N° /__/__/__/__/__/__/__/__/__/

THÈSE

pour obtenir le grade de

Docteur

de

l'Institut des Sciences et Industries du Vivant et de l'Environnement (Agro Paris Tech)

Spécialité : Chimie Analytique

présentée et soutenue publiquement
par

Rui José DOS SANTOS CLÍMACO PINTO

le 22 Juin 2009

DÉVELOPPEMENT DE NOUVELLES MÉTHODES CHIMIOMÉTRIQUES D'ANALYSE

APPLICATION À LA CARACTÉRISATION SPECTROSCOPIQUE DE LA QUALITÉ DES ALIMENTS

Directeur(s) de thèse : ***Douglas N. RUTLEDGE / António S. BARROS***

Travail réalisé : Laboratoire de chimie analytique, AgroParisTech 75005 Paris

Devant le jury :

M. El Mostafa QANNARI, Professeur, ENITIAA - Nantes □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ .**Rapporteur**
M. Jean-Michel ROGER, IGREF, CEMAGREF - Montpellier□ □ □ □ □ □ □ □ .□ ..□ □ □ ... **Rapporteur**
M. Dominique BERTRAND, Directeur de Recherche, INRA – Nantes□ ..□ □ □ □ □ □ ...**Examinateur**
M. Soren ENGELSEN, Professeur, University of Copenhagen (Danemark)□ □ □ □ ..□ ..**Examinateur**
M. Douglas N. RUTLEDGE, Professeur, AgroParisTech, INRA – Paris.....**Directeur de thèse**
M. António S. BARROS, Enseignant-chercheur, Univ. Aveiro (Portugal).....**Co-Directeur de thèse**

A mes parents
Pour le support total, toujours

To my parents
For the total support, always

Aos meus pais
pelo apoio total, sempre

Préface

Après deux ans de Master sur les sujets de l'analyse instrumentale et des données je n'étais pas satisfait avec ma connaissance de ces passionnants domaines. J'avais eu une première introduction à la chimiométrie avec le docteur António Barros à l'Université d'Aveiro au Portugal, qui était toujours très motivé sur ces sujets. Après aussi quelques mots de motivation des professeurs Ivonne Delgadillo, Manuel Coimbra et d'autres, je me suis décidé d'approfondir mes connaissances sur le sujet.

J'ai fais parti d'un projet européen pour préparer une base de données de spectres électromagnétiques de substances diverses où j'ai pas mal voyagé pour des réunions internationales et ou j'ai fait la connaissance des gens très intéressantes. Entre eux le professeur Douglas Rutledge, directeur du laboratoire de chimie analytique de l'AgroParisTech (ex – Institut National Agronomique Paris – Grignon) à qui, à la fin du projet j'ai proposé de m'encadrer une thèse de doctorat en collaboration avec António Barros de l'Université d'Aveiro. Passé le concours pour une bourse de doctorat avec la «Fundação para a Ciência e a Tecnologia (FCT)» du Portugal, c'était parti pour Paris…

Les méthodes multivariées et spectroscopiques présentées dans la suite de ce document ont été étudiés dans le cadre de cette thèse. Elles ont été appliquées, modifiées, améliorées, révisées, inventés ou même ont été détournées à des utilisations originales. La plupart des études présentées ici ont en commun le fait d'avoir donnée lieu à des articles acceptées ou soumis à publications dans des journaux internationaux à comité de lecture. D'autres études ont occupé l'esprit de l'auteur pas mal de temps, mais ne lui semblent pas avoir été encore suffisamment approfondies pour être publiées.

Je suis très fière que ce travail a donnée lieu à 7 articles dans des journaux internationaux à comité de lecture et que trois d'entre eux ont été considérés comme des «HOT ARTICLE» dans ces journaux. Il a été un grand honneur de gagner le prix du 1er meilleur poster à la conférence «Chimiométrie 2007» à Lyon. Et un vrai plaisir de présenter ma recherche oralement dans la conférence mondiale «Chemometrics in Analytical Chemistry - CAC2008» à Montpellier.

Je voudrais remercier spécialement à mes encadrants Douglas Rutledge et António Barros pour tout le support et l'amitié au cours de cette thèse.

Le support financier, la bourse de doctorat de la «Fundação para a Ciência e a Tecnologia (FCT)» du Portugal sans lequel cette thèse n'avais pas été possible de réaliser, a été beaucoup apprécié. Merci.

Un grand merci aux gens du laboratoire de chimie analytique d'AgroParisTech pour les bons moments passées ensemble, les discussions, l'aide et d'avoir ri même de mes blagues les moins réussis.

Merci à tous ceux qui ont partagé son travail, idées, connaissance et compétences avec moi au cours des diverses collaborations.

À tous les amis et camarades qui ont partagée les bons moments de mon séjour à Paris, merci beaucoup. Même sûr qu'on a eu aussi des mauvais moments, je ne me souviens d'aucun!

Paris a été un merveilleux période d'inspiration au niveau professionnel et personnel. Une période de découverte mais aussi de confirmation.

Sommaire

Les propriétés de compression de la transformation PCT permettent de réaliser et d'accélérer le calcul des méthodes multivariées même avec des matrices très larges où le nombre de variables est beaucoup plus grand que le nombre d'objets. Les résultats peuvent être retransformés vers le domaine des variables d'origine pour donner exactement les mêmes résultats qu'avec la méthode standard, sans transformation. Nous avons utilisé la PCT pour accélérer la régression PLS et dans le cadre de l'analyse du produit externe avec des très grandes matrices.

La méthode ANOVA-PCA facilite la comparaison de la variance des Facteurs responsables des différences entre les échantillons avec la variance résiduelle. Elle a été utilisée pour étudier par spectroscopie d'infrarouge proche la stabilité des matériaux de référence stockés dans différentes conditions. Nous l'avons aussi modifiée de façon à permettre la prédiction de nouveaux échantillons et pour améliorer ses capacités de trouver des facteurs signifiants et pour mieux comprendre les différentes sources de variabilité résiduelle.

En profitant d'une des idées de base de la méthode ANOVA-PCA, le calcul des niveaux, nous avons modifié une méthode d'analyse de tableaux multiples, la CCSWA, pour l'adapter à l'analyse des tableaux de niveaux de Facteurs provenant d'un plan d'expériences.

La spectroscopie MIR a été utilisée pour l'analyse d'aliments à l'aide d'ATR chauffante. Nous avons étudié des altérations de différentes huiles alimentaires par chauffage accélérée à différentes températures avec acquisition simultanée des spectres. Le même accessoire a été utilisé dans l'analyse de vin pour étudier les effets de différents traitements technologiques, tels que la micro-oxygénation et l'ajout de copeaux de bois. Après évaporation, sur le crystal, de l'eau et de l'éthanol, les constituants majoritaires, des pics correspondants aux autres composantes du vin, normalement cachées, ont pu donc se révéler.

Mots-clés

Chimiométrie, analyse de données, PLS, produit externe, PCT, Seg-PCT, ANOVA-PCA, analyse discriminante, ComDim, analyse multi-tableaux, CCSWA.

Spectroscopie, MIR, NIR, NMR, fluorescence, ATR, aliments, huiles, stabilité, oxydation, matériaux de référence, vin, micro-oxygénation, ajout de coupeaux.

Abstract

The compression properties of PCT allow one to perform or just accelerate the calculations of multivariate methods in the case of very wide matrices, where the number of variables is much greater than the number of objects. The results can be back-transformed into the original vectors base giving exactly the same results as those obtained when using the standard method, without transformation.

The ANOVA-PCA method facilitates the comparison of variance attributed to the Factors responsible for the differences between samples and the residual variance. It was used to study, by means of near infrared spectroscopy, the stability of reference materials stored under different conditions. The method was modified to allow the prediction of new samples, as well as to improve its ability to find significant Factors and understand sources of variability.
Using one of the main concepts of the ANOVA-PCA method, the calculation of factor levels, a method used for multi-table analysis – CCSWA - was modified for the analysis of the Factor level matrices resulting from an experimental design.

Mid-infrared spectroscopy was used for the analysis of food samples with the help of the heated ATR accessory. It allowed the study of the modifications of different edible oils after accelerated heating at different temperatures and simultaneous acquisition of spectra.
The same ATR apparatus was used in the analysis of wine to study the effects of different technological treatments, namely Micro-oxygenation and Oak addition. After the evaporation of the main constituents, water and ethanol, directly on the ATR crystal, the peaks corresponding to minor components could be uncovered.

Keywords

Chemometrics, data analysis, PLS, Outer product, PCT, Seg-PCT, ANOVA-PCA, discriminant analysis, ComDim, multi-table analysis, CCSWA.
Spectroscopy, MIR, NIR, NMR, fluorescence, ATR, food chemistry, edible oils, stability, oxydation, reference materials, wine, micro-oxygenation, Oak addition.

Liste et introduction aux publications, posters et présentations orales

Publications

I Barros, A.S.; Pinto, R.; Delgadillo, I.; Rutledge, D.N.; «Segmented Principal Component Transform–Partial Least Squares regression». Chemometrics and Intelligent Laboratory Systems 89 (**2007**) 59–68.

Les propriétés et les avantages de la méthode Seg-PCT sont étudiées dans le cadre de l'analyse de matrices très grandes (nombre de colonnes >>> nombre de lignes) par régression PLS. La méthode réduit beaucoup la taille de la matrice «de travaille» originale, ce que diminue aussi le temps et surtout les besoins de mémoire nécessaires pour les calculs. Les résultats obtenus sont exactement les mêmes qu'avec la méthode PLS standard.

II Sarembaud, J.; Pinto, R.; Rutledge, D.N.; Feinberg, M.; «Application of the ANOVA-PCA method to stability studies of reference materials». Analytica Chimica Acta 603 (**2007**) 147–154.

La stabilité des matériaux de référence externes (ERM) d'huile de tournesol et de farine de blé a été étudiée sous différentes conditions, suivant le plan d'expériences. L'effet de la température, la nature de l'atmosphère dans les emballages et le temps de stockage étaient les Facteurs. Les spectres NIR ont été acquis au cours du temps et la méthode multivariée ANOVA-PCA a été utilisée pour évaluer la signifiance des Facteurs par rapport à l'erreur résiduelle.

III Climaco Pinto, R.; Bosc, V.; Noçairi, H.; Barros, A.S.; Rutledge, D.N.; «Using ANOVA-PCA for discriminant analysis: Application to the study of mid-infrared spectra of carraghenan gels as a function of concentration and temperature». Analytica Chimica Acta 629 (**2008**) 47-55.

Un plan d'expériences a été établie pour étudier des gels de carraghénane à travers de la spectroscopie dans le moyen infrarouge. Les facteurs utilisés étaient la température, la concentration et le type de nettoyage du système ATR. La méthode ANOVA-PCA est utilisée pour la discrimination des groupes et évaluation de la signifiance des facteurs. Une nouvelle méthode de classification en utilisant l'ANOVA-PCA a été développée.

IV Barros, A.S.; Pinto, R.; Bouveresse, D.J-R.; Rutledge, D.N.; «Principal component transform — Outer product analysis in the PCA context» Chemometrics and Intelligent Laboratory Systems 93 (**2008**) 43–48.

Le paradigme PCT est revisité, cette fois avec une application sur l'analyse des produits externes. Cette méthode de combinaison de domaines peut créer des matrices énormes, et dans ces cas, l'application de PCT devient une aide précieuse pour accélérer les calculs et même pour les rendre faisables.

V Climaco Pinto, R.; Barros, A.S.; Locquet, N.; Schmidtke, ; Rutledge, D.N. ; «Improving the detection of significant factors using ANOVA-PCA by selective reduction of residual variability». Soumis à Analytica Chimica Acta le 10 février **2009**.

L'ANOVA-PCA présente des limitations qui peuvent réduire son pouvoir. La plus importante limitation est que, pour évaluer la signifiance des facteurs, elle ne compare que les deux composantes principales les plus importantes. Toutes les autres composantes ne sont pas prises en compte dans l'analyse. Une approche basée sur la réduction de la variabilité résiduelle a été proposée pour augmenter les capacités de la méthode, notamment son pouvoir discriminante et pour mieux comprendre les sources de variance résiduelle.

VI Jouan-Rimbaud Bouveresse, D. ; Climaco Pinto, R.; Schmidtke, L.; Locquet, N.; Rutledge, D.N.; «A multi-block extension of the APCA procedure for the identification of significant factors». Soumis à Analytica Chimica Acta le 10 Avril **2009**.

La méthode CCSWA a été développée en Sensométrie pour l'analyse des tableaux multiples. Elle cherche à trouver des dimensions communes à l'ensemble des tableaux et fournit des résultats qui incluent pour chaque Composantes Commune, les coordonnées des individus et des contributions factorielles pour chaque tableau, ainsi qu'un poids spécifique associé à chaque tableau pour cette Composante. La méthode a été modifiée pour rentrer dans le cadre des études du style ANOVA-PCA. Cette extension multi-tableaux de la méthode présente des propriétés intéressantes par rapport aux méthodes classiques.

VII Climaco Pinto, R.; Locquet, N.; Eveleigh, L.; Rutledge, D.N.; Preliminary studies on the Mid-Infrared analysis of edible oils by direct heating on an ATR diamond crystal. Soumis à Food Chemistry le 15 Mai **2009**.

Les altérations des huiles alimentaires peuvent inclure des substances nocives pour l'alimentation humaine. Plusieurs études existent sur le sujet, basées sur différentes techniques d'analyse, dont la spectroscopie infrarouge. Dans la plupart des travaux qui suivent, les huiles sont oxydés et des prélèvements utilisées pour acquérir les spectres. Dans cet article on présente une méthode rapide d'altération-acquisition simultanée des spectres avec le MIR-ATR chauffante, en regardant l'évolution à quelques longueurs d'onde connues.

Posters

VIII Climaco Pinto, R.; Barros, A.S.; Jouan-Rimbaud Bouveresse, D..; Rutledge, D.N.; «Using the Principal Component Transform (PCT) framework to enable Outer Product (OP) - Partial Least Squares (PLS) regression analysis in huge matrices». Chimiométrie 2005, Polytech'Lille – Cité scientifique, Villeneuve d'Ascq, France, 30 Novembre – 1 Décembre **2005**.

Le principe de la PCT est présenté et utilisé dans une application pratique, pour permettre l'utilisation de la méthode de combinaison des matrices par produit externe suivie par une régression PLS. La méthode conjointe Seg-PCT-OP-PLS est appliquée à des données RMN-2D acquises sur des vins. A cause de la taille des matrices, il aurait été impossible d'appliquer la méthode OP-PLS standard (sans compression par «Segmented PCT») à ces données.

IX Climaco Pinto, R.; Barros, A.S. ; Jouan-Rimbaud Bouveresse, D. ; Rutledge, D.N.; «Using Principal Component Transform to accelerate Outer Product – Partial Least Squares Regression calculations». Chimiométrie 2005, Polytech'Lille – Cité scientifique, Villeneuve d'Ascq, France, 30 Novembre – 1 Décembre **2005**.

Les calculs de la méthode OP-PLS sont longs et demandent beaucoup de mémoire de l'ordinateur, surtout dans la phase de validation croisée. La méthode PCT est donc utilisée

pour accélérer les calculs et réduire les besoins de mémoire. Ce travail présente une discussion de l'application de la méthode autour d'un exemple pratique.

X Climaco Pinto, R.; Bosc, V.; Barros, A.S.; Jouan-Rimbaud Bouveresse, D.; Rutledge, D.N.; «Using ANOVA-PCA for discrimination between FTIR-ATR spectra of Carragheenan gels in different concentrations and at different temperatures and classification of new samples. Comparison with other chemometric techniques». Chimiométrie 2006, École Nationale Supérieure, Paris, France, 30 Nov. – 1 Déc. **2006**.

La méthode ANOVA-PCA est appliquée à des données MIR–ATR pour étudier la rhéologie d'un gel alimentaire. Les échantillons sont préparés suivant un plan d'expériences simple et les longueurs d'onde responsables par la séparation des niveaux des facteurs sont identifiées. Une nouvelle méthode de prédiction basée sur l'ANOVA-PCA est présentée et discutée.

XI Climaco Pinto, R.; Noçairi, H.; Bosc, V.; Barros, A.S.; Rutledge, D.N.; «OSC-PCA : une modification de la méthode ANOVA-PCA». Chimiométrie 2007, École Supérieure de Chimie Physique Electronique de Lyon, France, 29-30 Novembre **2007**.

La méthode ANOVA-PCA a des limitations, notamment le fait qu'elle doit traiter les Facteurs quantitatifs comme s'ils étaient qualitatifs et qu'elle s'applique sur des données issues d'un plan d'expériences équilibré. De plus la première étape de génération des matrices par le calcul des moyennes des niveaux suivant la philosophe "ANOVA", est univariée. Une modification de l'ANOVA-PCA est proposée pour remplacer cette étape en utilisant l'OSC.

Présentations orales

XII Climaco Pinto, R.; Bosc, V.; Barros, A.S.; Noçairi, H.; Rutledge, D.N.; «Using ANOVA-PCA for variable selection and sample classification and replacing ANOVA by OSC to improve the detection of significant factors». 11th international conference on Chemometrics for Analytical Chemistry (CAC2008), SupAgro, Montpellier, France, 30 Juin – 4 Juillet **2008**.

La méthode ANOVA-PCA a été présentée et toutes les modifications et extensions proposées pour la méthode pendant les travaux de cette thèse ont été discutées et comparées.

Notation mathématique

Les matrices sont représentées par des lettres grasses majuscules (par exemple **X**).
Les vecteurs sont représentés par des lettres grasses minuscules (par exemple **x**).
Les nombres de lignes et de colonnes des matrices sont représentées en italique (*a, b*).
Les matrices transposées sont représentées par un T en exposant (par exemple $\mathbf{X}^T$)

Sigles et abréviations

ANOVA – Analyse de variance («Analysis of variance»)
ATR – Refléxion totale attenuée («Attenuated total reflection»)
CCSWA – Analyse des composantes communs et des poids spécifiques («Common components and specific weights analysis»)
ComDim – Dimensions communes («common dimensions»)
CP/MAS – («Cross polarization magic angle spinning»)
CV – Validation croisée («Cross-validation»)
EROS – («Error removal by orthogonal subtraction»)
EVD – («Eigenvalue decomposition»)
FTIR – Infrarouge avec transformée de Fourier («Fourier transform infrared spectroscopy»)
FTIR-PAS – Spectroscopie FTIR - photo acoustique («FTIR – photo acustic spectroscopy»)
ICA – Analyse en composantes indépendantes («Independent components analysis»)
IR – Infrarouge («Infrared»)
LC-MS – Chromatographie liquide – spectrométrie de masse («Liquid chromatography – mass spectrometry»)
LV – Variable latente (« Latent variable »)
MANOVA – ANOVA multivariée («Multivariate ANOVA»)
MIR – Moyen Infrarouge (Mid-infrared»)
MSC – («Multiplicative scatter correction»)
NIR – Proche Infrarouge («Near-infrared»)
NIPALS – («Nonlinear iterative partial least squares»)
NMR – Résonance magnétique nucléaire («Nuclear magnetic resonance»)
OP – Produit externe («Outer product»)
OPA – Analyse du produit externe («Outer product analysis»)

OSC – («Orthogonal signal correction»)

PARAFAC – Analyse de facteurs en parallèle («Parallel factor analysis»)

PC – Composante principale (« Principal Component »)

PCA – Analyse en composantes principales («Principal components analysis»)

PCR – Régression sur composantes principales («Principal components regression»)

PCT – Transformée en composantes principales («Principal components transform»)

PLS – Régression en moindres carrées partielles («Partial least squares» ou «Projection to latent structures»)

PLS-DA – Analyse discriminante par PLS («PLS discriminant analysis»)

PRESS – Somme des carrées des écarts de prédiction («Predicted residual error sum of squares»)

QSAR – Relation quantitative structure-activité («Quantitative structure-activity relationship»)

Seg-PCT – PCT segmentée («Segmented-PCT»)

SNV – («Standard normal variate»)

SVD - Décomposition en valeurs singulières («Singular value decomposition»)

SIMCA – («Soft independent modelling of class analogies»)

SIMPLS – («Statistically Inspired Modification of Partial Least Squares»)

⊗ - Opérateur du produit externe

Table de matières

ANNEXES

Figures

Introduction

Chimiométrie

Le terme «chimiométrie» («chemometrics») a été introduit en 1972 par S. Wold et B.R. Kowalski[1]. Plusieurs définitions de Chimiométrie existent, mais la plus communément acceptée est : «La Chimiométrie est la discipline de la chimie que utilise des méthodes mathématiques, statistiques et d'autres employant une logique formelle pour planifier ou sélectionner des expériences et des procédures expérimentaux optimales, pour extraire le maximum d'informations pertinentes à partir de l'analyse des données chimiques et pour comprendre des systèmes chimiques»[2].

Chimie et chimiométrie

La nature expérimentale de la chimie fait que depuis le début il y a eu besoin de planification expérimentale et d'analyse des données. Même si le nom de la discipline n'existe que depuis peu de temps, certaines des méthodes de la chimiométrie ont accompagné le développement de la chimie, sa croissance et ses mutations. Donc, aussi comme pour la chimie, la façon dont on regarde la chimiométrie aujourd'hui n'est pas la même que tout au début. Celle-ci a développé des outils pour exploiter des données provenant de la chimie organique et minérale, de la chimie "instrumentale" et de la chimie des solutions («wet chemistry»). On retrouve ces méthodes aussi depuis peu dans les domaines de la biologie, de la biochimie et de la métabonomique (et des autres «omiques». La chimiométrie moderne, basée sur des outils de l'analyse multivariée commune à d'autres sciences, a aidé au développement de nouvelles méthodes d'analyse chimique, pour optimiser les expériences, baisser les seuils de détection des méthodes, augmenter la précision des instruments, augmenter la vitesse des analyses et la densité de l'information générée, et pour automatiser des mesures et le contrôle des processus [2,3].

Méthodes chimiométriques

Il existe une grande gamme de méthodes multivariées, dont une liste non-exhaustive des plus populaires inclue les suivantes: PCA, Factor Analysis, ICA, PCR, PLS, PLS-DA, SIMCA, PARAFAC, Tucker, MANOVA, Réseaux de neurones, «Kohonen networks», «Fuzzy methods», «Warping», Algorithmes génétiques, «Support vector machines», «Multivariate curve resolution», Ondelettes …

Évolution de la chimiométrie

Certains associent le début de la chimiométrie avec les travaux d'analyse univariée de W.S. Gosset pour les brasseries Guinness vers 1908, d'autres pensent que la chimiométrie doit être considérée comme plus contemporaine et qu'elle est née avec l'utilisation de l'analyse multivariée au début des années 70. Dans notre opinion, c'est à ce moment, avec le développement de l'informatique, que les possibilités de computation et d'acquisition de données deviennent importantes et nécessitent une approche différente de la part des chimistes, ce qui entraine la création de cette nouvelle discipline, la chimiométrie[4]. Ceci est renforcé par l'augmentation du nombre et de la sophistication des instruments analytiques qui génèrent des quantités de données de plus en plus complexes, ce qui est encore plus vrai dans cette époque post-génomique[3,5,6]. Si l'apparition des ordinateurs personnels a beaucoup fait pour augmenter la capacité et la vitesse des calculs scientifiques, cela ne veut pas dire que tous les problèmes ont été résolus a ce niveau. Un des grands défis de l'analyse de données actuelle est de trouver de nouvelles approches rapides pour traiter des données provenant du suivi en continu de processus industrielles, ou des domaines d'étude, tels que la protéomique et la métabonomique où les signaux sont à très forte densité d'information. Avec la croissance exponentielle de la quantité de données, les anciennes approches chimiométriques ne permettent pas une exploitation complète à cause des problèmes computationnels ou même de compréhension par l'utilisateur humain final[5].

Volume de données

Il y a eu et il y aura encore dans les prochaines années une relation directe entre la vitesse de calcul des ordinateurs et de sa capacité de mémoire et la quantité de données sortant des instruments analytiques. Le développement des ordinateurs mène au développement des instruments analytiques, lesquels produisent des quantités de données de plus en plus énormes[5]. Ces données peuvent être tellement grandes que leur analyse chimiométrique deviens difficile voire impossible avec les ordinateurs courants. Donc, le fait que les ordinateurs puissent avoir des difficultés pour analyser toutes ces données n'est pas un problème récent et risque de ne jamais disparaître. Les données acquises par des techniques à «haut débit» ou à «large bande» (bio-puces, RMN-2D, GC-Masse, LC-Masse, Fluorescence-3D, …), ou résultant de la combinaison mathématique de différents types de spectres, peuvent aboutir à des matrices tellement énormes que cela crée des problèmes pour leur analyse multivariée avec des algorithmes traditionnelles et des ordinateurs habituels. Ces données

peuvent être appelées de méga-variées, dans le sens qu'ils peuvent posséder quelques millions de variables.

Il y donc intérêt de modifier les algorithmes utilisés pour des analyses multivariées, pour qu'ils puissent continuer de fonctionner même avec des matrices de grandes tailles.

De nombreuses approches ont été proposées pour essayer de traiter des matrices de données de grande taille. L'idée générale derrière la procédure présentée ici, «Principal Components Transform», est qu'une décomposition plein rang d'une matrice par Analyse en Composantes Principales donne origine à un nouveau jeu de variables, les Composantes Principales, qui contient nécessairement toute la variabilité (et donc toute l'information) contenue dans la matrice d'origine. Toute analyse multivariée appliquée à cette matrice de Composantes Principales doit donc aboutir aux mêmes résultats que l'analyse de la matrice d'origine. De la même façon, l'information contenue dans un ensemble de matrices est conservée dans l'ensemble des matrices de Composantes Principales créés par leur décomposition plein rang.

ANOVA-PCA

L'Analyse de variance (ANOVA)[1] permet de tester l'existence de différences significatives entre les groupes d'échantillons des différents niveaux d'un ou plusieurs facteurs, en comparant la variance inter-niveaux avec la variance intra-niveaux. Dans l'analyse univariée, pour décider sur l'acceptation de l'hypothèse nulle (qu'il n'y a pas de différences significatives entre les différents groupes d'individus, juste des différences aléatoires), on fait un test F de comparaison de ces deux variances. Dans le cas où la valeur de F calculée est plus petite que la valeur critique, le résultat du test n'est pas signifiant, et on en conclut donc que les groupes ne sont pas différents.

Dans le cas multivarié, une autre approche, l'ANOVA-PCA[7], a été proposée pour comparer les variances des facteurs par rapport à la variance résiduelle. La procédure est en deux étapes :

- une décomposition de la matrice de données en un ensemble de matrices représentant les facteurs faisant partie d'un plan d'expériences, et en une matrice de variances résiduelles.
- une analyse en composantes principales sur chacune des matrices des facteurs additionnée de la matrice de variances résiduelles.

Après avoir construit des ellipses de confiance autour des groupes de coordonnées factorielles correspondant aux niveaux pour le facteur, sur les deux premières PCs obtenues par la PCA, on peut décider de la signification de chacun des facteurs. Trois situations générales peuvent arriver :

- les niveaux se superposent, donc le facteur n'est pas considéré comme significatif.
- les niveaux sont bien séparés le long de PC1, ce facteur est alors la source dominante de variabilité, donc il est considéré comme significatif par rapport à la variance résiduelle.
- les niveaux sont séparés le long de PC2, ou même PC3, donc la variance résiduelle est supérieure à la variance attribuée au facteur, même si celui-ci contient de l'information.

Dans un premier travail au cours de cette thèse, la méthode ANOVA-PCA a été appliquée à des spectres NIR acquis sur des échantillons ayant été stockés sous différentes températures et conditions d'atmosphère selon un plan d'expériences afin de déterminer l'influence de ces différents facteurs sur la stabilité des échantillons au cours du temps.

Dans un deuxième travail, nous avons étudié des spectres MIR acquis sur des échantillons aqueux d'un polysaccharide naturel pour caractériser les propriétés d'absorption de ces gels. Au cours de cette étude, nous avons étendu la méthode pour permettre la prédiction de nouveaux échantillons et pour vérifier la validité des modèles.

Finalement, dans une dernière phase, nous avons utilisé une procédure de permutation aléatoire de l'affectation des individus aux niveaux de chaque Facteur pour tester la validité de la distinction entre Facteur significatif et bruit aléatoire indiquée par la méthode ANOVA-PCA.

Analyse de tableaux multiples

La méthode CCSWA est utilisée principalement dans l'analyse sensorielle pour l'étude de tableaux multiples. Cette méthode cherche à trouver des dimensions orthogonales communes à plusieurs matrices de données obtenues pour les mêmes échantillons. À chaque matrice est attribué un poids par rapport à son importance dans la construction de chaque dimension commune. La première Composante Principale obtenue à partir d'une matrice comprenant la variabilité pondérée de l'ensemble des différentes matrices des échantillons correspond à la première Composante Commune. On peut ensuite calculer les coordonnées des individus et les contributions des variables de chaque tableau ainsi qu'une mesure de la contribution de chaque tableau à cette Composante Commune. Après déflation de la matrice de variabilités pondérées, on peut calculer la Composante Commune suivante.

ANOVA-PCA est basée sur l'application d'une PCA à chacun des différents tableaux de «facteurs + résidus». Nous avons montré qu'en remplaçant l'étape PCA par une analyse simultanée de tous les tableaux par CCSWA, il est possible de mieux distinguer des Facteurs significatifs des Facteurs non-significatifs.

MIR-ATR chauffante

D'un autre coté, le développement des instruments de mesure sont aussi une source de créativité pour les chimistes analytiques. Des accessoires instrumentaux relativement récents permettent de développer des approches différentes mêmes pour des sujets d'étude considérés comme matures. C'est le cas de l'accessoire de Réflexion totale atténué (ATR) chauffante dans l'analyse des aliments, dont les nouvelles applications possibles sont nombreuses, y compris dans l'étude des huiles alimentaires et du vin.

Avec les huiles, nous avons profité de cet accessoire chauffant pour provoquer des modifications chimiques et simultanément acquérir des spectres, d'une façon bien plus rapide que dans les autres travaux comparables présentés dans la littérature. Quelques gouttes d'huile sont chauffées directement sur le cristal de l'ATR et des spectres moyen infrarouge (MIR) ont été acquis au cours du temps, à 3 températures différentes pour l'huile de tournesol et à la même température pour trois huiles différentes. Grâce au rapport surface/volume élevé pour les gouttes, cette façon de chauffer accélère les réactions d'oxydation et d'isomérisation des huiles, par rapport aux autres travaux publiés dans la littérature.

La spectroscopie moyen-infrarouge avec l'ATR chauffante a aussi été utilisée dans l'analyse de vin pour étudier les effets de différents traitements technologiques, dont la micro-oxygénation et l'ajout de copeaux de bois. A cause des difficultés d'analyser en MIR des échantillons avec un large pourcentage d'eau suite à sa forte absorption dans cette région spectrale, les échantillons de vin ont été séchés directement sur le cristal par chauffage doux avant l'acquisition du spectre. Après l'évaporation de l'eau et de l'éthanol majoritaires, d'autres pics correspondants aux autres composantes du vin normalement cachées ont pu être détectés.

1. Méthodes multivariées

1.1 Analyse en composantes principales (PCA)

Objectif: ***exploration, compression des données***

L'optique commune des méthodes multivariées est que l'on peut considérer les échantillons comme étant des points dans une espace définis par des variables et que les coordonnées d'un individu sont données par ses valeurs pour chacune de ces variables. L'analyse en composantes principales (PCA) [1,2,3] est à la base de la plupart des méthodes de l'analyse multivariée. C'est une méthode qui vise à trouver les directions de plus grande dispersion des individus dans cet espace, l'idée étant que les directions de plus grande dispersion sont les directions les plus intéressantes. Si les variables ne contiennent que du bruit, les individus seront dispersés de façon homogène et uniforme dans toutes les directions. Une direction qui s'écarte d'une telle répartition sphérique risque donc de contenir de l'information.
Mathématiquement, la PCA calcule des combinaisons linéaires des variables de départ donnant de nouveaux axes qui contiennent la plus grande partie de la variabilité de la matrice de données de départ[3]. La PCA est une méthode non-supervisée, où aucune hypothèse n'est fait concernant des relations éventuelles entre les individus et entre les variables. Elle fait simplement l'hypothèse optimiste, mais raisonnable, que les directions (Composantes Principales) de plus grandes dispersions des échantillons sont les directions les plus intéressantes et que la variabilité associée avec ces directions correspond à de l'information. De plus, pour éviter d'avoir la même «information» dans plusieurs Composantes Principales, elles doivent toutes être orthogonales.

La décomposition matricielle de la PCA permet d'obtenir des matrices des coordonnées factorielles (ou «scores») et des contributions factorielles (ou «loadings»), selon l'équation **(1)** et la Figure 1:

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P}^{\mathbf{T}} + \mathbf{E} \qquad \textbf{(1)}$$

avec $\mathbf{X}$ (n, p) la matrice de données originale, $\mathbf{T}$ (n, k) les coordonnées factorielles des individus sur les Composantes Principales, $\mathbf{P}^{\mathrm{T}}$ (k, p) les contributions factorielles des variables du départ aux Composantes Principales. Si l'on calcule k Composantes Principales (PC), on n'obtient qu'une approximation de la matrice de départ $\mathbf{X}$ où $\mathbf{E}$ (n, p) est la matrice des écarts entre les valeurs originales et cette approximation. La quantité de variance contenue dans

chaque PC est proportionnelle à sa valeur propre («eigenvalue») donnée dans la matrice diagonale **S** (*k*, *k*), calculée par $\mathbf{S} = \mathbf{T}^{\mathbf{T}}.\mathbf{T}$.

Dans beaucoup de cas simples, un maximum de 3-4 PCs est suffisant pour expliquer la variabilité intéressante des données. Cependant, le nombre de PCs peut être choisi par différentes procédures, notamment la validation croisée.

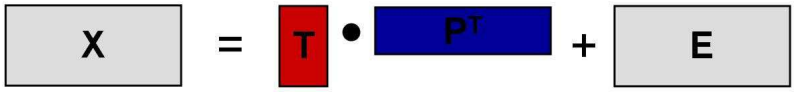


Figure 1 : Décomposition matricielle de la PCA. Les coordonnées **T** et les contributions factorielles **P** contiennent toute l'information importante par rapport aux objets et aux variables, respectivement. Les écarts **E** contiennent la variabilité résiduelle, pas important pour décrire le comportement des échantillons. Les dimensions des rectangles correspondent au cas des matrices avec $k < p$.

Les PCs sont simplement des entités mathématiques qui peuvent représenter, après un choix intelligent d'un ensemble représentatif, la matrice **X**. Au même temps, tout ce qui n'est pas important pour la description des données se trouve dans la matrice des résidus. Les PCs ne représentent pas forcément des entités physico-chimiques réelles présentes dans les données, mais plutôt des directions orthogonales de plus grande variabilité.

D'un point de vue géométrique, la PCA peut être plus facilement comprise comme une méthode de rotation des données pour que l'observateur soit le mieux placé pour comprendre les relations entre les individus. Les coordonnées factorielles permettent de projeter les individus sur des plans construits à partir des PCs, où l'on peut éventuellement détecter des repartions structurées des objets, la formation de groupes ou la présence d'individus aberrants. Comme on peut voir dans la Figure 2, la direction d'observation des données peut avoir une grande influence sur la compréhension de ce que l'on observe.

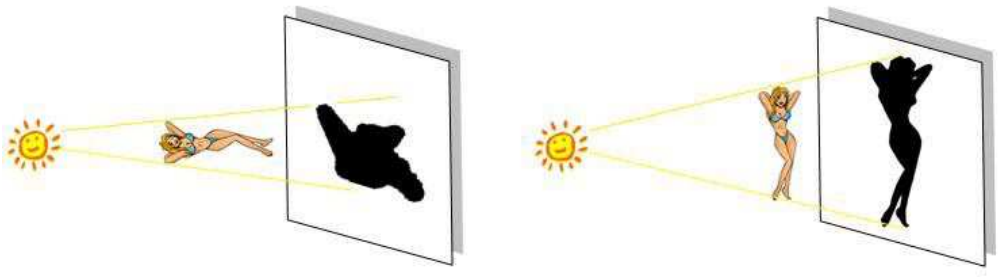


Figure 2 : Projection d'une figure sur un plan. La projection de gauche ne permet pas de comprendre ce qui est représenté, tandis qu'avec celle de droite, c'est évident.

Actuellement il existe plusieurs algorithmes pour le calcul de la PCA, dont les plus utilisées[8] sont le NIPALS[9,10] («Nonlinear iterative partial least squares») et le SVD[11] («Singular value decomposition») qui utilisent la matrice **X**; et les algorithmes POWER[12] et EVD[8] («eigenvalue decomposition») qui décomposent la matrice $\mathbf{X.X^T}$. Quand les données sont centrées par colonne, $\mathbf{X.X^T}$ est une matrice de variances-covariances et quand elles sont centrées et puis divisées par l'écart-type c'est une matrice de corrélations. Bien que les algorithmes soient différents (par exemple, SVD et EVD calculent toutes les Composantes Principales d'un seul coup, tandis que NIPALS et POWER les calculent séquentiellement), les résultats obtenus sont les mêmes (sauf, peut-être pour le signal).
La décomposition par PCA fait partie de la plupart des méthodes linéaires multivariées basées sur des projections, telles que la PCR[13], SIMCA[3,14,15], PLS[18,21] et Analyse Factorielle [16].

L'algorithme itératif NIPALS pour calculer la PCA [17] est présenté ci-dessous. Le vecteur des coordonnées factorielles **t** et les vecteurs propres **p** sont déterminés de façon à minimiser la somme des carrées des résidus **E**. Une contrainte doit être observée, c'est-à-dire que la norme de **p** soit égale à 1. Les PCs sont calculées successivement l'une après l'autre. En considérant une matrice de données **X** prétraitée de la façon souhaitée, et centrée par variable:

(1) $\mathbf{t} = \mathbf{x_i}$	Définir initialement **t** égal à une colonne de la matrice **X**
(2) $\mathbf{p} = (\mathbf{X^T.t}) / (\mathbf{t^T.t})$	Calculer le poids de chaque variable de **X** dans cette PC par projection de chaque colonne de **X** sur **t**
(3) $\mathbf{p} = \mathbf{p} / \|\mathbf{p}\|$	Normaliser le vecteur des poids pour chaque variable
(4) $\mathbf{t}_{ancien} = \mathbf{t}$	Garder cette valeur de **t** pour comparaison ultérieur
(5) $\mathbf{t} = (\mathbf{X.p}) / (\mathbf{p^T.p})$	Calculer la projection de chaque échantillon sur cette PC
(6) $d = \|\mathbf{t} - \mathbf{t}_{ancien}\|$	Vérifier la convergence. Si **t** ne change plus ($d < seuil$), on continue vers l'étape (7), sinon on revient à (2)
(7) $\mathbf{E} = \mathbf{X} - \mathbf{t.p^T}$	On enlève la partie de **X** expliquée par les vecteurs **t** et **p** calculées pendant cette itération
(8) $\mathbf{X} = \mathbf{E}$	Pour calculer un nouveau PC, on met **X** égal aux résidus, et on recommence à partir de (1)

Les variances des PCs peuvent finalement être obtenues sur la diagonale de $\mathbf{S} = \mathbf{T^T.T}$.

1.2 Régression au sens des moindres carrées partielles (PLS)

Objectif : ***régression multivariée***

La régression PLS, ou régression au sens des moindres carrées partielles ou régression par projection sur des structures latentes (PLS)[18] est l'outil standard pour faire des étalonnages-prédictions en chimiométrie. Cette méthode a son origine vers 1975 avec H. Wold et a été étudiée et développée par plusieurs autres auteurs, comme S. Wold[19,20], Martens et Naes[18] ou Tenenhaus[21].

La PLS cherche à trouver les relations entre deux matrices à travers un modèle linéaire multivariée. Elle permet l'analyse des données quand il y a beaucoup de variables colinéaires, bruitées ou incomplètes dans les deux matrices. Grâce à sa philosophie de «soft modelling», son application peut fournir des résultats même quand il n'existe pas une bonne connaissance des données, ou la théorie fondamentale à expliquer n'est pas connue.

Cette régression est appliquée depuis quelques années dans plusieurs domaines en chimiométrie, dont la calibration multivariée, les relations structure-activité («quantitative structure-activity relationship», QSAR) et le suivi de procédés et optimisation («process monitoring and optimization») sont les plus répandus[3]. Les variables prédictives (**X**) sont souvent des spectres mais peuvent aussi être des mesures bio-physico-chimiques, la plupart du temps avec une forte colinéarité entre variables. Les réponses (matrice **Y** ou vecteur **y**) peuvent être de nature variée, dont des concentrations des substances à prédire, mais aussi des propriétés physico-chimiques, des activités biologiques, des paramètres industrielles, des matrices spectrales d'un autre domaine.

La construction de la relation entre les variables prédictives et les réponses est actuellement faite à partir des algorithmes NIPALS[22] ou SIMPLS[22]. La relation obtenue minimise l'erreur de la prédiction tout en maximisant les covariances de **X** et **Y**. La modélisation conjointe est obtenue selon les relations :

$$\mathbf{X} = \mathbf{1}.\bar{\mathbf{x}}^{\mathbf{T}} + \mathbf{T}.\mathbf{P}^{\mathbf{T}} + \mathbf{E} \qquad \textbf{(2)}$$

$$\mathbf{Y} = \mathbf{1}.\bar{\mathbf{y}}^{\mathbf{T}} + \mathbf{U}.\mathbf{Q}^{\mathbf{T}} + \mathbf{F} \qquad \textbf{(3)}$$

où **X** et **Y** sont respectivement les variables prédictives et les réponses, $\bar{\mathbf{x}}$ et $\bar{\mathbf{y}}$ sont les moyennes des colonnes, **1** la matrice unitaire. **T** et **U** contiennent l'information (les coordonnées factorielles ou «scores») par rapport aux observations, **P** et **Q** l'information (les contributions factorielles ou «loadings») par rapport aux variables. **E** et **F** représentent les matrices résiduelles.

La régression (relation interne) entre les coordonnées factorielles **U** et **T** peut être représentée par

$$\mathbf{U} = \mathbf{T} + \mathbf{H} \quad \textbf{(4)}$$

et donc, pour des matrices **X** et **Y** centrées,

$$\mathbf{Y} = \mathbf{T}.\mathbf{Q}^T + \mathbf{F}^* \quad \textbf{(5)}$$

Avec **H** et **F*** comme des matrices résiduelles.

En projetant les matrices **X** et **Y** sur des espaces orthonormées définies par des matrices des coordonnées factorielles **T** et **U**, on élimine les problèmes de colinéarité normalement rencontrés lorsqu'on utilise les variables du départ.
L'algorithme pour le calcul de la régression PLS[23] est présenté ci-dessous. Les variables latentes (LV) sont calculées successivement jusqu'au nombre souhaitée:

(1) $\mathbf{u} = \mathbf{x_i}$	Définir initialement **u** égal à une colonne de la matrice **X**
(2) $\mathbf{w}^T = \mathbf{u}^T.\mathbf{X} / (\mathbf{u}^T.\mathbf{u})$	Les poids de chaque colonne de **X** dans cette LV sont calculés
(3) $\mathbf{w_{ancien}} = \mathbf{w}$	(alternativement, on peut garder **w** pour vérifier la convergence)
(4) $\mathbf{w} = \mathbf{w_{ancien}} / \|\mathbf{w_{ancien}}\|$	Les poids sont normalisés
(5) $\mathbf{t} = \mathbf{X}.\mathbf{w} / (\mathbf{w}^T.\mathbf{w})$	Projections des objets de **X** sur l'espace du vecteur **w**
(6) $\mathbf{q} = \mathbf{t}^T.\mathbf{Y} / (\mathbf{t}^T.\mathbf{t})$	Les poids de chaque colonne de Y dans ce LV sont calculés
(7) $\mathbf{u_{ancien}} = \mathbf{u}$	On garde le vecteur **u** pour comparaison
(8) $\mathbf{u} = \mathbf{Y}.\mathbf{q} / \|\mathbf{q}^T.\mathbf{q}\|$	Projections des objets de **Y** sur l'espace du vecteur **q**
(9) $d = \|\mathbf{u} - \mathbf{u_{ancien}}\|$	Vérifier la convergence. Si **u** (ou **w**) ne change plus ($d < seuil$), on continue vers l'étape (9) sinon on revient à (2)
(9) $\mathbf{p}^T = \mathbf{t_n}^T.\mathbf{X} / (\mathbf{t}^T.\mathbf{t})$	On calcule le vecteur **p** de contributions factorielles de **X**
(10) $\mathbf{E} = \mathbf{X} - \mathbf{t}.\mathbf{p}^T$	On soustraie la variance de **X** calculée pendant l'itération
(11) $\mathbf{F} = \mathbf{Y} - \mathbf{t}.\mathbf{q}^T$	On soustraie la variance de **Y** calculée pendant l'itération
(12) $\mathbf{X} = \mathbf{E}$ et $\mathbf{Y} = \mathbf{F}$	On garde tous les vecteurs **t**, **u**, **p**, **w**, **q**. On met les matrices **X** et **Y** égales aux résidus respectivement, et on recommence à partir de (1) pour une nouvelle variable latente

À partir des vecteurs obtenus, on peut calculer la matrice des coefficients de régression **B**, selon l'équation :

$$\mathbf{B} = \mathbf{W}.(\mathbf{P}^{T}.\mathbf{W})^{-1}.\mathbf{Q}^{T} \quad (6)$$

La prédiction des propriétés de nouveaux objets est obtenue à partir de

$$\mathbf{Y}_{\text{estimé}} = \mathbf{X} \,.\, \mathbf{B} + \mathbf{E} \quad (7)$$

Cette implémentation donne à la PLS une meilleure stabilité par rapport à d'autres modèles de régression et aussi une facilité d'interprétation due à la réduction du nombre de variables latentes nécessaires.

Les deux versions de la PLS, PLS1 et PLS2, présentent de petites différences, mais très importantes. Pour les cas où l'on veut modéliser une matrice **X** et sa relation avec un vecteur **y**, on utilise la PLS1. Si l'on a une matrice **Y** et on veut modéliser chacun des vecteurs, on peut prendre une colonne de **Y** à la fois et appliquer le PLS1, ce qui donne des résultats plus performantes que si on utilise la PLS2. Celle-ci est utilisée quand on veut modéliser toute une matrice **Y** d'un seul coup, mais la décomposition n'est pas optimisée pour un **y** particulier, ce qui peut diminuer la qualité des prédictions pour de nouveaux échantillons. La PLS2 trouve aussi une application énorme dans l'analyse discriminante, où elle est plus connue sous le nom de PLS-DA. En fait, dans la PLS-DA on applique la PLS2 à une matrice **Y** contenant des colonnes de valeurs binaires pour le classement des échantillons dans des groupes (1 si l'objet appartient au groupe représenté par cette colonne et 0 dans le cas contraire).

D'autres méthodes[24] étaient utilisées avant la généralisation de la PLS pour la modélisation des données avec plusieurs variables colinéaires. Les deux méthodes les plus utilisées étaient la régression sur composantes principales (PCR)[2] et la régression "Ridge" (RR) [3,18]. Dans la PCR, le but est d'appliquer une régression multiple aux coordonnées factorielles sélectionnées, suite à une PCA sur la matrice de données originales. Les variables de réponse ne participent donc pas dans la partie initiale de la méthode au calcul des variables latentes, ce qui fait qu'une partie de la variabilité apportée dans la phase de régression n'est pas de l'information explicative des variables de réponse. Dans la RR, certaines variables ne sont pas considérées et sont enlevées avant l'inversion de la matrice de variance-covariance. Ces deux méthodes ont presque toujours des performances inférieures à celles de la PLS.

Il existe de nombreuses variations et extensions de la méthode PLS standard, par exemples la «PLS-Discriminant Analysis» (PLS-DA)[25,26,27] pour l'analyse discriminante basée sur une matrice **Y** de valeurs binaires d'appartenance, «PLS-cluster»[28] pour la classification hiérarchique prédictive, la «INLR (-PLS)»[29] pour la modélisation des données légèrement

non-linéaires, la «GIFI-PLS»[30] pour des données fortement non-linéaires, la «Hierarchical-PLS»[31] pour les cas où on a beaucoup de variables provenant de blocs différents et la «Orthogonal-PLS» (OPLS)[32,33,34] ou la variabilité de X orthogonale à Y est éliminée avant de réaliser la régression PLS, ce qui facilite l'interprétation des résultats.

1.3 Validation-croisée

Objectif: ***détermination du nombre optimale de variables latentes***

D'une certaine façon, la régression PLS comme d'autres méthodes de projection des données, donne des approximations d'une façon similaire aux polynômes de Taylor. On peut construire des modèles de plus en plus proches des données simplement en augmentant le nombre de variables latentes utilisées. Ceci entraîne une diminution des résidus avec l'augmentation du nombre de variables latentes (ou des composantes principales dans le cas de la PCR). Dans le cas de l'utilisation de la PLS comme méthode de régression, on n'a pas comme objectif la diminution des écarts par rapport à l'étalonnage mais par rapport à la prédiction. On veut s'approcher le plus des valeurs de la variable réponse, mais on veut utiliser seulement l'information pertinente dans le jeu de données d'étalonnage pour construire un modèle de prédiction qui minimise les écarts pour de nouveaux échantillons. Tout ce qui est variabilité intrinsèque (bruit) des échantillons et qui n'a pas de rapport avec la variance intéressante pour la description des phénomènes étudiés ne doit pas être inclus dans le modèle, sous peine de sur-ajustement («overfitting») du modèle. Cet effet correspond à une augmentation de l'exactitude et de la précision de l'étalonnage, mais une diminution pour les prédictions. Le sous-ajustement correspond à un modèle avec un nombre de variables latentes inférieure à ce qu'il faut pour inclure toute la variabilité nécessaire pour minimiser les erreurs de prédiction. La méthode PLS est donc fortement dépendent d'un judicieux choix du nombre de variables latentes à inclure dans le modèle, ce qui implique un équilibre entre l'extraction de la variabilité des matrices (ajustement) et un bon pouvoir prédictif[2,19].

Plusieurs méthodes existent pour choisir le nombre de variables d'un modèle de régression et pour évaluer l'incertitude des modèles[35,36,37], dont la validation croisée[13,18] , le «jack-knifing»[38,39], la «bootstrapping»[40] et le test des permutations[3,41,42]. La plus utilisée, la validation croisée («cross-validation»), estime l'exactitude des prédictions pour des modèles avec un nombre de variables latentes croissant. Dans son implémentation la plus simple, la matrice est partitionnée plusieurs fois en deux parties, une (échantillons d'étalonnage) est

utilisée pour construire des modèles d'étalonnage avec différents nombres de variables latentes, et l'autre (échantillons de validation) pour déterminer les erreurs de prédiction. Après avoir appliqué cette procédure à tous les échantillons, on calcule une valeur moyenne pour la somme des carrées des erreurs résiduelles de prédiction, ou PRESS («Predicted residual error sum of squares») pour les modèles avec différents nombres de variables latentes. La dimensionnalité optimale pour le modèle de régression PLS est celle qui minimise cette valeur de PRESS.

Il y a plusieurs variantes de la validation croisée[18,43,44] par rapport à différentes facteurs, comme le nombre d'objets faisant partie des groupes d'étalonnage/validation, mais d'une façon générale, la méthode fonctionne comme indiqué dans la Figure 3. Pour simplifier, on considère le cas d'une seule variable réponse, donc $q = 1$ et chaque groupe de validation ne contient qu'un seul échantillon (procédure «leave-one-out»), donc $m = 1$.

Procédure pour la validation croisée

(1) les matrices **X** et **Y** commencent par être partitionnées horizontalement en deux jeux, dites d'étalonnage (ou d'entraînement) et de validation. Le nombre d'objets dans les deux jeux peut varier en fonction de l'algorithme utilisé. Le jeu de données d'entraînement $\mathbf{X_1}$ est utilisé pour construire des modèles PLS avec un certain nombre de variables latentes (entre 1 et *k*). On obtient donc une matrice $\mathbf{B_1}$ avec *k* vecteurs -colonne de coefficients de régression pour des modèles avec entre *1* et *k* variables latentes.

(2) Le jeu de données de validation $\mathbf{X_{p1}}$ est utilisé pour la prédiction avec chacun des modèles, donnant une valeur prédite pour la variable réponse ($\hat{\mathbf{y}}_{1k} = \mathbf{X_{p1}} \bullet \mathbf{b}$) pour chaque échantillon et chaque modèle.

(3) La différence entre les valeurs observées «réelles» et les valeurs prédites est calculée et mise au carré, pour donner des résidus carrés.

(4) On garde ces erreurs carrées pour chacun des modèles avec entre 1 et *k* variables latentes.

(5-6) On recommence la procédure avec une partition différente des échantillons et on la répète pour toutes les partitions possibles, jusqu'à ce que tous les objets aient servis pour la prédiction.

(7) À la fin, on obtient le PRESS, la somme des carrées des erreurs de prédiction pour chaque modèle. Le modèle donnant le PRESS minimum est retenu.

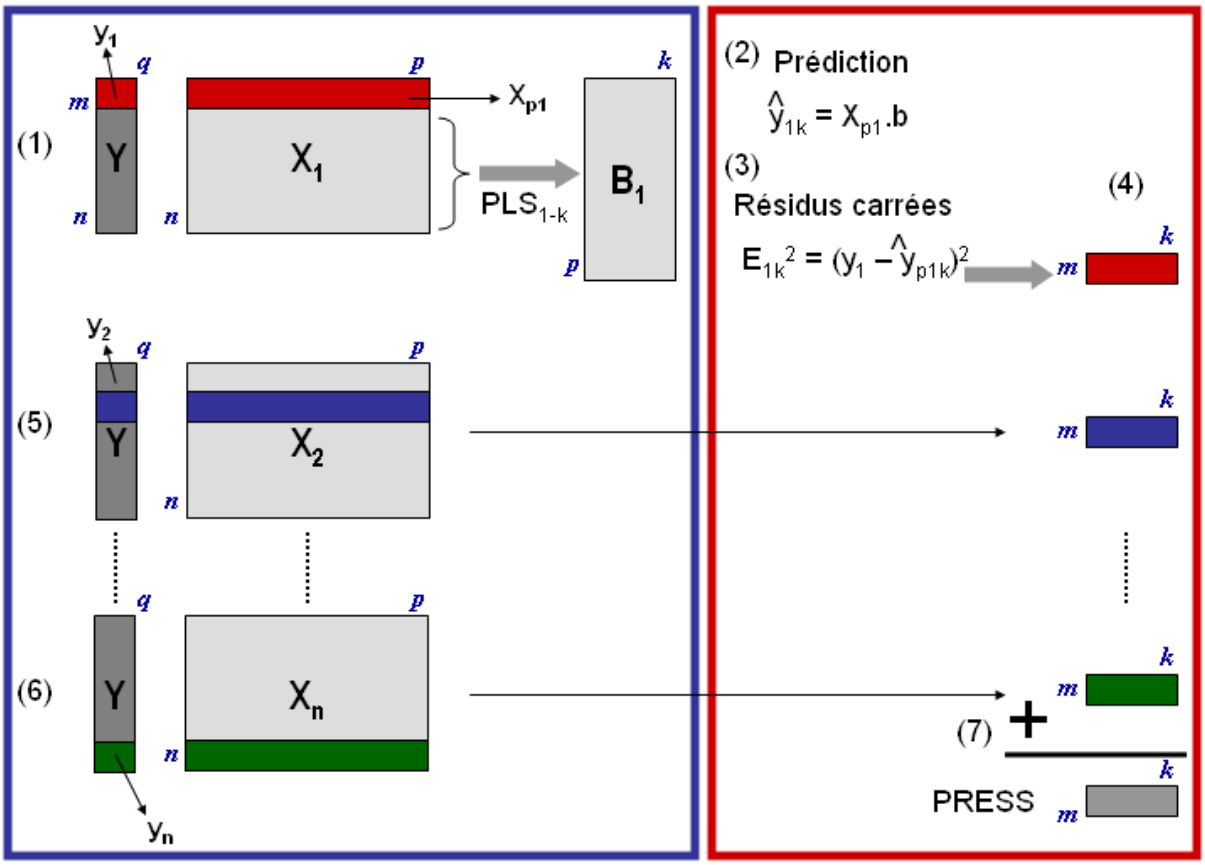


Figure 3 : Schéma général de la validation croisée. À gauche, en bleue, la partie de construction des modèles PLS avec entre 1 et k variables latentes et l'obtention des coefficients $\mathbf{b_k}$ pour chacun. À droite, en rouge, la prédiction de l'échantillon qui n'a pas fait partie de la calibration, le calcul des résidus carrées et du PRESS.

A cause du grand nombre de modèles construits au cours de la validation-croisée, les calculs peuvent prendre beaucoup de temps et nécessiter beaucoup de mémoire de l'ordinateur.
Pour éviter ce problème pour la PLS lors de la validation-croisée, nous avons démontré qu'il est possible et avantageux d'utiliser la PCT [45], ce qui réduit énormément le temps de calcul et les besoins de mémoire pour les cas où $n << p$.

1.4 Test des Permutations

Objectif: ***tester statistiquement des modèles***

La validation croisée, bien appliquée, donne une bonne estimation de l'erreur de prédiction de la régression d'un modèle. Pourtant, elle ne donne pas une indication sur la signification statistique du pouvoir prédictif estimé[3].
Pour avoir une estimation de la confiance sur un modèle de régression (avec un certain nombre de variables latentes dans le cas de la PLS), un test des permutations («permutation test») peut être utilisée. Donc, une fois la dimensionnalité optimale déterminée par validation

croisée, ce test, qui peut aussi être accéléré à l'aide de la méthode PCT, peut être appliqué pour estimer le niveau de signification du modèle.

Procédure pour le test de permutations

On décrit le fonctionnement du test des permutations en montrant l'exemple pour une régression PLS.

(1) on change aléatoirement toutes les positions des valeurs de **y** et on établit le modèle entre la matrice **X** et ce vecteur **y** perturbé pour un jeu d'échantillons d'étalonnage et on calcule les valeurs prédites avec le jeu de validation.

(2) on retient les erreurs de prédiction (PRESS) pour le meilleur modèle

(3) on répète la procédure de perturbation plusieurs fois (100 ou 10,000 fois, par exemple) et on retient les erreurs de tous les modèles pour chacun des couples de groupes d'étalonnage/prédiction

(4) on compare la valeur de PRESS du «vrai» modèle de validation croisée avec la distribution des erreurs des modèles «perturbés».

Dans le cas d'une discrimination, on peut utiliser la distance entre les barycentres des groupes, avec ou sans permutation des appartenances.

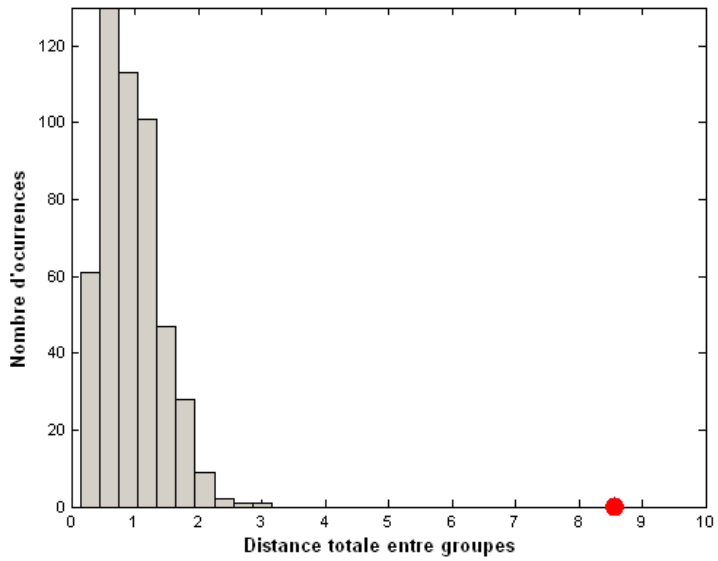


Figure 4: Exemple d'histogramme avec la valeur réelle de distance entre groupes (rouge) et la distribution obtenue pour les distances des modèles calculées après 500 permutations aléatoires des groupes.

On peut utiliser des histogrammes pour visualiser les proportions de prédiction pour chaque valeur de PRESS ou chaque distance entre barycentres. Cette méthode est un test unilatéral («one-sided test») pour des erreurs du type α, autrement dit, la possibilité, à un certain

pourcentage de confiance, de ne pas accepter H_0: *les modèles ne diffèrent pas*. Dans le cas où la valeur réelle est inférieure à 5% des valeurs «aléatoires», on n'accepte pas H_0, donc les prédictions du modèle sont significatives à 95%.

Le pouvoir de ce test est, évidemment, influencé par le nombre de modèles calculés et par le type de méthode sur lequel il est appliqué (régression, discrimination ou autre). A partir d'un certain nombre de modèles aléatoires, la distribution ne change plus significativement. Dans les cas où le nombre de variables latentes du modèle est évident, la distance entre l'erreur du «vrai» modèle et la limite de la distribution des modèles après permutation est visible, et cette distribution est mono-modale, il n'y a pas besoin de beaucoup de permutations. Dans les cas contraires, quand les valeurs de l'erreur du «vrai» modèle sont proche ou inclues dans les limites de la distribution ou que celles-ci ne sont pas simplement mono-modales où il y a des points isolés, il vaut mieux avoir une bonne estimation de cette distribution, donc il faudra réaliser un plus grand nombre de permutations. Certains auteurs[3] indiquent entre 25 et 100 permutations pour la PLS, tandis que d'autres, en appliquant le test à une autre méthode[46], considèrent que 100 permutations ne sont pas suffisantes.

Comme indiqué ci-dessus, le test de permutations peut être utilisé avec d'autres méthodes de prédiction que la régression[47]. Dans ce travail, nous l'avons utilisé pour vérifier la validité de la discrimination entre des groupes d'individus en utilisant une méthode modifiée à partir de l'ANOVA-PCA par la comparaison des distances «réelles» entre les barycentres de groupes et les distributions de distances après permutations aléatoires.

1.5 Analyse des produits externes

Objectif : combinaison de matrices correspondantes aux mêmes objets

L'analyse de produits externes («Outer Product Analysis», OPA)[23] permet de combiner mathématiquement des matrices de données caractérisant les mêmes objets, ce qui peut, après l'application des méthodes multivariées, faciliter la compréhension des liens entre les réponses dans ces deux domaines et les relations entre les objets. Cette combinaison de domaines peut s'avérer utile dans plusieurs situations, notamment dans l'analyse instrumentale.

L'OPA a été proposée en 1997 par Barros et al.[48] pour l'analyse conjointe des spectres moyen et proche infrarouge. Dans ce travail, les auteurs trouvent des relations entre les deux domaines en calculant les produits externes et en réalisant une analyse de variance $(ANOVA)^2$ fondée sur un critère de classification. Les relations entre les deux domaines sont observables

dans des tracés de variances inter- et intra-groupes. D'autres publications[49,50] ont suivi l'idée initiale en appliquant différentes méthodes multivariées à la matrice de produits externes, par exemple, la quantification de subérine dans le liège et sa caractérisation par étalonnage multivarié. Dans cette étude, le produit externe a été calculé entre des spectres de résonance magnétique nucléaire et infrarouge[51]. La combinaison OP a même été traitée par des méthodes d'analyse multivoie, par (PARAFAC)[52], aussi bien lors d'études en métabonomiques en combinant des signaux LC-MS avec des spectres 1H-NMR[53] que d'autres méthodes instrumentales[54].

Procédure pour l'OPA

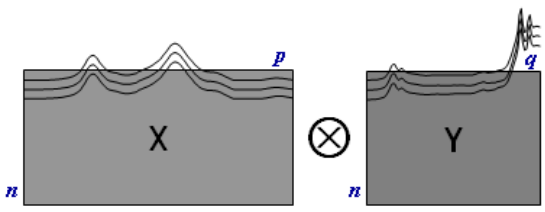
Pour calculer le produit externe OP entre deux matrices on suit la Figure 5.

(1) on définit deux matrices de données **X**(*n, p*) et **Y**(*n, q*), provenant de l'acquisition de 2 types de spectres (les mêmes ou différents) sur les mêmes produits. Les lignes *i* des deux matrices sont dans le même ordre pour qu'elles correspondent aux mêmes échantillons. L'opérateur pour le produit externe (OP) est représenté par ⊗. Notez que le nombre de colonnes dans les deux domaines peut être différent.

(2) on transpose chaque vecteur-ligne *i* de la matrice **X** et on le multiplie par tous les éléments de la même ligne *i* de la matrice **Y**. La matrice individuelle résultante correspond à une pondération des valeurs d'un vecteur par les valeurs de l'autre pour cet échantillon.

(3) après avoir calculé les matrices OP pour toutes les lignes, on peut les concaténer verticalement, pour obtenir un cube de matrices OP.

(4) ce cube de matrices OP peut être déplié de 3 façons différentes pour obtenir sur les lignes de la matrice dépliée soit les objets des matrices originaux, soit les variables originales de la matrice **X**, soit celles de la matrice **Y**. A ces 3 matrices **M** on peut appliquer des méthodes multivariées.



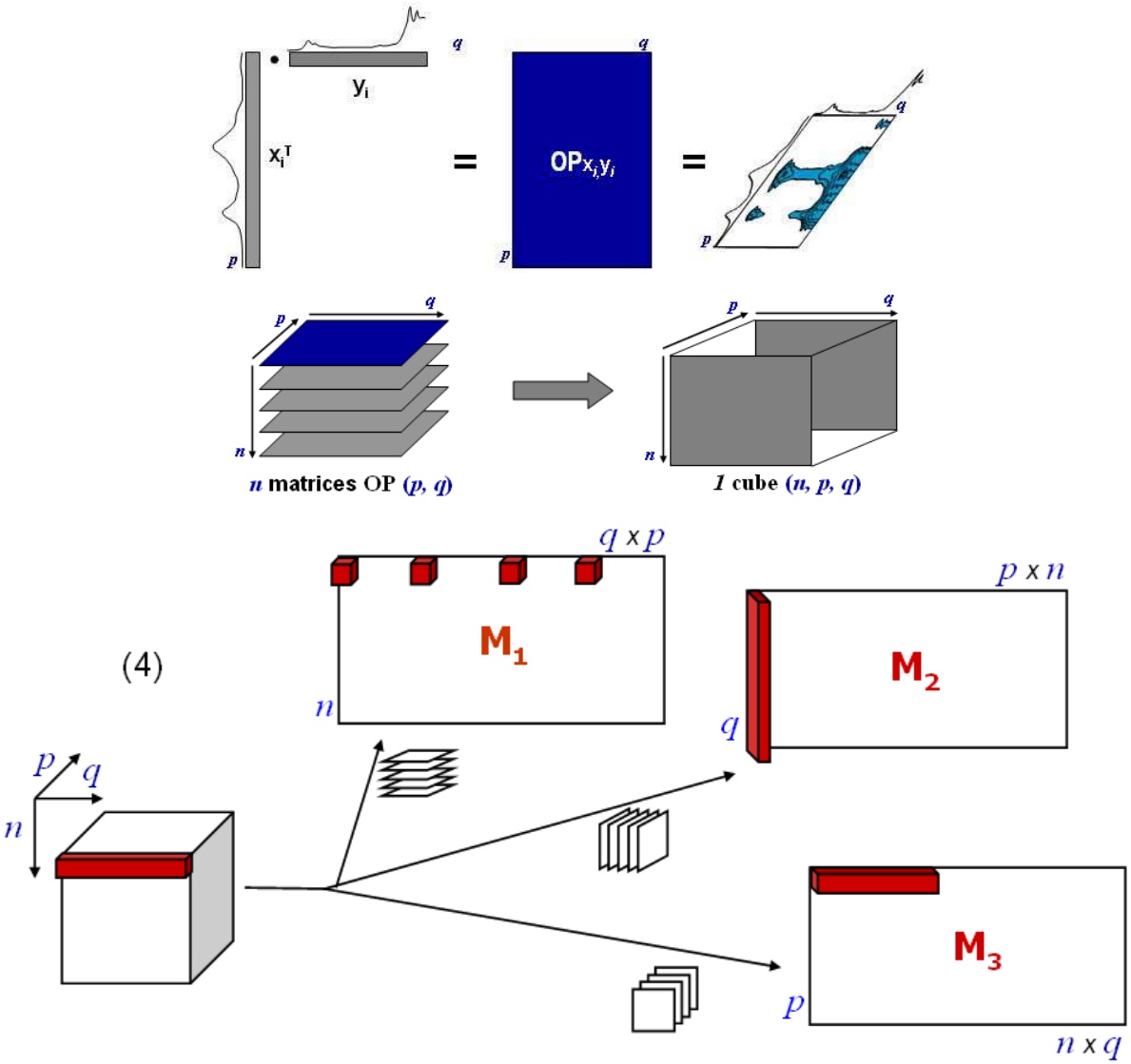


Figure 5: calcul d'un cube de matrices OP à partir de deux domaines et modes de dépliage.

Puisque ces 3 matrices dépliées, M1, M2 et M3, contiennent la même variabilité et décrivent différemment le même espace échantillons/variables, leur analyse par des méthodes multivariées fournit différentes informations complémentaires.

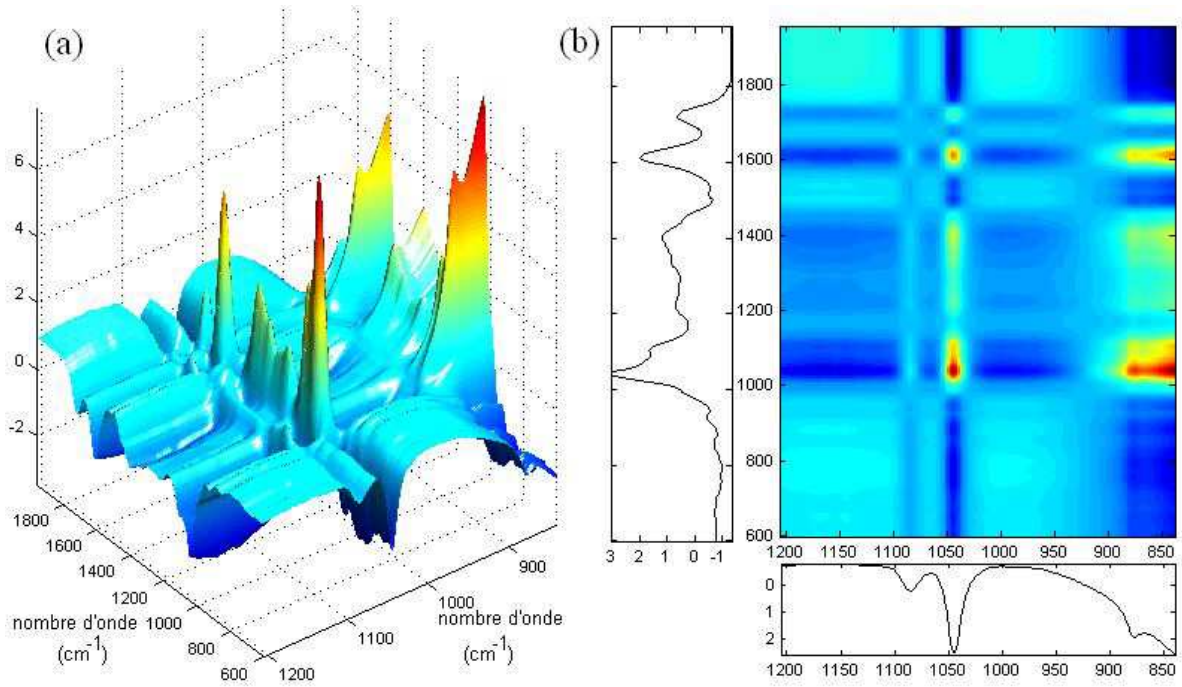


Figure 6 : Exemple de matrice OP calculée à partir de données MIR de vin obtenues par différentes méthodes. Représentation de la matrice OP par une (a) surface de réponse ; (b) les deux domaines d'origine (noir et blanc) et le graphique de contour de (a).

Pour la combinaison de plus de deux domaines, on calcule les matrices OP entre deux domaines et on les combine avec le domaine suivante. Un exemple de matrice OP est présenté dans la Figure 6. On peut immédiatement voir l'influence de cette pondération mutuelle des deux domaines sur la hauteur des pics.

1.6 Le paradigme PCT

Objectif : compression exacte, accélérer les calculs et diminuer besoins de mémoire

L'adjectif «multivariée» n'indique pas le nombre d'objets ou de variables contenus dans les matrices de données analysées, mais elles sont normalement de l'ordre de quelques centaines. Par rapport aux préfixes normalement utilisés dans les sciences, on doit pouvoir classifier les matrices analysées par rapport au nombre de variables. Donc, pour les cas où elles sont dans l'ordre des milliers, les matrices peuvent être appelées megavariées et dans le cas ou elles sont de l'ordre de millions, elles peuvent être appelés gigavariées, ce qui est encore accord avec des propositions d'autres chimiométriciens[3] cités ultérieurement. Ces données sont à l'origine des matrices tellement grandes qu'elles peuvent exiger d'énormes quantités de mémoire d'ordinateur ainsi que des temps de calcul excessifs. Les matrices peuvent même être si grandes que les ordinateurs ne disposent pas assez de mémoire pour le calcul des méthodes ou n'arrivent pas à faire les calculs dans une période de temps acceptable. Il y a donc besoin de procédures qui puissent réduire l'utilisation de la mémoire et les temps de calcul. Dans une première approche, plusieurs options sont possibles, comme la sélection de variables ou la sélection d'objets représentatifs de l'ensemble des données ou l'utilisation de transformations pour compresser les données.

Certaines méthodes permettent de travailler sur des jeux de données très larges, provenant par exemple de la génomique, la protéomique et la métabonomique, en recherchant un ensemble réduit de variables représentatives (après sélection par PCA, PLS ou classification hiérarchique). Ces méthodes ne visent pas une compression exacte des données, ce qui n'est pas le cas de la méthode qui sera présentée ici[5].

En ce qui concerne les méthodes à base de transformation, la Transformée de Fourier ou la Transformée en ondelettes («wavelets»)[55,56] ont deux défauts. Elles ne s'appliquent qu'à des données où les vecteurs-lignes sont des signaux. De plus, elles nécessitent le choix du nombre de coefficients (variables transformées) à retenir – si ce nombre est trop faible, on risque de perdre de l'information lors de la compression ("lossy data compression"). Il serait préférable de disposer d'une compression sans possibilité de perte d'information ("lossless data compression").

Les méthodes «kernel» ont aussi été utilisées pour faire la PCA, la PLS, etc. d'une façon plus rapide. Cette procédure décompose la matrice $\mathbf{X}.\mathbf{X}^T$ au lieu de $\mathbf{X}^T.\mathbf{X}$ ou de $\mathbf{X}$ ce qui, dans le cas de $p>>n$, réduit la taille de la matrice en mémoire et accélère les calculs. Les mêmes

vecteurs de coordonnées et contributions factorielles peuvent être obtenues, ainsi que la variance associée à chaque PC. Des études comparatives[8,57] sur la vitesse de plusieurs algorithmes PCA classiques modifiés pour travailler sur un "kernel" $\mathbf{X}.\mathbf{X}^T$ ont été réalisées. Elles ont démontré que dans la majorité des cas, le meilleur choix, quand le nombre de variables (p) >> nombre de objets (n), est d'utiliser l'algorithme kernel-EVD et que la différence pour le kernel-SVD n'est pas grande. L'algorithme kernel-NIPALS est le moins rapide. Ceci pose un problème pour l'analyse des grands tableaux avec valeurs manquantes, car les algorithmes itératifs comme le NIPALS sont ceux qui peuvent fonctionner sans problème avec une certaine proportion de valeurs manquantes. Quoi qu'il en soit, toutes ces méthodes nécessitent le calcul d'une matrice "kernel" à partir d'une très grande matrice, ce qui peut dans certain cas ne pas être possible ou peut nécessiter des temps de calcul très longs.

Au cours de cette thèse, nous avons publié deux articles au sujet de la PCT[58, 59].

1.6.1 PCT standard

Objectif: ***compression exacte des données***

La méthode PCT a été récemment proposée[45] pour accélérer les calculs et pour réduire les besoins en mémoire quand on applique des méthodes multivariées comme PLS et surtout lors de la validation croisée dans le cas de matrices énormes avec beaucoup plus de variables que d'objets (p >> n). Son utilité a aussi été démontrée plus récemment dans le cadre de régressions PLS2 entre deux matrices énormes[60]. La méthode est extrêmement simple à comprendre et à utiliser.

Procédure pour la PCT standard

(1) Les relations entre les p individus d'un jeu de données restent les mêmes lorsque l'on passe d'un espace multidimensionnel défini par les n variables de départ à un espace défini par les k Composantes Principales provenant d'une PCA à rang entier.

(2) Le nombre maximal de Composantes Principales k étant le plus petit de p et n, il est avantageux de remplacer la matrice de départ (taille $n*p$) par une matrice plus petite ($n * k$) de coordonnées factorielles sur les PCs.

(3) Dans le cas des matrices avec beaucoup plus de variables que d'objets, l'analyse multivariée de cette matrice de coordonnées factorielles est plus rapide et nécessite moins de mémoire centrale.
(4) Il est possible de recalculer toutes les variables et tous les critères que l'on aurait eu si l'on avait utilisé les variables de départ.
(5) On obtient exactement les mêmes résultats que si l'on avait appliqué la méthode multivariée sur les données originales

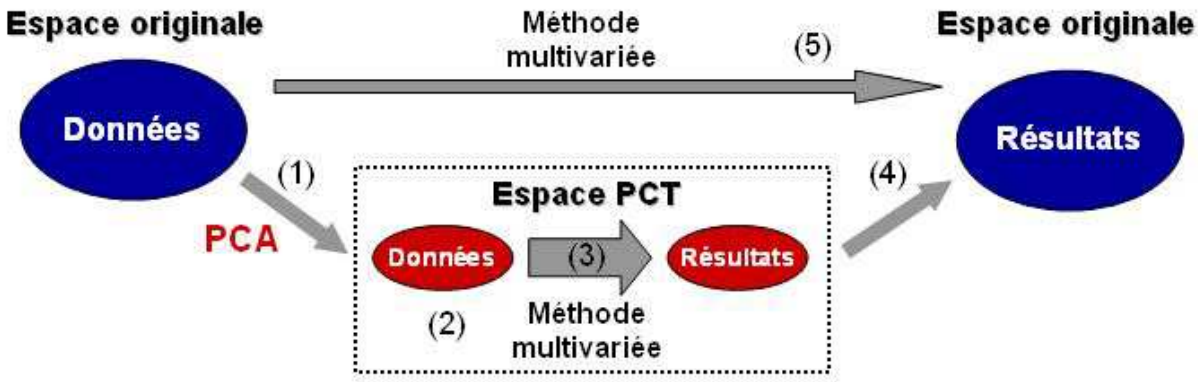


Figure 7 : schéma de l'application d'une méthode multivariée réalisée à l'aide de la seg-PCT (1-4) par rapport à une méthode multivariée réalisée de la façon standard (5)

Pour n'importe quelle méthode multivariée et n'importe quel type de données, on obtient exactement les mêmes résultats à condition de faire une décomposition rang entier de la matrice de départ lors de la transformation PCT. Par rapport à d'autres méthodes de compression, cette transformation a donc l'avantage d'être sans perte, car elle conserve toute la variabilité des données d'origine. Cette transformation PCT est réversible car, après avoir réalisé des calculs multivariés dans l'espace des Composantes Principales, il est possible de revenir vers l'espace des variables de départ. En cela, la méthode PCT est quelque peu comparable à la transformée de Fourier.

1.6.1.1 Transformation inverse

Après avoir appliqué la décomposition de la matrice **X** par PCA à rang entier selon l'équation **(7)**, on peut travailler dans la base vectorielle "PCT".

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P}^{\mathbf{T}} \qquad \textbf{(8)}$$

où **X** (*n, p*) est la matrice originale de données, **T** (*n, n*) les coordonnées factorielles associées aux objets et **P** (*p, n*) les contributions factorielles associées aux variables de départ. On utilise la matrice **T** pour effectuer la régression PLS avec un nombre optimal *f* de variables latentes. la PLS fournit plusieurs matrices de résultats, dont les coordonnées factorielles $\mathbf{T}_{\mathbf{PLS\text{-}T}}$ (*n, f*), les contributions factorielles $\mathbf{P}_{\mathbf{PLS\text{-}T}}$ (*n, f*) et les poids $\mathbf{W}_{\mathbf{PLS\text{-}T}}$ (*n, f*). Ces matrices représentent les données dans la base PCT (d'où l'indice $_{\mathbf{PLS\text{-}T}}$). Dans la plupart des cas, on voudrait étudier ces vecteurs/matrices dans la base des variables d'origine ($_{\mathbf{PLS\text{-}O}}$). Les transformations inverses se font de la façon suivante[45]:

$$\mathbf{T}_{\mathbf{PLS\text{-}O}}(n, f) = \mathbf{T}_{\mathbf{PLS\text{-}T}}(n, f) \quad \textbf{(9)}$$

$$\mathbf{P}_{\mathbf{PLS\text{-}O}}(p, f) = \mathbf{P}(p, n) \,.\, \mathbf{P}_{\mathbf{PLS\text{-}T}}(n, f) \quad \textbf{(10)}$$

$$\mathbf{W}_{\mathbf{PLS\text{-}O}}(p, f) = \mathbf{P}(p, n) \,.\, \mathbf{W}_{\mathbf{PLS\text{-}T}}(n, f) \quad \textbf{(11)}$$

$$\mathbf{b}_{\mathbf{PLS\text{-}O}}(p, f) = \mathbf{P}(p, n) \,.\, \mathbf{b}_{\mathbf{PLS\text{-}T}}(n, 1) \quad \textbf{(12)}$$

Toutes les valeurs calculées de cette façon sont exactement les mêmes que celles obtenues par une PLS directement sur la matrice **X**.

Parfois, surtout lorsque le nombre de variables est très élevé, il est préférable de ne pas conserver en mémoire la matrice **P** (*p, n*). Dans ce cas, on peut la calculer, en cas de besoin, à partir de :

$$\mathbf{P} = [(\mathbf{T}^{\mathbf{T}}.\mathbf{T})^{-1}.(\mathbf{T}^{\mathbf{T}}.\mathbf{X})]^{\mathbf{T}} \quad \textbf{(13)}$$

On peut donc obtenir toutes les matrices de la PLS dans la base vectorielle des variables d'origine à partir de la matrice originale **X** et des coordonnées factorielles **T** données par la PCT. Si le nombre de variables est trop grand pour calculer tout **P**, il est possible de le calculer par morceaux en n'utilisant que des segments de la matrice **X** dans l'équation 13.

1.6.2 Seg-PCT

Objectif : ***compression exacte des matrices par partition***

Dans le même contexte de la PCT, nous avons étudié une extension de la méthode, nommée Seg-PCT («Segmented-PCT»), proposée aussi par Barros et Rutledge[61]. Cette extension de la méthode permet d'analyser de matrices de n'importe quelle largeur en principe, qui ne pourraient pas être décomposées d'un seul coup par PCA. L'idée de la Seg-PCT est de

segmenter la matrice originale de données horizontalement en plusieurs sous-matrices, de réaliser une PCA sur chacune de ces matrices et de concaténer horizontalement l'ensemble des coordonnées factorielles.

Procédure pour la Seg-PCT

(1) La matrice de données originale **X** est segmentée *z* fois au niveau des variables, pour donner *z* sous-matrices **X***z* (pouvant contenir des nombres de variables différents)

(2) Une PCA (rang entier) est appliquée à chacune de ces *z* sous-matrices, pour donner des *z* matrices de coordonnées factorielles. Les données sont maintenant dans un ensemble d'espaces définis par des bases PCT individuelles.

(3) Les *z* matrices de coordonnées factorielles sont concaténées horizontalement pour donner une matrice beaucoup plus petite, mais renfermant toute la variabilité de la matrice d'origine.

(4) Différentes méthodes multivariées peuvent être appliquées à cette matrice concaténée. Il est ainsi possible d'étudier par PCA, PLS, validation-croisée, etc. des matrices encore plus grandes qu'avec la PCT standard.

(5) On peut retransformer les variables vers l'espace original à l'aide des contributions factorielles obtenues avec chacune des transformations PCT initiales.

(6) Les résultats obtenus sont exactement les mêmes qu'avec les données originales

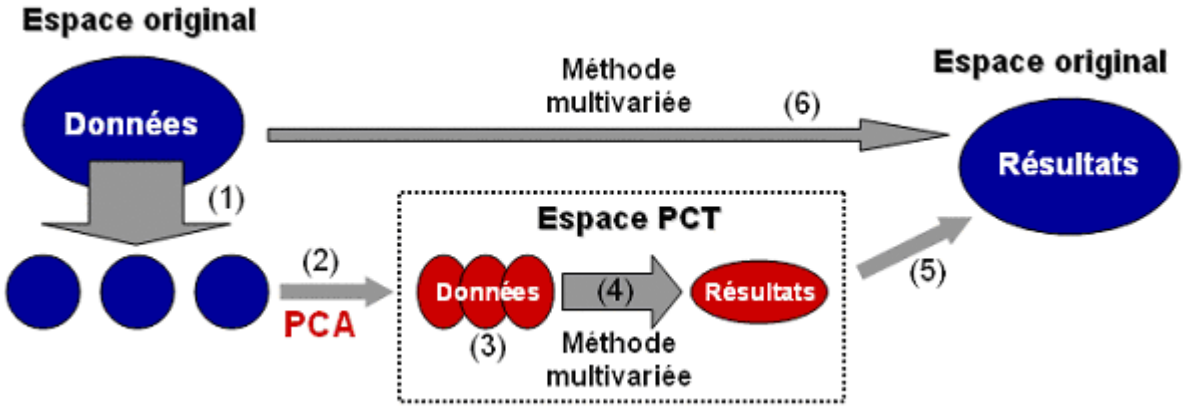


Figure 8 : Schéma de l'application d'une méthode multivariée appliquée à l'aide de la seg-PCT (1-5) par rapport à une méthode multivariée appliquée de la façon standard (6).

