Reduction of repeatability error for analysis of variance-Simultaneous Component Analysis (REP-ASCA): Application to NIR spectroscopy on coffee sample

Maxime Ryckewaert, Nathalie Gorretta, Fabienne Henriot, Federico Marini, Jean-Michel Roger

Reduction of repeatability error for Analysis of variance-Simultaneous Component Analysis (REP-ASCA): Application to NIR spectroscopy on coffee sample

Maxime Ryckewaert[a,b], Nathalie Gorretta[b], Fabienne Henriot[a], Federico Marini[c], Jean-Michel Roger[b]

[a] *Limagrain Europe, Chappes, France*
[b] *ITAP, Univ Montpellier, Irstea, Montpellier SupAgro, Montpellier, France*
[c] *Department of Chemistry, University of Rome "La Sapienza", P.le Aldo Moro 5, I-00185 Rome, Italy*

Abstract

A method to reduce repeatability error in multivariate data for Analysis of variance-Simultaneous Component Analysis (REP-ASCA) has been developed. This method proposes to adapt the acquisition protocol by adding a set containing repeated measures for describing repeatability error. Then, an orthogonal projection is performed in the row-space to reduce the repeatability error of the original dataset. Finally, ASCA is performed on the orthogonalized dataset. This method was evaluated on NIR spectral data of coffee beans. This study shows that the repeatability error due to physical variations between measurements can alter results of the analysis of variance. These effects are predominant in factors analysis and can be seen on spectra as constant or non-constant baselines. By reducing repeatability error

Email address: maxime.ryckewaert@limagrain.com (Maxime Ryckewaert)

with REP-ASCA, baselines are removed and factor analysis provides more information about chemical content of the factors of interest.

1. Introduction

When conducting measurements in experiments, measured values can differ from the true values of the samples. Measurement error is the difference between a measured quantity and its true value. For reliable measures, the measurement error must be relatively small compared to the variance of the measured samples [1].

The reproducibility error is the difference between two measurements of the same sample under two different conditions, for example, change of laboratory or operator. When conditions of acquisitions do not change, the reproducibility error is also called repeatability error [2]. In this paper, the repeatability error stands for differences observed between two measurements performed on the same physical sample.

In an experiment, the observed responses can contain repeatability error and this may affect the results of processing carried out on this data. This is especially true when we want to study the influence of independent variables, i.e., the so-called factors.

The use of design of experiments (DoE) allows defining the experiments to perform in an optimal way, in order to test the influence of the identified factors on the observed responses [3, 4]. The statistical analysis called the ANalysis Of Variance (ANOVA) of a dataset from an experimental design, was historically formalized in 1925 by Fischer with the first edition of [5]. The aim of this analysis is to split the total variance of the measured responses

into several sources of variances [6] However, the repeatability error may alter the results of ANOVA, even when the factors have a relevant influence on the response(s) of a studied system. Indeed, it may become more complicated to differentiate which part of the variance is explained by the factors and which part is due to the repeatability error.

In spectroscopy, measurement error may be due to both the variation of the measurement conditions (e.g. angle of view, reference, sensor temperature, etc) and the physical properties of the measured sample (e.g. particle size, roughness, physical texture, etc). In addition, a change in sample chemical composition modifies the light diffusion process. Measuring non-uniform samples adds a major difficulty compared to homogeneous powdered samples [7]. For example, variations in oil, water, fiber, protein, and mineral content of the same individual change the nature of the light-scattering medium. These chemical and physical variations can lead to unwanted variations in spectra. These variations can for example modify the spectra by adding them with baselines, which can be horizontal lines or slopes, even curves. When the changes are proportional to the intensity value, the effect is called multiplicative effect. These phenomena are well known in spectroscopy [8, 9] and are one of the sources of repeatability error.

Some sources (sensors, operator, etc.) of errors can be described as factors. If we consider these factors should have negligible effects on the total variance, we don't account for these factors in the analysis of variance and they are referred to as hidden. Some of these factors can be described as a random factor. In the case of univariate data, these random factors can be accounted for, using adapted methods such as Generalized Linear Mixed

Models (GLMMs) [10]. In the case of multivariate data, solutions lie in (1) reducing dimensionality or (2) modelling multiple responses (e.g., longitudinal analysis by using GLMMs; [11]). However, GLMMs are unsuitable in most cases, especially with spectral data [12].

