place them in the context of the lesser known methods, and the reader is referred to more detailed treatments in standard texts.[1,2,4] If the measurement data are represented by the matrix $\mathbf{X}$ ($m \times n$) consisting of n variables (measurement channels) for m objects (samples), then the PCA decomposition results in an orthogonal rotation of the original variable space such that the data matrix can be represented as

$$\mathbf{X} = \mathbf{TP} \quad (1)$$

where $\mathbf{T}$ ($m \times p$) is the scores matrix, which gives the coordinates of the objects in the new space, and the rows of $\mathbf{P}$ ($p \times n$) represent the eigenvectors (or loadings), which define the rotation of the original space (i.e., the linear combinations of the original variables giving rise to the new variables). The dimension p will be the smaller of m or n and defines the mathematical rank of $\mathbf{X}$. There are an infinite number of possible rotations of the original space; however, PCA provides the solution that maximizes the variance accounted for by each subsequent dimension (principal component or factor). The q-dimensional estimation of the data is given by

$$\hat{\mathbf{X}}_q = \mathbf{T}_q\mathbf{P}_q \quad (2)$$

where $q \leq p$, and $\mathbf{T}_q$ ($m \times q$) and $\mathbf{P}_q$ ($q \times n$) are the truncated scores and loadings matrices, respectively, consisting of the first q columns of $\mathbf{T}$ and the first q rows of $\mathbf{P}$. For a given q, the decomposition minimizes the sum of squared residuals, SSR_q:

$$SSR_q = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \hat{x}_{ij})^2 \quad (3)$$

where the notation "ij" indicates the measurement at row i and column j. Equivalently, this maximizes the amount of total variance retained in $\hat{\mathbf{X}}$. If q is chosen to be 2 or 3, the columns of $\mathbf{T}_q$ can be plotted against one another as a scores plot. Ideally, this yields an optimal visualization of the relationships among objects.

In HCA, the concept is to measure the distances among objects in the data set and group the objects (rows of $\mathbf{X}$) that are closest together. Starting with the same matrix, $\mathbf{X}$, the Euclidean distance between each pair of objects, i and j, is first calculated according to

$$d_{ij} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \quad (4)$$

This leads to a symmetric distance matrix, $\mathbf{D}$ ($m \times m$) with diagonal elements of zero. In the next step, the two objects with the shortest distance are identified and combined to form a new object that replaces the former objects, and a new distance matrix is calculated. Because the new object is a combination of the original objects, there is a variety of options (called linkage methods) to represent the new distance, such as using the average distance to the group or the distance to the nearest original object, but these will not be discussed in detail here. This process is then repeated, incrementally reducing the number of objects present at each iteration until only a single connection remains to be made. The hierarchy of these connections is finally displayed as a tree structure (a dendrogram) with the relationships between objects represented as a chain of branch points where the vertical height of each branch point represents the distance between the connected objects (a measure of "dissimilarity"). Those objects

(most often samples) emanating from a common branch point are considered to be most closely related (a cluster) with their similarity related to the height of the branch point.

Although they are different approaches, both PCA and HCA are based on measuring the squared differences among objects. These differences include both chemical variations and measurement noise. Both methods are designed to provide an optimal representation of the chemical variance when the measurement noise is independent and identically distributed with a normal distribution, often referred to as iid normal noise. This means that it is assumed that all of the measurements in the data set have the same error variance and there is no relationship among the errors for different variables (i.e., they are uncorrelated). While this is an implicit assumption in many data analysis methods (e.g., univariate regression), it is violated more often than not and can lead to suboptimal results.[23,32–34]

Measurement error structures

For univariate measurements, the uncertainty can be fully described by the error variance, σ^2, of the measurement. For multivariate measurements, it is necessary not only to provide the measurement variance for each variable, σ_i^2, but also to provide the covariance between measurement channels, σ_{ij}. Chemical measurement vectors are often heteroscedastic, meaning that different elements of the vector can exhibit different error variance. This non-uniform variance arises naturally from the measurement system.[32–35] For example, fundamental counting statistics, governed by the Poisson distribution, give rise to what is often referred to as shot noise, where the error standard deviation is proportional to the square root of the signal intensity. Such noise may be limiting in spectroscopic or mass spectrometric measurements where the signal amplitude is low. Proportional noise, where σ is proportional to the magnitude of the signal, is also commonly observed and typically associated with variations in a light source or ion source. Likewise, many measurement systems exhibit noise that is highly correlated. This includes baseline offset noise and multiplicative offset noise, the latter of which is typically the limiting noise source in NIR reflectance spectroscopy (vide infra), arising from variations in path length due to sample heterogeneity. Low frequency noise, also known as pink noise or 1/*f* noise, also falls into this category and is sometimes referred to as source flicker noise or drift noise in the context of analytical measurements.[35–41]

A common method to characterize multivariate measurement errors is the error covariance matrix (ECM).[23,24,32,33] If we consider a measurement vector, $\mathbf{x}$ $(1 \times n)$, which is an observation of a true (error-free) vector, $\boldsymbol{x}^o$, the error vector, $\boldsymbol{e}$, is defined as the difference between these vectors, $\boldsymbol{e} = \boldsymbol{x} - \boldsymbol{x}^o$. The error covariance between measurement channels i and j of the vector is defined as the expectation of the product of the corresponding errors:

$$\sigma_{ij} = E(\boldsymbol{e}_i \cdot \boldsymbol{e}_j) = \lim_{N\to\infty} \frac{\sum (x_i - x_i^o)(x_j - x_j^o)}{N} \quad (5)$$

Here the summation is over multiple realizations of measurement vector $\boldsymbol{x}$ and x_i and x_j are elements of that measurement vector. When $i = j$, the corresponding quantity is the error variance, signified as σ_i^2 rather than σ_{ii}. The collection of all of these error covariances is described by the ECM ($\boldsymbol{\Sigma}$) defined as the outer product of the expectation of the error vectors:

$$\boldsymbol{\Sigma} = E(\boldsymbol{e}^T \cdot \boldsymbol{e}) = E[(\boldsymbol{x} - \boldsymbol{x}^o)^T(\boldsymbol{x} - \boldsymbol{x}^o)] \quad (6)$$

The ECM is a symmetric $(n \times n)$ matrix, where the diagonal elements represent the error variance of each of the n variables and the off-diagonal represents the error covariances of the corresponding elements. The ECM is one of the most complete ways to describe the errors in a vectorial measurement with stationary characteristics. A related method is the error correlation matrix, **R**, which normalizes the off-diagonal elements by their corresponding standard deviations such that:

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \cdot \sigma_j} \quad (7)$$

This removes the effects of scale (diagonal elements are unity) and allows more direct visualization of correlation. Errors with $\rho_{ij} = 1$ are perfectly correlated.

In practice, the true measurement vector is unknown, so the experimental ECM is normally estimated by making replicate observations of the measurement vector and subtracting the sample mean vector ($\overline{\boldsymbol{x}}$). If r experimental replicates of the measurement vector (e.g., a spectrum) are made, the ECM is estimated as

$$\boldsymbol{\Sigma}_{expt} = \frac{1}{(r-1)} \sum_{k=1}^{r} (\boldsymbol{x}_k - \overline{\boldsymbol{x}})^T(\boldsymbol{x}_k - \overline{\boldsymbol{x}}) \quad (8)$$

It should be noted that the definition of the replicate is very important in this context, as it needs to capture all of the sources of variation one wishes to consider as measurement errors. Consequently, the ECM can be quite different depending on whether it is to include, for example, only technical replication or also sampling variability.