1.6.2.1 Transformation inverse

Comme pour la PCT-PLS dans la section précédente, on peut être intéressé d'avoir les matrices de résultats dans la base des variables d'origine. La procédure pour revenir aux

variables d'origine après la Seg-PCT-PLS est similaire à celle de la PCT-PLS. La seule différence est que l'on travaille à chaque fois sur une des partitions $\mathbf{X_Z}$ de la matrice originale et pas sur **X** complète. Donc, dans l'équation **(12)**, on utilise la partition $\mathbf{X_Z}$ pour laquelle on veut régénérer les coefficients $\mathbf{b}_{Z,\,PCT\text{-}O}$, les contributions factorielles $P_{Z,\,PCT\text{-}O}$ et $W_{Z,\,PCT\text{-}O}$ de façon analogue aux équations **(9–11)**. Comme pour la PCT standard, les coordonnées factorielles sont aussi les mêmes pour la Seg-PCT-PLS que pour la PLS standard, comme dans l'équation **(8)**.

Les méthodes hiérarchiques de Kettaneh et al.[6] et de Westerhuis et al.[62] proposent aussi une utilisation des coordonnées factorielles au lieu des données d'origine, sauf que l'ensemble de composantes principales possibles n'est pas utilisé dans sa totalité. Même si l'approche utilisée est similaire à la PCT, la philosophie et le but de son utilisation ne sont pas les mêmes.

Au cours de cette thèse nous avons publié un article sur la transformation Seg-PCT[59].

1.6.3 Procédure comparée entre l'OP-PLS standard et PCT-OP-PLS

(1) On applique une PCA individuelle (rang entier) sur chacun des domaines $\mathbf{X_1}$ et $\mathbf{X_2}$, pour obtenir les coordonnées factorielles $\mathbf{T_1}$ et $\mathbf{T_2}$, ainsi que les contributions factorielles $\mathbf{P_1}$ et $\mathbf{P_2}$ (mais dont on n'a pas besoin pour le moment). On utilise maintenant la base vectorielle PCT

(2) On calcule la matrice **PCT-OP** à partir de $\mathbf{T_1}$ et $\mathbf{T_2}$, beaucoup plus petite que la matrice **OP** obtenu à partir de $\mathbf{X_1}$ et $\mathbf{X_2}$.

(3) Lors d'une régression PLS entre la matrice **PCT-OP** dépliée et les variables à prédire, la détermination de la dimensionnalité optimale par validation croisée, «jack-knifing» ou «bootstrapping», les temps de calcul et les besoins en mémoire sont réduits de façon notable.

(4) On peut calculer un vecteur **b** de coefficients de régression fondée sur $\mathbf{T_1}$ et $\mathbf{T_2}$ pour le meilleur modèle PLS, et l'utiliser pour faire des prédictions.

(5) Si nécessaire, le vecteur **b** et chaque vecteur de contributions factorielles $\mathbf{l_i}$ aux variables latentes de la régression PLS peuvent être repliés pour donner de petites matrices $\mathbf{B_T}$ et $\mathbf{L_{Ti}}$ à partir desquelles on peut calculer les matrices $\mathbf{B_O}$ dans la base des variables d'origine.

(6) Cette transformation inverse se fait, comme indiquée dans la Figure 8, à l'aide des matrices $\mathbf{P_1}$ et $\mathbf{P_2}$, qui peuvent être recalculées par l'équation:

$$\mathbf{P}_A = [(\mathbf{T}_A^{\mathbf{T}}.\mathbf{T}_A)^{-1}.(\mathbf{T}_A^{\mathbf{T}}.\mathbf{X}_A)]^{\mathbf{T}} \qquad \textbf{(14)}$$

où A=1 ou 2, pour les domaines $\mathbf{X_1}$ ou $\mathbf{X_2}$.

(7) La matrice $\mathbf{B_O}$ permet d'évaluer les relations entre le vecteur **y** et les deux domaines originaux tandis que les matrices $\mathbf{L_{Oi}}$ permettent d'étudier les liens entre les combinaisons de variables de $\mathbf{X_1}$ et de $\mathbf{X_2}$ et la distribution des individus selon les Variables Latentes correspondantes.

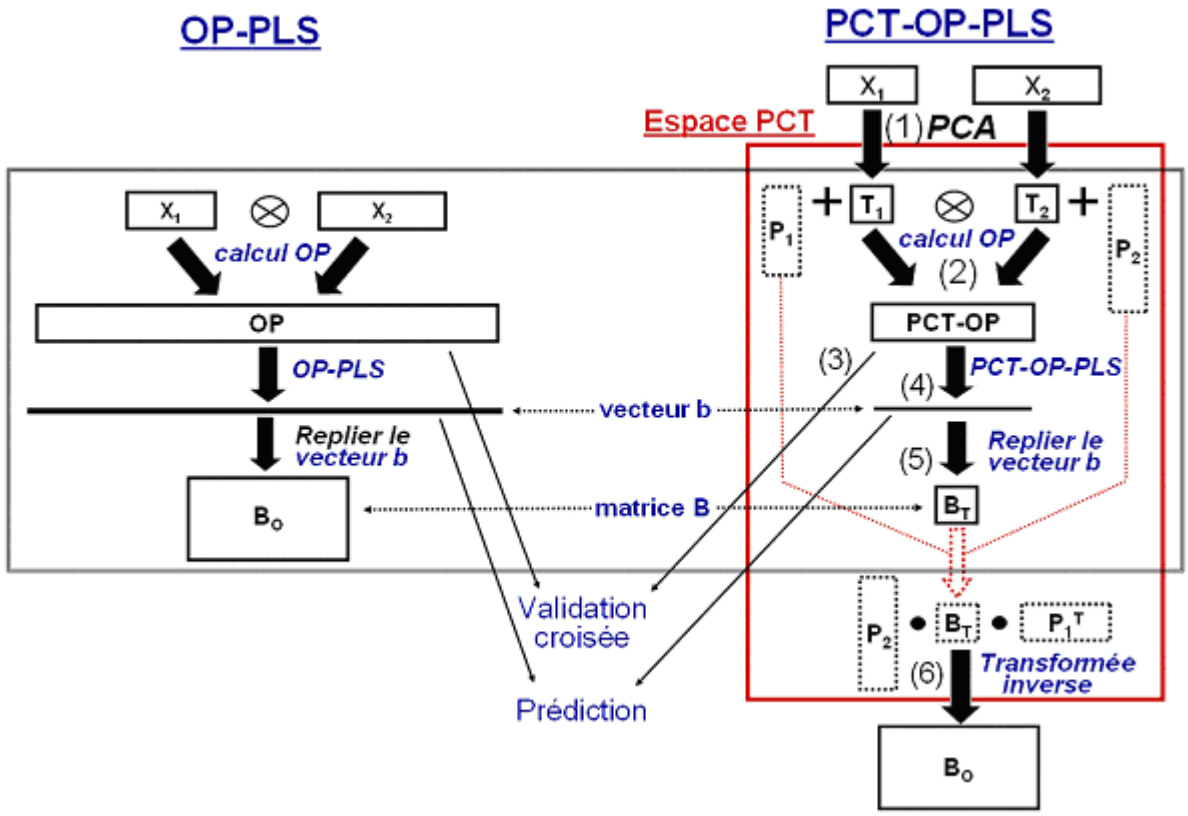


Figure 9: Schémas de l'OP-PLS standard (gauche) et de la PCT-OP-PLS (droite)

1.6.4 Exemples pratiques de la PCT

1.6.4.1 PCT-OP-PLS

La transformation PCT est spécialement adaptée pour utilisation avec la méthode de combinaison de matrices par le calcul du produit externe, dû au grand nombre de variables des matrices OP obtenu.

La façon de procéder pour l'utilisation de la PCT avec l'OP est montrée en pratique avec un exemple[63]. On va comparer les deux méthodes pour comprendre comment elles donnent exactement les mêmes résultats.

20 spectres de FTIR-PAS et de ^{13}C CP/MAS NMR de liège en poudre, avec différentes pourcentages de subérine, ont été acquis. Les deux domaines ont été combinés dans une matrice dépliée de produits externes et, par une régression PLS, mis en relation avec les concentrations de subérine, déterminées par une méthode gravimétrique. Le but de l'étude est de créer un modèle d'étalonnage pour la quantification de subérine dans le liège et aussi de relier les deux domaines spectraux ensemble pour mieux les comprendre.

Matrices d'origine

X1: FTIR-PAS avec une plage spectrale 1900-900 cm^{-1} (dimensions 20 x 260).

X2: ^{13}C CP/MAS NMR avec une plage spectrale 15.34-115.24 ppm (dimensions 20 x 256).

y: Pourcentage de subérine pour chaque échantillon (dimensions 20 x 1).

Les spectres sont centrés et réduits par ligne (prétraitement SNV[64]).

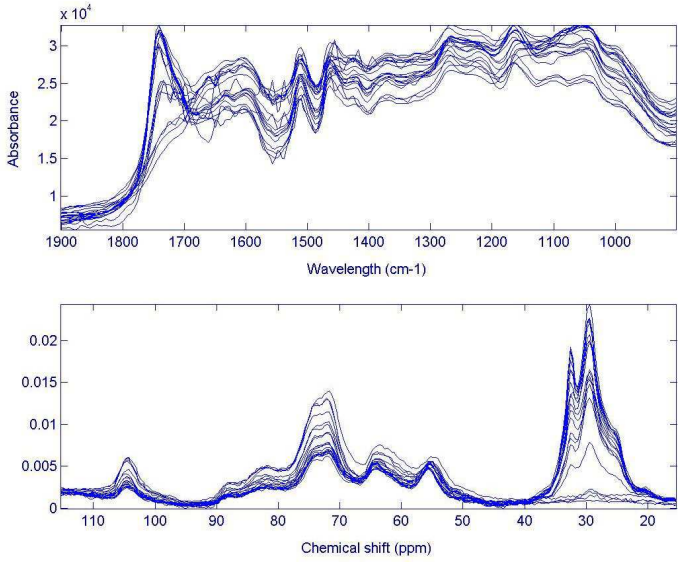


Figure 10 : Spectres de subérine acquis par FTIR-PAS (haut) et 13C CP/MAS NMR d'état solide (bas)

Procédure pour l'exemple Seg-PCT-OP-PLS

Les données ont été centrées par colonne et après une PCA sur chacun des domaines de départ, la matrice **PCT-OP** a été construite (dimension 20 x 400). Si ce n'était pas une matrice PCT, les dimensions seraient (20 x 66 560).

Les coefficients **b** (repliés) pour les modèles obtenus avec 3 variables latentes sont présentés dans la Figure 11, pour la méthode PCT-OP-PLS et pour la méthode OP-PLS standard.

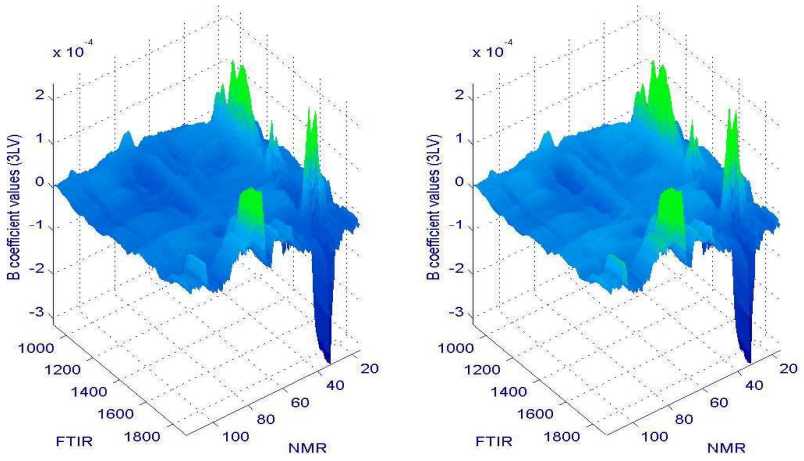


Figure 11 : Surfaces de coefficients **b** pour l'OP-PLS (gauche) et la PCT-OP-PLS (droite) après des modèles de régression avec 3 LVs. La corrélation entre les vecteurs obtenus par dépliage de ces matrices est 1.00.

Pour comprendre si les résultats obtenus avec les deux méthodes sont exactement les mêmes, on calcule la valeur absolue de la différence entre les deux matrices de coefficients de régression **b** et on présente les résultats dans la Figure 12. Les différences sont dues aux arrondis pendant les calculs et elles sont toutes inférieures à 10^{-16}, et sont donc négligeables.

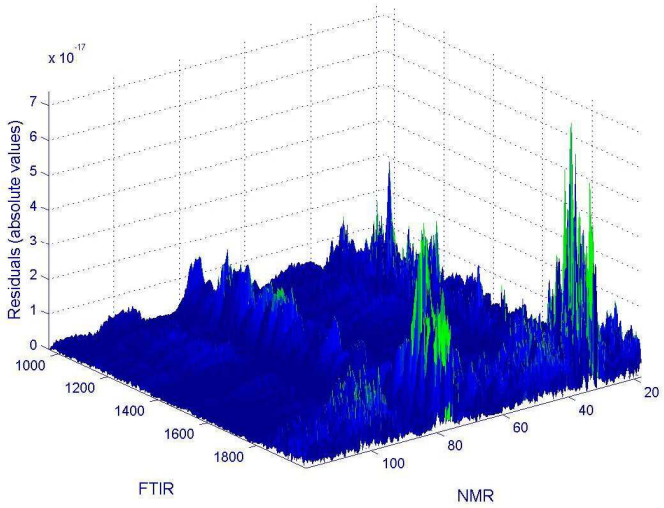


Figure 12 : Valeurs absolues pour la différence entre les valeurs des coefficients b des modèles OP-PLS standard et PCT-OP-PLS avec 3 variables latentes. Toutes les valeurs sont inférieures à 10^{-16}.

On a comparé les autres vecteurs obtenus par les deux méthodes et les résultats montrent que les différences sont complètement négligeables, et simplement dues aux erreurs d'arrondi.

Les deux techniques ont été comparées par rapport au temps de calcul et à la quantité de mémoire utilisée pour effectuer la validation croisée pour des modèles entre 1 et 10 variables latentes et pour ensuite construire un modèle avec 3 variables latentes. La procédure standard

a pris 206 secondes pour les calculs, tandis que la procédure OP l'a fait en 5 secondes. Ces résultats sont présentés dans la Figure 13.

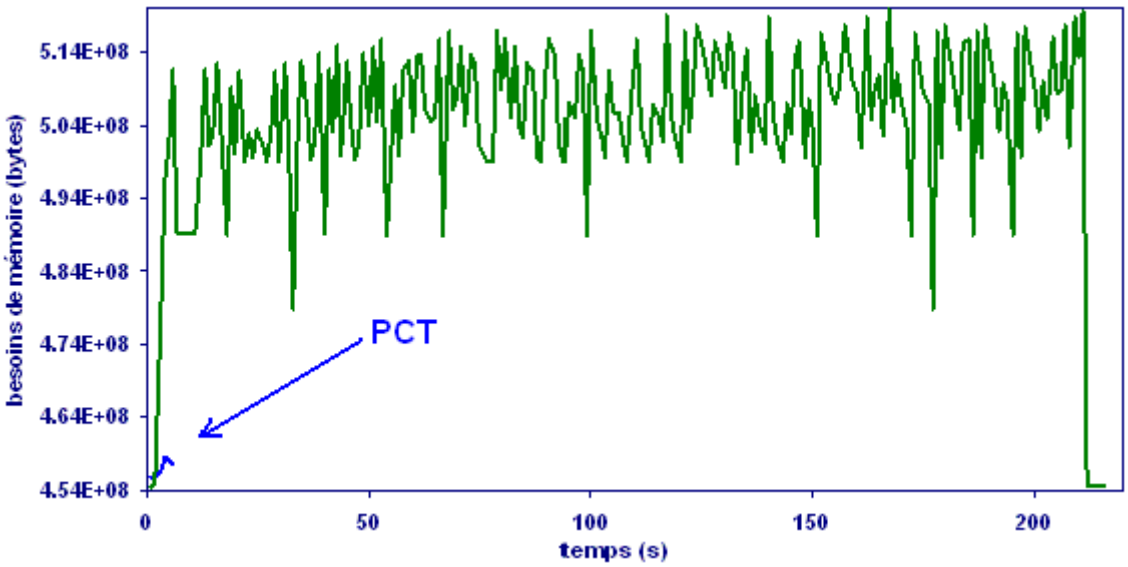


Figure 13 : Utilisation de la mémoire de l'ordinateur au cours du temps pendant la validation croisée («leave-1-out») jusqu'à 10 variables latentes par la méthode OP-PLS standard (ligne verre) et la méthode PCT-OP-PLS (ligne bleue, en bas à gauche). Chaque chiffre de la matrice correspond à 8 bytes.

1.6.4.2 Seg-PCT-OP-PLS

Comme pour la PCT-OP-PLS, les avantages pratiques et la façon de procéder pour la méthode Seg-PCT-OP-PLS sera illustrée avec un exemple[65].

Dans ce cas, le but est de produire une modèle d'étalonnage pour la discrimination de vins de cépages différents en appliquant la PLS-DA à une matrice OP obtenue en combinant des spectres NMR avec eux mêmes. Cela implique d'effectuer un validation croisée pour déterminer le meilleur modèle et ensuite prédire les échantillons de validation externe.

Matrices d'origine:

Etalonnage : **X1** et **X2** contiennent les mêmes 54 spectres RMN-2D de vins de 3 cépages différents.

Validation externe : 27 autres spectres acquis dans les mêmes conditions.

Y : 3 variables binaires (0 ou 1) d'appartenance aux 3 groupes de cépage.

Les spectres sont centrés et réduits en ligne (SNV).

Chaque spectre RMN-2D est une matrice de dimensions 874 x 2048, qui peut être dépliée pour donner un vecteur avec 1 x 1 789 952. La taille de la matrice initiale est donc de 54 x 1

789 952, ce qui correspond à plus de 700 Megaoctets. Cette matrice n'est pas facile à manipuler.

Si on pouvait calculer la matrice OP standard à partir de cette matrice originale, elle aurait comme dimensions 54 x (1 789 952 x 1 789 952), ce qui correspond à 161 000 Gigaoctets.

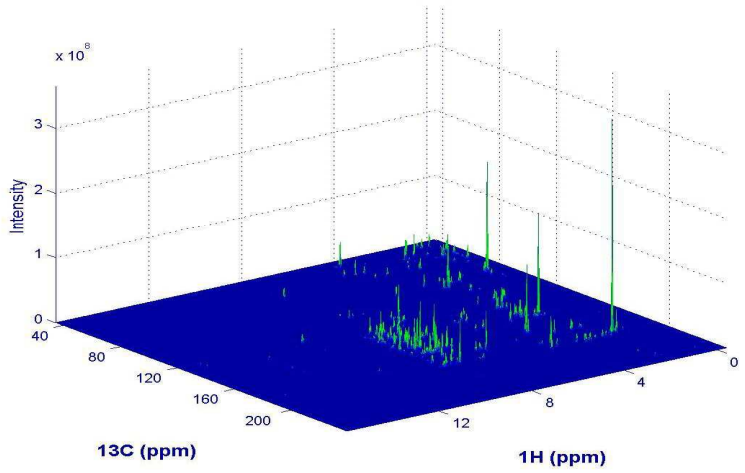


Figure 14 : Exemple d'un spectre 2D-NMR de vin (874 x 2048). Les pics les plus grands peuvent ne pas être les plus discriminants.

Procédure pour l'exemple Seg-PCT-OP-PLS

Le schéma de la figure suivante montre la procédure utilisée.

(1) Chaque spectre est déplié en un vecteur. La normalisation choisie, SNV, peut être appliqué à un vecteur à la fois, la matrice globale étant trop grande pour être normalisée d'un seul coup.

(2) Avec un ordinateur classique, il est impossible de construire une matrice OP de dimension (1 x 3 203 928 162 304). Il faut donc utiliser Seg-PCT

(3) La matrice d'origine est segmentée en 8 sous-matrices de dimensions 54x223 744 (92 Mb). On pouvait choisir d'autres nombres de fragments, avec différents nombres de colonnes chacun, mais sans beaucoup de conséquences au niveau des temps de calculs.

(4) PCA est appliquée à chaque sous-matrice et les coordonnées factorielles sont concaténées, pour donner une matrice de dimensions 54 x 432 (182 Kb)

(5) On calcule le produit externe de cette matrice avec elle-même et on obtient une matrice dépliée OP de dimensions 54 x 186 624 (77 Mb)

(6) Nous avons appliqué la PCT de nouveau pour réduire encore la taille de cette matrice et ainsi accélérer les calculs. On obtient une matrice de dimension 54 x 54 (23 Kb)

(7) La validation croisée est effectuée d'une façon extrêmement rapide entre cette petite matrice et une matrice de réponses binaires **Y**, et le meilleur modèle PLS-DA est retenu.

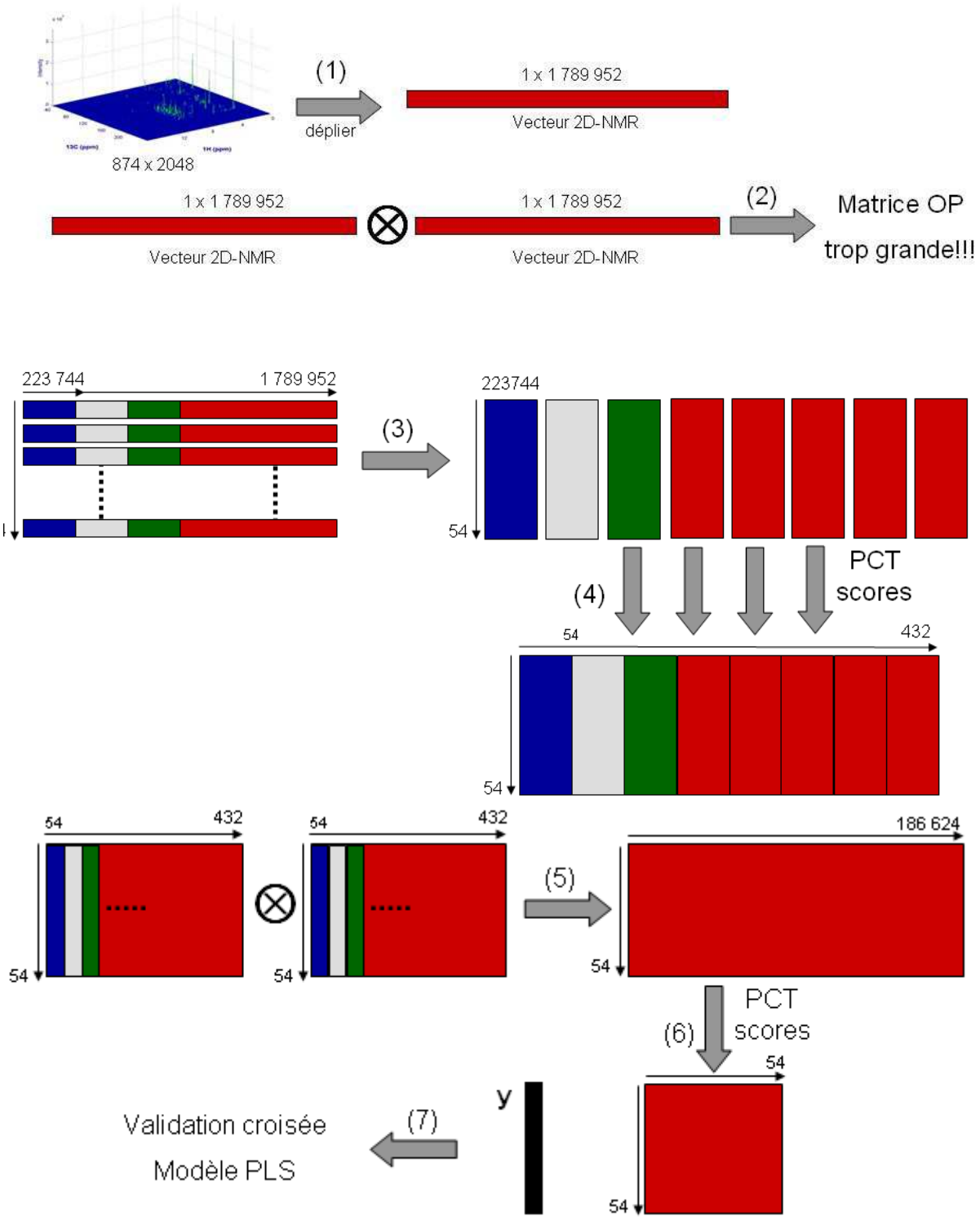


Figure 15: schéma de la Seg-PCT-OP-PLSDA pour les données RMN-2D de vin.

Validation externe et prédiction

Les données pour la validation externe et les nouveaux échantillons à prédire dans ce modèle doivent être traités de la même façon. En considérant un spectre à prédire avec le modèle trouvé, les différentes phases sont:

(1) Le spectre est déplié mais (2) la matrice OP résultant sera trop grande, donc il est (3) découpé en segments.

(4) Il est ensuite projeté sur les espaces PCT de chacun des segments à l'aide des contributions factorielles correspondantes. On obtient les coordonnées factorielles pour chaque segment et on les concatène.

(5) On calcule la matrice OP.

(6) On multiplie cette matrice OP dépliée par les contributions factorielles de la PCT-OP pour la projeter sur cette espace PCT. On obtient un vecteur que l'on multiplie par les coefficients **b** du modèle choisi pour donner la valeur prédite.

On aurait pu utiliser d'autres tailles pour les sous-matrices, le choix dépend toujours des capacités de l'ordinateur utilisé car il faut faire attention à ne pas avoir des matrices sur lesquelles on n'arrive pas à réaliser la PCT.

Il est toujours possible de reconstituer les coefficients **b** ou les différentes contributions factorielles dans l'espace original, simplement en pré et post-multipliant par les contributions factorielles de la PCT correspondante, suivant l'équation **(13)**. En même temps, si ces vecteurs sont trop grands pour pouvoir être manipulés en entier, on peut simplement reconstituer certaines parties séparément, comme précédemment.

1.7 ANOVA-PCA

Objectif : ***évaluation de la signifiance d'un facteur expérimentale***

1.7.1 Introduction

La méthode ANOVA-PCA (ou A-PCA) a été présentée par Harrington et al[7],[66] pour la détection de bio-marqueurs dans des grandes matrices composées de spectres de résonance magnétique nucléaire (NMR). C'est une méthode supervisée, où les objets sont représentatifs des différents niveaux des facteurs définis suivant un plan d'expériences équilibré. Dans une première phase, la variabilité est décomposée en contributions dues aux effets des facteurs et des interactions (ce qui explique l'utilisation du terme "ANOVA") et dans une deuxième phase, ces contributions sont comparées à la variabilité résiduelle (en utilisant la PCA).
D'autres méthodes suivant la même philosophie «ANOVA» ont été récemment développées pour l'application avec PCA[47,67,68]. Ces méthodes ne comparent pas les facteurs avec la variabilité résiduelle au moyen des distances entre les centroides des groupes et des dispersions à l'intérieur des groupes, mais cherchent plutôt à trouver des facteurs signifiants en utilisant les tests de permutations.

1.7.2 Procédure A-PCA

Dans le cas où il ya deux facteurs, le modèle mathématique de l'ANOVA est, pour un des vecteurs de la matrice de données:

$$\mathbf{x}_i = \bar{\mathbf{x}} + \bar{\boldsymbol{\alpha}}_j + \bar{\boldsymbol{\beta}}_k + \overline{\boldsymbol{\alpha\beta}}_{jk} + \boldsymbol{\varepsilon}_i \qquad \textbf{(15)}$$

où $\mathbf{x}_i$ est un spectre, $\bar{\mathbf{x}}$ est le spectre moyen (moyenne de chaque colonne), $\bar{\boldsymbol{\alpha}}_j$ et $\bar{\boldsymbol{\beta}}_k$ les effets pour deux facteurs, $\overline{\boldsymbol{\alpha\beta}}_{jk}$ l'interaction entres ces deux facteurs et $\boldsymbol{\varepsilon}_i$ les résidus ou variabilité pour ce spectre qui n'est pas due aux facteurs.

La procédure pour cette méthode est expliquée dans la figure suivante.
On commence par définir la matrice de données initiales **X** centrée par variable, pour enlever la moyenne globale.
(1) Dans la phase «ANOVA», on calcule les moyennes pour chaque niveau du facteur 1 et on crée une matrice où chaque objet contient la moyenne correspondant à son niveau. On soustrait cette matrice (Facteur 1) à la matrice précédente (dans ce cas, **X**) et on obtient les

Résidus 1. On procède de la même façon pour tous les autres facteurs et ensuite pour les interactions, jusqu'a obtenir la matrice représentative de la variance résiduelle **ε**.

(2) Cette matrice de résidus est ajoutée à chacune des matrices des facteurs et des interactions pour donner une matrice «Facteur + résidus».

(3) On applique la PCA à chaque matrice «Facteur + résidus». Après l'inspection des graphiques des projections des coordonnées factorielles 1 et 2, il est possible d'arriver à une conclusion sur la signifiance d'un facteur par rapport au bruit (résidus). Les contributions factorielles correspondantes sont utiles pour déterminer quelles variables sont responsables du comportement des objets pour le facteur en question.

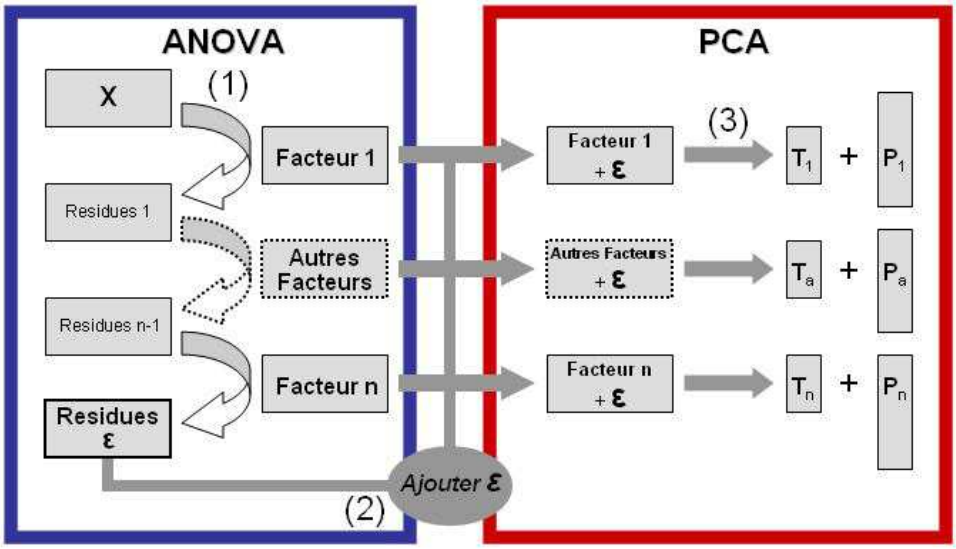


Figure 16 : schéma de l'A-PCA standard. Dans une première phase (bleue, phase ANOVA) on calcule les matrices avec les moyennes des niveaux pour chaque facteur et on obtient une matrice de résidus. Après avoir additionné la matrice des résidus à chacune des matrices des facteurs, on calcule dans une deuxième phase (rouge, phase PCA) une PCA individuelle pour chacune des matrices résultantes.

1.7.3 Résultats de l'A-PCA

Après le calcul de la PCA, en regardant les coordonnées factorielles des individus des différents groupes de niveaux, les trois situations qui peuvent se présenter et sont représentées dans la Figure 17. Dans le cas d'un facteur à 3 niveaux (trois années de fabrication différentes, par exemple), on peut observer les situations suivantes :

(1) Il n'y a pas de séparation des niveaux, donc le facteur n'est pas significatif par rapport à l'erreur résiduelle.

(2) La séparation est le long de l'axe PC1, donc la variance du facteur considéré est supérieure à la variance résiduelle. Le facteur est significatif. Basé sur les coordonnées factorielles, on

peut calculer des ellipses à un certain degré de confiance, 95% par exemple, autour de chaque groupe.

(3) La séparation est le long de l'axe PC2, donc la variance du facteur est inférieure à la variance résiduelle. Malgré cela, il y a de toute évidence de l'information dans la matrice de données, mais la variabilité associée à l'information (due au facteur étudié) est moins grande que la variabilité associée à l'erreur résiduelle. Ceci peut être le cas quand la variabilité résiduelle être structurée (variations de lignes de base, contaminants systématiques, …).

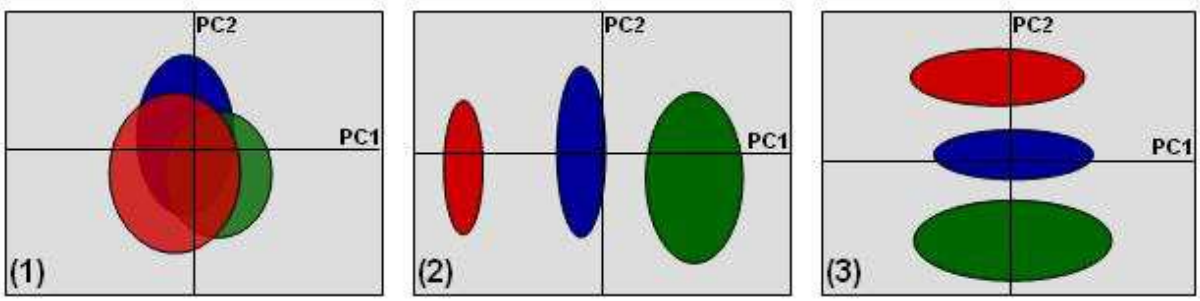


Figure 17: les trois types de résultats possibles pour les coordonnées factorielles de l'APCA pour un facteur à 3 niveaux. (1) le facteur n'est pas significatif par rapport à la variabilité résiduelle, (2) le facteur est significatif, (3) le facteur n'est pas significatif par rapport à la variabilité résiduelle, même s'il y a de l'information. La variance résiduelle est structurée.

Les avantages de cette méthode de détection de facteurs significatifs résident dans la nature même des calculs qui sont faciles à implémenter et pas trop exigeants au niveau de la mémoire.

A-PCA est plutôt une méthode simple et élégante qui fournit des résultats faciles à comprendre visuellement. En plus, il est possible d'utiliser la PCT et la Seg-PCT avec cette méthode, pour la rendre encore plus rapide.

1.7.4 Extensions et modifications de l'A-PCA

1.7.4.1 Analyse discriminante avec l'A-PCA

Une extension de la méthode A-PCA a été proposée au cours de cette thèse, de façon de l'utiliser comme méthode d'analyse discriminante et de classification. Dans ce cas, l'A-PCA est réalisée, fournissant les contributions factorielles pour les deux premiers PCs pour le facteur dont on souhaite tester la prédiction. Ensuite on projette le nouvel échantillon sur l'espace des coordonnées factorielles des deux premières PCs de ce facteur, en utilisant les vecteurs de contributions factorielles.

Procédure[69] pour la classification avec A-PCA

(1) On prétraite chaque échantillon à prédire ($\mathbf{x_i}$) de la même façon que les échantillons d'étalonnage, et donc on soustrait de l'échantillon la moyenne globale provenant de l'étalonnage, pour donner $\mathbf{x_{ic}}$.

(2) Quand on veut tester la classification de l'échantillon pour un facteur donné, on doit créer une matrice pour chacun des autres facteurs et interactions de façon à pouvoir obtenir toutes les combinaisons possibles pour les différents niveaux de ces dernières.

(3) On doit aussi créer une autre matrice ($\mathbf{X_{ic}}$), de mêmes dimensions, où toutes les lignes sont égales à l'échantillon à prédire.

(4) Finalement on doit soustraire de $\mathbf{X_{ic}}$ toutes les combinaisons possibles calculées en (2).

(5) Pour cet échantillon, plusieurs vecteurs sont obtenus, suivant les dimensions de $\mathbf{X_{ic}}$, mais seulement certains d'entre eux sont similaires au vecteur caractéristique du niveau auquel l'échantillon appartient, pour le facteur considéré.

(6) On projette la matrice résultante à l'aide des contributions factorielles de la PC1 et de la PC2 obtenues lors de la phase d'étalonnage. Les projections des vecteurs calculés dans (5) devraient tomber proche des groupes d'étalonnage. Par contre, d'autres vecteurs correspondant aux mauvais niveaux pour les différents facteurs, doivent tomber loin des groupes d'étalonnage. Ceci permet de visualiser la qualité de la prédiction, facilitant l'interprétation des résultats.

(7) On calcule des distances de Mahalanobis pondérées entre quelques projections et les centroides des différents groupes d'étalonnage. L'échantillon appartient au groupe le moins distant.

La distance de Mahalanobis entre deux vecteurs $\mathbf{x_i}$ et $\mathbf{x_j}$ est définie par[1]

$$\mathbf{D}_{ij}(\mathbf{x}) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{C}^{-1} (\mathbf{x}_i - \mathbf{x}_j)} \qquad \textbf{(16)}$$

où **C** comme la matrice de variance-covariance de la matrice **X**.

Un exemple de toutes les projections possibles pour un échantillon et les distances entre ces projections et les centroides est présenté dans la Figure 18. Dans ce cas particulier, 100 % des échantillons ont été correctement prédits.

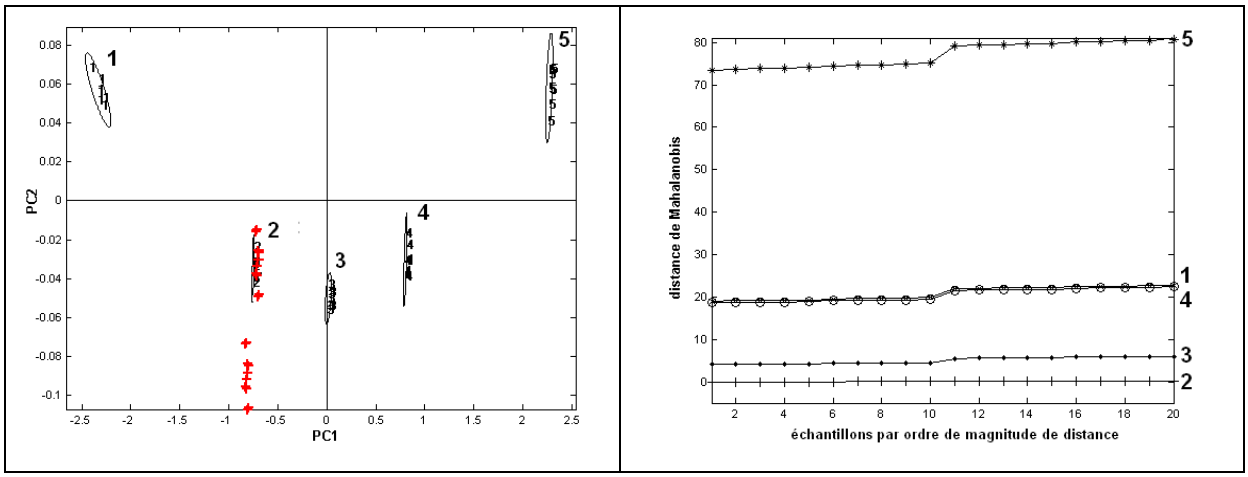


Figure 18: (gauche) En rouge, la projection de toutes les combinaisons pour un seul échantillon pour un facteur à cinq niveaux. (droite) Distances de chaque projection à chacun des cinq centroides (voire numéro de niveaux à droite). Les distances sont ordonnées par ordre de magnitude. Les distances des projections aux niveaux 1 et 4 sont similaires (noter que ce sont des distances de Mahalanobis). Les distances avec des magnitudes les plus petites (proches de zéro) ont été calculées par rapport au groupe 2, dont l'échantillon appartienne à ce niveau.

Cette adaptation de A-PCA pour en faire une méthode discriminante a fait l'objet d'une publication[69].

1.7.4.2 Réduction de la variabilité résiduelle

Objectif : ***augmenter les possibilités de détection de facteurs***

L'A-PCA présente quelques limitations, notamment en ce qui concerne la détection des facteurs intéressants car on suppose qu'ils se trouvent dans les deux premières Composantes Principales avec une relativement grande quantité de variance. Par contre, comme on le voit dans la Figure 17(3), la Composante Principale sur laquelle les niveaux d'un facteur sont séparés peut ne pas être celle avec la plus grande quantité de variance (PC1). On a donc de l'information intéressante, mais sans pouvoir dire pour autant que le facteur est significatif par rapport à l'erreur résiduelle. Afin de découvrir des facteurs significatifs mais avec des quantités de variance réduites, il peut être intéressant de réduire sélectivement et progressivement la variance résiduelle. Il serait ainsi possible aussi de mieux comprendre la nature de cette variance résiduelle. Pour ce faire nous avons proposé une nouvelle approche de l'A-PCA, inspirée par une méthode visant à enlever une partie de la variance résiduelle due aux répétitions. «Error removal by orthogonal subtraction (EROS)» a été récemment proposée[70] dans la littérature, avec le but de diminuer la variabilité intra-échantillon des mesures de spectroscopie en reflectance diffuse dans la région Visible-NIR. Cette méthode

simple est basée sur la détermination des plus grandes sources de variabilité d'une matrice **W** contenant toute la variabilité intra-échantillon. Après avoir appliqué une PCA à la matrice **W**, on détermine les directions de variabilité résiduelle les plus importantes dans la matrice. En sélectionnant quelques vecteurs propres **P** on peut soustraire sélectivement de la matrice des données originaux des parties de la variabilité résiduelle, suivant la relation :

$$\widetilde{\mathbf{X}} = \mathbf{X}(\mathbf{I} - \mathbf{P}.\mathbf{P}^T) \qquad \textbf{(17)}$$

La matrice **W** est obtenue à partir du calcul des sous-matrices ($\mathbf{Z_i}$), calculées par soustraction de la moyenne de la matrice des répétitions pour un même échantillon *i*. Une fois ces sous-matrices calculées pour tous les *I* échantillons, on calcule le produit $\mathbf{Z_i^T}.\ \mathbf{Z_i}$ pour chacune et on additionne toutes ces matrices en pondérant par les dégrées de liberté *r-I* (avec *r* le nombre totale de répétitions pour tous les échantillons), tel que :

$$\mathbf{W} = \sum_{i=1}^{I} \frac{\mathbf{Z}_i \mathbf{Z}_i^T}{(r - I)} \qquad \textbf{(18)}$$

En fait, la matrice de variances résiduelles de l'A-PCA est créée par la concaténation verticale des matrices $\mathbf{Z_i}$ de la méthode EROS. Il est donc clair qu'il a une relation entre les deux méthodes. Nous avons donc pensé à utiliser la méthode EROS pour réduire la variabilité résiduelle, de façon à augmenter la sensibilité de la méthode A-PCA. Après quelques essais, nous avons décidé de changer l'approche pour une procédure similaire mais plus simple. Dans le même esprit d'EROS, on peut diminuer progressivement la variabilité résiduelle de la matrice résiduelle en la reconstituant avec un nombre limité de PCs. Après ce pas, on peut comparer de nouveau les facteurs avec la variabilité résiduelle reconstituée. Si l'on n'arrive pas encore à détecter le facteur d'intérêt selon PC1 de l'A-PCA (Figure 17), on élimine encore un autre PC de la matrice résiduelle. On procède ainsi jusqu'au point où les niveaux du facteur étudié soient disposés le long de l'axe PC1. À ce moment-là, on applique un test des permutations pour comparer les distances «réelles» avec une distribution des distances pour des modèles avec permutations aléatoires (voir Figure 4), ce qui nous permet de savoir si le facteur est réellement significatif ou non.

Cette modification de la méthode, qui permet de mieux comprendre les sources de variabilité aléatoire présentes dans les données, a été soumise pour publication.[71].

1.8 A-ComDim

Objectif: ***identification de facteurs signifiants – perspective multitableaux***

Common Components and Specific Weights Analysis (CCSWA) est une méthode d'analyse multivariée de tableaux multiples, développée pour l'analyse de données en sensométrie[72,73,74]. Pendant ces travaux, nous avons utilisé une version Matlab intitulée "ComDim" («common dimensions») fourni dans la boîte à outils SAISIR[75] .

Dans le domaine de l'analyse sensorielle, les panels de dégustation sensorielle sont beaucoup utilisés pour qualifier des produits. La plupart du temps, les membres du panel sont entraînés pour décrire leurs expériences sensorielles par rapport aux sensations de saveur, toucher, odeur. Les panels sensoriels présentent normalement une variabilité considérable, dépendant des capacités naturelles, l'entraînement des membres, de la variabilité propre aux produits. Dans certains cas, à chaque membre du panel sensoriel est attribué le même groupe d'échantillons qu'il doit évaluer par rapport à un certain nombre de variables (éventuellement différentes pour tout le monde). Chaque membre du panel constitue donc une matrice de données, où chaque ligne représente un échantillon et chaque colonne une variable utilisée par ce juge. Vu que ce sont les mêmes échantillons avec les mêmes caractéristiques, bien que décrits avec des descripteurs différents par chaque juge, ces descripteurs peuvent être considérés comme simplement des vecteurs différents définissant le même espace multidimensionnel. Dans ce cas, il devrait être possible de trouver des directions communes dans cet espace commun. Il serait alors possible de déterminer quelles répartitions des échantillons sont communes à l'ensemble des juges, de hiérarchiser la "communalité" de ces répartitions et d'évaluer l'importance de chaque juge dans la définition d'une direction. L'existence d'un ou de plusieurs membres du panel avec une sensibilité différente ne devrait pas avoir beaucoup d'importance dans la définition de ces Composant Communes.

La méthode CCSWA-ComDim permet d'explorer plusieurs tableaux en même temps en cherchant à expliquer les dimensions communes de l'espace définies par les tableaux. Puis que le nombre et la nature des variables des différents tableaux peuvent varier, la méthode peut donc être appliquée à l'analyse simultanée de plusieurs domaines en spectroscopie. On peut donc chercher l'information commune aux mêmes échantillons analysés par différentes techniques ou méthodes ou dans des conditions différentes. On pourrait même l'utiliser pour la sélection de variables ou de régions spectrales, en découpant un ensemble de données

spectrales en différents tableaux en fonction des régions spectrales. La méthode peut calculer les coordonnées factorielles de chaque échantillon sur les Composantes Communes et les contributions des variables de départ à chacune de ces Composantes Communes. On peut aussi calculer l'importance («salience») de chaque tableau initial dans la définition de ces Composants.

L'algorithme itératif ComDim[75] est décrit dans le schéma de la Figure 19 pour le cas de deux matrices.

Procédure ComDim

(1) Pour chaque tableau **X** (*n* x *k*) et **Y** (*n* x *l*), on centre les variables et on divise chaque valeur par la norme de la matrice, ce qui donne les matrices $\mathbf{X_s}$ (*n* x *k*) et $\mathbf{Y_s}$ (*n* x *l*).

(2) On rentre dans un cycle pour le calcul itératif des Composantes Communes. Les matrices sont multipliées par leurs transposées, ce qui donne des matrices carrées avec les mêmes dimensions du nombre d'échantillons, donc les mêmes pour les deux domaines, **W** (*n* x *n*).

(3) Chaque matrice **W** est multipliée par un poids λ (pour la première itération λ_X et λ_Y = 1) et le résultat pour les deux matrices est additionné pour obtenir une matrice globale $\mathbf{W_G}$.

(4) On applique une PCA sur $\mathbf{W_G}$ pour calculer la dimension commune. Cette dimension commune est donnée par les coordonnées factorielles $\mathbf{U_W}$, des échantillons sur la première composante principale de $\mathbf{W_G}$, c'est-à-dire le vecteur **q** (*n* x *1*).

(5) On recalcule les poids λ_X et λ_Y en pré- et post- multipliant chaque matrice **W** par le même vecteur. Notez que si les corrélations entre **q** et les colonnes de **W** pour un tableau sont élevées, son poids λ sera également élevé. Ceci signifie que ce tableau aura beaucoup d'importance pour cette dimension commune.

(6) Pour savoir si le calcul des poids λ est fini ou non (convergence), on calcule une quantité «Dif» pour chaque itération. Ayant défini au préalable un seuil (10^{-10} par exemple), si on obtient $Dif_n^2 - Dif_{n-1}^2 >$ seuil, on continue dans ce cycle et on recalcule $\lambda_X.\mathbf{W_X}$ et $\lambda_Y.\mathbf{W_Y}$ (on rentre dans le cycle pour calculer le pas 3). Les deux matrices **I** sont des matrices identité.

(7) En revanche, le moment que $Dif_n^2 - Dif_{n-1}^2 < seuil$, le calcul de la première composante commune est fini, on garde le vecteur **q** et les poids λ_X et λ_Y pour chacune des matrices **X** et **Y** dans cette dimension commune.

(8) On recalcule les matrices $\mathbf{X_s}$ et $\mathbf{Y_s}$ a partir d'une matrice identité **I** (n x n), le vecteur **q**, les «anciennes» matrices $\mathbf{X_s}$ et $\mathbf{Y_s}$ et on recommence depuis le début pour la deuxième composante commune, jusqu'à obtenir le nombre de composantes communes souhaitées.

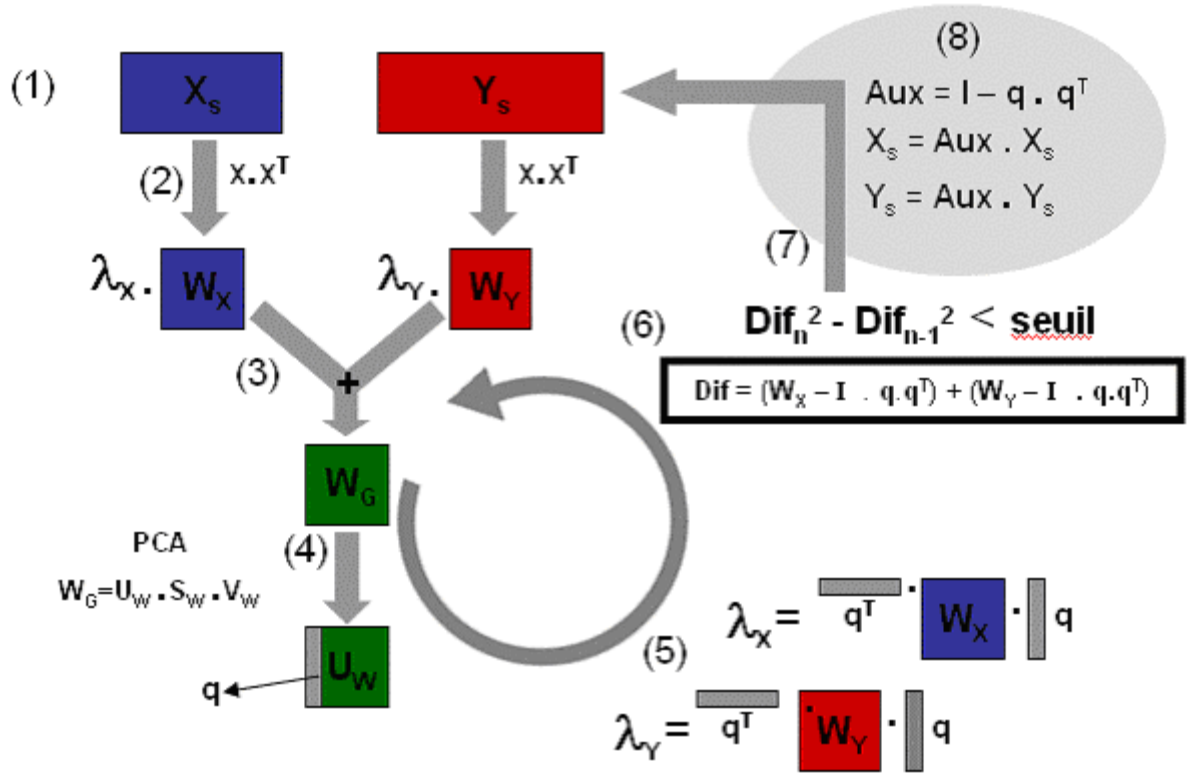


Figure 19: schéma de la CCSWA, implémenté dans l'algorithme ComDim. La composante commune provient toujours de la PC1 sur la matrice W_G courante.

A-ComDim

Une nouvelle application de la méthode ComDim a été développée à partir de la philosophie de l'A-PCA. On procède de la même façon que dans l'ANOVA-PCA pour trouver les niveaux des facteurs et les résidus par rapport à un plan d'expériences. Les différentes matrices de facteurs plus résidus, d'interaction plus résidus et le tableau de résidus seuls sont considérées comme un ensemble de «tableaux multiples». La valeur de «salience» la plus élevée pour chaque Composante Commune (CC) montre le ou les facteur(s) associé(s). Comme dans A-PCA, on peut représenter les coordonnées factorielles des CCs intéressantes contre la CC associée à la variance résiduelle. Puisque les résidus sont inclus dans chaque tableau, en plus d'être dans un tableau propre, la première Composante Commune est souvent associée à ces résidus et l'ensemble des tableaux ont donc une "salience" élevée pour CC1.

L'algorithme d' A-ComDim est le même que pour la méthode ComDim. L'idée pour l'A-ComDim est que les tableaux X et Y de la Figure 19 sont les matrices des facteurs + résidus de la Figure 16, calculées de la même façon que dans la méthode A-PCA. Un manuscrit a été soumis présentant la théorie de cette méthode et montrant ses avantages par rapport à A-PCA.[76].

2. Spectroscopie infrarouge[77, 78, 79]

Objectif : ***analyses rapides, multi-composées, non destructives***

Le rayonnement électromagnétique est la combinaison d'un champ électrique et un champ magnétique. Les différences de propriétés des différents types de rayonnement sont dues à l'énergie et donc à la longueur d'onde ou la fréquence.

Le spectre électromagnétique a été divisé en différentes régions, par rapport aux types de transitions atomiques ou moléculaires que le rayonnement provoque lors de son interaction avec la matière. L'infrarouge se situe entre les rayonnements visibles et des micro-ondes (Tableau 1). Plusieurs phénomènes peuvent subvenir lors de l'interaction entre ce type d'énergie et la matière, dont l'absorption, la diffraction, la réfraction et les réflexions spéculaires et diffuses. L'absorption d'énergie avec une fréquence spécifique pour les liaisons moléculaires est le phénomène le plus utilisé analytiquement, par transmission, dans la spectroscopie moyen infrarouge (MIR) et par réflexion diffuse dans le proche infrarouge (NIR).

Tableau 1 : longueurs d'onde (λ), fréquences (υ) et nombres d'onde ($\bar{\upsilon}$) pour les régions du proche et moyen infrarouge.

	λ (cm)	υ (Hz)	$\bar{\upsilon}$
Proche IR	$7.8\text{x}10^{-5} - 2.5\text{x}10^{-4}$	$3.8\text{x}10^{-14} - 1.2\text{x}10^{-14}$	12800 - 4000
Moyen IR	$2.5\text{x}10^{-4} - 5\text{x}10^{-3}$	$1.2\text{x}10^{-14} - 6\text{x}10^{-12}$	4000 - 200

Les molécules organiques sont dans un état constant de vibration. Quand au moins une de ses vibrations cause une variation de moment dipolaire, elles peuvent absorber le rayonnement infrarouge. Les différentes liaisons entre molécules ont des fréquences de vibration et de rotation caractéristiques et peuvent absorber de l'énergie à cette même fréquence, avec une augmentation de l'amplitude de vibration mais sans changer la fréquence.

L'intensité des bandes dépend de la nature et de la polarité des liaisons. Par exemple, dans le MIR, la liaison C=O, fortement polaire composée de deux atomes différents, absorbe fortement, tandis que la liaison C=C a une faible absorption. Plusieurs tables existent pour l'identification de bandes d'absorption, facilitant l'identification pour des cas simples et des mixtures peu complexes. Pourtant, les bandes d'absorption peuvent être superposées, ce qui est souvent le cas quand on analyse des mixtures complexes.

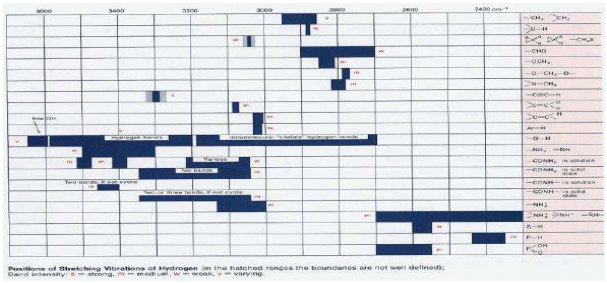
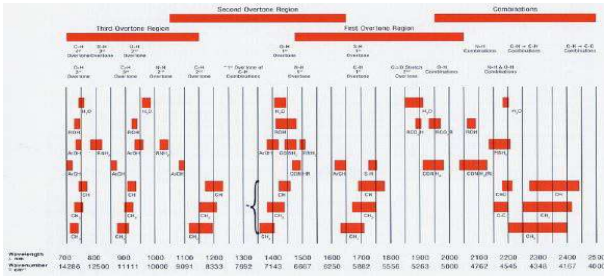



Figure 20 : exemples de tableaux pour l'identification des pics et des bandes d'absorption dans des spectres MIR (gauche) and NIR (droite)[80].

2.1.1 Spectroscopie MIR

En moyen infrarouge, on observe un grand nombre de pics assez bien résolus. Dans cette plage spectrale, la hauteur d'un pic d'absorption est proportionnelle à la concentration du groupement fonctionnel correspondant.

La plage spectrale en MIR, entre 4000-1300 cm^{-1} est appelée le zone des «fréquences de groupe» car des molécules avec le même groupe fonctionnel présentent des bandes d'absorption sensiblement aux mêmes fréquences, indépendamment de la chaine carbonée principale.

La région 1300-200 cm^{-1} contient des bandes vibrationnelles plus difficilement attribuables à des groupes fonctionnels spécifiques, du fait que les masses et les forces de liaison de chacune des espèces absorbantes sont très similaires. Pourtant, il y a dans cette région des absorptions importantes et souvent caractéristiques de ce qui est présent dans l'échantillon. C'est pour cette raison que cette région est intitulé le zone des «empreintes digitales».

2.1.2 La réflexion totale atténuée (ATR)

Il existe plusieurs façons de mesurer le spectre d'un échantillon en MIR. Les plus utilisées sont la mesure en cellules de transmission et sur un cristal d'ATR, c'est cette dernière méthode qui a été utilisée pendant les travaux de cette thèse. L'ATR a l'avantage de permettre l'acquisition rapide et presque sans préparation de spectres de qualité pour la plupart des échantillons solides, liquides et pâteux. L'épaisseur de l'échantillon ne change pas la quantité d'énergie absorbée car le phénomène est basé sur une réflexion atténuée sur la surface d'un cristal de diamant, qui est à l'origine d'une pénétration de la radiation d'une profondeur constante. Ceci peut être un avantage pour l'analyse de solutions aqueuses car l'eau absorbe

fortement dans la région moyen infrarouge et l'ATR ne laisse pas pénétrer trop profondément le rayonnement dans l'échantillon.

Pour solutions diluées, l'ATR respecte aussi la loi de Beer-Lambert :

$$A = k \cdot c$$

où A est l'absorbance et k la constante de proportionnalité par rapport à la concentration.

Le nettoyage du cristal est en général extrêmement facile et rapide. Il y a des appareils commerciaux qui donnent la possibilité d'acquérir des spectres pendant le chauffage de l'échantillon directement sur le cristal d'ATR, ce qui permet d'analyser des échantillons à différentes températures. De cette façon il est possible de étudier l'influence des réactions chimiques telle que l'oxydation ou l'isomérisation favorisées par le chauffage ou de suivre l'évolution des spectres pendant le séchage de l'échantillon directement sur le cristal. Ces deux types d'étude ont été réalisés au cours de cette thèse.

2.1.2.1 Méthode MIR-ATR chauffante: échantillons aqueux

Objectif : ***séchage rapide des échantillons aqueux pour révéler des autres composées***

La méthode ATR dans le MIR autorise des analyses simples et rapides avec une préparation réduite des échantillons. Si l'on utilise des dispositifs ATR appropriés pour l'analyse des échantillons aqueux, on n'a pas de problèmes de saturation du signal. Par contre une autre contrainte existe. L'eau absorbe fortement dans certaines régions de l'infrarouge, donc l'analyse des échantillons avec un pourcentage élevé d'eau reste difficile, voire impossible dans certains cas à cause des faibles différences entre des spectres d'échantillons dilués.

Les spectres de produits aqueux très différents ressemblent beaucoup au spectre d'eau, comme on peut le constater en regardant la Figure 21 où l'on voit, par exemple, que le grand pic d'absorbance vers 3300 cm^{-1} peut masquer les absorbance intéressantes des groupes C-H. De la même façon, d'autres pics dus à des solutés sont presque complètement cachés par la bande entre 1800 - 600 cm^{-1} qui recouvre une grande partie de la région des «empreintes digitales». Cette difficulté devient encore plus importante pour l'étude de mélanges à faibles concentrations.

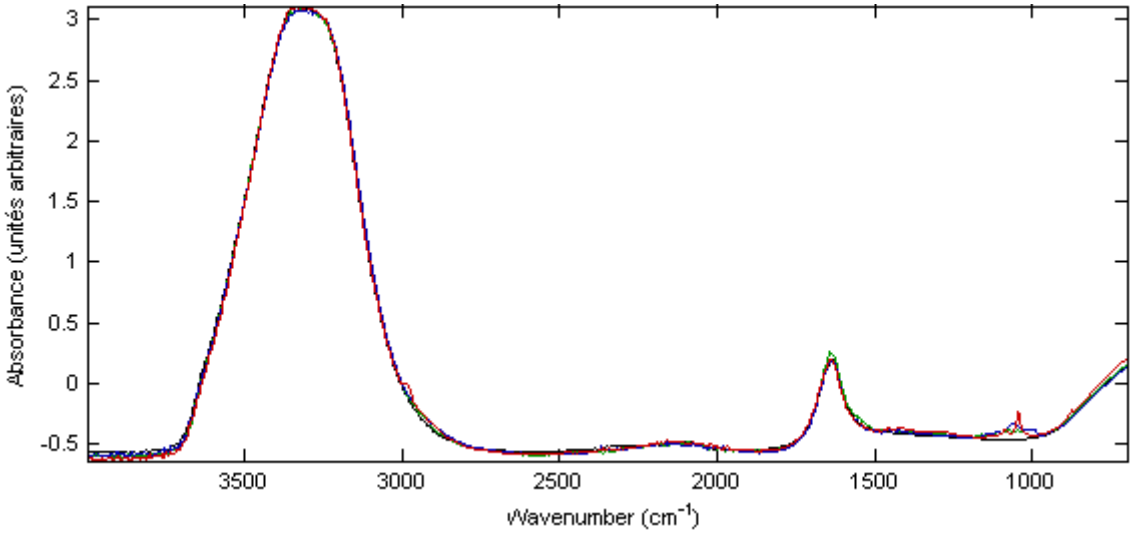


Figure 21 : spectres MIR-ATR de l'eau (noir), vin rouge (rouge), lait demi écrémé (vert) et du jus de pêche naturel (bleu). Les spectres se ressemblent tous et les plus grandes différences sont vers 1000 cm^{-1}.

Nous avons proposé une méthode pour éliminer ce problème fondée sur l'utilisation d'un dispositif d'ATR chauffante. Les spectres sont acquis après une courte période de chauffage, dont la durée et la température dépendent du produit analysé, et pendant laquelle la plupart de l'eau (et des autres substances volatiles, comme l'éthanol) s'évapore. Il ne reste sur le cristal d'ATR qu'un résidu visqueux fortement concentré, ce qui rend beaucoup plus visible les pics d'absorbance des solutés et permet une différentiation beaucoup plus facile des échantillons.
Pour déterminer le bon moment pour arrêter l'évaporation, on peut même acquérir les spectres en continu pendant le chauffage comme on peut le voir dans la Figure 22. L'évolution des pics observé dans les spectres lors du chauffage de vin (et c'est la même chose pour les autres produits présentés dans la Figure 21) ressemble quelque peu à une courbe de neutralisation :
(1) en fonction de la température de chauffage et de l'échantillon, les spectres n'évoluent pas beaucoup au début.
(2) dans une deuxième phase, on assiste au passage rapide d'un échantillon aqueux dilué vers un résidu visqueux, suite à une réduction considérable de la quantité l'eau, ce qui entraine une diminution de la contribution de l'eau aux spectres.
(3) finalement, dans une dernière phase, on observe un résidu presque solide sur le cristal. Les spectres n'évoluent presque plus car il ne reste plus que de l'eau liée et une mixture des solutés de départ. Dans le cas des boissons alcoolisées, presque toute l'eau et tout l'éthanol, ainsi qu'une partie des autres composés volatiles, sont ainsi éliminés.

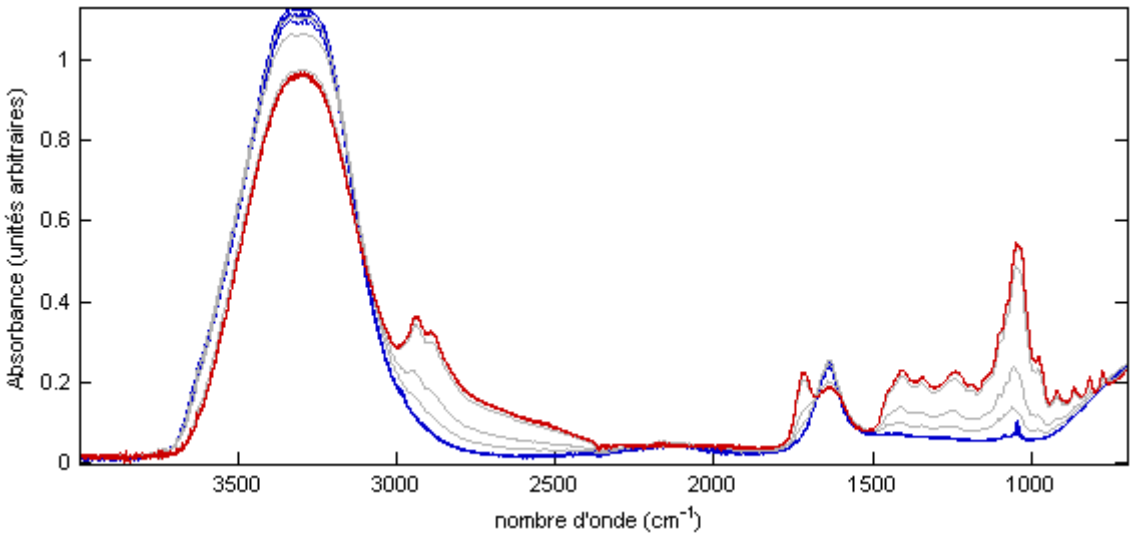


Figure 22 : Évolution de spectres de vin blanc acquis chaque minute au cours du chauffage/évaporation à 60ºC sur le cristal d'ATR. t_0-t_5 (bleu), t_6-t_9 (gris) et t_{10}-t_{15} (rouge).

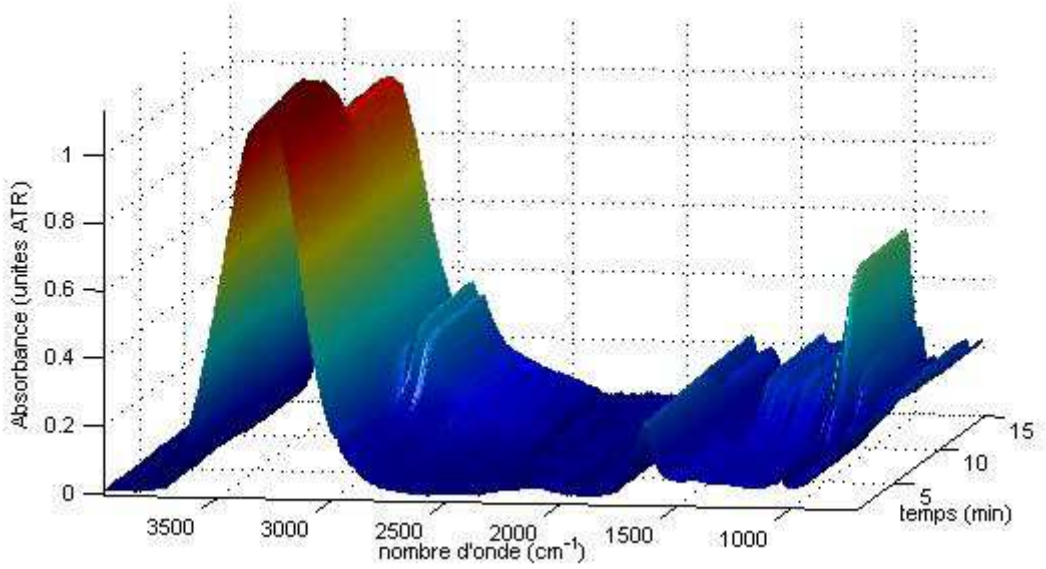


Figure 23 : Vue en perspective des mêmes spectres de vin blanc que dans la figure antérieure.

Dans la Figure 24 on peut voir l'effet de l'évaporation sur les spectres des mêmes produits que dans la Figure 21. Les différences entre les spectres sont maintenant évidentes, surtout dans la région des «empreintes digitales».

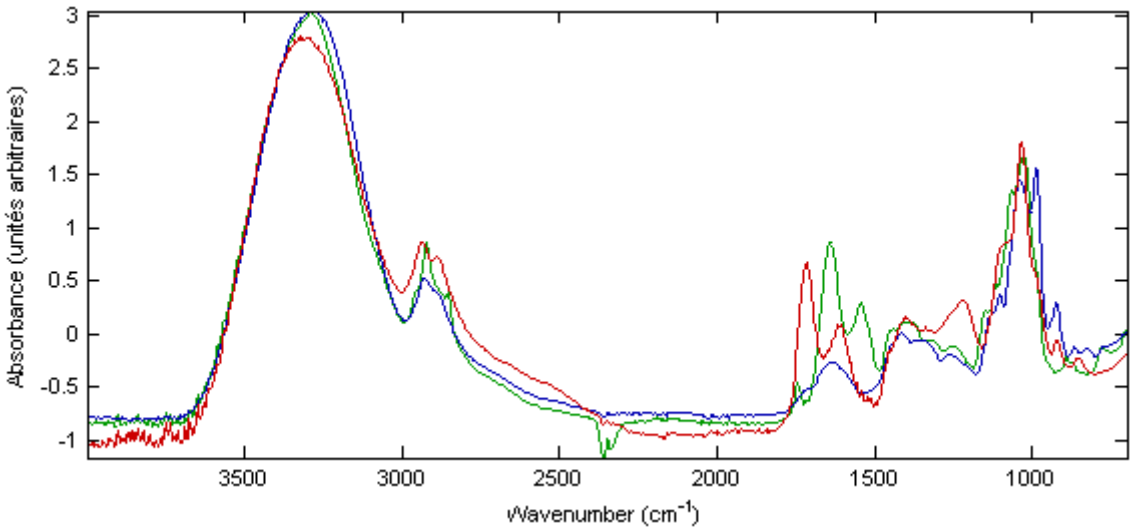


Figure 24: Spectres des mêmes produits de la Figure 21, après évaporation sur le cristal d'ATR. Vin rouge (rouge), lait demi écrémé (vert) et jus de pêche naturel (bleu). Les différences entre les spectres des différents produits sont évidentes.

Pour que la méthode fonctionne, il faut que les échantillons soient chimiquement stables à la température du chauffage et qu'ils forment un mélange visqueux après évaporation, ce qui est le cas pour beaucoup de produits alimentaires aqueux, comme le vin, la bière, les jus de fruits, le lait, entre autres. Deux articles basés sur l'utilisation de ce système ont été soumis pour publication au cours de cette thèse [71, 76].

2.1.2.2 Méthode MIR-ATR chauffante: altérations des huiles

Objectif: ***chauffage – acquisition de spectres en simultanée***

La composition des huiles alimentaires d'origine végétale est dominée par les triglycérides (95 – 98 %) avec différentes compositions en acides gras. Les composantes minoritaires (2 – 5 %) appartiennent à des groupes chimiques diverses, tels que alcools, esters, hydrocarbures, composées phénoliques, tocophérols, tocotriénols, pigments, phospholipides, entre autres[81].
Les acides gras des triglycérides les plus fréquents dans les huiles sont :

- poly-insaturées: linoléique (C18 :2) et linolénique (C18 :3)
- mono-insaturées: oléique (C18 :1)
- saturées: stéarique (C18 :0) et palmitique (C16 :0)

La variabilité de composition des huiles est surtout due à l'origine végétale. D'autres facteurs contribuent aussi à la variabilité de composition, dont la variété, le climat, le terrain et les conditions de production[82]. Cette variabilité est responsable des différentes caractéristiques physico-chimiques des huiles, notamment la résistance à l'oxydation.

L'altération des huiles alimentaires d'origine végétale a une très grande importance en ce qui concerne les aspects économiques, nutritionnels, organoleptiques, technologiques et surtout sanitaires. En effet, la consommation de certaines huiles avec un contenu élevé en acides poly-insaturées est bénéfique dû à son influence sur l'équilibre hormonal, le renforcement du système immunitaire et la prévention du cancer, du diabète, de l'obésité, de l'arthrite, entre autres. En revanche, la consommation des huiles peut être très mauvaise pour la santé quand elles ont été dégradées, soit naturellement lors du stockage dans le temps, soit suite à des procédés industriels ou culinaires qui peuvent accélérer des transformations néfastes.

Les altérations des huiles dépendent de la composition en acides gras et en antioxydants ainsi que de certains facteurs extérieurs, les plus importants étant la température et l'exposition à l'air (O_2), mais aussi le *p*H, la présence de catalyseurs et la lumière. Plusieurs mécanismes ont été proposés pour expliquer les altérations en fonction de différentes conditions chimiques et physiques de l'huile et de son environnement. Ces altérations, connues depuis longtemps, sont principalement des oxydations, des polymérisations, des isomérisations et des réactions de cyclisation[83,84,85].

L'oxydation non-enzymatique des huiles se produit à travers des mécanismes d'autoxydation ou de photo-oxydation. Même si dans certaines conditions il y a une contribution de la photo-oxydation, l'autoxydation est le mécanisme le plus important dans les processus technologiques et culinaires. Pour l'autoxydation, plusieurs mécanismes de réactions radicalaires ont été présentés pour les différents types d'huiles, impliquant les doubles liaisons des acides gras et la présence d'oxygène. Ils contiennent toujours des étapes d'initiation, de propagation et de terminaison :

- pendant l'initiation, il y a production des radicaux libres proches des doubles liaisons ;
- l'étape de propagation fait intervenir des réactions entre ces radicaux et d'autres molécules, principalement des acides gras, avec formation de hydroperoxydes et d'autres composés, résultant de rupture de chaines. De nouveaux radicaux se forment pendant cette étape.
- l'étape de terminaison résulte de réaction entre deux radicaux, ce qui fait qu'ils ne sont plus disponibles pour continuer la réaction en chaine.

L'oxydation des huiles rend possible d'autres types de altérations, telles que des isomérisations et des cyclisations, à partir des composés formés pendant l'oxydation.

Le chauffage des huiles entraine la formation de différents composés chimiques au cours du temps. On peut assister à la formation d'hydroperoxides, d'alcools, de cétones, d'aldéhydes, d'acides, d'esters, de lactones et d'hydrocarbures. Ce processus peut être observé grâce à différents types de spectroscopie, dont la spectroscopie MIR[86]. La Figure 25 montre

l'évolution d'une huile commerciale de tournesol au cours du chauffage pendant une heure à 150°C. On peut voir la disparition des bandes C=C-H -*cis* au cours du temps et en même temps l'apparition des pics associés aux hydroperoxides, alcools, et des C=C-H -*trans*.

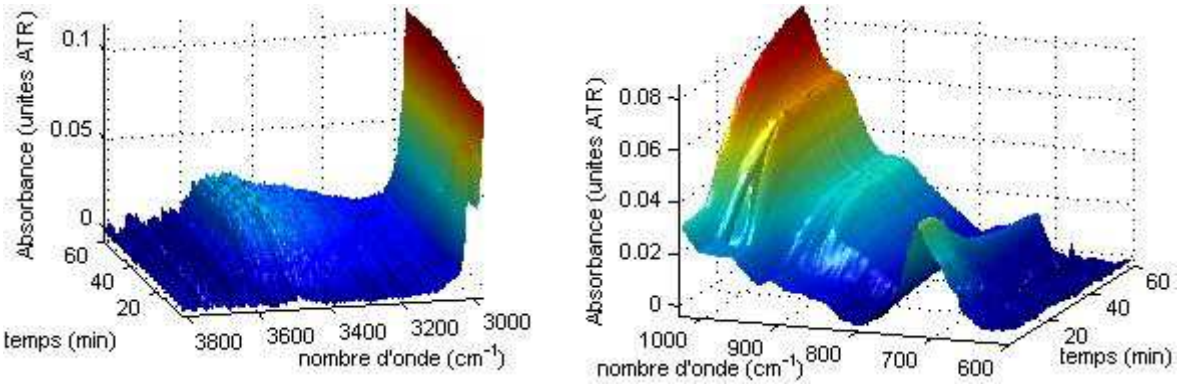


Figure 25 : Évolution des spectres de l'huile de tournesol au cours du chauffage (60 minutes) sur le cristal d'ATR à 150ºC. Détail de la région des peroxydes et produits de l'oxydation secondaire vers 3500cm^{-1} et des liaisons -*cis* vers 3008 cm^{-1} (gauche). Détail des régions -*trans* vers 980 cm^{-1} et –*cis* vers 700 cm^{-1} (droite).

Dans la plupart des études par spectroscopie infrarouge sur la dégradation des huiles, une méthode indirecte d'acquisition des spectres est utilisée. Les huiles sont chauffées dans un récipient sous certaines conditions, un aliquot est retiré et son spectre est acquis. Ces conditions de chauffage donnent normalement des vitesses d'oxydation faibles, et le degré d'oxydation est encore aussi limité, même après quelques jours.

Nous avons étudié une méthode rapide de chauffage-analyse spectrale en direct pour observer l'oxydation accélérée des huiles alimentaires. Dans cette méthode une petite quantité d'huile (75 µL) est déposée sur le cristal d'ATR chauffé à la température souhaitée pour promouvoir l'oxydation et les spectres sont acquis à des intervalles de temps bien définis (1 minute) pendant une certaine période (par exemple, 60 minutes). L'intérêt de cette méthode est clairement démontré dans un article soumis à publication[87].

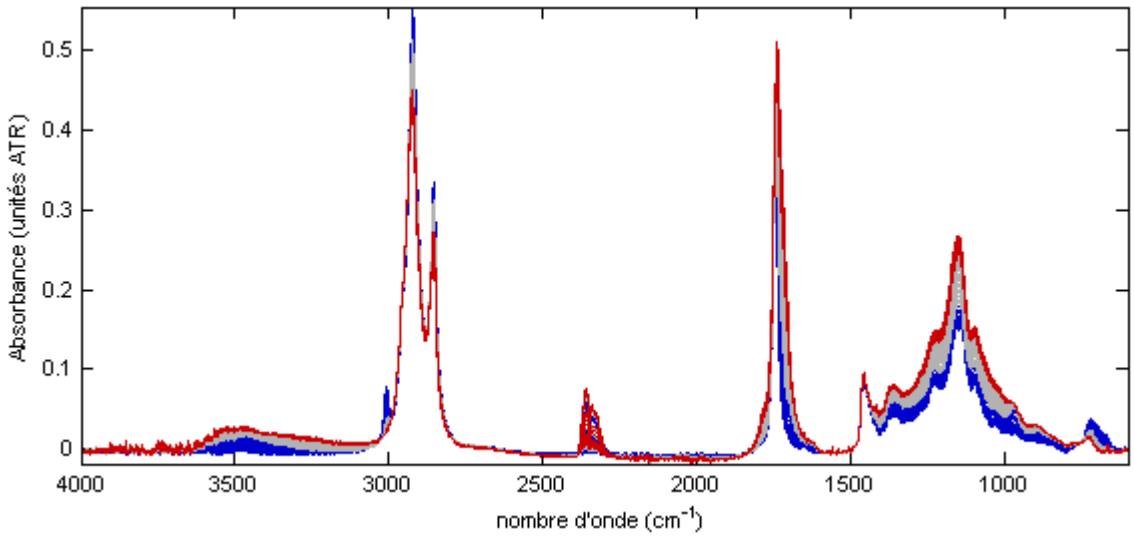


Figure 26 : Spectres d'huile de tournesol acquis toutes les minutes pendant 60 minutes à 150ºC. t_0-t_{20} (bleu), t_{21}-t_{40} (gris) et t_{41}-t_{60} (rouge).

Les avantages de cette méthode sont l'absence de manipulation au cours du chauffage (pas de problème d'échantillonnage) et une plus grande vitesse d'oxydation de l'huile (due à la surface exposée à l'air élevée par rapport au volume d'huile). Deux expériences ont été réalisées :

- 3 huiles – tournesol, olive et colza – ont été étudiés à la température de 150º.
- l'huile de tournesol avec et sans tocophérol a été étudié à 3 températures (130, 150, 170ºC).

Les résultats ont montré que cette méthode rapide donne des résultats intéressants. Nous avons ainsi pu confirmer les résultats déjà observés dans la littérature, comme une plus grande vitesse d'oxydation pour les huiles sans tocophérol et pour des températures les plus élevées. L'huile la plus stable est l'huile de colza, ensuite l'huile d'olive et finalement l'huile de tournesol.

3. Conclusions et perspectives

Ce travail a présenté des approches chimiométriques pour l'analyse de données en chimie, particulièrement pour l'analyse spectroscopique dans le domaine alimentaire. Néanmoins, toutes les méthodes chimiométriques présentées dans cette thèse sont aussi applicables dans d'autres domaines.

Depuis son introduction, la méthodologie de «Principal Components Transform» dans ses deux variantes (PCT standard et Seg-PCT) a continué à montrer son utilité à l'intérieur de beaucoup de calculs chimiométriques. La capacité de compression sans perte de la PCT et la possibilité de faire la transformation inverse font que les méthodes qui profitent le plus de son utilisation sont, soit celles qui ont de grands besoins en mémoire à cause de la taille des matrices, soit celles qui font intervenir beaucoup d'itérations ou de répétitions.

La PCT a été appliquée avec succès dans le cadre de plusieurs méthodes d'analyse multivariées. Dans le contexte actuel où des jeux de données de plus en plus énormes sont à analyser, son utilisation doit sans doute se généraliser.

L'A-PCA a montré son intérêt et applicabilité dans l'étude des données spectroscopiques acquises sur des aliments. La méthode est élégante, les calculs sont simples, les résultats sont visuellement faciles à exploiter et leur signification peut être évaluée statistiquement à l'aide de tests de permutations.

L'extension de la méthode pour permettre la classification de nouveaux échantillons peut être utile dans certains cas. Cette méthode permet une évaluation visuelle de la qualité des prédictions en regardant l'ensemble des possibilités des projections pour le même échantillon. Elle permet aussi une évaluation de la qualité du modèle en utilisant des méthodes classiques, comme le pourcentage de bons et mauvais classements, calculé suite à des validations croisées ou test de permutations.

La modification de la méthode en éliminant progressivement des parties de la variabilité résiduelle a permis d'avoir une meilleure compréhension des sources de variabilité et a augmenté sa capacité de détection des facteurs intéressants.

L'utilisation de la méthode multi-tableaux CCSWA pour l'analyse des facteurs et des interactions provenant de l'étape "ANOVA" de l'A-PCA a été testée et a donné des résultats intéressants. La méthode permet l'identification des facteurs significatifs, en détectant les tableaux "Facteur + Résidus" qui contribuent beaucoup à la définition des Composantes Communes.

Dans une perspective d'évolution de la méthode A-PCA, nous avons déjà fait des études sur la possibilité d'utilisation de l'OSC à la place de la décomposition "ANOVA". Les résultats ne sont pas encore définitifs et la méthode est encore en développement. Mais l'OSC peut apporter des propriétés intéressantes dans ce cadre, dont la facilité d'interprétation des résultats par l'élimination des sources de variabilité orthogonales à la variation d'intérêt et la possibilité de l'appliquer à des données aussi bien quantitatives que qualitatives.

La spectroscopie infrarouge a été largement utilisée dans le cadre de cette thèse et des données de proche et moyen infrarouge ont été exploitées dans multiples situations. D'un point de vue expérimental, nous avons pu spécialement exploiter les capacités de la spectroscopie MIR en utilisant les particularités de l'ATR chauffante. Cette technique spectroscopique à été appliquée avec succès à l'étude d'échantillons alimentaires aqueux et des huiles végétales. Dans le premier cas, elle a permis d'améliorer la résolution des bandes d'absorption des composées présentes dans le vin après évaporation de l'eau. Dans le second cas, elle a permis l'accélération du processus d'altération des huiles à différentes températures et l'acquisition simultanée des spectres MIR pour suivre ces réactions.

4. Bibliographie

[1] Otto, M.; «Chemometrics: statistics and computer application in analytical chemistry», Wiley-VCH, Weinheim, **1999**.

[2] Massart, D.L.; Vandeginste, B.G.M.; Buydens, L.M.C.; De Jong, S.; Lewi, P.J.; Smeyers-Verbeke, J.; «Handbook of Chemometrics and Qualimetrics – Part A». Elsevier, Amsterdam, **1997**.

[3] Eriksson L, Johansson E, Kettaneh-Wold N; Wold S.; «Multi- and Megavariate Data Analysis. Part I – basic principles and applications». Umetrics AB, Umea, Sweden. 1-527, **2001**.

[4] Brown, S.D.; «Has the chemometrics revolution ended? Some views on the past, present and future of chemometrics». Chemometrics and Intelligent Laboratory Systems, 30 (**1995**) 49-58.