For multivariate data, chemometric methods can be used to take advantage of the DoE [13]. Among those methods, the Analysis of Variance - Simultaneous Component Analysis (ASCA) [14] is the most used, but the repeatability error is not taken into account in ASCA. This method systematically averages the repetitions of measurements on an individual. Averaging observations helps to reduce the repeatability error but requires increasing the number of measurements. Climaco-Pinto et al. [15] propose a method consisting in reducing the residual variance after each step of the ANOVA-PCA. This method makes the separation of factors clearer by reducing within-variance. This method does not focus on the repeatability error.

In this study, a method named Reduction of the error of repeatability-Analysis of variance-Simultaneous Component Analysis (REP-ASCA) is proposed to reduce repeatability error by adding additionnal dataset. REP-ASCA is evaluated on NIR spectral data collected on coffee beans.

The objectives of this paper are: (1) to suggest a method (REP-ASCA) to reduce the repeatability error (2) to apply REP-ASCA to real data from NIR spectroscopy (3) to study the influence of the number of measurement in the additionnal set on the reduction of the repeatability error.

2. Theory

2.1. ASCA

Multivariate data can be represented as a matrix $\mathbf{X}$ of size (N, P) where N is the number of observations and P the number of the measured variables. The analysis of variance establishes a model in order to decompose the matrix $\mathbf{X}$ into different sources of variation. By doing so, relationships between the factors and the measured variables can be studied.

The ASCA method includes of three main steps: The first one consists of partitioning the observation matrix $\mathbf{X}$ into matrices associated with the effect of the studied factors and their interactions. For example, in a DoE involving two studied factors A and B, $\mathbf{X}$ can be written according to the ASCA approach as follows:

$$\mathbf{X} = \mu + \mathbf{X}_A + \mathbf{X}_B + \mathbf{X}_{AB} + \mathbf{E} \tag{1}$$

With μ the mean of all observations, $\mathbf{X}_A$ and $\mathbf{X}_B$ the terms associated respectively to the main effects of the factors A and B, $\mathbf{X}_{AB}$, the interaction term between A and B and $\mathbf{E}$ the residuals. Matrices $\mathbf{X}_A$, $\mathbf{X}_B$ and $\mathbf{X}_{AB}$ contain identical replicates of the average spectra collected at each level of the particular factor (A, B and AB, respectively). For example, for a given level of the factor A, all the rows of X_A corresponding to this level will contain the same spectrum.

For a complete and balanced design, factors A, B and AB are independent of one another [16]. More specifically, matrices $\mathbf{X}_A$, $\mathbf{X}_B$ and $\mathbf{X}_{AB}$ are all orthogonal to one another [17]. Thus: $var(\mathbf{X}) = var(\mathbf{X}_A) + var(\mathbf{X}_B) + var(\mathbf{X}_{AB}) + var(\mathbf{E})$. The term $var(\mathbf{E})$ contains the within-level variance and

the other terms the between-level variance for the corresponding factor. In practice, the variances of the factors are replaced by the sum of the squares of the elements of the associated matrix. For example, with $\|\mathbf{X}_A\|^2$ corresponding to the square Frobenius' norm of $\mathbf{X}_A$, the variance of $\mathbf{X}_A$ is replaced by [18, 17, 19]:

$$\text{SSQ}(\mathbf{X}_A) = \|\mathbf{X}_A\|^2 \tag{2}$$

The second step of ASCA tests if factors have a significant effect on the total variance. To do that, a permutation test is performed and a p-value is estimated by comparing variance value on the original data with the distribution of variance obtained by permutations. The permutation test can be performed for data which does not necessarily respect the conditions of normality [12]. For a factor i and its associated matrix $\mathbf{X}_i$, the p-value is calculated according to the following equation [20]:

$$\text{p-value}(\mathbf{X}_i) = \frac{\text{nbr}(\text{SSQ}(\mathbf{X}_{i,perm} \geq \text{SSQ}(X_i))}{k} \tag{3}$$

Where nbr() calculates the number of occurrences, k is the number of permutations and $\mathbf{X}_{i,perm}$ the matrix obtained after a random row permutations. The p-value is then calculated by counting the number of cases where the variance of the studied factor is lower than the variances resulting from the permutations. Indeed, by doing so, the effect of the studied factor is compared to its distribution under the null hypothesis as estimated by the permutations.