The ECM estimated by the replication procedure above is likely to be quite noisy itself when the number of replicates is relatively small[23,32] and therefore of limited practical utility. Two approaches are commonly used to improve the quality of the ECM. The first is to pool (average) the ECMs obtained for different measurement vectors, each with a limited number of replicates.[32] This results in an averaging effect that leads to a smoother ECM but makes the implicit assumption that measurement vectors for different samples have the same ECM. Although not strictly valid, this assumption is reasonable where measurements exhibit similar characteristics. The second approach is to develop an empirical model of the ECM.[32,33,42] For many kinds of measurements, the ECM can be represented using a model characteristic of that particular technique using a limited number of parameters. Where this can be done, the result is a smoother, more reliable ECM that can be calculated separately for each measurement vector.

Knowledge of the measurement error characteristics through the ECM is key to improving data analysis methods, because it allows better extraction of the chemical variance from the associated noise variance. By implicitly describing the information associated with each measurement, the ECM allows more optimal results to be obtained.

Maximum likelihood principal components analysis (MLPCA)

MLPCA was developed as a tool to provide better subspace estimation for multivariate data when assumptions of iid normal errors are no longer valid.[23–25] It can be viewed as a more generalized form of PCA in which the ECM is incorporated into the decomposition procedure to yield a more optimal solution. Rather than simply minimizing the residual variance of the truncated q-dimensional solution, MLPCA uses a weighted objective function that attempts to match the residual variance to the characteristics of the ECM for each measurement vector. The approach is analogous to using weighted least squares in univariate regression. The specific objective function used depends on the complexity of the error structure and there are six general categories, ranging in complexity from the trivial case of iid normal errors (where MLPCA and PCA are equivalent) to general error heterosce-

dasticity and correlation that can exist within both the rows and the columns of a data matrix. One of the most common implementations is where error correlation exists only within the rows of a data matrix. Under these conditions, the objective function to be minimized is defined as

$$S^2_{\text{obj}} = \sum_{i=1}^{m} (\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i) \quad (9)$$

where, $\boldsymbol{x}_i$ represents measurement vector i (row i of $\mathbf{X}$), $\hat{\boldsymbol{x}}_i$ is the estimate of the vector based on the MLPCA decomposition, and $\boldsymbol{\Sigma}_i$ is the ECM for the vector. In the general case, this objective function is optimized by an alternating least squares algorithm, but in the special case where $\boldsymbol{\Sigma}_i$ is the same for all row vectors, a direct solution can be obtained through rotation and scaling of the original data. Another difference between MLPCA and PCA is that, where PCA uses an orthogonal projection of the measurement vector onto the subspace to obtain the scores vector ($\boldsymbol{t}_q = (\mathbf{P}_q\, \mathbf{x}^{\mathrm{T}})^{\mathrm{T}}$), MLPCA employs a maximum likelihood projection:

$$\mathbf{t}_q = \mathbf{x}\boldsymbol{\Sigma}^{-1}\mathbf{P}_q^{\mathrm{T}}\left(\mathbf{P}_q\boldsymbol{\Sigma}^{-1}\mathbf{P}_q^{\mathrm{T}}\right)^{-1} \quad (10)$$

This oblique projection uses the information in the error covariance matrix to ensure that the projection uses the measurements in $\boldsymbol{x}$ that minimize the uncertainty in the low dimensional projection.

In principle, MLPCA should result in the optimal subspace estimation assuming that the intrinsic dimensionality of the data (also called the pseudorank, q) and the ECM are exactly known. In practice, q is often uncertain and only an estimated ECM is available, so this can limit the optimality of the solution. There can also be complications from rank deficiency of the ECM (which needs to be inverted) when it is estimated from a limited number of replicates, although there are strategies to address this.[25,43] Despite these limitations, however, MLPCA has demonstrated superior performance to PCA in a variety of applications ranging from multivariate calibration[44,45] to curve resolution.[46,47]

Projection pursuit analysis (PPA)

An inherent limitation of PCA and HCA is an assumption that the largest source of chemical variation in a data set is associated with the characteristic we are interested in, specifically, in the current context, the classification of samples into two or more groups. For example, in the detection of a disease state, it is hoped that the dominant source of difference is in a set of chemical compounds that are associated with the presentation of the disease, often referred to as biomarkers. However, the differences among these compounds may be obscured by other natural variations in the data set, resulting in an exploratory visualization that does not reveal clustering according to the anticipated characteristics. To overcome these limitations, it is necessary to use visualization methods that do not rely solely on variance metrics.

The concept of projection pursuit was first advanced nearly five decades ago, originally proposed by Kruskal[26] and named by Friedman and Tukey,[27] who further developed the idea. The strategy proposed is simply to look for linear projections of the multivariate data that are interesting based on a measure of "interestingness" as quantified by a projection index. Although the concept is simple, implementation has been complicated by (*i*) how to define "interesting", (*ii*) how to quantify a projection index consistent with this definition, and (*iii*) how to optimize the projection index once it is defined. A common criterion for interesting projections is those that exhibit non-Gaussian behavior, but this can be difficult to quantify, especially in chemical applications where the number of samples tends to be small. Consequently, PPA has not gained much traction in chemical research. Recently, however, PPA algorithms have been developed that are both effective and efficient for chemical data[28–31] and have been applied to problems that include forensics, metabolomics, and provenance.[28–30,48] These algorithms are based on the use of kurtosis, the fourth statistical moment, as the projection index. For univariate measurements, including projections into a one-dimensional space, the kurtosis can be defined as

$$\kappa = \frac{1/N \sum (x_i - \bar{x})^4}{\left[1/N \sum (x_i - \bar{x})^2\right]^2} \quad (11)$$

where κ is used to represent the kurtosis and the summations are over N measurements. Alternative definitions are also employed for multivariate kurtosis and the reader is referred to the original reference for a more complete description.[28] Kurtosis is a simple and useful measure for assessing the normality of data, taking on a value of 3 for a normal distribution, with higher values for heavily tailed distributions and lower values for flatter distributions. In particular, minimizing the kurtosis of projected data will emphasize naturally occurring clusters. Minimization of the projection index is a nonlinear optimization problem but can be performed efficiently through the use of a quasi-power method.[28]

Although a variety of PPA algorithms have been developed based on these principles, the most effective for clustering is often the stepwise univariate kurtosis PPA algorithm, which successively partitions the data into binary groups. Because it is not based on variance, PPA can often reveal clusters in the data that are not apparent with PCA and HCA, as will be demonstrated in this work.

Experimental

Computational aspects

All calculations were carried out within the MatLab programming environment (Mathworks, Natick, MA). Programs for carrying out MLPCA and PPA were written in-house and are available from the corresponding author, as are the data.