[5] Kettaneh, N.; Berglund, A.; Wold, S.; «PCA and PLS with very large data sets». Computational statistics & data analysis 48 (**2005**) 69-85.

[6] Wold, S.; Kettaneh, N.; Tjessem, K.; «Hierarchical multi-block PLS and PC models, for easier interpretation, and as an alternative to variable selection». Journal of Chemometrics, 10 (1996) 463-482.

[7] Harrington, P.; Vieira, N.; Espinoza, J.; Nien, J.; Romero, R.; Yergey, A.; «Analysis of variance-principal component analysis: A soft tool for proteomic discovery». Analytica Chimica Acta, 544 (**2005**) 118-127.

[8] Wu, W.; Massart, D.L.; de Jong, S.; «The kernel PCA algorithms for wide data. Part I: theory and algorithms». Chemometrics and Intelligent Laboratory Systems, 36 (**1997**) 165 – 172.

[9] http://statistics4u.info/fundstat_eng/dd_nipals_algo.html (04-01-2009)

[10] http://folk.uio.no/henninri/pca_module/pca_nipals.pdf (04-01-2009)

[11] http://publications.csail.mit.edu/lcs/pubs/pdf/MIT-LCS-TM-641.pdf (04-01-2009)

[12] http://math.fullerton.edu/mathews/n2003/PowerMethodMod.html (03-01-2009)

[13] Vandeginste, B.G.M.; Massart, D.L.; Buydens, L.M.C.; De Jong, S.; Lewi, P.J.; Smeyers-Verbeke, J.; «Handbook of Chemometrics and Qualimetrics – Part B. Elsevier, Amsterdam, **1998**.

[14] Wold, S.; «Pattern recognition by means of disjoint principal component models». Pattern recognition, 8 (**1976**) 127-139.

[15] Kvalheim, O.M.; Karstang, T.V.; «SIMCA – classification by means of disjoint cross validated principal components models», in *Multivariate pattern recognition in chemometrics illustrated by case studies*, Brereton, R.G.; ed., 209 – 245. Elsevier, Amsterdam, **1992**.

[16] Cudeck, R.; «Exploratory factor analysis», in *Handbook of applied multivariate statistics and multivariate modelling,* Tinsley, H.E.A.; Brown, S.D.; eds., 265 – 295. Academic press, San Diego, **2000**.

[17] http://www.statistics4u.info/fundstat_eng/dd_nipals_algo.html

[18] Martens, H.; Naes, T.; «Multivariate Calibration», John Wiley & Sons, **1989**.

[19] Wold, S.; Sjostrom, M.; Eriksson, L.; «PLS-regression: a basic tool for chemometrics», Chemometrics and intelligent laboratory systems, 58 (**2001**) 109-130.

[20] Wold, S.; Trygg, J.; Berglund, A.; Antti, H.; «Some recent developments in PLS modelling», Chemometrics and intelligent laboratory systems, 58 (**2001**) 131-150.

[21] Tenenhaus, M.; «La regression PLS», Editions Technip, Paris, **1998**.

[22] http://www.statsoft.com/textbook/stpls.html (05-01-2009)

[23] Barros, A.S.; «Contribution à la sélection et la comparaison de variables caractéristiques», Doctoral thesis, Institut National Agronomique Paris-Grignon, Paris, **1999**.

[24] www.math.su.se/matstat/reports/seriea/2007/rep8/report.pdf (05-01-2009)

[25] Indahl, U.G.; Martens, H.; Naes, T.; «From dummy regression to prior probabilities in PLS-DA», Journal of Chemometrics 21 (**2007**) 529-536.

[26] Chevallier, S.; Bertrand, D.; Kohler, A.; Courcoux, P. ; «Application of PLS-DA in multivariate image analysis», Journal of Chemometrics, 20 (**2006**) 221-229.

[27] Barker, M.; Rayens, W.; «Partial least squares for discrimination», Journal of Chemometrics, 17 (**2003**) 166-173.

[28] Barros, A.S.; Rutledge, D.N.; «PLS_Cluster: a novel technique for cluster analysis», Chemometrics and intelligent laboratory systems, 70 (2) (**2004**) 99-112.

[29] Berglund, A.; Wold, S.; «INLR, implicit non-linear latent variable regression», Journal of Chemometrics, 11 (**1997**) 141-156.

[30] Berglund, A.; Kettaneh, N.; Uppgard, L.-L.; Wold, S.; Bendwell, N.; Cameron, D.R.; «The GIFI approach to non-linear PLS modelling», Journal of Chemometrics, 15 (**2001**) 321-336.

[31] Wold, S.; Kettaneh, N.; Tjessem, K.; «Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection», Journal of Chemometrics, 10 (**1996**) 463-482.

[32] Trygg, J.; Wold, S.; «Orthogonal projections to latent structures (O-PLS)», Journal of Chemometrics, 16 (**2002**) 119-128.

[33] Trygg, J.; «O2-PLS for qualitative and quantitative analysis in multivariate calibration», Journal of Chemometrics, 16 (2002) 283-293.

[34] Trygg, J.; Wold, S.; O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter», Journal of Chemometrics, 17 (**2003**) 53-64.

[35] Hoskudsson, A.; «Dimension of linear models», Chemometrics and intelligent laboratory systems, 1 (**1996**) 37-55.

[36] Zhang, L.; Garcia-Munoz, S.; «A comparison of different methods to estimate prediction uncertainty using partial least squares (PLS): a practitioner's perspective», Chemometrics and intelligent laboratory systems (**2009**), doi: 10.1016/j.chemolab.2009.03.007.

[37] Faber, N.M.; «Uncertainty estimation for multivariate regression coefficients», Chemometrics and intelligent laboratory systems, 64 (**2002**) 169-179.

[38] Martens, H.; Martens, M.; «Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR)», Food quality and preference, 11 (**2000**) 5-16.

[39] Anderssen, E.; Dyrstad, K.; Westad, F.; Martens, H.; «Reducing over-optimism in variable selection by cross-model validation», Chemometrics and intelligent laboratory systems, 84 (**2006**) 69-74.

[40] Wehrens, R.; Putter, H.; Buydens, L.M.C.; «The bootstrap: a tutorial», Chemometrics and laboratory systems, 54 (**2000**) 35-52.

[41] Dijksterhuis, G.B.; Heiser, W.J.; «The role of permutation tests in exploratory multivariate data analysis», Food quality and preference, 6 (**1995**) 263-270.

[42] Anderson, M.J.; «Permutation tests for univariate or multivariate analysis of variance and regression», Canadian journal of fisheries and aquatic sciences, 58 (**2001**) 626-639.

[43] Filzmoser, P.; Liebmann, B.; Varmuza, K.; «Repeated double cross validation», Journal of Chemometrics (**2009**) (www.interscience.wiley.com) DOI: 10.1002/cem.1225

[44] Xu, Q.S.; Liang, Y.Z.; Du, U.P.; «Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration, Journal of Chemometrics, 18 (**2004**) 112-120.

[45] Barros, A.S.; Rutledge, D.N.; «Principal Components Transform – Partial Least Squares (PCT-PLS): A novel method to accelerate Cross-Validation in PLS regression». Chemometrics and Intelligent Laboratory Systems, 73 (**2004**) 245– 255.

[46] Meyners, M.; «Permutation tests: are there differences in product liking?», Food Quality and Preference, 12 (**2001**) 345-351.

[47] Vis, D.J.; Westerhuis, J.A.; Smilde, A.K.; van der Greef, J.; «Statistical validation of megavariate effects in ASCA», BMC Bioinformatics 8 (**2007**) 322.

[48] Barros, A.S.; Safar, M.; Devaux, M.F.; Rober, P.; Bertrand, D. ; Rutledge, D.N. ; «Relations between mid-infrared and near-infrared spectra detected by analysis of variance of an intervariable data matrix», Applied spectroscopy, 51 (**1997**) 1384-1393.

[49] Rutledge, D.N.; Barros, A.S.; Vackier, M.C.; Baumberger, S.; Lapierre, C.; «Analysis of time domain NMR and other signals», in *Advances in Magnetic Resonance in Food Science*. Belton, P.S.; Hills, B.P.; Webb, G.A.; eds., 203-216. The Royal Society of Chemistry, Cambridge, **1999**.

[50] Rutledge, D.N.; Barros, A.S.; Giangiacomo, R.; «Interpreting near infrared spectra of solutions by outer product analysis with time domain NMR», in *Magnetic Resonance in Food Science – A view to the future*. Webb, G.A.; Belton, P.S.; Gil, A.M.; Delgadillo, I., eds., 179-192, Royal Society of Chemistry, Cambridge, **2001**.

[51] Lopes, M.H.; Barros, A.S.; Pascoal Neto, C.; Rutledge, D.; Delgadillo, I.; Gil, A.M.; «Variability of cork from Portuguese *Quercus suber* studied by solid-state ^{13}C-NMR and FTIR spectroscopies», Biopolymers (Biospectroscopy), 62 (**2001**) 268-277.

[52] Smilde, A.; Bro, R.; Geladi, P.; «Multi-way analysis: applications in the chemical sciences», Wiley, **2004**.

[53] Forshed, J.; Stolt, R.; Idborg, H.; Jacobsson, S.P.; «Enhanced multivariate analysis by correlation scaling and fusion of LC/MS and 1H NMR data», Chemometrics and Intelligent Laboratory Systems, 85 (**2007**) 179-185.

[54] Rutledge, D.N.; Bouveresse, D.J.-R.; «Multi-way analysis of outer-product arrays using PARAFAC», Chemometrics and Intelligent Laboratory Systems, 85 (**2007**) 170-178.

[55] Nounour, M.N.; Bakshi, B.R.; «Multiscale methods for denoising and compression», in *Wavelets in chemistry. Walczak, B.; ed., 119-148, Elsevier, **2000**.*

[56] Vogt, F.; Tacke, M.; «Fast principal component analysis of large data sets», Chemometrics and intelligent laboratory systems, 59 (**2001**) 1-18.

[57] Wu, W.; Massart, D.L.; de Jong, S.; «Kernel-PCA algorithms for wide data. Part II: Fast cross-validation and application in classification of NIR data». Chemometrics and Intelligent Laboratory Systems 37 (**1997**) 271-280.

[58] Barros, A.S.; Pinto, R.; Jouan-Rimbaud Bouveresse; D.; Rutledge, D.; «Principal component transform – Outer product analysis in the PCA context», Chemometrics and Intelligent Laboratory Systems, 93 (**2008**) 43-48.

[59] Barros, A.S.; Pinto, R.; Delgadillo, I.; Rutledge, D.N.; «Segmented principal component transform–Partial Least Squares regression». Chemometrics and Intelligent Laboratory Systems, 89 (**2007**) 59-68.

[60] Bouveresse, D.J.-R.; Rutledge, D.N.; «Two new extensions of principal component transform to compute a PLS2 model between two wide matrices: PCT-PLS2 and segmented PCT-PLS2». Analytica Chimica Acta, publié sur le web le 17 Janvier **2009**.

[61] Barros, A.S.; Rutledge, D.N.; «Segmented principal component transform–principal component analysis». Chemometrics and Intelligent Laboratory Systems, 78 (**2005**) 125-137.

[62] Westerhuis, J.A.; Kourti, J.A.; MacGregor, J.F.; «Analysis of multiblock and hierarchical PCA and PLS models». Journal of Chemometrics, 12 (5) (**1998**) 301-321.

[63] Climaco Pinto, R.; Barros, A.S.; Bouveresse, D. J.-R.; Rutledge, D.; «Using Principal Component Transform to accelerate Outer Product – Partial Least Squares Regression calculations». Poster presentation, Chimiométrie 2005, Villeneuve d'Ascq, France, **2005**.

[64] Fearn, T.; Riccioli, C.; Garrido-Varo, A.; Guerrero-Ginel, J.E.; «On the geometry of SNV and MSC, Chemometrics and Intelligent Laboratory Systems, 96 (**2006**) 22-26.

[65] Pinto, R.; Barros, A.S.; Bouveresse, D. J.-R.; Rutledge, D.; «Using the Principal Component Transform (PCT) framework to enable Outer Product (OP) - Partial Least Squares (PLS) regression analysis in huge matrices». Poster presentation, Chimiométrie 2005, Villeneuve d'Ascq, France, **2005**.

[66] Harrington, P.; Vieira, N.; Chen, P.; Espinoza, J.; Nien, J.; Romero, R.; Yergey, A.; «Proteomic analysis of amniotic fluids using analysis of variance-principal component analysis and fuzzy rule-building expert systems applied to matrix-assisted laser desorption/ionization mass spectrometry». Chemometrics and Intelligent Laboratory Systems, 82 (**2006**) 283-293.

[67] Smilde, A.K.; Jansen, J.J.; Hoefsloot, H.C.J.; Lamers, R.-J.A.N.; van der Greef, J.; Timmerman, M.E.; «ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data», Bioinformatics, 21 (13) (**2005**) 3042-3048.

[68] Jansen, J.J.; Hoefsloot, H.C.J.; van der Greef, J.; Timmerman, M.E.; Westerhuis, J.A.; Smilde, A.K.; «ASCA: analysis of multivariate data obtained from an experimental design», Journal of Chemometrics, 19 (**2005**) 469-481.

[69] Climaco Pinto, R.; Bosc, V.; Noçairi, H.; Barros, A.S.; Rutledge, D.N.; «Using ANOVA-PCA for discriminant analysis: Application to the study of mid-infrared spectra of carraghenan gels as a function of concentration and temperature». Analytica chimica acta 629 (**2008**) 47-55.

[70] Zhu, Y.; Fearn, T.; Samuel, D.; Dhar, A.; Hameed, O.; Bown, S.; Lovat, L.B.; «Error removal by orthogonal subtraction (EROS): a customised pre-treatment for spectroscopic data», Journal of Chemometrics, 22 (**2008**) 130-134.

[71] Climaco Pinto, R.; Barros, A.S.; Locquet, N.; Schmidtke, ; Rutledge, D.N. ; «Improving the detection of significant factors using ANOVA-PCA by selective reduction of residual variability». Soumis à Analytica Chimica Acta le 10 Février **2009**.

[72] Qannari, E.M.; Wakeling, I. ; Halliday, J.H.; MacFie, J.H.; "A hierarchy of models for analysing sensory data", Food quality and preference 6 (**1995**) 309-314.

[73] Qannari, E.M.; Wakeling, I.; Courcoux, P.; MacFie, H.J.H.; "Defining the underlying sensory dimensions", Food quality and preference 11 (**2000**) 151-154.

[74] Hanafi, M.; Mazerolles, G.; Dufour, E.; Qannari, E.M. ; « Common components and specific weight analysis and multiple co-inertia analysis applied to the coupling of several measurement techniques", Journal of Chemometrics, 20 (**2006**) 172-183.

[75] SAISIR (**2008**). Package of function for chemometrics in the Matlab (R) environment. Dominique Bertrand coordinator (bertrand@nantes.inra.fr). Unité "Biopolymères, Interactions, Assemblage". INRA, rue de la Géraudière - BP 71627 - 44316 Nantes Cedex 3 France.

[76] Bouveresse, D. J.-R. ; Climaco Pinto, R. ; Schmidtke, L.; Locquet, N. ; Rutledge, D.N.; «A multi-block extension of the APCA procedure for the identification of significant factors». Soumis à Analytica Chimica Acta le 10 Avril **2009**.

[77] Bertrand, D. ; Dufour, E. ; eds. «La spectroscopie infrarouge *et ses applications analytiques- 2ème édition*. Editions Tec & Doc, Paris, **2006**.

[78] Ismail, A.A.; van de Voort, F.R.; Sedman, J.; «Fourier Transform Infrared Spectroscopy : Principles and applications», in *Instrumental methods in food analysis*, Paré, J.R.J.; Bélanger, J.M.R.; eds., 93-140. Elsevier, **1997**.

[79] Stuart, B. ; «Infrared spectroscopy : fundamentals and applications», Wiley, **2004**.

[80] Guide for infrared spectroscopy, Bruker Optics (www.brukeroptics.com/downloads) (02-05-2009)

[81] Cert, A.; Moreda, W.; Pérez-Camino, M.C.; «Chromatographic analysis of minor constituents in vegetable oils». Journal of Chromatography A, 881 (**2000**) 131-148.

[82] Belitz, H.D.; Grosch, W.; «Food chemistry», Springer-Verlag, Berlin, **1986**.

[83] Cowan, J.C.; «Polymerization, copolymerization and isomerisation», Journal of the American oil chemist's society, 31 (11) (**1954**) 529-534.

[84] Gercar, N.; Smidovnik, A.; «Kinetics of geometrical isomerisation of unsaturated FA in soybean oil», Journal of the American oil chemist's society, 79 (5) (2002) 495-500.

[85] Guillén M. D.; Cabo, N.; «Fourier transform infrared spectra data versus peroxide and anisidine values to determine oxidative stability of edible oils», Food Chemistry, 77 (**2002)** 503-51.

[86] Van de Voort F.R.; Ismail A.A.; Sedman J.; Emo G.; «Monitoring the oxidation of edible oils by Fourier Transform infrared Spectroscopy», Journal of the American oil chemist's society, 71 (**1994)** 243-253.

[87] Locquet, N. ; Climaco Pinto, R.; Eveleigh, L.; Rutledge, D.N.; «Preliminary studies on the Mid-Infrared analysis of edible oils by direct heating on an ATR diamond crystal». Soumis à Food Chemistry le 15 Mai **2009**.

ANNEXES I – VII

Publications

ANNEXE I

Available online at www.sciencedirect.com
ScienceDirect
Chemometrics and Intelligent Laboratory Systems 89 (2007) 59–68
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

Segmented Principal Component Transform–Partial Least Squares regression

António S. Barros [a,*], Rui Pinto [b], Ivonne Delgadillo [a], Douglas N. Rutledge [b]

[a] *Departamento de Química, Universidade de Aveiro, 3810-193 Aveiro, Portugal*
[b] *Laboratoire de Chimie Analytique, AgroParisTech, 16 rue Claude Bernard, 75005 Paris, France*

Received 24 January 2007; received in revised form 31 May 2007; accepted 31 May 2007
Available online 9 June 2007

Abstract

An approach for doing PLS on very wide datasets is proposed in this work. The method is based on the decomposition, by means of a SVD, of non-superimposed segments of the original data matrix. It is shown that this approach uses less computer resources compared to SIMPLS and PCT–PLS1. Furthermore, it is also shown that the results obtained by this approach are the same as those obtained by other regression methods (PLS and SIMPLS). The method implementation is simple and can be done in a distributed environment.

Keywords: PLS; Cross validation; Segmented PLS; Principal Component Transform

1. Introduction

Partial Least Squares regression (PLS) is a widely used method to perform data analysis in a vast range of applications [1–4]. The use of PLS regression on very wide data sets, which is becoming more and more common when dealing with imaging, 2D spectroscopy, GC-MS, Outer-Product analysis, and the vast field of GPM (genomic, proteomics and metabonomics) [5], increases the demand on computer resources such as calculation time and memory [6]. The subject of computer resources requirements is even more important when cross-validation [7,8] is used to assess model dimensionality, particularly when using the highly demanding validation methods such as Monte Carlo [9] or bootstrapping [10]. This follows from the fact that the number of variables (m) in $\boldsymbol{X}_{(n,m)}$ and consequently in the $\mathbf{b}_{(m,1)}$ vector can be very large, where normally for very wide data matrices $m \gg n$. In the computational context, the operations on such large vectors can be very time consuming due to the frequent swapping of elements of a given vector between the main and the virtual memory.

Nevertheless, there are many regression methods that can be used to build regression models for megavariate datasets. PLS kernel [11] and its improved variants [12] are one group of such methods that has been developed in two versions, one for tall matrices and one for wide matrices. The main idea of this approach is to build, in one initial step, smaller (kernel or cross-product) matrices and to then extract the PLS model parameters from these small matrices. Another interesting approach was published by Wu and Manne [13] which provides a fast method of doing PLS regression, as well as PCR, based also on smaller tridiagonal matrices. An interesting point about this work is that one could probably avoid the cross-validation step, but that subject is still under investigation according to the authors. Wavelets could be also an alternative to deal with large datasets by compressing the original space before doing the regression modeling [14–16]. Finally, an important and a very useful approach for data treatment of very wide or tall data sets was proposed by Wu *et al.* [17], based on the concept of the kernel matrix ($\mathbf{XX}^T$). The kernel approach solves, to some extent, the eigen decomposition problems of very wide matrices. Hence, due to the dual properties of PCA, one can perform the PCA on $\mathbf{XX}^T$, if the number of object is much smaller than the number of variables, or on $\mathbf{X}^T\mathbf{X}$, if the opposite is true. However, it should be emphasized that depending on the amount of the main computer memory, even doing these matrix products can be very time consuming due to the high number of memory page faults (due to the lack of locality of the values in memory) that occur if the matrices are not fully loaded into the main memory. However, this concept will very useful in the context of this work, as it will be shown in Section 2.3.

In a previous work [18], an approach was proposed to build PLS regression models based on the calculations of the PLS

* Corresponding author. Tel.: +351 234 372 581; fax: +351 234 370 084.
E-mail address: antbarros@dq.ua.pt (A.S. Barros).

0169-7439/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2007.05.009

models in a loss-less compressed space, the Principal Components space. The procedure, defined as Principal Component Transform–Partial Least Squares regression (PCT–PLS), was based on the full eigen decomposition of the **X** matrix before proceeding to the PLS regression. The scores are first calculated from the **X** matrix and the PLS regression is then applied to the scores in this compressed space. This approach reduced the time for PLS modeling and at the same time decreased the amount of memory needed to perform the calculations. However, the full eigen decomposition of very wide **X** matrices, the limiting step in PCT–PLS, can be also very computing resource demanding. The present work proposes the combination of these two techniques (SegPCT–PCA and PCT–PLS) to perform PLS in very wide data sets, defined as Segmented Principal Component Transform–Partial Least Squares regression (SegPCT–PLS).

2. Theory

2.1. Notations

Matrices are shown in bold uppercase (**X**), column vectors in bold lowercase (**x**) and row vectors ($\mathbf{x}^T$) — transposed. To facilitate comprehension, matrix dimensions are shown as $\mathbf{X}_{(n,m)}$, where n is the number of rows and m is the number of columns.

The symbol | is used to show concatenation or segmentation points, depending on the context. For instance, if $\mathbf{X}=[\mathbf{X}_1|\mathbf{X}_2|\ldots \mathbf{X}_q]$, matrix **X** is segmented into q sub-matrices.

2.2. Partial Least Squares regression

In the context of multivariate regression, one has the following equation:

$$\mathbf{y}_{(n,1)} = \mathbf{X}_{(n,m)}\mathbf{b}_{(m,1)} + \mathbf{f}_{(n,1)} \quad (1)$$

where n is the number of rows (objects), m is the number of columns (variables), **y** is the vector of dependent responses, **X** the matrix of independent variables, **b** the regression vector (b coefficients) and the **f** vector is the variability not accounted for by the regression model.

The PLS1 solution to Eq. (1) is given by:

$$\mathbf{b}_{(m,1)} = W_{(m,k)}\left[\mathbf{P}^T_{(k,m)}\mathbf{W}_{(m,k)}\right]^{-1} c^T_{(k,1)} \quad (2)$$

where k is the number of Latent Variables (LVs), **W** is the loadings weights, **P** the **X** loadings and **c** the **y** loadings.

2.2.1. General description of the PLS1 algorithm

The PLS1 bilinear decomposition algorithm can be generally described as follow:

Let $\mathbf{X}_{(n,m)}$ and $\mathbf{y}_{(n,1)}$ be the independent matrix and dependent vector, respectively. Starting with a vector $\mathbf{u}_{(n,1)}=\mathbf{y}_{(n,1)}$ vector:

1. $\mathbf{w}^T_{(1,m)}=\mathbf{u}^T_{(1,n)}\,\mathbf{X}_{(n,m)}$
2. $\mathbf{t}_{(n,1)}=\mathbf{X}_{(n,\mathrm{m})}\,\mathbf{w}_{(m,1)}$
   $t=t/\|t\|$
3. $\mathbf{c}^T_{(1,1)}=\mathbf{t}^T_{(1,n)}\,\mathbf{y}_{(n,1)}$
4. $\mathbf{u}_{(n,1)}=\mathbf{y}_{(n,1)}\,\mathbf{c}_{(1,1)}$
   steps 1 to 4 are iterated until convergence of the **t** vector, then
5. $\mathbf{p}^T_{(1,m)}=\mathbf{t}^T_{(1,n)}\,\mathbf{X}_{(n,m)}$
6. $\mathbf{E}_{(n,m)}=\mathbf{X}_{(n,m)}-\mathbf{t}_{(n,1)}\,\mathbf{p}^T_{(1,m)}$
7. $\mathbf{f}_{(n,1)}=\mathbf{y}_{(n,1)}-\mathbf{t}_{(n,1)}\,\mathbf{c}^T_{(1,1)}$

2.2.2. Segmented Principal Component Transform–Partial Least Squares regression

The present work is based on a suggest approach (Segmented Principal Component Transform–Principal Component Analysis — SegPCT–PCA) [19] to overcome the problem of performing the eigen decomposition of very wide matrices. The same concept could be used in order to improve the performance of PCT–PLS. The increase in performance in SegPCT–PCA for very wide matrices compared to PCA results from performing PCA on column-wise segments of the initial wide matrix. For each small segment, the scores and loadings are recovered. Concatenating the scores, recovered for each segment, and then performing a second PCA on those concatenated scores yields the same scores as if one had done a PCA on the initial wide matrix. This property was shown in Eq. (11) of the SegPCT–PCA work [19], where the scores of the original space can be easily recovered from a smaller set of matrices and vectors. An interesting and useful side effect of this segmentation approach is that it is not necessary to optimize the segment width. Moreover, if needed, the original space loadings are also easily reconstructed by calculating the products between the loadings from each segment and those from the PCA on the concatenated scores (following Eq. (12) of the SegPCT–PCA work).

Having recovered the scores by means of the SegPCT–PCA procedure, one will use them as input to PCT–PLS1 [18] to build a PLS1 model. This approach will accelerate the calculation of PLS1 models, especially in cross-validation scenarios.

The implementation of this approach is straightforward and is shown as pseudo-code in Table 1. A Matlab® implementation detail of SegPCT–PCA is described in Appendix A of [19].

If one needs to have the original space **b** vector for interpretation, one can recover it by pre-multiplying the $\mathbf{b}_{\mathrm{PCT}}$ vector by the original space loadings: $\mathbf{b}=\mathbf{P}\,\mathbf{b}_{\mathrm{PCT}}$ as shown in [18], or alternatively, by using the segmented approach as described below.

For completeness, even though it is not necessary for implementation purposes, the demonstration of the segmentation approach in the PLS1algorithm is shown in Appendix A, where one can see some interesting properties of this procedure. At the same time, Appendix B shows the calculation of the **b** vector in the original space using the segmented approach. As shown, Eq. (B.5) represents the **b** vector in the PC-space ($\mathbf{b}_{\mathrm{PCT}}$).

Table 1
Pseudo-code of the SegPCT–PLS1 procedure

X, y	Input of X and y
[**T**, **P**] ⟵ SegPCT-PCA(**X**, segw)	**T** : scores; **P** : loadings (optional); *segw* : segment width
$\mathbf{b}_{\mathrm{PCT}}$ ⟵ PLS1(**T**, **y**)	$\mathbf{b}_{\mathrm{PCT}}$: PCT **b** vector (**b** vector in the PC space)

Table 2
Optimum models using the SIMPLS and PLS-based methods, cross-validation (leave-one-out) for data set 1

Method	LV	RMSECV	Time (s)	Comments
PLS1	4	0.250	43	
SIMPLS	4	0.250	8	
PCT–PLS1	4	0.250	15	Eigen decomposition of $\mathbf{X}$
PCT–PLS1	4	0.250	13	Eigen decomposition of $\mathbf{XX}^T$

Since PLS1 cross-validation in the original space can be very computationally demanding, the segmented PCT regression matrices and vectors could be used instead, as this requires much less memory.

2.3. SegPCT–PLS performance considerations

The implementation of SegPCT–PLS1 has an important advantage for the calculation and use of scores of a wide matrix $\mathbf{X}$. However, the segmenting approach of a wide $\mathbf{X}$ matrix shows that the limiting step is the eigen decomposition of each matrix segment ($\mathbf{X}_i$ − segment $i=\{1,...,q\}$) (Eq. (A.1)), or more precisely how the eigen decomposition of the matrix segments is done. In fact there are several ways to perform the eigen decomposition of these sub-matrices. One of them is to do the Singular Values Decomposition (SVD) [20] of each X_i sub-matrix. This approach, however, can be time consuming for the cases where each of the i segments in $\{1,...,q\}$ has a large number of variables (m_i) compared to the number of objects (n). Therefore, one could take advantage of the kernel matrices [17,21] for a fast eigen decomposition of the segmented matrices.

The SVD of an $\mathbf{X}$ matrix is given by:

$$\mathbf{X}_{(n,m)} = \mathbf{U}_{(n,k)}\mathbf{S}_{(k,k)}\mathbf{P}^T_{(k,m)}$$

where $\mathbf{U}$ and $\mathbf{P}$ are the eigenvectors matrices (orthonormal) and $\mathbf{S}$ is a diagonal matrix whose diagonal elements are the singular values.

The scores ($\mathbf{T}$) are obtained by:

$$\mathbf{T}_{(n,k)} = \mathbf{U}_{(n,k)}\mathbf{S}_{(k,k)} \quad (3)$$

On the other hand, and considering that m is much larger than n, one could perform the SVD on the much smaller kernel matrix of the form $\mathbf{X}_{(n,m)}\mathbf{X}^T_{(m,n)}$.

The SVD of $\mathbf{XX}^T$ is given by:

$$\mathbf{XX}^T_{(n,n)} = \mathbf{U}_{(n,k)}\mathbf{S}^2_{(k,k)}\mathbf{P}^T_{(k,n)} \quad (4)$$

which is a much smaller matrix than $\mathbf{X}_{(n,m)}$ provided that m is much larger than n.

The scores ($\mathbf{T}$) of the original $\mathbf{X}$ matrix are obtained by:

$$\mathbf{T}_{(n,k)} = \mathbf{U}_{(n,k)}\mathbf{S}^{1/2}_{(k,k)} \quad (5)$$

Therefore the segmentation in Eq. (A.1):

$$\mathbf{X}_{(n,m1)} = \left[\mathbf{X}_{1(n,m1)}|\mathbf{X}_{2(n,m2)}|\ldots|\mathbf{X}_{q(n,mq)}\right]$$

can be replaced by the following expression:

$$\left[\mathbf{X}_{1(n,m1)}\mathbf{X}^T_{1(m1,n)}|\mathbf{X}_{2(n,m2)}\mathbf{X}^T_{2(m2,n)}|\ldots|\mathbf{X}_{q(n,mq)}\mathbf{X}^T_{q(mq,n)}\right]$$

where, according to Eq. (4), the segmented eigen decomposition yields:

$$\left[\mathbf{U}_{1(n,k1)}\mathbf{S}^2_{1(k1,k1)}\mathbf{P}^T_{1(k1,n)}|\mathbf{U}_{2(n,k2)}\mathbf{S}^2_{2(k2,k2)}\mathbf{P}^T_{2(k2,n)}|\ldots \ldots|\mathbf{U}_{q(n,kq)}\mathbf{S}^2_{q(kq,kq)}\mathbf{P}^T_{q(kq,n)}\right]$$

which can be further simplified, according to Eqs. (4) and (5), into:

$$\left[\mathbf{T}_{1(n,k1)}\mathbf{P}^T_{1(k1,n)}|\mathbf{T}_{2(n,k2)}\mathbf{P}^T_{2(k2,n)}|\ldots|\mathbf{T}_{q(n,kq)}\mathbf{P}^T_{q(kq,n)}\right]$$

It is important to highlight the fact that although the kernel matrices of the form $\mathbf{XX}^T$ are already used to accelerate the eigen decomposition of wide matrices, in many cases, the calculation of the $\mathbf{XX}^T$ can be time consuming and very memory demanding, whereas the calculations of $\mathbf{XX}^T_i$ where $i\in\{1,...,q\}$ is much faster and uses less memory as one has just to load the small segments ($\mathbf{X}_{i(n,mi)}$) into the main memory, one at a time.

3. Illustrations

3.1. Data set 1 — NIR spectra of gasoline samples

Sixty near-infrared spectra of gasoline samples with known octane numbers were used. Samples were measured using diffuse reflectance between 900 and 1700 nm, at 2 nm intervals (401 points). This classical test data set was obtained at

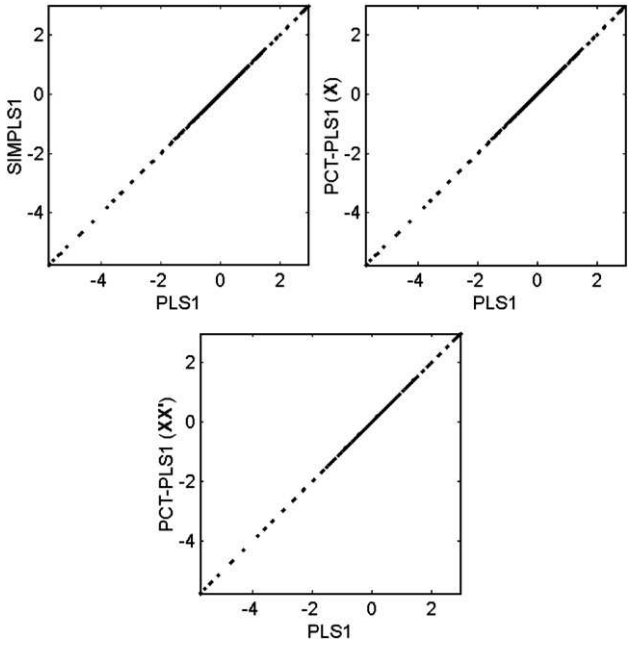


Fig. 1. **b** vector scatter plots of PLS1 vs. SIMPLS, PLS1 vs. PCT–PLS1 based on the SVD of $\mathbf{X}$ and PLS1 vs. PCT–PLS1 based on the SVD of $\mathbf{XX}^T$ (data set 1).

Table 3
Optimum SegPCT–PLS1 models, determined by cross-validation (leave-one-out) for data set 1 as a function of the segment width

Segment width	LV	RMSECV	Time (s)
Eigen decomposition of $\mathbf{X}_q$			
60	4	0.250	14
80	4	0.250	16
100	4	0.250	17
120	4	0.250	16
140	4	0.250	15
160	4	0.250	14
180	4	0.250	15
200	4	0.250	15
Eigen decomposition of $\mathbf{X}\mathbf{X}_q^T$			
60	4	0.250	17
80	4	0.250	15
100	4	0.250	18
120	4	0.250	14
140	4	0.250	18
160	4	0.250	14
180	4	0.250	13
200	4	0.250	14

ftp://www.ftp.clarkson.edu/pub/hopkepk/chem-data/kalivas and has been discussed by Kalivas [22]. The data set is composed of an **X** matrix with 60 samples and 401 variables and a **y** vector (60 elements) with the octane numbers. The data set was analysed as downloaded.

3.2. Data set 2 – outer product between FTIR and NMR of cork samples

This data set is composed of 20 cork samples for which the suberin content was determined [23]. For each sample, Fourier Transform Infrared spectroscopy (FTIR) and solid state ^{13}C-Nuclear Magnetic Resonance (NMR) spectra were obtained. Both sets of spectra were combined by means of an Outer Product matrix [23] in order to analyze the relationships between the two domains. The resulting Outer Product matrix (FTIR ⊗ ^{13}C-NMR) consisted of 20 objects and 450,702 (882 × 511) variables. The main goal was to establish a PLS calibration model between the OP matrix of spectra and the suberin content.

3.3. Data set 3 — outer product between MIR and NIR of oils samples

The data set is composed of 45 spectra of 15 samples of different oils measured by Mid Infrared (MIR) and Near Infrared (NIR) in triplicate. The two domains were combined by

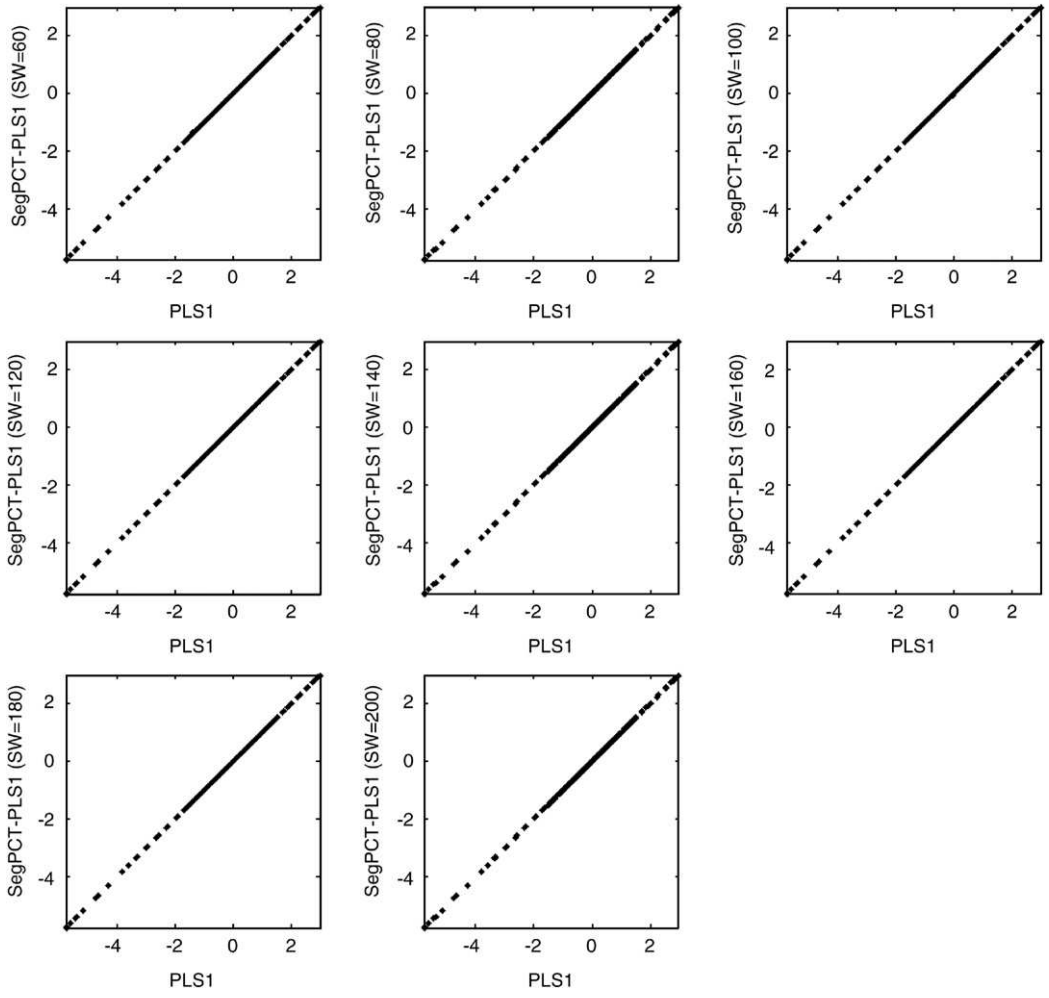


Fig. 2. **b** vector scatter plots of PLS1 vs. SegPCT–PLS1 as a function of the segment width (SW = 60, 80, 100, 120, 140, 160, 180, 200), based on the SVD of **X** (data set 1).

means of an OP and the resulting OP matrix of 45 objects by 490,625 variables was used to build a PLS regression model to predict the iodine content in order to study the relationships between the two domains.

3.4. Data set 4 — random simulated data

Several (**X**) matrices with 100 objects and from 10^5 to 10^6 variables were built using numbers uniformly pseudo-random distributed. For each generated **X** matrix, a **y** vector was calculated by a linear combination of a sub-set of the randomly generated **X** variables — including at least the square root of the total number of variables. The purpose of this data set was to evaluate the performance of the proposed method (SegPCT–PLS) as a function of increasing data matrix size.

4. Results and discussion

The outline of the following of this section is as follows. A comparison will be made for each type of data set between the performance and characteristics of models calculated using PLS, SIMPLS [24], PCT–PLS and SegPCT–PLS. The use of kernel matrices of the form $\mathbf{XX}^{\mathrm{T}}$ will also be studied. All calculations were done on a Pentium IV 1.6 GHz with 256 Mbytes of RAM.

4.1. Data set 1 — NIR spectra of gasoline samples

This sub-section concerns the comparison of several multivariate PLS-based regression methods applied to a reference data set in order to assess the model characteristics of SegPCT–PLS in different contexts. The model's predictive ability was evaluated by internal cross-validation using the Root Mean Square Error of Cross-Validation (RMSECV).

Table 2 shows the comparison of different PLS-based procedures for the determination of the octane number. As can be seen, all methods have found the same model parameters of 4 Latent Variables (LVs) and a prediction error of 0.250. These results are confirmed by the linear relationships in Fig. 1 comparing the **b** vector profiles between standard PLS1 and the other PLS-based methods. The different calculation speeds presented in Table 2 show that the SIMPLS1 method is fastest. The PCT–PLS1 procedures have similar performances and are slightly slower than SIMPLS1. However, the purpose of these particular calculations is not to compare performances but to confirm that all of them give the same results.

The essential feature of this section is shown in Table 3 showing the application of SegPCT–PLS1 based on the eigen decomposition of $\mathbf{X}_q$ and of $\mathbf{XX}_q^T$, as a function of the segments widths. As can be clearly seen, all models gave the same statistical results, error of prediction and model dimensionality, meaning that optimization of the segment width (optimal SW) is not a major concern. Moreover, the calculation speeds are very similar. The statistical results shown in Table 3 are comparable

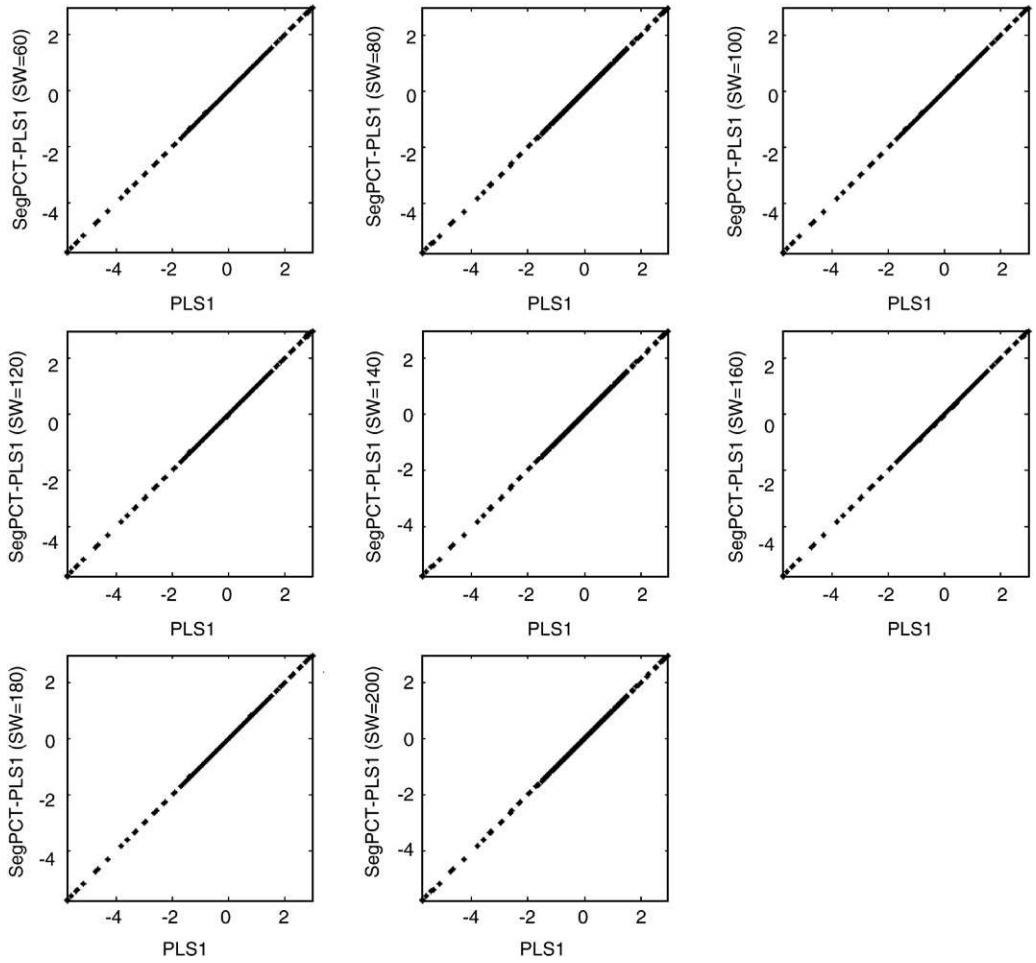


Fig. 3. **b** vector scatter plots of PLS1 vs. SegPCT–PLS1 as a function of the segment width (SW = 60, 80, 100, 120, 140, 160, 180, 200), based on the SVD of $\mathbf{XX}^T$ (data set 1).

Table 4
Optimum models given by cross-validation (leave-one-out) using the SIMPLS and PLS-based methods for data set 2

Method	LV	RMSECV	Time (s)	Memory(Mb)	Comments
SIMPLS	3	2.272	6324	215	
PCT–PLS1	3	2.272	578	114	Eigen decomposition of $\mathbf{X}$
PCT–PLS1	3	2.272	42	36	Eigen decomposition of $\mathbf{XX}^T$

to those of Table 2, which suggests that SegPCT–PLS1 gives the same results as PLS1 or SIMPLS. This is confirmed by the comparison of the **b** vectors for PLS1 and SegPCT–PLS1 as a function of the segments widths based on the eigen decomposition of $\mathbf{X}_q$ (Fig. 2) and of $\mathbf{XX}^T$ (Fig. 3). Figs. 2 and 3 show that the **b** vectors of both SegPCT–PLS1 methods are equal to the PLS1 **b** vector.

The lower speed performances of the SegPCT–PLS1 and PCT–PLS1 methods are not an issue in this case as they are most suitable for very wide data sets, which is the subject of the next two sub-sections. As such, the important conclusion of these results is that SegPCT–PLS1 methods give the same models as standard PLS1 and SIMPLS.

4.2. Data set 2 — outer product between FTIR and NMR of cork samples

The results of the previous section showed that both approaches used for SegPCT–PLS1, *i.e.* based on the SVD of $\mathbf{X}$ and on the SVD of $\mathbf{XX}^T$ give the same models as the PLS1, SIMPLS and PCT–PLS1 procedures.

In this section a performance comparison is done using a wide data set (data set 2). Once more, Table 4 shows that the regression models from the different approaches give the same results. However, this time, SIMPLS is the slowest method (*ca.* 105 min), whereas the PCT–PLS1 procedure based on the eigen decomposition of the $\mathbf{XX}^T$ kernel matrix is the fastest. Additionally, the latter method takes approximately 6 times less memory than the former procedure. This is obvious as the SIMPLS1 method used the original variable space, which corresponds to vectors with 450,702 elements, whereas the PCT–PLS1 works on smaller ones, vectors with 20 elements (the theoretical maximum rank of the data set).

Table 5 shows that, once more, the SegPCT–PLS1 gives the same statistical results as the SIMPLS and PCT–PLS methods (optimal model dimensionality and calibration error). It is also to be noted that the eigen decomposition of the segmented $\mathbf{X}_q$ matrices is slower than the eigen decomposition of the $\mathbf{XX}_q^T$ kernel matrices, which is logical as $\mathbf{XX}_q^T$ kernel matrices are smaller than the $\mathbf{X}_q$ matrices. Concerning the SegPCT–PLS1 based on the segmented kernel matrices, one can see that it takes approximately the same time to perform the cross-validation as the PCT–PLS1 based on the kernel matrix. However, Tables 4 and 5 show clearly that the former approach requires much less memory than the latter. For example, one can see that PCT–PLS1 based on the kernel matrix takes 36 Mb of memory while SegPCT–PLS1 based on the segmented kernel matrices takes less than 1 Mb in the case of segments with 4096 variables (Table 5).

It is important to note that there is no need to optimize for the segment size for the SegPCT–PLS1 procedure as all the segment widths gave the same models. Therefore, the results suggest that by using small to medium size segments (*e.g.*, one hundredth of the original-variable matrix size) one can determine the regression models in a reasonable amount of time and using less computer memory because one just needs to load into the main memory a small portion (segment) of the $\mathbf{X}$ matrix, as opposed to loading all the $\mathbf{X}$ matrix into the memory as with the other PLS1-based methods.

Table 6
Optimum SegPCT–PLS1 models determined by cross-validation (leave-three-out), based on the eigen decomposition of the segmented kernel matrices ($\mathbf{XX}_q^T$) *as a function of the segment width for data set 3*

Segment width	LV	RMSECV	Time(s)	Memory(Mbytes)
1024	2	12.161	124	4.5
2048	2	12.161	109	2.7
4096	2	12.161	105	3.4
8192	2	12.161	100	2.5
16384	2	12.161	100	6.7
32768	2	12.161	99	13
65536	2	12.161	100	17
131072	2	12.161	98	41
262144	2	12.161	102	87

4.3. Data set 3 — outer product between MIR and NIR of oils samples

Having shown the advantage (of SegPCT–PLS1 based on the eigen decomposition of the segmented kernel matrices ($\mathbf{XX}_q^T$)) in terms of memory usage when compared to the other methods, this section compares SegPCT–PLS1 and PCT–PLS1, both based on the SVD of the kernel matrices, in the context of an internal cross-validation with leave-three-out and using an even larger data set.

The application of PCT–PLS1 based on the eigen decomposition of the kernel matrix to this data set yielded a regression model with 2 LVs and with a calibration error of 12.161, took 105 s to complete the cross-validation and required 87 Mb of memory. The results in Table 6 show that the SegPCT–PLS1 method takes approximately the same time to perform the calculations as the PCT–PLS1. Nevertheless, the memory

Table 5
Optimum SegPCT–PLS1 models, determined by cross-validation (leave-one-out) as a function of the segment width for data set 2

Segment width	SegPCT–PLS1 (based on the SVD of X_q)				SegPCT–PLS1 (based on the SVD of XX_q^T)			
	LV	RMSECV	Time(s)	Memory (MBytes)	LV	RMSECV	Time(s)	Memory (MBytes)
1024	3	2.272	108	1.6	3	2.272	42	1.7
2048	3	2.272	126	1.3	3	2.272	42	1.1
4096	3	2.272	165	1.7	3	2.272	40	<1.0
8192	3	2.272	209	3	3	2.272	43	1.1
16384	3	2.272	360	6	3	2.272	40	2.3
32768	3	2.272	431	14	3	2.272	40	4.7
65536	3	2.272	494	31	3	2.272	39	9.8
131072	3	2.272	489	54	3	2.272	43	15.0
262144	3	2.272	564	126	3	2.272	39	35.0

requirements are significantly lower. The SegPCT–PLS1 using, for instance, segments with 4096 variables shows a 26-fold decrease in memory compared to the PCT–PLS1. These results shows that using small segments one could have the same time performance as PCT–PLS1 with the advantage of using much less memory to perform the calculations.

4.4. Data set 4 — simulated data sets

The previous two sections showed that for the case of very wide matrices the SegPCT–PLS procedure is comparable in terms of computation speed to the PCT–PLS method. On the other hand, the SegPCT–PLS had a significantly lower demand on computer memory. The essential advantage of SegPCT–PLS compared to PCT–PLS is only highlighted when the data matrices are in fact huge. The previous examples showed matrices that, although very large, could still be fully loaded into the main memory (256 Mb in the case of the computer used for these calculations), which explains why both approaches have similar computation speeds, although the memory requirements are very different.

Since the SegPCT–PLS and PCT–PLS approaches based on the eigen decomposition of kernel matrices are the most efficient in terms of computational resources, an evaluation of both methods was done as a function of increasing matrix size.

The results for matrix sizes ranging from 10^5 to 10^6 variables and 100 objects are presented in Table 7. In the case of SegPCT–PLS1, each matrix was split into 100 segments. Calibration models were built for each case using internal cross-validation to assess the model dimensionality. For all tested cases the calibration error and the model complexity were the same.

The results of Table 7 show that in terms of computation speed the PCT–PLS1 is faster than SegPCT–PLS1 for moderately wide matrices, whereas for very wide matrices the opposite is observed. This behavior is due to the fact that as long as the matrices can be fully loaded into the main memory there is no computation speed advantage for using SegPCT–PLS1, as there is no need for memory swapping. However, if the initial matrix cannot be completely loaded into the main memory, the PCT–PLS1 procedure starts using virtual memory, which slows down the calculations. This effect can be clearly observed for the last two rows of Table 7. PCT–PLS1 required 292 and 391 Mb of virtual memory (values between parentheses) to perform the calculations, whereas for SegPCT–PLS1 this value is only about 15 Mb. Therefore, for very wide matrices, SegPCT–PLS1 is more computer-resource efficient (speed and memory) than PCT–PLS1, while for moderately wide matrices PCT–PLS1 should be used.

Table 7
Performance comparison of PCT–PLS1 and SegPCT–PLS1 methods as a function of the original matrix sizes. Cross-validation leave-5-out for data sets 4

Matrix size	PCT–PLS1		SegPCT–PLS1		
	Time (s)	Memory* (Mb)	Time (s)	Memory* (Mb)	Segment size
[100, 100 000]	49	39 (65)	106	7.6 (15.4)	1000
[100, 250 000]	132	98 (99)	194	5.4 (15.4)	2500
[100, 500 000]	298	195 (197)	302	6.3 (15.4)	5000
[100, 750 000]	891	221 (292)	413	7.3 (15.4)	7500
[100, 1000 000]	2185	220 (391)	518	8.4 (15.4)	10 000

(*) Memory values refer to the working set of the algorithms (usage of the main memory to perform the calculations), whereas memory values between parentheses are due to the amount of allocated virtual memory.

These results suggest that as function of the initial data set size one should choose the more appropriate method to efficiently determine the regression models.

5. Conclusions

The present work proposes a new approach to perform PLS regression modeling based on a combination of the PCT–PLS and SegPCT–PCA methods in order to accelerate the cross-validation computations and to reduce the amount of memory necessary for very wide datasets. This methodology will facilitate the application of PLS-based methods to very wide datasets not only for regression but also for classification/discriminant purposes such as PLS Discriminant Analysis [25–27], PLS_Cluster analysis [28] etc.

Several aspects can be further explored based on the segmentation of the initial matrix, namely, determine methods to select the most important segments for a given purpose, *e.g.* selection of regions of interest [29,30]. Furthermore, the use of a full eigen decomposition for the cases where one has a large number of objects can slow down the calculations, and as such, the evaluation of partial-eigen decomposition of each segment will be an interesting path to follow. Another interesting path will be the evaluation and application of this methodology in distributed environments, as the algorithm is inherently parallel. This could provide new ways to analyze huge datasets using computer clusters.

Appendix A

Starting with a vector $\mathbf{u}(n,1)=\mathbf{y}(n,1)$ and considering the first step in the PLS1 algorithm (Section 2.2.1):

1. $\mathbf{w}^T_{(1,m)}=\mathbf{u}^T_{(1,n)}\,\mathbf{X}_{(n,m)}$

$\mathbf{w}$ and $\mathbf{X}$ can be split into q segments or bands as:

$$\left[\mathbf{w}^T_{1(1,m1)}|\mathbf{w}^T_{2(1,m2)}|\ldots|\mathbf{w}^T_{q(1,mq)}\right] = \mathbf{u}^T_{(1,n)}\left[\mathbf{X}_{1(n,m1)}|\mathbf{X}_{2(n,m2)}|\ldots|\mathbf{X}_{q(n,mq)}\right]$$

where q is the number of segments, mq is the number of variables in segment q, and $m1+m2+\ldots+mq=m$ rearranging:

$$\left[\mathbf{w}^T_{1(1,m1)}|\mathbf{w}^T_{2(1,m2)}|\ldots|\mathbf{w}^T_{q(1,mq)}\right] = \left[\mathbf{u}^T_{(1,n)}\mathbf{X}_{1(n,m1)}|\mathbf{u}^T_{(1,n)}\mathbf{X}_{2(n,m2)}|\ldots|\mathbf{u}^T_{(1,n)}\mathbf{X}_{q(n,mq)}\right]$$

which is the same as:

$$\mathbf{w}^T_{i(1,mi)} = \mathbf{u}^T_{(1,n)}\mathbf{X}_{i(n,mi)}, \text{where } i=\{1,\ldots,q\}$$

decomposing the $\mathbf{X}_i$ sub-matrices into $\mathbf{X}_i=\mathbf{T}_{Xi}\mathbf{P}_{Xi}{}^T$, where $i=\{1,..,q\}$, one has:

$$\mathbf{w}^T_{i(1,mi)} = \mathbf{u}^T_{(1,n)}\mathbf{T}_{Xi(n,hi)}\mathbf{P}^T_{Xi(hi,mi)} \quad (A.1)$$

where hi is the number of principal components of the i segment.

Post-multiplying by $\mathbf{P}_{Xi}$, and since $\mathbf{P}_{Xi}^T\mathbf{P}_{Xi}=\mathbf{I}$, *then Eq. (A.1)* can be rearranged as:

$$\mathbf{w}_{i(1,mi)}^T\mathbf{P}_{Xi(mi,hi)} = \mathbf{u}_{(1,n)}^T\mathbf{T}_{Xi(n,hi)}, \text{where } i=\{1,\ldots,q\} \quad (A.2)$$

The left side of Eq. (A.2) can be thought of as the contribution of the partitioned loadings matrix ($\mathbf{P}_{Xi}$) of the original space to the segmented portion i of the vector weight loadings ($\mathbf{w}_i$). Assuming that this relationship can be expressed in simplified notation as:

$\mathbf{w}_{i(1,mi)}^T\,\mathbf{P}_{Xi(mi,hi)}=\mathbf{w}_{\mathrm{PCT}i}^T$, where $i=\{1,\ldots,q\}$, then Eq. (A.2) can be expressed as:

$$\mathbf{w}_{PCTi(1,hi)}^T = \mathbf{u}_{(1,n)}^T\mathbf{T}_{Xi(n,hi)} \quad (A.3)$$

Moreover, Eq. (A.3) can then be seen as:

$$\begin{aligned}&\left[\mathbf{w}_{\mathrm{PCT1}(1,h1)}^T|\mathbf{w}_{\mathrm{PCT2}(1,h2)}^T|\ldots|\mathbf{w}_{\mathrm{PCTq}(1,hq)}^T\right]\\&=\left[\mathbf{u}_{(1,n)}^T\mathbf{T}_{X1(n,h1)}|\mathbf{u}_{(1,n)}^T\mathbf{T}_{X2(n,h2)}|\ldots|\mathbf{u}_{(1,n)}^T\mathbf{T}_{Xq(n,hq)}\right]\\&=\mathbf{u}_{(1,n)}^T\left[\mathbf{T}_{X1(n,h1)}|\mathbf{T}_{X2(n,h2)}|\ldots|\mathbf{T}_{Xq(n,hq)}\right]\end{aligned}$$

or

$$\mathbf{w}_{PCT(1,h1+h2+\ldots+hq)}^T = \mathbf{u}_{(1,n)}^T\left[\mathbf{T}_{X1(n,h1)}|\mathbf{T}_{X2(n,h2)}|\ldots|\mathbf{T}_{Xq(n,hq)}\right]$$

which is equivalent to step 1 of the PLS1 algorithm, while reducing the size of the $\mathbf{w}$ vector from m to $h1+h2+\ldots+hq$, where $h1+h2+\ldots+hq<<m$.

Concerning PLS1 step 2, one has:

2. $\mathbf{t}_{(n,1)}=\mathbf{X}_{(n,\ \mathrm{m})}\,\mathbf{w}_{(\mathrm{m},1)}$

again segmenting $\mathbf{X}$ and $\mathbf{w}$ into q bands along the variables direction one obtains:

$$\begin{aligned}\mathbf{t}_{(n,1)} &= \left[\mathbf{X}_{1(n,m1)}|\mathbf{X}_{2(n,m2)}|\ldots|\mathbf{X}_{q(n,mq)}\right]\\&\times\left[\mathbf{w}_{1(m1,1)}|\mathbf{w}_{2(m2,1)}|\ldots|\mathbf{w}_{q(mq,1)}\right]\end{aligned}$$

where $[\mathbf{w}_{1(m1,1)}|\mathbf{w}_{2(m2,1)}|\ldots|\mathbf{w}_{q(mq,1)}]$ is viewed in this case as segmentation by rows.

Rearranging:

$$\begin{aligned}\mathbf{t}_{(n,1)} &= \mathbf{X}_{1(n,m1)}\mathbf{w}_{1(m1,1)}+\mathbf{X}_{2(n,m2)}\mathbf{w}_{2(m2,1)}+\ldots\\&+\mathbf{X}_{q(n,mq)}\mathbf{w}_{q(mq,1)}\end{aligned}$$

and decomposing $\mathbf{X}$:

$$\begin{aligned}\mathbf{t}_{(n,1)} &= \mathbf{T}_{X1(n,h1)}\mathbf{P}_{X1(h1,m1)}^T\mathbf{w}_{1(m1,1)}\\&+\mathbf{T}_{X2(n,h2)}\mathbf{P}_{X2(h2,m2)}^T\mathbf{w}_{2(m2,1)}+\ldots\\&+\mathbf{T}_{Xq(n,hq)}\mathbf{P}_{Xq(hq,mq)}^T\mathbf{w}_{q(mq,1)}\end{aligned} \quad (A.4)$$

and since $\mathbf{P}_{Xq(hq,mq)}^T\ \mathbf{w}_{q(mq,1)}=\mathbf{w}_{\mathrm{PCT}q(hq,1)}$, then Eq. (A.4) can be further simplified to:

$$\begin{aligned}\mathbf{t}_{(n,1)} &= \mathbf{T}_{X1(n,h1)}\mathbf{w}_{PCT1(h1,1)}+\mathbf{T}_{X2(n,h2)}\mathbf{w}_{PCT2(h2,1)}+\ldots\\&+\mathbf{T}_{Xq(n,hq)}\mathbf{w}_{PCTq(hq,1)}\end{aligned}$$

or

$$\begin{aligned}\mathbf{t}_{(n,1)} &= \left[\mathbf{T}_{X1(n,h1)}|\mathbf{T}_{X2(n,h2)}|\ldots|\mathbf{T}_{Xq(n,hq)}\right]\\&\times\left[\mathbf{w}_{PCT1(h1,1)}|\mathbf{w}_{PCT2(h2,1)}|\ldots|\mathbf{w}_{PCTq(hq,1)}\right]\end{aligned} \quad (A.5)$$

As the steps 3 and 4 of PLS1 algorithm are related to the object space they remain unchanged.

Steps 1 to 4 are iterated until convergence of the $\mathbf{t}$ vector, then considering step 5:

5. $\mathbf{p}_{(1,m)}^T=\mathbf{t}_{(1,n)}^T\,\mathbf{X}_{(n,m)}$

Once more, the segmentation of the variable-space given for $\mathbf{p}$ and $\mathbf{X}$ is:

$$\begin{aligned}&\left[\mathbf{p}_{1(1,m1)}^T|\mathbf{p}_{2(1,m2)}^T|\ldots|\mathbf{p}_{q(1,mq)}^T\right]\\&=\mathbf{t}_{(1,n)}^T\left[\mathbf{X}_{1(n,m1)}|\mathbf{X}_{2(n,m2)}|\ldots|\mathbf{X}_{q(n,mq)}\right]\end{aligned}$$

Decomposing $\mathbf{X}_i$ where $i=\{1,\ldots,q\}$ then:

$$\begin{aligned}&\left[\mathbf{p}_{1(1,m1)}^T|\mathbf{p}_{2(1,m2)}^T|\ldots|\mathbf{p}_{q(1,mq)}^T\right]\\&=\mathbf{t}_{(1,n)}^T\left[\mathbf{T}_{X1(n,h1)}\mathbf{P}_{X1(h1,m1)}^T|\mathbf{T}_{X2(n,h2)}\mathbf{P}_{X2(h2,m2)}^T|\ldots\right.\\&\left.|\mathbf{T}_{Xq(n,hq)}\mathbf{P}_{Xq(hq,mq)}^T\right]\end{aligned}$$

and rearranging:

$$\begin{aligned}&\left[\mathbf{p}_{1(1,m1)}^T|\mathbf{p}_{2(1,m2)}^T|\ldots|\mathbf{p}_{q(1,mq)}^T\right]\\&=\left[\mathbf{t}_{(1,n)}^T\mathbf{T}_{X1(n,h1)}\mathbf{P}_{X1(h1,m1)}^T|\mathbf{t}_{(1,n)}^T\mathbf{T}_{X2(n,h2)}P_{X2(h2,m2)}^T|\ldots\right.\\&\left.|\mathbf{t}_{(1,n)}^T\mathbf{T}_{Xq(n,hq)}\mathbf{P}_{Xq(hq,mq)}^T\right]\end{aligned}$$

which is the same as:

$$\mathbf{p}_{i(1,mi)}^T = \mathbf{t}_{(1,n)}^T\mathbf{T}_{Xi(n,hi)}\mathbf{P}_{Xi(hi,mi)}^T, \text{where } i=\{1,\ldots,q\} \quad (A.6)$$

Post-multiplying by $\mathbf{P}_{Xi}$ and since $\mathbf{P}_{Xi}^T\mathbf{P}_{Xi}=\mathbf{I}$:

$$\mathbf{p}_{i(1,mi)}^T\mathbf{P}_{Xi(mi,hi)} = \mathbf{t}_{(1,n)}^T\mathbf{T}_{Xi(n,hi)} \quad (A.7)$$

The left side of Eq. (A.7) can be viewed as the contribution of the segmented loadings ($\mathbf{P}_{Xi}$) to the segmented portion i of loading $\mathbf{X}(\mathbf{p}_i)$. Using the notation $\mathbf{p}_i^T\,\mathbf{P}_i=\mathbf{p}_{PCTi}^T$, where $i=\{1,\ldots,q\}$ then Eq. (A.7) can be further simplified into:

$$\mathbf{p}_{PCTi(1,hi)}^T = \mathbf{t}_{(1,n)}^T\mathbf{T}_{Xi(n,hi)}, \text{where } i=\{1,\ldots,q\} \quad (A.8)$$

Eq. (A.8) is the same as:

$$\begin{aligned}&\left[\mathbf{p}_{PCT1(1,h1)}^T|\mathbf{p}_{PCT2(1,h2)}^T|\ldots|\mathbf{p}_{PCTq(1,hq)}^T\right]\\&=\left[\mathbf{t}_{(1,n)}^T\mathbf{T}_{X1(n,h1)}|\mathbf{t}_{(1,n)}^T\mathbf{T}_{X2(n,h2)}|\ldots|\mathbf{t}_{(1,n)}^T\mathbf{T}_{Xq(n,hq)}\right]\end{aligned}$$

or

$$\mathbf{p}_{PCT(1,h1+h2+\ldots+hq)}^T = \mathbf{t}_{(1,n)}^T\left[\mathbf{T}_{X1(n,h1)}|\mathbf{T}_{X2(n,h2)}|\ldots|\mathbf{T}_{Xq(n,hq)}\right] \quad (A.9)$$

This is the PCT equation equivalent to step 5 of the PLS1 algorithm.

Step 6 of the PLS1 algorithm concerns the update of the $\mathbf{X}$ matrix and $\mathbf{f}$ vector before the calculation of the next k Latent Variable (LV).

6. $\mathbf{E}_{(n,m)}=\mathbf{X}_{(n,m)}-\mathbf{t}_{(n,1)}\,\mathbf{p}_{(1,m)}^{T}$

The segmentation of $\mathbf{E}$, $\mathbf{X}$ and $\mathbf{p}$ gives:

$$\begin{aligned}&\left[\mathbf{E}_{1(n,m1)}|\mathbf{E}_{2(n,m2)}|\ldots|\mathbf{E}_{q(n,mq)}\right]\\&=\left[\mathbf{X}_{1(n,m1)}|\mathbf{X}_{2(n,m2)}|\ldots|\mathbf{X}_{q(n,mq)}\right]-\mathbf{t}_{(n,1)}\left[\mathbf{p}_{1(1,m1)}^{T}|\mathbf{p}_{2(1,m2)}^{T}|\ldots|\mathbf{p}_{q(1,mq)}^{T}\right]\end{aligned}$$

Rearranging:

$$\begin{aligned}&\left[\mathbf{E}_{1(n,m1)}|\mathbf{E}_{2(n,m2)}|\ldots|\mathbf{E}_{q(n,mq)}\right]\\&=\left[\mathbf{X}_{1(n,m1)}|\mathbf{X}_{2(n,m2)}|\ldots|\mathbf{X}_{q(n,mq)}\right]\\&\quad-\left[\mathbf{t}_{(n,1)}\mathbf{p}_{1(1,m1)}^{T}|\mathbf{t}_{(n,1)}\mathbf{p}_{2(1,m2)}^{T}|\ldots|\mathbf{t}_{(n,1)}\mathbf{p}_{q(1,mq)}^{T}\right]\end{aligned}$$

Decomposing $\mathbf{X}_i$, where $i=\{1,\ldots,q\}$:

$$\begin{aligned}&\left[\mathbf{E}_{1(n,m1)}|\mathbf{E}_{2(n,m2)}|\ldots|\mathbf{E}_{q(n,mq)}\right]\\&=\left[\mathbf{T}_{X1(n,h1)}\mathbf{P}_{X1(h1,m1)}^{T}|\mathbf{T}_{X2(n,h2)}\mathbf{P}_{X2(h2,m2)}^{T}|\ldots|\mathbf{T}_{Xq(n,hq)}\mathbf{P}_{Xq(hq,mq)}^{T}\right]\\&\quad-\left[\mathbf{t}_{(n,1)}\mathbf{p}_{1}^{T})|\mathbf{t}_{(n,1)}\mathbf{p}_{2(1,m2)}^{T}|\ldots|\mathbf{t}_{(n,1)}\mathbf{p}_{q(1,mq)}^{T}\right]\end{aligned}$$

which is the same as:

$$\mathbf{E}_{i(n,mi)}=\mathbf{T}_{Xi(n,hi)}\mathbf{P}_{Xi(hi,mi)}^{T}-\mathbf{t}_{(n,1)}\mathbf{p}_{i(1,mi)}^{T}$$

Post-multiplying by $\mathbf{P}_{Xi}$ and since $\mathbf{P}_{Xq}^{T}\mathbf{P}_{Xq}=\mathbf{I}$ *then:*

$$\mathbf{E}_{i(n,mi)}\mathbf{P}_{Xi(mi,hi)}=\mathbf{T}_{Xi(n,hi)}-\mathbf{t}_{(n,1)}\mathbf{p}_{i(1,mi)}^{T}\mathbf{P}_{Xi(mi,hi)}$$

Since $\mathbf{p}_{i}^{T}\,\mathbf{P}_{Xi}=\mathbf{p}_{PCTi}^{T}$ and using the notation $\mathbf{E}_i\,\mathbf{P}_i=\mathbf{E}_{\mathrm{PCT}i}$, where $i=\{1,\ldots,q\}$ then:

$$\mathbf{E}_{\mathrm{PCT}i(n,hi)}=\mathbf{T}_{Xi(n,hi)}-\mathbf{t}_{(n,1)}\mathbf{p}_{PCTi(1,hi)}^{T}$$

which shows that the segmented matrix updating can also be done in the PC-space.

The updating for the $\mathbf{y}$ vector is not modified as explained previously.

Appendix B

For prediction purposes, one needs to calculate the $\mathbf{b}$ vector. This vector is expressed in the original space by:

$$\mathbf{b}_{(m,1)}=\mathbf{W}_{(m,k)}\left[\mathbf{P}_{(k,m)}^{T}\mathbf{W}_{(m,k)}\right]^{-1}\mathbf{c}_{(k,1)}^{T}\tag{B.1}$$

for k Latent Variables (LVs).

From Eq. (A.1) one knows that for a segment i among $\{1,\ldots,q\}$ segments:

$$\mathbf{w}_{i(1,mi)}^{T}=\mathbf{u}_{(1,n)}^{T}\mathbf{T}_{Xi(n,hi)}\mathbf{P}_{Xi(hi,mi)}^{T}$$

and since $\mathbf{w}_{\mathrm{PCT}i}{}^{T}=\mathbf{u}_{(1,n)}{}^{T}\,\mathbf{T}_{\mathrm{Xi}(n,hi)}$ then:

$$\mathbf{w}_{i(1,mi)}^{T}=\mathbf{w}_{\mathrm{PCT}i(1,hi)}^{T}\mathbf{P}_{Xi(hi,mi)}^{T}$$

Therefore for k LVs and for a segment $i=\{1,\ldots,\mathrm{q}\}$ segments, one has:

$$\mathbf{W}_{i(k,mi)}^{T}=\mathbf{W}_{PCTi(k,hi)}^{T}\mathbf{P}_{Xi(hi,mi)}^{T}\tag{B.2}$$

In the same way and from Eq. (A.6):

$$\mathbf{p}_{i(1,mi)}^{T}=\mathbf{t}_{(1,n)}^{T}\mathbf{T}_{Xi(n,hi)}\mathbf{P}_{Xi(hi,mi)}^{T}\text{where }i=\{1,\ldots,q\}$$

and since $\mathbf{p}_{\mathrm{PCT}i}{}^{T}=\mathbf{t}_{(1,n)}{}^{T}\,\mathbf{T}_{Xi(n,hi)}$ then:

$$\mathbf{p}_{i(1,mi)}^{T}=\mathbf{p}_{PCTi}^{T}\mathbf{P}_{Xi(hi,mi)}^{T}$$

Again and for k LVs and for a segment $i=\{1,\ldots,q\}$ segments, one has:

$$\mathbf{P}_{i(k,mi)}^{T}=\mathbf{P}_{PCTi(k,hi)}^{T}\mathbf{P}_{Xi(hi,mi)}^{T}\tag{B.3}$$

Therefore Eq. (B.1) can be seen as a segmented $\mathbf{b}$ vector:

$$\mathbf{b}_{i(mi,1)}=\mathbf{W}_{i(mi,k)}\left[\mathbf{P}_{i(k,mi)}^{T}\mathbf{W}_{i(mi,k)}\right]^{-1}\mathbf{c}_{(k,1)}^{T},\text{where }i=\{1,\ldots,q\}\tag{B.4}$$

Using Eqs. (B.2) and (B.3), Eq. (B.4) can be expanded to:

$$\begin{aligned}\mathbf{b}_{i(mi,1)}&=\mathbf{P}_{Xi(mi,hi)}\mathbf{W}_{PCTi(hi,k)}\\&\quad\times\left[\mathbf{P}_{PCTi(k,hi)}^{T}\mathbf{P}_{Xi(hi,mi)}^{T}\mathbf{P}_{Xi(mi,hi)}\mathbf{W}_{PCTi(hi,k)}\right]^{-1}\mathbf{c}_{(k,1)}^{T}\end{aligned}$$

Since $\mathbf{P}_{\mathrm{X}i}{}^{T}\,\mathbf{P}_{\mathrm{X}i}=\mathbf{I}$ then:

$$\mathbf{b}_{i(mi,1)}=\mathbf{P}_{Xi(mi,hi)}\mathbf{W}_{PCTi(hi,k)}\left[\mathbf{P}_{PCTi(k,hi)}^{T}\mathbf{W}_{PCTi(hi,k)}\right]^{-1}\mathbf{c}_{(k,1)}^{T}$$

Pre-multiplying by $\mathbf{P}_{\mathrm{X}i}{}^{T}$ and as $\mathbf{P}_{\mathrm{X}i}{}^{T}\,\mathbf{P}_{\mathrm{X}i}=\mathbf{I}$ then:

$$\mathbf{P}_{Xi(hi,mi)}^{T}\mathbf{b}_{i(mi,1)}=\mathbf{W}_{\mathrm{PCT}i(hi,k)}\left[\mathbf{P}_{\mathrm{PCT}i(k,hi)}^{T}\mathbf{W}_{\mathrm{PCT}i(hi,k)}\right]^{-1}\mathbf{c}_{(k,1)}^{T}$$

Using the notation that $\mathbf{P}_{\mathrm{X}i}{}^{T}\,\mathbf{b}_i=\mathbf{b}_{\mathrm{PCT}i}$ one obtains:

$$\mathbf{b}_{PCTi(hi,1)}=\mathbf{W}_{PCTi(hi,k)}\left[\mathbf{P}_{PCTi(k,hi)}^{T}\mathbf{W}_{PCTi(hi,k)}\right]^{-1}\mathbf{c}_{(k,1)}^{T}$$

which can be concatenated by rows to give:

$$\begin{aligned}&\mathbf{b}_{\mathrm{PCT}(h1+h2+\ldots+hq,1)}\\&\quad=\left[\mathbf{b}_{\mathrm{PCT}1(h1,1)}|\mathbf{b}_{\mathrm{PCT}2(h2,1)}|\ldots|\mathbf{b}_{\mathrm{PCT}q(hq,1)}\right]\end{aligned}\tag{B.5}$$

Finally and only if needed, the $\mathbf{b}$ vector in the original space can be recovered by:

$$\mathbf{b}_{i(mi,1)}=\mathbf{P}_{Xi(mi,hi)}\mathbf{b}_{\mathrm{PCT}i(hi,1)},\text{ for segment }i=\{1,\ldots,q\}$$

References

[1] I.S. Helland, Chemom. Intell. Lab. Syst. 58 (2001) 97–107.
[2] H. Martens, Chemom. Intell. Lab. Syst. 58 (2001) 85–95.
[3] S. Wold, M. Sjöström, L. Eriksson, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.
[4] S. Wold, H. Martens, H. Wold, in: A. Ruhe, B. Kägström (Eds.), Lecture Notes in Mathematics 1983. Proceedings of the Conference Matrix Pencils, March 1982, Springer-Verlag, Heidelberg, 1983, pp. 286–293.
[5] L. Eriksson, H. Antti, J. Gottfries, E. Holmes, E. Johansson, F. Lindgren, I. Long, T. Lundstedt, J. Trygg, S. Wold, Anal. Bioanal. Chem. 380 (2004) 419–429.
[6] S. Wold, A. Berglund, N. Kettaneh, J. Chemom. 16 (2002) 377–386.
[7] D.W. Osten, J. Chemom. 2 (1988) 39–48.
[8] S. Lanteri, Chemom. Intell. Lab. Syst. 15 (1992) 159–169.

[9] Q.-S. Xu, Y.-Z. Liang, Chemom. Intell. Lab. Syst. 56 (2001) 1–11.
[10] R. Wehrens, H. Putter, L.M.C. Buydens, Chemom. Intell. Lab. Syst. 54 (2000) 35–52.
[11] S. Rännar, F. Lindgren, P. Geladi, S. Wold, J. Chemom. 8 (1994) 111–125.
[12] B.S. Dayal, J.F. MacGregor, J. Chemom. 11 (1997) 73–85.
[13] W. Wu, R. Manne, Chemom. Intell. Lab. Syst. 51 (2000) 145–161.
[14] F. Vogt, M. Tacke, Chemom. Intell. Lab. Syst. 59 (2001) 1–18.
[15] J. Trygg, S. Wold, Chemom. Intell. Lab. Syst. 42 (1998) 209–220.
[16] P. Teppola, P. Minkkinen, J. Chemom. 14 (2000) 383–399.
[17] W. Wu, D.L. Massart, S. De Jong, Chemom. Intell. Lab. Syst. 36 (1997) 165–172.
[18] A.S. Barros, D.N. Rutledge, Chemom. Intell. Lab. Syst. 73 (2004) 245–255.
[19] A.S. Barros, D.N. Rutledge, Chemom. Intell. Lab. Syst. 78 (2005) 125–137.
[20] G.E. Forsythe, A. Michael, C.B. Moler, Computer Methods for Mathematical Computations, Prentice-Hall, Englewood Cliffs, NJ, 1977.
[21] W. Wu, D.L. Massart, S. De Jong, Chemom. Intell. Lab. Syst. 37 (1997) 271–280.
[22] J.H. Kalivas, Chemom. Intell. Lab. Syst. 37 (1997) 225–259.
[23] M.H. Lopes, A.S. Barros, C. Pascoal Neto, D.N. Rutledge, I. Delgadillo, A.M. Gil, Biopolymers (Biospectroscopy) 62 (2001) 268–277.
[24] S. De Jong, Chemom. Intell. Lab. Syst. 18 (1993) 251–263.
[25] J. Arunachalam, S. Gangadharan, Anal. Chim. Acta 157 (1984) 245–260.
[26] B.K. Alsberg, D.B. Kell, R. Goodacre, Anal. Chem. 70 (1998) 4126–4133.
[27] M. Barker, W. Rayens, J. Chemom. 17 (2003) 166–173.
[28] A.S. Barros, D.N. Rutledge, Chemom. Intell. Lab. Syst. 70 (2004) 99–112.
[29] L. Norgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Appl. Spectrosc. 54 (2000) 413–418.
[30] C.H. Spielgelman, M.J. McShane, M.J. Goetz, M. Motamedi, Q.L. Yue, G.L. Coté, Anal. Chem. 70 (1998) 35–44.

ANNEXE II

ANALYTICA CHIMICA ACTA 603 (2007) 147–154

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/aca

Application of the ANOVA-PCA method to stability studies of reference materials

Julien Sarembaud[a,1], *Rui Pinto*[b], *Douglas N. Rutledge*[b], *Max Feinberg*[c,*]

[a] *Bureau Interprofessionnel d'étude Analytique (Bipea), 6-14 Avenue Louis Roche, 92230 Gennevilliers, France*
[b] *Laboratoire de Chimie Analytique, AgroParisTech, 16 rue Claude Bernard, 75005 Paris, France*
[c] *Institut National de la Recherche Agronomique (INRA), 16 rue Claude Bernard, 75231 Paris, Cedex 05, France*

ARTICLE INFO

Article history:
Received 8 June 2007
Received in revised form
10 September 2007
Accepted 17 September 2007
Published on line 29 September 2007

Keywords:
Near infrared spectroscopy
Stability study
External reference materials
ANOVA-PCA

ABSTRACT

Near infrared spectroscopy (NIRS) is an analytical technique that can be very useful for stability studies in particular because of its non destructive analytical capability. However, the spectral interpretation and treatment of this kind of multivariate data remains difficult without the use of chemometrics. In this article, a recent chemometrics method, analysis of variance – principal component analysis (ANOVA-PCA), was used for NIRS stability studies of sunflower and bread wheat external reference materials (ERM). It provided a practical tool for the study of the significance of various storage conditions according to an experimental design. Thus, the effect of the temperature, the nature of the atmosphere in the packaging and the storage duration were tested. ANOVA-PCA highlighted the influence of temperature and storage duration on the stability of the sunflower materials. For the bread wheat materials, the storage conditions did not have a significant effect on stability. Consequently, by applying ANOVA-PCA to near infrared spectral data, the sunflower materials were found to be considered stable for the time length of the study, i.e. 18 months stored in a cold room, while the bread wheat materials were found to be considered stable for the time length of the study, i.e. 12 months under the same conditions.

1. Introduction

Near infrared spectroscopy (NIRS) has gained widespread acceptance in recent years in many different fields of analytical chemistry, including the quality control of food and drugs [1–3]. NIRS analysis is considered as a rapid, non-destructive and non-pollutant technique compared to classical wet analysis techniques [4]. Moreover, it allows to analyze several constituents of the product at the same time [5,6].

Stability studies aim to investigate the changes that may occur during storage of a given product or material. It allows to define the effects of degradation factors on the material stability and to assess shelf-life. In this context, the term factor must be understand as it is used in experimental design methodology, i.e. a controlled independent variable whose levels are set by the experimenter. The materials used for our studies are reference material called external reference material (ERM) which are used by laboratories to control method trueness, as explained by the French standards association (AFNOR) in the AFNOR V03-115 Guide [7]. They are agricultural ERM and the degradation of such biological products is known to be strongly influenced by environmental conditions. Thus, King and Bolin demonstrated that the microbial growth depends on intrinsic properties such as water activity, pH

* *Corresponding author*. Tel.: +33 1 44 08 16 52; fax: +33 1 44 08 72 76.
E-mail addresses: jsarembaud@bipea.org (J. Sarembaud), rui.pinto@agroparistech.fr (R. Pinto), rutledge@agroparistech.fr (D.N. Rutledge), feinberg@inapg.inra.fr (M. Feinberg).
[1] Tel.: +33 1 47 33 91 66; fax: +33 1 40 86 92 59.
0003-2670/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2007.09.046

and oxygen [8]. Rodriguez et al. highlighted the best storage conditions for wheat flour to be with a small percent of oxygen compared to the normal atmosphere [9]. Microbial activity is considered as the main source of degradation as it can modify the material's global composition. As these ERM are biological material, the material stability may be reduced by the action of specific influences such as temperature or light.

NIRS may in many cases be used for these stability studies notably by its non-destructive properties and the fact that the spectrum represents a fingerprint of the global composition of the product [10,11]. However, the exploitation of the spectra requires the use of chemometrics.

Nearly all stability studies using NIR spectroscopy are based on the quantitative monitoring of particular analytes using predictive models established by multivariate calibration methods, such as PLS, to relate the spectra to the analyte concentration. In the method presented here, there is no calibration step. The evolution of the predicted analyte concentrations is not used. The changes in the spectra themselves are used directly to evaluate, qualitatively, the stability of the samples. The spectrum is considered as a fingerprint of the global detectable composition of the product. In this context it was assumed that:

- if the signal changes it means that the materials have changed;
- if the signal does not change, two things may arise:
  - the materials did not change;
  - the materials have changed but the method can not detect these changes.