For the last step, a Simultaneous Component Analysis (SCA) [21, 22] is performed on all the terms of the matrices belonging to equation 1. Thus,

for a given factor i, the decomposition of the corresponding matrix $\mathbf{X}_i$ is written:

$$\mathbf{X}_i = \mathbf{T}_i \mathbf{P}_i^t \tag{4}$$

Where $\mathbf{T}_i$ and $\mathbf{P}_i$ correspond respectively to the scores and the loadings of the principal components. $\mathbf{T}_i$ and $\mathbf{P}_i$ are respectively matrices of size (N,L) and (P,L) with L is the number of levels of factor i minus one.

This analysis reduces the representation space. These loadings define a subspace spanned by $\mathbf{X}_i$ and highlight the spectral directions related to the factors studied. The scores are the new coordinates of the observations on these principal components.

2.2. Repeatability error

When a sample measurement is repeated several times, measurements are never identical. Observed variation between measurements is due to the repeatability error. When measurements are performed on different samples, variance differences are due to the variance between samples but also contain repeatability error. The greater the repeatability error, the greater the within-variance increases at the expense of between-variance. In fact, the variance of the factor becomes smaller relatively to the total variance. This tends to decrease the significance of the factors.

Some of the error can be due to a random phenomenon whose average is null. In ASCA, the repeatability error is reduced because repeated measures are systematically averaged over every sample. As a consequence, when the repeatability error is high, one solution is to reduce its variance by increasing

the number of observations. As a result, the number of acquisitions increases considerably. In addition, it is impossible to link the repeatability error to a specific factor in ASCA. Indeed, it is a nested factor [23]. The solution proposed in this study is to use an additional set of repeated measures to take away the repeatability error of the observations before analysis of variance. [24]

2.3. Proposed method: REP-ASCA

The method proposed in this study consists in identifying the spectral directions responsible for the repeatability error and removing them from $\mathbf{X}$. By definition, $\mathbf{X}$ can be represented in two dual spaces: the column-space of dimension R^N spanned by the observations and the row-space of dimension R^P spanned by the variables. In ASCA, the decomposition part of the variance is performed in the column-space. Equation 1 can be written as a series of orthogonal projections [17]:

$$\mathbf{X} = \mathbf{M}_1\mathbf{X} + \mathbf{M}_A\mathbf{X} + \mathbf{M}_B\mathbf{X} + \mathbf{M}_{AB}\mathbf{X} + \mathbf{E} \tag{5}$$

Where each matrix $\mathbf{M}_i$ is an orthogonal projection matrix in the column-space (R^N) defined by the following equation:

$$\mathbf{M}_i = \mathbf{D}_i(\mathbf{D}_i^t\mathbf{D}_i)^{-1}\mathbf{D}_i^t \tag{6}$$

Where $\mathbf{D}_i$ called the dummy-matrix or design matrix is a binary matrix encoding the class belonging of the factor i.

The variance of repeatability error can be included into all factors. This variance cannot be expressed into a unique factor and therefore represented

by an orthogonal projection. This error cannot be reduced in the column-space.

The spectral directions of the repeatability error can be described in the row-space through the space W spanned by the repeatability error. Let $\mathbf{X}_s$ be a matrix containing the repetitions of measurements performed on a set of samples representative of $\mathbf{X}$. Each repeated set is centered and all the centered sets are stored in a matrix $\mathbf{W}$. The matrix $\mathbf{W}$ contains only the effects related to the repetition of measurements. A Principal Component Analysis (PCA) is performed on $\mathbf{W}$ and provides loadings $\mathbf{P}$ which define the repeatability error subspace. The matrix $\mathbf{X}$ is then projected orthogonally to $\mathbf{P}$ using the following formula:

$$\mathbf{X}_{\perp} = \mathbf{X}(\mathbf{I} - \mathbf{P}\mathbf{P}^t) \tag{7}$$

ASCA is then performed on $\mathbf{X}_{\perp}$ instead of $\mathbf{X}$. This method contains two parameters to adjust: the number k of dimensions removed by the projection, i.e. the number of rows of $\mathbf{P}$ and the minimum number of spectra required in $\mathbf{W}$, i.e. the number of rows of $\mathbf{W}$.