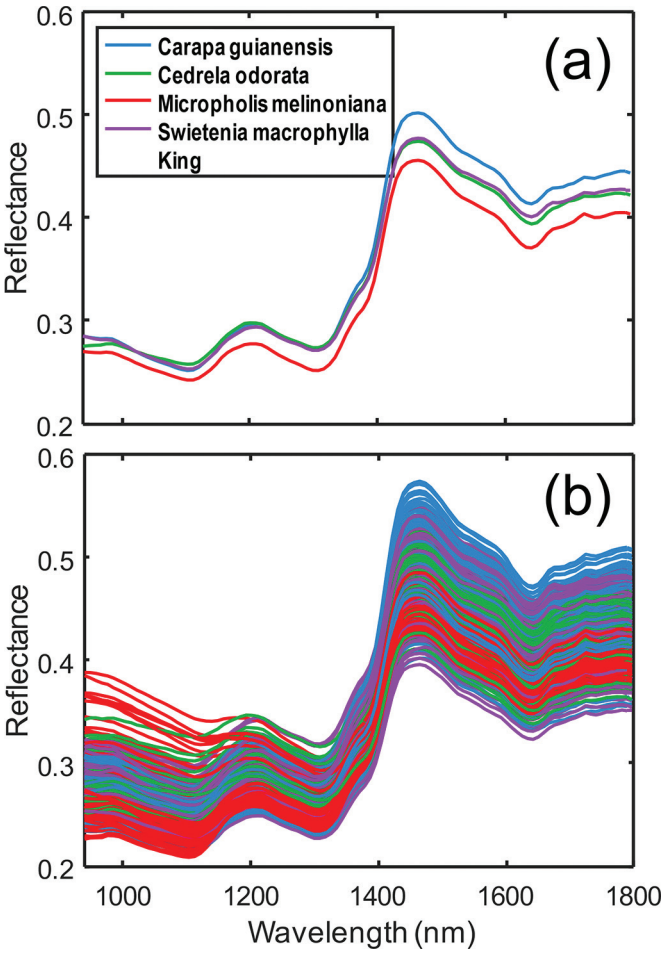
Species selection

The broad objective of this research, of which this study is a part, is the development of instrumental methods to distinguish wood species, with a particular emphasis on discriminating high value species such as mahogany. Species were selected based on the book "Similar woods to mahogany (*Swietenia macrophylla* King.): An illustrated key for anatomical field identification",[49] edited by the Brazilian Forest Service. From the 15 species listed, the three species that were the most difficult to distinguish, based on the appearance and macroscopic wood characteristics, were chosen for this study. These were *Carapa guianensis* Aubl., *Cedrela odorata* L., and *Micropholis melinoniana* Pierre, along with mahogany itself, *Swietenia macrophylla* King.

Sampling and sample preparation

Each sample of crabwood (*C. guianensis*), cedar (*C. odorata*), and curupixa (*M. melinoniana*) was obtained from an individual disk located at the base of a tree trunk. The samples of *S. macrophylla* were collected in authorized forestry exploitation areas in Para state, Brazil. Mahogany samples were obtained from tips of seized boards coming from the state of Mato Grosso, Brazil. Altogether, 108 solid samples were measured, 26 of crabwood, 28 of cedar, 29 of curupixa, and 25 of mahogany. Besides alleged species, all samples were identified by a wood anatomist of the Forest Products Laboratory in Brasilia, registered as FPBw in the Index Xylariorum.[50]

Fig. 1. Near-infrared reflectance spectra of wood samples. (*a*) Mean spectra of four species, as indicated in the legend. (*b*) Full set of 432 spectra. [Colour online.]



Samples were dried in open air conditions and cut into blocks of approximately 2 cm^3 with oriented faces according to wood growth directions. Surfaces were made uniform with 80 grit sandpaper.

Acquisition of spectra

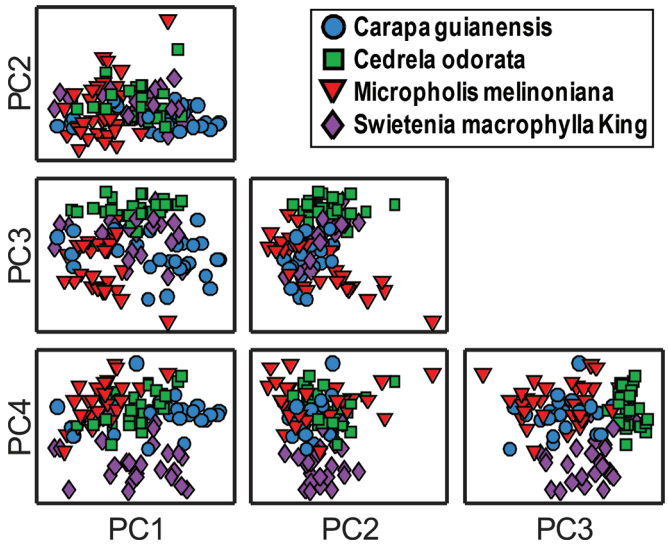
Samples were measured on a handheld spectrometer, Phazir RX (Polychromix). Four replicate spectra were obtained for each sample, two on each radial face, measured on distinct spots, resulting in a total of 432 spectra. Spectra were measured in the diffuse reflectance mode between 939.5 and 1796.6 nm with 9 nm of resolution. Resultant spectra (Fig. 1) consisted of 100 data points per spectrum and were converted to log(1/*R*) scale for the data analysis. Figure 1*a* shows the mean spectrum for each of the four species, and Fig. 1*b* shows all 432 spectra, with each of the replicates displayed individually. The data used in this study are provided in the Supplementary data.

Results and discussion

PCA and HCA of NIR spectra

It is clear from Fig. 1 that spectra of the four species exhibit a strong similarity and that the variation between individual samples is quite large, making the discrimination of the four classes a challenging problem. To determine if the usual data visualization methods would be able to distinguish the classes, PCA and HCA were applied to the NIR spectra. To improve the quality of the measurements and simplify the visualization, the mean of the four replicate spectra were used for each sample, resulting in a data matrix of 108 samples by 100 wavelength channels. Initially, only column mean centering was applied to the data. The paired scores plots for the first four principal components from PCA are presented in Fig. 2, where the different species are represented by different symbols as indicated in the legend. Based on the distribution of samples in the scores plots, there is no apparent separation of the species based on the NIR spectra. Although there is some suggestion of separation of classes 2 and 4 (*C. odorata* and *S. macrophylla*) using the third and fourth PCs, there is still strong overlap and no clear clustering is evident. Higher PCs did not improve separation.

Fig. 2. Paired scores plots from principal components analysis of sample mean spectra after column mean-centering, with species identified as in the legend. [Colour online.]



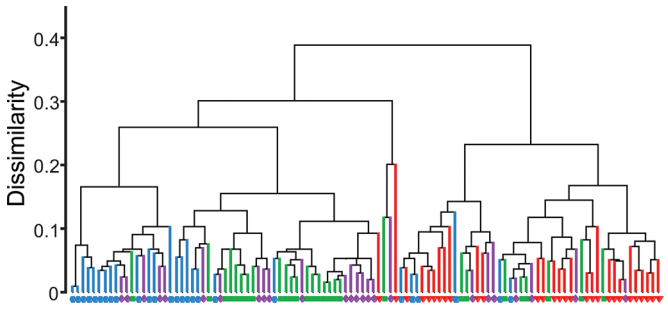
In many cases of multivariate analysis, it is necessary to preprocess data to obtain satisfactory results, so a variety of common preprocessing methods were employed here to see if the class separation could be improved. These included autoscaling, multiplicative signal correction, and the standard normal variate. Multiplicative signal correction and the standard normal variate are widely used in NIR spectroscopy to account for multiplicative offset noise.[17,21] None of the methods implemented resulted in any improvement of the PCA results.

The application of HCA did not improve on these results. HCA was implemented through algorithms in the Statistics and Machine Learning Toolbox of MatLab. A variety of linkage (average, complete, median, etc.) and preprocessing (autoscaling, multiplicative signal correction, standard normal variate) options were applied to both the full set of 432 spectra and the set of 108 sample mean spectra. Results for the latter are shown in Fig. 3, with the symbols and the colour of the bottom branches representing the species. The results shown are for an average distance calculation and mean-centering as the only preprocessing. Although some local groupings are evident in the tree structure, no consistent clusters are observed that would strongly support the hypothesis that the species can be distinguished on the basis of their NIR spectra. Although different preprocessing and linkage options produced changes in the tree structure, similar irregular class distributions were observed in all cases with no strong evidence of clusters related to species.