Since NIR spectroscopy has intrinsic detection limitations, such as the detection of traces, it is possible that some minor constituents of the material may change without this being reflected in the spectrum. This problem would arise whether the predicted analyte concentrations were used or the whole spectrum. This drawback is a real limitation of the method but it must be remembered the importance of water in the stability of biological materials and the global role of moisture on each component.

The recently developed analysis of variance – principal component analysis (ANOVA-PCA) method was first used for proteomic applications [12,13]. This method has no link with the widely used procedure for variable selection, which consists in using an analysis of variance (ANOVA) to detect variables which do not vary significantly as a function of the factors being studied, in order to eliminate them before proceeding with a multivariate analysis of the resulting reduced dataset using PCA. In a way somewhat similar to ANOVA, the ANOVA-PCA method looks for the variations in the variables associated with each factor of an experimental design and then uses a principal component analysis (PCA) to assess the significance of each factor by comparing the corresponding variations to the residual error.

In this paper, we used the ANOVA-PCA method to distinguish the effect of storage conditions used in the stability studies from the effect of storage time.

Table 1 – List of materials used and their storage conditions

Type of ERM	Experimental factors tested	Duration of the stability study in months (factor *t*)
Bread wheat kernels	Factor *T* – temperature Factor *A* – atmosphere Factor *D* – duration Factor *S* – samples	6,7,8,9,10,11,12
Sunflower kernels	Factor *T* – temperature Factor *A* – atmosphere Factor *D* – duration Factor *S* – samples	1,9,12,18

2. Experimental

2.1. *Reference materials used for the stability studies*

The agricultural external reference materials used for these stability studies were prepared by the Bureau Interprofessionnel d'Études Analytiques (Bipea) within the framework of its traditional activity of proficiency testing schemes (PTS) organizer. Homogeneity checking was carried out on these ERMs and the stability is assumed to be the same for all ERM of identical material resulting from the same manufacture and storage. Table 1 shows the different types of ERMs. According to previous experiments on the ERM stability described in [14], the sunflower ERMs are known to evolve rapidly and thus studies began just one month after their manufacture. On the other hand, the bread wheat ERMs are known to be more stable and studies began 6 months after their manufacture.

2.2. *Experimental design*

Four experimental factors were evaluated on sunflower and bread wheat ERMs according to an experimental design for stability studies.

Two environmental factors, temperature (*T*) and composition of packaging atmosphere (*A*), were tested.

For factor *T*, two storage conditions were applied:

(1) In cold room at +4 °C, noted CR.
(2) At room temperature (+20 ± 5 °C), noted RT.

For factor *A*, two levels were assigned for checking the influence of the atmosphere:

(1) Normal atmospheric air pressure, noted NOR.
(2) Under vacuum (0.2 bar), noted VAC.

The structure of the packaging medium used for these solid unground ERMs is a multi-layer Bernhard® packaging, thermally-sealed at the end of the ERM preparation process.

A third factor, coded factor *D*, is the duration of storage of the materials.

The fourth factor, *S*, is related to the reproducibility of the experiment. Several samples were taken for each storage condition which allowed us to obtain more robust averages.

Table 1 summarises all the storage conditions for the different ERMs used.

2.3. Method of analysis and chemometrics

Materials were analysed by NIRS using a Bruker® Vector 22N/C spectrophotometer over a wavelength range of 10,500–4524 cm^{-1} and a resolution set at 8 cm^{-1}. The spectra were obtained using diffuse reflection on 200 g of sample placed in the cell cup for the analysis. Each spectrum corresponds to the average of 32 scans.

When ERMs were stored in cold room, stabilization at room temperature for a minimal duration of 12 h is necessary before analysis by NIRS according to the procedure described in [11] and which is routinely used in the laboratory for checking material homogeneity. The materials were divided with a rifle divisor into samples with masses around 200 g.

At each storage period, 1–3 samples were analyzed by NIRS and measurements were done with i repetitions ($1 < i < 6$) under repeatability conditions (same operator, same spectrophotometer, same quartz cup and same delay between analyses). Table 1 summarises the different characteristics of the ERMs used for the stability studies.

2.4. ANOVA-PCA

The objective of the ANOVA-PCA method is to compare the variability of the data due to the levels of each factor of the experimental design to the variability of the residual error. The procedure consists of two parts:

(1) A sequential decomposition of the initial data matrix into a set of matrices based on the different factors of the experimental design.
(2) A principal components analysis of each factor matrix to which the residual error matrix has been added back.

Initially, the matrix is centred by subtracting the average spectrum. As shown in Fig. 1, this column-centred matrix is then used to calculate a new matrix of the same size containing, for each sample and each variable, the average for each level of the first factor (Fig. 2). This first factor matrix is then subtracted from the centred matrix to give a first residuals matrix which is then treated in the same way to calculate a matrix containing the averages for each level of the second factor. This second factor matrix is subtracted from the first residuals matrix to give a second residuals matrix. The procedure is repeated for the remaining factor matrices until the final residual error matrix is obtained. The matrices corresponding to the interactions between factors can also be calculated. The residual error matrix is then added back to each of the factor matrices.

It can be seen that this decomposition of the original matrix into a series of matrices based on the levels of the factors is in some ways comparable to an ANOVA.

The PCA part of the procedure consists in subjecting each of the new factor plus residual error matrices to PCA. The basic hypothesis of the ANOVA-PCA method is that if an experimental factor is a dominant source of variation compared to the residual error, then the first principal component (PC1) will

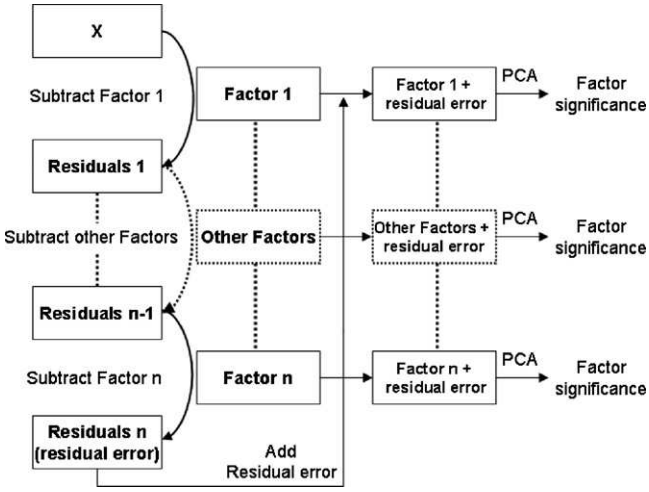


Fig. 1 – A schematic representation of the different steps of the ANOVA-PCA method. *X* is the initial spectral data matrix already centred by columns. The number of "Other Factors" depends on the experimental design and for each of these factors there is a corresponding factor matrix.

mainly characterize this variation and the second principal component (PC2) will mainly reflect random variations.

The Hotelling T^2 distribution [15] may be used in order to generate the 95% confidence intervals around the clusters in the principal component scores plot corresponding to the samples at each level for the factor.

In practice, the hypothesis that PC1 will contain the information and PC2 the noise is seen to be an over-simplification. The study of the principal component scores and loadings reveals several situations:

(1) The experimental design factor tested is in fact the dominant source of variation compared with the residual error, in which case PC1 will characterize principally the variation due to this factor. The scores plot will separate the samples along PC1 into clusters that correspond to the factor levels if the levels are significantly different. The residual error is then mainly to be found on PC2 and, depending on the data dispersion, the separation of the scores along this axis will be more or less important.
(2) There is no separation of the predefined clusters, neither along PC1 nor PC2. Hence, the experimental factor appears

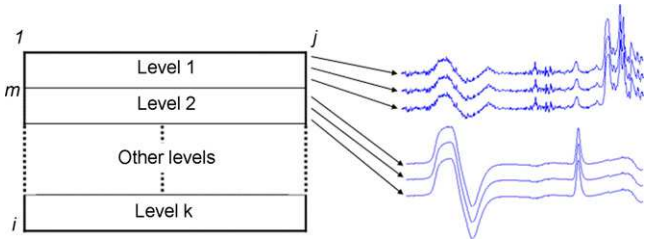


Fig. 2 – Example of factor matrix. Each of the *k* levels of the factor is composed of *m* lines, which are all equal to the respective level mean.

not to be a dominant source of variation compared to the residual error. In this case, the principal source of variation may be either an uncontrolled factor or random error.

(3) There is no separation of the predefined cluster along the PC1 axis, but there is along PC2. In this case, the experimental factor is not the dominant source of variation compared to the residual error but it does have an effect, albeit small compared to the noise.

(4) The scores of the samples along PC1 and/or PC2 are not aligned along the corresponding axis in a way that is linearly related to the factor levels, meaning that another source of variation may exist such as an interaction between two factors.

2.5. *Standard normal variate (SNV)*

All spectra were normalized by standard normal variate (SNV) pre-treatment before the application of the ANOVA-PCA method. The SNV pre-treatment is widely applied to NIR spectra to reduce baseline shifts and variations of global intensity. The mean of the signal intensities is subtracted from each point in the signal and the result is divided by the standard deviation of the complete signal [16,17].

2.6. *Software*

The software used for these stability studies are Opus® version 5.5 (Bruker) and MATLAB version 7 (Mathworks).

3. Results and discussion

The experimental designs for the sunflower and bread wheat ERMs were constructed to distinguish the effects of temperature, the atmosphere in the packaging and the storage duration. The initial data matrix was used to generate five new matrices representing the experimental design factors Temperature, Atmosphere, Duration, Samples and the residual error.

3.1. *Sunflower ERMs*

Fig. 3 presents the spectra for the sunflower materials. First of all a four-factor model with no interaction is established in vector form in order to divide the initial data matrix into a new set of matrices. Therefore, this model is given by the following Eq. (1):

$$\mathbf{X}_{(i,j)} = \bar{\mathbf{X}}_{(i,j)} + \bar{\mathbf{T}}_{(i,j)} + \bar{\mathbf{A}}_{(i,j)} + \bar{\mathbf{D}}_{(i,j)} + \bar{\mathbf{S}}_{(i,j)} + \boldsymbol{\varepsilon}_{(i,j)} \quad (1)$$

where i is the total number of spectra obtained and j the spectral wavelengths. $\mathbf{X}$ corresponds to the initial data matrix, $\bar{\mathbf{X}}$ is a matrix whose lines are all equal to the global average of $\mathbf{X}$. $\bar{\mathbf{T}}$, $\bar{\mathbf{A}}$, $\bar{\mathbf{D}}$ and $\bar{\mathbf{S}}$ represent the factor matrices according to the experimental design. In these factor matrices, each line contains the average of its corresponding level, as depicted in Fig. 2. They are respectively related to the experimental design temperature ($\bar{\mathbf{T}}$), composition of the atmosphere ($\bar{\mathbf{A}}$), storage duration ($\bar{\mathbf{D}}$) and sample (or replicates group, $\bar{\mathbf{S}}$). Finally, $\boldsymbol{\varepsilon}$ corresponds to the residual error matrix.

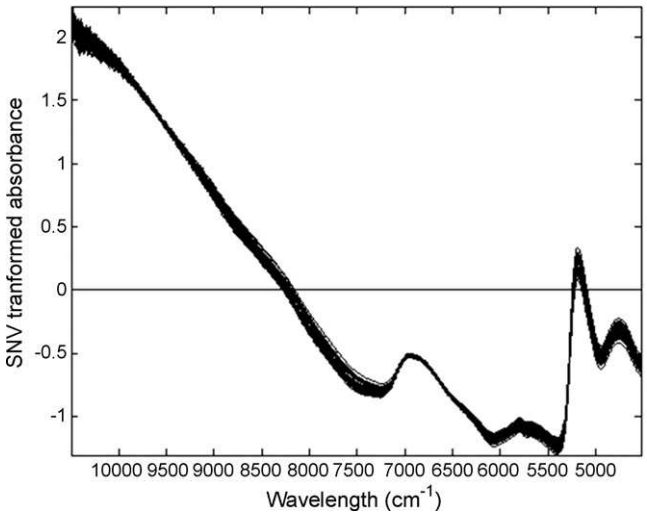


Fig. 3 – Initial data for the sunflower materials. 64 NIR spectra acquired according to the experimental design. Spectral region 10,500–4524 cm^{-1}.

The residual error matrix is added back to each factor matrix and the PCA is then applied to each of these factor + residual error matrices.

Then, the scores plot for the PCA for each factor + residual error matrix is studied. The scores plot for the PCA on the matrix Temperature + residual error is given in Fig. 4. PC1 divides the spectral points into separate clusters corresponding to the pre-defined factor levels, which indicates the existence of significant changes due to the effect of the temperature. PC2 shows that the variation within each temperature level is related to the residuals matrix as explained earlier. This PCA result corresponds to the first case explained above where the experimental factor is the main source of variation compared to the residual error. Moreover, a PCA on the residuals matrix (Fig. 5) shows that its PC1 loadings are quite similar to the PC2 loadings of the matrix Temperature + residual error. Consequently, PC1 represents the variation due to the factor *T* which is considered as the main source of variation and PC2 represents mostly the variation due to the residual error.

These results show that the conservation of the sunflower ERMs depends on the storage temperature. Nevertheless, it was not possible at this point to conclude which storage temperature is best for the sunflower ERMs. However, it would be expected that lower temperatures are to be preferred for conservation.

Fig. 6 shows the PCA results for the factor matrix Atmosphere in the packaging + residual error. The scores clusters due to the two factor levels overlap, which indicates that this factor appears not to be a dominant source of variation compared to the residual error. Contrary to what was expected, the presence of oxygen which is usually thought to be one of the most active sources of degradation, may have a minor role in the evolution of this type of material.

As the storage temperature was shown to have an effect on the sunflower ERMs, the same calculation was done separately for each temperature. The results obtained (not shown

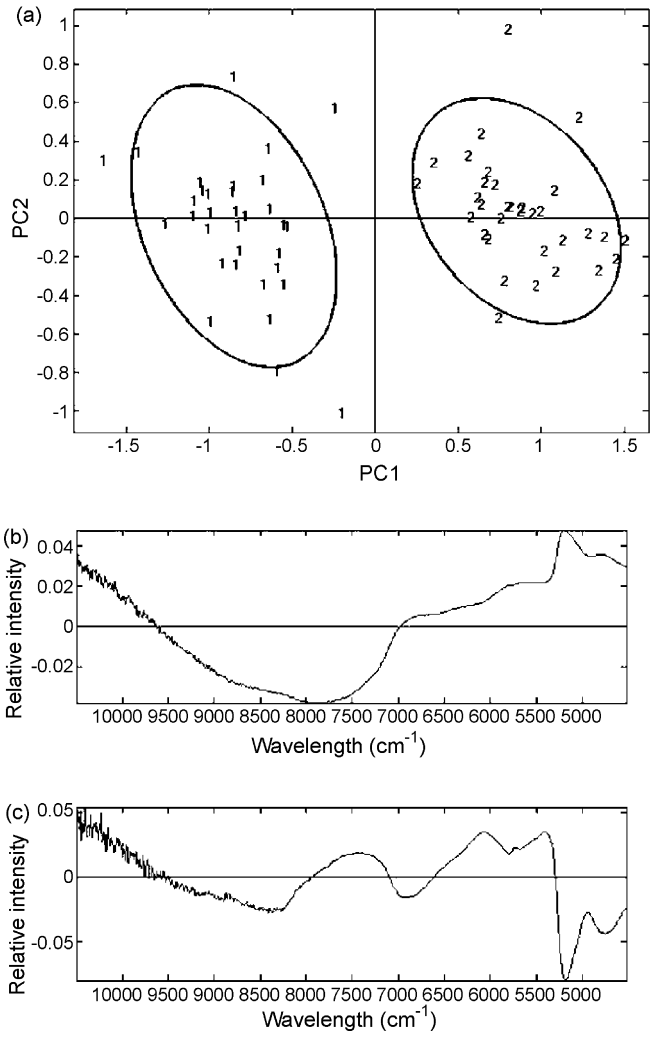


Fig. 4 – PCA scores plot (a), PC1 loadings (b) and PC2 loadings (c) for the matrix Temperature + residual error for the sunflower materials. Scores plot: the 95% confidence intervals are calculated for each level. PC1 loadings are related to the factor and PC2 loadings to the residual error. Cluster 1: ERM stored in room temperature; cluster 2: ERM stored in cold room.

here) were the same for each temperature and confirmed that the experimental factor *A* was not significant compared to the residual error.

Fig. 7a shows the scores plot of the PCA on the matrix Duration + residual error. Fig. 7b and c show the corresponding

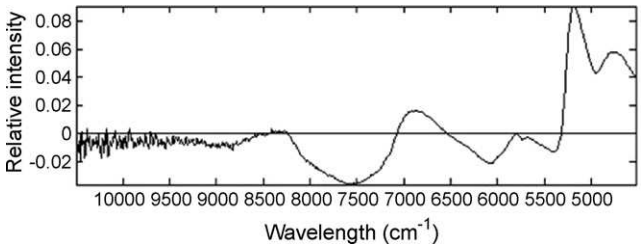


Fig. 5 – PC1 Loadings of the PCA on the residual error matrix for the sunflower materials.

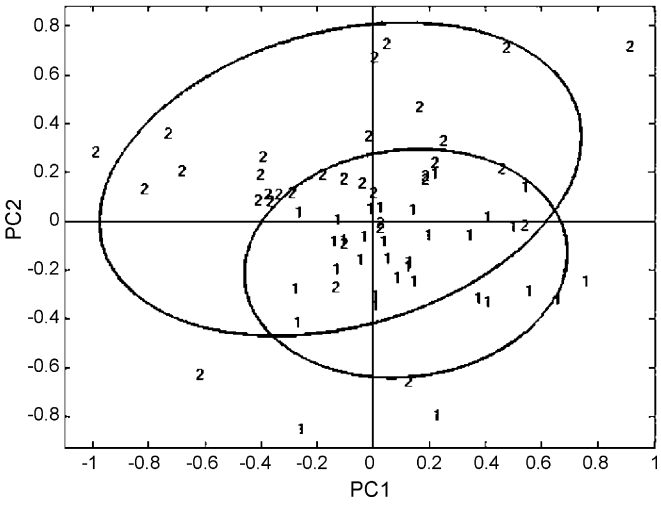


Fig. 6 – PCA scores plot for the matrix Atmosphere + residual error for the sunflower materials. The 95% confidence intervals are calculated for each level. Cluster 1: VAC; cluster 2: NOR.

loadings. Although there is not a complete separation between successive level clusters, it is possible to observe a trend along PC1 axis scores, which may characterize the factor *D*. Clusters 1 and 4 are, indeed, well distinguished along PC1. As previously for the Temperature analysis, the PC2 corresponds to the residual error variability. The PCA on the residuals matrix (Fig. 5) shows that its PC1 loadings are quite similar to these PC2 loadings. Therefore, PC1 represents the storage duration variation which is dominant compared to the residual error variation.

The scores plot of the PCA on the matrix Samples + residual error, showed that all the groups overlapped (results not shown). This means that the factor *S* is not significant.

According to these results, the sunflower ERMs may change with the factors *T* and *D*. As both these factors are found to be a dominant source of variation when individually compared to the residual error variability, the interaction between them may also be important. Therefore, a new analysis was performed in which a new factor *TD* (Temperature × Duration, with eight different levels) was defined by the combination of these two factors (four Duration levels for each of the two temperatures). This would enable to compare the evolution of materials as a function of storage duration for each Temperature level. The initial data matrix is now decomposed according to the following Eq. (2):

$$\tilde{X}_{(i,j)} = \bar{X}_{(i,j)} + \overline{TD}_{(i,j)} + \tilde{S}_{(i,j)} + \varepsilon_{(i,j)} \quad (2)$$

where *i* is the total number of spectra and *j* the number of spectral wavelengths. $\boldsymbol{X}$ corresponds to the initial data matrix, $\bar{\boldsymbol{X}}$ is a matrix which lines are all equal to the global average of $\boldsymbol{X}$. $\overline{TD}$ represents the matrix defined by this new factor Temperature × Duration, $\tilde{\boldsymbol{S}}$ the matrix due to the samples and $\boldsymbol{\varepsilon}$ the residual error matrix.

Fig. 8a shows the scores plot obtained for the PCA on the matrix Temperature × Duration + residual error, confirming the result obtained earlier while also giving other extra

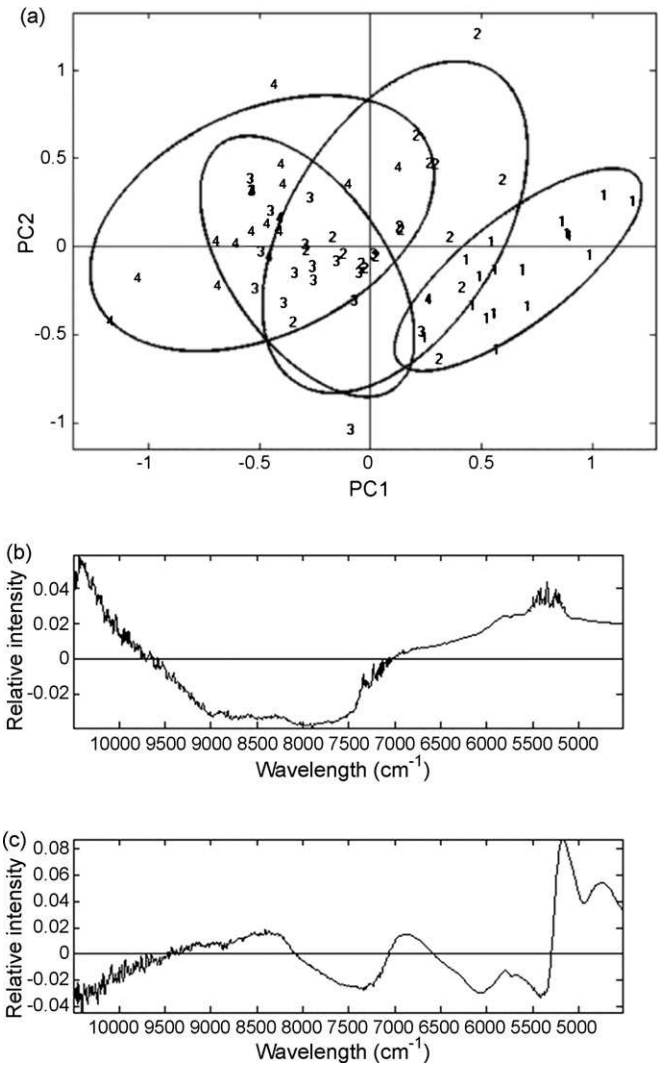


Fig. 7 – PCA scores plot (a), PC1 loadings (b) and PC2 loadings (c) for the matrix Duration × residual error for the sunflower materials. Scores plot: the 95% confidence intervals are calculated for each level. Cluster 1: t1; cluster 2: t2; cluster 3: t3; cluster 4: t4. PC1 loadings are related to factor Duration and PC2 loadings to the residual error.

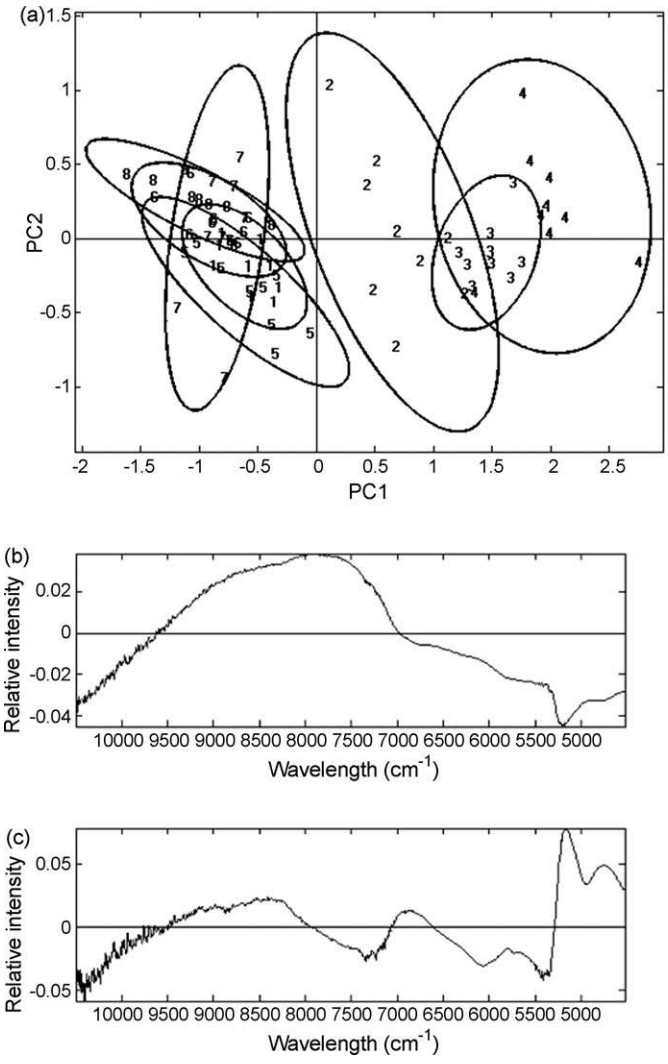


Fig. 8 – PCA scores plot (a), PC1 loadings (b) and PC2 loadings (c) for the matrix Temperature × Duration + residual error for the sunflower materials. Scores plot: the 95% confidence intervals are calculated for each level. Clusters 1–4: ERMs stored at room temperature; clusters 5–8: ERM stored in cold room. Clusters 1 and 5: t1; clusters 2 and 6: t2; clusters 3 and 7: t3; cluster 4 and 8: t4. PC1 loadings are related to factor Duration and PC2 loadings to the residual error.

information. Clusters 5–8 (cold room) were not separated along the PC1 axis contrary to clusters 1–4 (room temperature) which were clearly separated and show a trend. Thus, the sunflower ERMs stored at room temperature evolved significantly as a function of the storage duration. The other interesting information was that cluster 1 was equivalent to cluster 5 which means that changes due to storage at room temperature appeared between the first and the second time duration of the experiment. Fig. 8b and c are the corresponding loadings. As expected, the PC2 loadings represented in Fig. 8c are similar to the PC1 loadings of the residual matrix on Fig. 5.

The second tested factor S for this new decomposition of the initial data matrix gives the same results as those of the previous analysis, showing no separation between the levels (results not shown).

3.2. *Bread wheat ERMs*

Fig. 9 presents the spectra for the bread wheat materials.

The interpreting strategy used here is the same as for the sunflower ERMs. First, a model is constructed for the decomposition of the initial data matrix then the scores plots are investigated in order to determine if the experimental factors are significant or not. The decomposition model is the same as for the sunflower ERMs as given in Eq. (1).

For comparison purposes, residual error loadings were calculated and are presented on Fig. 10.

Sample degradation was expected to be relatively limited as a function of the temperature since wheat kernels are usu-

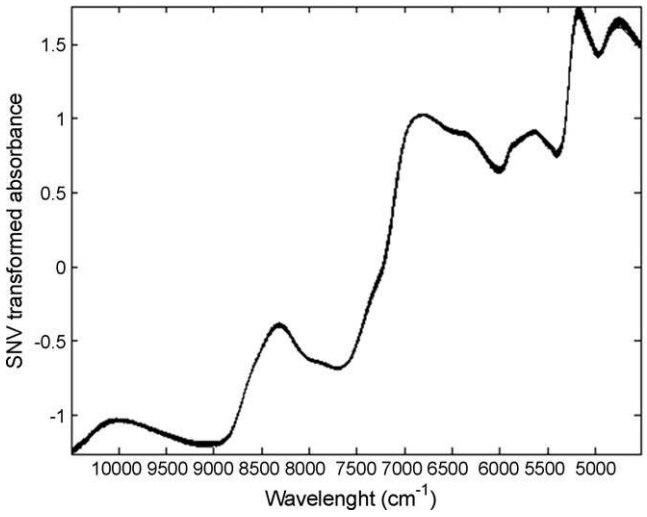


Fig. 9 – Initial data for the sunflower materials. 336 NIR spectra taken according to the experimental design. Spectral region 10,500–4524 cm^{-1}.

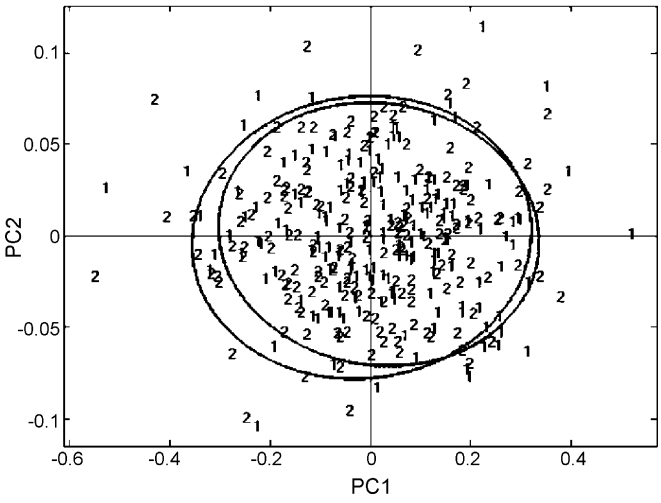


Fig. 11 – PCA scores plot for the matrix Temperature + residual error for the bread wheat materials. The 95% confidence intervals are calculated for each levels. Cluster 1: ERM stored in room temperature; cluster 2: ERM stored in cold room.

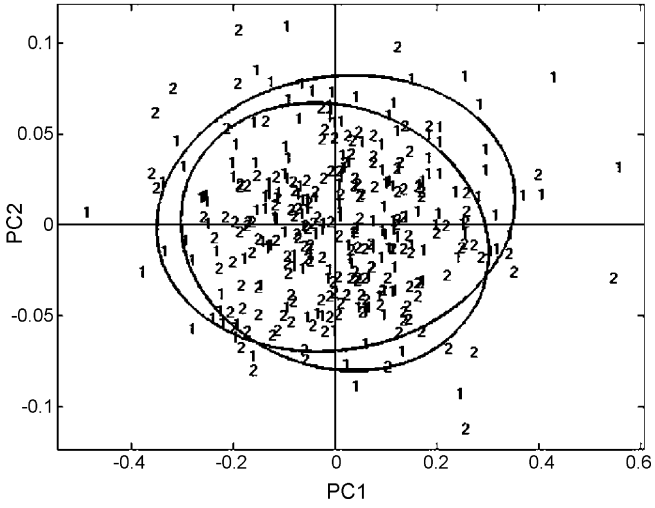


Fig. 12 – PCA scores plot for the matrix Atmosphere + residual error for the bread wheat materials. The 95% confidence intervals are calculated for each levels. Cluster 1: VAC; cluster 2: NOR.

ally stored in silos where the temperatures may be higher than those used in these experiments. The results of PCA on the matrix Temperature + residual error (Fig. 11) show that the experimental factor *T* was not the main source of variation compared to the residual error. One of the most expected changes for the bread wheat ERMs was water loss during the storage but this was avoided by sealing the packages. Apparently, proteins and saccharides such as starch were not influenced by this experimental factor.

The scores plot for the PCA on the matrix Atmosphere + residual error are given in Fig. 12. The two different storage levels are not distinguished along PC1. Consequently, the experimental factor *A* appears not to be a dominant source of variation compared with the residual error.

For this case, it seemed that this experimental factor *A* has no influence on the evolution of this material. Consequently, the oxidation reaction appears to have little effect. As the two temperatures of storage gave similar results, no calculations were done with each temperature condition.

Fig. 13a describes the scores plot of the PCA on the matrix Duration + residual error. Several observations can be made. First of all, the different clusters were not separated according to the PC1 axis but along the PC2 axis. Fig. 13b and c show that the PC1 loadings correspond to the PC1 loadings for the PCA on the residual error matrix (see Fig. 10). Therefore, the factor *D* was not considered as the main source of variation compared to the residual error. Nevertheless, a separation of the various clusters exists on PC2. This is an example of the third case presented in Section 2.4, as there is interesting variability associated with the factor (along PC2), but it is not as great as the random error variability (along PC1) and at the same time the PC2 scores plot does not show a simple time-related trend for the clusters. The clusters corresponding to the durations 1, 2 and 3 are superposed; whereas cluster 4 is well separated and cluster 5 even further away. A non-linear process may be involved as cluster 6 is between clusters 5 and 4, while cluster 7 is superposed on clusters 1–3.

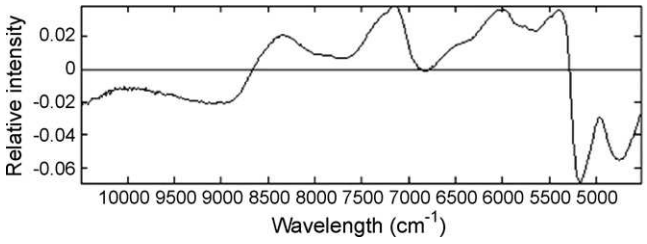


Fig. 10 – PC1 loadings of the PCA on the residual error matrix for the bread wheat materials.

The scores plot for the PCA on the matrix Samples + residual error shows that all groups overlapped, which

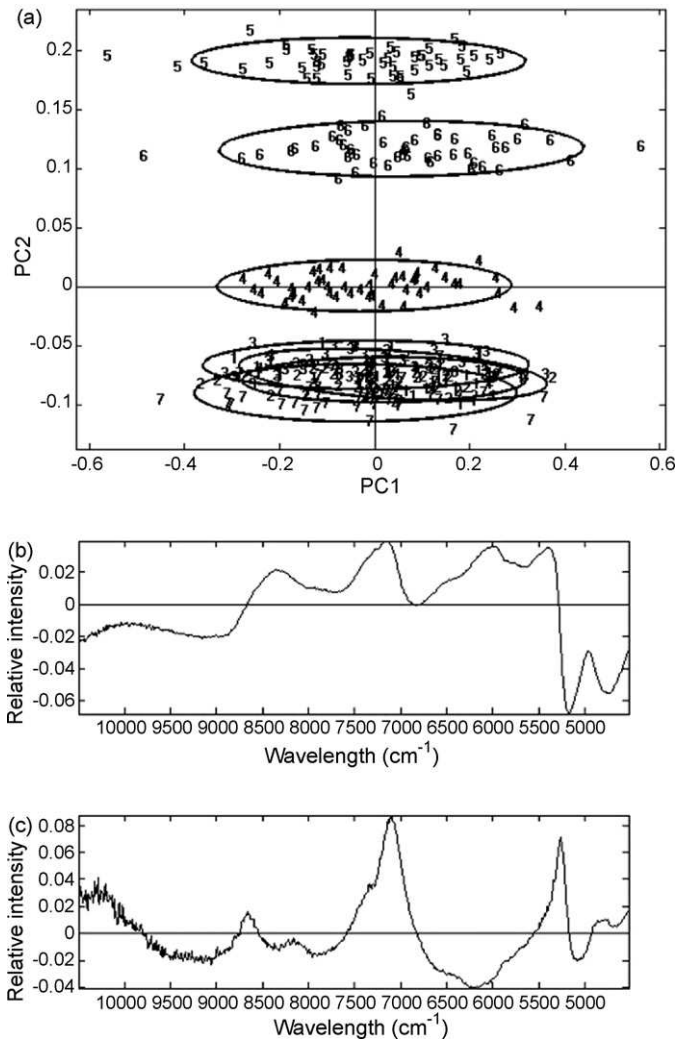


Fig. 13 – PCA scores plot (a), PC1 loadings (b) and PC2 loadings (c) for the matrix Duration + residual error for the bread wheat materials. Scores plot: the 95% confidence intervals are calculated for each levels. Cluster 1: t1; cluster 2: t2; cluster 3: t3; cluster 4: t4; cluster 5: t5; cluster 6: t6; cluster 7: t7. PC1 loadings are related to factor Duration and PC2 loadings to the residual error.

means that the experimental factor *S* is not to a dominant source of variation compared to the residual error (results not shown).

4. Conclusion

The aim of this paper was to use a practical tool – ANOVA-PCA – on NIR spectral data to distinguish the effect of storage conditions and duration for specific stability studies on external reference materials. Hence, this method provides information on the significance of the principal experimental factors such as temperature.

Contrary to what was expected, storing under vacuum for these particular products was found not to be significant which means that it does not improve the global stability of the materials.

The temperature had an influence only for the sunflower ERMs which shows the importance of the ERM composition. Consequently, the best storage conditions correspond to a storage in a cold room with or without packaging under vacuum. The factor Duration was not significant for the best conditions of storage (cold room) which indicates that, according to the ANOVA-PCA method used on NIR spectral data, the sunflower ERMs are considered stable for the duration length of the study, i.e. 18 months. For the bread wheat ERMs, they are also considered stable for the duration length of the study, i.e. 12 months. However, if no different evolutions are noticed for the bread wheat ERMs for the factor Temperature, this was not the case for the sunflower ERMs where the combination of the effects of Duration and Temperature lead to its modification.

REFERENCES

[1] M. Cocchi, C. Durante, G. Foca, A. Marchetti, L. Tassi, A. Ulrici, Talanta 68 (2006) 1505–1511.
[2] T.G. Axon, R. Brown, S.V. Hammond, S.J. Maris, J. Near Infrared Spectrosc. 6 (1998) 13–19.
[3] M. Blanco, I. Villarroya, Trends Anal. Chem. 21 (2002) 240–250.
[4] H. Büning-Pfaue, Food Chem. 82 (2003) 107–115.
[5] D. Cozzolino, M.J. Kwiatkowski, M. Parker, W.U. Cynkar, R.G. Dambergs, M. Gishen, M.J. Herderich, Anal. Chim. Acta 513 (2004) 73–80.
[6] A. Fassio, D. Cozzolino, Ind. Crops Prod. 20 (2004) 321–329.
[7] NF FD V03-115 (1996), Analyse des produits agricoles et alimentaires-Guide pour l'utilisation des matériaux de référence, AFNOR Paris la Défense.
[8] A.D. King, H.R. Bolin, Food Technol. 43 (1989) 132–139.
[9] M. Rodriguez, L.M. Medina, R. Jordano, Nahrung 44 (2000) 247–252.
[10] H. Cen, Y. He, Trends Food Sci. Technol. 18 (2007) 72–83.
[11] M.E. Lafargue, M.H. Feinberg, J.J. Daudin, D.N. Rutledge, J. Near Infrared Spectrosc. 11 (2002) 109–121.
[12] P. Harrington, N. Vieira, J. Espinoza, J. Nien, R. Romero, A. Yergey, Anal. Chim. Acta 544 (2005) 118–127.
[13] P. Harrington, N. Vieira, P. Chen, J. Espinoza, J. Nien, R. Romero, A. Yergey, Chemometrics Intell. Lab. Syst. 82 (2006) 283–293.
[14] J.A. Sarembaud, M. Feinberg, Accred. Qual. Assur. 12 (2007) 75–83.
[15] H.A. Hotelling, Generalized *T* test and measure of multivariate dispersion, Second Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, 1951, p. 23–41.
[16] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Appl. Spectrosc. 43 (1989) 772–777.
[17] A. Garrido-Varo, R. Carrette, V. Fernandez-Cabanas, J. Near Infrared Spectrosc. 6 (1998) 89–95.

ANNEXE III

ANALYTICA CHIMICA ACTA 629 (2008) 47–55

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/aca

Using ANOVA-PCA for discriminant analysis: Application to the study of mid-infrared spectra of carraghenan gels as a function of concentration and temperature

Rui Climaco Pinto[a,c], *Véronique Bosc*[b], *H. Noçairi*[c], *António S. Barros*[d], *Douglas N. Rutledge*[a,c,*]

[a] *Laboratoire de Chimie Analytique, AgroParisTech, 16 rue Claude Bernard, 75005 Paris, France*
[b] *UMR Scale 1211, AgroParisTech, 1 avenue des Olympiades, 91744 Massy, France*
[c] *UMR INRA/AgroParisTech 214 "IAQA", 16 rue Claude Bernard, 75005 Paris, France*
[d] *Departamento de Química, Universidade de Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal*

ARTICLE INFO

Article history:
Received 26 May 2008
Received in revised form
7 August 2008
Accepted 9 September 2008
Published on line 17 September 2008

Keywords:
Analysis of variance-principal component analysis (ANOVA-PCA)
Prediction
Discriminant analysis
Mid-infrared spectroscopy
Carraghenan

ABSTRACT

In this work the ANOVA-PCA method is applied to a MIR spectroscopy dataset of carraghenan in order to evaluate which of the factors within its fixed effects experimental design are significant in relation to the residual error. The factors defined in the experimental design are concentration (1% and 2%), temperature (30, 40, 45, 50, and 60 °C), day (1 and 2) and sample (20 samples, 3 repetitions). The two factors, concentration and temperature, were considered as significant and the main features related with its physico-chemical properties were identified. It is also of interest to acquire a better understanding of the interaction between concentration and temperature and its effect on the adhesion of gels onto the surface of contact. In fact, no significant interaction was found between the two factors, but it was shown that the factor temperature behaves in a non-linear way.

As classification using the ANOVA-PCA procedure has not been developed until now, a new method is proposed for the classification of new samples in respect to the levels of each significant factor.

1. Introduction

Analysis of variance-principal component analysis (ANOVA-PCA) [1] was first used in proteomics for the detection of biomarkers in high dimensional data sets [1,2] and more recently to assess the stability of reference materials [3]. This method uses the ANOVA paradigm to separate variations into main effects, interaction and noise followed by PCA [4] to evaluate the significance of each of these effects against the residual error. The PCA loadings can be interpreted to understand the origin of the dispersion of the samples.

Mid-infrared (MIR) spectroscopy has been widely used to study polysaccharides [5–7].

At different concentration and temperature values, the adhesion of gels varies. It is still not well understood why this happens and if there is an interaction between these two factors. The present experiment has been designed to study these effects. Carraghenan is a well characterised gel and so

* *Corresponding author at*: Laboratoire de Chimie Analytique, AgroParisTech, 16 rue Claude Bernard, 75005 Paris, France. Tel.: +33 1 44 08 16 48; fax: +33 1 44 08 16 53.
E-mail address: rutledge@agroparistech.fr (D.N. Rutledge).
0003-2670/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2008.09.024

was used as the model in this study. The temperature covers a range in which one always has a gel and not a solid, and where water evaporation is not a problem. The concentration range was set to low and well defined values compatible with maintaining a gel form at all the temperature values used.

2. Theory

2.1. ANOVA-PCA

The ANOVA-PCA method has no link with the widely used procedure for variable selection or elimination which consists in using an analysis of variance (ANOVA) to detect variables which do not vary significantly as a function of the factors being studied, in order to eliminate them before proceeding with a multivariate analysis of the resulting reduced dataset using PCA.

In fact, ANOVA-PCA (Fig. 1) successively calculates a series of matrices corresponding to the means of the variables at each level of each factor in an experimental design, and then subtracts them from the original matrix to get the matrix of residual error. In vector form, a two-factor model with interaction is given by

$$\mathbf{x}_i = \bar{\mathbf{x}} + \bar{\boldsymbol{\alpha}}_j + \bar{\boldsymbol{\beta}}_k + \overline{\boldsymbol{\alpha\beta}}_{jk} + \boldsymbol{\varepsilon}_i \quad (2)$$

In which $\mathbf{x}_i$ is a spectrum, $\bar{\mathbf{x}}$ is the average spectrum for the whole data set, $\bar{\boldsymbol{\alpha}}_j$ and $\bar{\boldsymbol{\beta}}_k$ are the effects for factors 1 and 2, respectively, $\overline{\boldsymbol{\alpha\beta}}_{jk}$ is the interaction between these two factors and $\boldsymbol{\varepsilon}_i$ are the residuals for the spectrum.

After centering the data (subtracting $\bar{\mathbf{x}}$ from $\mathbf{x}_i$), one obtains "Initial data" (Fig. 1). According to the schema, the averages of each of the levels for Factor 1 are then calculated. Matrix "Factor 1" has exactly the same number of lines as the initial matrix and each line is the average corresponding to the level to which it belongs for Factor 1. After subtracting this matrix from the matrix "Initial data", one obtains the "Residuals 1" matrix. The procedure continues by calculating the averages of the "Residuals 1" matrix for Factor 2 and subtracting it to give the "Residuals 2" matrix. Finally, the averages of the "Residuals 2" matrix for the Factor 1 × Factor 2 interaction are calculated and subtracted to give the final matrix of "residual error".

After adding the "residual error" matrix back to each of the "Factor" average matrices, there is a second, multivariate step, in which a principal component analysis (PCA) is used to assess the significance of each factor by comparing the factor variance to the residual error. The hypothesis of the ANOVA-PCA method as proposed by Harrington et al. [1] is that if an experimental factor is a dominant source of variation compared to the residual error, then the first principal component (PC1) will mainly characterise this variation and the second principal component (PC2) will mainly reflect random variations.

The Hotelling T2 distribution [8] may be used to generate the 95% confidence intervals around the clusters in the PC1–PC2 principal component scores plot corresponding to the samples at each level for the factor.

Other methods, such as ANOVA simultaneous component analysis (ASCA) [9–11], use some of the same principles as ANOVA-PCA. In the case of ASCA, the calculation of the averages of the levels for each of the factors is done in the same way as in ANOVA-PCA. The difference is:

- In ASCA the averages are calculated and then PCA is performed directly on them, without including the residuals. Therefore there is no way to directly estimate the significance of the factor, as there is no reference for comparison. It is necessary to use bootstrapping or similar resampling methods in order to estimate the significance of the grouping of samples for the factors.
- In ANOVA-PCA the calculated residuals are added back to each of the matrices of averages and then PCA is performed

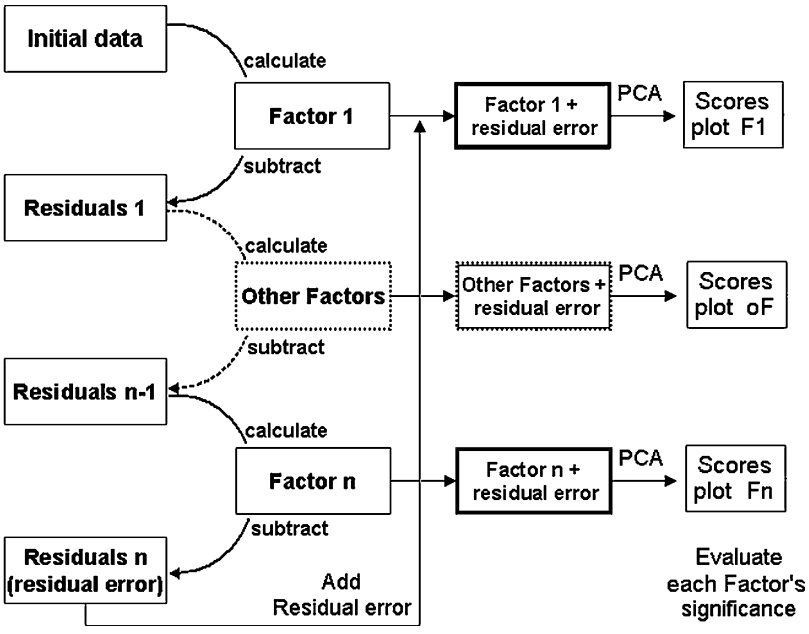


Fig. 1 – Scheme of the ANOVA-PCA method.

to evaluate the significance of the factor (represented by the means) by comparison to the residual error. Finally, one can calculate Hotelling confidence ellipses and determine if the levels for a certain factor are well separated by looking at the centroid/ellipse distance on the scores maps containing contributions from the factor levels and the residuals. Although it is not necessary, it is also possible to perform the same type of non-parametric resampling techniques as in ASCA.

In the original paper of Harrington et al. [1], significant factors are assumed to contribute to just one component (PC1), while the residual error contributes to PC2. However we have observed datasets where this is not the case. In the present paper for example, the factor temperature contributes with relevant information to both PC1 and PC2. The residuals have smaller variance and contribute to PC3.

While significant PCs can be detected in ASCA by the results of bootstrapping, in ANOVA-PCA it can be done by simply comparing the inter-group distances to the residual variability for each PC.

2.2. *Prediction with ANOVA-PCA*

The objective here is to show that the ANOVA-PCA procedure is not limited to detecting significant factors and that it may also be used to predict the factor levels for new samples. In this sense, it may be considered as a method for discriminant analysis. A detailed description of the many other well-known methods [4], such partial least squares-discriminant analysis, factorial discriminant analysis, linear discriminant analysis, is beyond the scope of this paper.

To understand the proposed extension of ANOVA-PCA for classification of new samples, consider as an example the prediction of a new sample $\mathbf{x}_n$ for Factor 1. First the new spectrum is pretreated in the same way as the calibration group spectra. Then, the average of the "calibration set" ($\bar{\mathbf{x}}$) is subtracted from $\mathbf{x}_n$, giving $\mathbf{x}_{nc}$. At this phase, a series of $K \times L$ vectors are calculated by subtracting from this sample vector $\mathbf{x}_{nc}$, each possible combination of level averages for Factor 2 (K levels) and for the Factor 1 × Factor 2 interaction ($L = J \times K$ levels). In matrix form one has:

$$\mathbf{C} = \mathbf{X}_{nc} - \mathbf{B} - \mathbf{AB} \tag{3}$$

where

- $\mathbf{C}$ ($L \times V$) has L lines and the same number of columns (V) as the "calibration" matrix and is defined as the matrix of all possible combinations of subtractions of level averages of all factors;
- $\mathbf{X}_{nc}$ is a matrix whose lines are all equal to $\mathbf{x}_{nc}$;
- $\mathbf{B}$ is a matrix with the level averages for Factor 2 repeated in order of the levels;
- $\mathbf{AB}$ is a matrix with the level averages for the interaction, obtained in a similar way to $\mathbf{B}$.

The matrix **B,** for example, is calculated as follows:

- $\mathbf{Z}$ ($K \times V$) is a matrix in which each line is a level average for Factor 2 (K levels);
- $\mathbf{F}$ ($L \times K$) is a transformation matrix of 0 and 1 s such that Eq. (4) generates repetitions of the averages.

$$\mathbf{B} = \mathbf{F} \cdot \mathbf{Z} \tag{4}$$

This is better understood with an example. Consider a case where Factor 1 has three levels and Factor 2 has two levels, so that $L = J \times K = 6$. Eq. (4) then gives:

$$\begin{bmatrix} \bar{\mathbf{x}}_1^{\mathrm{T}} \\ \bar{\mathbf{x}}_1^{\mathrm{T}} \\ \bar{\mathbf{x}}_1^{\mathrm{T}} \\ \bar{\mathbf{x}}_2^{\mathrm{T}} \\ \bar{\mathbf{x}}_2^{\mathrm{T}} \\ \bar{\mathbf{x}}_2^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \bar{\mathbf{x}}_1^{\mathrm{T}} \\ \bar{\mathbf{x}}_2^{\mathrm{T}} \end{bmatrix}$$

where $\bar{\mathbf{x}}_1^{\mathrm{T}}$ and $\bar{\mathbf{x}}_2^{\mathrm{T}}$ are the average spectra for the two levels of Factor 2.

The number of possible combinations to subtract from the vector being tested is, as said above ($K \times L$). Only some of the vectors in $\mathbf{C}$ will be similar to the vectors describing the average of a level for the desired factor. For the tested sample, all the possible projections onto the PC1–PC2 space for a factor is given by:

$$\mathbf{R} = \mathbf{C} \cdot \mathbf{P}_{12} \tag{5}$$

where

- $\mathbf{R}$ is the matrix with all the possible projections for this sample;
- $\mathbf{C}$ is the matrix of vectors calculated as above;
- $\mathbf{P}_{12}$ is the matrix of PC1 and PC2 loadings for a given factor.

To class the sample, the Mahalanobis distance [12,13] of each of these projections to the barycentres of all the factor levels is calculated. It is expected that the projections will form N clusters, N being the number of levels of the factor with highest amount of variance (in this case, either factor two or the interaction). The classification of the sample is not done just by determining which group is the closest to one of the possible projected vectors as this may be affected by noise. The classification is based on the minimal average distance of a cluster of possibilities. After having sorted the values of all projections in decreasing order of magnitude, one calculates the average distance of the first $(J \times K)/N$ samples to the barycentres of the different levels. The attribution is made to the group with the smallest average distance.

3. Experimental

3.1. *Materials and methods*

3.1.1. *Carraghenan dataset*

1% and 2% solutions of carraghenan (from red seaweed *Euchema denticulatum*, Degussa SA France) were prepared in 0.1 M NaCl and analysed at 30, 40, 45, 50 and 60 °C on two

Table 1 – Number and description of levels for each factor of the carraghenan data experimental design

Factors	Number of levels	Levels
Concentration (C)	2	1; 2 ($g L^{-1}$)
Temperature (T)	5	30, 40, 45, 50, and 60 (°C)
Day (D)	2	1; 2
Sample (S)	20	1–20 (three repetitions each)
Total spectra	60	–

different days in the mid-infrared region using a Fourier Transform Mid-Infrared Spectrometer (Bruker Vector 33) with a thermostated Golden Gate Attenuated Total Reflection (ATR) sampling device (Specac). The solutions were kept in a hot water bath at 78 °C. Then some droplets were placed on the ATR single reflection diamond crystal. Spectra were acquired after a 1 min delay to stabilize the temperature of the sample. The procedure was repeated in triplicate for each sample. 64 scans were collected and averaged over the region 4050–600 cm^{-1}, at 4 cm^{-1} resolution. The experimental design is summarized in Table 1.

All calculations were performed in MATLAB 7.0 (R14), using a computer with an Intel Pentium processor operating at 1.60 GHz and 992 MB RAM.

The software used in spectral data acquisition and analysis was OPUS (version 3.1 build 3, 0, 17) and MATLAB (version 7.0.4.365 (R14) Service Pack 2).

4. Results and discussion

4.1. ANOVA-PCA on the carraghenan data

A four-factor model with no interaction is used to model the data according to the procedure explained above. Other more complete models were also studied but as these showed that interaction were not significant, only the simplified model is presented here. The model is given by:

$$\mathbf{X}_{(i,j)} = \bar{\mathbf{X}}_{(i,j)} + \bar{\mathbf{C}}_{(i,j)} + \bar{\mathbf{T}}_{(i,j)} + \bar{\mathbf{D}}_{(i,j)} + \bar{\mathbf{S}}_{(i,j)} + \boldsymbol{\varepsilon}_{(i,j)}. \quad (6)$$

The factors considered are concentration ($\bar{\mathbf{C}}$), temperature ($\bar{\mathbf{T}}$), day ($\bar{\mathbf{D}}$) and sample ($\bar{\mathbf{S}}$), according to a balanced, fixed effects model. The order of subtraction of factors is just of practical implementation, as it does not change the results. Note however that interactions, when considered, must be subtracted after the main effects. The factor "Sample" should be subtracted last, as it corresponds to the repetitions and so, after its subtraction, only the residuals are left. Moreover, *i*, *j* are the matrix dimensions, respectively number of spectra and number of spectral wavenumbers, $\mathbf{X}$ is the initial data matrix, $\bar{\mathbf{X}}$ is a matrix in which its lines are equal to the grand column-average of $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ is the residual error matrix.

4.1.1. Initial data

The 60 spectra used to build the dataset are presented in Fig. 2. The major band at around 3320 cm^{-1} and another at 1635 cm^{-1} are well known and ascribed to water (–OH group). A closer inspection of the data shows a band centred at 2350 cm^{-1} due to atmospheric CO_2 and a group of bands in the "fingerprint" region of 1300–1000 cm^{-1}, attributed to the carraghenan.

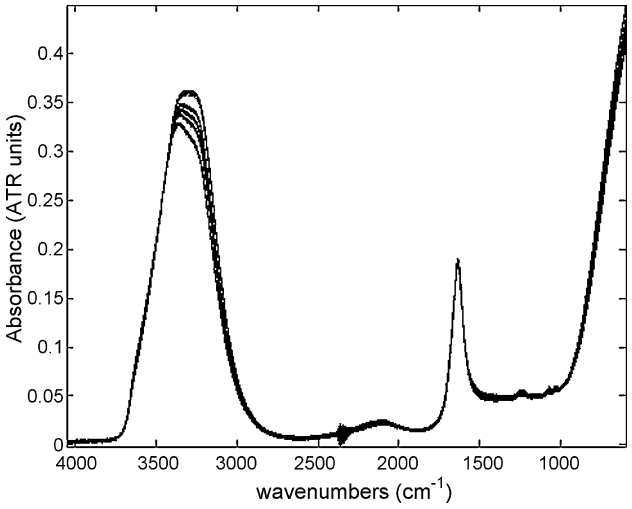


Fig. 2 – Plot of carraghenan spectra dataset.

As the interfering CO_2 peak centered at 2350 cm^{-1} (see Fig. 8) presented high random variance, it was removed. The region 2400–2250 cm^{-1} was simply truncated from the data along with another high residual variance area at 775–600 cm^{-1}. The spectra were then pre-processed using standard normal variates (SNV) [14].

4.1.2. ANOVA-PCA procedure

The first step of ANOVA-PCA is centering the variables, by subtracting the global average ($\bar{\mathbf{X}}$) from $\mathbf{X}$. Some features within the data become much more visible (Fig. 3). After centering, the univariate part of the ANOVA-PCA method is performed, i.e. obtaining the averages for the each of the factor levels (named factor matrices). Then, once all the factor levels are calculated, they are successively subtracted from the initial data, to give the residual error matrix. After adding this residual error matrix back to each of the fac-

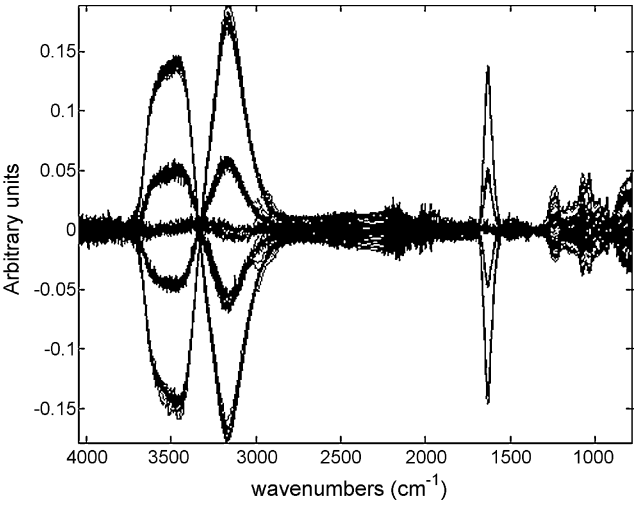


Fig. 3 – Plot of the dataset of centered carraghenan spectra.

tor matrices, the second, multivariate part of the method is performed, i.e. applying PCA to each of the final factor + residual error matrices, in order to compare the variance due to the factors with the variance due to the residual error.

4.1.3. *Scores and loadings of ANOVA-PCA*

4.1.3.1. *Factor concentration* ($\bar{C}$). After applying ANOVA-PCA, by looking at the scores in Fig. 4a one can see that the ANOVA-PCA method shows that the factor concentration is interesting for the characterisation of the dataset. PC1 scores for the two concentrations levels are clearly separated and by looking at the loadings profiles one can see which wavenumbers are important for that separation. PC1 is related to the concentration of carraghenan in the samples, as can be seen by comparing its loadings profiles (Fig. 4b) to the spectra of a very concentrated solution of carraghenan (see Fig. 8). Samples spread out across PC2, due to the influence of the main component present in the residual error, whose spectral features may be seen in its loadings plot (Fig. 4c).

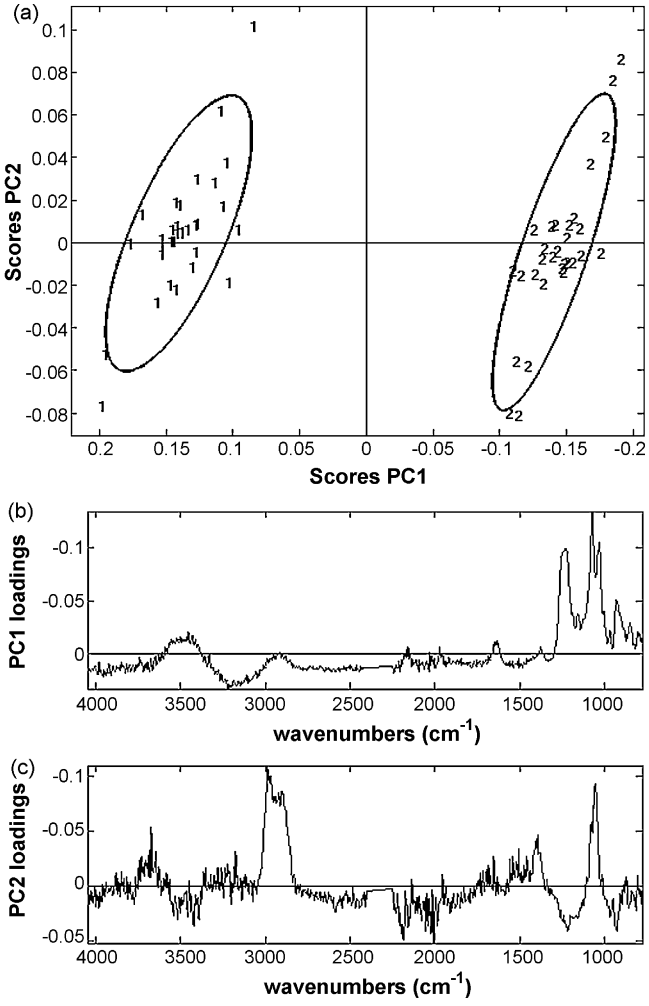


Fig. 4 – Results of PCA on factor concentration plus residual error; scores plot PC1 × PC2 (a) in which numbers 1–2 are for 1 and 2% of carraghenan concentration; loadings PC1 (b) and loadings PC2 (c).

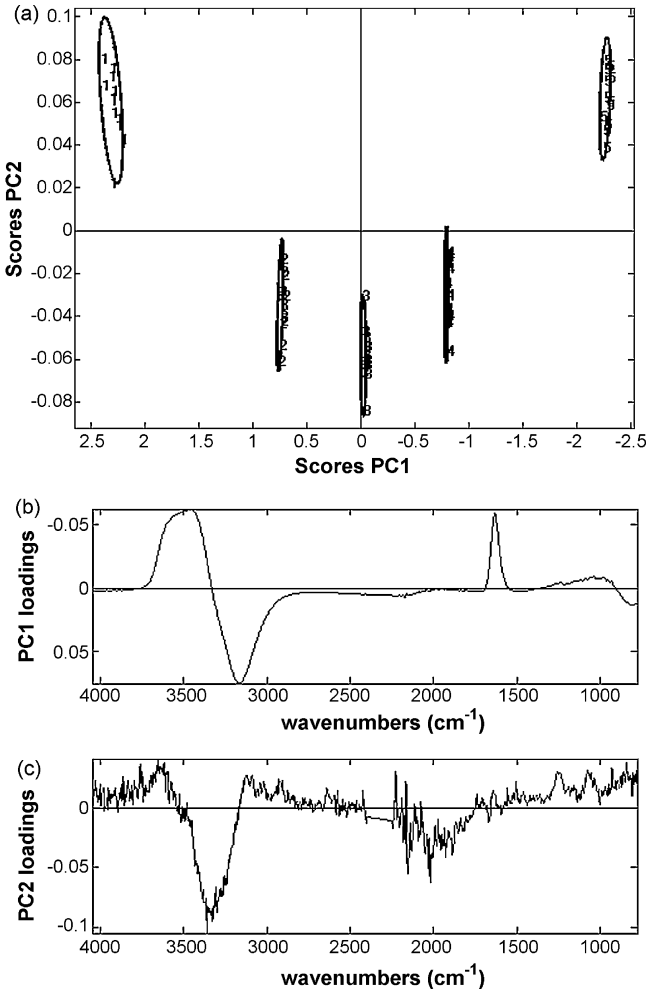


Fig. 5 – Results of PCA on factor temperature plus residual error; scores plot PC1 × PC2 (a) in which numbers 1–5 are for 30, 40, 45, 50, and 60 °C; loadings PC1 (b) and loadings PC2 (c).

4.1.3.2. *Factor temperature* ($\bar{T}$). Fig. 5 shows that the levels for the factor temperature are very well separated into clusters along the PC1 axis. PC1 loadings indicate that the frequencies that are responsible for this separation are mainly related to the absorption wavelengths of water. As found in literature [15], at higher temperatures there is a shift to higher wavenumbers for the peak centered at 3320 cm^{-1} and a decrease in the –OH peak around 3150 cm^{-1}. There is also an increase in the peak centered at 1635 cm^{-1}.

The spread of the samples along PC2 is not only due to residual error but shows a grouping where intermediate temperatures are in opposition to the maximum and minimum temperatures. This may be due to a non-linear behavior of the spectra as a function of temperature. As PC1 clearly separates the clusters, the factor temperature is significant compared to the residual error.

4.1.3.3. *Factor day* ($\bar{D}$). The next factor to analyse is the day during which the spectra were acquired. This was in order to check reproducibility, to understand if different operators

obtain the same results using slightly different procedures. Sample preparation, handling, instrumental behavior and especially the way the ATR crystal was cleaned, were all evaluated together. Therefore, in each of the 2 days different operators cleaned the crystal in a different way, which would also help to determine the most efficient way to do so, as the carraghenan was thought to stick to the crystal. On the first day, ethanol was poured over the crystal and wiped with absorbent paper. On the second day, paper was moistened with ethanol, which was then used to clean the crystal, so that a much smaller quantity of solvent was used.

In relation to the factor day, it looks like it cannot separate the samples into two groups at the 95% confidence level (Fig. 6). It means this can be considered as a fairly homogeneous dataset in respect to that factor, although we can see that the spectra acquired on the first day present more variability (being more spread out on PC1–PC2) than those acquired on the second day. This variability is related to the cleaning procedure used, as one can see by comparing the PC1 loadings to the spectrum of ethanol vapor of Fig. 8. The interference of ethanol was greatly reduced on day 2 by having much less ethanol vapor near the instrument.

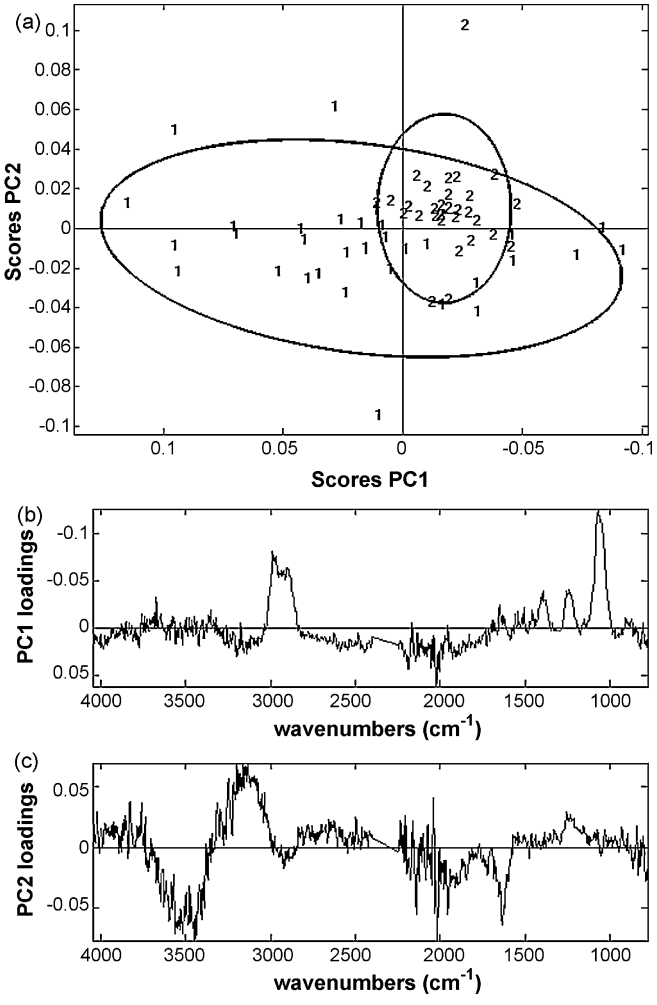


Fig. 6 – Results of PCA on factor day plus residual error; scores plot PC1 × PC2 (a) in which numbers 1–2 are for day 1 and day 2; loadings PC1 (b) and loadings PC2 (c).

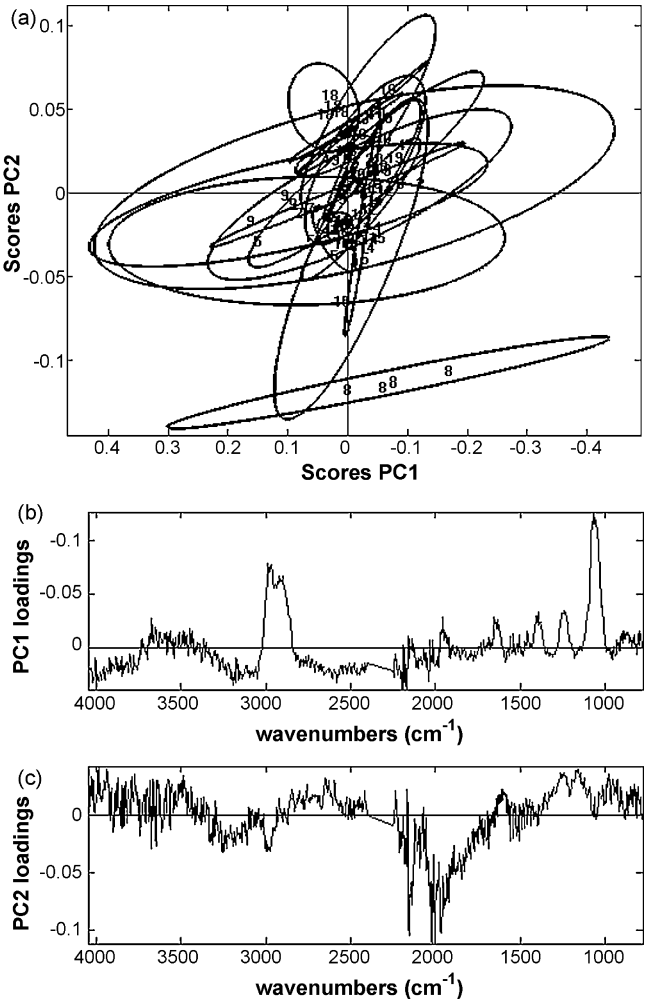


Fig. 7 – Results of PCA on factor sample plus residual error; scores plot PC1 × PC2 (a) in which numbers 1–20 are for the different samples' groups (each three replicates compose a group); loadings PC1 (b) and loadings PC2 (c).

4.1.3.4. *Factor sample* ($\bar{S}$). Finally all clusters of three replicates – factor sample – are superimposed (Fig. 7), which means there is no additional factor present in the data and it is just residual error which is governing the behavior of the scores. It would appear that main variability of the groups is again due to the ethanol vapor (PC1), as the scores are dispersed horizontally along the PC1 axis. On the other hand, some of the groups are distributed along the vertical PC2 axis, due mainly to changes in the baseline, as in the case of group 8.

4.1.3.5. *Interactions*. One of the objectives of the work was to determine the importance of the interaction between the factors concentration and temperature. Other interactions between the factors were also investigated, namely concentration × day, temperature × day and concentration × temperature × day. In fact, no significant interactions were found between the factors.

4.1.3.6. *Constituents spectra*. The spectra of a concentrated solution of carraghenan, of CO_2 and of ethanol vapor are pre-

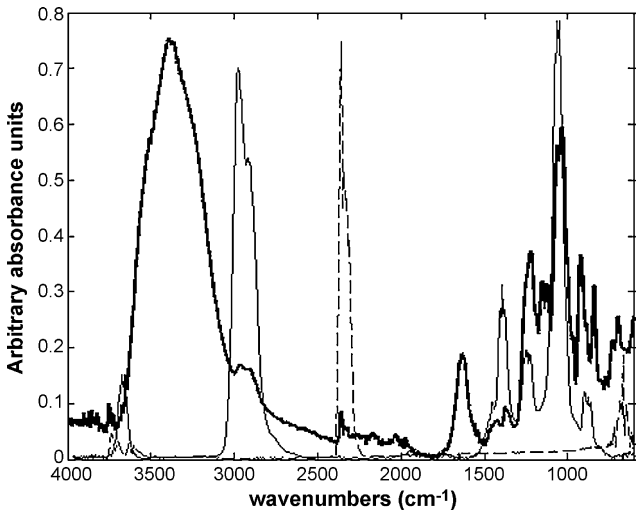


Fig. 8 – Spectra of carraghenan (thick), ethanol vapor (thin) [16] and CO_2 vapor (dashed) [16].

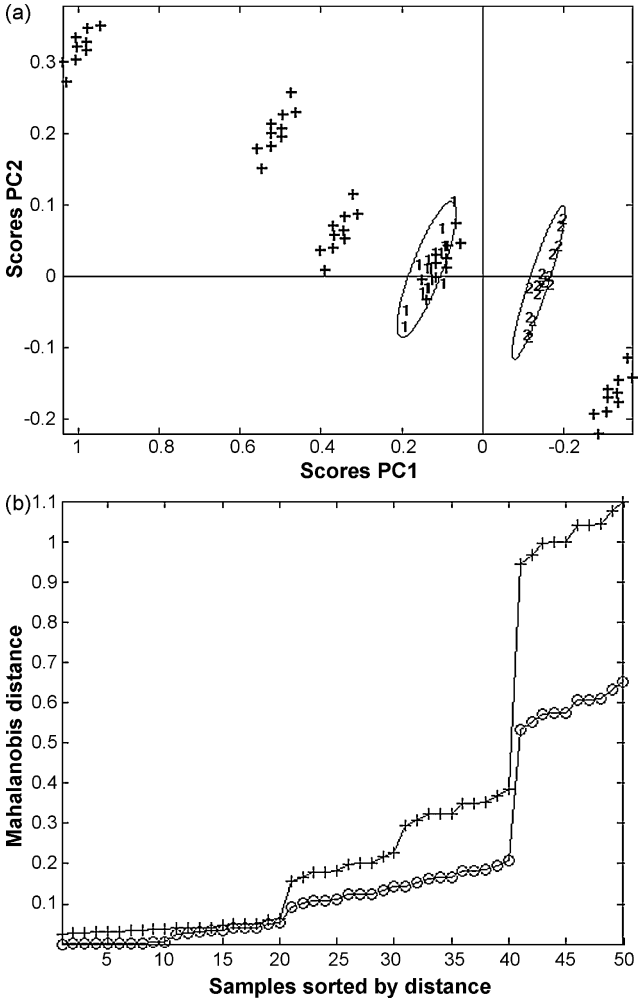


Fig. 9 – Calibration set of concentration using day 1 samples; (a) projection of all possible combinations for the prediction of one sample belonging to level 1 (1%) on factor concentration; (b) sorted Mahalanobis distances in relation to both levels' barycentre (○, distances to group 1; +, distances to group 2).

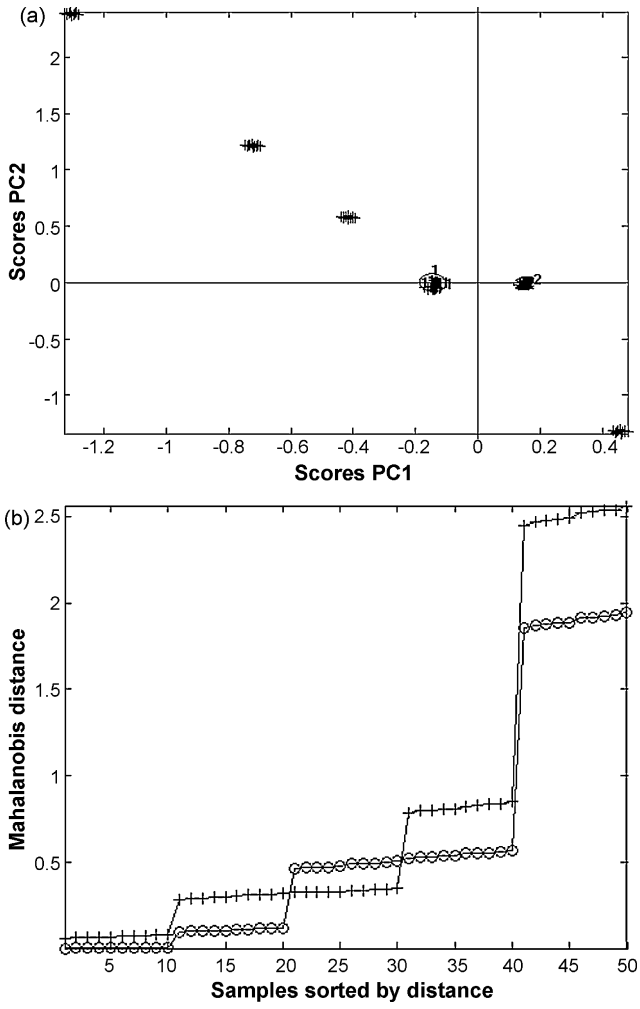


Fig. 10 – Calibration set of concentration using day 2 samples; (a) projection of all possible combinations for the prediction of one sample belonging to level 1 (1%) on factor concentration; (b) sorted Mahalanobis distances in relation to both levels' barycentre (○, distances to group 1; +, distances to group 2).

sented in Fig. 8. They allow us to confirm the relationship between these compounds and the principal component loadings found with the ANOVA-PCA method. As the carragheenan spectrum was obtained after evaporation of a concentrated sample in the ATR crystal, it does not correspond exactly to a spectrum of pure carragheenan. The interesting comparable region is between 1800 and 700 cm^{-1}.

4.2. *Prediction with ANOVA-PCA*

As factor day was seen not to be significant compared to the residual error, the spectra were separated in two sub-groups, one for each day. Then, one of the days was used for the calibration and the other one for prediction using the selected region of the spectra. Since day 1 has more residual error than day 2, it is interesting to compare models made using each of them as the calibration subset.

4.2.1. *Prediction of concentration*

In order to predict the concentration level for a new sample, one subtracts from the sample spectrum all combinations of level means of temperature and sample, resulting in $L = 5 \times 10 = 50$ different possible combinations. Each of them is then multiplied by the matrix of the concentration PC1 and PC2 loadings from the calibration dataset, to give 50 projections (see Fig. 9a). The Mahalanobis distances from these projections to each group centroid is calculated and are sorted by magnitude (see Fig. 9b) to decide to which level the sample belongs.

The pattern of the predicted scores for each sample is similar for all samples and depends on the type and quantity of variance associated with the subtracted factors. The pattern of variability for the day 1 calibration model does not look very promising for the prediction of new samples, as some of the projections may fall inside the ellipses of both groups. This is because the projections are spread along a direction

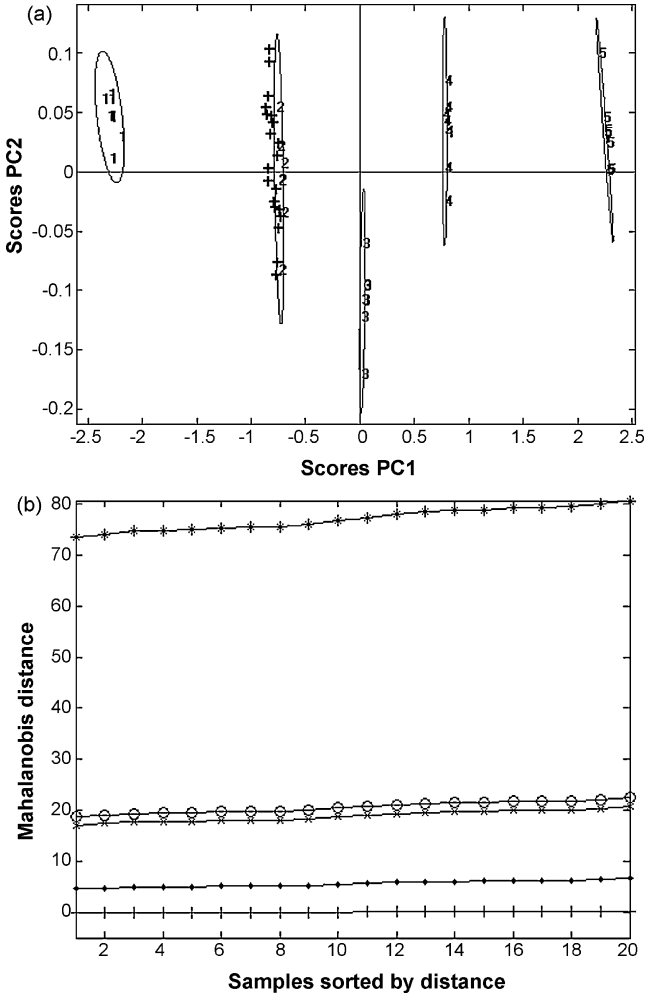


Fig. 11 – Calibration set of temperature using day 1 samples; (a) projection of all possible combinations for the prediction of one sample belonging to level 2 (40 °C) on factor temperature, (b) sorted Mahalanobis distances in relation to all levels' barycentre; Distances to: ○, group 1 (30 °C); +, group 2 (40 °C); ♦, group 3 (45 °C); ×, group 4 (50 °C); and *, group 5 (60 °C).

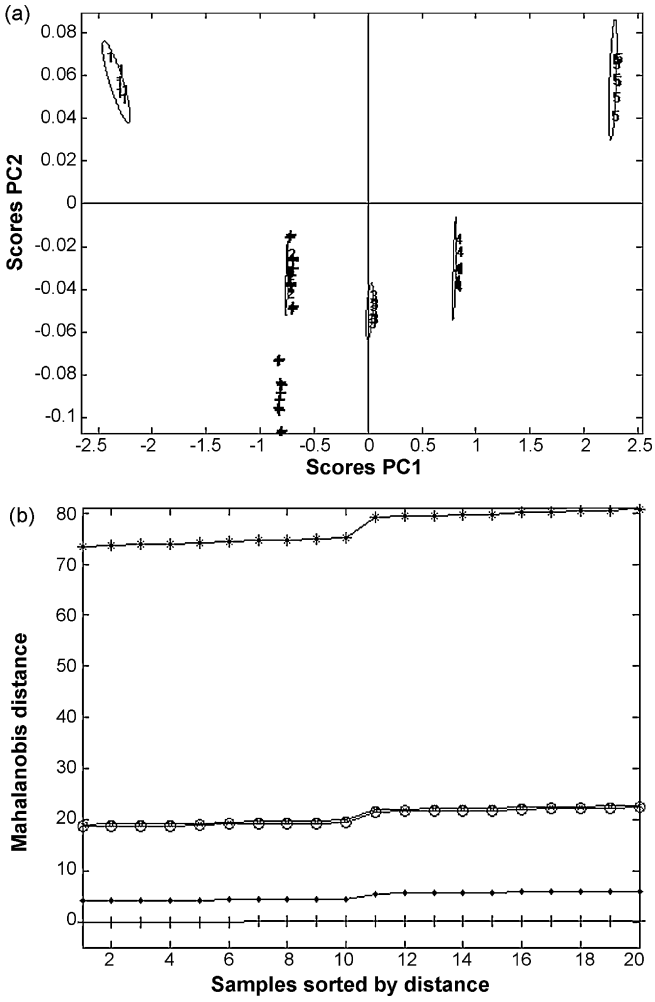


Fig. 12 – Calibration set of temperature using day 2 samples; (a) projection of all possible combinations for the prediction of one sample belonging to level 2 (40 °C) on factor temperature; (b) sorted Mahalanobis distances in relation to all levels' barycentre; Distances to: ○, group 1 (30 °C); +, group 2 (40 °C); ♦, group 3 (45 °C); ×, group 4 (50 °C); and *, group 5 (60 °C).

which is mainly related to the factor temperature, which has the highest variance. As a result, the samples may be wrongly classed, and in fact, for the calibration with day 1, one sample is. This does not happen in the model based on day 2 samples. On this subset, the prediction scores for each of the samples are spread along a distinct axis, which cannot cross over the ellipse of the other group. All day 1 samples were correctly classified using this model.

The use of the average distance of several projected points to the barycentre of the levels gives better results than considering only the distance of the closest prediction to each of the levels. In fact, one sample which was misclassified using the latter method is well classified using the average distance method.

Fig. 10a shows all the possible predictions for one of the samples at 1% concentration using day 2 samples for the calibration. Although the projection pattern for the sample is

similar for both models (day 1 and day 2, see also Fig. 9a), this one has much less dispersion. This is due to the lower residual error in the calibration model using day 2 samples, which leads to better prediction models. In fact, the concentration calibration model obtained using day 2 is better than the one obtained with day 1. Although day 1 samples presented more residual variability than the ones from day 2, all day 1 samples were correctly predicted on the day 2 calibration model. In Fig. 10b, one can see that group 1 is the closest to the average of the first L/N points (50/5 = 10 projections).

4.2.2. *Prediction of temperature*

Fig. 11a shows all the possible temperature projections for one of the samples from day 2 at 40° temperature in the calibration model made with day 1 samples. There are two concentration levels and 10 sample levels, so there are 20 possible combinations for this sample. The variances associated with the factors concentration and sample are much smaller than that of the factor temperature, so the prediction scores pattern are much more compact than in the prediction of concentration. Correct classifications for 100% of the samples were obtained with this model.

Due to the variability associated with the factor concentration, one can see two clusters in the projections for one of the samples of day 2 in the calibration model for temperature made with the samples from day 2 (Fig. 12a). This dispersion does not change the results for the prediction (Fig. 12b), because the projections spread along the PC2 axis but the factor levels spread along the PC1 axis. All samples were correctly predicted using this model.

5. Conclusions

ANOVA-PCA was applied to carraghenan data in order to evaluate the significance of the factors when compared to the residual error. Two factors, temperature and concentration, were found to be significant at 95% when compared to the residual error. The factor temperature indicated also another feature which separated the extreme temperatures from intermediate ones. Interactions between the factors were also investigated, but none was found to be significant. Although the variability for the two levels of factor day is different, with day 1 presenting more variability due to the variation of ethanol vapor in the spectrometer after cleaning the crystal, the two levels could not be separated, so factor day was considered as not significant.

A new prediction method with ANOVA-PCA was proposed. For each factor, the method calculates all the projections (scores) for a given factor after subtracting all possible combinations of levels of the other factors. An average of selected Mahalanobis distances from projections of its scores to the barycentre of the levels is calculated, in order to decide to which level the new sample belongs. It is possible to look at all the possibilities of prediction for the calibration data, which gives a visual picture of whether the model has predictive power.