The setting of the parameter k can be done according to a set of criteria. First, the scree plot of the PCA eigenvalues can be examined, and the value of k that corresponds to a net break is the selected. Another criterion is to study the evolution of the ratio between-variance/total-variance for the design factors and identify a break, a plateau. When the data are signals as it is the case with spectral data, studying shape of loadings (break, baseline) can provide information about the nature of repeatability errors and thus allow to adjust the k parameter accordingly. The REP-ASCA method assumes

that the repeatability error is structured, i.e. a large part of this error has a subspace of limited size, which can be removed by orthogonal projection into the row-space. When the number k is too high, the loadings become noisy.

When building $\mathbf{W}$, that is to say acquiring $\mathbf{X}_s$, it is necessary to check that enough repeated measures are available to describe the repeatability error. This can be verified by monitoring the evolution of the structure of the subspace spanned by $\mathbf{W}$. Let $\mathbf{J}_n$ be the inertia matrix of $\mathbf{W}$ when it contains n spectra. The RV coefficient [25, 26] between $\mathbf{J}_n$ and $\mathbf{J}_{n-1}$ is calculated for each addition of elements. The evolution of this coefficient as a function of n reflects the evolution of the subspace structure generated by $\mathbf{W}$. In practice, the filling of $\mathbf{W}$ is stopped when this coefficient no longer changes.

3. Materials

3.1. Experimental data

The REP-ASCA was tested on data resulting from near infrared spectroscopy of coffee beans [18]. The aim of the experiment was to study species and roasting times influences on spectra. Spectra were recorded over a spectral range of 1000 nm to 2500 nm with a resolution of 1.2 nm and transformed in pseudo-absorbance ($-\log 1/R$). Initially, ten geographical origins were compared representing a total of 800 spectra. For this study, only one origin (80 spectra) was selected.

Two species of coffee (Robusta and Arabica) were roasted during four different periods (0mn, 25mn, 50mn, 75mn) at a constant temperature of 180 °C. Seven successive spectral measurements were collected on each sam-

ple. This DoE yielded a matrix $\mathbf{X}$ of 56 spectra (2 species x 7 spectral measurement x 4 duration times). In addition, 3 repetitions of measurements have been made on each of the 8 samples leading to 24 spectra in the matrix $\mathbf{X}_s$.

3.2. REP-ASCA model

The matrix $\mathbf{X}$ was projected orthogonally to the k first loadings extracted from X_s yielding $\mathbf{X}_\perp$. The value of k was varied from 0 (classical ASCA) to 12. In accordance with ASCA modelling, $\mathbf{X}_\perp$ was decomposed into the following terms:

$$\mathbf{X}_\perp = \mu + \mathbf{X}_{species} + \mathbf{X}_{time} + \mathbf{X}_{speciesXtime} + \mathbf{E} \tag{8}$$

With: $\mathbf{X}_{species}$ and $\mathbf{X}_{time}$ factors related respectively to species and roasting duration time. $\mathbf{X}_{speciesXtime}$ is the interaction term and $\mathbf{E}$ the residuals.

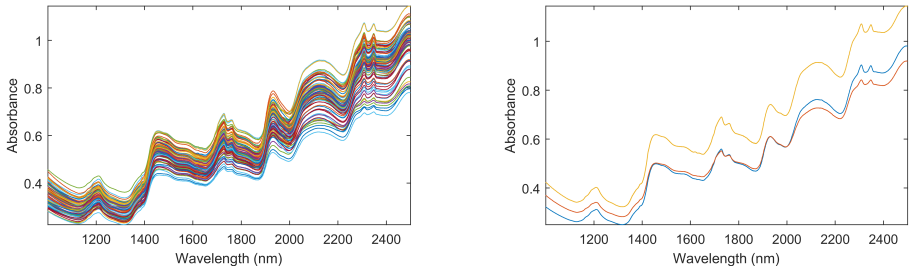
4. Results and discussion

4.1. Preliminary analysis of spectra

:

The $\mathbf{X}$ spectra are shown in 1a. We can observe some characteristic peaks of coffee spectra according to the literature [27, 28, 29]. Positions and characteristics of these peaks are presented in Table 1:

However, some differences between spectra can be observed. There is a shift in the baseline. This shift increases with the wavelength. This is frequently observed in NIR spectroscopy when the spectra are acquired in



(a) (b)

Figure 1: (a) All spectra forming by $\mathbf{X}$, (b) 3 spectra measured on one sample

Table 1: Position and peak characteristic of coffee spectra

Wavelength (nm)	Bond vibration	Assignment
1212	C−H stretching 2^{nd} overtone	Caffeine
1450	C−H stretching 1^{st} overtone	H_2O and Cellulose
1725/1765	C-H stretching 1^{st} overtone CH_2	Fatty acids
1940	O−H stretching + O−H deformation	H_2O
2080-2150	C=O (1^{st} overtone) and O–H (combination)	Protein/Fatty acids/Carbohydrates
2309/2350	C–H + C=C (combination)	Carbohydrates : Cellulose

reflectance [30, 31, 9, 32]. This phenomenon is due to differences in the light scattering.

Figure 1 b shows three spectra of $\mathbf{X}_s$ measured on the same sample and illustrating the repeatability error. These spectra present the same kind of variation observed on $\mathbf{X}$ spectra (Fig. 1a). A vertical translation as a shift in a constant baseline appears between spectra at all wavelengths. This is because each spectrum was measured on a different area of the same sample. In each area, the reflectance intensity varies according to the geometric configuration of the measured grains in relation to the sensor. These aspects

are also observed over the whole databases $\mathbf{X}$ and $\mathbf{X}_s$.

The baselines also vary in slopes (see Fig. 1b). This is due to a variation of particle size and surface roughness of the product [30]. It can therefore be assumed that the internal structure of coffee beans varies within the same sample.

A more detailed analysis of 1b shows three distinct spectral regions for baseline variations:

- Between 1000 nm and 1400 nm, a negative slope is observed, which varies for each repetition.
- Between 1400 nm and 2000 nm, baselines are increasing, but appear to have identical slopes to ones observed at other spectral ranges.
- Between 2000 and 2500, baselines are increasing, but appear to have different slopes. It can be hypothesized that all of these baseline variations are due to several scattering regimes (e.g., Mie and Rayleigh). These scattering regimes are governed by the relationship between particle size and light wavelength [7].

All these baselines deform the information contained in the spectra. These undesired effects should be removed to better describe the factors.

4.2. ASCA on $\mathbf{X}$

The Explained variance and the p-value for each factor are obtained from ASCA on $\mathbf{X}$ and summarized in Table 2. We can then observe that the largest source of variance is explained by the residuals. 82.01% of the total variance could not be related to a studied factor. 5.93% of the variance is

Table 2: Explained variances and p-values of the factors

Factor	Explained Variance (%)	p-value
$\mathbf{X}_{species}$	5.93	0.056
$\mathbf{X}_{time}$	8.61	0.128
$\mathbf{X}_{speciesXtime}$	3.45	0.548
$\mathbf{E}$	82.01	

explained by the species, 8.61% of the total variance is explained by the roasting duration time and 3.45% by the interaction term.

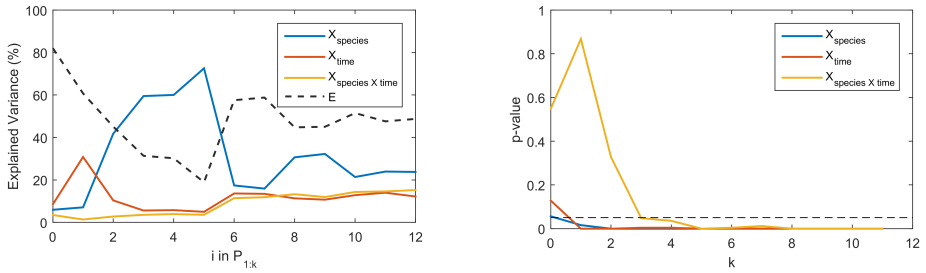
In addition, the p-values obtained after the permutation test, show that none of the factors is statistically significant at 5%. These results can be explained in part by a large part of variance not explained by the model. The factor variances are too low in comparison to the residual variance. Here, the case study shows that the repeatability error is so large that none of the identified factors are expressed. Within-variance of repeated measures are more important than variances between-level of factors. It is then very difficult in this particular case to draw conclusions with ASCA.