Error structure of NIR spectra

A central premise of this work is that exploratory analysis by HCA and PCA can be adversely affected by non-iid error structures. It is therefore necessary to examine the measurement error

Fig. 3. Dendrogram resulting from hierarchical clustering of sample mean spectra after mean-centering. Species are colour coded in the same manner as Figs. 1 and 2. Mean distance was used in the clustering algorithm. [Colour online.]



characteristics of the NIR spectra that are the focus of this study. For each of the 108 samples examined, replicate spectra were obtained from four different physical locations and should reflect the within-sample error variance. On the basis of these four replicates, an ECM can be calculated for each sample using eq. 8, leading to 108 individual ECMs.

Unfortunately, ECMs calculated on the basis of a small number of replicates are very noisy and unreliable.[42] For example, an error variance estimated from four replicates is expected to have a relative standard deviation of about 82%. This high level of noise in the individual ECMs makes their visual interpretation difficult and precludes their use in any advanced data analysis strategy. One solution to this problem, as noted earlier, is to pool (average) individual sample ECMs. This is valid in cases where the spectral characteristics of the samples are very similar, and therefore, the improved precision gained by pooling outweighs any between-sample differences. The similarity of the spectra in this study is evident from Fig. 1, so pooling was a viable option.

Initial pooling of the ECMs was carried out within each of the four species investigated. This was done as a preliminary evaluation to confirm the similarity of the ECMs within each group prior to global pooling, which is normally done. It was anticipated that the four ECMs would show very similar characteristics, which were consistent with NIR spectra. Although this was true for three of the groups (classes 1, 2, and 4), the remaining group (class 3) was distinctly different from the others, as shown in Fig. 4. For classes 1, 2, and 4 (Figs. 4*a*, 4*b*, and 4*d*), the ECMs are typical for NIR spectra,[23,32] showing heteroscedastic noise (non-uniform variance) along the diagonal and, more importantly, structured covariance (off-diagonal elements) that is consistent with offset and multiplicative offset noise. The latter is a dominant noise source in NIR reflectance measurements, arising from differences in the effective path length of scattered photons caused by changes in the scattering characteristics of the sampled region. The result is a shift in the spectral intensity proportional to its magnitude (hence the term multiplicative offset noise). The direction of the shift from the mean is random but is consistent within a spectrum, leading to highly correlated noise in which variance and covariance are directly related to the signal magnitude, as is evident in Figs. 4*a*, 4*b*, and 4*d*. Figure 4*c* is anomalous in this regard, however. Although the correlated and heteroscedastic noise is still evident, the magnitude of the measurement errors is largest in the shorter wavelength regions, where the signal is the lowest. This is confirmed through an examination of Fig. 1*b*, which shows a substantially greater variation of *M. melinoniana* in this region. The reason for the anomalous behavior of the third class is unclear, but it may be due to different physical properties that lead to different scattering characteristics by these samples.

Although the differences observed above suggest that a global pooling of ECMs from each class may not be representative of all samples, global pooling was nevertheless carried out and the results are shown in Fig. 5*a*. As anticipated, the globally pooled ECM (calculated using a weighted average reflecting the number of samples in each group) reflects the characteristics of the dominant classes (1, 2, and 4) but with a higher variance and (or) covariance in the short wavelength region due to the contribution of class 3. Despite its inaccuracy in its universal representation of all errors, the globally pooled ECM can still give improved results, because it is still superior to the assumption of iid normal errors, which is made in the usual applications of HCA and PCA. This is further explored in the section that follows. Also shown in Fig. 5*b* is the error correlation matrix (**R**) corresponding to the ECM (**Σ**) in Fig. 5*a*, calculated using eq. 7. The error correlation matrix removes the effects of the magnitude of the error that are evident in the ECM, showing only how they are related. Figure 5*b* shows almost perfect correlation (same direction and relative magnitude change in the errors) within three regions (<1148 nm, 1166–1343 nm, >1395 nm) but a smaller degree of correlation between these regions. This is typical for offset and (or) multiplicative offset noise in NIR spectra and shows a strong interdependence of measurement errors.

MLPCA of NIR spectra

For a matrix of chemical measurements, the intrinsic rank (pseudorank, chemical rank) is defined as the dimensionality of the space needed to account for all of the chemical variation in the absence of measurement error, and for linear systems, this is typically equal to the number of independently observable chemical components. When the intrinsic rank is well-defined and the ECMs of the measurement vectors are accurately known, MLPCA should yield the optimal estimate of the chemical subspace. For exploratory data analysis, however, decomposition by MLPCA is only guaranteed to provide an optimal visualization of the data when the intrinsic rank is equal to the dimensionality of the space into which the data are projected (called the projection rank), which can only be realized when the intrinsic rank is less than or equal to three.[51] In cases where the intrinsic rank exceeds the projection dimensionality, the advantages of MLPCA are less certain, but its application may provide a more useful visual projection of the data than PCA. In general, a definitive determination of the intrinsic rank (q) is difficult, so the application of MLPCA is typically carried out using different values to assess the projections empirically.

The application of MLPCA requires a specification of the data matrix, the corresponding ECMs, and the dimensionality of the subspace to be estimated. Based on the results of the previous section, which showed that the ECMs were not homogeneous among the different sample classes, it was decided to assign the ECM for each measurement vector based on its class membership (species), using the pooled ECM for the corresponding class. This error structure is representative of case E for the MLPCA algorithms[23,25] for general row-correlated errors. The objective function in this case is given by eq. 9 and is minimized through the alternating least squares method. The data matrix consisted of 108 rows corresponding to the sample mean spectra (column mean-centered) and an initial rank of two was selected. Although the alternating least squares algorithm is slower than the direct solution, which can be obtained when all of the ECMs can be assumed to be the same, it is considered to be more reliable when this assumption is violated, and the execution time was only about 20 s in this case.

The scores plot obtained through the application of MLPCA(E) (with a specified rank of 2) in this manner is shown in Fig. 6*a*. The results show a clear clustering of the samples into separate groups corresponding to the individual species, with the exception of one point from class 3 (*M. melinoniana*; it is noted that this does not correspond to the extreme point in the upper left panel of Fig. 2). This supports the hypothesis that there is sufficient information in the NIR spectra to distinguish among the four species. More

rich2/ccj-cjc/ccj-cjc/ccj99914/ccj1054d14z | xppws | S=3 | 6/5/18 | 8:18 | 4/Color Fig: F1-F8 | Art: **cjc-2017-0730** | Input-1st disk, 2nd ??
Pagination not final (cite DOI) / Pagination provisoire (citer le DOI)

Fig. 4. Pooled error covariance matrices of the near-infrared spectra for each of the four species examined: (*a*) class 1: *C. guianensis*, (*b*) class 2: *C. odorata*, (*c*) class 3: *M. melinoniana*, (*d*) class 4: *S. macrophylla*. [Colour online.]