Some other adaptations of the ANOVA-PCA method are being studied in order to address issues related with the use of non-equilibrated models and to treat qualitative and quantitative factors differently.

REFERENCES

[1] P. Harrington, N. Vieira, J. Espinoza, J. Nien, R. Romero, A. Yergey, Anal. Chim. Acta 544 (2005) 118–127.
[2] P. Harrington, N. Vieira, P. Chen, J. Espinoza, J. Nien, R. Romero, A. Yergey, Chemometr. Intell. Lab. Syst. 82 (2006) 283–293.
[3] J. Sarembaud, R. Pinto, D.N. Rutledge, M. Feinberg, Anal. Chim. Acta 603 (2007) 147–154.
[4] D.L. Massart, B. Vandeginste, L. Buydens, S. De Jong, P. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics, vol. 20A, Elsevier, Amsterdam, 1997, pp. 519–557.
[5] J. Copikova, A.S. Barros, I. Smidova, M. Cerna, D.H. Teixeira, I. Delgadillo, A. Synystsya, M.A. Coimbra, Carbohydr. Polym. 63 (2006) 355–359.
[6] P. Volery, R. Besson, C. Schaffer-Lequart, J. Agric. Food Chem. 52 (2004) 7457–7463.
[7] E.J. van Velzen, J. van Duynhoven, P. Pudney, P. Weegels, J. van der Maas, Cereal Chem. 80 (4) (2003) 378–382.
[8] H.A. Hotelling, Generalized *T* test and measure of multivariate dispersion, in: Second Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, 1951, pp. 23–41.
[9] A.K. Smilde, J.J. Jansen, H.C.J. Hoefsloot, R-J.A.N. Lamers, J. Greef, M.E. Timmerman, Bioinformatics 21 (13) (2005) 3048–4043.
[10] J.J. Jansen, H.C.J. Hoefsloot, J. Greef, M.E. Timmerman, J.A. Westerhuis, A.K. Smilde, J. Chemometr. 19 (9) (2005) 469–481.
[11] D.J. Vis, J.A. Westerhuis, A.K. Smilde, J. Greef, BMC Bioinform. 8 (2007) 322.
[12] P.C. Mahalanobis, On the generalised distance in statistics, in: Proceedings of the National Institute of Science of India, vol. 12, 1936, pp. 49–55.
[13] R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart, Chemometr. Intell. Lab. Syst. 50 (1) (2000) 1–18.
[14] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Appl. Spec. 43 (5) (1989) 772–777.
[15] D. Bertrand, in: D. Bertrand, E. Dufour (Eds.), Spectroscopie de l'eau (La spectroscopie infrarouge et ses applications analytiques, 2nd ed.), Lavoisier Tec&Doc, 2006, pp. 102–103.
[16] NIST Mass Spec Data Center, S.E. Stein (director), in: P.J. Linstrom, W.G. Mallard (Eds.), Infrared Spectra (NIST Chemistry WebBook, NIST Standard Reference Database Number 69), National Institute of Standards and Technology, Gaithersburg, MD, 2005, http://webbook.nist.gov.

ANNEXE IV

Chemometrics and Intelligent Laboratory Systems 93 (2008) 43–48

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Principal component transform — Outer product analysis in the PCA context

A.S. Barros [a,*], R. Pinto [b], D. Jouan-Rimbaud Bouveresse [b], D.N. Rutledge [b]

[a] Departamento de Química, Universidade de Aveiro, 3810-193 Aveiro, Portugal
[b] INRA/AgroParisTech, UMR Ingénierie Analytique pour la Qualité des Aliments (IAQA), 16, rue Claude Bernard, 75005 Paris, France

ARTICLE INFO

Article history:
Received 27 April 2007
Received in revised form 13 March 2008
Accepted 26 March 2008
Available online 7 April 2008

Keywords:
Outer product analysis
PCA
Data fusion

ABSTRACT

Outer product analysis is a method that permits the combination of two spectral domains with the aim of emphasizing co-evolutions of spectral regions. This data fusion technique consists in the product of all combinations of the variables that define each spectral domain. The main issue concerning the application of this technique is the very wide data matrix obtained which can be very hard to handle with multivariate techniques such as PCA or PLS, due to computer resources constraints. The present work presents an alternative way to perform outer product analysis in the PCA context without incurring into high demands on computational resources. This works shows that by decomposing each spectral domain with PCA and performing the outer product on the recovered scores, one can obtain the same results as if one calculated the outer product in the original variable space, but using much less computational resources. The results show that this approach will make possible to apply outer product analysis to very wide domains.

1. Introduction

Outer product analysis is a method of data fusion which combines two (or even more) datasets by first calculating an outer-product matrix from each pair of sample vectors. Each matrix is then unfolded to produce a very wide matrix containing all possible product combinations between the two datasets. This operation will, for instance, emphasize co-evolutions of spectral regions in signals acquired in two different domains (heterospectral) or even within the same domain (homospectral). This data fusion technique has been applied in different contexts of instrumental analysis. The seminal work was published in 1997 by Barros et al. [1] for the analysis of different oils measured by Mid Infrared (MIR) and Near Infrared (NIR) spectroscopy. This technique was also used for the quantification and characterization of suberin in cork by Fourier Transform Infrared (FT-IR) spectroscopy and ^{13}C NMR [2]. Many other works have been published using this approach. More recently, the outer product was analysed from a multi-way data perspective, where signals were combined by means of outer product and analysed using Parallel Factor Analysis (PARAFAC) [3,4].

Since the outer product matrix represents all possible product combinations of the variables of both domains, the resulting matrix can be very wide and hence, time consuming to build and analyse by means of multivariate methods. Therefore, it is important to find ways to analyse these matrices more quickly, especially, for instance, when one needs to perform exploratory analyses and cross-validations, as these approaches are very demanding in terms of computer resources (time and memory). There are many ways to perform PCA [5–8] on very wide matrices (megavariate matrices) [8–10], but all are time consuming. On the other hand, although the drawbacks of the Lanczos or Arnoldi methods to perform the Singular Value Decomposition (SVD) of large matrices [11] can be solved by general Krylov decomposition [12,13], these approaches are not straightforward to implement and in most cases they only recover the most important eigenvectors (larger eigenvalues). Alternatively, one can use wavelets decomposition [14,15] and then perform PCA on the most important wavelet coefficients. The idea of compressed spaces has also been used in other methods for matrix decomposition [16,17]. The postponed basis matrix multiplication (PBM) method developed by Alsberg and Kvalheim [21,22] projects the data onto a particular base space (much smaller than the initial data space) chosen *a priori*. This operation will give a compressed data matrix which is much easier to handle computationally. Still, the bases onto which the initial space is to be projected must be chosen in such a way that little of the original internal structure is lost. This issue was later solved by Kiers and Harshman [18] by using columnwise orthonormal basis matrices (by performing SVD on the chosen base spaces) to perform the projection and hence ensuring that the scores and loadings obtained were the true ones. However, one still has to find the subspace to be used in the projection.

Another interesting approach for very wide matrices is to use the kernel method [19] to decompose a matrix of the form $\mathbf{XX}^T$, rather than $\mathbf{X}^T\mathbf{X}$ which can be a very fast alternative to perform PCA in the original variable space. However, despite the fact that the kernel method is very efficient for tall or wide matrices, the main issue in outer product analysis, is (or can be) to the calculation of the initial outer product matrix and the recovery of the original space loadings,

* Corresponding author. Tel.: +351 234 372 581; fax: +351 234 370 084.
E-mail address: antonio.barros@ua.pt (A.S. Barros).

0169-7439/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2008.03.009

which, in many cases, can be very computer intensive. Recently, a method for matrix segmentation (SegPCT-PCA) [20] was proposed that can be used to decompose very wide matrices, but, in this case, one would still have the same constraints for the building of the initial outer product matrix. This work proposes to avoid the construction of the outer product matrix in the original space and instead, to build an outer product matrix based on the scores calculated for each domain and then to perform the PCA on this smaller outer product scores matrix. This approach is based on the Principal Component Transform (PCT) concept already used in the PLS1 regression [21] to speed up the cross-validation process.

2. Theory

2.1. Notations

Matrices are shown as bold uppercase (**X**), column vectors bold lowercase (**x**) and row vectors as $\mathbf{x}^T$ (transposed). To facilitate comprehension, matrix dimensions are often shown as $\mathbf{X}_{(n,\ m)}$, where n is the number of rows (objects) and m is the number of columns (variables). An important concept in standard outer product analysis is the vec operator which is used to unfold a matrix to give a vector. As an example, given a $\mathbf{X}_{(n,\ m)}$ matrix, the *vec* operator transforms (unfolds) the **X** matrix into a vector, i.e., $\mathbf{x}_{(n\cdot m,\ 1)}=\mathrm{vec}(\mathbf{X}_{(n,\ m)})$. The operator ⊗ stands for the Kronecker operator or direct tensor product.

The outer product operator is represented by Θ. Considering two domains, if matrix $\mathbf{X}_{(n,\ m)}$ represents one domain and matrix $\mathbf{Y}_{(n,\ p)}$, represents the other domain, then the outer product of these two domains can be represented by $\mathbf{K}_{(n,\ m\cdot p)}=\mathbf{X}_{(n,\ m)}\Theta\mathbf{Y}_{(n,\ p)}$.

2.2. Outer product analysis

In many cases, one may wish to determine the relationships that exist between two types of signal such as Near Infrared (NIR) and Mid Infrared (MIR), MIR and Raman, UV–Visible and Nuclear Magnetic Resonance (NMR), etc. To do this, it may be useful to acquire the two sets of signals for the same samples and analyse how they vary simultaneously as a function of some property, such as concentration. Outer Product Analysis does this by applying chemometric techniques to the n outer product matrices calculated, for each of the n samples, by outer product multiplication of the two signal vectors.

Essentially, the procedure starts by calculating the products of the intensities in the two signal domains for each sample. All the intensities of one domain are multiplied by all intensities in the other domain, resulting in a data matrix containing all possible combinations of the intensities in the two domains. The Outer Product of two signal-vectors of lengths m and p for the n samples gave n (m rows by p columns) matrices which are then unfolded to give n ($m\cdot p$)-long row-vectors. This procedure corresponds to a mutual weighting of each signal by the other:

i) if the intensities are simultaneously high in the two domains, the product is higher;
ii) if the intensities are simultaneously low in the two domains, the product is lower;
iii) if one of the intensities is high and the other low, the resulting product tends to an intermediate value.

The resulting outer product matrix, $\mathbf{K}_{(n,\ m\cdot p)}$ can be analysed by many chemometric methods such as: Principal Component Analysis (PCA), Partial Least Squares regression (PLS), Factorial Discriminant Analysis (FDA), etc. After the chemometric analysis of the outer product matrix, each latent vector (from PCA, PLS, FDA etc.) can be folded back to give a (m rows by p columns) matrix (latent matrix), which may be easily examined to detect the relations between the two domains, based on the properties of the applied chemometric technique.

An obvious computational problem of this approach is that, since the outer product represents all possible products combination between the variables of both domains, the outer product matrices are often very wide, which in many situations makes impossible the application of chemometric techniques, mainly due to limitations of computer resources.

2.3. Outer product in the context of Principal Component Analysis

Let us consider two domains represented by the $\mathbf{X}_{(n,\ m)}$ and $\mathbf{Y}_{(n,\ p)}$ matrices. As previously seen (*c.f.* Section 2.2), the outer product (OP) matrix is defined as the product of all possible combinations between the $\mathbf{X}_{(n,\ m)}$ and $\mathbf{Y}_{(n,\ p)}$ columns. The resulting OP matrix is defined as $\mathbf{K}_{(n,\ m\cdot p)}$, where n represents the number of rows, m and p the number of columns (variables) of **X** and **Y**, respectively. It is clear that building the OP matrix (**K**) can, in many cases, be very demanding (mainly due to computer memory limitations), and consequently, the application of PCA would be very time and memory consuming, or even not feasible.

One possible approach is to do the eigen decomposition of the kernel matrix $\mathbf{K}_{(n,\ m\cdot p)}\mathbf{K}^T_{(m\cdot p,\ n)}$, however, the limiting step is, in fact, not the eigen decomposition of $\mathbf{K}\mathbf{K}^T$ but the process of building the outer product matrix. Therefore, it would be important to find new ways to build and analyse such very wide matrices without the storage and computation time constraints.

As the OP operation is performed on an object (row) basis, one can see that for a given object i one has:

$$\mathbf{K}_{i(m,p)} = \mathbf{x}_{(m,1)}\mathbf{y}^T_{(1,p)}$$

or equivalently, the unfolding of the $\mathbf{K}_i$ matrix gives:

$$\mathbf{k}^T_{i(1,m\cdot p)} = \left[\mathrm{vec}\left(\mathbf{x}_{(m,1)}\mathbf{y}^T_{(1,p)}\right)\right]^T \tag{1}$$

Since by definition of tensor vector products [22], the outer product (K) can be seen as:

$$\mathbf{K}_{(n,m\cdot p)} = \begin{bmatrix} (\mathbf{x}^T\otimes\mathbf{y}^T)_1 \\ (\mathbf{x}^T\otimes\mathbf{y}^T)_2 \\ \vdots \\ (\mathbf{x}^T\otimes\mathbf{y}^T)_n \end{bmatrix} \tag{2}$$

The PCA-NIPALS algorithm decomposition of an **X** matrix has generally five steps, starting with a random **t** vector:

1. start with a random **t** vector
2. $\mathbf{p}^T_{(1,\ m)}=\mathbf{t}^T_{(1,n)}\mathbf{X}_{(n,\ m)}$
3. $\mathbf{t}_{(n,\ 1)}=\mathbf{X}_{(n,\ m)}\mathbf{p}_{(m,1)}$iterate steps 2–3 until convergence
4. then deflate **X**

$$\mathbf{X}_{(n,m)} = \mathbf{E}_{(n,m)} = \mathbf{X}_{(n,m)} - \mathbf{t}_{(n,1)}\mathbf{p}^T_{(1,m)}$$

5. go to step 1 to recover the next Principal Component.

In the outer product analysis case, the **X** matrix in the NIPALS decomposition is replaced by the **K** matrix (outer product matrix). Therefore one starts at step 1 with a random **t** vector and one follows to step 2, by calculating the loadings:

$\mathbf{p}^T_{(1,\ m\cdot p)}=\mathbf{t}^T_{(1,\ n)}\mathbf{K}_{(n,\ m\cdot \mathrm{p})}$, then, from expression (2) one gets:

$$\mathbf{p}^T_{(1,m\cdot p)} = \mathbf{t}^T_{(1,n)}\begin{bmatrix} (\mathbf{x}^T\otimes\mathbf{y}^T)_1 \\ (\mathbf{x}^T\otimes\mathbf{y}^T)_2 \\ \vdots \\ (\mathbf{x}^T\otimes\mathbf{y}^T)_n \end{bmatrix}_{(n,m\cdot p)} \tag{3}$$

On the other hand, each object i of **x** and **y** domains, can be decomposed as:

$$\mathbf{x}^T_{i(1,m)} = \mathbf{t}^T_{iX(1,kX)}\mathbf{P}^T_{X(kX,m)} \text{ and } \mathbf{y}^T_{i(1,p)} = \mathbf{t}^T_{iY(1,kY)}\mathbf{P}^T_{Y(kY,p)}$$

where kX and kY are the number of Principal Components (PCs) of $\mathbf{X}$ and $\mathbf{Y}$ matrices, respectively.

Therefore, Eq. (3) can be further developed to give:

$$\mathbf{p}_{(1,m\cdot p)}^{\mathrm{T}} = \mathbf{t}_{(1,n)}^{\mathrm{T}} \begin{bmatrix} (\mathbf{t}_{1X}^{\mathrm{T}}\mathbf{P}_X^{\mathrm{T}}) \otimes (\mathbf{t}_{1Y}^{\mathrm{T}}\mathbf{P}_Y^{\mathrm{T}}) \\ (\mathbf{t}_{2X}^{\mathrm{T}}\mathbf{P}_X^{\mathrm{T}}) \otimes (\mathbf{t}_{2Y}^{\mathrm{T}}\mathbf{P}_Y^{\mathrm{T}}) \\ \vdots \\ (\mathbf{t}_{nX}^{\mathrm{T}}\mathbf{P}_X^{\mathrm{T}}) \otimes (\mathbf{t}_{nY}^{\mathrm{T}}\mathbf{P}_Y^{\mathrm{T}}) \end{bmatrix}_{(n,m\cdot p)} \tag{4}$$

Now, considering a property of the Kronecker product, in which $(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D})=(\mathbf{AC})\otimes(\mathbf{BD})$, Eq. (4) can be rearranged as:

$$\mathbf{p}_{(1,m\cdot p)}^{\mathrm{T}} = \mathbf{t}_{(1,n)}^{\mathrm{T}} \begin{bmatrix} (\mathbf{t}_{1X}^{\mathrm{T}} \otimes \mathbf{t}_{1Y}^{\mathrm{T}})(\mathbf{P}_X^{\mathrm{T}} \otimes \mathbf{P}_Y^{\mathrm{T}}) \\ (\mathbf{t}_{2X}^{\mathrm{T}} \otimes \mathbf{t}_{2Y}^{\mathrm{T}})(\mathbf{P}_X^{\mathrm{T}} \otimes \mathbf{P}_Y^{\mathrm{T}}) \\ \vdots \\ (\mathbf{t}_{nX}^{\mathrm{T}} \otimes \mathbf{t}_{nY}^{\mathrm{T}})(\mathbf{P}_X^{\mathrm{T}} \otimes \mathbf{P}_Y^{\mathrm{T}}) \end{bmatrix}_{(n,m\cdot p)} \tag{5}$$

and since this expression is a partitioned matrix, it can be further rearranged to:

$$\mathbf{p}_{(1,m\cdot p)}^{\mathrm{T}} = \mathbf{t}_{(1,n)}^{\mathrm{T}} \begin{bmatrix} \mathbf{t}_{1X}^{\mathrm{T}} \otimes \mathbf{t}_{1Y}^{\mathrm{T}} \\ \mathbf{t}_{2X}^{\mathrm{T}} \otimes \mathbf{t}_{2Y}^{\mathrm{T}} \\ \vdots \\ \mathbf{t}_{nX}^{\mathrm{T}} \otimes \mathbf{t}_{nY}^{\mathrm{T}} \end{bmatrix}_{(n,kX\cdot kY)} (\mathbf{P}_X^{\mathrm{T}} \otimes \mathbf{P}_Y^{\mathrm{T}})_{(kX\cdot kY,m\cdot p)} \tag{6}$$

Post-multiplying both sides by $(\mathbf{P}_X\otimes\mathbf{P}_Y)$, and since $(\mathbf{P}_X^{\mathrm{T}}\otimes\mathbf{P}_Y^{\mathrm{T}})(\mathbf{P}_X\otimes\mathbf{P}_Y)=(\mathbf{P}_X^{\mathrm{T}}\mathbf{P}_X)\otimes(\mathbf{P}_Y^{\mathrm{T}}\mathbf{P}_Y)$ and $\mathbf{P}^{\mathrm{T}}\mathbf{P}=\mathbf{I}$, then $(\mathbf{P}_X^{\mathrm{T}}\otimes\mathbf{P}_Y^{\mathrm{T}})(\mathbf{P}_X\otimes\mathbf{P}_Y)=\mathbf{I}\otimes\mathbf{I}=\mathbf{I}$, so we can have Eq. (6) simplified to:

$$\mathbf{p}_{(1,m\cdot p)}^{\mathrm{T}}(\mathbf{P}_X \otimes \mathbf{P}_Y)_{(m\cdot p,kX\cdot kY)} = \mathbf{t}_{(1,n)}^{\mathrm{T}} \begin{bmatrix} \mathbf{t}_{1X}^{\mathrm{T}} \otimes \mathbf{t}_{1Y}^{\mathrm{T}} \\ \mathbf{t}_{2X}^{\mathrm{T}} \otimes \mathbf{t}_{2Y}^{\mathrm{T}} \\ \vdots \\ \mathbf{t}_{nX}^{\mathrm{T}} \otimes \mathbf{t}_{nY}^{\mathrm{T}} \end{bmatrix}_{(n,kX\cdot kY)} \tag{7}$$

Defining $\mathbf{p}_{\mathrm{PCT}(1,\ kX\cdot kY)}^{\mathrm{T}}=\mathbf{p}_{(1,\ m\cdot p)}^{\mathrm{T}}(\mathbf{P}_X\otimes\mathbf{P}_Y)_{(m\cdot p,\ kX\cdot kY)}$, where the PCT subscript represents the vector in the PC-space [20,21], Eq. (7) can be further simplified by representing it in a compressed space (PCT-space) as:

$$\mathbf{p}_{PCT(1,kX\cdot kY)}^{\mathrm{T}} = \mathbf{t}_{(1,n)}^{\mathrm{T}} \begin{bmatrix} \mathbf{t}_{1X}^{\mathrm{T}} \otimes \mathbf{t}_{1Y}^{\mathrm{T}} \\ \mathbf{t}_{2X}^{\mathrm{T}} \otimes \mathbf{t}_{2Y}^{\mathrm{T}} \\ \vdots \\ \mathbf{t}_{nX}^{\mathrm{T}} \otimes \mathbf{t}_{nY}^{\mathrm{T}} \end{bmatrix}_{(n,kX\cdot kY)} \tag{8}$$

Eq. (8), is equivalent to step 2 of NIPALS algorithm, while drastically reducing the size of the loadings vector from $m\cdot p$ to $kX\cdot kY$.

In order to calculate the scores, step 3 of the above algorithm, one has:

$$\mathbf{t}_{(n,1)} = \mathbf{K}_{(n,m\cdot p)}\mathbf{p}_{(m\cdot p,1)}$$

Then, again from expression (2), one gets:

$$\mathbf{t}_{(n,1)} \begin{bmatrix} (\mathbf{x}^{\mathrm{T}} \otimes \mathbf{y}^{\mathrm{T}})_1 \\ (\mathbf{x}^{\mathrm{T}} \otimes \mathbf{y}^{\mathrm{T}})_2 \\ \vdots \\ (\mathbf{x}^{\mathrm{T}} \otimes \mathbf{y}^{\mathrm{T}})_n \end{bmatrix} \mathbf{p}_{(m\cdot p,1)} = \begin{bmatrix} (\mathbf{t}_{1X}^{\mathrm{T}} \otimes \mathbf{t}_{1Y}^{\mathrm{T}})(\mathbf{P}_X^{\mathrm{T}} \otimes \mathbf{P}_Y^{\mathrm{T}}) \\ (\mathbf{t}_{2X}^{\mathrm{T}} \otimes \mathbf{t}_{2Y}^{\mathrm{T}})(\mathbf{P}_X^{\mathrm{T}} \otimes \mathbf{P}_Y^{\mathrm{T}}) \\ \vdots \\ (\mathbf{t}_{nX}^{\mathrm{T}} \otimes \mathbf{t}_{nY}^{\mathrm{T}})(\mathbf{P}_X^{\mathrm{T}} \otimes \mathbf{P}_Y^{\mathrm{T}}) \end{bmatrix} \mathbf{p}_{(m\cdot p,1)}$$

$$= \begin{bmatrix} \mathbf{t}_{1X}^{\mathrm{T}} \otimes \mathbf{t}_{1Y}^{\mathrm{T}} \\ \mathbf{t}_{2X}^{\mathrm{T}} \otimes \mathbf{t}_{2Y}^{\mathrm{T}} \\ \vdots \\ \mathbf{t}_{nX}^{\mathrm{T}} \otimes \mathbf{t}_{nY}^{\mathrm{T}} \end{bmatrix}_{(n,kX\cdot kY)} (\mathbf{P}_X^{\mathrm{T}} \otimes \mathbf{P}_Y^{\mathrm{T}})_{(kX\cdot kY,m\cdot p)}\mathbf{p}_{(m\cdot p,1)} \tag{9}$$

As by definition, $\mathbf{p}_{\mathrm{PCT}(kX\cdot kY,\ 1)}=(\mathbf{P}_X^{\mathrm{T}}\otimes\mathbf{P}_Y^{\mathrm{T}})_{(kX\cdot kY,\ m\cdot p)}\ \mathbf{p}_{(m\cdot p,\ 1)}$, then expression (9) is simplified to:

$$\mathbf{t}_{(n,1)} \begin{bmatrix} \mathbf{t}_{1X}^{\mathrm{T}} \otimes \mathbf{t}_{1Y}^{\mathrm{T}} \\ \mathbf{t}_{2X}^{\mathrm{T}} \otimes \mathbf{t}_{2Y}^{\mathrm{T}} \\ \vdots \\ \mathbf{t}_{nX}^{\mathrm{T}} \otimes \mathbf{t}_{nY}^{\mathrm{T}} \end{bmatrix}_{(n,kX\cdot kY)} \mathbf{p}_{\mathrm{PCT}(kX\cdot kY,1)}$$

After convergence (step 4), the $\mathbf{K}$ matrix is deflated as:

$$\mathbf{E}_{(n,m\cdot p)} = \mathbf{K}_{(n,m\cdot p)} - \mathbf{t}_{(n,1)}\mathbf{p}_{(1,m\cdot p)}^{\mathrm{T}} = \begin{bmatrix} (\mathbf{x}^{\mathrm{T}} \otimes \mathbf{y}^{\mathrm{T}})_1 \\ (\mathbf{x}^{\mathrm{T}} \otimes \mathbf{y}^{\mathrm{T}})_2 \\ \vdots \\ (\mathbf{x}^{\mathrm{T}} \otimes \mathbf{y}^{\mathrm{T}})_n \end{bmatrix} - \mathbf{t}\mathbf{t}^{\mathrm{T}} \begin{bmatrix} \mathbf{t}_{1X}^{\mathrm{T}} \otimes \mathbf{t}_{1Y}^{\mathrm{T}} \\ \mathbf{t}_{2X}^{\mathrm{T}} \otimes \mathbf{t}_{2Y}^{\mathrm{T}} \\ \vdots \\ \mathbf{t}_{nX}^{\mathrm{T}} \otimes \mathbf{t}_{nY}^{\mathrm{T}} \end{bmatrix} (\mathbf{P}_X^{\mathrm{T}} \otimes \mathbf{P}_Y^{\mathrm{T}})$$

Post-multiplying by $(\mathbf{P}_X\otimes\mathbf{P}_Y)$ one gets:

$$\mathbf{E}(\mathbf{P}_X \otimes \mathbf{P}_Y) = \begin{bmatrix} (\mathbf{x}^{\mathrm{T}} \otimes \mathbf{y}^{\mathrm{T}})_1 \\ (\mathbf{x}^{\mathrm{T}} \otimes \mathbf{y}^{\mathrm{T}})_2 \\ \vdots \\ (\mathbf{x}^{\mathrm{T}} \otimes \mathbf{y}^{\mathrm{T}})_n \end{bmatrix} (\mathbf{P}_X \otimes \mathbf{P}_Y) - \mathbf{t}\mathbf{t}^{T} \begin{bmatrix} \mathbf{t}_{1X}^{\mathrm{T}} \otimes \mathbf{t}_{1Y}^{\mathrm{T}} \\ \mathbf{t}_{2X}^{\mathrm{T}} \otimes \mathbf{t}_{2Y}^{\mathrm{T}} \\ \vdots \\ \mathbf{t}_{nX}^{\mathrm{T}} \otimes \mathbf{t}_{nY}^{\mathrm{T}} \end{bmatrix} (\mathbf{P}_X^{\mathrm{T}} \otimes \mathbf{P}_Y^{\mathrm{T}})(\mathbf{P}_X \otimes \mathbf{P}_Y)$$

Again, since $(\mathbf{P}_X^{\mathrm{T}}\otimes\mathbf{P}_Y^{\mathrm{T}})(\mathbf{P}_X\otimes\mathbf{P}_Y)=\mathbf{I}\otimes\mathbf{I}=\mathbf{I}$, then:

$$\mathbf{E}(\mathbf{P}_X \otimes \mathbf{P}_Y) = \begin{bmatrix} (\mathbf{x}^{\mathrm{T}} \otimes \mathbf{y}^{\mathrm{T}})_1 \\ (\mathbf{x}^{\mathrm{T}} \otimes \mathbf{y}^{\mathrm{T}})_2 \\ \vdots \\ (\mathbf{x}^{\mathrm{T}} \otimes \mathbf{y}^{\mathrm{T}})_n \end{bmatrix} (\mathbf{P}_X \otimes \mathbf{P}_Y) - \mathbf{t}\mathbf{t}^{\mathrm{T}} \begin{bmatrix} \mathbf{t}_{1X}^{\mathrm{T}} \otimes \mathbf{t}_{1Y}^{\mathrm{T}} \\ \mathbf{t}_{2X}^{\mathrm{T}} \otimes \mathbf{t}_{2Y}^{\mathrm{T}} \\ \vdots \\ \mathbf{t}_{nX}^{\mathrm{T}} \otimes \mathbf{t}_{nY}^{\mathrm{T}} \end{bmatrix}$$

can be further simplified to:

$$\mathbf{E}(\mathbf{P}_X \otimes \mathbf{P}_Y) \begin{bmatrix} \mathbf{t}_{1X}^{\mathrm{T}} \otimes \mathbf{t}_{1Y}^{\mathrm{T}} \\ \mathbf{t}_{2X}^{\mathrm{T}} \otimes \mathbf{t}_{2Y}^{\mathrm{T}} \\ \vdots \\ \mathbf{t}_{nX}^{\mathrm{T}} \otimes \mathbf{t}_{nY}^{\mathrm{T}} \end{bmatrix} - \mathbf{t}\mathbf{t}^{\mathrm{T}} \begin{bmatrix} \mathbf{t}_{1X}^{\mathrm{T}} \otimes \mathbf{t}_{1Y}^{\mathrm{T}} \\ \mathbf{t}_{2X}^{\mathrm{T}} \otimes \mathbf{t}_{2Y}^{\mathrm{T}} \\ \vdots \\ \mathbf{t}_{nX}^{\mathrm{T}} \otimes \mathbf{t}_{nY}^{\mathrm{T}} \end{bmatrix} = (\mathbf{I} - \mathbf{t}\mathbf{t}^{\mathrm{T}}) \begin{bmatrix} \mathbf{t}_{1X}^{\mathrm{T}} \otimes \mathbf{t}_{1Y}^{\mathrm{T}} \\ \mathbf{t}_{2X}^{\mathrm{T}} \otimes \mathbf{t}_{2Y}^{\mathrm{T}} \\ \vdots \\ \mathbf{t}_{nX}^{\mathrm{T}} \otimes \mathbf{t}_{nY}^{\mathrm{T}} \end{bmatrix}$$

which is equal to:

$$\begin{bmatrix} \mathbf{t}_{1EX}^{\mathrm{T}} \otimes \mathbf{t}_{1EY}^{\mathrm{T}} \\ \mathbf{t}_{2EX}^{\mathrm{T}} \otimes \mathbf{t}_{2EY}^{\mathrm{T}} \\ \vdots \\ \mathbf{t}_{nEX}^{\mathrm{T}} \otimes \mathbf{t}_{nEY}^{\mathrm{T}} \end{bmatrix} = (\mathbf{I} - \mathbf{t}\mathbf{t}^{\mathrm{T}}) \begin{bmatrix} \mathbf{t}_{1X}^{\mathrm{T}} \otimes \mathbf{t}_{1Y}^{\mathrm{T}} \\ \mathbf{t}_{2X}^{\mathrm{T}} \otimes \mathbf{t}_{2Y}^{\mathrm{T}} \\ \vdots \\ \mathbf{t}_{nX}^{\mathrm{T}} \otimes \mathbf{t}_{nY}^{\mathrm{T}} \end{bmatrix} \tag{10}$$

Expression 10 shows that the deflating process is also performed in a compressed space of size (n, $kX\cdot kY$), instead of the original size (n, $m\cdot p$).

The core expression used in the NIPALS algorithm:

$$\begin{bmatrix} \mathbf{t}_{1X}^{\mathrm{T}} \otimes \mathbf{t}_{1Y}^{\mathrm{T}} \\ \mathbf{t}_{2X}^{\mathrm{T}} \otimes \mathbf{t}_{2Y}^{\mathrm{T}} \\ \vdots \\ \mathbf{t}_{nX}^{\mathrm{T}} \otimes \mathbf{t}_{nY}^{\mathrm{T}} \end{bmatrix}$$

shows that, in order to perform a PCA on an outer product matrix ($\mathbf{K}$), it is sufficient to perform a full rank PCA on both $\mathbf{X}$ and $\mathbf{Y}$ matrices, then recover the correspondent scores matrices $\mathbf{T}_X$ and $\mathbf{T}_Y$, and afterward do an outer product between those scores. Applying a PCA to this scores outer product matrix will recover the scores on the original space, avoiding the usage of very wide vectors.

The next step is to recover the loadings in the original variable space. To perform this operation without very demanding computational operations, one takes advantage of the loadings calculated in the compressed space ($\mathbf{p}_{\mathrm{PCT}}$). Considering the left side of Eq. (3) where, by definition:

$$\mathbf{p}_{(1,m\cdot p)}^{\mathrm{T}} = \mathbf{p}_{\mathrm{PCT}(1,kX\cdot kY)}^{\mathrm{T}} \left(\mathbf{P}_{X(kX,m)}^{\mathrm{T}} \otimes \mathbf{P}_{Y(kY\cdot m)}^{\mathrm{T}}\right)_{(kX\cdot kY,m\cdot p)}$$

which transposing gives:

$$\mathbf{p}_{(m\cdot p,1)} = \left(\mathbf{P}_{X(m\cdot kX)} \otimes \mathbf{P}_{Y(p,kY)}\right)_{(m\cdot p,kX\cdot kY)} \mathbf{p}_{\mathrm{PCT}(kX\cdot kY,1)} \tag{11}$$

Then if one defines:

$$\mathrm{vec}\left(\mathbf{P}_{\mathrm{PCTa}(kY,kX)}\right) = \mathbf{p}_{\mathrm{PCTa}(kX\cdot kY,1)}$$

where the a subscript is one of the PCs extracted from the outer product $\mathbf{K}$ matrix, one can write Eq. (11) as:

$$\mathrm{vec}\left(\mathbf{P}_{a(p,m)}\right) = \left(\mathbf{P}_{X(m,kX)} \otimes \mathbf{P}_{Y(p,kY)}\right)_{(m\cdot p,kX\cdot kY)} \mathrm{vec}\left(\mathbf{P}_{\mathrm{PCT}a(kY,kX)}\right) \tag{12}$$

Table 1
High-level description for the OP-PCT-PCA algorithm

Step	Computation	Comments
1	**X**,**Y**	Input of **X** and **Y** matrices
2	$[\mathbf{T}_X, \mathbf{P}_X] \longleftarrow$ PCA(X)	Full rank PCA of **X**
3	$[\mathbf{T}_Y, \mathbf{P}_Y] \longleftarrow$ PCA(Y)	Full rank PCA of **Y**
4	$\mathbf{K}=\mathbf{OP}(\mathbf{T}_X, \mathbf{T}_Y)$	Outer product (OP) between $\mathbf{T}_X$ and $\mathbf{T}_Y$
5	$[\mathbf{T}, \mathbf{P}_{PCT}]=$PCA($\mathbf{K}$)	Full rank PCA of **K**. **T** represents the scores of the original space. PPCT represents the PCT loadings (compressed space)
6	for $a=1$:PC unfold $\mathbf{P}_{PCTa}$ $\mathbf{P}_a=\mathbf{P}_Y\,\mathbf{P}_{PCTa}\,\mathbf{P}_X^T$ end for	Rebuild each Principal Component's loadings (a) — Eq. (13)

One of the relations concerning the vec operator states that [23]:

$$\mathrm{vec}(\mathbf{AWB}) = (\mathbf{B}^T \otimes \mathbf{A})\mathrm{vec}(\mathbf{W})$$

If $\mathbf{A}=\mathbf{P}_Y$, $\mathbf{W}=\mathbf{P}_{PCTa}$ and $\mathbf{B}=\mathbf{P}_X^T$, then Eq. (12) can be reorganized as:

$$\mathrm{vec}(\mathbf{P}_Y\mathbf{P}_{PCTa}\mathbf{P}_X^T) = (\mathbf{P}_X \otimes \mathbf{P}_Y)\mathrm{vec}(\mathbf{P}_{PCTa})$$

and thus, comparing to Eq. (12), one has:

$$\mathbf{P}_{a(p,m)} = \mathbf{P}_{Y(p,kY)}\mathbf{P}_{PCTa(kY,kX)}\mathbf{P}^T_{X(kX,m)} \quad (13)$$

Eq. (12) shows that the loadings of the original space can be computed in a straightforward manner by simply pre-multiplying and post-multiplying the PCT loadings (compressed space) by the **Y** and **X** matrix loadings, respectively, avoiding, once more, the use of very wide vectors.

To summarize the steps needed to perform a PCA of an outer product matrix, a high level description of the OP-PCT-PCA algorithm is shown in Table 1. The algorithmic description shows that the calculation of the scores of the original space (steps 2 to 5) is independent of the calculation of the loadings in the original space (step 6). Moreover, as shown in step 6, each loadings matrix in the original space can be calculated independently — i.e. if needed, one just has to rebuild the desired loadings matrix in the original space. Finally, it must be noted that to perform the PCAs, be it in the form of OP-PCA or OP-PCT-PCA, one has used as input to the SVD procedure the matrix of the form $\mathbf{KK}^T$ or $\mathbf{K}^T\mathbf{K}$, depending of the case, in order to optimize the computation speed.

3. Experimental

3.1. Dataset 1 — FTIR and NMR data of cork samples

The dataset is composed of 20 cork samples for which the suberin content was determined [2]. For each sample, Fourier Transform Infrared spectroscopy (FTIR) and solid state ^{13}C Nuclear Magnetic Resonance (NMR) spectra were obtained. Both domains were combined by means of an outer product matrix [2] in order to analyse the relationships between the two domains. The obtained outer product matrix (FTIR Θ ^{13}C NMR) consisted of 20 objects by 450702 (882×511) variables.

3.2. Dataset 2 — MIR and NIR data of oils

72 samples of oils acquired both in the Mid Infrared (MIR) domain between 4050 and 700 cm^{-1} (1738 variables) and in the Near Infrared (NIR) domain between 6100 and 4000 cm^{-1} (1090 variables). Once again, both domains were combined by the outer product operation, giving an outer product matrix (MIRΘNIR) with 72 rows (objects) by (1738×1090) 1,894,420 (variables).

3.3. Algorithm instrumentation

The algorithm was developed in Microsoft® Visual C++ 2003 and the memory profiles were obtained using the Performance Monitor (Perfmon) program of Windows® XP whereas the time was obtained programmatically. All computations were done in a Pentium IV 3.4 GHz with 1 GByte of memory.

4. Results and discussion

The main objective of this section is to prove, experimentally, that applying the Principal Component Transform (PCT) framework to do a PCA on an outer product matrix, gives the same results as applying

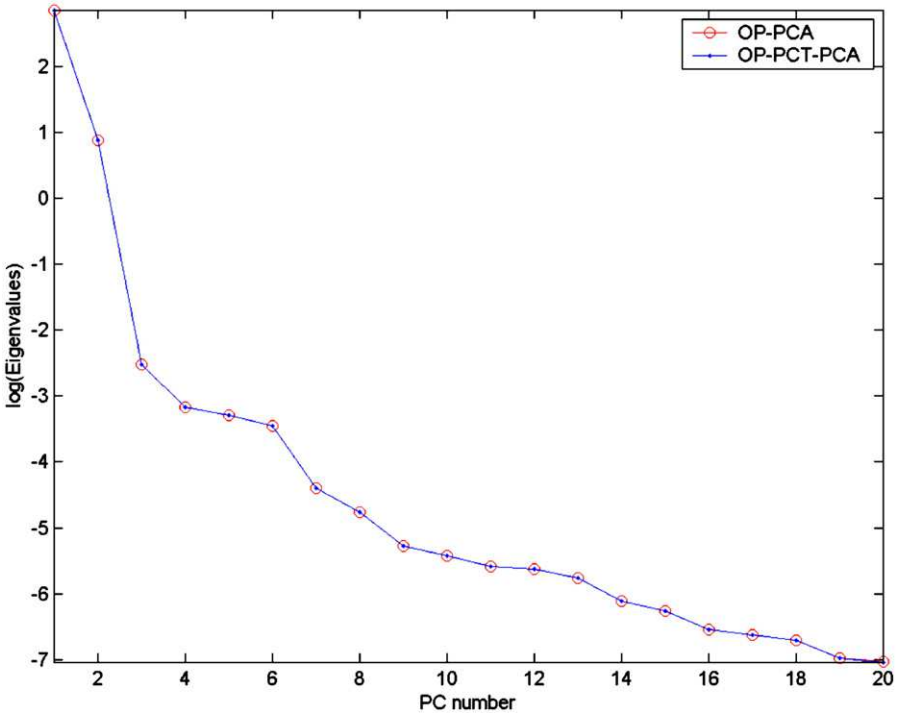


Fig. 1. Logarithm of eigenvalues of the 20 PCs from OP-PCA and OP-PCT-PCA of data set 2.

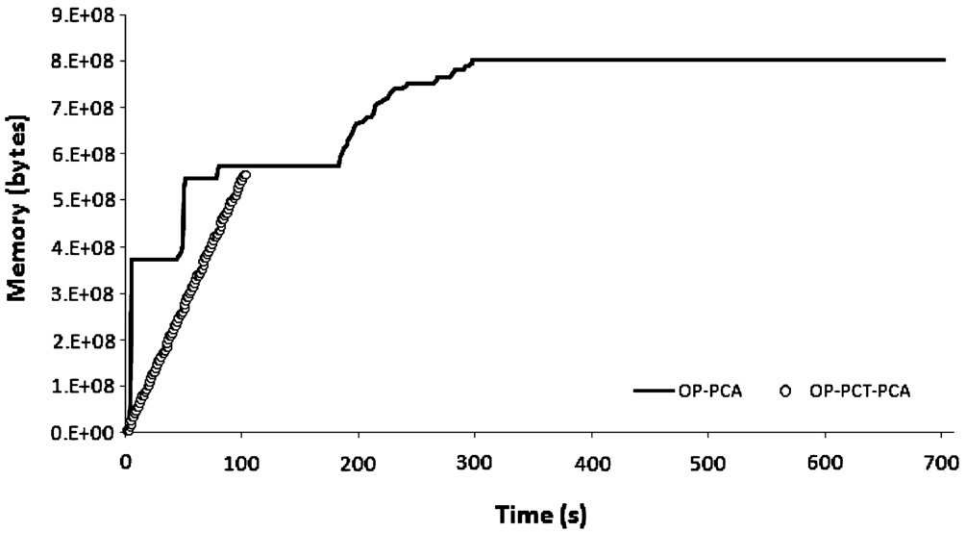


Fig. 2. Memory profile for OP-PCA and OP-PCT-PCA of dataset 2.

directly the PCA to the explicitly calculated outer product matrix, showing at the same time that the former approach is faster and uses less computer resources than the latter.

4.1. Data set 1 — FTIR & ^{13}C NMR

A PCA was applied to the outer product matrix representing the data fusion of FTIR and ^{13}C NMR of cork samples. The size of the explicitly built outer product matrix was 20 objects and 450,702 variables. The application of PCA to this outer product matrix, recovering 20 Principal Components (PCs), took around 5.2 s and used approximately 129 Mb of memory. On the other hand, using the proposed approach of PCT (OP-PCT-PCA), about 0.5 s and only 39 Mb of memory was needed, an approximately 10-fold decrease in time of computation and more than 3 times less memory. To assess the performance of the OP-PCT-PCA algorithm it was found that steps 1 to 5 (Table 1) – calculation of the core matrices (original space scores and PCT loadings) – required 11 ms, whereas, rebuilding all the 20 loadings in the original space (step 6 of Table 1), took 532 ms.

Fig. 1 shows the logarithm transformed eigenvalue profiles of both approaches to perform a PCA, showing that they are the same. In fact, and in order to compare the scores and loadings calculated using both approaches to perform PCA, the Root Mean Square Error (RMSE) was computed. An RMSE of 3×10^{-6} was found for the scores and 9×10^{-4} for the loadings, showing that both procedures gave virtually the same results, the differences being due to round-off errors.

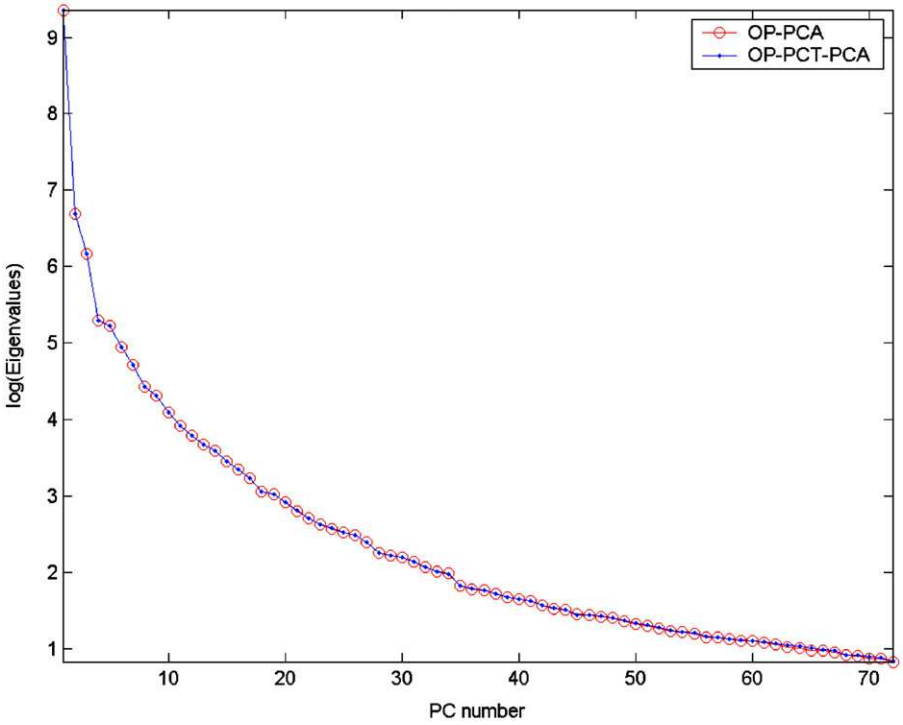


Fig. 3. Logarithm eigenvalues profiles of the 72 PCs from OP-PCA and OP-PCT-PCA of data set 2.

4.2. Dataset 2 — MIR and NIR data of oils

The application of PCA to the outer product matrix $\mathbf{K}_{(72,\ 1{,}894{,}420)}$ (OP-PCA) recovering 72 PCs, took approximately 12 min and required around 803 Mb of memory. The OP-PCT-PCA procedure was computed in less than 2 min and required around 556 Mb of total memory.

It is worth detailing the performance of the OP-PCT-PCA algorithm as it gives a chance to analyse the key points of this procedure. Actually, as described earlier, the OP-PCT-PCA calculates the scores and the loadings in the original space in an independent manner (Table 1). For that matter, the computation of the original scores (steps 1 to 5 of Table 1), took just 596 ms, using approximately 9 Mb of memory. On the other hand, to calculate the loadings in the original space (step 6 of Table 1) took around 1.5 s and about 8 Mb for the computation of each of the 72 loadings recovered. This is important, as one can compute just the desired loadings. In Fig. 2, the memory allocation profiles of both methods are shown. The OP-PCT-PCA memory allocation profile shows a steady increase of memory which mainly represents the allocation memory for each loadings calculation (72 matrices — step 6 of Table 1).

In Fig. 3, the almost identical logarithm transformed eigenvalue profile of both approaches is shown. In fact, the RMSE between both approaches was, for the scores, less than 10^{-6} and for the loadings less than 10^{-4}, once more, ascribed to round-off errors.

5. Conclusions

The results of this work demonstrate that the proposed approach for doing PCA on outer product matrices, defined as OP-PCT-PCA, is more efficient than to perform directly the PCA (OP-PCA). This was achieved by doing the outer product, not in the original space, where the resulting data fusion matrices can be very wide, but by doing it in a "compressed" space (the Principal Component Transform framework). The outer product in this "compressed" space is built by using the scores of each domain. Furthermore, it was shown that the calculations of the loadings in the original space are straightforward, and are done independently from the scores. This characteristic can be used to minimise even more the impact in terms of computer resources since one could recover only a small number of loadings, if desired. A summary of the computational resources for the application of the proposed method is shown in Table 2. On the other hand, and due to the fact that the scores and loadings can also be calculated independently as stated above, a summary of the computational resources needed to achieve this alternative way of applying the OP-PCT-PCA is shown in Table 3.

This proposed approach (OP-PCA-PCA) will allow the exploratory analysis of matrices obtained by OP fusion of very wide datasets.

Appendix A

Table 2
Summary of computational resources

Dataset	Method	Time	Memory (Mb)	RMSE		PCs
				Scores	Loadings	
1	OP-PCA	5.2 s	129	–	–	20
	OP-PCT-PCA	0.5 s	39	3×10^{-6}	9×10^{-4}	20
2	OP-PCA	12 min	803	–	–	72
	OP-PCT-PCA	2 min	556	10^{-6}	10^{-4}	72

Table 3
Summary of computational resources required independent calculation of scores and loadings using OP-PCT-PCA

Dataset	Table 1 steps	Time (ms)	Memory	Comments
1	1–5	11	~2 Kb	Scores and PCT loadings calculation
	6	532	~2 Mb	Computation of each original loadings
2	1–5	596	~9 Mb	Scores and PCT loadings calculation
	6	1500	~8 Mb	Computation of each original loadings

References

[1] A.S. Barros, M. Safar, M.F. Devaux, P. Robert, D. Bertrand, D.N. Rutledge, Appl. Spectrosc. 51 (1997) 1384–1393.
[2] M.H. Lopes, A.S. Barros, C. Pascoal Neto, D.N. Rutledge, I. Delgadillo, A.M. Gil, Biopolymers (Biospectroscopy) 62 (2001) 268–277.
[3] D.N. Rutledge, D.J.-R. Bouveresse, Chemom. Intell. Lab. Syst. 85 (2007) 170–178.
[4] J. Forshed, R. Stolt, H. Idborg, S.P. Jacobsson, Chemom. Intell. Lab. Syst. 85 (2007) 179–185.
[5] I.T. Jolliffe, Principal Component Analysis, Springer-Verlag, New York, 1986.
[6] E.R. Malinowski, D.G. Howery, Factor Analysis in Chemistry, J. Wiley & Sons, New York, 1980.
[7] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part A, Elsevier, The Netherlands, 1997, pp. 519–556.
[8] M. Partridge, R. Calvo, Intell. Data Anal. 2 (1998) 203–214.
[9] E. Oja, ICANN'95, Paris, 1995, pp. 89–94.
[10] A. Weingessel, K. Hornik, IEEE Trans. Neural Netw. 11 (2000) 1242–1250.
[11] R.B. Lehoucq, D.C. Sorensen, C. Yung, ARPACK users guide: Solution of large scale eigenvalues problems with implicity restarted Arnoldi methods. Technical report from http://www.caam.rice.edu/software/arpack; computational and applied mathematics, Rice University.
[12] G.W. Steward, SIAM J. Matrix Anal. Appl. 23 (2001) 601–614.
[13] G.W. Stewart, SIAM J. Matrix Anal. Appl. 24 (2002) 599–601.
[14] F. Vogt, M. Tacke, Chemom. Intell. Lab. Syst. 59 (2001) 1–18.
[15] F. Vogt, M. Tacke, J. Chemom. 16 (2002) 562–575.
[16] B.K. Alsberg, O.M. Kvalheim, Chemom. Intell. Lab. Syst. 24 (1994) 31–42.
[17] B.K. Alsberg, Chemom. Intell. Lab. Syst. 30 (1995) 223–225.
[18] H.A.L. Kiers, R.A. Harshman, Chemom. Intell. Lab. Syst. 36 (1997) 31–40.
[19] W. Wu, D.L. Massart, S. de Jong, Chemom. Intell. Lab. Syst. 36 (1997) 165–172.
[20] A.S. Barros, D.N. Rutledge, Chemom. Intell. Lab. Syst. 78 (2005) 125–137.
[21] A.S. Barros, D.N. Rutledge, Chemom. Intell. Lab. Syst. 73 (2004) 245–255.
[22] D.S. Burdick, Chemom. Intell. Lab. Syst. 28 (1995) 229–237.
[23] C.D. Meyer, Matrix analysis and applied linear algebra, SIAM, 2000.

ANNEXE V

Improving the detection of significant factors using ANOVA-PCA by selective reduction of residual variability

R.Climaco-Pinto[1,2] A.S.Barros[2] N.Locquet[3] L. Schmidtke[4] D.N.Rutledge[1,3]*

[1] Laboratoire de Chimie Analytique, AgroParisTech. 16, rue Claude Bernard. 75005 Paris, France.

rpinto@agroparistech.fr

[2] Departamento de Química, Universidade de Aveiro.Campus Universitário de Santiago. 3810-193 Aveiro, Portugal.

antonio.barros@ua.pt

[3] INRA/AgroParisTech UMR 214 "IAQA".16, rue Claude Bernard. 75005 Paris, France.

rutledge@agroparistech.fr

[4] National Wine and Grape Industry Centre, Charles Sturt University, Boorooma Street, Wagga Wagga, NSW 2650, Australia.

LSchmidtke@csu.edu.au

Keywords: ANOVA-PCA, ASCA, Error removal, Discrimination.

Abstract

Selective elimination of residual error can be used when applying Harrington's ANOVA-PCA in order to improve the capabilities of the method. ANOVA-PCA is sometimes unable to discriminate between levels of a factor when sources of high residual variability are present. In some cases this variability is not random, possesses some structure and is large enough to be responsible for the first principal components calculated by the PCA step in the ANOVA-PCA. This fact sometimes makes it impossible for the interesting variance to be in the first two PCA components. By using the proposed selective residuals elimination procedure, one may improve the ability of the method to detect significant factors as well as have an understanding of the different kinds of residual variance present in the data.

Two data sets are used to show how the method is used in order to iteratively detect variance associated with the factors even when it is not initially visible. A permutation method is used to confirm that the observed significance of the factors was not accidental.

1 Introduction

1.1 ANOVA-PCA (or APCA)

Analysis of variance – principal components analysis (ANOVA-PCA) has been used for the detection of biomarkers [1,2], to assess the stability of reference materials [3] and to evaluate the significance of factors of an experimental design, as well as for prediction of new samples [4]. This supervised method uses the ANOVA paradigm to create a series of matrices containing the means for the different levels of the main effects and interactions of the factors of an experimental design to which are added the residual errors. PCA is then applied to each of these mean plus error matrices in order to evaluate the significance of the effects against the residual error. As usual with PCA, scores and loadings are obtained, which may be used to study the existence of groupings of individuals and to evaluate the importance of the initial variables in the definition of the effects and the sources of residual variation and to compare it to the different factors in the experimental design.

It is clear from the above description that this procedure is not related to the ANOVA-based method that is often used to detect significant variables prior to a multivariate analysis such as PCA. It is in fact very similar to ASCA where similar matrices, but without the residual errors added back, are analyzed by Simultaneous Components Analysis [5, 6, 7]. To avoid confusion and underline this similarity, we prefer to use the term APCA throughout this paper, rather than ANOVA-PCA. The most important difference between APCA and ASCA is that with the latter method the multivariate analysis is performed on the matrices of level means of the factors without the residual errors having been added back, which means that it is necessary to use a resampling procedure such as bootstrapping in order to be able evaluate the significance of the factors in comparison to the residual error. With standard APCA, resampling is not necessary as the significance of the factors can be estimated by examining the scores plots. But, although resampling is not required by APCA, it may of course be applied in a similar way to gain further insight into the characteristics of the factors and samples.

Depending on the data being analyzed, problems may arise with APCA. One clear limitation of the method is when there is a large amount of structured residual variance, due for example to an interfering substance with specific absorbance peaks. The variability due to this interference may give rise to principal components with high variance, which will make it difficult to reach a conclusion. In the original APCA method, it was considered that if the first principal component is not due to the variability of the factor being tested, that factor is considered non-significant.

By eliminating part of the variability from the residual error matrix, it may be possible to make the spectra of the replicates more comparable so that the principal components which are related to the factor become the ones with largest variance, instead of those related to the more or less structured variability present in the residuals. By applying PCA to the residuals matrix calculated in the first step of the APCA, and eliminating a certain number of Principal Components, it is possible to selectively reduce the residual variability. Then, a less noisy version of the initial matrix, with reduced intra-sample variability, can be rebuilt and analyzed again by APCA.

A similar method where some of the variability between replicates is removed from the data by PCA to perform error removal by orthogonal subtraction (EROS) when using replicates has been presented in the literature [8].

1.2 APCA with selective residuals variance reduction

The ability of the APCA method may be improved in cases where there is only a separation of the scores of the samples for the levels of the Factor under consideration along PC2 or higher, indicative of a situation where the Factor is significant even though its variability is less than that of the residual errors.

The modification of the APCA method consists in rebuilding a series of increasingly less-noisy initial matrices by eliminating increasing numbers of Principal Components from the residual error matrix, thus reducing the inter-repetition variability. These less-noisy matrices are then analyzed by APCA to determine at what point it is possible to observe a separation of the scores for the samples belonging to different levels of the considered Factor.

To verify that the separation is not simply an artifact due to the elimination of too much residual error, the Factor levels are randomized and the APCA is repeated, in order to compare the results of the tested Factor levels with those of the random levels. If the distances between group centroids for the tested Factor levels are superior to the distances for the random levels, the Factor may be considered as significant compared to that reduced residual error. This procedure makes it easier to study the relationship between the information due to a Factor and the different parts of the residual variance.

The present work shows how the selective reduction of residual variance can help the standard APCA method to overcome one of its limitations. Two real datasets are used to illustrate the proposed procedure.

2 Theory

2.1 APCA

APCA can be seen as a supervised method to test whether a data matrix contains information related to the various Factors of an experimental design. Each of the samples (rows) of the matrix is attributed to a level for each of the Factors and Interactions of the design. APCA successively calculates a series of matrices corresponding to the means of the variables at each level of each Factor, and then subtracts them from the original matrix to give a final matrix of residual errors. In this matrix, the vector of residuals of each sample is different.

In the simple case of an experimental design with two Factors with j and k levels, this decomposition can be written in vector form as:

$$\mathbf{x}_i = \bar{\mathbf{x}} + \bar{\boldsymbol{\alpha}}_j + \bar{\boldsymbol{\beta}}_k + \overline{\boldsymbol{\alpha\beta}}_{jk} + \boldsymbol{\varepsilon}_i \quad (1)$$

$\mathbf{x}_i$ being the vector of responses for sample i, $\bar{\mathbf{x}}$ the global average vector for the whole data matrix $\mathbf{X}$ (of dimensions I x F). $\bar{\boldsymbol{\alpha}}_j$ and $\bar{\boldsymbol{\beta}}_k$ are the effects for Factors one and two respectively, $\overline{\boldsymbol{\alpha\beta}}_{jk}$ is the interaction between these Factors and $\boldsymbol{\varepsilon}_i$ the residuals for the vector of responses for the sample. The length of all these vectors is F.

The residuals matrix is then added back to each of the factor matrices and a PCA is performed on each one of these "factor + residuals" matrices and the scores and loadings are examined. The original APCA procedure only considered the first two PCs. Conclusions may be drawn concerning each Factor by examining the Hotelling T^2 [9] ellipses (at 95% confidence) for each level. Four situations may arise:

1. If there is a separation of the levels along PC1, the Factor is significant compared to the residual error.
2. If the separation is along PC2, there is information related to the Factor levels, but the noise variance contained in PC1 related to the residual error is greater.
3. If there is no separation in the first two components, it may mean that the data matrix contains no information related to the different levels of the Factor.
4. If there is no separation in the first two components, it may mean that although there is information in the data matrix related to the different levels of the Factor, the residual error variability is so much greater that it is only visible in later PCs.

In order to be able to decide between the situations 3 and 4, a method is proposed here to progressively reduce the part of the variance of the residual error matrix to be added back to the level means matrices. Once a separation is observed along PC1, its validity is verified by using a permutation method. The proposed method is therefore an intermediate along the continuum between APCA and ASCA.

2.2 Improved detection of significant factors by selective elimination of residual variability

In some situations the variability between replicates behaves like white noise, in which case simply averaging the replicates may reduce it sufficiently for the differences among different samples to become apparent. However, this is very often not the case, and it is then necessary to selectively eliminate certain parts of the variability. In the case of APCA, this may be done very simply by reducing the variance of the residual errors matrix.

The way the residual variability is extracted from the residuals matrix is succinctly described below, in equation (2). This intra-sample (or inter-replicates) variability is the residual error of APCA $\boldsymbol{\varepsilon}$ (of dimensions I x F), calculated according to equation (1). The residual variability may be progressively reduced by applying a PCA to $\boldsymbol{\varepsilon}$ and subtracting n PCs from the residuals, by using its scores **T** and eigenvectors **P** according to:

$$\boldsymbol{\varepsilon}_r = \boldsymbol{\varepsilon} - \mathbf{T}\mathbf{P}^T \tag{2}$$

where $\boldsymbol{\varepsilon}_r$ is the matrix of reduced residuals after elimination of the selected variance, T (of dimensions n x I) and **P** (of dimensions F x n) matrices composed of, respectively, the n selected PC scores and eigenvectors from the PCA on $\boldsymbol{\varepsilon}$.

A reduced residual error version of the initial data set can then be "rebuilt" to give $\tilde{\mathbf{X}}$ by summing all the matrices of the Factor levels and the matrix of reduced residuals, as in equation (4):

$$\tilde{\mathbf{X}} = \overline{\mathbf{X}} + \boldsymbol{\alpha} + \boldsymbol{\beta} + \boldsymbol{\alpha\beta} + \boldsymbol{\varepsilon}_r \tag{3}$$

where $\overline{\mathbf{X}}$ is a matrix in which all lines are equal to the global average, $\boldsymbol{\alpha}$ is a matrix in which each line is the average of the group to which the sample belongs for that Factor (the same for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha\beta}$) and $\boldsymbol{\varepsilon}_r$ is the matrix of reduced residuals.

2.3 APCA with selective reduction of residual variability

If no separation of the levels is achieved when performing the standard APCA method on the data, one starts to eliminate residual variability from the data. This is done iteratively by subtracting increasing proportions of the variance in $\boldsymbol{\varepsilon}$ using increasing numbers of PCs as $\mathbf{P}$ in equation (3) to then calculate $\tilde{\mathbf{X}}$ as in equation (4). The iterations are continued until a good separation of the groups of the levels of the desired Factor is attained when APCA is applied to $\tilde{\mathbf{X}}$. When the separation is attained for n PCs, the corresponding residuals $\boldsymbol{\varepsilon}_n$ will be retained for subsequent analyses.

In fact, whether a Factor is significant or not, a separation of the groups will eventually be found, because the residuals tend to zero as more Principal Components are eliminated. In the extreme case of completely eliminating the residual variability, it is certain that even small differences in level averages will give rise to a separation of the groups. To verify that the observed separation of the levels is not due to chance (just because the residuals are too small), as in ASCA, a resampling procedure based on permutation of group memberships is used [7].

2.4 Permutation procedure

The samples are randomly attributed to Factor levels and the level averages are calculated for this random group classification. The residuals previously calculated for the standard analysis ($\boldsymbol{\varepsilon}_n$) are added to the resulting matrix of Factor averages and the PCA is performed. The total or partial Mahalanobis [10] inter-group distances corresponding to each of the levels is calculated. The permutation procedure is repeated hundreds or thousands of times and the distances are added and averaged in order to obtain a sampling distribution. A histogram of all the distances is plotted, to be compared with the inter-group distance obtained the by the non-random APCA inter-group distances.

Two situations may be observed when examining the histogram of inter-group distances :

1. The randomly permuted levels are also separated with the tested matrix of residuals $\boldsymbol{\varepsilon}_n$, meaning that the Factor is not significant. So much variability has been deleted from the residuals using n PCs that even random averages separate the level groups.
2. The random level groups are not separated while the non-random ones are, in which case the Factor may be considered as significant.

The reason that the residuals added to both the normal and the random permutation models are those from the calculation of the normal classification of samples is that if a Factor is

significant, the residuals from the permutation models are necessarily larger than those of the standard method. So, by using the smaller residuals of the standard model, the probability of observing a significant separation of the random level groups is increased, thus reducing the probability of false positives, i.e. accepting as significant a Factor which is not.

3 Material and methods

Two datasets (herein named "Carraghenan" and "Wine") were used to demonstrate the method. Both data sets were acquired using a Fourier Transform Mid-Infrared Spectrometer (Bruker Vector 33) with a thermostated Attenuated Total Reflection (ATR) sampling device ("Golden Gate", Specac).

Carraghenan data: This data set has already been studied and described in detail elsewhere [4]. Mid-infrared spectra of carraghenan solutions were collected and 64 scans averaged over the region 4050-600 cm^{-1}, at 4 cm^{-1} resolution. The diamond ATR crystal was thermostated at the temperatures defined by the experimental design.

Factors: Concentration (1 and 2 %), Temperature (30, 40, 45, 50, 60 ºC) and Day (spectra measured on 2 different days by different operators). Each sample was measured in triplicate. 2x5x2x3 = 60 spectra in all.

Wine data: Mid-infrared spectra of model of red wines. 32 scans were collected and averaged over the region 4000-600 cm^{-1}, at 4 cm^{-1} resolution after 20 minutes of evaporation on the ATR crystal warmed at 70ºC.

Factors: Year (A, B, C), Oak chips addition (Yes, No), Micro-oxygenation (Yes, No). Three samples for each type of wine, each sample measured in triplicate. 3x2x2x3x3 = 108 spectra in all.

4 Results and discussion

4.1 Carraghenan data:

As the study of factor Temperature is not interesting in the framework of this method, because of the results already obtained, the study of this data set will focus on the factors Concentration and Day.

4.1.1 Factor Concentration

4.1.1.1 Standard APCA

In a previous article [4], the same set of mid-infrared spectra of carraghenan was studied using APCA. In that study, there was a large variance due to the CO_2 peak which gave rise to PC1 when studying the Factor Concentration. PC1 was therefore related to this peak and not to the Factor (Concentration) being studied, which appeared in PC2 (Figure 1). That CO_2 variance, of no interest for the study, was not random but structured and sufficiently large to create PC1. Based on the original philosophy of APCA, the tested factor would be considered as not significant compared to the residual error. When the CO_2 region at around 2350 cm^{-1} was deleted, the factor Concentration then became significant compared to the residual error. Deleting parts of the spectra is, however, not optimal as in some cases there may be information in that same region of the spectrum.

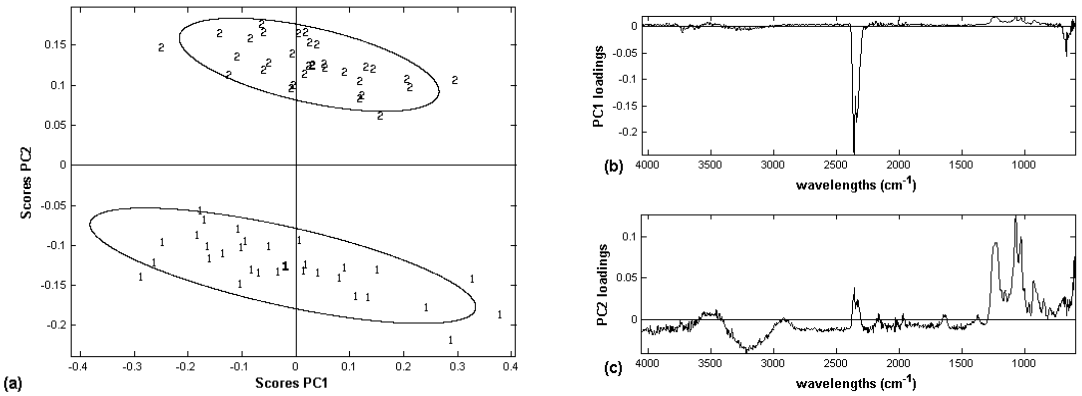


Figure 1: (a) APCA scores and (b, c) loadings using APCA on the initial data for factor Concentration. Although there is a separation of the two concentration levels, it is not on PC1, which is associated to the CO_2 present in the residuals.

4.1.1.2 Selective reduction of residual variability-APCA

The carraghenan data set was studied by selective reduction of residual variability before APCA, using the complete spectrum. The distances between the centroids of the two levels were calculated for the real and for the random groupings. The results presented in Table 1 show that the distance between the two groups is already slightly larger for the normal than for the random calculation, indicating that the variability that separates the two groups is due to a significant factor even though it is contributing to PC2. To investigate what would happen if part of the residual error matrix variability is eliminated, residual error reduction was applied. **T** and **P** were calculated using $n = 1$ and then APCA was applied to the resulting

matrix $\widetilde{\mathbf{X}}$. As can be seen in Figure 2, after this treatment, the contribution to the variability of the CO_2 peak centered at 2350cm^{-1} has almost completely disappeared.

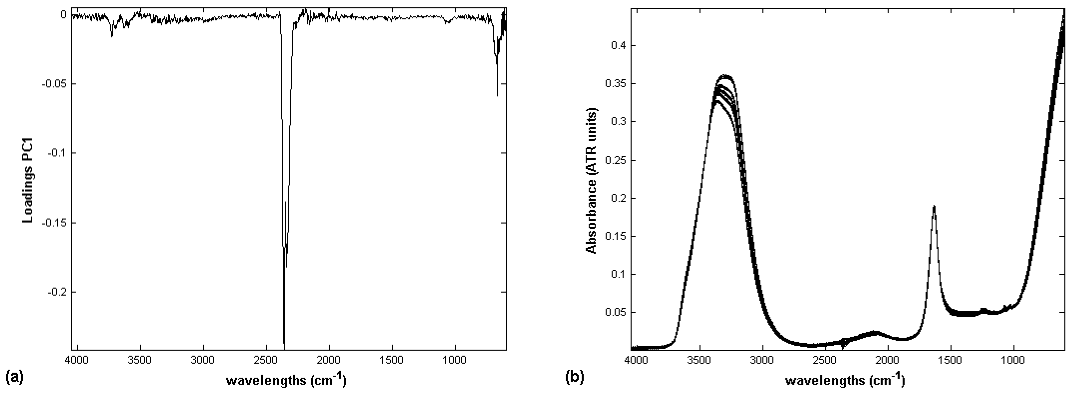


Figure 2: (a) Loadings of PC1 extracted from the residual variance matrix; (b) Resulting data matrix $\widetilde{\mathbf{X}}$ after eliminating PC1.

The concentration levels are now separated by the PC1 of APCA (Figure 3) and the distances between the real groups is much larger than between the random permutation groups, so the factor may now be considered as significant compared to the reduced residual error.

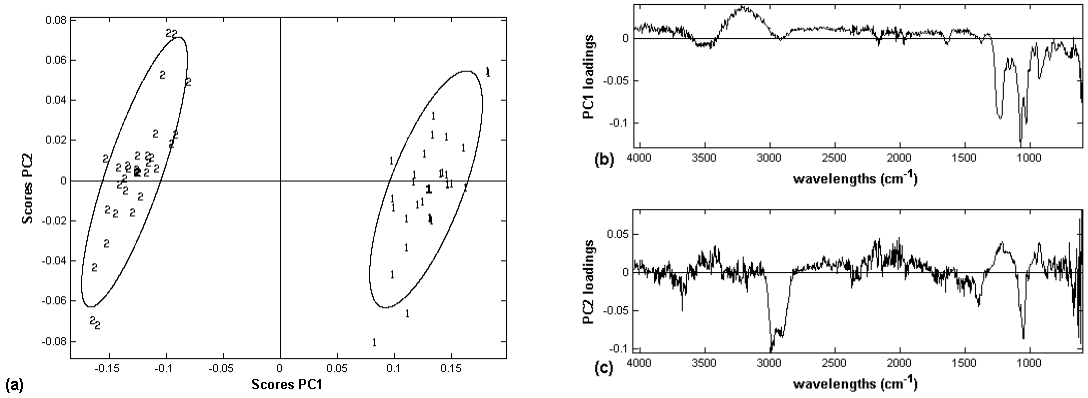


Figure 3: (a) APCA scores and (b) loadings using APCA for Factor Concentration of $\widetilde{\mathbf{X}}$ after eliminating PC1 from the residual variance matrix. This reduction in the residual variance has eliminated the structured CO_2 related variance leading to the Factor Concentration being significant compared to the reduced residual error.

The separation along PC1 having been achieved, the distance between the two groups is calculated. Also 500 random permutations are performed and APCA applied to the permutated data sets. The level averages and the distances histogram obtained using the initial

data are presented in Figure 4, for comparison with the ones of Figure 5, obtained after elimination of PC1 from the residual variance matrix.

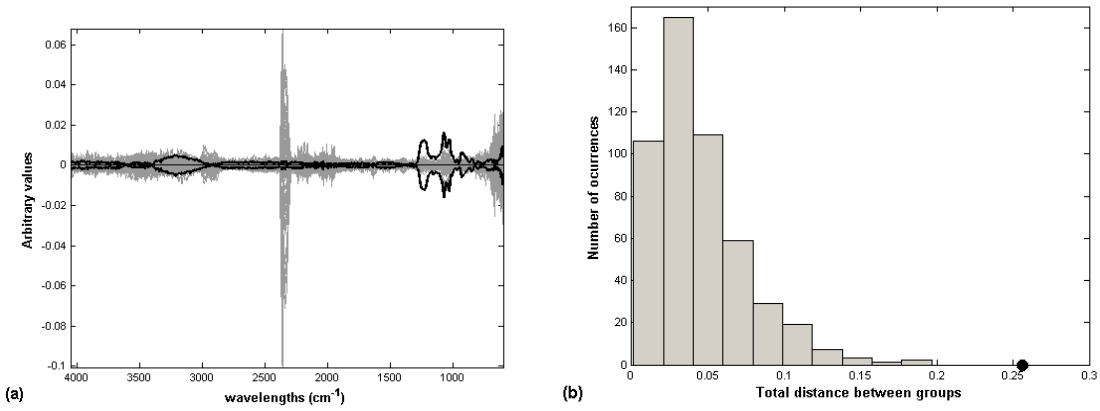


Figure 4: Data related to the results obtained in Figure 1, before eliminating any residual variance. (a) Averages of the levels for the normal data set **X** (black) and plot of residuals (grey). (b) Mahalanobis distances between the two levels for 500 random permutations (grey) and for the normal classification (black dot).

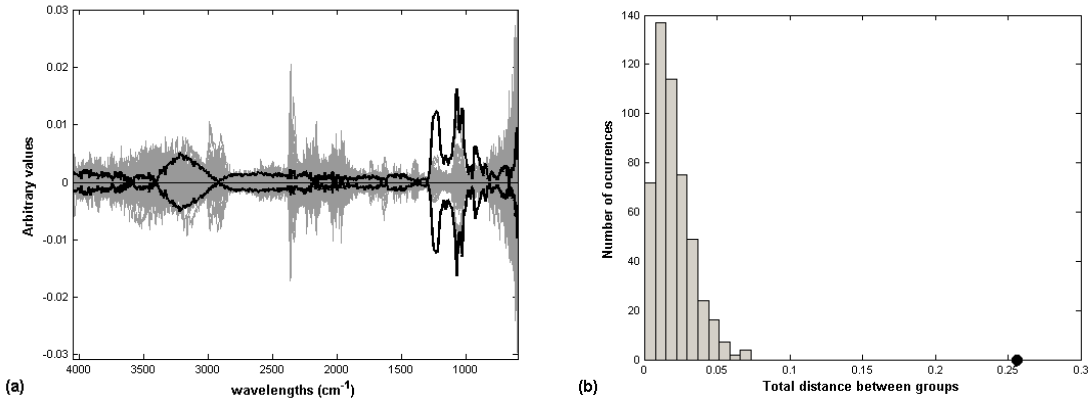


Figure 5: Data related to the results obtained in Figure 3, after eliminating PC1 from the residual variance matrix. (a) Averages for the concentration levels of the data set $\widetilde{\mathbf{X}}$ after eliminating PC1 from the residuals matrix, (black) and residuals (grey). (b) Mahalanobis distances for 500 random permutations (grey) and for the normal classification (black dot).

The elimination of just one component from the residuals matrix is enough to provide the separation of the two concentration levels along PC1 of the APCA.

Although the real distances between the groups decreases slightly when PC1 is eliminated (Table 1), the random distances decreases much more due to smaller residuals, amplifying the relative distance between the real and the random distances.

Table 1: Mahalanobis distances between the 2 group centroids on PC1 and PC2 of the Factor Concentration.

Number of PCs eliminated from **ε**	Maximum random distance(500 iterations)	Real distance
0	0.197	0.256
1	0.074	0.256
60 (all)	0.060	0.254

It is interesting to note that if a Factor is significant compared to the residual error, its loadings hardly change even though the separation of the levels increases with increasing number of PCs eliminated from the residual variance. Also interesting to note is that the standard normal variates (SNV) pre-treatment was applied to the data after the residual variance was extracted, not before. This can make a significant difference for the case when different types of residual variability are present in the data.

4.1.2 Factor Day

4.1.2.1 Standard APCA

Spectra were acquired on two different days, when different ATR crystal cleaning procedures were used, resulting in different quantities of residual ethanol in the system. The idea was to check if the different cleaning procedures would give different results and to compare the residual variance for each of the procedures.

Looking at Figure 6 one can see that the two days are not well separated along PC1 (which is again due to CO_2) and not completely separated along PC2, which would appear to be due to ethanol and to a baseline shift in the right-hand side of the spectra. Figure 7 shows that the distance between the two days is comparable to the distance when groups are attributed randomly.

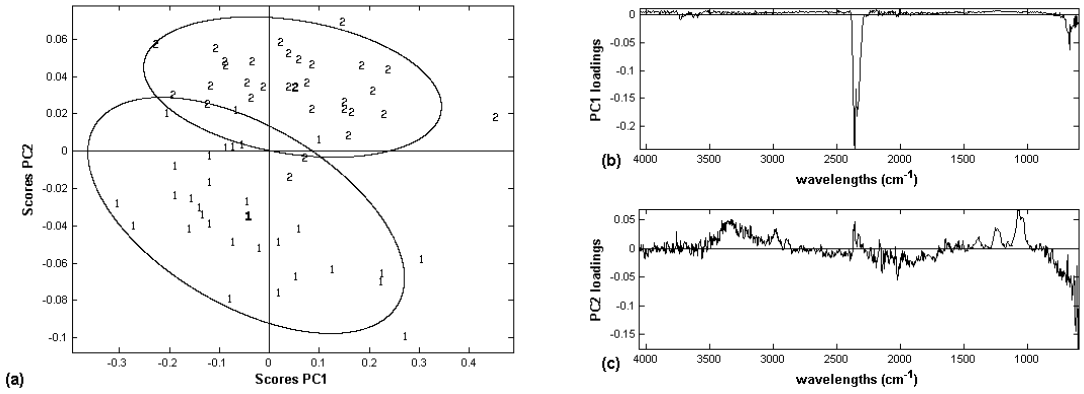


Figure 6: (a) APCA scores and (b, c) loadings for the factor Day on the initial Carraghenan dataset **X**. PC1 is related to CO_2 and does not separate the groups, while there is some separation along PC2.

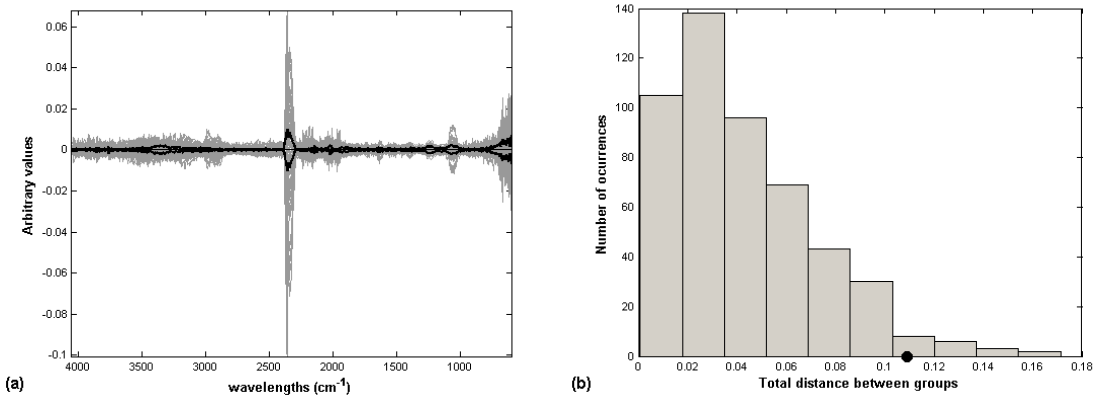


Figure 7: (a) Averages for the two days when random (grey) and normal classifications (black) are calculated on the initial data set, **X**. (b) Mahalanobis distances for 500 group permutations (grey) and for the normal case (black dot).

4.1.2.2 Selective reduction of residual variability-APCA

The procedure for selective reduction of residual variability is applied to the data and PC1 is removed. Then APCA is applied to the $\tilde{\mathbf{X}}$ matrix. As can be seen in, Figure 8b, c , the separation of the two days along PC1 seems to be due to a combination of the spectral features present in the previous PC1 (atmospheric CO_2) and PC2 (ethanol near 3300, 1000 – 1200 cm^{-1} and baseline shift). PC2 reflects differences in the variability of the quantity of ethanol present in the spectra on the two days. In Figure 8a, the variability within day 1 along PC2 is much greater than that for day 2, which is explained by the way the crystal was washed, using much more ethanol, and its loadings resembles the ethanol vapor spectrum as shown in a previous article [4]. Figure 9b*Figure 12* shows that the distance between the centroids for the two days is larger than that due to the randomly permuted levels.

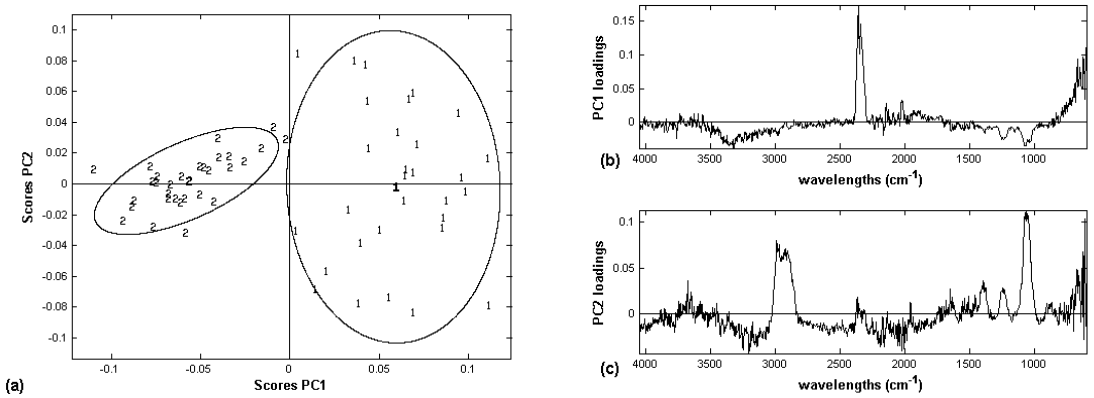


Figure 8: (a) APCA scores and (b) loadings for the factor Day on the Carraghenan data set $\widetilde{\mathbf{X}}$ after eliminating PC1.

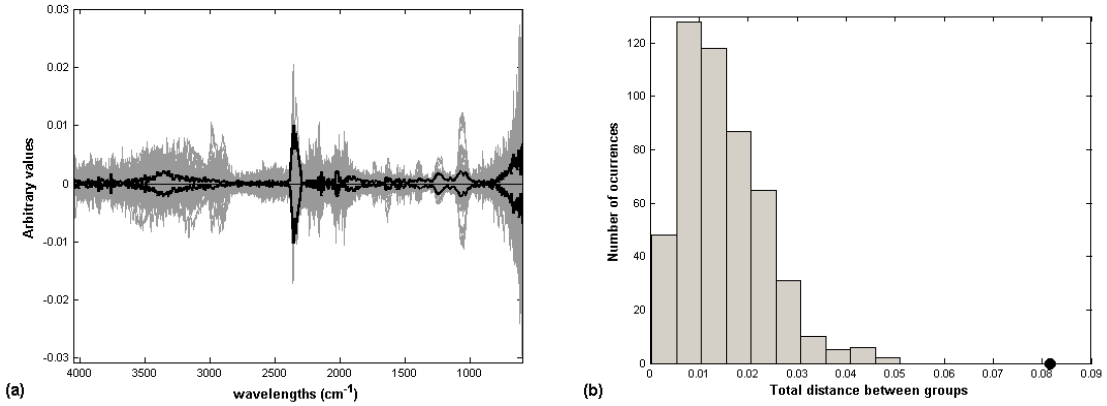


Figure 9: (a) Averages for the two days when random (grey) and normal classifications (black) are calculated on the Carraghenan data set $\widetilde{\mathbf{X}}$ after eliminating PC1. (b) Mahalanobis distances for 500 group permutations (grey) and for the normal one (black dot).

The results obtained after eliminating increasing number of PCs from the residuals matrix are presented in Table 2. This table shows that after the elimination of PC1, the distance between the centroids of the two days is greater for the normal calculation than for the randomly permuted ones. The real and the maximum random distances do not change significantly with increasing amount of variance eliminated and are more or less stable after eliminating PC1 from the residual variance.

Table 2: Mahalanobis distances between the 2 centroids due to Factor Day. 500 iterations were calculated.