4.3. REP-ASCA

4.3.1. Selection of the number k of component removed

The parameter k, which corresponds to the number of components of $\mathbf{W}$ to remove, must be set. For this purpose, consequences on ASCA performed on $\mathbf{X}_{\perp}$in terms of explained variances and p-values are studied according to k.

Fig. 2a shows, that part of variance carried by $\mathbf{E}$ decreases up to the



(a) (b)

Figure 2: Evolution of: (a) explained variances by factors, (b) p-value for each factor ; according to the projection number.

5th component. Then, it subsequently increases. Inversely, the term $\mathbf{X}_{species}$ explains more and more variance and that up to the 5th component. The effects related to the repeatability error hide the expression of the term $\mathbf{X}_{species}$ to the 5th component. For the $\mathbf{X}_{time}$ term, projecting orthogonally to the 1^{st} component increases the variance part from 8.61% to 30.81%. This large increase shows that the 1^{st} component of $\mathbf{W}$ mainly masked variance due to roasting time. Nevertheless, the variance explained by this term decreases thereafter. This means that from the second component, the repeatability error is very much related to roasting. For the interaction term $\mathbf{X}_{speciesXtime}$ explained variance is very small. It increases very slowly as the repeatability error is removed.

By reducing the variance due to the repeatability error, the variances of the factors can be modified in different proportions. Information is extracted in different proportions from the different factors. The significativity of each factor can then potentially change. The evolution of p-value (Fig. 2b) shows that the repeatability error masks the effects of the factors. By removing the

information carried by the first component of $\mathbf{W}$ (i.e., projection orthogonally to the 1st component of the repeatability error), the $\mathbf{X}_{time}$ and $\mathbf{X}_{species}$ factors become both significant, in agreement with what reported in the original study [18]. The interaction term becomes significant after a projection orthogonal to the first three components.

4.3.2. Analysis of the repeatability error

Let's examine in detail the results of REP-ASCA for k=5. The sources of repeatability can be explored through the analysis of the loadings of the first 5 component (Fig 3).

Fig. 3a shows the percentages of explained variance by component. The first component explains 91.6% of the total variance. The loadings of the first component (Fig. 3b) are all positive. This indicates a systematic variation across all variables. The shape of these loadings is the same as the average spectrum of the individuals of $\mathbf{X}_s$. This reflects the vertical translation observed in (Fig. 1a).

The second component explains 6.53% (Fig. 3a) of the total variance. Surprisingly, the loadings of the second component (Fig. 3c) are similar to the loadings of the first component (Fig. 3b). However, slight differences can be observed: first, the loadings of the second component are centered on zero. This indicates an overall slope variation. Secondly, these loadings show a zero slope between 1000 nm and 1400 nm, whereas it is negative on the loadings of the first component of this same spectral range. As a result, the deflection is not homogeneous over all wavelengths.

The third component explains 1.13% (Fig. 3a) of the total variance. The loadings of the third component (Fig. 3d) show two slopes: one between 1000 nm and 1400 nm and another between 2000 nm and 2500 nm. This corresponds to different scattering regimes related to the size of the scattering media in respect to the wavelength. This scattering phenomenon is finer than what we saw before. This phenomenon results in a non-constant curving baseline effect. The fourth component explains 0.2% of the total variance (Fig. 3a). On the loadings (Fig. 3e), a first slope is visible between 1000 nm and 1300 nm and a second between 1900 nm and 2300 nm. Unlike previous components, slopes are less visible. This component is therefore less related to physical phenomena. This gives information related to the chemistry of the