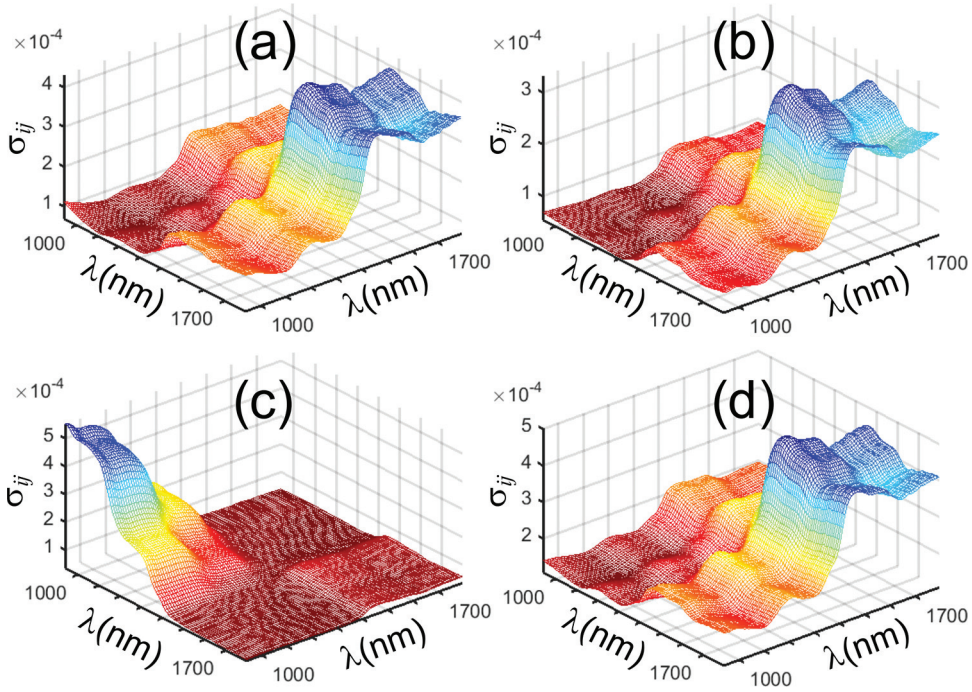
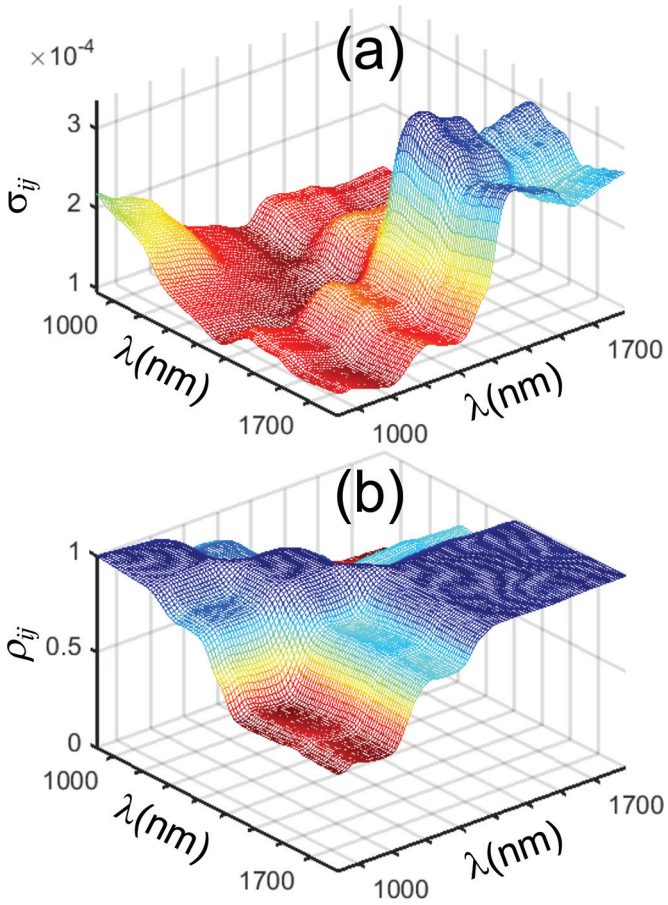


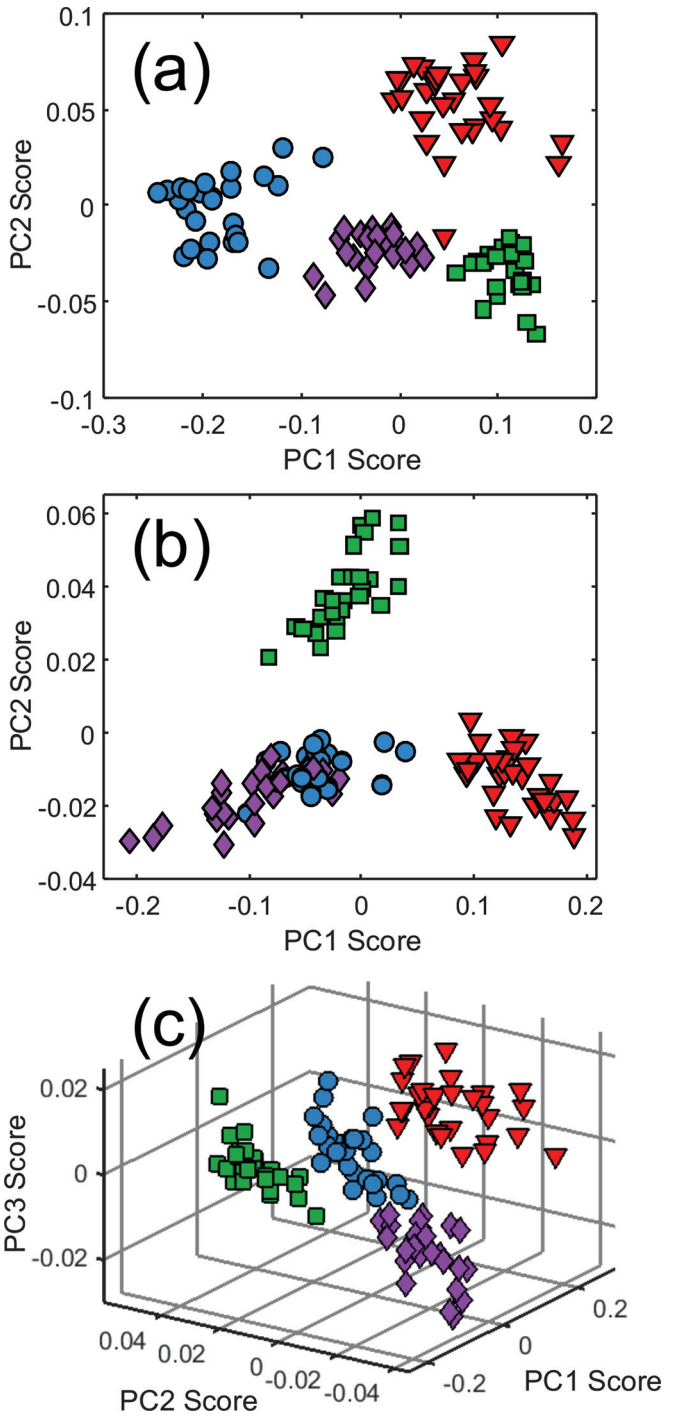
Fig. 5. Pooled error structure of near-infrared spectra based on 108 samples: (*a*) pooled error covariance matrix and (*b*) pooled error correlation matrix. [Colour online.]



importantly, in the context of the current work, it supports the broader hypothesis that the visualization of data by PCA can be impeded by non-iid measurement error structures and that this problem can be mitigated through the application of MLPCA. By incorporating information about the measurement error variance and covariance into the decomposition of the data, MLPCA can more effectively separate the variability originating from chemical differences from that arising from measurement noise, thereby giving a more useful picture of the relationships among samples.