Number of PCs eliminated from ε	Maximum random distance(500 iterations)	Real distance
0	0.171	0.109
1	0.051	0.082
60 (all)	0.053	0.078

4.2 Wine data:

In the previous data set, the elimination of part of the residual variance is determinant to change the results for the significance of the tested factors. In the following example this effect is even more evident. The initial data and after column-centering are presented in Figure 10. There is a large quantity of residual variance in the fingerprint region (far right) of the spectra as the reproducibility of the evaporation method was not very good.

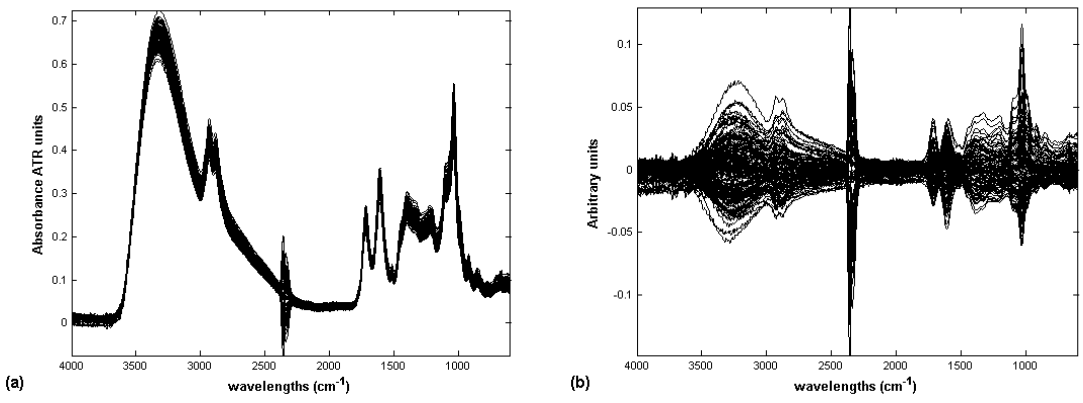


Figure 10: (a) Initial wine raw data and (b) after column-centring

4.2.1 Factor Year

4.2.1.1 Standard APCA

The APCA analysis for the whole spectrum does not show any separation of the levels for the Factor Year (Figure 11a and Figure *12*b). There is again a strong influence of atmospheric

CO_2 in the spectra (Figure 11b), which accounts for most of the variability of the spectra and so is present in PC1. In PC2 there is no real separation of the levels for any of the Factor levels. Examining the loadings in Figure 11, one can have a better understanding of the type of residual variance and note the regions which interfere in the detection of the interesting factors.

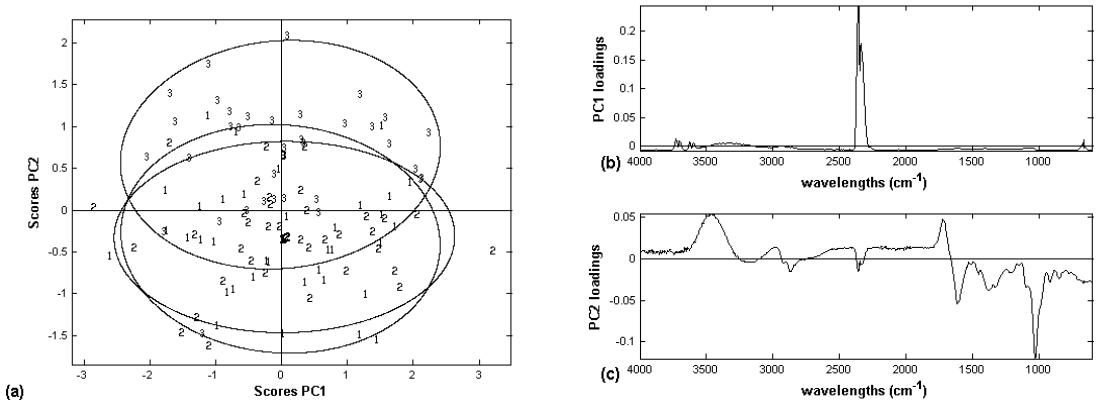


Figure 11: (a) Scores and (b, c) loadings for the Factor Year following APCA on the initial wine data.

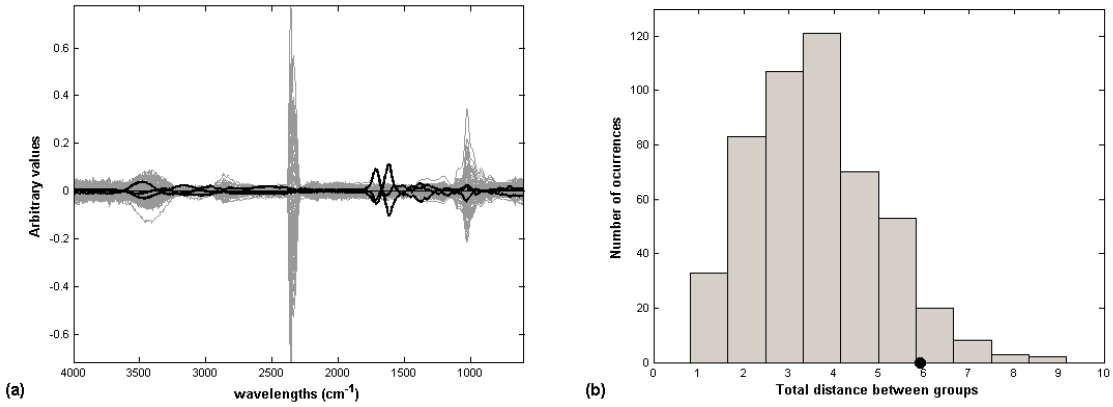


Figure 12: (a) Averages for the three years when random (grey) and normal classifications (black) are calculated on the wine data set. (b) Mahalanobis distances for 500 group permutations (grey) and for the normal one (black dot). There is no separation of Year with the initial data using the standard APCA method.

The results above would indicate that it is necessary to test the effect of eliminating part of the residual error on APCA. Principal Components are eliminated successively from the intra-sample residual error matrix, until the APCA results show a separation of the factor levels.

4.2.1.2 Selective reduction of residual variability-APCA

After eliminating PC1

The elimination of PC1 from the residual variance leads to a slight separation of the 3 years, as can be seen in Figure 13. This separation is mainly along PC2, but includes some variation also in PC1. There is still a large amount of residual variability in the fingerprint region of the spectra (Figure 14b,c), which contributes to the observed results.

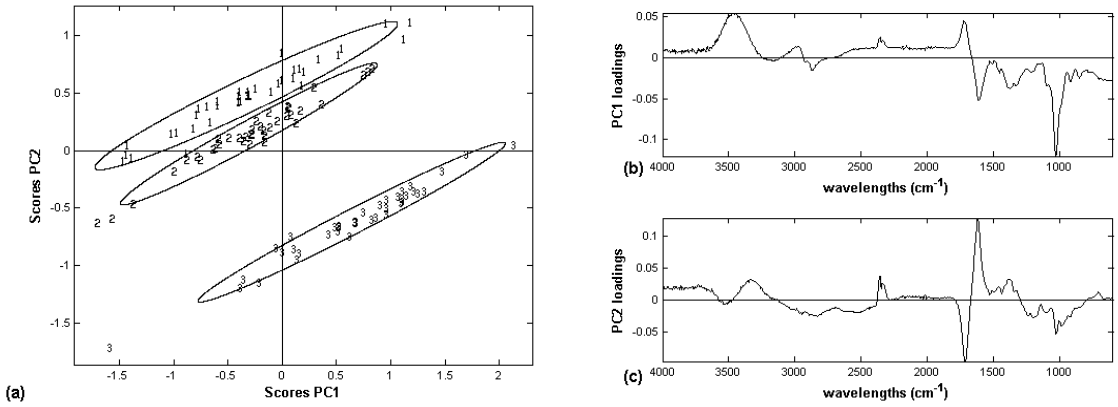


Figure 13: (a) Scores and (b, c) loadings for APCA after eliminating PC1.

Although the separation of the groups is not along PC1, the real distance between the three levels is already larger than the random distances (Figure 14b), which means that the Factor is significant. As this separation is due to both PC1 and PC2 (Figure 14a), it is not clear which spectral features are responsible for it. It is therefore of interest to continue reducing residual variance in order to obtain a better separation of the groups along PC1 so as to understand better the origin of the residual variability still present in the data.

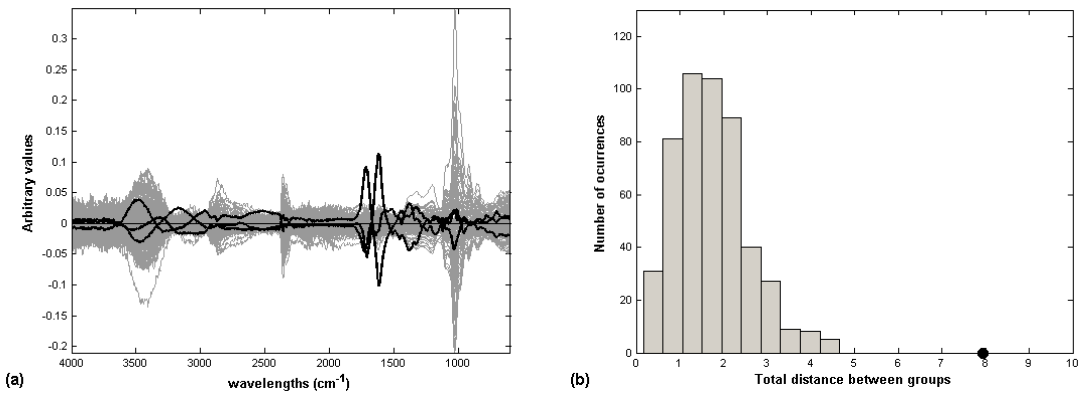


Figure 14: (a) Averages for the three years when random (grey) and normal classifications (black) are calculated on the wine data set $\tilde{\mathbf{X}}$ after eliminating PC1. (b) Mahalanobis distances for 500 group permutations (grey) and for the normal one (black dot).

After elimination of PCs 1-2

After the elimination of the first two Principal Components of the residual variance matrix, a much better separation of the groups is obtained, as can be seen in Figure 15. The 3 groups are separated by both PCs, which may have an interesting spectral interpretation. There may be two types of effects in the wines due to the Factor Year:

- the conditions inherent to each year of harvest (climate, growing conditions, etc.) which result in physico-chemical differences in the wines;

- an effect due to ageing of the wine, even with appropriate storage, between vinification and analysis.

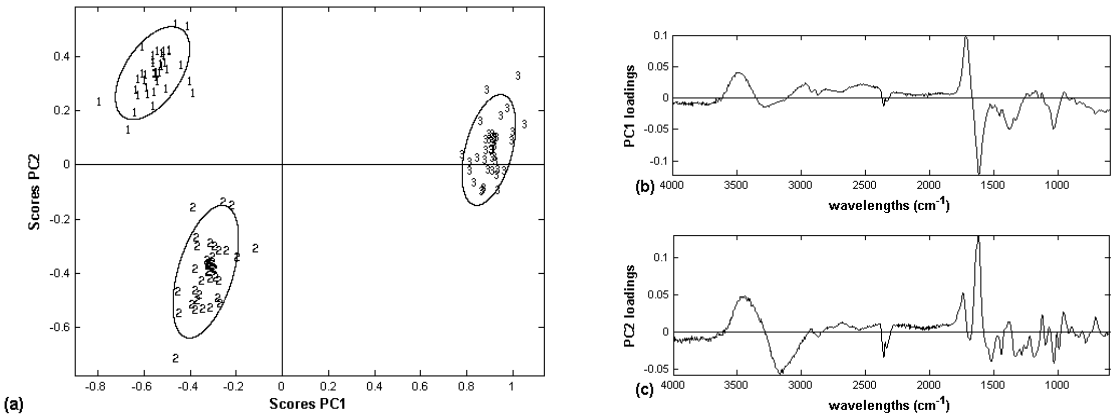


Figure 15: (a) Scores and (b, c) loadings for APCA on the Factor Year after eliminating the first two PCs.

In Figure 16a it can be seen that the residual variability was greatly reduced in relation to the real averages. The consequence of this is that the relative difference between real and random distances becomes much larger (Figure 16b).

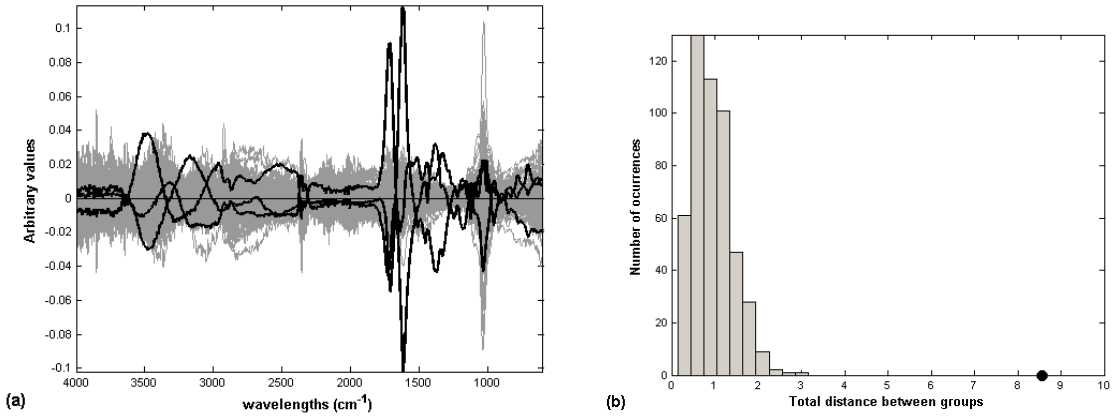


Figure 16: (a) Averages for the three years when random (grey) and normal classifications (black) are calculated on the wine data set $\tilde{\mathbf{X}}$ after eliminating PC1-2. (b) Mahalanobis distances for 500 group permutations (grey) and for the normal one (black dot).

After randomization, the total inter-group Mahalanobis distances obtained are shown in Table 3 where it can be seen that the results do not change substantially after the first two PCs are eliminated from the residual variation.

Table 3: Mahalanobis distances between the 3 centroids due to factor Year

Number of PCs eliminated from ε	**Maximum random distance(500 iterations)**	**Real distance**
0	8.083	5.914
1	4.658	7.952
2	3.163	8.555
108 (all)	2.882	8.544

4.2.2 Factor Oak

It was not possible to separate the two levels for the Factor Oak using the standard APCA method. Therefore, residual variance was eliminated according to the procedure proposed until a separation was obtained. This required removing the first three PCs from the residual variance matrix.

After extracting PCs1-3

The residual variance matrix for the Factor Oak is by definition the same as that for the Factor Year. The variability eliminated in PC1-3 for the Factor Oak is similar to that eliminated by PC1-2 for the Factor Year, i.e. mainly related to CO_2 and to variability in the fingerprint region of the spectra.

Once a separation was obtained (Figure 17a), it was necessary to test the significance of the Factor by calculating the distances for the permutations. Those distances are presented below and, although only 19 out of 1000 permutations gave values higher than the real distance (Figure 18b), the Factor Oak can not be considered as significant as the Factor Year, probably because of the presence of the CO_2 peak (Figure 17b).

This highlights a difficulty inherent to APCA and which is not addressed by any treatment of the residual error matrix. If an interfering spectral feature is not uniformly present at the different Factor levels, it will contribute to the average vectors and so to the Factor matrix, instead of to the residual error matrix. Such is the case for this data set, where the CO_2 peak varied uniformly over the levels for Year, but not for Oak.

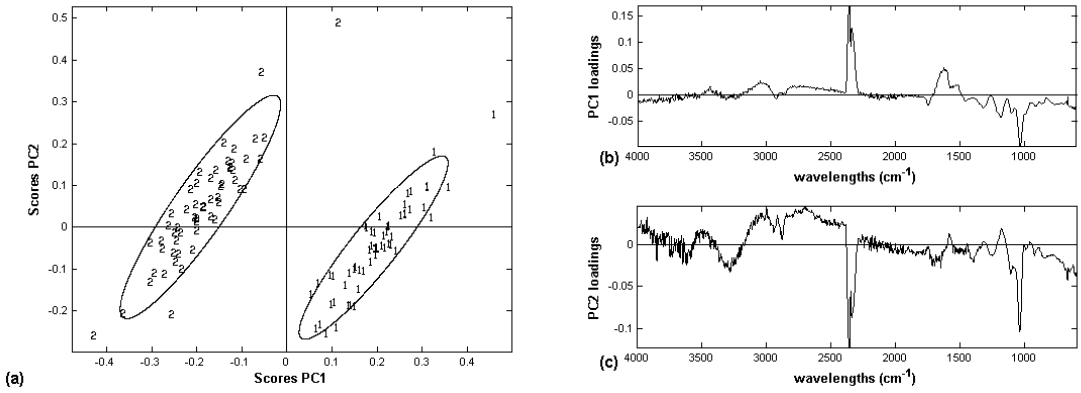


Figure 17: (a) Scores and (b,c) loadings for APCA in the Factor Oak after eliminating PCs 1-3

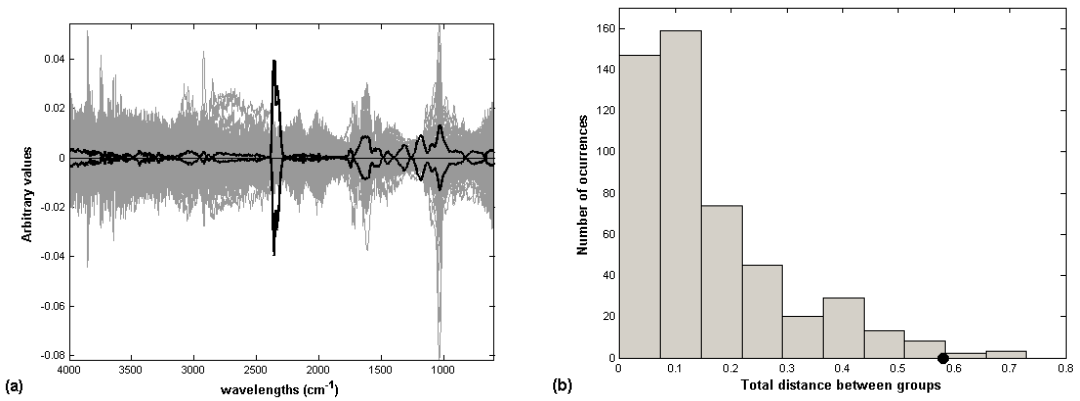


Figure 18: (a) Averages for the two Oak levels when random (grey) and normal classifications (black) are calculated on the wine data set $\tilde{\mathbf{X}}$ after eliminating the first three PCs. (b) Mahalanobis distances for 500 group permutations (grey) and for the normal one (black).

Table 4: Mahalanobis distances between the 2 centroids due to the Factor Oak levels

Number of PCs eliminated from $\boldsymbol{\varepsilon}$	Maximum random distance(500 iterations)	Real distance
0	3.015	1.223
3	0.730	0.580
108 (all)	0.973	0.572

4.2.3 Factor Micro-oxygenation (Mox)

The analysis of the Factor Mox was performed in the same way as for the Factor Oak. Here, 5 PCs had to be eliminated in order to have a separation of the levels, but the real distance between these factors was much less than the randomly calculated distances, indicating that the Factor Mox is not significant. The distances for different numbers of eliminated PCs are present in Table 5.

Table 5: Mahalanobis distances between the 2 centroids due to factor Mox

Number of PCs eliminated from $\boldsymbol{\varepsilon}$	Maximum random distance (500iterations)	Real distance
0	2.848	0.186
4	1.016	0.272
108 (all)	0.796	0.265

5 Conclusion

The procedure for the iterative use of residual error reduction with APCA is shown. PCA is used to eliminate successive numbers of principal components from the residual error, or intra-sample variance matrix, and APCA is applied to the reconstituted data matrix to evaluate the significance of the Factors compared to the reduced residual error. A permutation

procedure was used to verify the validity of the separation of groups. The procedure enables one to understand in more detail the sources of residual variation within the data.

In the first data set studied, Carragheenan, one of the Factors, Concentration, was related to PC2. By using the proposed procedure, it was possible to reduce the contribution of the intra-sample variance so that the tested factor is represented by PC1, and so is significant in comparison to the residuals. Similarly, the factor Day was not considered as significant when using the initial data but after elimination of PC1 from the residual variance matrix, it also became significant.

As for the second data set, Wine, there was no separation of the factor levels on PCs 1 or 2 using standard APCA. After applying the proposed sequential residual variability reduction, the Factor Year became significant, as confirmed by the permutation procedure.

The other two Factors, Oak and Mox, were not significant compared to the reduced residual error using the whole spectral region.

The method presented here improves the detection capabilities of APCA and the study of the sources of variance that may interfere in the results.

One advantage of the method is that it is possible eliminate some sources of residual variability prior pre-treatment of the spectra, which can make a difference in the results.

Acknowledgements

Rui Clímaco Pinto acknowledges a PhD. grant from the "Fundação para a Ciência e Tecnologia" (FCT) - Portugal

References

[1] P. Harrington, N. Vieira, J. Espinoza, J. Nien, R. Romero, A. Yergey, *Anal. Chim. Acta*, 544 (2005) 118-127,

[2] P. Harrington, N. Vieira, P. Chen, J. Espinoza, J. Nien, R. Romero, A. Yergey *Chemometr. Intell. Lab. Syst.*, 82 (2006) 283-293.

[3] J. Sarembaud, R. Pinto, D.N. Rutledge, M. Feinberg, *Anal. Chim. Acta*, 603 (2007) 147-154.

[4] R. Climaco-Pinto, V. Bosc, H. Noçairi, A.S. Barros, D.N. Rutledge, Anal. Chim. Acta, 629 (2008) 47-55.

[5] A.K. Smilde, J.J. Jansen, H.C.J. Hoefsloot, R-J.A. N. Lamers, J.Greef, M.E. Timmerman, Bioinformatics 21 (13) (2005) 4043-3048.

[6] J.J.Jansen, H.C.J. Hoefsloot, J.Greef, M.E.Timmerman, J.A. Westerhuis, A.K.Smilde, J. Chemometr. 19 (9) (2005), 469-481, 2005.

[7] D.J.Vis, J.A.Westerhuis, A.K.Smilde, J.Greef, BMC Bioinform. 8 (2007) 322.

[8] Y.Zhu, T.Fearn, D.Samuel, A.Dhar, O.Hameed, S.G.Bown, L.B.Lovat; J. Chemometrics 22 (2008) 130-134.

[9] H.A. Hotelling, Generalized T test and measure of multivariate dispersion, in: Second Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, (1951), 23–41.

[10] P.C. Mahalanobis, On the generalised distance in statistics, in: Proceedings of the National Institute of Science of India, vol. 12, (1936), 49–55.

ANNEXE VI

A multi-block extension of the APCA procedure for the identification of significant factors

D. Jouan-Rimbaud Bouveresse[1], R. Climaco Pinto[2,3], L.M. Schmidtke[4], N. Locquet[1,2], D.N. Rutledge*[1,2]

1) INRA/AgroParisTech, UMR 214 "IAQA" 16, rue Claude Bernard, 75005, Paris, Franc.

2) Laboratoire de Chimie Analytique, AgroParisTech. 16, rue Claude Bernard. 75005 Paris, France

3) Departamento de Química, Universidade de Aveiro Campus Universitário de Santiago. 3810-193 Aveiro, Portugal

4) National Wine and Grape Industry Center, School of Agriculture and Wine Sciences, Charles Sturt University, Wagga Wagga, NSW 2650, Australia

ABSTRACT

A modification of the ANOVA-PCA method, proposed by Harrington *et al.* to identify significant factors and Interactions in an experimental design, is presented in this article. The modified method uses the idea of multiple table analysis, and looks for the common dimensions underlying the different data tables generated by the "ANOVA-step" of the ANOVA-PCA method, in order to identify the significant factors. In this paper, the "Common Component and Specific Weights Analysis" method is used to analyse the calculated Multiblock data set. This new method, which was called AComDim, was compared to the standard ANOVA-PCA method, by analysing four real data sets, and was often found to be more sensitive than this latter method, either by better separating some groups (less overlap), or even by separating groups which were not separated by ANOVA-PCA.

KEYWORDS

Common Component and Specific Weights Analysis, ComDim; ANOVA-PCA

*Corresponding author. Tel.: +33 1 44 08 16 39. Fax: +33 1 44 08 16 53.
E-mail address: douglas.rutledge@agroparistech.fr

1. INTRODUCTION

Several multi-block analysis procedures exist for the simultaneous study of multiple sets of matrices with different variables describing the same samples (for example, see [1-4]). These methods are very often used in sensometry for the analysis of sensory data. For example, a series of samples can be judged by a number of assessors, each using several (possibly different) characteristics to describe them. These methods may also be useful in chemometrics to combine information about the same set of samples contained in signals acquired using different techniques (IR spectroscopy; Raman spectroscopy; physico-chemical analyses; etc.). One such multi-block technique is "Common Component and Specific Weights Analysis" - CCSWA [5]. The objective of CCSWA is to find the directions describing common distributions of the samples in the spaces defined by the different data blocks (hence the name *Common Component*, abbreviated CC). A parameter, called the *salience*, indicates the importance of each block in the construction of the common dimension, and a "percentage of explanation" of each dimension can be computed. "ComDim", the particular implementation of CCSWA used in this work, was developed and coded in Matlab [6] by D. Bertrand [7].
The work presented in this article shows that an interesting extension of ComDim (and possibly of other multi-block techniques) is to use it in the analysis of sets of tables calculated from a single initial data matrix. One such application, AComDim, is based on replacing the many separate PCAs performed in the ANOVA-PCA method [8], also abbreviated APCA, by a single analysis using ComDim. In this case, the various "Factor matrices" and "Interaction matrices" calculated from the initial data matrix are all analysed simultaneously, resulting in a series of "Common Components" along which the samples are distributed, each associated with a vector of "saliences" reflecting the importance of the contribution of each data block to the "Common Component".

After a brief presentation of both the ComDim and APCA methods, this article will present several real case studies, showing the interest of this new method, particularly compared to the standard APCA method.

2. THEORY

2.1. Notation

Matrices will be denoted by bold uppercase letters (*e.g.*, **X**), column vectors will be denoted by bold lowercase letters (*e.g.*, **u**), and row vectors by bold lowercase letters followed by the uppercase symbol T (*e.g.*, $\mathbf{u}^T$), standing for "transposed". Scalars will be indicated by a letter in italics (*e.g.*, *N* or *n*).

2.2. ComDim [5, 9-10]

The main idea of the Common Dimension procedure, ComDim, is to calculate a weighted sum of the sample variance-covariance matrix of each table, and then extract its first (normed) Principal Component, or "common dimension" (also called "Common Component", CC).
The algorithm then iteratively finds the weight of each table in the previously calculated CC. Finally, the percentage of variability explained by the CC can be calculated. After the computation of the first CC, each original table matrix is deflated, and the procedure is repeated for the calculation of the second Common Component , and so forth. Each Common Component is the first PC of a weighted sum of deflated matrices.
In order to present ComDim from an algorithmic point of view, one assumes a set of n samples is described by p sets of different variables. Hence, p matrices $\mathbf{X}_i$ of sizes $n \times k_i$ ($i = 1$ to p) are available, for which one wants to determine the common components. One first column-centers each matrix (to obtain $\mathbf{X}_{ic}$), and then divides it by its Frobenius norm (to obtain the scaled matrix $\mathbf{X}_{is}$). For each $\mathbf{X}_{is}$, a matrix $\mathbf{W}_i$ of dimensions $n \times n$ can be computed as:

$$\mathbf{W}_i = \mathbf{X}_{is} \cdot \mathbf{X}_{is}^T \qquad (1)$$

The common components are computed in an iterative fashion. At each iteration, the weighted sum of the p $\mathbf{W}_i$ matrices is computed, resulting in a global $\mathbf{W}_G$ matrix. In the first iteration, all the weights λ_i are set to 1.

$\mathbf{W}_G$=0;
for i = 1 to p
    $\lambda_i = 1$;
    $\mathbf{W}_G = \mathbf{W}_G + \lambda_i \cdot \mathbf{W}_i$ (2)
end

$\mathbf{W_G}$ is then decomposed by singular value decomposition (SVD), yielding $\mathbf{U_W}$ (matrix of row-singular vectors), $\mathbf{S_W}$ (square matrix of zeros, except for the singular values on the diagonal), and $\mathbf{V_W}$ (matrix of column-singular vectors):

$$\mathbf{W_G} = \mathbf{U_W} \cdot \mathbf{S_W} \cdot \mathbf{V_W}^T \quad (3)$$

The first column of $\mathbf{U_W}$ (*i.e.*, the normed score vector of $\mathbf{W_G}$ associated with the largest singular value) is chosen as the first estimation of the "common component score" of $\mathbf{W_G}$, denoted as $\mathbf{q}$. A new estimation of λ_i is calculated using $\mathbf{q}$ and $\mathbf{W_G}$ and an *unfit* value is then determined as a function of the updated λ_i values:

$$unexpl = p$$

(p being the total variance of the p normalised data tables before calculating any CCs).

unfit=0;
for i = 1 to p
$\quad \lambda_i = \mathbf{q}^T \cdot \mathbf{W_i} \cdot \mathbf{q}$ (4)
$\quad \mathbf{Aux} = \mathbf{W_i} - \lambda_i \times \mathbf{q} \cdot \mathbf{q}^T$ (5)
$\quad$ for j = 1 to n
$\quad\quad$ for k = 1 to n
$\quad\quad\quad unfit = unfit + \mathbf{Aux}(j,k)*\mathbf{Aux}(j,k)$ (6)
$\quad\quad$ end
$\quad$ end
end

Aux is a "residuals" matrix of the variability unaccounted for by the Common Components calculated up to that point. The *unfit* value is the variance of all the tables unexplained by all those CCs.

The calculation of $\mathbf{W_G}$ (from the updated λ_i values), then $\mathbf{q}$ (after SVD of the updated $\mathbf{W_G}$), and then λ_i is iterated as in (Eqs. 4-6) until convergence of *unfit*. The final $\mathbf{q}$ vector is the first common component (its elements being equivalent to score values). The final λ_i value

indicates the weight of the original $\mathbf{X_i}$ in the common component ("salience"). A percentage of variance contained in the common components is given by :

$$expl = 100 \times (unexpl - unfit°) / p \quad (7)$$

unexpl is then updated as:

$$unexpl = unfit \quad (8)$$

Each $\mathbf{X_{is}}$ data matrix is then "deflated" (Eqs. 9-10), new estimations of $\mathbf{W_i}$ computed from these "deflated" $\mathbf{X_{is}}$ matrices, and the next common components computed as before.

$$\mathbf{Aux} = \mathbf{I} - \mathbf{q} \,.\, \mathbf{q}^T \quad (9)$$

(where **I** is the $n \times n$ Identity matrix)

$$\mathbf{X_{is}} = \mathbf{Aux} \,.\, \mathbf{X_{is}} \quad (10)$$

2.3. APCA [8]

Analysis of variance – principal component analysis (APCA) was introduced in 2005 by Harrington *et al.* [8] for the detection of biomarkers in high dimensional proteomic data sets. Since then, it has been applied in several other situations [11-13]. The aim of the method is to determine whether known characteristics of the samples (or "factors", in the Experimental Design terminology) produce a variation in the data which is significantly larger than the variations due to noise. A clear explanation of the method is given in [13]. To briefly summarise the method here, one assumes that f factors, each described by l_f levels, are known for matrix **X**. First, **X** is column-centered, yielding $\mathbf{X_c}$. A "Factor 1 matrix" $\mathbf{M_1}$ is created from $\mathbf{X_c}$, by replacing each row by the average vector of all rows whose level for factor 1 is the same as the row being replaced. Hence $\mathbf{M_1}$ is constituted by l_1 replicated rows. A first residual matrix can then be computed as:

$$\mathbf{X_{R1}} = \mathbf{X_c} - \mathbf{M_1} \quad (11)$$

A "Factor 2 matrix" $\mathbf{M_2}$ is then created from $\mathbf{X_{R1}}$, by replacing each row by the average vector of all rows whose level for factor 2 is the same as the row being replaced. $\mathbf{X_{R2}}$ is then computed as:

$$\mathbf{X_{R2}} = \mathbf{X_{R1}} - \mathbf{M_2} \qquad (12)$$

This procedure is repeated for each factor.

After the calculation of all $\mathbf{M_i}$ matrices, the 2-factor "Interaction matrices" $\mathbf{M_{ij}}$ (i,j = 1 to *f*), as well as higher order interactions, are computed in a similar fashion.

After subtracting all these factor and interaction matrices, a final residual matrix remains, denoted as **R**, which is then added to each $\mathbf{M_i}$ and $\mathbf{M_{ij}}$ matrix, yielding $\mathbf{M_{iR}}$ and $\mathbf{M_{ijR}}$ (for i,j = 1 to *f*). These matrices contain variability due to the differences between the averages for each factor level and noise. A graphical representation of the APCA procedure is presented in Figure 1.

Figure 1

A PCA analysis of each $\mathbf{M_{iR}}$ and $\mathbf{M_{ijR}}$ is then performed. Harrington proposed that if the factor *i* (or the interaction of factors *i* and *j*) is significant compared to noise, its variation will be described by PC1, and the residual variation will be described by PC2. However, Sarembaud *et al.* [12] showed that in many cases, although the factor does separate the samples as a function of the factor levels, its variability is less than that of the noise and so the sample groups are not distributed along PC1 but along a later PC.

APCA has the advantage over the more commonly used Simultaneous Components Analysis (ASCA) [14, 15, 16] in that with the latter method the PCA is done on the Factor matrices *without* the residual errors having been added back, which means that it is necessary to use a resampling procedure such as bootstrapping in order to be able evaluate the significance of the factors in comparison to the residual error. With APCA, the significance of the factors can be estimated by examining the scores plots.

2.4. AComDim

The idea of AComDim is to replace the several PCAs in the second step of APCA by a single ComDim analysis. In this method, instead of performing a PCA separately on each mean plus residuals matrix, $\mathbf{M_{iR}}$ and $\mathbf{M_{ijR}}$, the ComDim procedure is applied to all of them, as well as the residuals matrix, simultaneously, in order to find the underlying common components.

By examining the calculated saliences, it is possible to determine which Common Component is related to which factor. By examining the 2D plots of the scores associated with that Common Component against the scores of the Common Component related to the residual error matrix, it is possible to evaluate the significance of the factor.

In this paper, the procedure is applied to a number of data sets (see section 3), and has produced interesting results, which are compared to the results obtained with APCA. All these results will be described and commented in section 4.

3. EXPERIMENTAL

3.1. The Data

Apple data

This dataset was based on experiments organized in Norwich (UK) from 24/02/1999 to 25/2/1999 within the framework of the European Concerted Action "ASTEQ".

Two apple cultivars (*Cox* and *Jonagold*) at three different maturity levels (fresh, ripe, and over-ripe) are studied. Duplicate reflectance spectra were measured on each intact fruit using a spectrophotometer (Optical Spectrum Analyser (OSA) 6602, Rees Instruments Ltd., Goldalming, UK) in a 0°/45°-configuration, where the bundled detecting fibres and the bundled source fibres were placed in a black holder (type 6151) at an angle of 45°. The light source (Dual Light Source, type 6290) consists of a 12 V/100 W Tungsten halogen source of the type Philips 7724.M/28. The spectrum (380 - 2000 nm) was collected in two parts. A Si-detector (type 6611, 380 - 1080 nm) and an InGaAs-detector (type 6614, 1080 - 2000 nm): each measures one part of the whole spectrum and the two spectra are concatenated at 1080 nm. Every spectrum was divided by a reference spectrum taken on a $BaSO_4$-plate. Each reflection spectrum was the average of 5 individual optical scans from 380 to 2000 nm with 7.5 nm increments. SNV transformation was applied to the 94 spectra before analysis. Both the red side and the green side of the apples were measured.

Lignin data

Time-Domain (TD-) NMR relaxation curves of starch-lignin mixtures [15] are studied. After pre-extrusion of starch in presence of water, different amounts of lignin were added to some samples, leading to 5 concentrations (0%, 5%, 10%, 15% and 30%). From these 5 mixtures, canes were produced, out of which films were obtained by compression, yielding $2 \times 5 = 10$

samples. These samples were maintained at two relative humidity levels, by keeping them above a saturated solution of either $MgCl_2$ (humidity level (HL) = 33%) or NaCl (humidity level = 75%), yielding 20 samples. Therefore, each sample can be described by 3 characteristics, namely humidity level, form (cane / film), and lignin content. 20 different samples were produced (5 concentrations × 2 humidity levels × 2 sample forms), and duplicate TD-NMR measurements were done on each sample.

Wine data

Wine samples from three consecutive vintages were subject to oak chip maceration and/or micro-oxygenation with each treatment performed in triplicate with appropriate controls. Thus the number of wines in this study is 3 × 2 × 2 × 3 = 36 (vintage × mOx × Oak × replicates). Further details are given in ref. [16]. The wine samples were measured by fluorescence and Mid-InfraRed (MIR) spectroscopies:

1) Fluorescence data: Synchronous front-face fluorescence spectra of the 36 wine samples were measured on a Xenius spectrofluorometer (Safas), with a Xenon lamp. The analysed samples are placed in a quartz cuvette with two optical faces. The excitation wavelength (λ_{ex}) range varies from 250 to 650 nm, and the emission wavelengths are equal to $\lambda_{ex} + \Delta\lambda$, $\Delta\lambda$ varying between 20 and 200 nm with a step of 4 nm (resulting in 46 synchronous spectra for each sample).

2) MIR data [17]: The 36 wine samples were measured in triplicates on a Vector 33 (Bruker) Fourier-Transform mid-infrared spectrometer (resulting in 108 spectra). 32 scans were collected and averaged over the region 2071-600 cm^{-1}, at 4 cm^{-1} resolution after 20 minutes of evaporation on a Golden Gate ATR crystal (Specac) warmed at 70°C.

3.2. Software

All computations were performed using Matlab 7.6.0 (R2008a) [6], and in-house code available from the authors upon request. The ComDim procedure was adapted from the free toolbox SAISIR [7].

4. Results and Discussion

In all the case studies presented in this section, 10 common dimensions are computed. The reason for this is that this number was sufficient to be sure that all important sources of

variation in the data would be taken into account, and as will be seen in the different cases, the 10th CC explains so little variability that it did not seem necessary to take more CCs into account in the models.

4.1. Apple data

The first step of both AComDim and APCA is the "ANOVA-step", which yields the following matrices:

- Block 1 = $\mathbf{M_{1R}}$ = Factor "Variety" + Residuals matrix
- Block 2 = $\mathbf{M_{2R}}$ = Factor "Colour of the face measured" + Residuals matrix
- Block 3 = $\mathbf{M_{3R}}$ = Factor "Maturity" + Residuals matrix
- Block 4 = $\mathbf{M_{12R}}$ = Interaction Variety × Colour + Residuals matrix
- Block 5 = $\mathbf{M_{13R}}$ = Interaction Variety × Maturity + Residuals matrix
- Block 6 = $\mathbf{M_{23R}}$ = Interaction Colour × Maturity + Residuals matrix
- Block 7 = $\mathbf{M_{123R}}$ = Interaction Variety × Colour × Maturity + Residuals matrix
- Block 8 = $\mathbf{R}$ = Residuals matrix

AComDim

The percentage of variance contained in each of the ten dimensions is given in Table 1.

Table 1

One can notice that the common components are not exactly sorted in decreasing values of the percentage of variance (as is the case with Principal Components Analysis).

Figure 2 shows the salience values of these blocks on each Common Component. CC1 was considered as representing the residual noise as its saliences are high and almost equal for all blocks (except for blocks 2, 3, and to a lesser extent block 1), indicating that no block predominates the others in the computation of the common dimension, and hence no grouping can be observed on the scores. Since all blocks contain a contribution from the residual error matrix, it is to be expected that the first Common Component would be due to the added residual error. This would explain why no groupings can be observed in its score plots.

On the other hand, CCs 2, 4, 7, 8 and 10 have one salience value significantly larger than all other ones, and are therefore highly influenced by that block, which one expects to be the block defining groups which can be observed in the one-dimensional score plots.

Figure 2

Figure 3 presents the scores of the samples on CC 1 *versus* the scores on CC 2, 4, 7, 8 and 10.

Figure 3

The study of the "significant" CCs is as follows:

Common Component 2 contains 10% of the variability. The largest salience value is that of $\mathbf{M_{3R}}$, and as can be seen on Figure 3a, the scores of the samples are sorted according to their maturity level, the freshest samples having the smallest score values, while the "Over-ripe" samples have the largest scores. One can note that the intermediate maturity level (which is called "Ripe") has larger scores than the "Fresh" samples, but is relatively overlapped with the "Over-ripe" level, which is due to the difficulty of obtaining in practice a given maturity level.

As for Common Component 4, which contains 7.5% of the variability, its largest salience corresponds to $\mathbf{M_{2R}}$, and, the scores of the samples on this dimension are related to the colour of the face being measured by the NIR spectrometer, the higher scores corresponding to the green face, and the lower scores to the red face.

Common Component 7 is clearly related to the Variety of the apples as its largest salience value is for the corresponding matrix, $\mathbf{M_{1R}}$. The higher scores are for *Cox* apples, while the lower scores are for *Jonagold* apples. This Common Component only contains 0.14% of the variability present in the data blocks.

Common Component 8 (0.01% of the variability) is largely due to $\mathbf{M_{12R}}$ (largest salience), which corresponds to the Interaction between the Variety and the Face measured. Figure 3d, which represents the scores on Common Component 8, shows that the lower values correspond to samples {*Cox*-Green + *Jonagold*-Red}, while the higher values correspond to samples {*Cox*-Red + *Jonagold*-Green}.

Common Component 10, (0.004% of the variability) is highly influenced by $\mathbf{M_{23R}}$, which corresponds to the interaction Face × Maturity. As can be seen on Figure 3e, the scores increase with Maturity for samples whose red face is measured, while when the green face is measured, the scores decrease with increasing maturity.

To summarise the interpretation of applying AComDim to the Apple data, one can say that this method helps distinguish each main Effect (Variety, Colour of the face measured, and Maturity level), and two interaction effects (Variety × Side and Colour × Maturity).

Comparison with APCA

The APCA method was also applied to the Apple data and the PC1-PC2 score plots obtained after PCA of $\mathbf{M_{1R}}$, $\mathbf{M_{2R}}$ and $\mathbf{M_{3R}}$ are respectively presented in Figures 4a to 4c.

Figure 4

The two apples varieties are not well separated along PC1 (Figure 4a), which was not the case with AComDim (Figure 3c). As to the colour of the measured face, the two groups which can

be seen on the PC1-PC2 score plot of Figure 4b overlap whereas with AComDim (Figure 3b), the two groups were well separated. The three levels of maturity are also not well separated (Figure 4c), and although the Fresh apples do not overlap with the Over-ripe apples, there is a slight overlap between Fresh and Ripe samples, which was not the case with AComDim (Figure 3a). Hence, the comparison between APCA and AComDim on the first level of this case study shows that AComDim outperforms APCA.

The study of the second-order interactions shows that none of them could be distinguished with APCA, whereas AComDim separated of groups of samples with the Variety × Side and Colour × Maturity interactions, which is not illogical given the significant separations observed for the corresponding Main Effects.

The comparison of these two methods on the Apple Data has shown that in this case, AComDim outperforms APCA.

4.2. Lignin Data

The AComDim and APCA procedures were applied to the Lignin data. First, the "ANOVA-step" yields the 8 following matrices:

- Block 1 = $\mathbf{M_{1R}}$ = Factor "Moisture" matrix + Residuals matrix
- Block 2 = $\mathbf{M_{2R}}$ = Factor "Shape" matrix + Residuals matrix
- Block 3 = $\mathbf{M_{3R}}$ = Factor "Lignin concentration" matrix + Residuals matrix
- Block 4 = $\mathbf{M_{12R}}$ = Interaction "Moisture × Shape" + Residuals matrix
- Block 5 = $\mathbf{M_{13R}}$ = Interaction "Moisture × Lignin" + Residuals matrix
- Block 6 = $\mathbf{M_{23R}}$ = Interaction "Shape × Lignin" + Residuals matrix
- Block 7 = $\mathbf{M_{123R}}$ = Interaction "Moisture × Shape × Lignin" + Residuals matrix
- Block 8 = $\mathbf{R}$ = Residuals matrix

The AComDim procedure was first applied to these data blocks. The salience values for each dimension are presented on Figure 5.

Figure 5

From these saliences, it seems that the important blocks, hence the important characteristics on each dimensions are as follows:

- The humidity level shows a high salience on CC2 (20.73% of the variability), and the shape of the sample on CC3 (11.79% of the variability). On these two CCs, the interaction Moisture × Shape has the second largest value

- The lignin concentration in the samples is important for CC5 (4.45% of variability).
- The second-order interaction Shape × Lignin has the highest salience value on CC6 (0.68% variability), while the interaction Moisture × Lignin seems significant on CC7 (0.23% of variability).

Similarly to the previous case study, all informative CCs are plotted *versus* CC1 representing the noise. These score plots were compared to the PC1-PC2 scores plots obtained by the APCA method. All these plots are presented on Figure 6.

Figure 6

Both AComDim and APCA discriminate the samples according to the moisture level (Figures 6 a and b), although AComDim has the additional advantage that within each moisture level cluster, a second discrimination according to the samples shape is obtained. The same occurs for the discrimination of the samples according to the samples shape (Figures 6 c and d) with AComDim, each Shape-cluster can be split into two sub-clusters related to the humidity level. The discrimination of the samples according to the lignin concentration is also possible with both methods. However, with APCA, the 0% and 5% groups overlap more than with AComDim.

Concerning the two-factor interactions, with AComDim, the Moisture level × Shape interaction was detected and discussed earlier on CC2 and CC3. Figure 6g presents the PC1-PC2-PC3 three-dimensional score plot obtained by applying APCA, where two main groups of samples are distinguished, corresponding to the humidity level. Within each group, two subgroups corresponding to the form of the samples are separated. Therefore, contrarily to the assumptions of APCA, the group distinction requires more than PC1 to be clearly identified. As to the other interactions (Moisture × Lignin, Shape × Lignin, Moisture × Shape ×Lignin), neither AComDim nor APCA gives any separation.

4.3. Wine Data

The AComDim and APCA procedures were applied to both Wine data sets. First, the "ANOVA-step" yields the 8 following matrices:

- Block 1 = $\mathbf{M_{1R}}$ = Factor "Year" matrix + Residuals matrix
- Block 2 = $\mathbf{M_{2R}}$ = Factor "mOx" matrix + Residuals matrix
- Block 3 = $\mathbf{M_{3R}}$ = Factor "Oak" matrix + Residuals matrix
- Block 4 = $\mathbf{M_{12R}}$ = Interaction Year × mOx + Residuals matrix
- Block 5 = $\mathbf{M_{13R}}$ = Interaction Year × Oak + Residuals matrix
- Block 6 = $\mathbf{M_{23R}}$ = Interaction mOx × Oak + Residuals matrix

- Block 7 = $\mathbf{M_{123R}}$ = Interaction Year × mOx × Oak + Residuals matrix
- Block 8 = **R** = Residuals matrix

Fluorescence Data

In this case, each sample is described by an excitation-emission matrix (EEM). Therefore, the data set is a three-way array, of dimension 36 × 100 × 46. In order to apply AComDim and APCA, the array is unfolded to give a matrix (two-dimensional array), of dimensions 37 × 4600.

The saliences obtained on the 10 dimensions are presented on Figure 7.

Figure 7

Clearly, the interesting dimensions are dimensions number 2, 3, 4, 8 and 9. The corresponding score plot CC1 *vs* CC_i (i = 2, 3, 4, 8, 9) were investigated, and one could see that:

(i) The scores on CC2 (11.9 % of the variability) and on CC9 (0.12% of explained variability) discriminate the samples according to the Year. The CC9 vs CC2 score plot is presented on Figure 8a.

Figure 8

Both CC2 and CC9 discriminate the samples according to the same factor. This probably means that different parts in the EEM are related to the factor Year. In order to verify this, the AComDim loadings on dimensions 2 and 9 were computed. The calculation of the "loadings" is done by multiplying the original block by the corresponding score matrix. As the data were unfolded before analysis, unfolded loading vectors are obtained, which can be folded back to give an EEM-like surface. The surfaces of loadings on CC2 and CC9 are presented on Figure 8b, and clearly, they represent different peaks of the fluorescence EEMs. These peaks represent different polyphenols contained in the wine. The samples discrimination along CC2 is mainly due to differences in flavonolic compounds, while on CC9, anthocyanins are responsible for the separation.

(ii) The scores on CC4 (0.86% of variability) discriminate the samples according to the presence or absence of Oak chips;

(iii) the scores on CC3 (0.81% of the variability) should group the samples according to their value in the Year × Oak interaction. However, the CC1 *vs* CC3 score plot (Figure 8c), does not lead to this conclusion in a straightforward way. In fact, although one cannot say that two groups can be distinguished, it is interesting to note that the negative scores are those of samples of group {2004 - No Oak ∪ 2005 – Oak ∪ 2006 - Oak}, while the positive scores correspond to samples of the group {2004 – Oak ∪ 2005 – No Oak ∪ 2006 – No Oak}. If one

considers only years 2005 and 2006, one can see that samples with positive scores correspond to samples with no addition of oak chips, while samples with negative scores correspond to samples with addition of oak chips. In fact, samples collected in 2004 are spread all over CC3, while the other samples are more "compact".

(iv) The scores on CC8 should be influenced by the 3-factors interaction, as indicated by the saliences on this dimension. This interaction has 12 levels. However, on the score plot (Figure 8d), one can see that the situation is not as evident. The different data points are labelled according to their level. Two main groups can be distinguished, one with positive score values and one with negative score values. Within each of these groups, subgroups can be seen:

- $\{1 \cup 4\}$
- $\{6 \cup 7 \cup 10 \cup 11\}$
- $\{5 \cup 8 \cup 9 \cup 12\}$
- $\{2 \cup 3\}$

The extreme samples on this plot (1, 2, 3 and 4) correspond to 2004. This year put apart, the two other groups correspond to samples with either {micro-Oxygenation and no oak chip addition} or {no micro-oxygenation and oak chip addition} (positive scores), or to samples with either both {micro-oxygenation and oak chip addition} or {no micro-oxygenation and no oak chip addition} (negative scores).

The score plots obtained by APCA presented in Figures 9a, 9b and 9c (blocks $\mathbf{M_{1R}}$, $\mathbf{M_{3R}}$ and $\mathbf{M_{13R}}$ respectively), give some sample discrimination according to the investigated factor (or interaction).

Figure 9

The year of production introduces the most variability in the data, thus splitting the samples into three groups along PC1. However, the separation of the samples according to the year can also be seen on PC2, which must then correspond to a variability source different from just noise (containing more than 10% of the variability). The loadings on PC1 and on PC2 (Figure 9b) show that two main peaks in the EEM contribute to these PCs, namely the flavonolic compounds on PC1, and the anthocyanins on PC2. These zones are the same as those highlighted in CC2 and CC9. Figure 9c presents the PC1-PC2 score plot obtained after PCA of $\mathbf{M_{3R}}$. The presence or not of Oak chips has an influence on the EEMs, and therefore, the samples are separated into two groups, but this separation occurs on PC2, and not on PC1. As to the PCA of $\mathbf{M_{13R}}$, it leads to a score plot where, with the exception of 2004, the positive

scores correspond to oak chip addition, while the negative scores correspond to no oak chip addition.

MIR Data

The AComDim analysis of the Wine MIR data differentiates the samples according to the production year on two common dimensions, CC2 (4.6% of the variability), and CC6 (0.1% of the variability), as can be seen on Figure 10a. The loadings on CC2 and CC6 are displayed in Figure 10b, and show the two spectral regions varying with the production year. The two bands which can be distinguished in these loading plots are the bands corresponding to the elongation vibration of the C=O ester group (around 1750 cm^{-1}) and the elongation vibration of the aromatic C=C group (around 1620 cm^{-1}).

Figure 10

The salience plots indicate two other dimensions largely influenced by one block: *(i)* CC9, (0.02% of the variability) is mostly based on the Year × Oak interaction. However, on the score plot displayed on Figure 10c, three groups are separated along CC9: the middle group corresponds to samples produced in 2004, while the negative-scores group corresponds to the sample from the group {2005 – Oak2 ∪ 2006 – No Oak}, and the positive-scores group corresponds to the group {2006 – Oak 2 ∪ 2005 – No Oak}. *(ii)* CC10, contains 0.01% of the variability and its largest salience corresponds to the Oak data block. The score plot displayed in Figure 10d shows that the samples are split into two groups according to the addition or not of Oak chips.

With APCA, the only plot where the factor has more variability than noise is Year. Here again, the variability related to Year is not limited to PC1 (Figure 11).

Figure 11

It is interesting to note that PC1 contains "only" around 46.5% of the variance in the data, while PC2 contains almost as much (41.5%), and PC3 still contains 7.5% of variance. The contributions of the spectral zones to these 3 PCs are presented on Figure 11b. Here again, the elongation vibrations of C=O ester groups and C=C aromatic groups are significant, as well as a band around 1020 cm^{-1}, corresponding to the in-plane deformation of the aromatic C–H group.

5. CONCLUSION

This article has presented a novel method for the detection of significant Factors, AComDim, and has compared it to the classical APCA method for several different data sets. The first conclusion which can be drawn from these comparisons is that AComDim is often more sensitive than APCA. The samples are more clearly grouped according to the Factors levels or to the levels of Interactions between Factors, than with APCA (less overlap). In some cases, APCA does not even indicate any significant groups.

Moreover, contrarily to the assumptions on APCA, and confirming a previous study, it has been shown that the effect of the Factor is sometimes not detected in PC1, but in later PCs (up to PC3 in the cases presented here).

Similarly to APCA, AComDim can be used to calculate loadings, which help in the interpretation of the results, by highlighting the links between a studied Factor (or Interaction) and the variables affected by it.

ACKNOWLEDGEMENTS

Jin Chen is gratefully acknowledged for providing the data and for her help with the interpretation of the Wine fluorescence data.

BIBLIOGRAPHY

[1] D. Plaehn, D.S. Lundahl, "An L-PLS preference cluster analysis on French consumer hedonics to fresh tomatoes", *Food Quality and Preferences* **17** (2006), 243

[2] D.C. Plaehn, D.S. Lundahl, "Regression with multiple regressor arrays", *Journal of Chemometrics*, **21** (2007), 621

[3] K. Muteki, J.F. MacGregor, "Multi-block PLS modeling for L-shape data structures with applications to mixture modeling", *Chemometrics and Intelligent Laboratory Systems* **85** (2007), 186

[4] A. Höskuldsson, "Multi-block and path modelling procedures", *Journal of Chemometrics* **22** (2008), 571

[5] E.M. Qannari, I. Wakeling, P. Courcoux, H.J.H MacFie, "Defining the underlying sensory dimensions", *Food Quality and Preferences* **11** (2000), 151

[6] The MathWorks Inc., Natick (MA, USA), 2008

[7] SAISIR (2008). Package of function for chemometrics in the Matlab (R) environment. Dominique Bertrand coordinator (bertrand@nantes.inra.fr). Unité "Biopolymères, Interactions, Assemblage". INRA, rue de la Géraudière - BP 71627 - 44316 Nantes Cedex 3 France.

[8] P. de B. Harrington, N.E. Vieira, J. Espinoza, J.K. Nien, R. Romero, A.L. Yergey, "Analysis of variance-principal component analysis: a soft tool for proteomic discovery", *Analytica Chimica Acta* **544** (2005), 118

[9] E. M. Qannari, I. Wakeling, H.J.H. MacFie, "A hierarchy of models for analysis sensory data", *Food Quality and Preference* **6** (1995), 309

[10] M. Hanafi, G. Mazerolles, E. Dufour, E.M. Qannari, "Common components and specific weight analysis and multiple Co-inertia analysis applied to the coupling of several measurement techniques", *Journal of Chemometrics* **20** (2006), 1

[11] P. de B. Harrington, N.E. Vieira, P. Chen, J. Espinoza, J.K. Nien, R. Romero, A.L. Yergey, "Proteomic analysis of amniotic fluids using analysis of variance-principal component analysis and fuzzy rule-building expert systems applied to matrix-assisted laser desorption/ionization mass spectrometry", *Chemometrics and Intelligent laboratory Systems* **82** (2006), 283

[12] J. Sarembaud, R. Pinto, D.N. Rutledge, M. Feinberg, "Application of the ANOVA-PCA method to stability studies of reference materials", *Analytica Chimica Acta* **603** (2007), 147

[13] R. Climaco-Pinto, V. Bosc, H. Noçairi, A.S. Barros, D.N. Rutledge, "Using ANOVA-PCA for discriminant analysis: Application to the study of mid infrared spectra of Carraghenan gels as a function of concentration and temperature", *Analytica Chimica Acta* 629 (2008), 47

[14] A.K. Smilde, J.J. Jansen, H.C.J. Hoefsloot, R-J.A. N. Lamers, J.Greef, M.E.

Timmerman, *Bioinformatics* 21 (13) (2005) 4043

[15] J.J.Jansen, H.C.J. Hoefsloot, J.Greef, M.E.Timmerman, J.A. Westerhuis, A.K.Smilde, *J.*

Chemometrics 19 (9) (2005), 469

[16] D.J.Vis, J.A.Westerhuis, A.K.Smilde, J.Greef, BMC *Bioinform*. 8 (2007) 322

[15] D.N. Rutledge, A.S. Barros, F. Gaudard, "ANOVA and factor analysis applied to time domain NMR signals", Magnetic Resonance in Chemistry 35, S13 (1997)

[16] A. Rudnitskaya, L.M. Schmidtke, I. Delgadillo, A. Legin, G. Scollary, "Study of micro-oxygenation and oak ch²ip maceration on wine composition using an electronic tongue and chemical analysis", *Analytica Chimica Acta* (2009), in press

[17] R. Climaco Pinto, A.S. Barros, N. Locquet, L. Schmidtke, D.N. Rutledge, "Improving the detection of significant factors using ANOVA-PCA by selective reduction of residual variability", submitted to *Analytica Chimica Acta*

TABLE 1: AComDim applied to the Apple data: Percentage of variance explained by each common dimension

Common Dimension	% explained information
1	80.63
2	9.69
3	1.72
4	7.52
5	0.18
6	0.09
7	0.14
8	0.01
9	0.003
10	0.004

FIGURE CAPTIONS

Figure 1

Graphical representation of the APCA procedure

Figure 2

Apple Data: Saliences versus Block index, for Common Dimensions 1 to 10

Figure 3

AComDim applied to the Apple data: Scores on CC1 *versus* scores on (a) CC2; (b) CC4; (c) CC7; (d) CC8; and (e) CC10

Figure 4

APCA on Apple data: PC1-PC2 score plots obtained after PCA of (a) $\mathbf{M_{1R}}$, (b) $\mathbf{M_{2R}}$, and (c) $\mathbf{M_{3R}}$

Figure 5

Lignin data: Saliences versus Block index, for Common Dimensions 1 to 10

Figure 6

Comparison between AComDim and APCA for Lignin data: (a) CC1 *vs* CC2 scores plot; (b) APCA of $\mathbf{M_{1R}}$ PC1-PC2 scores plot; (c) CC1 *vs* CC3 scores plot; (d) APCA of $\mathbf{M_{2R}}$ PC1-PC2 scores plot; (e) CC1 *vs* CC5 scores plot; (f) APCA of $\mathbf{M_{3R}}$ PC1-PC2 scores plot; (g) APCA of $\mathbf{M_{12R}}$ PC1-PC2-PC3 scores plot;

Figure 7

Wine_Fluorescence data: Saliences versus Table index, for Common Dimensions 1 to 10

Figure 8

(a) CC9 *vs* CC2 for the AComDim analysis of the Wine_Fluorescence data; (b) AComDim-loadings on CC2 and CC9; (c) CC1 *vs* CC3 score plot; (d) CC1 *vs* CC8 score plot.

Figure 9

APCA on Wine_Fluorescence data: (a) PC1-PC2 score plots obtained after PCA of $\mathbf{M_{1R}}$; (b) Loadings on PC1 and on PC2 after PCA of $\mathbf{M_{1R}}$; (c) PC1-PC2 score plots obtained after PCA of $\mathbf{M_{3R}}$; (d) PC1-PC2 score plots obtained after PCA of $\mathbf{M_{13R}}$ (g1: 2004-No oak, g2: 2004-oak; g3: 2005-No oak; g4: 2005-oak; g5: 2006-No oak; g6: 2006-oak)

Figure 10

Wine_MIR data: (a) CC6 *vs* CC2; (b) AComDim loadings on CC2 and on CC6; (c) CC1 *vs* CC9; (d) CC1 *vs* CC10

Figure 11

(a)APCA of Wine_MIR data: PC1-PC2-PC3 score plot; (b) Loadings on the first 3 PCs after PCA of $\mathbf{M_{1R}}$

Figure 1

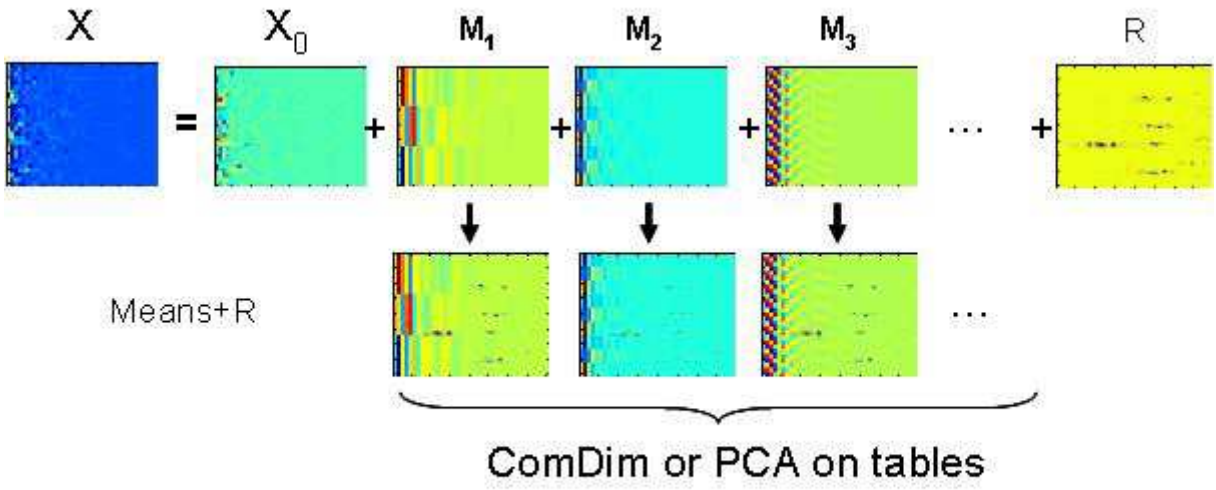


Figure 2

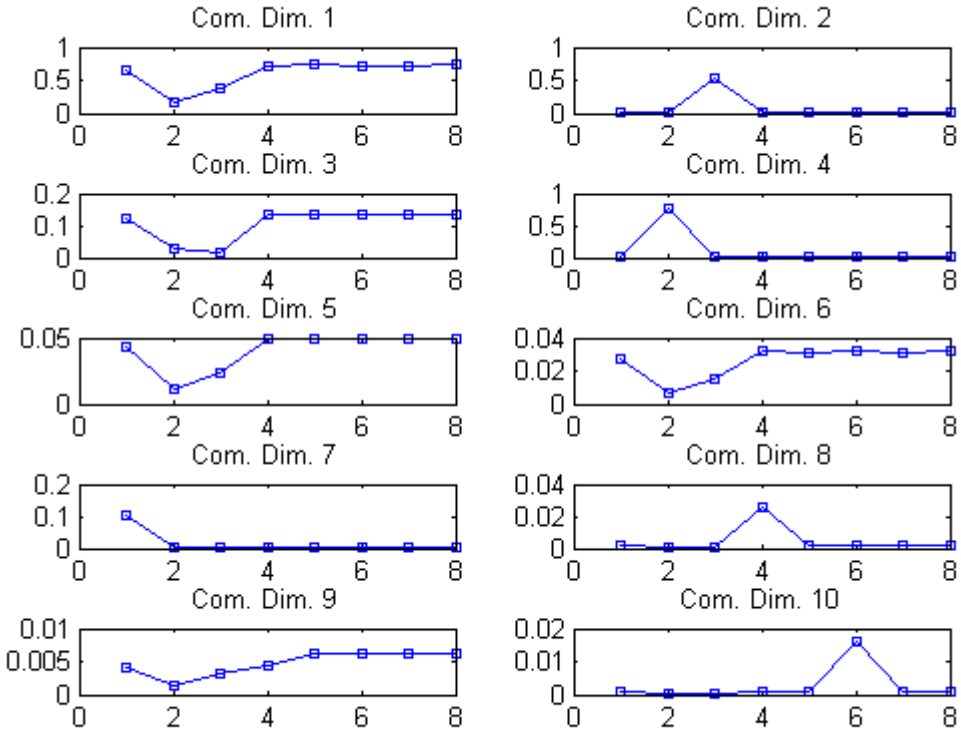
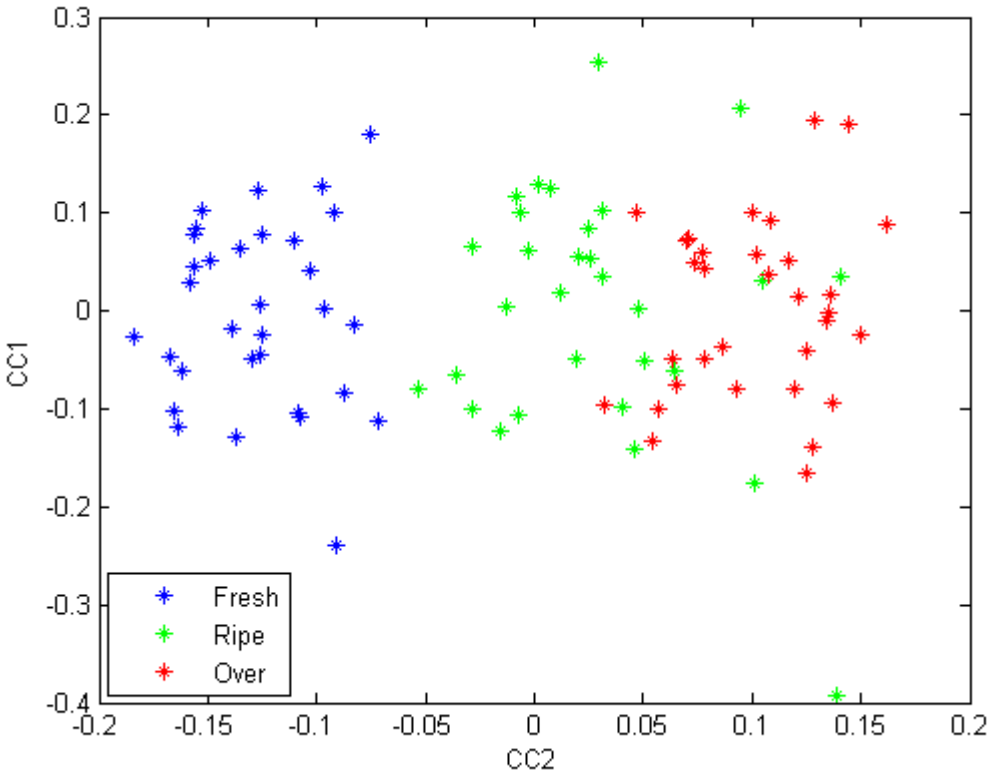
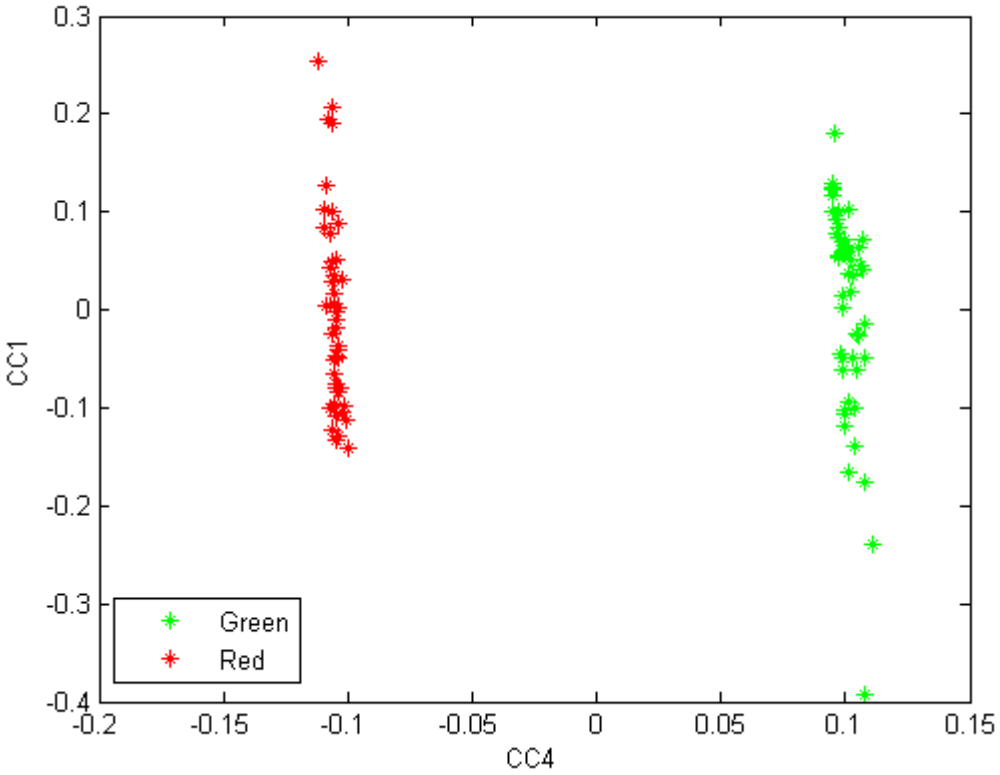

Com. Dim. 1
Com. Dim. 2
Com. Dim. 3
Com. Dim. 4
Com. Dim. 5
Com. Dim. 6
Com. Dim. 7
Com. Dim. 8
Com. Dim. 9
Com. Dim. 10

Figure 3

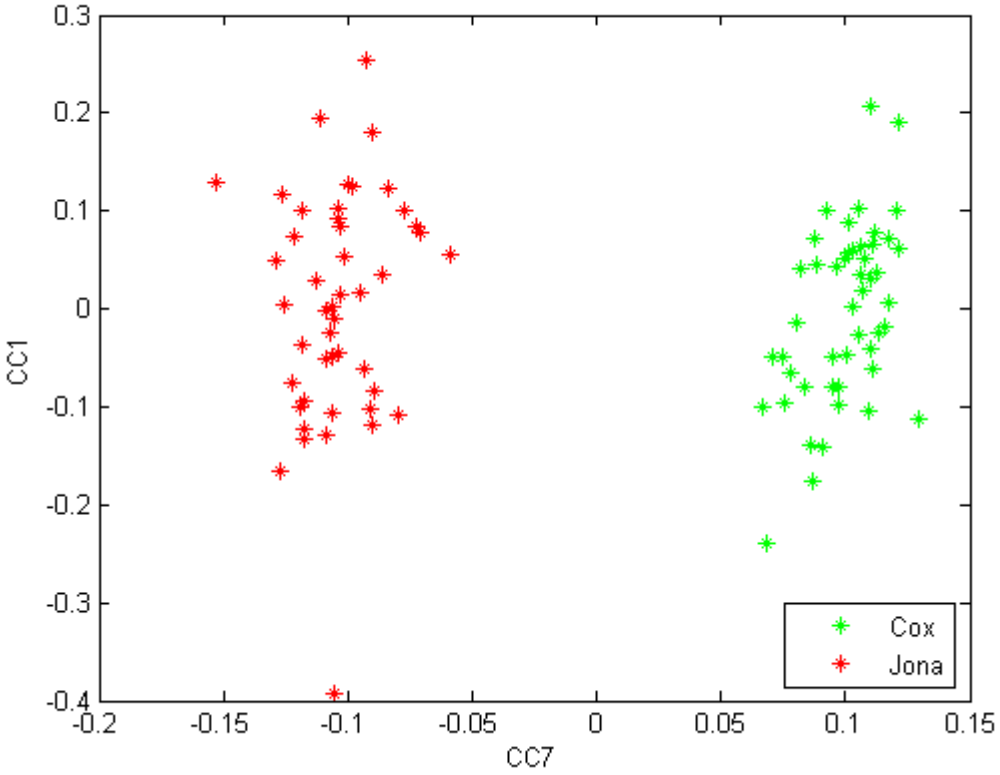
(a)


0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
CC1
-0.2
-0.15
-0.1
-0.05
0
0.05
0.1
0.15
0.2
CC2
Fresh
Ripe
Over

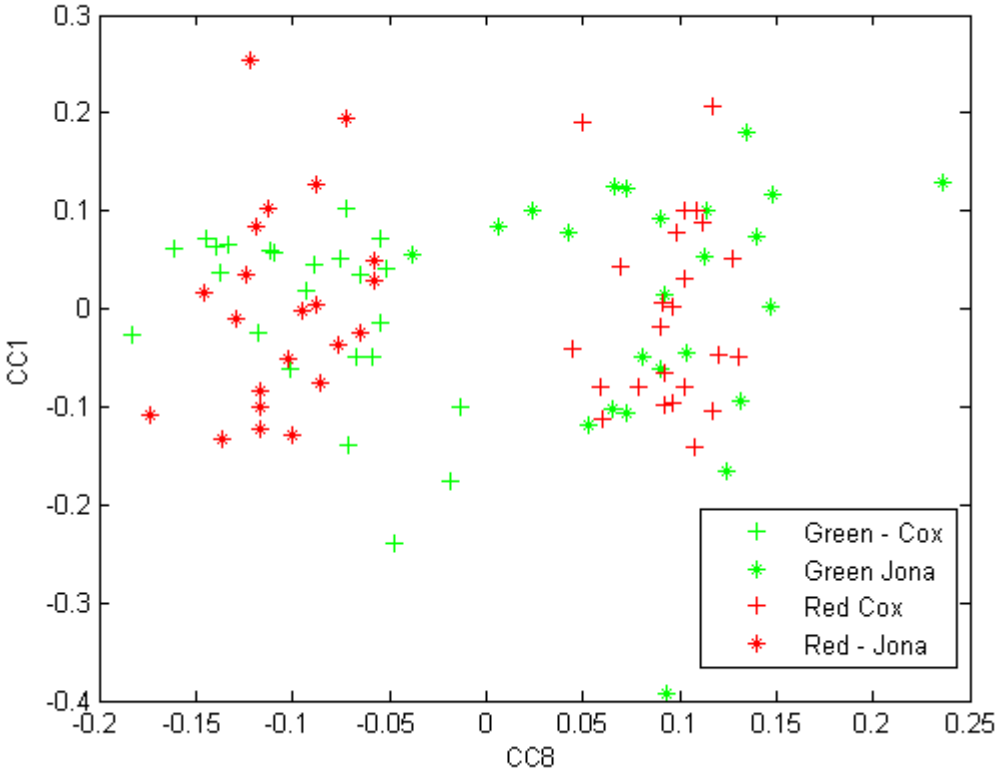
(b)


0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
CC1
-0.2
-0.15
-0.1
-0.05
0
0.05
0.1
0.15
CC4
Green
Red

(c)


0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
CC1
-0.2
-0.15
-0.1
-0.05
0
0.05
0.1
0.15
CC7
Cox
Jona

(d)


0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
CC1
-0.2
-0.15
-0.1
-0.05
0
0.05
0.1
0.15
0.2
0.25
CC8
Green - Cox
Green Jona
Red Cox
Red - Jona

(e)

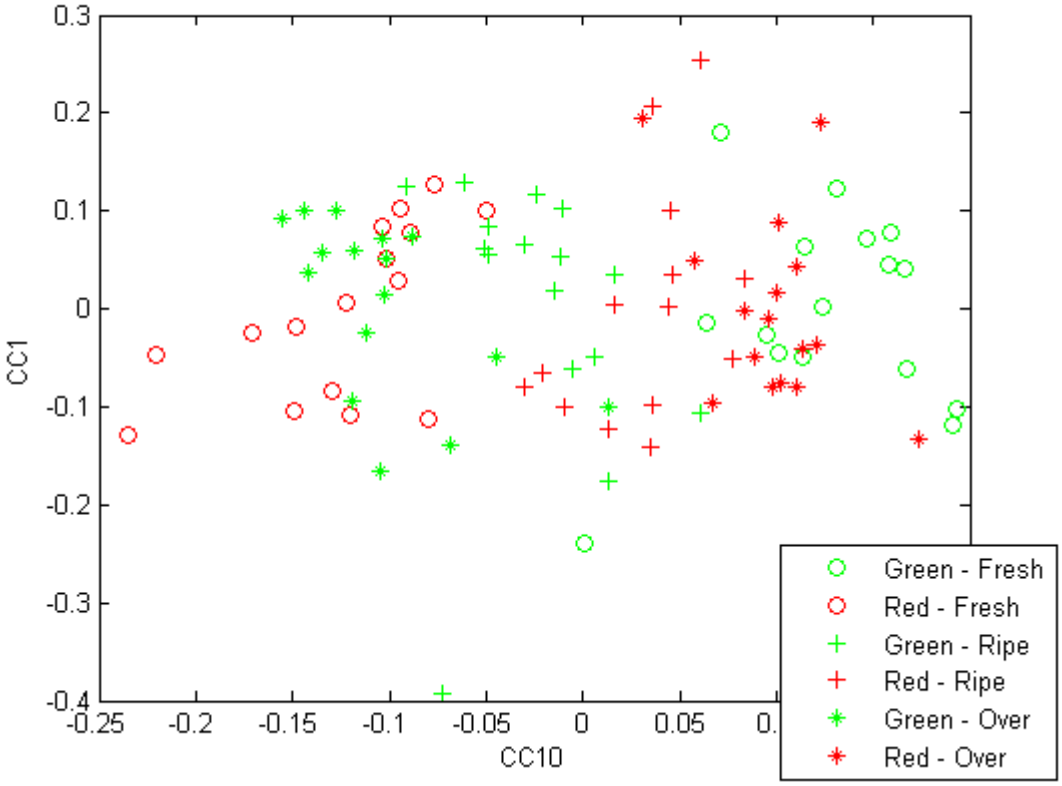
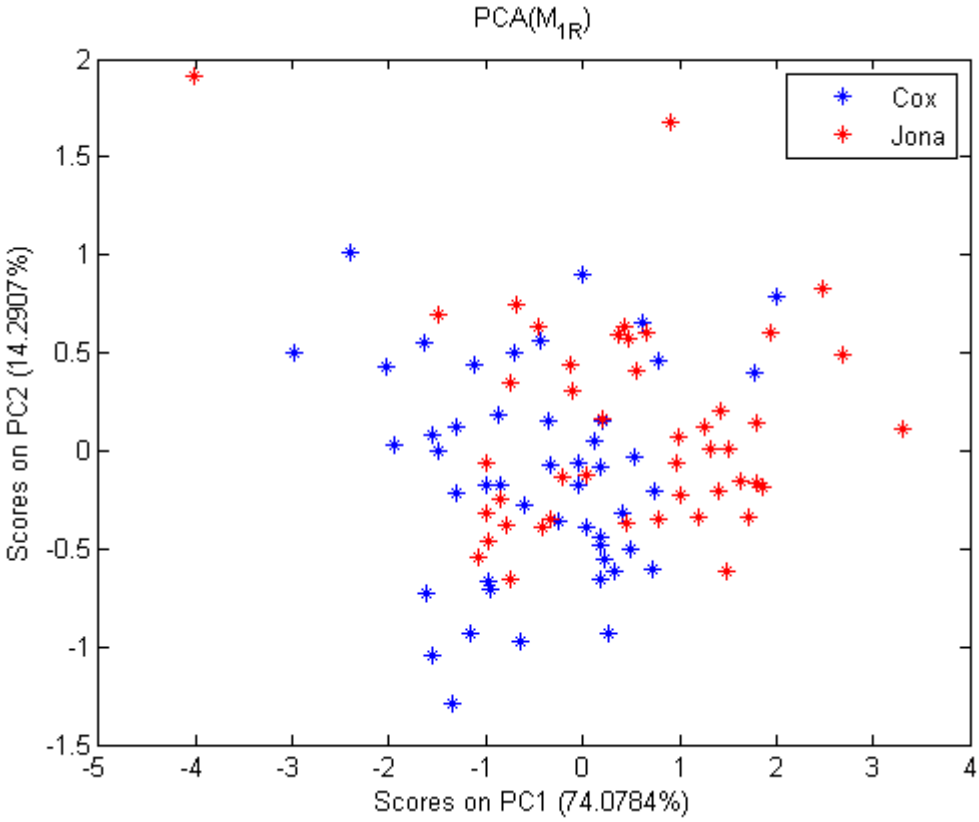
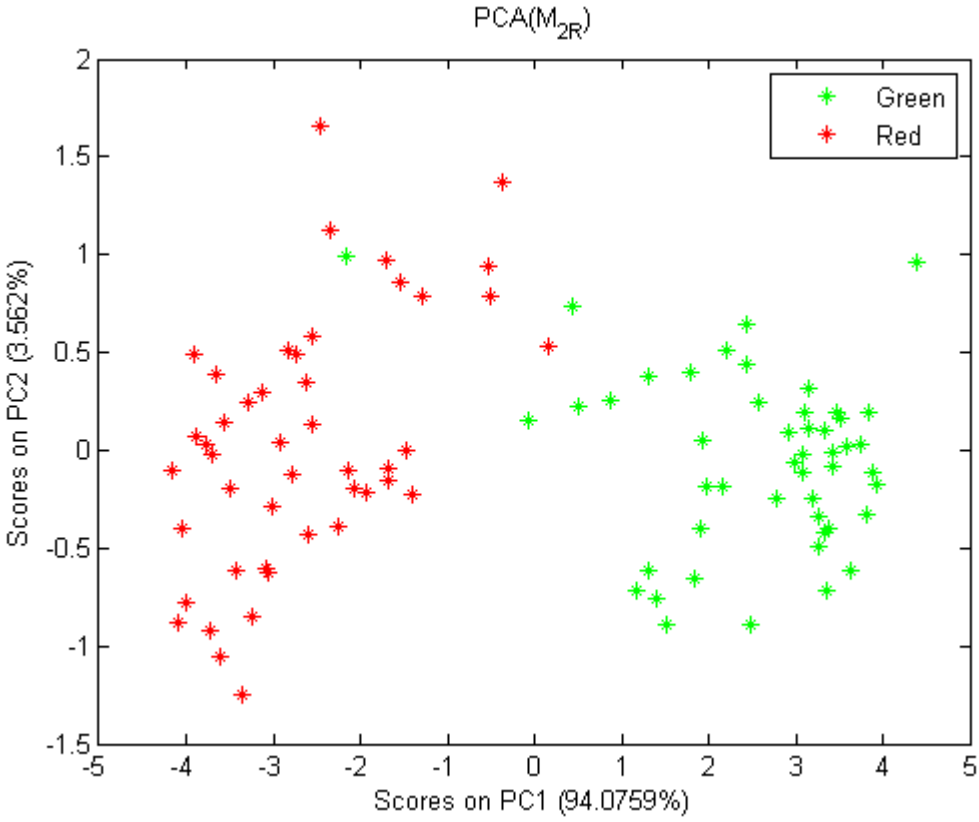

0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
CC1
-0.25
-0.2
-0.15
-0.1
-0.05
0
0.05
CC10
Green - Fresh
Red - Fresh
Green - Ripe
Red - Ripe
Green - Over
Red - Over

Figure 4

(a)


PCA(M1R)
Cox
Jona
Scores on PC2 (14.2907%)
Scores on PC1 (74.0784%)

(b)


PCA(M2R)
Green
Red
Scores on PC2 (3.562%)
Scores on PC1 (94.0759%)

(c)

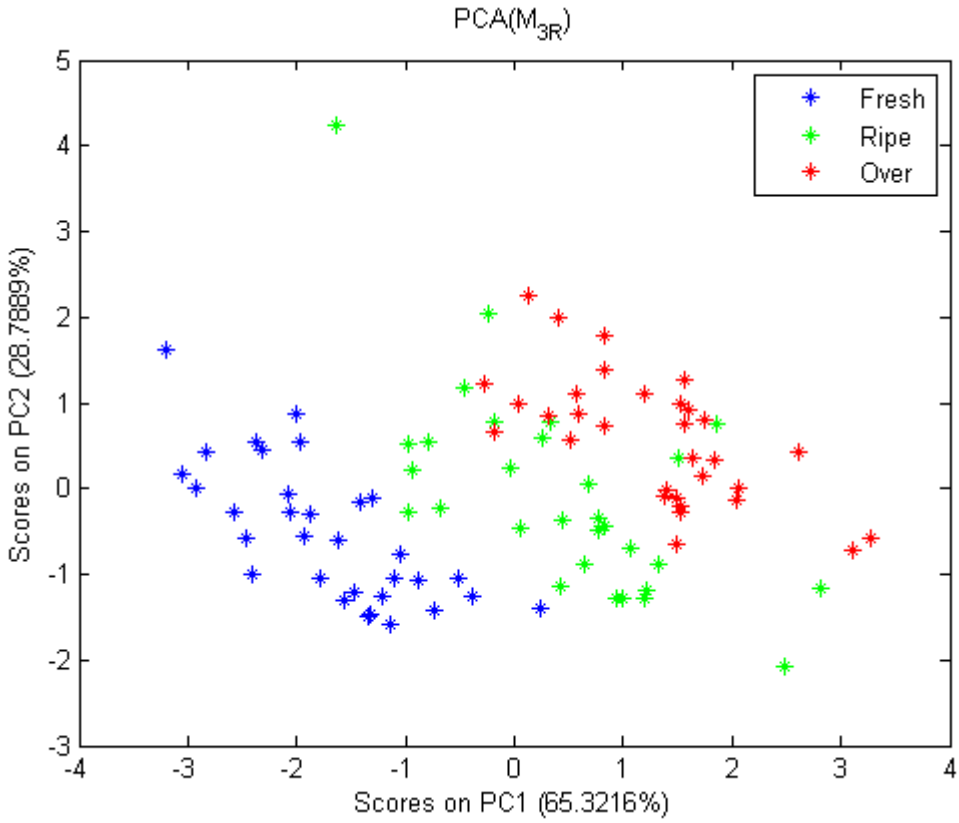

PCA(M3R)
Fresh
Ripe
Over
Scores on PC2 (28.7889%)
Scores on PC1 (65.3216%)
5
4
3
2
1
0
-1
-2
-3
-4
-3
-2
-1
0
1
2
3
4

Figure 5

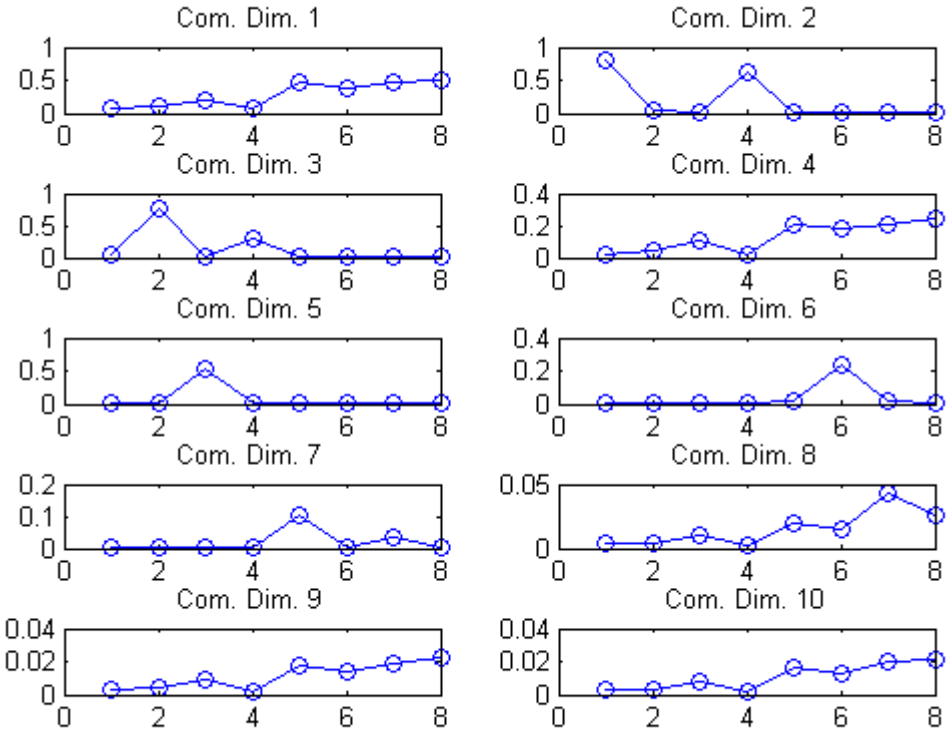
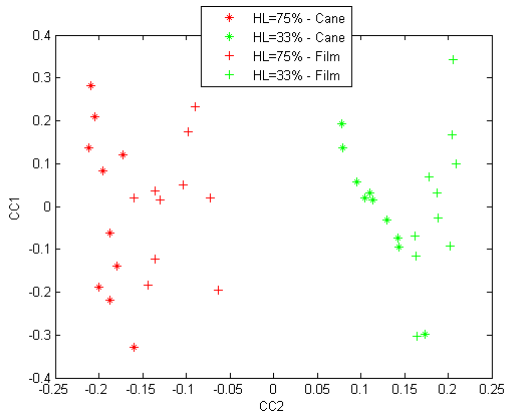
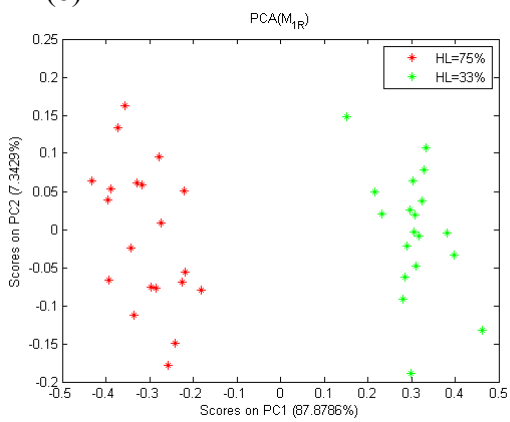

Com. Dim. 1
Com. Dim. 2
Com. Dim. 3
Com. Dim. 4
Com. Dim. 5
Com. Dim. 6
Com. Dim. 7
Com. Dim. 8
Com. Dim. 9
Com. Dim. 10

Figure 6

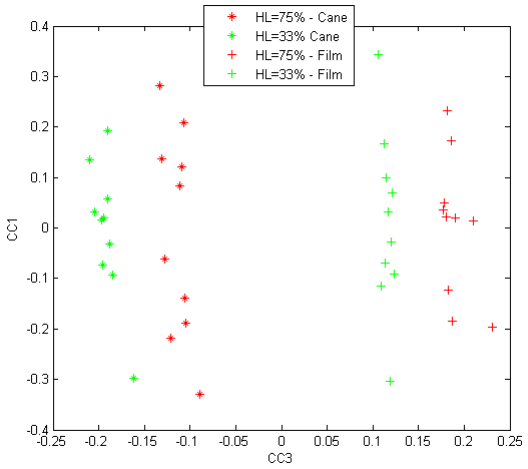
(a)



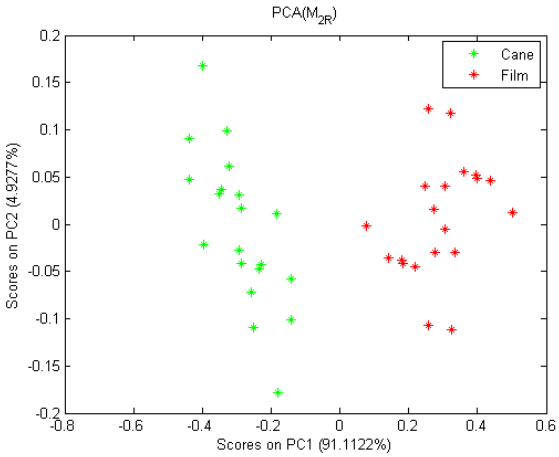
(b)



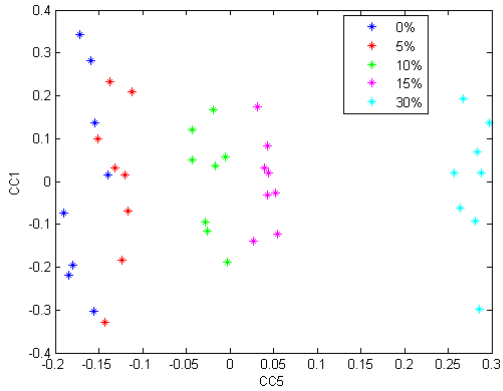
(c)



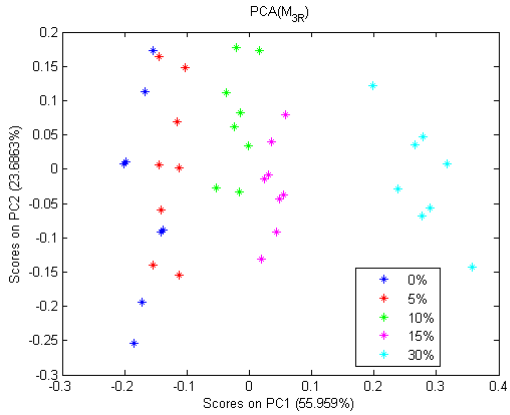
(d)



(e)



(f)



(g)

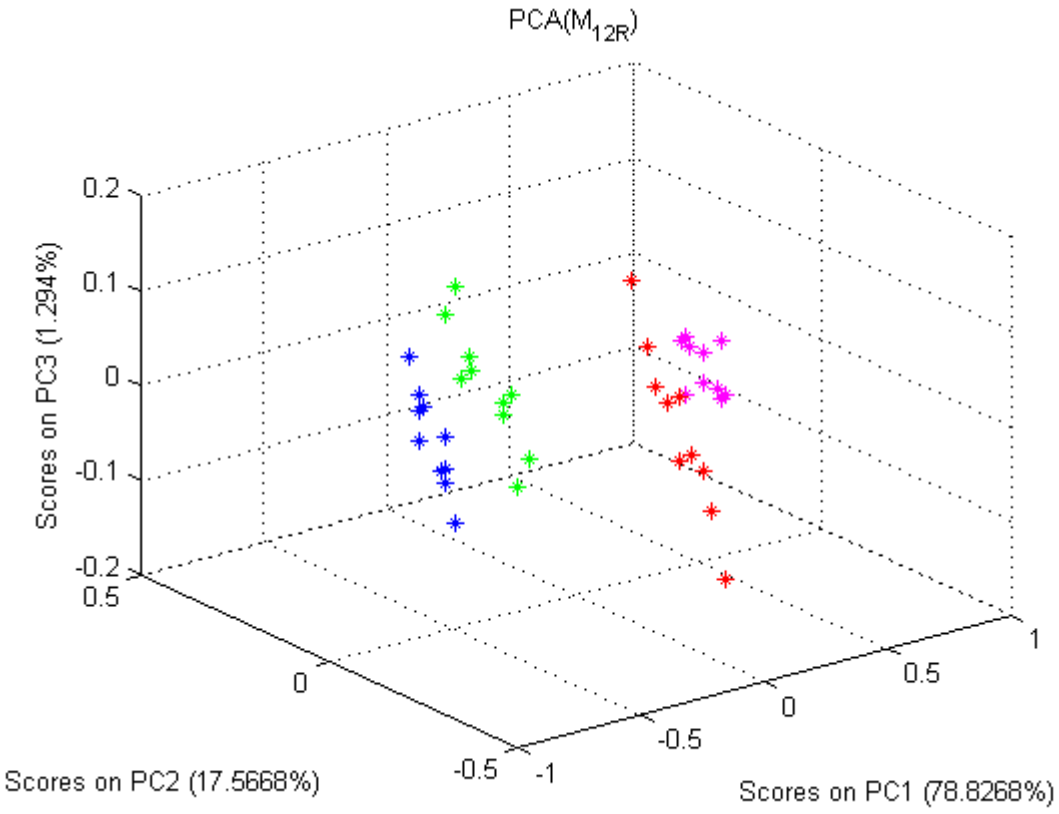

PCA(M12R)
Scores on PC3 (1.294%)
0.2
0.1
0
-0.1
-0.2
0.5
0
-0.5
Scores on PC2 (17.5668%)
-1
-0.5
0
0.5
1
Scores on PC1 (78.8268%)

Figure 7

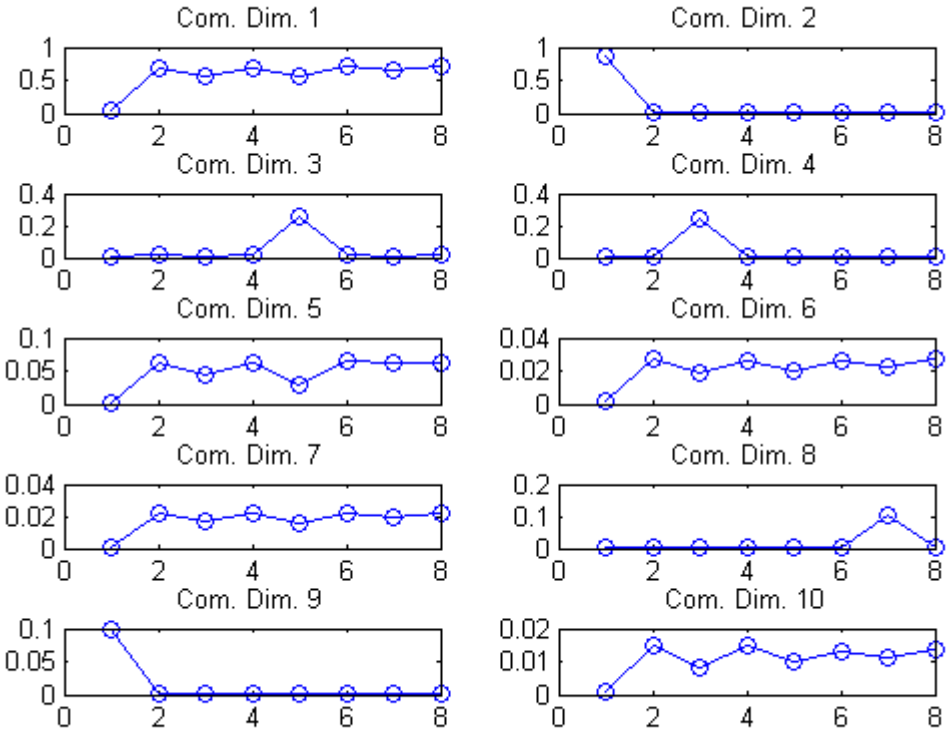
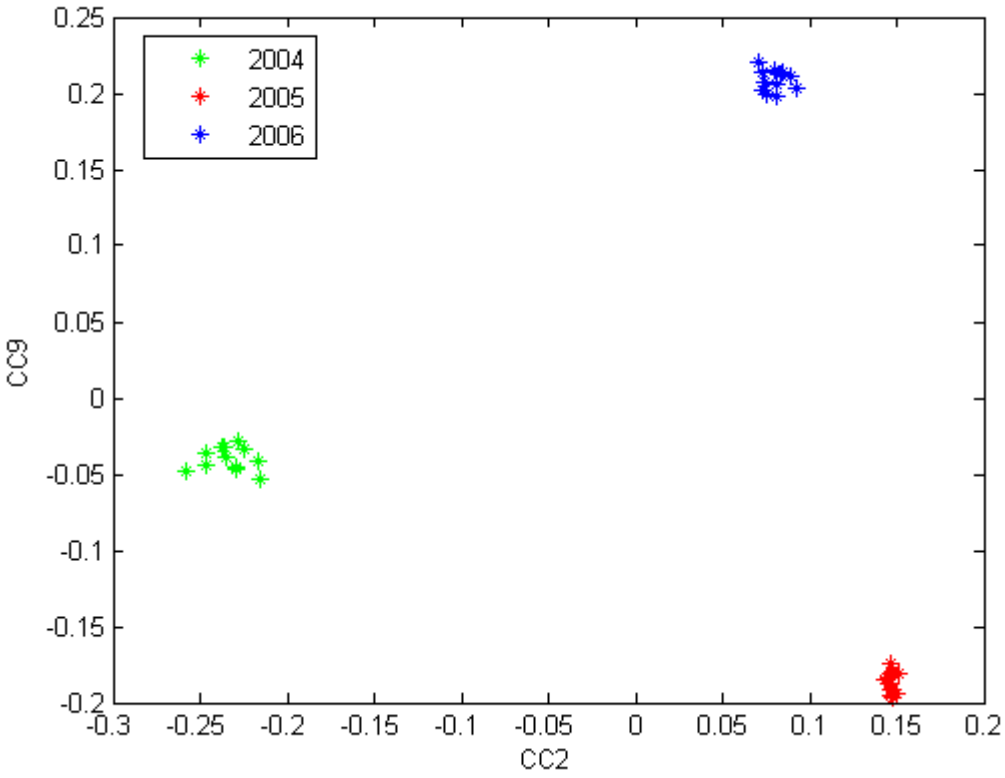
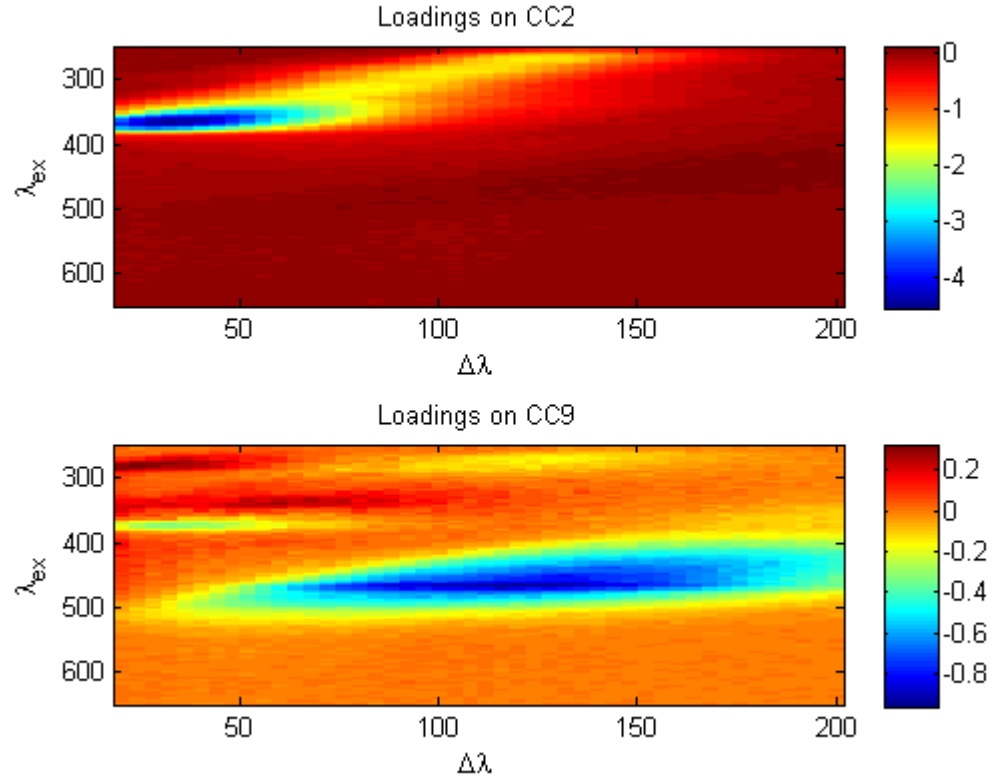

Com. Dim. 1
Com. Dim. 2
Com. Dim. 3
Com. Dim. 4
Com. Dim. 5
Com. Dim. 6
Com. Dim. 7
Com. Dim. 8
Com. Dim. 9
Com. Dim. 10

Figure 8

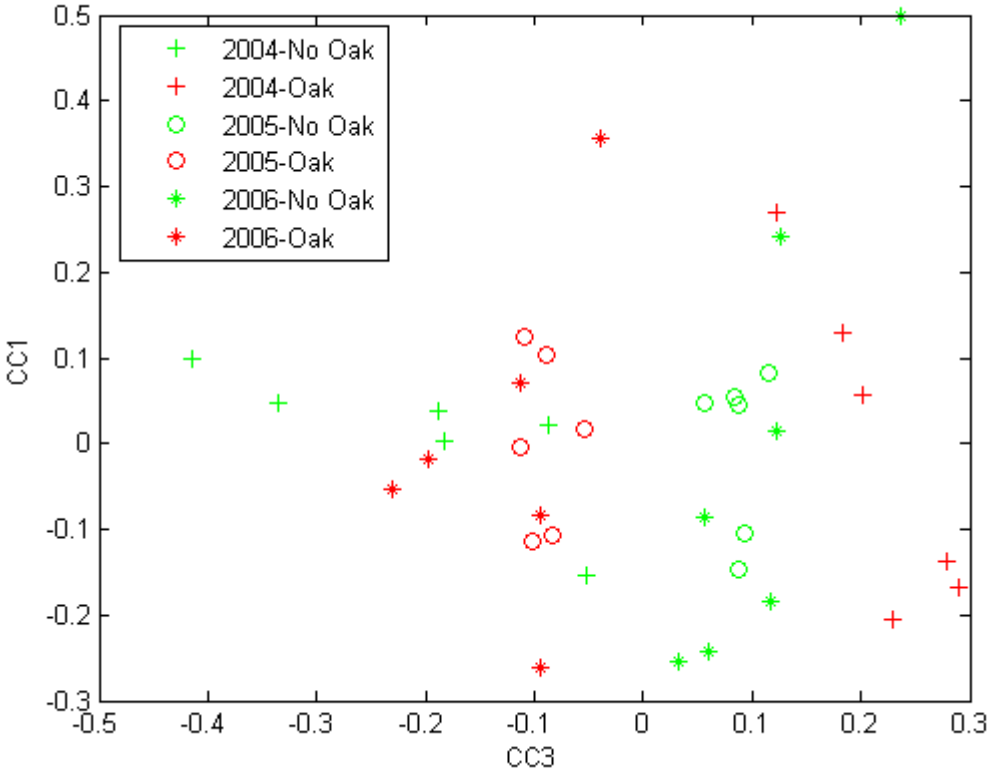
(a)


2004
2005
2006
CC9
CC2
0.25
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
-0.3
-0.25
-0.2
-0.15
-0.1
-0.05
0
0.05
0.1
0.15
0.2

(b)


Loadings on CC2
λex
Δλ
300
400
500
600
50
100
150
200
0
-1
-2
-3
-4
Loadings on CC9
λex
Δλ
0.2
0
-0.2
-0.4
-0.6
-0.8

(c)


2004-No Oak
2004-Oak
2005-No Oak
2005-Oak
2006-No Oak
2006-Oak
CC1
CC3

(d)

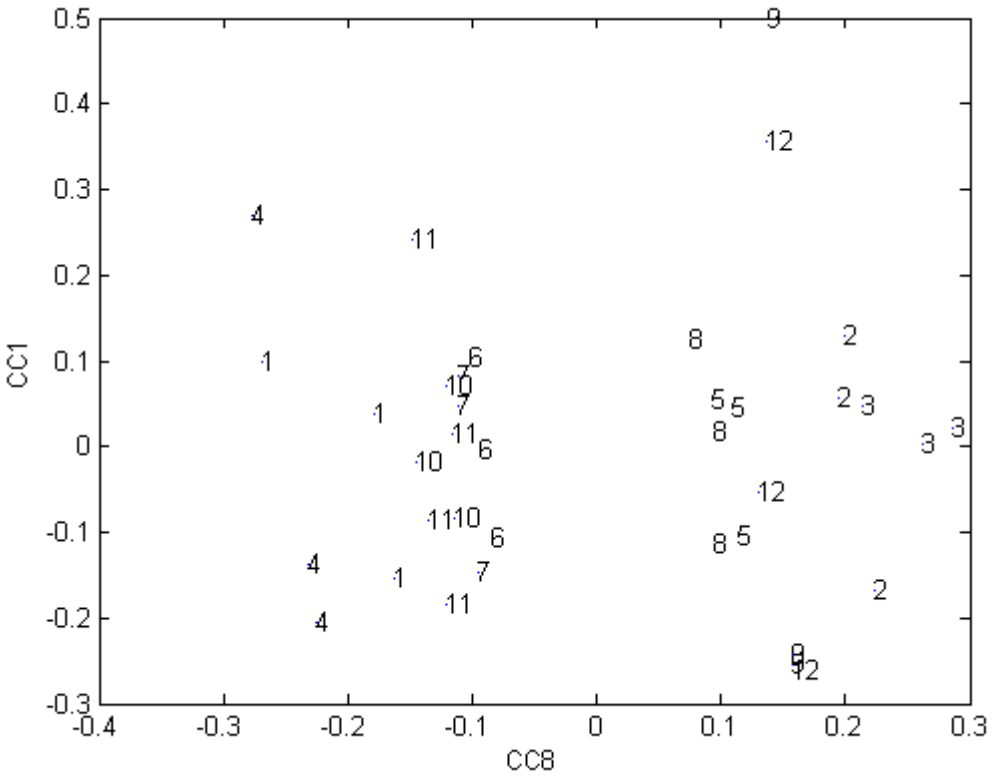
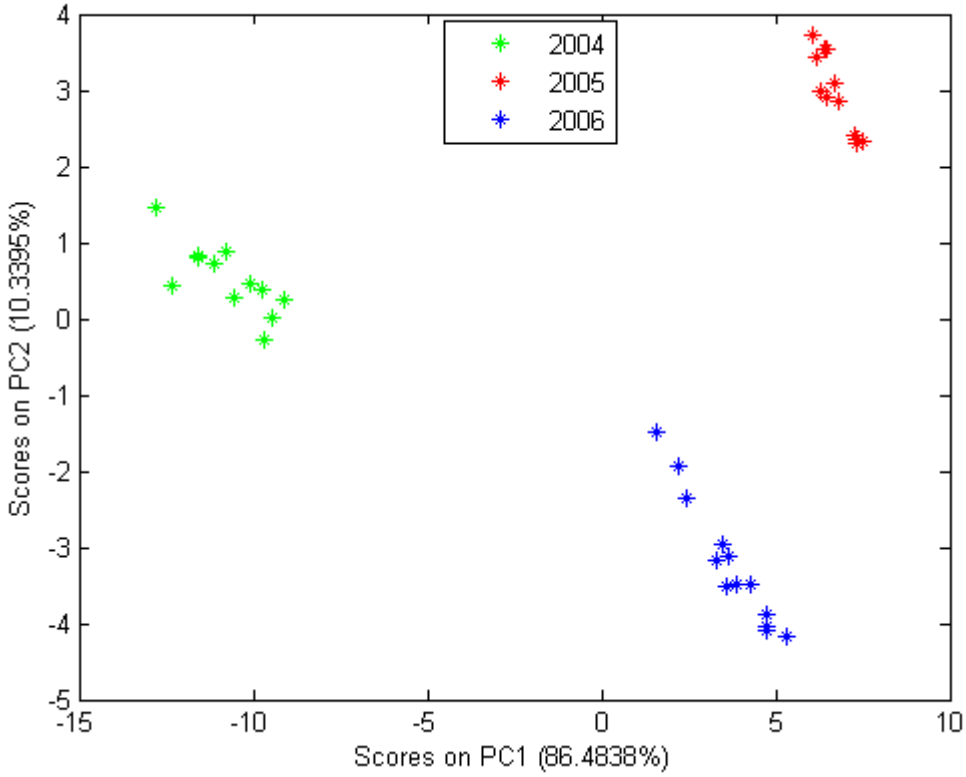
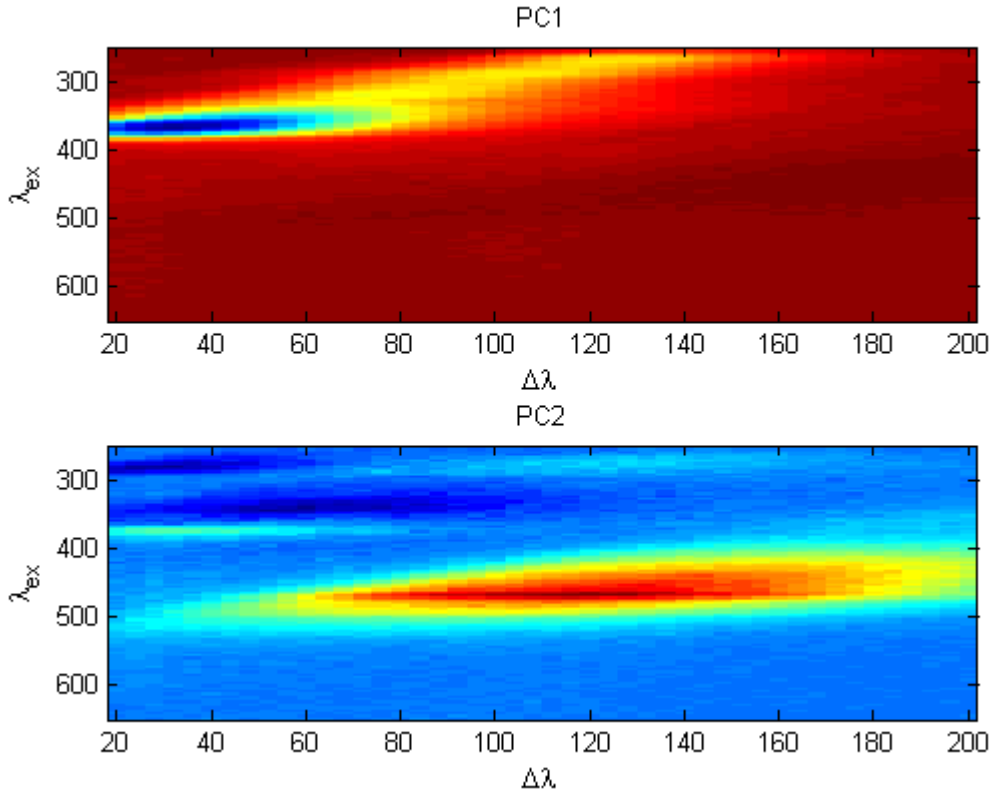

CC1
CC8

Figure 9

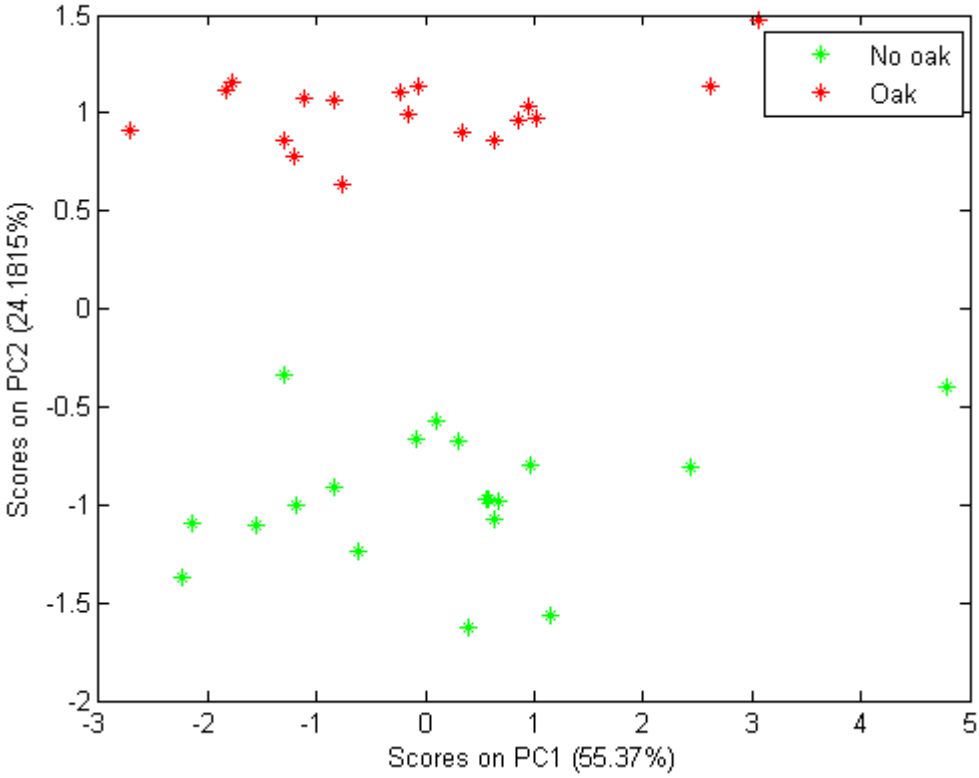
(a)


2004
2005
2006
Scores on PC2 (10.3395%)
Scores on PC1 (86.4838%)
4
3
2
1
0
-1
-2
-3
-4
-5
-15
-10
-5
0
5
10

(b)


PC1
PC2
λex
Δλ
300
400
500
600
20
40
60
80
100
120
140
160
180
200

(c)


No oak
Oak
Scores on PC2 (24.1815%)
Scores on PC1 (55.37%)
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-3
-2
-1
0
1
2
3
4
5

(d)

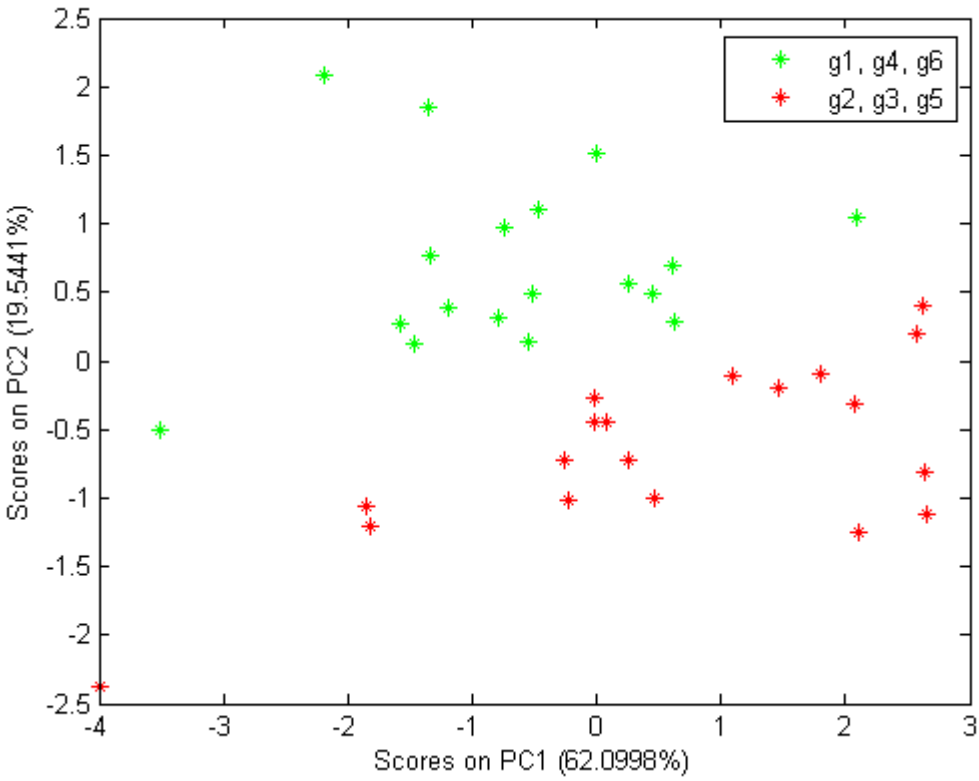
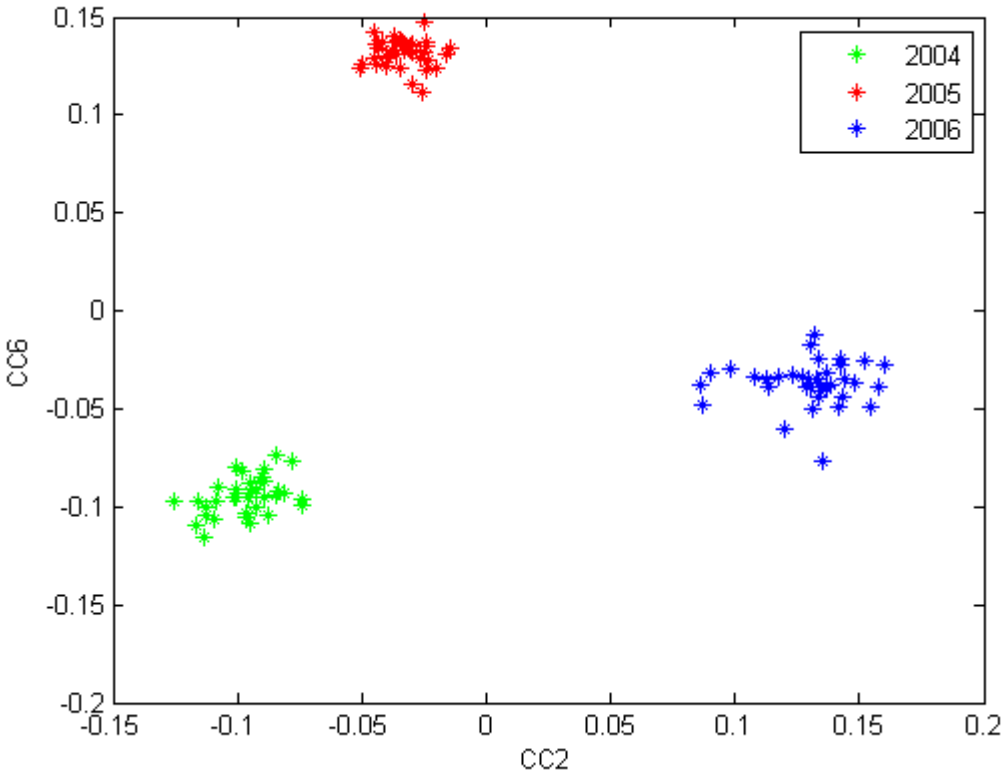
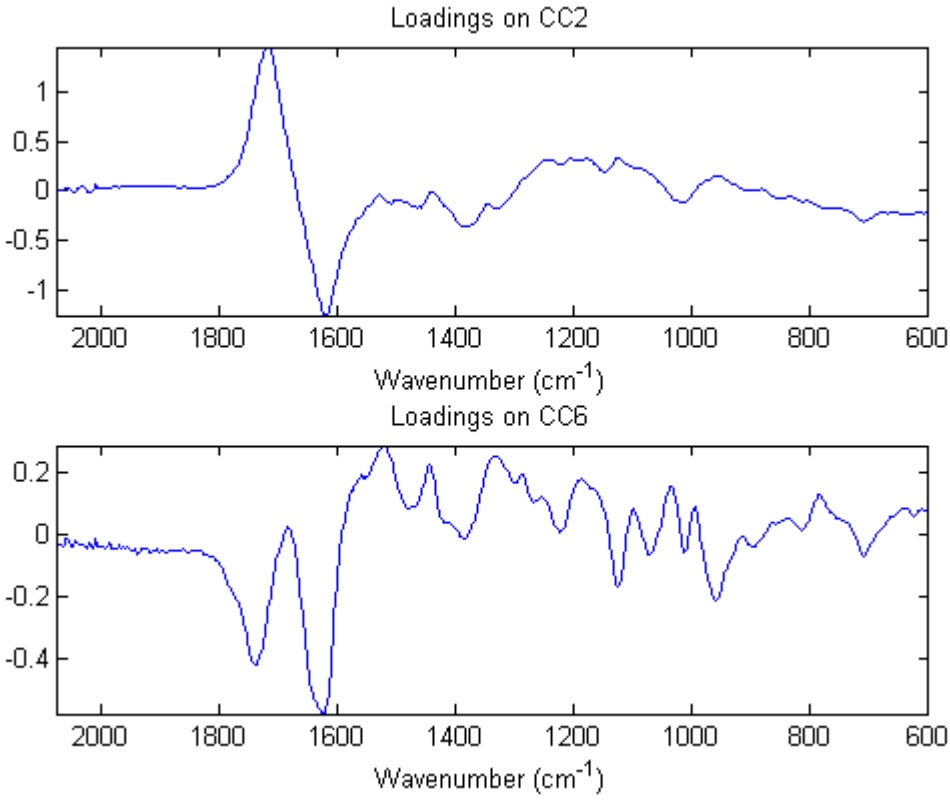

g1, g4, g6
g2, g3, g5
Scores on PC2 (19.5441%)
Scores on PC1 (62.0998%)
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-4
-3
-2
-1
0
1
2
3

Figure 10

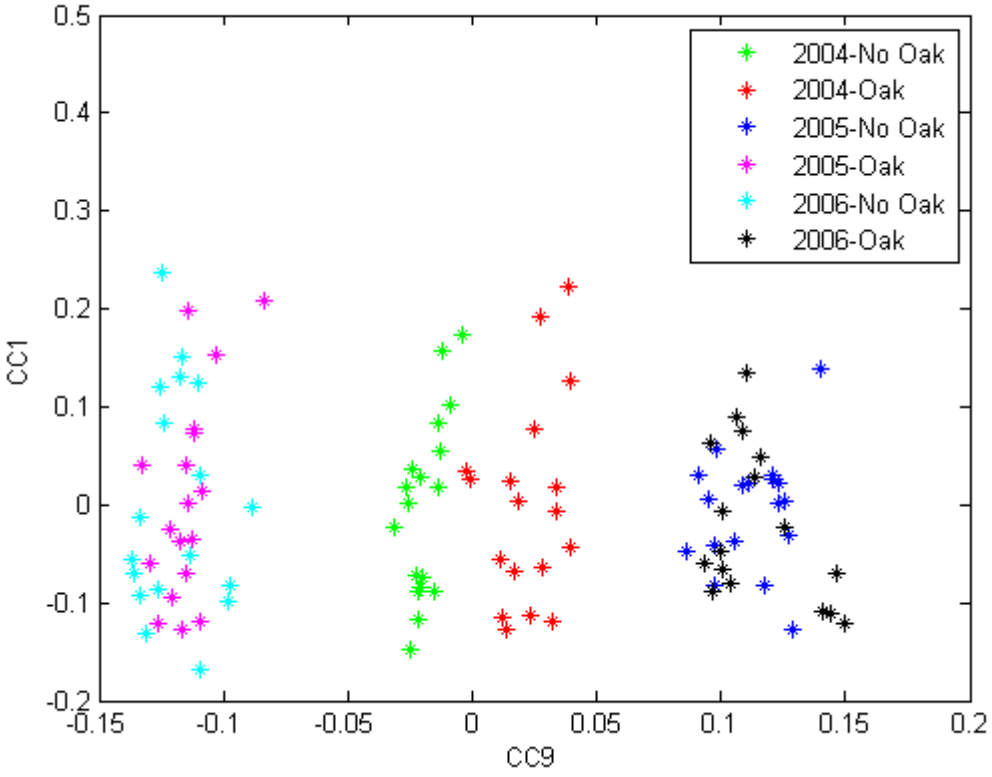
(a)


2004
2005
2006
CC6
CC2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
-0.15
-0.1
-0.05
0
0.05
0.1
0.15
0.2

(b)


Loadings on CC2
1
0.5
0
-0.5
-1
2000
1800
1600
1400
1200
1000
800
600
Wavenumber (cm-1)
Loadings on CC6
0.2
0
-0.2
-0.4
2000
1800
1600
1400
1200
1000
800
600
Wavenumber (cm-1)

(c)


2004-No Oak
2004-Oak
2005-No Oak
2005-Oak
2006-No Oak
2006-Oak
CC1
CC9
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.15
-0.1
-0.05
0
0.05
0.1
0.15
0.2

(d)

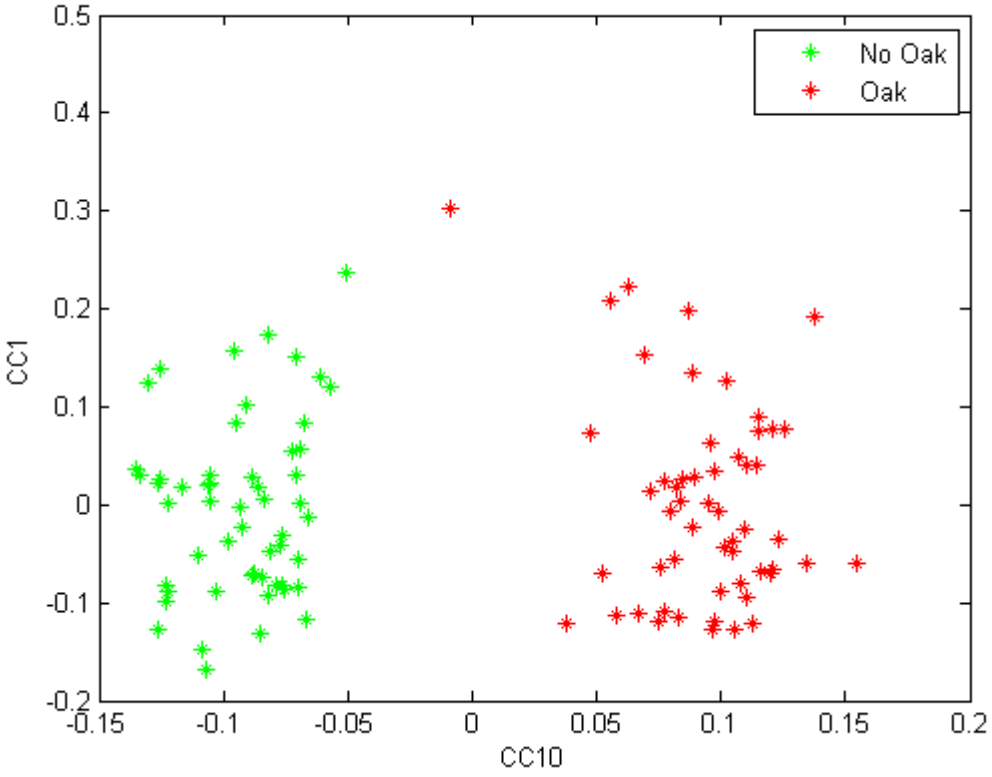
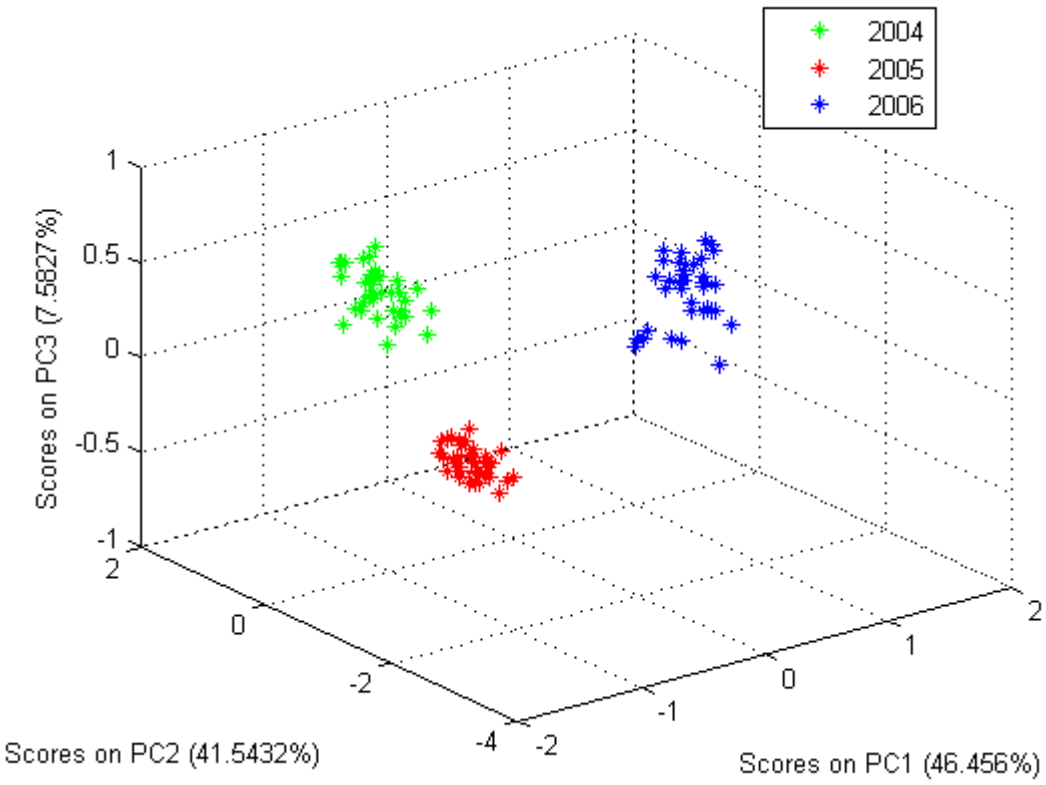
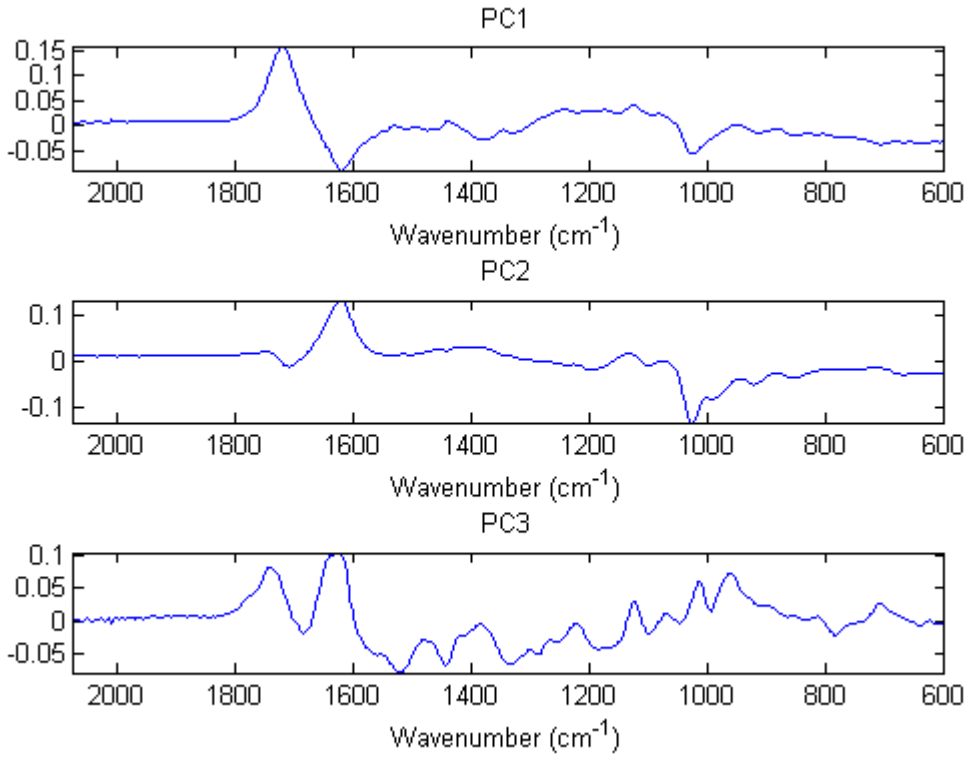

No Oak
Oak
CC1
CC10
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.15
-0.1
-0.05
0
0.05
0.1
0.15
0.2

Figure 11

(a)


2004
2005
2006
Scores on PC3 (7.5827%)
Scores on PC2 (41.5432%)
Scores on PC1 (46.456%)

(b)


PC1
PC2
PC3
Wavenumber (cm-1)

ANNEXE VII

Preliminary studies on the Mid-Infrared analysis of edible oils by direct heating on an ATR diamond crystal

R. Climaco Pinto[1,3]; N. Locquet[2]; L. Eveleigh[1,2]; D.N. Rutledge[1,2]

1) Laboratoire de Chimie Analytique, AgroParisTech. 16, rue Claude Bernard. 75005 Paris, France

2) INRA/AgroParisTech, UMR 214 "IAQA" 16, rue Claude Bernard, 75005, Paris, France

3) Departamento de Química, Universidade de Aveiro Campus Universitário de Santiago. 3810-193 Aveiro, Portugal

Abstract

In this work a new, easy and rapid MIR-ATR technique to monitor the thermal stability of oils is presented. The method uses a heated ATR apparatus set at selected temperatures to thermally modify the oils and acquire spectra, simultaneously. Three different edible oils sunflower, olive and canola, are subjected to the method. Sunflower oil with or without tocopherol and at three different temperatures was also analyzed. Wavelengths known to be relevant to oil degradation processes are selected to show the modifications in the spectra over time.

1. Introduction

MIR

Mid-Infrared spectroscopy (MIR) is well suited to the analysis of edible oils without prior treatment of the sample, resulting in an easy to use, rapid spectroscopic technique[1]. The technique is well adapted to the study of organic compounds as the characteristic vibrational mode of each functional group gives rise to bands in the infrared spectrum at specific frequencies, influenced by surrounding functional groups [2,3,4].

Oils

Among other applications, MIR has been used to discriminate oil samples from different regions [5] and of different botanical origins[6], to detect adulteration[7], monitor "freshness"[8], study oxidation processes and other degradation reactions [3,15,18,19,22,23,24], and to quantify different functional groups [9,10,11].

Thermal Degradation of oils

Though it even takes place at room temperature, oil degradation is much faster during frying due to the activation energies of the various elementary reactions involved [13]. Oil oxidation is the most noticeable degradation process and has consequences on the organoleptic and quality characteristics of food samples, with consequences on health, nutritional quality and consumer choice. Fried products absorb a large quantity of its frying oil, thus eventually accumulating degradation products. [2, 12, 13, 14, 15]

Edible oils are almost entirely constituted of triglycerides (95-98%) with different fatty acid compositions, varying in length, unsaturation type and relative percentage. Oils also contain many other families of compounds at low concentrations (2-5%), including alcohols, esters, hydrocarbons, phenolic compounds, tocopherols, tocotrienols, pigments and phospholipids[16].

There have been several studies on the constitution and oxidation of edible oils. Oxidation of oils is mainly linked to two types of parameters - chemical composition and the physico-chemical conditions to which they are subjected:

- The chemical composition depends on the fatty acid composition, as well as the presence of antioxidants, mainly tocopherols but, depending on the oil family, also polyphenols [14, 17].
- Physico-chemical conditions such as temperature, presence of oxygen, light, catalysts, oxidation surface area and time all influence the rate of thermal degradation reactions.

ATR

Attenuated Total Reflection (ATR) is a well known technique [18,19,20,21] to analyze solid, liquid and gel samples. In general it does not require preparation of the sample and it enables one to analyze very small amounts of sample, on the order of the tens of microliters, making it an easy and rapid spectroscopic technique.

Analysis of oils on a heated ATR

Heat-controlled ATR can mimic a flat-frying system, which may allow to directly monitor the oxidation of oil samples in an easy and very rapid way. The samples are placed on the ATR crystal and heated at the desired temperature. Due to the smaller volumes of sample compared to other experimental set-ups, the ratio of surface-area/volume is much larger. This enhances the contact with air, leading to much faster oxidation rates, while following similar oxidation mechanisms. Examples of other experimental procedures:

- 10g of oil in Petri dishes heated at 70°C for 8-56 days. Transmission spectra on KBr disks[3]
- Heating at 80 to 300°C in porcelain capsules for 20 and 40 min. Transmission spectra on KBr disks [24]
- 1 liter during 4x8h at 147, 171 and 189°C, 100 µl are deposited on the ATR crystal [18]
- samples from fast food restaurants with 4 cycles frying cycle, kept at 4°C, ATR analysis [19]
- Heating at 130-275ºC of 50 ml of oil in metallic beakers for 30min. Exposition for 1, 2 and 3h to UV radiation. 2 µl analysed by transmission on two KBr disks [14]

Difference to other methods

Earlier works by several authors[3, 14,15 ,22,23,24] use different procedures to assess the oxidation state of oils. In general these procedures involve the oxidation of large volumes of oil at the desired temperature, which leads to long experiment times. Small aliquots of the oil are then used to acquire the spectra during the experiment, which entails significant sampling variability and also implies the continual presence of an operator.

In this study, we present a new method for fast simultaneous oxidation-analysis of oils using a heated MIR-ATR. Results are presented for three different oils – sunflower, olive and canola - as well as for sunflower heated at three different temperatures.

2. Experimental

2.1. Samples:

We choose three kind of edible oils which are the most consumed .

Theses oils are very different about their composition of fatty oils and tocopherols, the most important components in oils.

Tableau 1 : Percentage of fatty acids in some vegetable oils (%)[25]

Fatty acids		Canola	Olive oil	Sunflower oil
Myristic acid	**C14:0**	0,1-0,2	≤ 0,05	0-0,1
Palmitic acid	**C16:0**	3,0-5,0	7,5-20,0	5,5-7,7
Palmitoleic acid	**C16:1**	0,2-0,6	0,3-3,5	0-0,3
Stearic acid	**C18:0**	1,0-2,0	0,5-5,0	2,8-6,5
Oleic acid	**C18:1**	52,0-67,0	55,0-83,0	14,0-38,0
Linoleic acid = ω6	**C18:2**	16,0-24,8	3,5-21,0	48,2-74,2
Linolenic acid = ω3	**C18:3**	6,5-14,0	≤ 0,9	0-0,1
Arachidic acid	**C20:0**	0,2-0,8	≤ 0,6	0,2-0,4

Eicosenoïc acid	**C20:1**	0,9-2,4	≤ 0,4	0-0,2
Behenic acid	**C22:0**	0,1-0,5	≤ 0,2	0,7-1,3
Lignoceric acid	**C24:0**	0-0,2	≤ 0,2	0-0,4

Table 2 : Tocopherol of some vegetable oils (mg / kg)[25]

Oils	Canola	Olive	Sunflower
α-tocopherol	100 –400	63 –227	400 –1000
β-tocopherol	0 – 150	0 –2	0 –60
γ-tocopherol	180 –780	5 –15	0 –60
total amount	**400 –2700**	**68 –244**	**400 –1600**

Sunflower, olive and canola oils were acquired in the local supermarket. The three oils were chosen in respect to their usual proportions of fatty acids and tocopherols. Of the three, sunflower oil is the richest in α-tocopherol and linoleic acid (C18:2). It is the most consumed if France, its main fatty acid is subject to oxidation [26] although it has high antioxydant content. Olive oil is rich in oleic acid (C18:1) and is therefore less oxidable despites its low tocopherol content. Canola oil is rich in γ-tocopherol and contains both some highly oxidisable linolenic acid (C18:3) and much oleic acid, that is an oxidation resistant fatty acid. Its behavior was expected to be particularly interesting as, upon heating, it rapidly produces a grassy, fishy odour.

The tocopherol measurements were made by HPLC-UV (Perkin-Elmer) following ISO9936 in the laboratory and we obtained the results showed in the table 3.

Table 3 : Experimentally determined tocopherol content of vegetable oils (mg / kg).

Oils	**Tocopherol** (α and γ) **(mg/kg)**
Canola	**460**
Olive	**187**
Sunflower	**509**

Samples of the sunflower oil were also stripped of tocopherol using the method described by Yoshida *et al.*[27]. Alumina (WN3 neutral, Sigma) was activated at 200ºC for four hours. The oil samples and activated alumina in a ratio of 1:1 (v:w) were then filtered on a fritted glass disc using a vacuum pump.

2.2. Thermal treatment :

75 µL of oil were heated directly on the ATR crystal for 60 minutes at 3 different temperatures (130°, 150°, 170°C). To contain the oil on top of the ATR crystal, a miniature flat-frying system was developed using a Viton® fluoroelastomer gasket with 11.5 mm internal diameter within which 75µL of the sample were deposited using a micropipette. The degradation reactions are very fast under these conditions due to the large sample surface to volume ratio.

2.3. MIR spectra acquisition and data analysis:

Spectra were acquired against air reference using a Vector 33 spectrophotometer (Bruker Optiks GmbH, Ettlingen, Germany), equipped with a KBr beamsplitter and DTGS detector. The sample was deposited on a temperature controlled "Golden Gate" ATR apparatus (Specac, London, UK), regulated to the desired constant temperature. The wavelength range measured was 4000-600 cm^{-1} with 4cm^{-1} resolution and 32 scans accumulated, taking approximately 32 seconds for each spectrum. Spectral acquisition started every minute, giving 60 spectra for each sample in one hour. OPUS® 6.5 software (Bruker) was used for instrument control and data acquisition.
The ATR crystal and its surroundings were thoroughly cleaned between samples using ethyl acetate, to avoid that any of the sample sticks to the crystal.
Data analysis was carried out using Matlab® version 7.6.0.324 (R2008a).

3. Results and discussion :

In different studies [15,18,22] , one can see the appearance of hydroperoxides around 3500 cm^{-1} and aldehydes around 1728 cm^{-1} [15,22]. One can observe other modifications around 1000 cm^{-1}, in particular increase of trans isomers at 967 cm^{-1} [15]. One can also see the decrease of cis C=C-C-H in 3008 cm^{-1} [22].

As our objective was to show the feasibility of studying the thermal degradation of edible oils using the proposed setup, particular wavelengths representative of functional groups known to be modified were selected for presentation.
Figure 1a shows the infrared spectra of sunflower oil heated at 150ºC for 60 minutes. The selected wavelengths are indicated with vertical dotted line and are presented in Table 4. These bands were

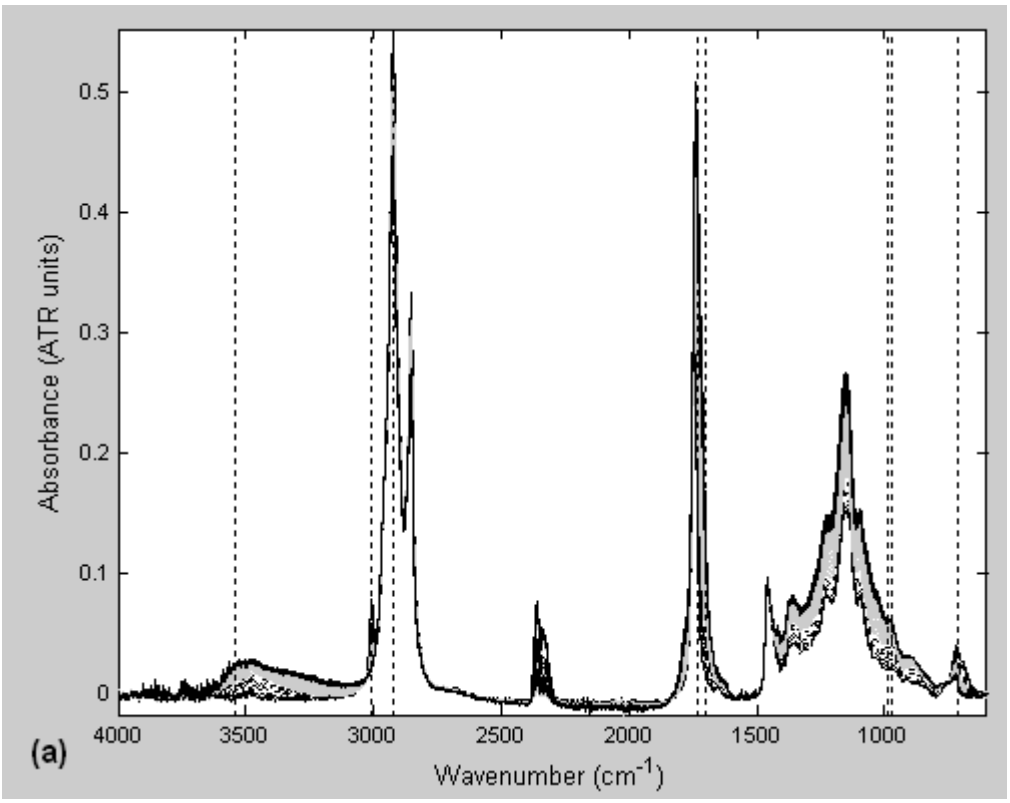
chosen because of their relation to reaction mechanisms proposed in the literature and their attributions are as shown in Table 4.

Table 4: Selected wavelengths for the oxidation experiments. Observed experimental wavelength, literature wavelength, functional group attribution and references.

Experimental Wavenumber (cm^{-1})	Wavenumber (cm^{-1})	Functional group	Reference
3539	**3360- 3560**	Hydroperoxides	Guillén M.D. et al., (2002)[22], (1997)[2]
	3530	alcohols, secondary oxidation products hydroperoxides	Guillen M.D. (2002) [22]
	3444	Hydroperoxides	Guillen M.D. et al (2002) [22]
	3020	assymmetrical stretching C=C–H "*cis*"	Guillen M.D. et al (1999)[3]
	3018	assymmetrical stretching C=C–H "*cis*"	Magdi M.M. *et al*., (1990) **[28]**
	3008	C=C-H (cis) stretching	Moya Moreno MCM et al (1999)[23]
3008	**3007**	symetrical stretching C=C–H "*cis*"	Vlachos N. *et al*., (2006) [144]
	3007	C=C-H (cis) stretching	Guillen M.D. et al (1997)[2]
	2935	Asymetrical stretching CH2	Magdi M.M. et al., (1990)[28]
	2929	Asymetrical stretching CH2	Guillen M.D. et al (1999) [2]
	2929	Asymetrical stretching CH2	Innawong B. (2004) [19]
2922	**2924**	Asymetrical stretching CH2	Guillen M.D. et al (1997) [2]
	2922	Asymetrical stretching CH2	Safar M. and al (1994) [30]
	2922	Asymetrical Stretching CH2	Van de voort et al (1994) [15]
	1749	C=O ester stretching	Yang H. et al. (2005) [6]
1732	**1746**	C=O ester stretching	Guillen M.D. (1997[2], 1999[3], 2002[22], 2007[29])
	1746	elongation C=O ester	Vlachos N. *et al*., (2006) [14]
	1743	elongation C=O ester	Safar M. and al (1994) **[30]**
1699	**1728**	aldehydes, ketones	Guillen M.D. et al (2002) [22]
	1658	elongation C=C (*cis*-olefins)	Guillen M.D. et al (1999) [3]
	1654	elongation C=C (*cis*-olefins)	Vlachos N. *et al*., (2006) [14]
985	**988**	C=C-H trans-trans conjugated dienic system or cis-trans conjugated double bond	Guillen M.D. et al (2007)
	987	C=C-H conjugated trans double bond	Van de voort et al (1994) [15]
	974	C=C-H Isolated trans double bond	Guillen M.D. and al. (1999) [3]
970	**967**	bending C=C–H "trans"	Christy A.A. et al (2003) [4]
	967	bending C=C–H "trans"	Guillen M.D. et al (1999) [3]
	967	bending C=C–H "*trans*"	Belton P.S. *et al*., (1988) [31]

	966	bending C=C–H "*trans*"	Safar M. and al (1994) [30]
	950	C=C-H cis-trans conjugated double bond	Guillen M.D. et al (2007)[29]
710	**723**	bending CH_2 ("rocking") and out-of-plane bending =C-H "*cis*"	Vlachos N. *et al*., (2006) [14]

Figure 1b shows the same spectra after subtraction of the spectrum at t_0. The modifications due to the thermal degradation of the oil are more evident than in Figure 1a.



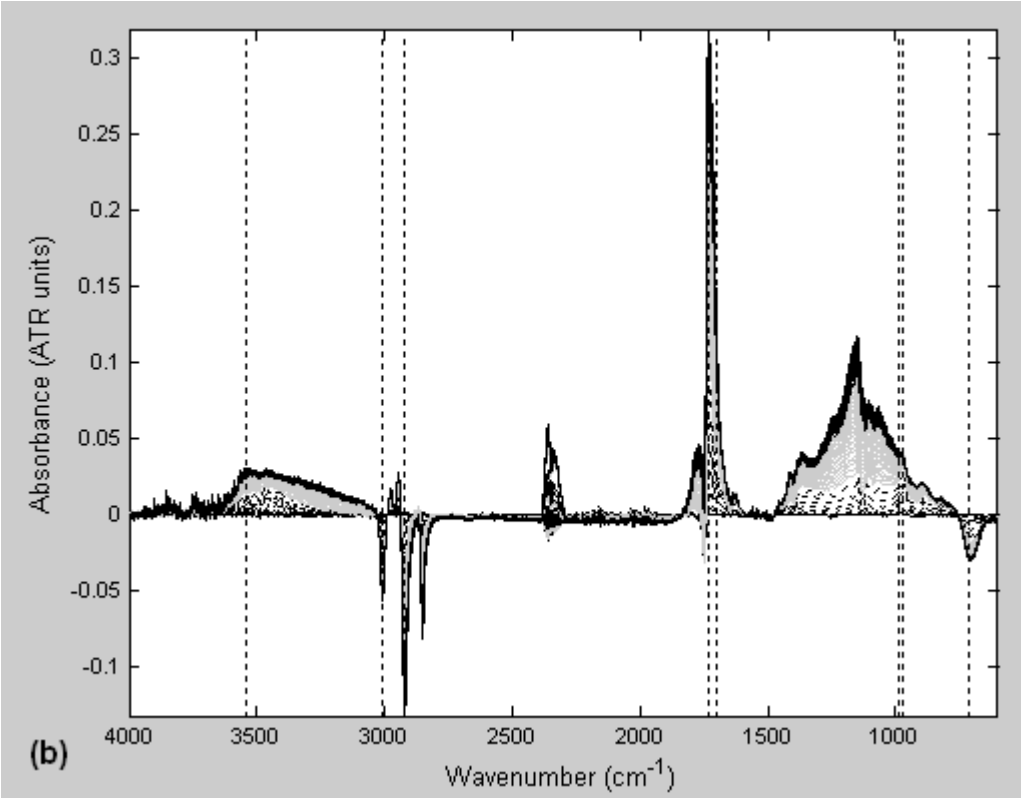


Figure 1 : (a) 60 baseline-corrected spectra of Sunflower oil submitted to 150ºC temperature in the ATR crystal during 1 hour. t_0 - t_{20} (---); t_{21} - t_{40} (grey); t_{41} - t_{60} (black). (b) the same data subtracting the spectrum at t_0. Vertical lines indicate the wavelengths which were selected for detailed analysis.

3.1. Evolution of selected wavelengths in 3 different oils at 150ºC

The evolution of all selected wavelengths is shown individually in Figure 2. As expected from the literature, there is an increase in absorbance with time for all wavelengths except for 3008 and 710cm^{-1}. The increase in bands is obviously due to the formation of products while the decreases at 3008 and 710cm^{-1} are related to the disappearance of –cis bonds due to isomerisation or oxidation. Most of the curves for olive and especially sunflower oil observed in Figure 2 have three phases:

- They start with different latency periods during which there is little variation in the intensity.
- They then increase/decrease at different rates for the different functional groups and oils.
- The rate decreases until it eventually comes to zero (depending on the band and type of oil).

Canola oil behaves differently to the others, indicating a greater stability towards the reactions giving rise to these spectral changes. Most canola curves do not show the same behaviour as the other two oils, as 60 minutes at 150ºC is not sufficient to provoke a comparable transformation.

It is possible to observe a latency period in the beginning of the reaction, which generally increases in the following order Canola > Olive > Sunflower, confirming Canola oil as the most stable of the three oils at 150ºC.

The rates of change of the oils are the inverse of that of the latency period, being greatest for sunflower oil.

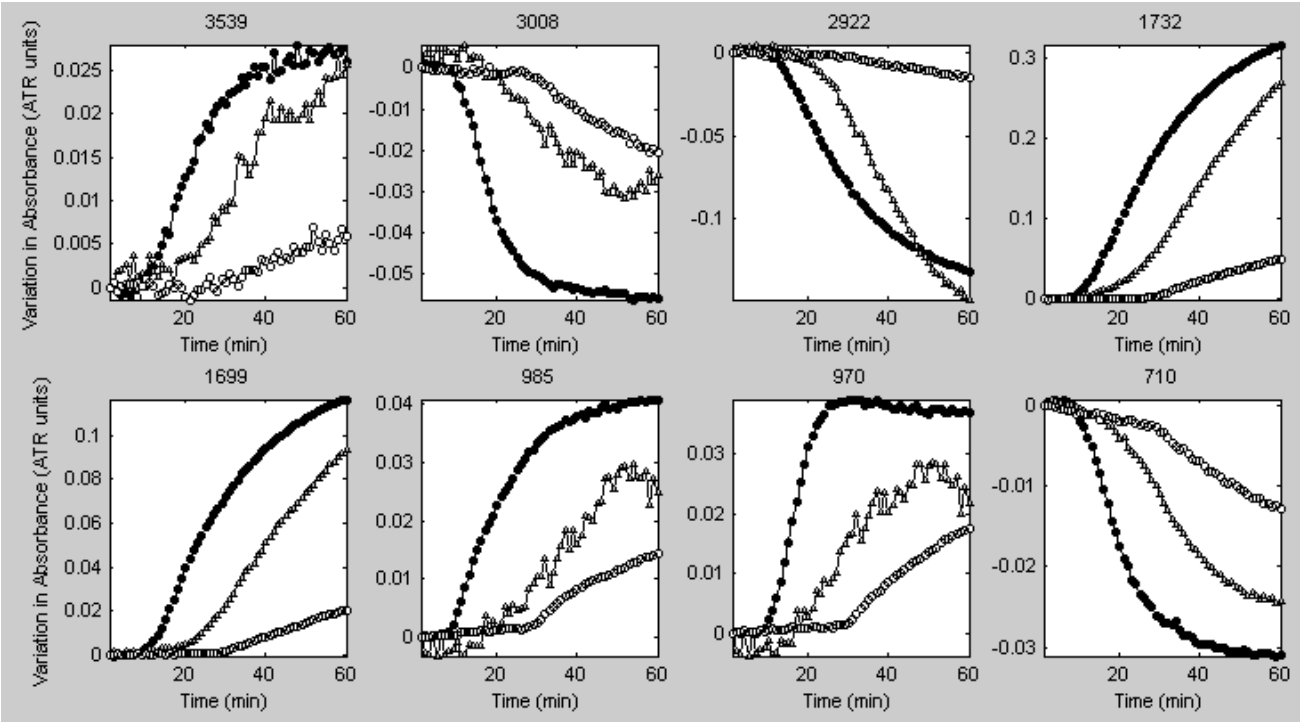


Figure 2 : evolution of individual selected wavelengths during time for the experiment with the sunflower (black circles), olive (grey triangles) and canola (white circles) oils.

Figure 2 shows an increasing tendency for the two bands at 985 and 970 cm^{-1} followed by stabilization. The absorption at 970 appears to have a higher rate of change than that at 985 cm^{-1}, although in the end both present more or less the same intensity. However this apparent behaviour is an artifact due to the broad band to their left (centered at 1155cm^{-1}) continuously increasing, and which forces those two bands to increase too. In Figure 3a shows the same region in which those bands absorb, after subtracting only the first spectrum. A spline[32] baseline correction of each spectrum using a polynomial passing through three points in the region 1051-926 cm^{-1}, gives the results in Figure 3b.

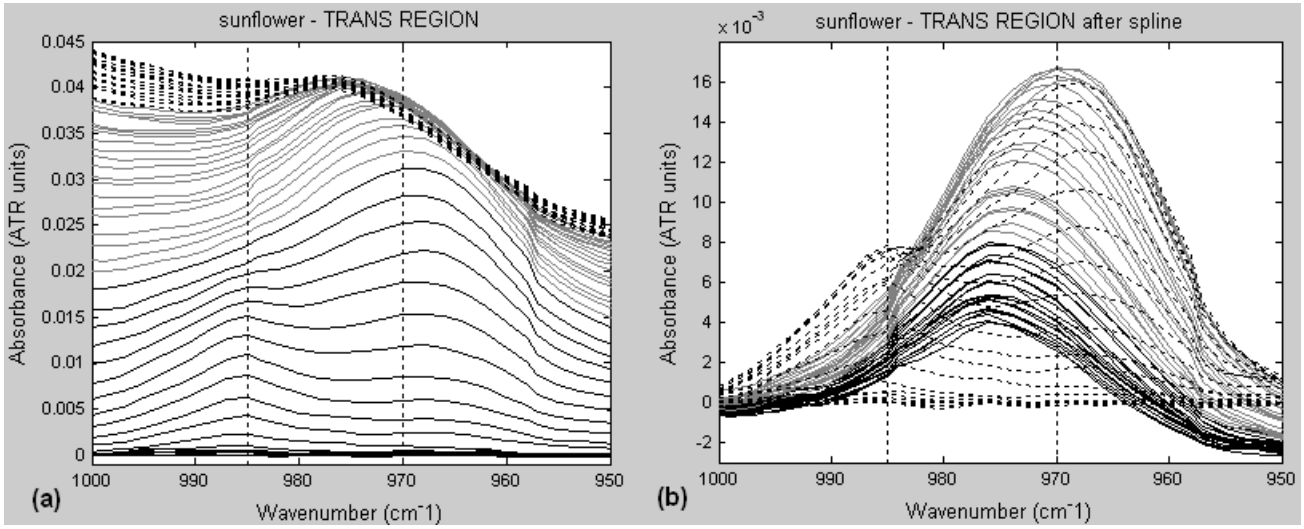


Figure 3 : Detail of the 1000-950 cm^{-1} region before (a) and after (b) correction with spline. t_0 - t_{20} (---); t_{21} - t_{40} (grey); t_{41} - t_{60} (black).

Plotting the maxima of the bands at 985 and 970 cm^{-1} over time gives the results shown in Figure 4a and Figure 4b. Both bands in fact present a period of latency, followed by an increase to reach a maximum and finally a decrease to more or less the original position. Canola oil does not present exactly the same behaviour in the first 60 minutes as the other two oils, because of its lower reaction rate, but there is no reason to suppose it will not evolve in the same after some more time.

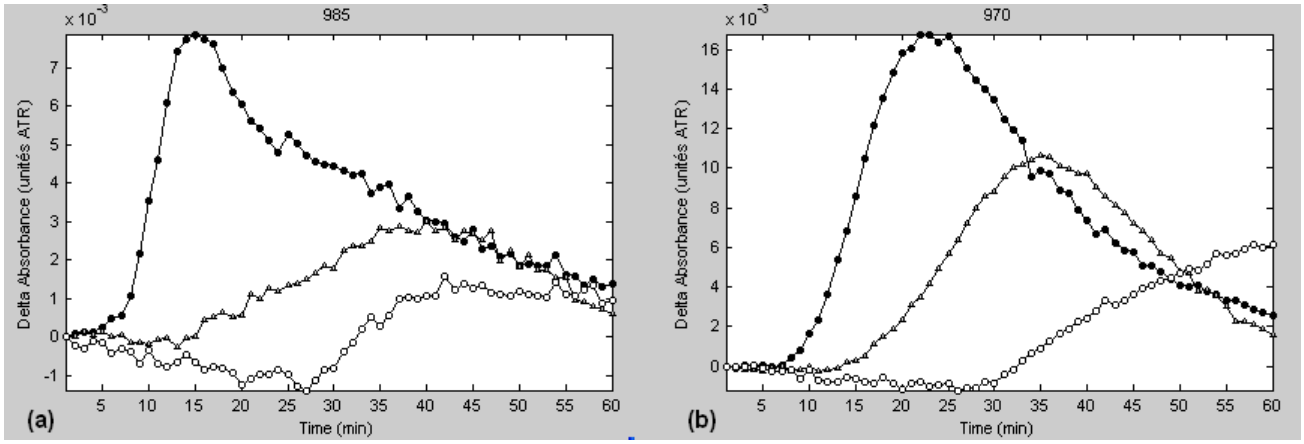


Figure 4 : Evolution of wavelengths 985 (a) and 970 cm^{-1} (b) after baseline correction using a spline function for the experiment with the sunflower (black circles), olive (grey triangles) and canola (white circles) oils.

The bands present absorption maxima at different times. For sunflower oil the maximum of absorption for the band at 985 cm^{-1} is at 15 minutes and for the band at 970 cm^{-1} it is after 23 minutes. Also, the absorption maximum of the band at 970cm^{-1} is approximately twice that of the band at 985 cm^{-1}.

The same type of phenomenon was already observed by van de Voort *et al.* [15]. As mentioned in Table 4, the peak at 985 cm^{-1} can be assigned to trans conjugated double bonds, present in either

isomeration products such as CLA (conjugated linoleic acid) or oxidation products such as 2,4-decadienal. Similarly, the peak at 970 cm^{-1} can be attributed to non conjugated trans double bonds present in either isomerisation products such as elaidic acid, or oxidation products such as 3-nonenal or 2-heptenal. There are at least two reasons that explain the decrease of these two peaks : on the one hand, isomerisation products are subject oxidation, that is degradation. On the other hand, oxidation products are volatile and tend to escape from the oil especially at the temperatures of this experiment. The difference in the time of heating at which the maximum peak intensity appears may be related to the speed of formation of the conjugated or non-conjugated products, this speed itself being related to the energy necessary to form the radical involved in the reaction. The mechanism illustrated in Figure 5 shows that 2,4-decadienal is formed together with an alkyl radical during a single final step, while 3-nonenal is formed by a more complex scheme which involves first a vinyl radical formation and then combination with an hydroxyl radical. Since the energy required to form a vinyl radical is approximately 35 kJ/mol higher than the energy required to form an alkyl radical and since hydroxyl capture cannot be instantaneous, 3-nonenal is expected to appear more slowly than 2,4-decadienal. Another tentative of explanation is that though it is not favorable compared to hydrogen removal at carbon 11 because approximately 85 kJ/mol higher, hydrogen removal at carbon 8 or carbon 14 may occur. Generated radicals would then produce non-conjugated trans fatty acids or degradation products.

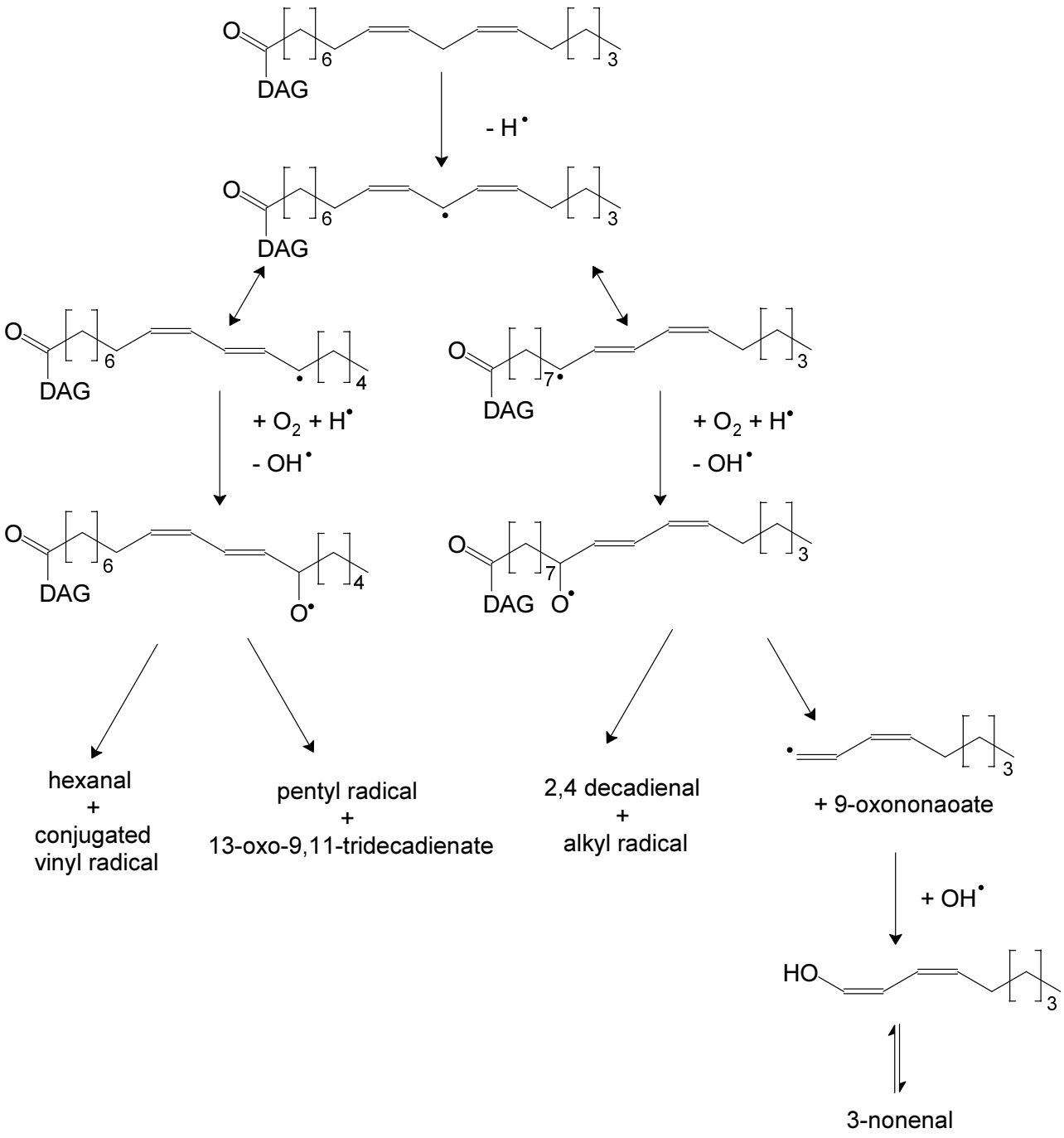


Figure 5 : Volatile compound formation from linoleic acid during autoxidation. DAG : diacylglycerol

3.2. Evolution of selected wavelengths in sunflower oil with and without tocopherol at 3 different temperatures

Several conclusions may be drawn from an examination of the curves acquired at different temperatures.

The thermal degradation of sunflower oil with and without tocopherol at 3 different temperatures is in general agreement with the literature. Higher temperatures reduce the latency times and accelerate the reactions, while the presence of tocopherol increases the latency period and slows down the reactions.

As can be seen in Figure 6, all the curves belonging to oil without tocopherol undergo modifications before the ones with tocopherol. This difference in behaviour is larger at lower temperatures.

Sunflower oil without tocopherol at 130ºC does not change to a large extent and with tocopherol does not appear to change at all after 60 minutes.

The curves in Figure 6 for the bands at 985 and 970 cm^{-1} again being influenced by the broad band at higher wavenumbers nearby, the same spline baseline correction as before was applied, giving the results shown in Figure 7.