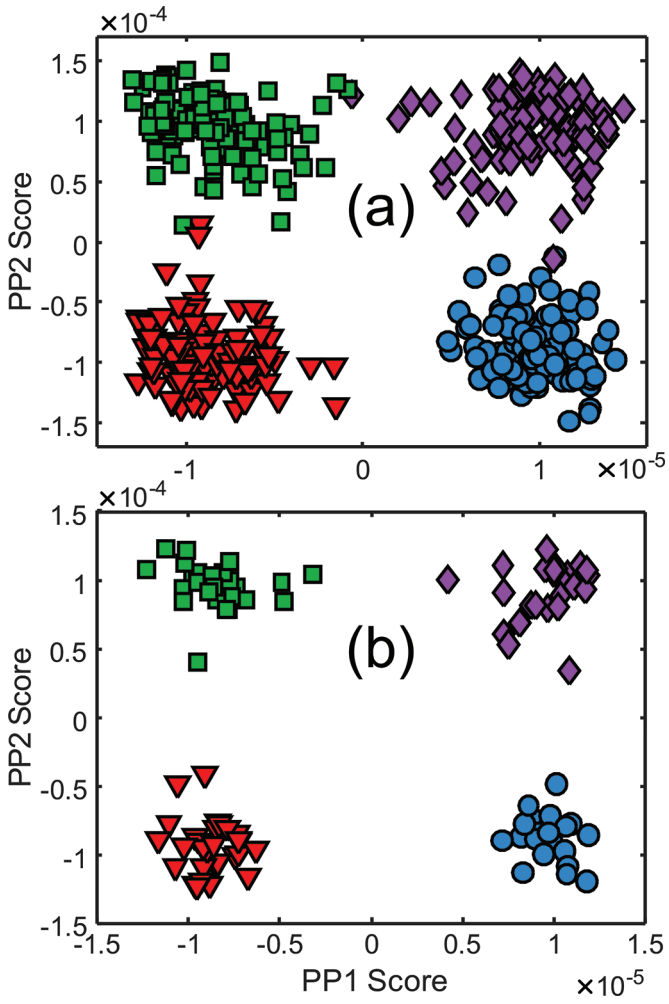
A potential argument that can be made to counter the conclusions drawn above is that, by defining the ECM according to class membership, indirect information related to class membership is being provided to the MLPCA algorithm and therefore biasing the outcome. This is a legitimate argument, as a truly unsupervised method should not include any information that could indirectly be associated with class membership. Although it cannot be concluded that the results in Fig. 6*a* are biased, this possibility cannot be excluded, so further evidence is needed. There are three possible ways to exclude bias. The first would be to provide an individual ECM for each sample based on its replicates. However, because there are only four replicates measured for each sample, the ECMs would be unreliable, as well as rank deficient due to the small number of replicates (rank = 3). Under these circumstances, anomalously small variances (due to limited replication) tend to drive the optimization, giving excessive weight to a few samples. This was confirmed by using the individual ECMs, resulting in a scores plot with a tight central cluster and a few dispersed samples (results not shown). A second possibility is to use a parameterized model for the ECM developed from multiple samples.[32,42] This can then be employed to calculate individual ECMs with greater reliability. In this case, however, it is clear that the same model could not be applied to all samples due to the differing characteristics of one of the classes. The third option would be to employ the globally pooled ECM, shown in Fig. 5*a*, to all of the samples. Although it is expected that the MLPCA solution obtained in this way would be suboptimal, it eliminates the possibility of bias and may produce projections superior to PCA.

Fig. 6. Scores plots from maximum likelihood principal components analysis (MLPCA) of near-infrared spectra. (*a*) Rank 2 MLPCA results using class-specific error covariance matrices (ECMs). (*b*) Rank 2 MLPCA results using a global average ECM. (*c*) Rank 3 MLPCA results using a global average ECM. Symbols correspond to the legend in Fig. 2. [Colour online.]



To implement this third option, MLPCA (case D, common row covariance) was applied to the 108 sample mean spectra (column mean-centered) using the globally pooled ECM with a specified rank of two and three. The scores plot for the rank two solution is shown in Fig. 6*b*. This result shows a clear separation of classes 2 and 3 (*C. odorata* and *M. melinoniana*) but strong overlap of the other two classes. However, the three-dimensional projection,

Fig. 7. Scores plots for the projection pursuit analysis of near-infrared spectra. (*a*) Scores plot from the analysis of all 432 spectra. (*b*) Scores plot resulting from the projection of sample means into the same space. Symbols correspond to the legend in Fig. 2. [Colour online.]



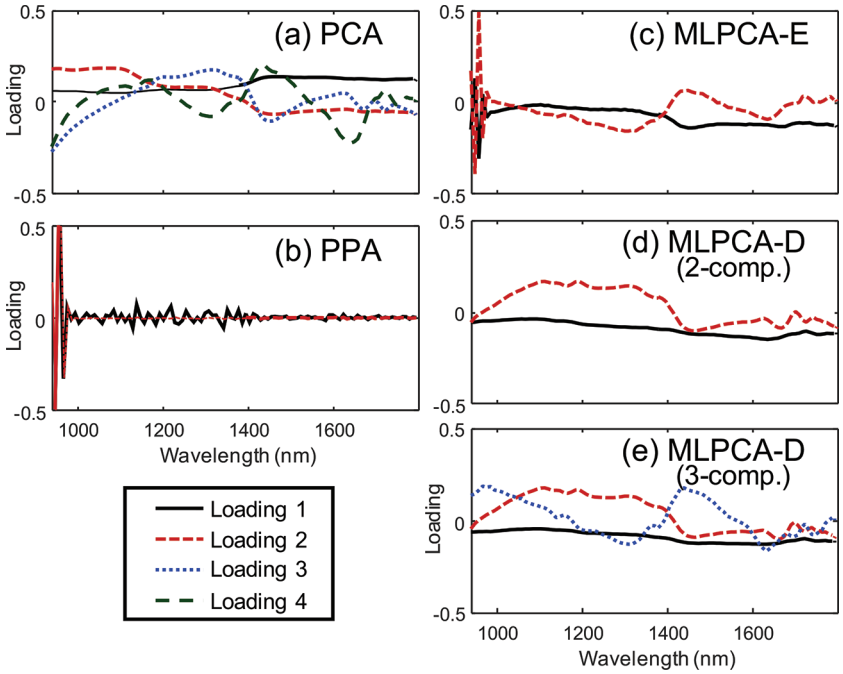
presented in Fig. 6*c*, shows a distinct separation of all four classes. As might be expected, the separation observed here is not as clear as for Fig. 6*a*, as a common ECM is erroneously assumed for all samples, but the results are far more informative than PCA. These results also exclude the possibility of an unintended bias and support the premise that class information can be more clearly extracted by incorporating measurement error information into the data analysis.

It should be noted that, in all of these cases, higher rank MLPCA solutions were also investigated. Separation of classes was still observed with increasing dimension, although the quality was diminished in the case of MLPCA(E) and slightly improved in the case of MLPCA(D) (results not shown).

PPA of NIR spectra

A weakness of all three methods investigated so far (HCA, PCA, and MLPCA) is that they rely on an assumption that the dominant sources of chemical variance are associated with the classes of interest; however, even when error variance is removed, other sources of chemical variance may eclipse the factors of interest. For example, in biological samples, variation in chemical species among individuals in a population or due to diurnal rhythms may mask smaller effects of interest. In principle, some of these variations can be built in to the description of measurement errors; however, in practice, this can be difficult to do. Projection pursuit

Fig. 8. Loadings plots for the various methods employed in this work, as labeled. [Colour online.]



approaches can avoid this problem by examining other criteria to obtain the optimal low dimensional projection.

In this work, kurtosis-based PPA was implemented using a stepwise univariate algorithm and orthogonal scores, with a two-dimensional projection space. This algorithm uses a stepwise procedure that first minimizes the univariate kurtosis along one projection dimension, optimally resulting in a binary separation of the data. After "deflation" of the data to remove the extracted dimension, the process is repeated in an attempt to provide a binary separation in subsequent dimensions, ultimately resulting in scores and loadings of selected dimensions analogous to PCA (although the loadings are not required to be orthogonal in this case). In this application, all 432 spectra (mean centered) were employed, as PPA works best when the ratio of samples to variables is high. The algorithm uses a nonlinear optimization method that is significantly slower than PCA, and random initial starting points are used to ensure a global optimum. In this implementation, 1000 initial guesses were used and the execution time was about 20 min.