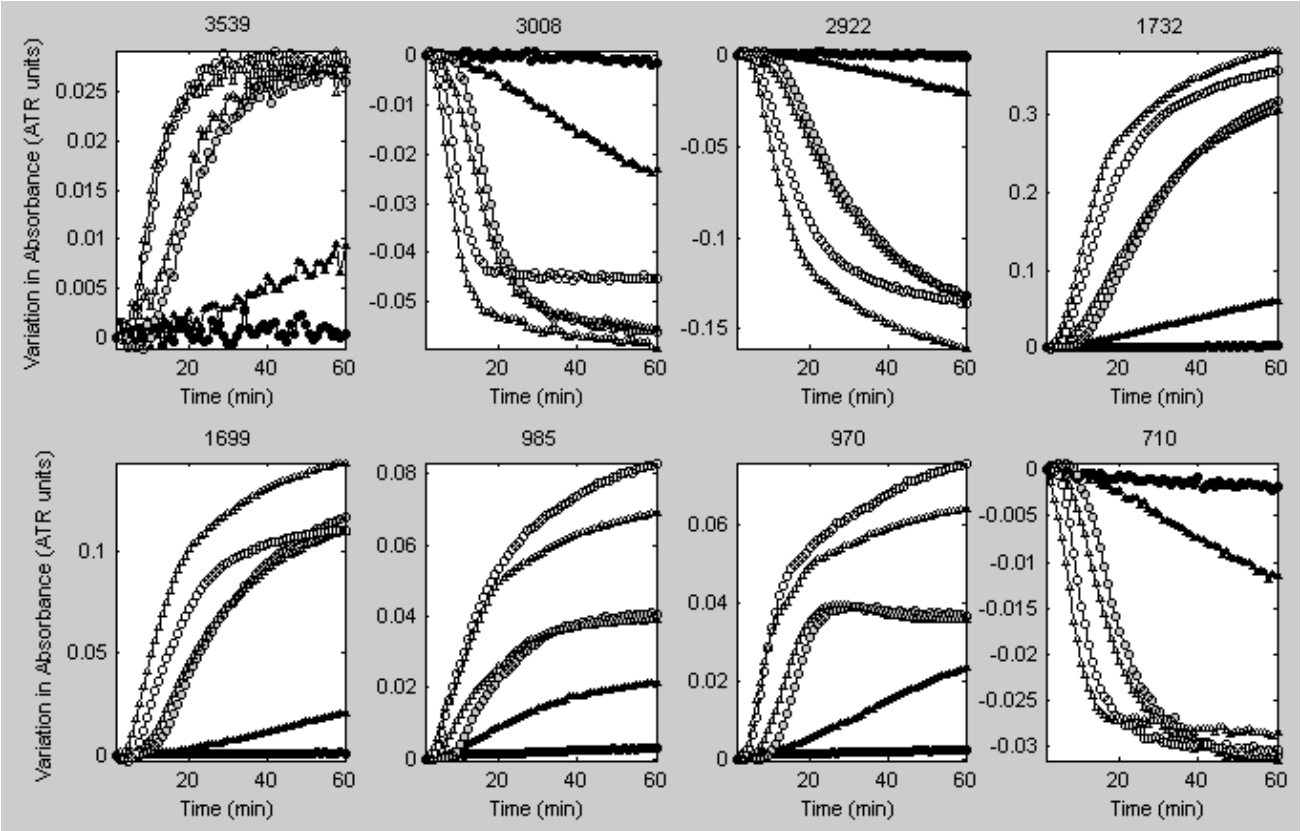


Figure 6: Evolution of individual selected wavelengths during time for the experiment with the sunflower oil with (circles) and without tocopherol (triangles) submitted to 130 (black), 150 (grey) and 170ºC (white).

After correction, the curves show the same behavior as before. Higher temperatures and oils without tocopherol present shorter latency periods and faster reactions. In Figure 7a and Figure 7b, the

sunflower oil at 170ºC does not even show a latency period and even at 130ºC, with tocopherol, no significant changes are visible.

The same pattern (at 170 and 150ºC) of increasing-maximum-decreasing absorbance is observed for both bands as before.

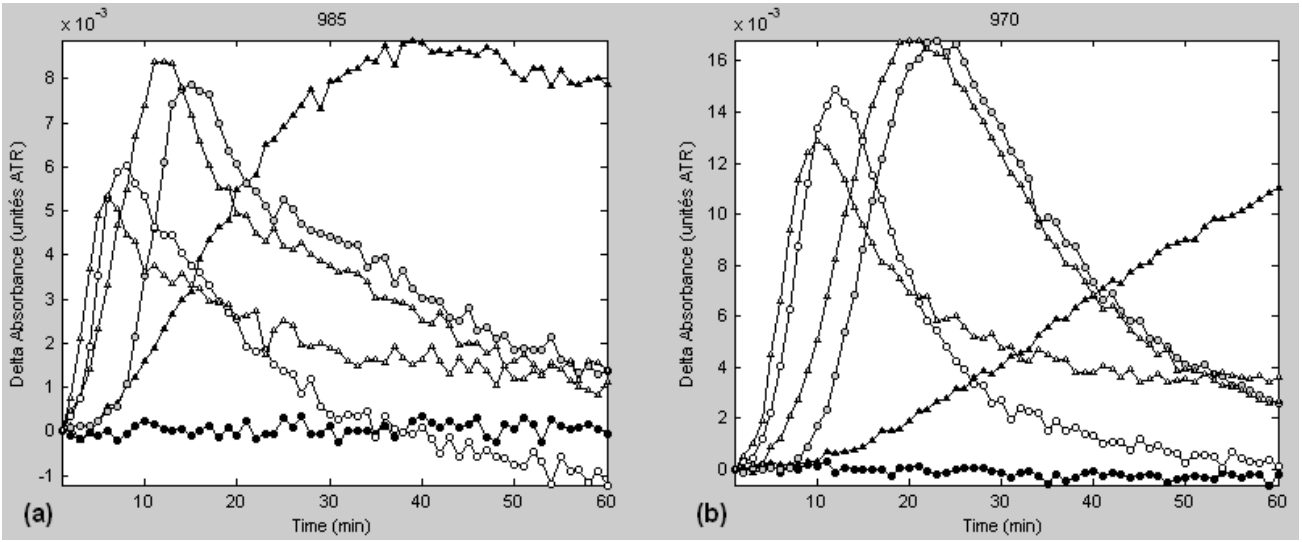


Figure 7 : Evolution of wavelengths 985 (a) and 970 cm^{-1} (b) after baseline correction using a spline function for the experiment with the sunflower oil with (circles) and without tocopherol (triangles) submitted to 130 (black), 150 (grey) and 170ºC (white).

Oxidation and isomerisation

In all cases, curves corresponding to *cis-trans* isomerisation, (3008 and 970 cm^{-1}) on the one hand and curves for oxidation reactions (3539 cm^{-1}, 1732 cm^{-1} and 1628 cm^{-1}) on the other all show the same type of increasing or decreasing evolution. This would tend to indicate that both the isomerisation and oxidation reactions have the same rate-determining step. It is known[12] that the first step of oil oxidation, i.e. radical formation, can lead to both reactions. The present results indicate that this first step is rate-determining and so the other steps in the two reaction schemes do not significantly influence the total reaction rates. The hypothesis of a single fatty free radical rate-determining step is comforted by the fact that the antioxidant tocopherol, acting as a radical scavenger, has the same effect on the latency period for isomerisation and oxidation.

Canola behaviour

According to these mid-infrared studies, canola oil would appear to be the most stable although when heated it produces more volatile oxidation products than does olive oil[33].

Thus, it must be noticed that the chemical information provided by the present approach is global. On the one hand, canola is globally the most stable oil because of its high oleic acid content and presence of a noticeable quantity of tocopherols. On the other hand, the linolenic acid in canola oil oxidizes quickly, generating strong off-flavours.

4. Conclusion

A new, easy-to-use and rapid method for monitoring the thermal degradation of edible oils using MIR-ATR is presented. The heated ATR apparatus simultaneously allows the fast oxidation-isomerisation of oils and acquisition of spectra, facilitating the study of their modifications over time. Some characteristic wavelengths were selected to follow the heat-induced modifications.

The method was applied to three different edible oils: sunflower, olive and canola. Sunflower is the least resistant to oxidation at 150ºC, exhibiting shorter latency periods and faster reaction curves for the selected wavelengths. Canola is, as expected, the most stable of the three, due to its fatty acid composition.

The relationship of oxidation to temperature and to the absence of tocopherol was also studied, confirming faster reactions at higher temperatures and in absence of tocopherol.

The method looks promising for the rapid analysis of the thermal degradation of oils.

Acknowledgements

The authors thank Thi Huyen Tram Nguyen for the determination of tocopherols.

References

[1] Riaublanc, A.; Bertrand, D.; Dufour, E . Chap 6 lipides. In. *La spectroscopie infrarouge et ses applications analytique,* Tec & Doc. ed. 2e ,Lavoisier Paris, Bertrand, D.; Dufour, E . **2006,**141-174.

[2] Guillen, M.D.; Cabo, N. Infrared Spectroscopy in the study of edible oils and fats . *J. Sc. Food Agric*; **1997**, 78, 1-11.

[3] Guillen, M.D. ; Cabo, N. Usefulness of the frequency data of the Fourier transform infrared spectra to evaluate the degree of the oxidation of edible oils. *J. Agric. Food Chem*. **1999**, *47*, 709-719.

[4] Christy, A.A. ; Egeberg, P..K. Ostensen, E.T. Simultaneous quantitative determination of isolated trans fatty acids and conjugated linoleic acids in oils and fats by chemometric analysis of the infrared profiles. *Vibrational Spectroscopy.* **2003,** *33*, 37-48.

[5] Tapp, H.S.; Defernez, M.; Kemsley, E.K. FTIR spectroscopy and multivariate analysis can distinguish the geographic origin of extra virgin olive oils. *Journal of agricultural and food chemistry*. **2003**, *51* (21), 6110-6115.

[6] Yang, H.; Irudayaraj, J.; Paradkar, M.M. Discriminant analyis of edible oils and fats by FT IR, FT-NIR and FT-Raman spectroscopy. *Food chemistry*. **2005**, *93* (1), 25-32.

[7] Ozen, B.F.; Mauer, L.J. Detection of hazelnut oil adulteration using FT-IR spectroscopy. *Journal of agricultural and food chemistry.* **2002**, *50* (14), 3898-3901.

[8] Sinelli, N.; Cosio, M.S.; Gigliotti, C.; Casiraghi, E. Preliminary study on application of mid infrared spectroscopy for the evaluation of the virgin olive oil freshness. *Analytica chimica acta***. 2007,** *598* (1), 128-134.

[9] Al-Alawi, A.; van de Voort, F.R.; Sedman, J.; Ghetler, A. Automated FTIR analysis of free fatty acids or moisture in edible oils*. Journal of the association for laboratory automation.* **2006,** *11* (1), 23-29.

[10] Ahmed, M.K.; Daun, J.K.; Przybylski, R.; FT-IR based methodology for quantitation of total tocopherols, tocotrienols and plastochromanol-8 in vegetable oils*. Journal of food composition and analysis.* **2005**, *18* (5), 359-364.

[11] Sherazi, S.T.H.; Kandhro, A.; Mahesar, S.A.; Bhanger, M.I.; Talpur, M.Y.; Arain, S. Application of transmission FT-IR spectroscopy for the trans fat determination in the industrially processed edible oils. *Food chemistry*. **2009**, *114* (1), 323-327.

[12] Choe, E.; Min, D.B.; Mechanisms and factors for edible oil oxidation. *Comprehensive Reviews in food science and food safety*. **2006,** *5*, 169-186.

[13] Choe, E.; Min, D.B.; Chemistry of deep fat frying oils. *Journal of Food Science*. **2007,** *72*, 77-86.

[14] Vlachos, N; Skopelitis, Y.; Psaroudaki, M.; Konstantinidou, V.; Chatzilazarou A.; Tegou, E. Applications of Fourier transform-infrared spectroscopy to edible oils. *Analytica Chimica Acta,* Elsevier. **2006**, *573-574*, 459-465.

[15] Van de Voort, F.R.; Ismail, A.A.; Sedman, J.; Emo, G. Monitoring the oxidation of edible oils by Fourier Transform infrared Spectroscopy*. J.Am.Oil Chem.S.***1994**. *71*, 243-253.

[16] Cert, A.; Moreda, W.; Pérez-Camino, M.C.; Chromatographic analysis of minor constituents in vegetable oils*. Journal of Chromatography A..* **2000**, *881*, 131-148.

[17] Kamal-Eldin Afaf ; Effect of fatty acids and tocopherols on the oxidative stability of vegetable oils. *Eur.J. Lipid Sci.Technol*. **2006**, *58*, 1051-1061.

[18] Moros, J.; Roth, M.; Garrigues, S.; de la Guardia M.; Preliminary studies about thermal degradation of edible oils through attenuated total reflectance mid-reflectance mid-infrared spectrometry. *Food Chemistry*. **2009,** *114*, 1529-1536.

[19] Innawong, B.; Mallikarjunan, P.; Irudayaraj,J.; Marcy, J. E. The determination of frying oil quality using Fourrier transform infrared attenuated total reflectance**.** *Lebensm-Wiss. u-Technol*. **2004**, 37, 23-28.

[20] Inon, F.A.; Garrigues, J. M.; Garrigues, S.; Molina, A.; de la Guardia, M. Selection of calibration ste samples in determination of olive oil acidity by Partial least squares attenuated total reflectance Fourier transform infrared spectroscopy. *Analytica chimica acta*. **2003**, *489*, 59-75.

[21] Maggio, R.M. ; Kaufman, T.S.; Del Carlo, M. ; Cerretani, L.; Bendini, A.; Cichelli, A.; Compagnone, D. Monitoring of fatty acid composition in virgin olive oil by Fourier transformed infrared spectroscopy coupled with partial least squares ; *Food Chemistry*. **2009,** *114*, 1549-1554.

[22] Guillén, M. D.; Cabo, N. Fourier transform infrared spectra data versus peroxide and anisidine values to determine oxidative stability of edible oils. *Food Chemistry*. **2002**, *77*, 503-510.

[23] Moya Moreno, M.C.M.; Mendoza Olivares, D.; Amezquita Lopez, F.J.;Gimeno Adelanto, J.V.; Bosch Reig, F. Analytical evaluation of polyunsaturated fatty acids degradation during thermal oxidation of edible oils by Fourier transform infrared spectroscopy. *Talanta*. **1999**, *50*, 269-275.

[24] Moya Moreno, M.C.M.; Mendoza Olivares, D.; Amezquita Lopez, F.J.; Gimeno Adelanto, J.V.; Bosch Reig, F. Determination of insaturation grade and trans isomers generated during thermal oxydation of edible oils and fats by FTIR. *Journal of molecular structure*. **1999**, *482-483*, 551-556.

[25] Harwood, J.; Aparicio, R.; Handbook of olive oil : Analysis and properties, An Aspen publication. Aspen publisher, Inc. Gaithersburg, Maryland, 2000, chap. 6, 130-134

[26] Kenaston, C.B.; Wilbur, K.M.; Ottolenghi, A.; Bernheim, F. Comparison of methods for determining fatty acid oxidation produced by ultraviolet irradiation. *Journal of the American Oil Chemists' Society*. **1955**, *55*, 33-35.

[27] Yoshida, H.; Kondo, I.; Kajimoto, G. Participation of free fatty acids in the oxidation of purified soybean oil during microwave heating. *Journal of the American Oil Chemists' Societ.* **1992**, *69*, 1136-1140.

[28] Magdi, M.M.; Richard, E.; Ecdonald, J.T.; David, J.A.; Samuel, W. Identification and quantification of 9-trans, trans-12-octadecadienoic acid methyl ester and related compounds in hydrogenated soybean oil and margarines by capillary gas chromatography/matrix isolation/Fourier infrared spectroscopy. *J. Agric. Food Chem*. **1990**, *38*, 86-92.

[29] Guillen, M.D. ; Goicoechea, E. Detection of primary and secondary oxidation products by Fourier Transform Infrared spectroscopy (FTIR) and ^{1}H nuclear Magnetic resonance (NMR) in sunflower oil during storage. *J. Agric. Food Chem*. **2007**, *55*, 10729-10736

[30] Safar M. Chap 2.1 : étude des lipides in *Comparaison des plages spectrales de l'infra rouge proche et moyen pour l'étude des produits agro-alimentaires*. PhD thesis . **1995**. 49-59

[31] Belton, P.S.; Wilson, R.H;, Sadeghi, H.; Orabchi, J.; Peers, K.E. A rapid method of the estimation of isolated trans double bands in oils and fats using Fourier transform infrared spectroscopy combined with attenuated total reflectance. *Lebensm. Wiss. U, Technol*. **1988**, *21*, 153-157.

[32] Zeaiter, M.; Rutledge, D. Preprocessing methods. in Comprehensive Chemometrics, volume 3, 121-231, Elsevier, Oxford (UK), **2009**.

[33] : Jeleń, H.H.; Obuchowska M.; Zawirska-Wojtasiak R.; Wsowicz E.; Headspace Solid-Phase Microextraction Use for the Characterization of Volatile Compounds in Vegetable Oils of Different Sensory Quality. *Journal of agricultural and food chemistry.***2000,** *48* (6), 2360–2367.

ANNEXES VIII - XI

Posters

ANNEXE VIII

Using Principal Component Transform to accelerate Outer Product – Partial Least Squares Regression calculations

Pinto, Rui[1]**; Barros,** António S.[2]**; Bouveresse,** Delphine. [1]; **Rutledge,** Douglas N.[1]**;**

[1]*Laboratoire de Chimie Analytique, Institut National Agronomique Paris-Grignon.*
16, rue Claude Bernard. 75005 Paris, France.

[2]*Departamento de Química, Universidade de Aveiro.*
Campus Universitário de Santiago. 3810-193 Aveiro, Portugal.

Corresponding author: rutledge@inapg.inra.fr

KEYWORDS
Principal Component Transform (PCT), Megavariate data, Outer-Product Analysis (OPA), Partial Least Squares Regression (PLS), Cross-Validation, Spectroscopy, Chemometrics.

ABSTRACT
Principal Component Transform (PCT) [1] is used to accelerate calculations and to reduce memory needs when applying multivariate methods like Partial Least Squares regression (PLS) to huge matrices in which the number of variables is much larger than the number of objects. In the PCT method, the objects within a matrix are transformed into a new smaller vectorial space by means of PCA. If the full-rank is used, these new matrices retain the same relations between objects and variables. Multivariate methods can then be applied using the (much) smaller matrix of scores, instead of the matrix of original variables. Just as in Fourier Transform, after obtaining the results in the new space one can inverse-transform the objects back to the original variables space. PCT has already been used to accelerate PLS regression model building [1].
Outer Product Analysis (OPA) [2,3] combines two matrices describing the same set of objects, in order to emphasize simultaneous changes in variables. Used with Partial Least Squares Regression (OP-PLS), it can relate these changes to a vector of chemical, physical or other properties [4].

The selection of the number of latent variables to use in the PLS model depends on its predictive ability for new samples, which is in general calculated by means of procedures like cross-validation. These methods calculate many PLS models with an increasing number of latent variables, which can be very time-consuming if one uses big data matrices.

Given the size of the OP matrices, common in spectroscopy, one can see the obvious advantage of using PCT with OP-PLS, in terms of both memory and time required to perform the calculations, particularly when doing cross-validation as one uses only a square matrix of scores with the same dimensions as the number of objects in the original matrix.

This poster explains the PCT-OP-PLS method and shows, by means of an example, that the results obtained by OP-PLS and PCT-OP-PLS are exactly the same. By using these relatively small matrices it is possible to obtain the same results as with the normal OP-PLS method. In the case of very large matrices, it may even be impossible to perform the calculations using the standard OP-PLS method.

REFERENCES

[1] Barros, A.S.; Rutledge, D.N.; "Principal Components Transform – Partial Least Squares (PCT-PLS): A novel method to accelerate Cross-Validation in PLS regression". Chemometrics and Intelligent Laboratory Systems, 2004, 73, p. 245– 255

[2] Di Natale, C.; Zude-Sasse, M.; Macagnano A.; Paolesse, R.; Herold, B.; D'Amico, A.; "Outer product analysis of electronic nose and visible spectra: application to the measurement of peach fruit characteristics". Analytica Chimica Acta, 2002, 459, p. 107–117.

[3] Jaillais, B.; Pinto, R.; Barros, A.S.; Rutledge, D.N.; "Outer-Product Analysis (OPA) using PCA to study the influence of temperature on NIR spectra of water". Vibrational Spectroscopy, 2005, Vol. 39 , 1 , p. 50 – 58.

Using Principal Component Transform to accelerate Outer Product –Partial Least Squares Regression calculations

Rui Pinto[1], António Barros[2], Delphine Jouan-Rimbaud Bouveresse[1], Douglas N. Rutledge[1]
[1]Laboratoire de Chimie Analytique, Institut National Agronomique Paris-Grignon, France
[2]Departamento de Química, Universidade de Aveiro, Portugal

INTRODUCTION

PCT

Principal Component Transform (PCT) [1] is used to accelerate calculations and to reduce memory needs when applying multivariate methods like Partial Least Squares regression (PLS) to huge matrices in which the number of variables is much larger than the number of objects. In the PCT method, the objects within a matrix are transformed into a new smaller vectorial space by means of PCA. If the full-rank is used, these new matrices retain the same relations between objects and variables. Multivariate methods can then be applied using the much smaller matrix of scores, instead of the matrix of original variables. Just as in Fourier Transform, after obtaining the results in the new space, one can inverse-transform the objects back to the original variables space. PCT has already been used to accelerate PLS regression model building [1].

OP-PLS

Outer Product Analysis (OPA) [2,3] combines two matrices describing the same set of objects, in order to emphasize simultaneous changes in variables. The Outer Product is the multiplication of each element of one vector by all the elements in the other vector. The resulting matrices are unfolded and concatenated row-wise, producing very wide matrices. Partial Least Squares Regression, for example, can then be used (OP-PLS) to relate these simultaneous changes to a vector of chemical, physical or other properties.
The selection of the number of latent variables to use in the PLS model depends on its predictive ability for new samples, and is in general calculated by means of procedures like cross-validation. These methods calculate many PLS models with increasing numbers of latent variables, which can be very time-consuming for big data matrices.

PCT-OP-PLS

Given the size of the OP matrices common in spectroscopy, one can see the obvious advantage of using PCT with OP-PLS, in terms of both memory and time required to perform the calculations, particularly when doing cross-validation as one uses only a square matrix of scores with the same dimensions as the number of objects in the original matrix.

OBJECTIVES

This poster presents how to apply the PCT-OP-PLS method and show, with a simple example, that the results obtained by OP-PLS and PCT-OP-PLS are exactly the same. The full work is used and also show that the PCT method is less time and memory consuming than the normal method. Relatively small matrices are used due to practical problems - in the case of very large matrices, it may even be impossible to perform the calculations using the standard OP-PLS method.

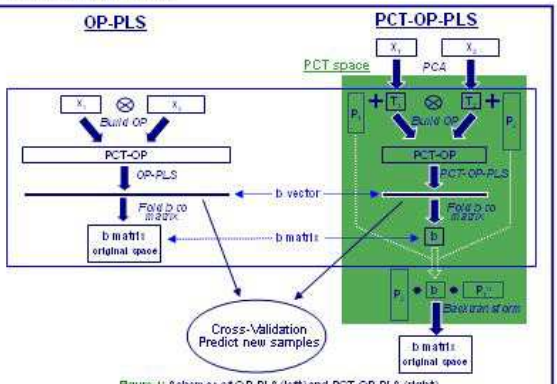
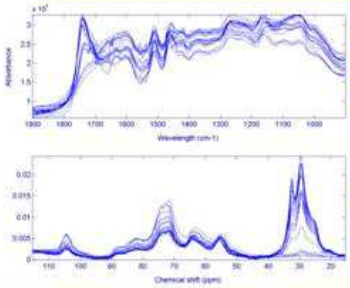


Figure 1: Schemas of OP-PLS (left) and PCT-OP-PLS (right).



Initial Data

20 FTIR-PAS and solid-state ^{13}C CP/MAS NMR spectra of cork powder with different percentages of suberine.

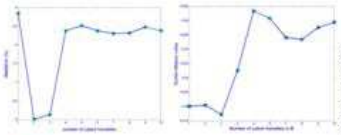
Matrices

FTIR-PAS spectral region: 1900-900 cm^{-1} (matrix = 20x260)
NMR spectral region: 15.34-115.24 ppm (matrix = 20x256)
Percentage of suberine for each sample (matrix = 20x1).

OP-PLS

Both sets of spectra were column-centred.
OP-PLS (OP matrix dimension 260x256)
PCT-OP-PLS (PCT-OP matrix dimension 20x400)
Both methods were performed and gave matrices of b-coefficients, loadings, scores and eigenvalues.

Figure 2: Initial FTIR-PAS (top) and solid-state ^{13}C CP/MAS NMR (bottom) spectra

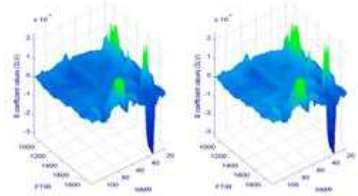


NOTE: The Durbin-Watson (DW) criterion [4] is given by:

$$DW = \frac{\sum_{i=2}^{n}(x_i - x_{i-1})^2}{\sum_{i=1}^{n}(x_i)^2}$$

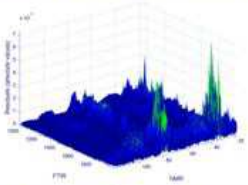
x_i and x_{i-1} being successive values in a vector and n the number of values. In the majority of cases, if a vector contains information, adjacent intensities are highly correlated and its DW tends to zero. "Noisy" vectors gives higher values. DW can therefore be used to estimate signal/noise of b-coefficients and loadings vectors [4].

Figure 3:
(Left) Root Mean Square Error of Cross-Validation (RMSECV, leave-1-out, in %)
(Right) Durbin-Watson criterion for the b-coefficients vectors of models with the first 10 Latent Variables for the normal OP-PLS (□) and the PCT-OP-PLS methods (○). The correlation between the results of the two methods is 1.00 for all the variables. 3 LVs were selected to build the model.



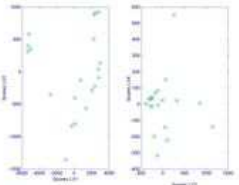
NOTE: Combining two spectral regions by means of OP and relating it to a property vector by means of PLS, one can sometimes obtain better models than with only one of the domains. Also, one can find links between well known variables in one of the domains (like MIR) and unknown variables in the other (like NIR, NMR).

Figure 4: Surfaces of b-coefficients for OP-PLS (left) and PCT-OP-PLS (right) with 3 Latent Variables. The correlation between the unfolded b vectors for the two methods is 1.00.



NOTE: To compare the b surfaces obtained by both methods, the absolute values of the differences between each point were calculated (Residuals). The maximum square value obtained is smaller than 4×10^{-6} while the residuals are all smaller than 10^{-12}. These values, 12 orders of magnitude smaller than the values of b, are due to rounding errors of the computer and are not significative.

Figure 5: Residuals (absolute values of the difference between each point) of b-coefficients (of Figure 4) of models with 3 latent variables calculated by the normal OP-PLS and the PCT-OP-PLS methods. The values are all smaller than 10^{-12}.

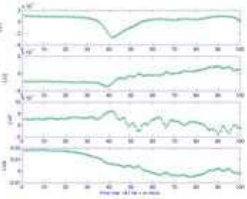


NOTE: The two methods are compared by superimposing the scores obtained by both techniques and by calculating the correlations between them.

OP-PLS and PCT-OP-PLS give the same results.

Only the scores for the first four Latent Variables are presented.

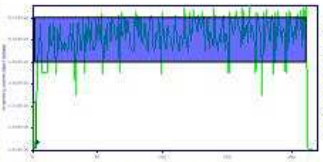
Figure 6: Scores plots of Latent Variables 1-4 for the normal OP-PLS (X) and the PCT-OP-PLS methods (+). The correlation of scores obtained by the two methods is 1.00 between all Latent Variables.



NOTE: The two methods are compared by superimposing the loadings obtained by both techniques and by calculating the correlations between them.

OP-PLS and PCT-OP-PLS give the same results.
Only the first 100 loadings for first four Latent Variables are presented.

Figure 7: Plots of loadings for the normal OP-PLS (X) and the inverse-PC-transformed loadings for the PCT-OP-PLS methods (+) for the first 100 points of Latent Variables 1-4.
The correlation of loadings obtained by the two methods is 1.00 for all Latent Variables



NOTE: The two techniques were compared by evaluating the time and memory needed to process Internal Cross-Validation with 10 Latent Variables and to build the chosen model with 3 Latent Variables. PCT-OP-PLS did it in about 1 second, while normal OP-PLS needed 206 seconds to perform. One can see the huge difference in memory needs by each method.

Figure 8: Memory usage during cross-validation (leave-1-out) for 10 Latent Variables using PCT-OP-PLS (—) and the normal OP-PLS method (—). The number of columns in the OP matrices used by each method is proportional to the length of the rectangles in blue (normal OP-PLS, top) and green (PCT-OP-PLS, bottom left)

CONCLUSIONS

The PCT-OP-PLS method gives exactly the same results as the OP-PLS method.

Given the size of the OP matrices common in spectroscopy, one can see the obvious advantage of PCT with OP-PLS, for both memory and time required to perform the calculations, particularly when doing cross-validation.

REFERENCES:

[1] Barros, A.S.; Rutledge, D.N.; "Principal Components Transform – Partial Least Squares (PCTPLS): A novel method to accelerate Cross-Validation in PLS regression", Chemometrics and Intelligent Laboratory Systems, 2004, 73, p. 245- 255
[2] Di Natale, C.; Surdo-Scipio, M.; Macagnano, A.; Esposito, R.; Agrosì, B.; D'Amico, A.; "Outer product analysis of electronic nose and visible spectra: application to the measurement of peach fruit characteristics", Analytica Chimica Acta, 2002, 459, p. 107–117.
[3] Jallais, B.; Pinto, R.; Barros, A.S.; Rutledge, D.N.; "Outer-Product Analysis (OPA) using PCA to study the influence of temperature on NIR spectra of water", Vibrational Spectroscopy, 2006, Vol. 39, 1, p. 60 – 65.
[4] Rutledge, D. N.; Barros A. S.; "The Durbin-Watson statistic as a morphological estimator of information content", Analytica Chimica Acta, 2002, 460, 279-294.

ACKNOWLEDGEMENTS:

Rui Pinto acknowledges a PhD. grant from the "Fundação para a Ciência e Tecnologia" (FCT) – Portugal

ANNEXE IX

Using the Principal Component Transform (PCT) framework to enable Outer Product (OP) - Partial Least Squares (PLS) regression analysis in huge matrices

Pinto, Rui[1]**; Barros,** António S.[2]**; Bouveresse,** Delphine. [1]; **Rutledge,** Douglas N.[1]**;**

[1]*Laboratoire de Chimie Analytique, Institut National Agronomique Paris-Grignon.*
16, rue Claude Bernard. 75005 Paris, France.

[2]*Departamento de Química, Universidade de Aveiro.*
Campus Universitário de Santiago. 3810-193 Aveiro, Portugal.

Corresponding author: rutledge@inapg.inra.fr

KEYWORDS
Principal Component Transform (PCT), Segmented Principal Component Transform (Seg-PCT), Megavariate data, Outer-Product Analysis (OPA), Partial Least Squares Regression (PLS), Cross-Validation, Spectroscopy, Chemometrics.

The basic idea of the Principal Components Transform (PCT) [1] method is to compress the original matrix into a new vectorial space with fewer variables, then use this new matrix to apply multivariate methods, obtain the results in the new space and finally transform the objects back into the original variables space.

If the number of variables in the studied matrix is so large that it is impossible even to create an OP matrix let alone perform PCA, then Segmented-PCT (Seg-PCT), which has already been used to analyse data sets too large to perform normal PCA [2], may be the method to apply.

In this approach, one segments the initial data matrix column-wise into several sub-matrices, then performs PCT on each of them and concatenate the resulting scores. These can then be used to calculate an OP matrix before using multivariate methods such as Partial Least Squares regression (PLS) to study it [3]. The number of objects in the original matrix must be

much smaller than the number of variables and the number of segments such that the matrix of concatenated scores is much smaller than the original one.

After obtaining the results in the new space one can back-transform the objects to the original variables space.

This poster demonstrates with a practical example how to use the PCT framework to perform OP-PLS. The Seg-PCT method is applied to an unfolded two dimensional Nuclear Magnetic Resonance (2D-NMR) dataset to perform discriminant Seg-PCT-OP-PLS of wine samples.

Because of the size of each spectrum and the finite memory of personal computers, it is impossible to directly create an OP matrix, let alone analyse it using PLS. With the use of Seg-PCT, it becomes possible. Internal cross-validation in the PCT space is used to choose the correct number of latent variables to use in the PLS model. Finally, some new samples are used to test the discriminant model.

REFERENCES

Barros, A.S.; Rutledge, D.N.; "Principal Components Transform – Partial Least Squares (PCT-PLS): A novel method to accelerate Cross-Validation in PLS regression". Chemometrics and Intelligent Laboratory Systems, 2004, 73, p. 245– 255

Barros, A.S.; Rutledge, D.N.; "Segmented principal component transform-principal component analysis". Chemometrics and intelligent laboratory systems, 2005, Vol. 78 , 1-2 , p. 125 – 137.

Jaillais, B.; Pinto, R.; Barros, A.S.; Rutledge, D.N.; "Outer-Product Analysis (OPA) using PCA to study the influence of temperature on NIR spectra of water". Vibrational Spectroscopy, 2005, Vol. 39 , 1 , p. 50 – 58.

Using the Principal Component Transform (PCT) paradigm to allow Outer Product (OP)-Partial Least Squares (PLS) regression analysis of datasets

Rui Pinto[1], António Barros[2], Delphine Jouan-Rimbaud Bouveresse[1], Douglas N. Rutledge[1]
[1]Laboratoire de Chimie Analytique, Institut National Agronomique Paris-Grignon, France
[2]Departamento de Química, Universidade de Aveiro, Portugal

Imagine you have matrices with so many columns you cannot even load them into memory or, when you can, your computer crashes when you try to do a multivariate statistical analysis. The PCT framework may solve your problem.

INTRODUCTION

PCT

The basic idea of the Principal Components Transform (PCT) [1] method is to transform the original matrix into a new vectorial space with fewer variables, then use this new matrix to apply multivariate methods, obtain the results in the new space and finally, if necessary, inverse-transform back into the original variables space, just as in Fourier Transform.

OP

An OP matrix [2,3] combines two matrices describing the same set of objects, in order to emphasize simultaneous changes in the two sets of variables. The OP matrix is the multiplication of each element of one row-vector by all the elements in the correspondent row-vector in the second domain, followed by column-wise concatenation of these vectors. After doing the same for all row-vectors, these are concatenated row-wise.

Segmented PCT

If the number of variables in the studied matrix is so large that it is impossible even to create an OP matrix let alone perform PCA, then Segmented-PCT (Seg-PCT), which has already been used to analyse data sets too large to perform normal PCA [4], may be the method to apply.

In this approach, the initial data matrix is segmented column-wise into several sub-matrices, PCT is performed on each of them and the resulting scores concatenated. These scores can then be used to calculate an OP matrix before using multivariate methods such as Partial Least Squares regression (PLS) to study it [3]. The number of objects in the original matrix must be much smaller than the number of variables and the number and size of the segments such that the matrix of concatenated scores is much smaller than the matrix.

After obtaining the results in the new space one can inverse-transform the objects into the original variables space.

OBJECTIVES

Demonstrate, by means of a schema and a practical example, how to use the PCT paradigm to perform OP-PLS. The Seg-PCT method is applied to an unfolded 2-Dimensional Nuclear Magnetic Resonance (2D-NMR) dataset to perform Seg-PCT-OP-PLS Discriminant Analysis of wine samples.

Perform internal cross-validation in the PCT space to choose the correct number of latent variables to use in the PLS Discriminant Analysis model.

Show how to predict new samples, in the PCT space, in order to test the Discriminant model.

Show how the PCT paradigm can be used in distributed processing.

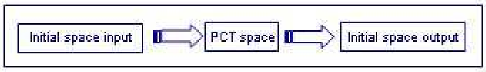
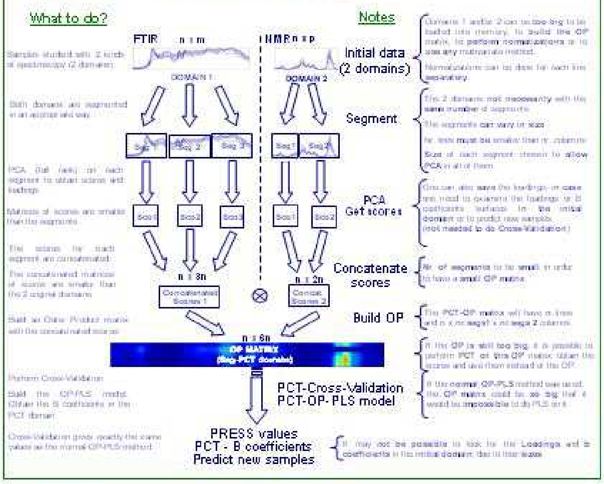


Figure 1: Schema of the PCT concept: after obtaining the results in the PCT space, the objects can be inverse-transformed to the original variables space.

THEORY



Prediction

1- Starting with the two vectors for a new sample, do the same segments of the vectors as when creating the model.

2- Multiply each segment by the corresponding loadings of the initial PCAs.

3- Concatenate the obtained scores for the two domains into two different vectors.

4- Create an OP vector for each sample using these two concatenated vectors.

5- Multiply the OP vector of each sample by the unfolded B coefficients vector from the PCT-OP-PLS model

Using distributed processing

Distributed processing can be used to save variables on other disks or to calculate PCA of the segments on different computers.

Showing parts of huge Loadings and B coefficients in the initial domain

Although the PCT-OP-PLS *Scores* and *Eigenvalues* are the same as those to be had using the standard method, PCT *Loadings* and *B coefficients* are not. They are in a different domain to the original data (the PCT domain).

As these two statistics may be much too big to load into memory, it is only possible to examine relatively small portions of them at a time, extracted from the PCT domain.

For B coefficient surfaces:

1- In the initial PCA loadings, select the Latent Variables and the variables you want to visualize in the two domains.

2- These new sub-matrices of loadings (L) are multiplied by the B coefficients surface in the PCT domain (B_{PCT}):

$$B_{surface} = L_{x2} \times B_{PCT}^{T} \times L_{x1}^{T}$$

3- The procedure is similar for the loadings X and W.

PRACTICAL EXAMPLE

Data:

Calibration: 54 2D-NMR spectra of wines of different types

External validation: 27 similar spectra

The OP matrix was calculated between the matrix and itself (Homo-OPA)

Data first column-centered and standardized.

Objective:

Perform PCT-OP-PLS Discriminant Analysis on the wine 2D-NMR data using the same matrix in both domains.

Perform Cross-Validation, choose the best model and predict new samples.

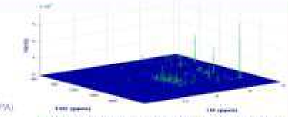


Figure 2: Example of 2D-NMR spectra of wine

Experimental details

Each spectrum is a matrix with dimensions 874 x 2048, which can be unfolded to give a vector with 1 789 952 values. The size of the matrix is 54 x 1 789 952 (corresponding to 773.26 Megabytes).

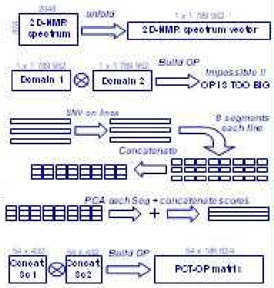
If the normal OP method was used, the resulting Homo-OP matrix to be used in the calibration would have dimensions 54 x (1 789 952 x 1 789 952), or $54 \times 3.2 \times 10^{12}$, corresponding to 173 000 Gigabytes.

On the dataset used, it is not even possible to perform any normalization of the initial domains, so SNV was performed individually on each line (spectrum). The standardised lines were then divided into 8 segments and each part of these line segments was concatenated vertically to give 8 matrices.

PCA was performed on each of the 8 segments. The scores for these 8 PCAs were concatenated to produce a concatenated scores matrix of dimensions 54 x 432, corresponding to only 186.62 Kilobytes.

Using Seg-PCT with 8 segments, we obtain from the two matrices of concatenated scores an (Homo) PCT-OP matrix of size 54 x 186 624 corresponding to 80.62 Megabytes.

The Cross-Validation for the calibration took 1.5 hours to perform.



CONCLUSIONS

- The PCT paradigm was used to perform OP-PLS on huge initial dataset.
- In the example shown, data size made it impossible to calculate an OP matrix on the original variables.
- Seg-PCT-OP-PLS has been compared with OP-PLS using small datasets and gives identical results.
- It is shown here how to perform cross-validation in the PCT space and select the best PLS model accordingly.
- To predict new samples, the objects must be transformed into the PCT space, before calculations.
- Even the huge matrices obtained after PLS (B coefficients surfaces, Loadings) may be visualized using PCT.
- PCT can be used in distributed processing, allowing several computers to do the calculations simultaneously.
- PCT can be successively applied to allow reduction of data in different phases of the method.
- Given the size of the OP matrices used in spectroscopy, one can see the obvious advantage of using Seg-PCT to perform OP-PLS, for both memory and time required to perform the calculations, particularly when doing cross-validation.

REFERENCES:

[1] Barros, A.S.; Rutledge, D.N.; "Principal Components Transform – Partial Least Squares (PCT-PLS): A novel method to accelerate Cross-Validation in PLS regression". Chemometrics and Intelligent Laboratory Systems, 2004, 73, pp. 245–255

[2] Di Natale, C.; Zude-Sasse, M.; Macagnano, A.; Paolesse, R.; Herold, B.; D'Amico, A.; "Outer product analysis of electronic nose and visible spectra: application to the measurement of peach fruit characteristics". Analytica Chimica Acta, 2002, 459, pp. 107–117.

[3] Jaillais, B.; Pinto, R.; Barros, A.S.; Rutledge, D.N.; "Outer-Product Analysis (OPA) using PCA to study the influence of temperature on NIR spectra of water". Vibrational Spectroscopy, 2005, Vol. 39, 1, pp. 50–58.

[4] Barros, A.S.; Rutledge, D.N.; "Segmented principal component transform – principal component analysis". Chemometrics and Intelligent Laboratory Systems, 2005, 78, pp. 125–137.

ACKNOWLEDGEMENTS:

Rui Pinto acknowledges a PhD. grant from the "Fundação para a Ciência e Tecnologia" (FCT) – Portugal

ANNEXE X

Using ANOVA-PCA for discrimination between FTIR-ATR spectra of Carragheenan gels in different concentrations and at different temperatures and classification of new samples. Comparison with other chemometric techniques.

Pinto, Rui[1]**; Bosc,** Véronique[2]**; Barros,** António S.[3]**; Bouveresse,** Delphine. [1]; **Rutledge,** Douglas N.[1]**;**

[1]*Laboratoire de Chimie Analytique, Institut National Agronomique Paris-Grignon.*
16, rue Claude Bernard. 75005 Paris, France.

[2]*École Nationale Supérieure des Industries Agricoles et Alimentaires-UMR Scale 1211. 1, avenue des Olympiades. 91744 Massy, France.*

[3]*Departamento de Química, Universidade de Aveiro.*
Campus Universitário de Santiago. 3810-193 Aveiro, Portugal.

Corresponding author: rutledge@inapg.inra.fr

KEYWORDS
ANOVA-PCA, discrimination, Principal Component Transform (PCT), other chemometrics discrimination techniques, FTIR Spectroscopy, Chemometrics.

ABSTRACT
Analysis of variance –principal component analysis (ANOVA-PCA) [1 Harrington] has been used with proteomics for the detection of biomarkers in high dimensional data sets [1,2 Harrington]. This supervised classification method uses ANOVA to separate covariations into main effects and interaction and PCA to evaluate the significance of each of the effects against the residual error. When applied to spectroscopy, the obtained loadings can be interpreted and it is possible to classify new samples in one of the groups within a determined confidence interval.
In this work the method is applied to a set of mid-infrared spectra of carragheenan gels with 1 and 2% concentrations at 4 different temperatures to separate the samples in different groups

related with the two factors and to understand which wavelengths are related with each of the factors. Subsequently, new samples are classified to evaluate the reproducibility of the method.

It is discussed the possibility and the advantages of the use of the Principal Component Transform (PCT) [3] method to accelerate the calculations if higher dimensional data sets are analysed.

The ANOVA-PCA method is also compared with other chemometric classification methods.

REFERENCES

[1] Harrington, P.; Vieira, N.; Espinoza, J.; Nien, J.; Romero, R.; Yergey, A.; "Analysis of variance-principal component analysis: A soft tool for proteomic discovery". Analytica chimica acta, 2005, 544, p. 118-127.

[2] Harrington, P.; Vieira, N.; Chen, P.; Espinoza, J.; Nien, J.; Romero, R.; Yergey, A.; "Proteomic analysis of amniotic fluids using analysis of variance-principal component analysis and fuzzy rule-building expert systems applied to matrix-assisted laser desorption/ionization mass spectrometry". Chemometrics and Intelligent Laboratory Systems, 2006, 82, p. 283-293.

[3] Barros, A.S.; Rutledge, D.N.; "Principal Components Transform – Partial Least Squares (PCT-PLS): A novel method to accelerate Cross-Validation in PLS regression". Chemometrics and Intelligent Laboratory Systems, 2004, 73, p. 245– 255

Using ANOVA-PCA to discriminate Mid-IR spectra of carragheenan gels at different concentrations and temperatures and to class new samples

Rui Pinto[1], Véronique Bosc[2], António Barros[3], Delphine Jouan-Rimbaud Bouveresse[1], Douglas N. Rutledge[1]
[1] Laboratoire de Chimie analytique, Institut National Agronomique Paris-Grignon, France.
[2] École Nationale Supérieure des Industries Agricoles et Alimentaires-UMR Sciale 121, Massy, France.
[3] Departamento de Química, Universidade de Aveiro, Portugal

INTRODUCTION

Analysis of variance–principal component analysis (ANOVA-PCA) [1 Harrington] was introduced in proteomics to detect biomarkers in high dimensional data sets [1,2 Harrington].
This supervised classification method uses ANOVA to separate covariations into main effects and interactions and then uses PCA to evaluate the significance of each of these effects against the residual error.
When applied to spectroscopic data, the loadings obtained can be interpreted and it is possible to class new samples into one of the groups within a determined confidence interval.
In this work the method is applied to a set of mid-infrared spectra of carrageenan gels at 1% and 2% concentrations at 4 different temperatures in order to separate the samples into different groups related with the two factors and to understand which wavelengths are related to each of the factors.
Mid-IR spectroscopy using an ATR was applied to characterise the surface interaction properties of carrageenan gels with different rheologies due to concentration and temperature.

OBJECTIVES

- Apply ANOVA-PCA to spectroscopic data
- Develop a method based on ANOVA-PCA for prediction of group membership of new samples
- Test the use of the segmented-Principal Component Transform (seg-PCT) [3] method to accelerate the calculations when huge datasets are analysed
- Compare results obtained when replacing PCA by ICA and PLS (ANOVA-ICA and ANOVA-PLS)

MATERIALS AND METHODS

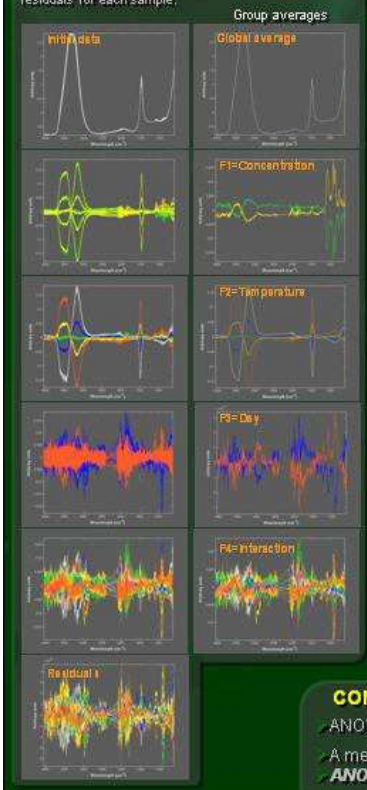
1 and 2% solutions of iota-carrageenan were prepared and analysed at 30, 40, 45, 50 and 60°C in the mid-infrared region using a Fourier Transform Infrared spectrometer (Bruker Vector 33) with a temperature-controlled "Golden Gate" Attenuated Total Reflection (ATR).
Solutions were kept at 78°C then droplets were placed on the ATR single reflection diamond crystal and left 1 minute to stabilize the temperature before acquiring the spectrum.
The procedure was repeated on two different days and in triplicate for each sample. 64 scans were collected and averaged in the region 4000-600 cm-1, at 4 cm-1 resolution, resulting in a matrix of dimensions 60x1790.
ANOVA-PCA analysis was applied to the data as described in the literature [1,2], considering the factors concentration, temperature and day of the analysis in order to study the significance of the factors when compared to the residual error.
All calculations were performed in MATLAB 7.0 (R14).

ANOVA

Group averages for each factor calculated.

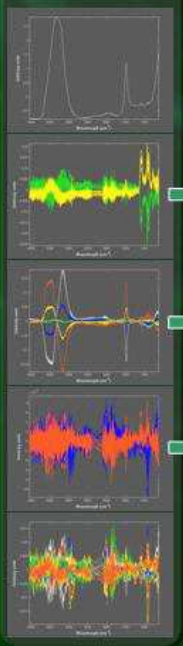
Subtract the averages to the respective group samples.

If done sequentially for all the factors, one obtains the residuals for each sample.



Factor + residuals

The residuals for each sample are added to the respective group average in each of the factors.



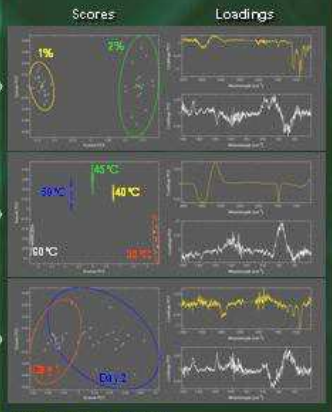
PCA

PCA is applied to each Factor + residuals matrix.

If the factor is significant in relation to the residuals (pure error), it will be described on PC1. Pure error will be described on PC2.

It is possible to draw confidence interval ellipses on the scores.

Loadings can be interpreted.



ANOVA-PCA discussion

The factors Concentration and Temperature are significant.

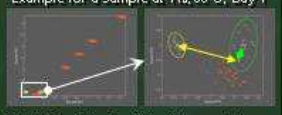
Day is not significant as its scores overlap.

The interesting discriminating wavelengths can be observed in the loadings.

PREDICTION

Prediction of the group membership for a new sample is possible by subtracting from its spectrum all possible combinations of the averages other than the one for the factor being tested and calculating its scores.

Example for a sample at 1%, 30°C, Day 1



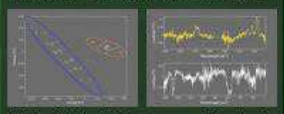
Calculation of the Euclidian distances between each of the possible predicted scores and the different group averages.

Prediction of concentration (all samples)

OTHER RESULTS

ANOVA-ICA (Independent Component Analysis) gave very similar results.

For ANOVA-PLS (Partial Least Squares) the separation of the scores for the two days in the factor "Day" became significant (below).



PCT (Principal Components Transform), which can be used as a loss-less data reduction pretreatment for huge datasets, was successfully applied.

CONCLUSIONS

- ANOVA-PCA: 2 factors – concentration and temperature - are significant at the 95% level.
- A method was developed to use ANOVA-PCA for ***prediction***, using the Loadings.
- ***ANOVA-PLS*** was tested and showed that the Factor "day" is in fact important.
- ***PCT*** and ***Seg-PCT*** can be used to perform ANOVA-PCA if the dataset is huge.
- Other multivariate methods (PLS, ICA) can be used instead of PCA, usually giving similar results.

REFERENCES:
[1] Harrington, P.; Vieira, N.; Espinoza, J.; Nien, J.; Romero, R.; Yergey, A.; *"Analysis of variance–principal component analysis: A soft tool for proteomic discovery"*, Analytica Chimica Acta, 2005, 544, 118-127.
[2] Harrington, P.; Vieira, N.; Chen, P.; Espinoza, J.; Nien, J.; Romero, R.; Yergey, A.; *"Proteomic analysis of amniotic fluids using analysis of variance-principal component analysis and fuzzy-rule-building expert systems applied to matrix-assisted laser desorption/ionization mass spectrometry"*, Chemometrics and Intelligent Laboratory Systems, 2006, 82, 283-293.
[3] Barros, A. S.; Rutledge, D.N.; *"Principal Components Transform – Partial Least Squares (PCT-PLS): A novel method to accelerate Cross-Validation in PLS regression"*, Chemometrics and Intelligent Laboratory Systems, 2004, 73, 245–255.

ACKNOWLEDGEMENTS: Rui Pinto acknowledges a PhD. grant from the "Fundação para a Ciência e Tecnologia" (FCT) - Portugal

ANNEXE XI

OSC-PCA : une modification de la méthode ANOVA-PCA

Pinto, Rui[1, 4]**; Noçairi,** Hicham[2]**; Bosc,** Véronique[3]**; Barros,** António S.[4]**; *Rutledge,** Douglas N. [1, 2]

[1]*Laboratoire de Chimie Analytique, AgroParisTech.*
16, rue Claude Bernard. 75005 Paris, France.

[2] *UMR INRA/AgroParisTech 214 "IAQA"*
16, rue Claude Bernard. 75005 Paris, France.

[3]*École Nationale Supérieure des Industries Agricoles et Alimentaires-UMR Scale 1211.* 1, avenue des Olympiades. 91744 Massy, France.

[4]*Departamento de Química, Universidade de Aveiro.*
Campus Universitário de Santiago. 3810-193 Aveiro, Portugal.

*Corresponding author: rutledge@inapg.inra.fr

MOTS CLÉS
ANOVA-PCA, OSC, Correction orthogonal de signal, Prédiction, Classification, Facteur

Description de la méthode ANOVA-PCA

La méthode ANOVA-PCA a été développée en proteomique pour la détection des bio-marqueurs dans de grandes bases de données [1,2]. Il est important de préciser que cette méthode n'est pas une procédure pour filtrer des variables. Dans cette méthode, l'Analyse de Variances (ANOVA) n'est pas utilisée, comme c'est souvent le cas [3], simplement pour détecter les variables qui ne contiennent pas d'information afin de les éliminer du tableau de données. En fait, ANOVA-PCA [4] utilise le paradigme de l'ANOVA pour décomposer la matrice de données d'origine en un ensemble de matrices de mêmes tailles contenant chacune les moyennes pour chaque niveau des effets principaux et leurs interactions, ainsi qu'une matrice de d'erreur résiduelle. Après cette première phase univariée, la matrice résiduelle est additionnée aux matrices des moyennes des effets. Ensuite, on applique l'Analyse en

Composantes Principales (PCA) [5] sur ces matrices pour évaluer si les différents effets sont significatifs par rapport à l'erreur résiduelle.

Limitations de la méthode ANOVA-PCA classique

Avec la méthode classique d'ANOVA- PCA, les facteurs quantitatifs sont traités comme s'ils étaient qualitatifs et l'information sur les quantités n'est pas utilisée. Par exemple, quand on utilise différentes procédures d'extraction, les groupes pour cet effet sont logiquement définis comme qualitatives. Par contre, quand les échantillons sont à différentes concentrations, les groupes pour l'effet concentration peuvent être définies quantitativement. Du coup, la démarche classique n'est pas bien adaptée à ce type mixte de facteurs quantitatifs/qualitatifs. D'où, l'intérêt d'introduire une autre démarche qui peut tenir compte de ces deux types de facteurs.

De plus, ANOVA-PCA ne s'applique qu'à des données acquises selon un plan d'expériences équilibrées, aussi bien en ce qui concerne les niveaux des facteurs que le nombre d'échantillons à chaque niveau.

Ayant été développée pour la détection de bio-marqueurs, ANOVA-PCA ne permet pas actuellement la prédiction pour de nouveaux échantillons.

Une dernière limitation de la méthode classique est qu'elle est basée sur une première étape de décomposition univariée suivi par une analyse multivariée. Il serait plus logique d'utiliser des méthodes multivariées dans les deux étapes.

Modifications de la démarche

Pour palier ces différentes limitations, nous proposons de remplacer l'étape ANOVA par la correction orthogonale du signal (Orthogonal Signal Correction –OSC). OSC [6] est une méthode de prétraitement qui permet d'éliminer des variations quantitatives ou qualitatives orthogonales à la variation d'intérêt, et qui pourraient diminuer la qualité du modèle de prédiction.

Nous allons utiliser cette technique pour retirer successivement des données d'origines les variations orthogonales à chaque effet principal et à leurs interactions, afin d'obtenir une matrice résiduelle. En même temps, nous générons une série de matrices qui correspondent aux données d'origines filtrées par rapport à chaque facteur. Ensuite, comme dans la procédure classique, la matrice résiduelle est ajoutée à chaque matrice filtrée.

L'application de PCA à ces différentes matrices (OSC-PCA) évite toutes ces limitations de la démarche ANOVA-PCA classique.

Applications de OSC-PCA

Dans cette présentation, nous montrerons les avantages de OSC-PCA par rapport à ANOVA-PCA à travers l'application des deux méthodes à des spectres en Moyen Infrarouge acquis sur des échantillons de gels de carraghénanes à différentes concentrations et températures.

Références

1. **Harrington, P.; Vieira, N.; Espinoza, J.; Nien, J.; Romero, R.; Yergey, A.;** *"Analysis of variance-principal component analysis: A soft tool for proteomic discovery"*. Analytica Chimica Acta, 2005, 544, 118-127.
2. **Harrington, P.; Vieira, N.; Chen, P.; Espinoza, J.; Nien, J.; Romero, R.; Yergey, A.;** *"Proteomic analysis of amniotic fluids using analysis of variance-principal component analysis and fuzzy rule-building expert systems applied to matrix-assisted laser desorption/ionization mass spectrometry"*. Chemometrics and Intelligent Laboratory Systems, 2006, 82, 283-293.
3. **Nicola, C.; Barros, A. S.; Rutledge, D. N.; Hossenloop, J.; Trystram, G.; Emonet, C.;** *"Detecting information in gas sensor responses using analysis of variance"*, Analusis, 1998, 26, 235-141.
4. **Sarembaud, J.; Pinto, R.; Rutledge, D.N.; Feinberg, M.;** *"Application of the ANOVA-PCA to stability studies of reference materials"*. *Analytica Chimica Acta*, (2007 sous presse).
5. **Massart, D. L.; Vandeginste, B.; Buydens, L.; De Jong, S.; Lewi, P.; smeyers-Verbeke, J.;** "Handbook of chemometrics and qualimetrics", vol. 20A, Elsevier, Amsterdam (1997), 519-557.
6. **Olivieri, A. C.; Goicoechea, H. C.;** "A comparison of orthogonal signal correction and net analyte preprocessing methods. Theoretical and experimental study". Chemometrics and Intelligent Laboratory Systems, 2001, 56, 2, 73-81.

OSC-PCA : une modification de la méthode ANOVA-PCA

[1]Pinto, Rui; [2]Noçairi, Hicham; [3]Barros, António S.; [1, 2] Rutledge, Douglas N.

1 Laboratoire de Chimie Analytique, AgroParisTech. 16, rue Claude Bernard. 75005 Paris, France.
2 UMR INRA/AgroParisTech 214 MIA-QA n° 16, rue Claude Bernard. 75005 Paris, France.
3 Departamento de Química, Universidade de Aveiro. Campus Universitário de Santiago. 3810-193 Aveiro, Portugal.

INTRODUCTION

ANOVA-PCA a été développée en protéomique pour la détection des bio-marqueurs dans de grandes bases de données [1,2].
Ici, l'Analyse de Variances (ANOVA) n'est pas utilisée, comme c'est souvent le cas [3], simplement pour détecter les variables qui ne contiennent pas d'information afin de les éliminer du tableau de données.
ANOVA-PCA [4] utilise le paradigme de l'ANOVA pour décomposer la matrice de données d'origine en un ensemble de matrices de mêmes tailles contenant chacune les moyennes pour chaque niveau des effets principaux et leurs interactions, ainsi qu'une matrice de d'erreur résiduelle.
Ensuite la matrice résiduelle est ajouté aux matrices des moyennes des effets.
On applique l'Analyse en Composantes Principales (PCA) [5] à ces matrices pour évaluer si les différents effets sont significatifs par rapport à l'erreur résiduelle (voir schéma).

LIMITATIONS DE LA MÉTHODE ANOVA-PCA

- Certains facteurs, comme des procédures d'extraction, sont qualitatives. D'autres, comme la concentration, sont quantitatives.
- ANOVA-PCA traite les facteurs quantitatifs comme s'ils étaient qualitatifs et l'information sur les quantités n'est pas utilisée. D'où, l'intérêt d'introduire une autre démarche qui permet tenir compte de ces deux types de facteurs.
- ANOVA-PCA n'est applique que sur des plans d'expériences équilibrés, aussi bien en ce qui concerne les niveaux des facteurs que le nombre d'échantillons à chaque niveau.
- La première étape d'ANOVA-PCA est une décomposition univariée, ce qui n'est peut-être pas optimal pour des données multivariées.

MODIFICATIONS PROPOSÉES (OSC-PCA)

- Remplacer l'étape ANOVA par la correction orthogonale du signal (Orthogonal Signal Correction – OSC). OSC [6] est une méthode de prétraitement qui permet d'éliminer des variations quantitatives ou qualitatives orthogonales à la variation d'intérêt.
- La matrice des résidus est obtenue après « filtrage » OSC de la matrice X initiale par rapport à l'ensemble des facteurs.
- Ensuite on applique une OSC pour chaque facteur à la matrice X initiale pour obtenir plusieurs matrices filtrées.
- On ajoute les résidus à chacune de ces matrices et on applique PCA.
- Les calculs sont plus lourds que pour la méthode ANOVA-PCA standard.

OBJECTIFS

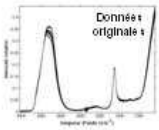
- Discussion des avantages et inconvénients de l'utilisation de la méthode ANOVA-PCA standard.

- Implementation et discussion de la modification de la méthode ANOVA-PCA.

- Étudier les facteurs quantitatifs "Temperature" et" Concentration" dans l'exemple.

- Comparer les résultats obtenus avec les deux méthodes.

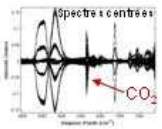
Example: données FTIR avec 4 facteurs



Solutions 1 et 2 % de carraghenane analysées par FTIR-ATR à 30, 40, 45, 50 et 60°C en deux jours différentes.

3 repetitions pour chaque échantillon.

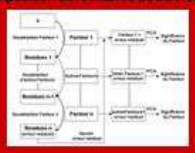
4 Facteurs: Temperature, Concentration, jour et échantillon.



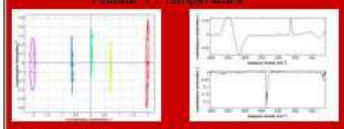
Après centrage des données, la variation aléatoire du pic de CO2 à 2400 cm-1 devient très importante.

Schéma ANOVA-PCA

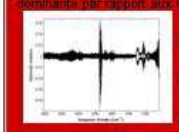
Décomposition de la matrice de données et PCA



Facteur 1 : Température

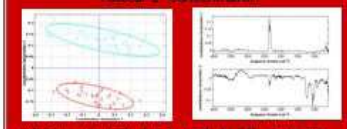


- Variance la plus importante due aux plus reliés à la température et pas au pic du CO2.
- Niveaux de température separées sur l'axe PC1.
- Le Facteur température est une source de variation dominante par rapport aux résidues.



- Facteur concentration + erreur
- Le pic du au CO2 a une influence très importante.
- PCA nécessaire à cette matrice

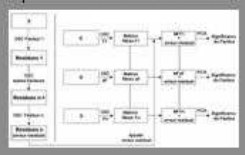
Facteur 2: Concentration



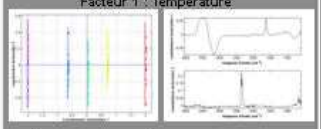
- Variance principale due au pic du CO_2 et pas aux niveaux de concentration.
- Le Facteur concentration n'est pas une source dominante de variation par rapport aux résidues.

Schéma OSC-PCA

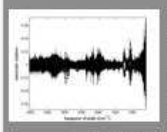
Décomposition de la matrice de données et PCA



Facteur 1 : Température

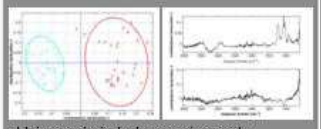


- Résultats similaires à l'ANOVA-PCA.
- Ellipses à 95% encore plus serrées que pour l'ANOVA-PCA.



- Matrice filtrée pour le facteur concentration + erreur residuel.
- Le pic du CO2 n'est plus visible.
- Sur cette matrice, on applique la PCA.

Facteur 2: Concentration



- Variance principale due aux niveaux de concentration et pas au pic du CO2
- Le facteur concentration est donc une source dominante de variation par rapport aux résidues

Discussion de l'exemple

ANOVA – PCA	OSC – PCA
QUALITATIVE/QUANTITATIVE	
• Les facteurs température et concentration sont quantitatives. • On a 5 niveaux de température donc on pourrait utiliser un modèle du type regression au lieu du modèle qualitatif standard.	• La temperature est un facteur quantitatif - utilisation des valeurs de temperature pour le filtrage OSC. • Modèle du type regression.
MODÈLES NON-ÉQUILIBRÉS	
• Facteurs bien equilibrés. • Deux concentrations et cinq températures bien distribuées, centrées à 45°C avec le même nombre de répétitions. • On peut utiliser le modèle ANOVA-PCA sans problèmes.	• Facteurs bien equilibrés. • Néanmoins, la méthode OSC-PCA n'a pas besoin des modèles équilibrés. • Les échantillons ne sont pas considerées en groupes (pas de calcul des moyennes).
UNIVARIÉE/MULTIVARIÉE	
• ANOVA-PCA ne trouve pas la concentration comme une source de variation importante par rapport à l'erreur résiduel. • Nature univariée de la première phase de la méthode – calcul des moyennes des niveaux – donne trop d'importance au pic du CO2. • Variation due à la concentration est faible, donc la contribution du CO2 est maximale la plus importante. • Changer la méthode pour résoudre le problème de l'utilisation de la phase univariée dans la méthode ANOVA-PCA.	• OSC-PCA a trouvé la température et la concentration comme sources de variations importantes par rapport à l'erreur résiduel. • Nature multivariée de la première phase de la méthode – filtrage de la variance orthogonale aux 4 facteurs – enlève presque complètement la contribution du CO_2. • Pic du CO_2 n'a pas une contribution importante pour les premières Composantes Principales.

CONCLUSIONS

- La nouvelle mèthode OSC-PCA est presentée comme une solution pour quelques cas où ANOVA-PCA ne marche pas bien.
- Un example est presenté, avec des facteurs quantitatifs.
- La concentration ne semble pas être un facteur significatif avec la méthode ANOVA-PCA mais le devient avec OSC-PCA.

REFERENCES:

[1] Harrington, P.; Vieira, N.; Espinoza, J.; Nien, J.; Romero, R.; Yergey, A.; *"Analysis of variance–principal component analysis: A soft tool for proteomic discovery"*. Analytica Chimica Acta, 2005, 544, 118-127.
[2] Harrington, P.; Vieira, N.; Chen, P.; Espinoza, J.; Nien, J.; Romero, R.; Yergey, A.; *"Proteomic analysis of amniotic fluids using analysis of variance-principal component analysis and fuzzy rule-building expert systems applied to matrix-assisted laser desorption/ionization mass spectrometry"*. Chemometrics and Intelligent Laboratory Systems, 2006, 82, 283-293.
[3] Nicola, C.; Barros, A. S.; Rutledge, D. N.; Hosseinpour, J.; Tryfram, G.; Pommet, C.; *"Detecting information in gas sensor responses using analysis of variance"*, Analisis, 1998, 26, 235-141.
[4] Saremaud, J.; Pinto, R.; Rutledge, D.N.; Feinberg, M.; *"Application of the ANOVA-PCA to stability studies of reference materials"*. Analytica Chimica Acta, 2007, 603, 147-154.
[5] Massart, D. L.; Vandeginste, B.; Buydens, L.; De Jong, S.; Lewi, P.; Smeyers-Verbeke, J.; *"Handbook of chemometrics and qualimetrics"*, Vol. 20A, Elsevier, Amsterdam (1997), 519-557.
[6] Wold, S.; Antti, H.; Lindgren, F.; Ohman, J.; *"Orthogonal signal correction of near-infrared spectra"*. Chemometrics and Intelligent Laboratory Systems, 1998, 44, 175-185.

ACKNOWLEDGEMENTS: Rui Pinto acknowledges a PhD. grant from the "Fundação para a Ciência e Tecnologia" (FCT) - Portugal

ANNEXE XII

Présentation orale

Using ANOVA-PCA for variable selection and sample classification and replacing ANOVA by OSC to improve the detection of significant factors

R.Pinto[1, 3] V. Bosc[2] A.S.Barros[3] H.Noçairi[4] D.N.Rutledge[1, 4]

[1] Laboratoire de Chimie Analytique, AgroParisTech. 16, rue Claude Bernard. 75005 Paris, France.
rpinto@agroparistech.fr

[2] École Nationale Supérieure des Industries Agricoles et Alimentaires - UMR 1211 (SCALE). 1, avenue des Olympiades. 91744 Massy, France. bosc@ensia.fr

[3] Departamento de Química, Universidade de Aveiro.Campus Universitário de Santiago. 3810-193 Aveiro, Portugal.

antonio.barros@ua.pt

[4] UMR INRA/AgroParisTech 214 "IAQA".16, rue Claude Bernard. 75005 Paris, France.

nocairi@agroparistech.fr

rutledge@agroparistech.fr

Keywords: ANOVA-PCA, OSC, OSC-PCA, Discrimination, Classification.

1. Introduction

1.1 ANOVA

Analysis of variance – principal component analysis (ANOVA-PCA) [1, 2, 5] has been used in proteomics for the detection of biomarkers in high dimensional data sets [1, 2]. This supervised method uses the ANOVA paradigm to separate covariations into main effects and interaction and PCA to evaluate the significance of each of the effects against the residual error within a determined confidence interval. As with PCA, scores and loadings are obtained, which may be used to study the existence of groups of individuals and to evaluate the importance of the initial variables in the definition of the effects and the main sources of residual variation.

Depending on the data being analyzed, problems may arise with ANOVA-PCA, because of some weaknesses of the method. In this work we will present the motivations to modify the standard ANOVA-PCA method and propose solutions that overcome some of its weaknesses and limitations.

1.2 Motivation 1 – qualitative / quantitative

In practice, for each of the factors, groups may actually be defined either in a qualitative or in a quantitative way. For example, when different extraction procedures are used, groups for that effect are defined in a qualitative way. On the other hand, when samples have different concentrations, the groups for the concentration effect are quantitatively related. With ANOVA-PCA any information about the quantities is lost so it may be interesting to modify the procedure in order to include both qualitative and quantitative factors in the same analysis.

1.3 Motivation 2 – non-balanced models

When the experimental design is not correctly balanced (numbers of samples for each level or distances between levels) an ANOVA-like procedure may not give a correct description of the data. In such cases, a regression type of method would be more applicable.

1.4 Motivation 3 – multivariate/univariate:

One fundamental difference exists between the two phases of the ANOVA-PCA method. The first part of the procedure is univariate as the group means for the different factors and the residuals are calculated for each variable one at a time. The PCA step however is multivariate. When applied to spectroscopy, in which the data are multivariate by nature, it may be preferable to use a multivariate method in place of the initial univariate step of calculating the means. In the seminal ANOVA-PCA papers by Harrington et al [1, 2], a large number of samples are used, which diminishes the effect of random variations in the first part of the method – univariate averaging. When using small numbers of samples, the negative and positive random interference signals may not equal out, giving rise to important residual variation. An alternative way should be proposed to eliminate this problem.

1.5 Addressing the motivations - OSC

To overcome the problems given above, we propose to change the first part of the method, the calculation of the level means, by replacing ANOVA by Orthogonal Signal Correction (OSC) [3]. This is a multivariate pre-treatment method that eliminates the systematic qualitative or quantitative variations orthogonal to the variation of interest, resulting in more parsimonious, and occasionally better, prediction models. Since OSC may be used with qualitative and quantitative data, it responds to motivation 1. As it has no problem with non-equilibrated experimental design data, it is also responds to motivation 2. Also, in response to motivation 3, it is a multivariate method, which is more coherent

when analyzing multivariate data. It reduces the problem arising with the use of small numbers of samples because it will consider this variation as another orthogonal factor to the variation of interest.

1.6 Advances - classification

The main use of the original ANOVA-PCA method was for the detection of biomarkers [1, 2]. It has been recently used to assess the stability of reference materials [4]. Until now, the method has not been used for classification of samples. In the present paper a procedure is proposed for the classification of new samples in respect to the levels of each significant factor. The standard and new methods will both be applied to three different datasets and the results will be discussed and compared with those obtained with other chemometric methods used for similar purposes.

2. Theory

2.1 ANOVA-PCA

The ANOVA-PCA method (figure 1, left) has no link at all with the widely used procedure for variable selection, which consists in using an analysis of variance (ANOVA) to detect variables which do not vary significantly as a function of the factors being studied, in order to eliminate them before proceeding with a multivariate analysis of the resulting reduced dataset using PCA.

In fact, ANOVA-PCA calculates successively a series of matrices corresponding to the means of the variables at each level of each factor in an experimental design, and then subtracts them from the original matrix to get the matrix of residuals . After adding the residuals to each one of the factor matrices, there is a second, multivariate phase, in which a principal component analysis (PCA) is used to assess the significance of each factor by comparing the variations due to the factor to that due to the residuals. The basic hypothesis of the ANOVA-PCA method is that if an experimental factor is a dominant source of variation compared to the residuals , then the first principal component (PC1) will mainly characterize this variation and the second principal component (PC2) will mainly reflect random variations.

For a simple two-factor model with no interaction, the decomposition of the data matrix would be given by equation (1), where each data point $\mathbf{X}_{(i,j)}$ is equal to the columns mean $\overline{\mathbf{X}}_{(i,j)}$, plus the mean of its level for factors 1 and 2 ($\overline{\mathbf{F1}}_{(i,j)}$ and $\overline{\mathbf{F2}}_{(i,j)}$) plus a "Sample" factor($\overline{\mathbf{S}}_{(i,j)}$), in which the levels are the groups of replicates and finally, the residuals $\boldsymbol{\varepsilon}_{(i,j)}$.

$$\mathbf{X}_{(i,j)} = \overline{\mathbf{X}}_{(i,j)} + \overline{\mathbf{F1}}_{(i,j)} + \overline{\mathbf{F2}}_{(i,j)} + \overline{\mathbf{S}}_{(i,j)} + \boldsymbol{\varepsilon}_{(i,j)}. \qquad (1)$$

The Hotelling T^2 distribution may be used in order to generate the 95% confidence intervals ellipse around the clusters in the principal component scores plot corresponding to the samples at each level for the factor.

2.2 Prediction with ANOVA-PCA

The method proposed here for using ANOVA-PCA to perform prediction is based on a simple concept. Consider the prediction of a new sample for a given factor. By subtracting separately from the values of the sample to predict all the possible combinations of level averages of all factors, except the ones for the factor being tested, one obtains a series of all the possible vectors of values for that sample. Only some of these vectors will be similar to the vectors describing the average of a level for the factor desired. Multiplying all those possible vectors by the PC1 and PC2 loadings of a "calibration" ANOVA-PCA for that factor, one obtains the projection of all the possibilities for the sample. Finally, the level to which the sample belongs is determined by the minimum distance of each of these projections to the center of each cluster of samples for the different levels of the factor being considered.

2.3 OSC-PCA

The OSC-PCA method (figure 1, right) is performed in a similar way to the standard ANOVA-PCA method. In a first, multivariate phase of the method, the matrix of residuals are obtained by subtracting sequentially from the data one latent variable (LV) orthogonal to each of the factors, using one of the manly available OSC algorithms [3]. Separately, each of the factor matrices is calculated by extracting simultaneously from the initial data the information orthogonal to the factor in question, using an optimized number of LVs. As in ANOVA-PCA, the residuals matrix is added to the factor matrices and in the second multivariate phase of the method, PCA is performed on each one of these new matrices to evaluate the significance of each of the factors against the residual error. The number of orthogonal LVs to use for each factor of interest must, as always with OSC, be optimized.

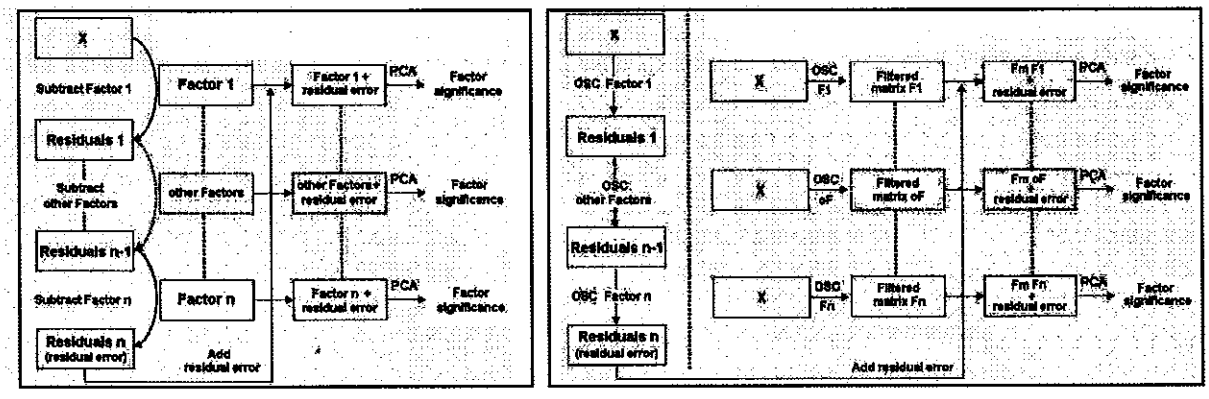


Figure 1 : ANOVA-PCA (left) and OSC-PCA (right) procedures

3. Material and methods

Three datasets were used in this work, which were named Carragheenan, Valium and Lignin datasets. ***Carragheenan data***: Mid-infrared spectra. Factors = Concentration (1 and 2 %), Temperature (30, 40, 45, 50, 60 ºC) and Day (spectra measured on 2 different days) [5]. ***Valium data***: Near-infrared spectra. Factors = Concentration (2, 5, 10 mg) and Side (3 different measurements) [6]. ***Lignin data***: Time-domain NMR signals. Factors = Moisture (2 contents), Concentration (0, 5, 10, 15, 30 %) and Shape (cane or film) [7]

4. Results and discussion

4.1 Prediction with ANOVA-PCA

The group prediction for one sample from the lignin dataset is presented as an example. Figure 2 shows all the projections calculated for the prediction vectors for that sample with respect to water content (left) and shape (right). It can be seen that the prediction is much more unequivocal for "Moisture" than for "Shape". In the former case, the possible prediction values are aligned along an axis which does not approach group 2. For the "Shape" prediction, although most predictions are nearer to group 2, the distribution is so wide that the closest prediction may be nearer the center of the wrong group.

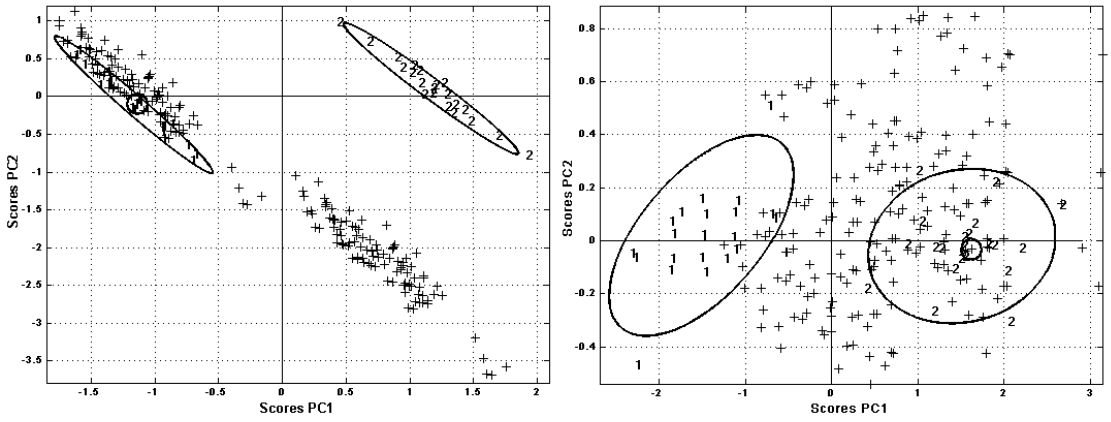


Figure 2 : All possible vectors for prediction of "Moisture" (left) and "Shape" (right) of one lignin data sample. The prediction is much better for "Moisture" than "Shape".

4.2 OSC-PCA

ANOVA-PCA may be used as a visual variable selection tool, by exploring the loadings plots. In the case of the carragheenan data (figure 3), it is obvious that the CO_2 peak at around 2400 cm^{-1} should be deleted prior to analysis as it is at the origin of PC1. Since the factor "Concentration" is not responsible for the distribution of the samples along PC1, one could erroneously conclude that this factor is not significant compared to the residual error.

The application of the OSC-PCA method to the dataset changes the situation without having to delete the CO_2 zone from the spectra, due to the elimination of interfering variations by OSC. In the case of the carragheenan dataset, the factor "Concentration" now becomes significant in comparison to the residual error. OSC-PCA aligns the samples along PC1 as a function of the factor levels (figure 4).

5. Conclusion

Examples are presented to show that ANOVA-PCA may have a role as a method for visual variable selection and for classification of new samples. OSC-PCA is shown to be an interesting method which may overcome some of the weaknesses identified when using the ANOVA-PCA method. In particular, it addresses the problems of ANOVA-PCA when considering quantitative as qualitative factors, and the use of unbalanced experimental designs.

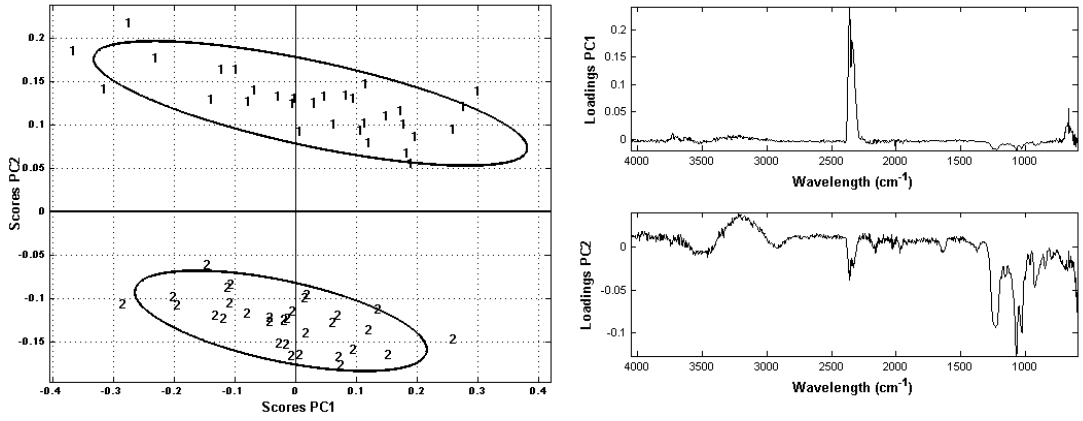


Figure 3 : Scores and loadings for the carragheenan data using ANOVA-PCA (factor "Concentration"). Although there is a separation of the two concentrations, PC1 is associated to the CO_2 present in the residuals. The Factor "Concentration" would appear not to be significant compared to the residual error.

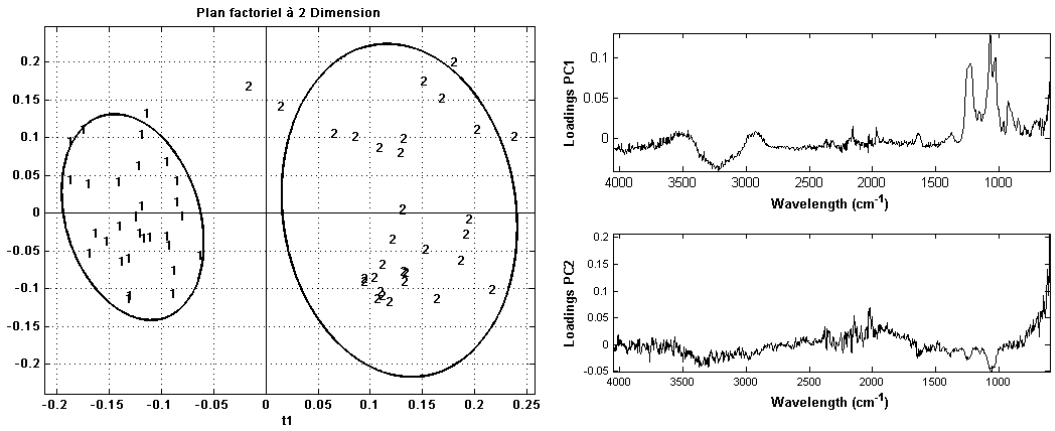


Figure 4 : Scores and loadings using OSC-PCA (factor "Concentration"). The separation of the two concentrations occurs along PC1. "Concentration" is now significant when compared to the residuals.

6. References

[1] P. Harrington, N. Vieira, J. Espinoza, J. Nien, R. Romero, A. Yergey: Analysis of variance-principal component analysis: A soft tool for proteomic discovery. *Analytica chimica acta*, 544, 118-127, 2005.

[2] P. Harrington, N. Vieira, P. Chen, J. Espinoza, J. Nien, R. Romero, A. Yergey: Proteomic analysis of amniotic fluids using analysis of variance-principal component analysis and fuzzy rule-building expert systems applied to matrix-assisted laser desorption/ionization mass spectrometry. *Chemometrics and Intelligent Lab Systems*, 82, 283-293, 2006.

[3] A.C. Olivieri, H.C. Goicoechea: A comparison of orthogonal signal correction and net analyte preprocessing methods. Theoretical and experimental study. *Chemometrics and Intelligent Laboratory Systems*, 56, 2, 73-81, 2001.

[4] J. Sarembaud, R. Pinto, D.N. Rutledge, M. Feinberg: Application of the ANOVA-PCA method to stability studies of reference materials. *Analytica chimica acta*, 603, 147-154, 2007.

[5] R. Pinto, V. Bosc, A. Barros, D. Jouan-Rimbaud Bouveresse, D. N. Rutledge, Using ANOVA-PCA to discriminate Mid-IR spectra of carragheenan gels at different concentrations and temperatures and to class new samples, *Chimiométrie 2006*, Paris.

[6] D. Jouan-Rimbaud Bouveresse, A.S. Barros, D. N. Rutledge, Generalised PLS_Cluster: an extension of PLS_Cluster for interpretable hierarchical clustering of multivariate data, *Sens. & Instrumen. Food Qual*. 1:79–90, 2007

[7] D.N. Rutledge, A.S. Barros, F. Gaudard, ANOVA and Factor Analysis applied to Time Domain NMR signals, *Magnetic Resonance in Chemistry* 35, S13, 1997.