The scores plot resulting from PPA of the raw data are shown in Fig. 7*a* and shows clear clustering of the four species, although there are a few samples that are grouped incorrectly. For a more direct comparison with earlier figures (Figs. 2 and 6), the 108 sample mean spectra have been projected into the same subspace in Fig. 7*b* and exhibit no overlap, as might be expected due to the smaller error variance. It is important to note that no class information was provided implicitly or explicitly to the algorithm, so the natural clustering on the basis of species was discovered solely on the basis of the observed spectra, supporting the hypothesis that the multivariate information available in the NIR spectra can be used to distinguish among the classes. No preprocessing of the data was necessary other than column mean-centering, and no measurement error information was provided to the algorithm.

Although PPA is an extremely powerful tool for exploratory studies, it is not without its limitations. Current algorithms are best suited for balanced data sets (approximately equal numbers of samples in each class) with more samples than variables and are most effective for 2, 4, or 8 classes. Ongoing work is directed at removing some of these limitations.

Loading vectors

Although it is sometimes asserted that loading vectors associated with scores plots can be interpreted to provide information on which variables are most important for classification or regression, such interpretation has been shown to be dubious at best because of the complexity of the linear relationships embodied in the loading vectors.[52] In the current work, comparisons are further limited by the nature of the methods used, as MLPCA does not use orthogonal projections of the data to generate scores and PPA does not result in orthogonal loading vectors. Nevertheless, for the sake of completeness, the loading vectors generated by the various methods are presented for qualitative comparison in Fig. 8. Although differences are readily apparent, the most noteworthy contrast is in the low wavelength regions (<1000 nm) for PPA and MLPCA(E) (Figs. 8*b* and 8*c*) when compared with the other cases. These were the most effective techniques for separation of the species in two dimensions, implying that these methods are able to effectively exploit information in this region. However, further attempts at interpretation would be purely speculative.

Conclusions

The results of this study can be summarized by five main conclusions. First, even when chemical information related to classification is present in a data set, traditional exploratory methods such as HCA and PCA may be incapable of revealing it. This is demonstrated by juxtaposing Figs. 2 and 3, which show no clear organization of the samples, with Figs. 6 and 7, which clearly show division of the samples based on biological species. A second conclusion, derived from the results shown in Fig. 6, is that inclusion of measurement error information into the data analysis, in this case through the application of MLPCA, can greatly improve the visualization of chemical information by more effectively separating the chemical variation from the noise variance. Thirdly, it can be further inferred from this that the limiting factor in the effective implementation of PCA (and likely HCA) was the presence of heteroscedastic and correlated errors (i.e., a non-iid error structure), suggesting that a better understanding of measurement errors should be a key component in the analysis of any

multivariate data set. Fourth, it was clearly demonstrated through Fig. 7 that the implementation of data visualization methods such as PPA that do not rely strictly on variance as a criterion for low dimensional projection could be extremely beneficial in studies involving multivariate data. Finally, with regard to the specific experimental data employed in this work, there is clear evidence that NIR spectroscopy has the capability to distinguish similar species of wood using the procedures described.

The alternative methods described here are not without their limitations. The application of MLPCA requires the availability of information on the measurement error structure, which may be difficult to obtain in certain studies where extensive replication is challenging. However, recent work has demonstrated that it is possible to develop measurement error models for analytical systems that minimize or eliminate the need for replication.[32,34,42] A better understanding of measurement error structures will certainly be advantageous in developing improved multivariate tools. In contrast, PPA does not require measurement error information but is less susceptible to non-iid error structures than PCA or HCA because it is not variance based. Nevertheless, PPA can be challenged by data sets that have a low sample to variable ratio or have unbalanced classes. The former problem has been addressed through variable compression, selection, and regularization,[28,30,31] and a re-centering strategy can mitigate the latter issue.[29] Further algorithmic developments will no doubt extend the applications of PPA to exploratory analysis.

Many areas of modern scientific discovery are initiated by testing an initial hypothesis that a complex multivariate data set contains information relevant to a desired goal such as disease detection or forensic classification. Such studies often involve a limited number of samples and a large number of variables. Although supervised classification methods (by design) are well suited to building classification models, they are poorly suited to test an initial hypothesis based on limited samples due to their need for extensive validation. Unsupervised (exploratory) methods play a key role in this workflow, because they do not have such strict validation requirements, but they are currently limited to two dominant techniques, HCA and PCA. As demonstrated here, these methods can fail to reveal important information in certain circumstances, and failure to support an initial hypothesis can impede the advance of research. Therefore, it is important to expand the toolbox available to researchers for exploratory analysis, and the alternative methods described here, MLPCA and PPA, are two approaches that can contribute in this regard.

Supplementary data

Supplementary data are available with the article through the journal Web site at http://nrcresearchpress.com/doi/suppl/10.1139/cjc-2017-0730.

Acknowledgements

The authors gratefully acknowledge the financial support of the Natural Sciences and Engineering Research Council (NSERC) of Canada and the ITTO-CITES Program and CNPq (process 308748/2015-8) of Brazil.

References

(1) Massart, D. L.; Vandeginste, B. G. M.; Deming, S. M.; Michotte, Y.; Kaufman, L. *Chemometrics: A Textbook*; Elsevier: Amsterdam, 1988.
(2) Beebe, K. R.; Pell, R. J.; Seasholtz, M. B. *Chemometrics: A Practical Guide*; Wiley: New York, 1998.
(3) Auf der Heyde, T. P. E. *J. Chem. Educ.* **1990**, *67*, 461. doi:10.1021/ed067p461.
(4) Malinowski, E. R. *Factor Analysis in Chemistry*, 3rd ed.; Wiley: New York, 2002.
(5) Bro, R.; Smilde, A. K. *Anal. Methods* **2014**, *6*, 2812. doi:10.1039/C3AY41907J.
(6) Meyer-Baese, A.; Wildberger, J.; Meyer-Baese, U.; Nilsson, C. L. *Electrophoresis* **2014**, *35*, 3452. doi:10.1002/elps.201400219.
(7) Peng, W.; Zhang, Y.; Zhu, R.; Mechref, Y. *Electrophoresis* **2017**, *38*, 2124. doi:10.1002/elps.201700027.
(8) Pontes, J. G. M.; Brasil, A. J. M.; Cruz, G. C. F.; de Souza, R. N.; Tasic, L. *Anal. Methods* **2017**, *9*, 1078. doi:10.1039/C6AY03102A.
(9) Yi, L.; Dong, N.; Yun, Y.; Deng, B.; Ren, D.; Liu, S.; Liang, Y. *Anal. Chim. Acta* **2016**, *914*, 17. doi:10.1016/j.aca.2016.02.001.
(10) Hendriks, M. M. W. B.; van Eeuwijk, F. A.; Jellema, R. H.; Westerhuis, J. A.; Reijmers, T. H.; Hoefsloot, H. C. J.; Smilde, A. K. *TrAC, Trends Anal. Chem.* **2011**, *30*, 1685. doi:10.1016/j.trac.2011.04.019.
(11) Laghi, L.; Picone, G.; Capozzi, F. *TrAC, Trends Anal. Chem.* **2014**, *59*, 93. doi:10.1016/j.trac.2014.04.009.
(12) Bosque-Sendra, J. M.; Cuadros-Rodríguez, L.; Ruiz-Samblás, C.; de la Mata, A. P. *Anal. Chim. Acta* **2012**, *724*, 1. doi:10.1016/j.aca.2012.02.041.
(13) Martín-Alberca, C.; Ortega-Ojeda, F. E.; García-Ruiz, C. *Anal. Chim. Acta* **2016**, *928*, 1. doi:10.1016/j.aca.2016.04.056.
(14) Muro, C. K.; Doty, K. C.; Bueno, J.; Halámková, L.; Lednev, I. K. *Anal. Chem.* **2015**, *87*, 306. doi:10.1021/ac504068a.
(15) Calcerrada, M.; García-Ruiz, C. *Anal. Chim. Acta* **2015**, *853*, 143. doi:10.1016/j.aca.2014.10.057.
(16) Old, O. J.; Fullwood, L. M.; Scott, R.; Lloyd, G. R.; Almond, L. M.; Shepherd, N. A.; Stone, N.; Barr, H.; Kendall, C. *Anal. Methods* **2014**, *6*, 3901. doi:10.1039/c3ay42235f.
(17) Byrne, H. J.; Knief, P.; Keating, M. E.; Bonnier, F. *Chem. Soc. Rev.* **2016**, *45*, 1865. doi:10.1039/C5CS00440C.
(18) Kangas, M. J.; Burks, R. M.; Atwater, J.; Lukowicz, R. M.; Williams, P.; Holmes, A. E. *Crit. Rev. Anal. Chem.* **2017**, *47*, 138. doi:10.1080/10408347.2016.1233805.
(19) Gromski, P. S.; Muhamadali, H.; Ellis, D. I.; Xu, Y.; Correa, E.; Turner, M. L.; Goodacre, R. *Anal. Chim. Acta* **2015**, *879*, 10. doi:10.1016/j.aca.2015.02.012.
(20) Brereton, R. G.; Lloyd, G. R. *J. Chemom.* **2014**, *28*, 213. doi:10.1002/cem.2609.
(21) Rinnan, Å.; van den Berg, F.; Engelsen, S. B. *TrAC, Trends Anal. Chem.* **2009**, *28*, 1201. doi:10.1016/j.trac.2009.07.007.
(22) Engel, J.; Gerretzen, J.; Szymańska, E.; Jansen, J. J.; Downey, G.; Blanchet, L.; Buydens, L. M. C. *TrAC, Trends Anal. Chem.* **2013**, *50*, 96. doi:10.1016/j.trac.2013.04.015.
(23) Wentzell, P. D. *J. Braz. Chem. Soc.* **2014**, *25*, 183. doi:10.5935/0103-5053.20130293.
(24) Wentzell, P. D.; Andrews, D. T.; Hamilton, D. C.; Faber, K.; Kowalski, B. R. *J. Chemom.* **1997**, *11*, 339. doi:10.1002/(SICI)1099-128X(199707)11:4<339::AID-CEM476>3.0.CO;2-L.
(25) Wentzell, P. D. In *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*; Brown, S. D., Tauler, R., Walczak, B., Eds.; Elsevier: Amsterdam, 2009; Vol. 2, p. 507. doi:10.1016/B978-044452701-1.00057-0.
(26) Kruskal, J. B. In *Statistical Computation*; Milton, R. C., Nelder, J. A., Eds.; Academic Press: New York, 1969; p. 427. doi:10.1016/B978-0-12-498150-8.50024-0.
(27) Friedman, J. H.; Tukey, J. W. *IEEE Trans. Comput.* **1974**, *C-23*, 881. doi:10.1109/T-C.1974.224051.
(28) Hou, S.; Wentzell, P. D. *Anal. Chim. Acta* **2011**, *704*, 1. doi:10.1016/j.aca.2011.08.006.
(29) Hou, S.; Wentzell, P. D. *J. Chemom.* **2014**, *28*, 370. doi:10.1002/cem.2568.
(30) Hou, S.; Wentzell, P. D. *Metabolomics* **2014**, *10*, 589. doi:10.1007/s11306-013-0612-z.
(31) Wentzell, P. D.; Hou, S.; Silva, C. S.; Wicks, C. C.; Pimentel, M. F. *Anal. Chim. Acta* **2015**, *877*, 51. doi:10.1016/j.aca.2015.03.006.
(32) Leger, M. N.; Vega-Montoto, L.; Wentzell, P. D. *Chemom. Intell. Lab. Syst.* **2005**, *77*, 181. doi:10.1016/j.chemolab.2004.09.017.
(33) Karakach, T. K.; Wentzell, P. D.; Walter, J. A. *Anal. Chim. Acta* **2009**, *636*, 163. doi:10.1016/j.aca.2009.01.048.
(34) Wentzell, P. D.; Tarasuk, A. C. *Anal. Chim. Acta* **2014**, *847*, 16. doi:10.1016/j.aca.2014.08.007.
(35) Ingle, J. D., Jr.; Crouch, S. R. *Spectrochemical Analysis*; Prentice Hall: Englewood Cliffs, N.J., 1988.
(36) Ince, A. T.; Williams, J. G.; Gray, A. L. *J. Anal. At. Spectrom.* **1993**, *8*, 899. doi:10.1039/ja9930800899.
(37) Hayashi, Y.; Matsuda, R. *Anal. Chem.* **1994**, *66*, 2874. doi:10.1021/ac00090a013.
(38) Hayashi, Y.; Matsuda, R. *Anal. Sci.* **1995**, *11*, 929. doi:10.2116/analsci.11.929.
(39) Van Vliet, C. M. *Sens. Actuators, B* **1995**, *24*, 6. doi:10.1016/0925-4005(95)85006-6.
(40) Hayashi, Y.; Matsuda, R.; Poe, R. B. *J. Chromatogr. A* **1996**, *722*, 157. doi:10.1016/0021-9673(95)00437-8.
(41) Mittermayr, C. R.; Lendl, B.; Rosenberg, E.; Grasserbauer, M. *Anal. Chim. Acta* **1999**, *388*, 303. doi:10.1016/S0003-2670(99)00083-5.
(42) Wentzell, P. D.; Cleary, C. S.; Kompany-Zareh, M. *Anal. Chim. Acta* **2017**, *959*, 1. doi:10.1016/j.aca.2016.12.009.
(43) Wentzell, P. D.; Lohnes, M. T. *Chemom. Intell. Lab. Syst.* **1999**, *45*, 65. doi:10.1016/S0169-7439(98)00090-2.
(44) Wentzell, P. D.; Andrews, D. T.; Kowalski, B. R. *Anal. Chem.* **1997**, *69*, 2299. doi:10.1021/ac961029h.
(45) Schreyer, S. K.; Bidinosti, M.; Wentzell, P. D. *Appl. Spectrosc.* **2002**, *56*, 789. doi:10.1366/00037020276076857.
(46) Wentzell, P. D.; Karakach, T. K.; Roy, S.; Martinez, M. J.; Allen, C. P.; Werner-Washburne, M. *BMC Bioinf.* **2006**, *7*, 343. doi:10.1186/1471-2105-7-343.
(47) Tauler, R.; Viana, M.; Querol, X.; Alastuey, A.; Flight, R. M.; Wentzell, P. D.;

Hopke, P. K. *Atmos. Environ.* **2009**, *43*, 3989. doi:10.1016/j.atmosenv.2009.05.018.
(48) Pereira, J. F. Q.; Silva, C. S.; Braz, A.; Pimentel, M. F.; Honorato, R. S.; Pasquini, C.; Wentzell, P. D. *Microchem. J.* **2017**, *130*, 412. doi:10.1016/j.microc.2016.10.024.
(49) Coradin, V. T. R.; Camargos, J. J. A.; Marques, L. F.; da Silva, E. R., Jr. *Madeiras similares ao mogno (Swietenia macrophylla King.): chave ilustrada para identificação anatômica em campo*; Serviço Florestal Brasileiro: Brasília, 2009.
(50) Stern, W. L. *IAWA Bull.* **1988**, *9*, 209.
(51) Wentzell, P. D.; Hou, S. *J. Chemom.* **2012**, *26*, 264. doi:10.1002/cem.2428.
(52) Brown, C. D.; Green, R. L. *TrAC, Trends Anal. Chem.* **2009**, *28*, 506. doi:10.1016/j.trac.2009.02.003.