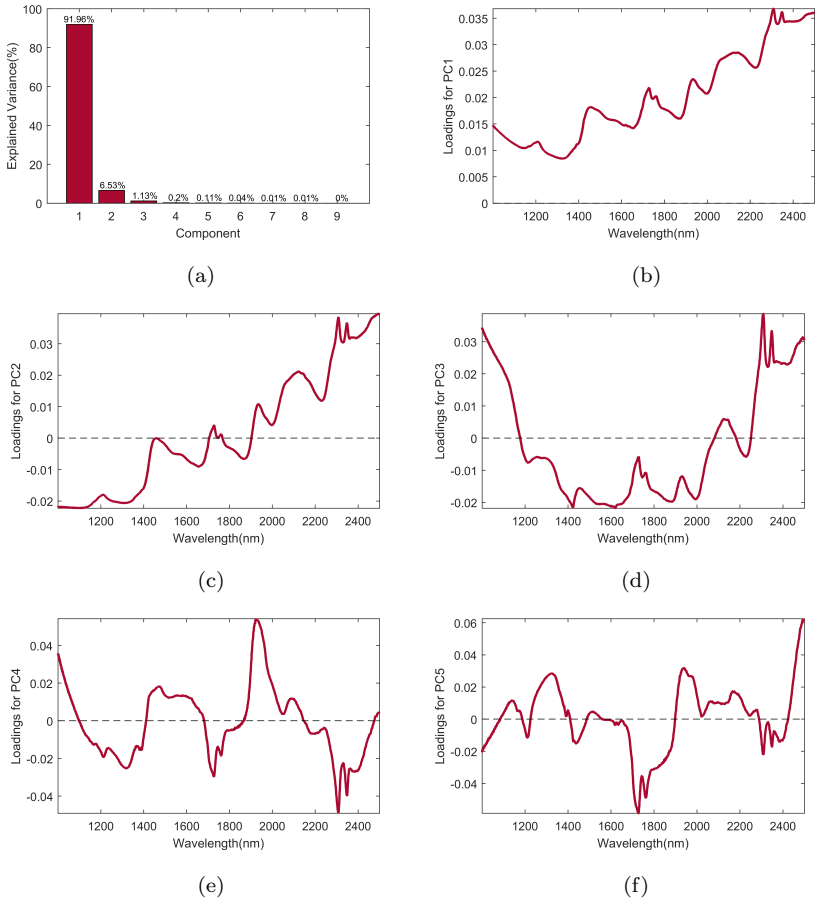


Figure 3: PCA performed on W: (a) Part of variances explained by principal components (b) Loadings on PC1, (c) Loadings on PC2 (d) Loadings on PC3, (e) Loadings on PC4, (f) Loadings on PC5.

medium. A characteristic peak at 1940 nm is related to the presence of water [33] as a result of the combination of the O-H stretch band and the O-H_2 torsion band. Two peaks are located at 1710 nm and at 2307 nm respectively linked to lipids and proteins content. The fifth component explains 0.11% of the total variance (Fig. 3a). As shown in (Fig. 3f), the loadings of this component show slopes between 1000 nm and 1300 nm and 2400 nm and 2500 nm. Peaks of 1730 nm and 1762 nm correspond to the lipids.

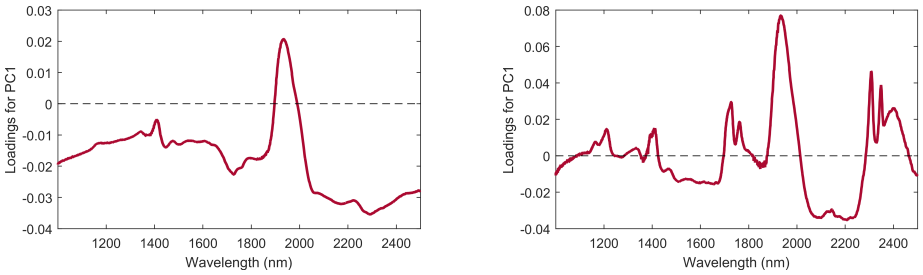
The description of the loadings of the first five components highlights that the variance of the repeated measures is related to physical phenomena. These physical phenomena induce essentially shifts in baselines on spectra. In the next components, chemical information appears. The repeatability error is therefore mainly of a physical nature, and to a lesser extent also due to a chemical difference, changing the diffusion of light in the medium.

4.4. Factor analysis after REP-ASCA (with k =5)

4.4.1. Species

Without or after orthogonal projection (Fig. 4a and 4b), the water content differentiates the two species by the peak of water located at 1930 nm. A second smaller peak located at 1398 nm corresponds to the hydrogen bond. It is also called weakly-hydrogen-bonded water [34]. Note that without orthogonal projection the loadings are all negative except the peak located at 1930 nm (Fig. 4a). After orthogonal projection according to the first 5 components of $\mathbf{W}$, the loadings (Fig. 4b) are centered on zero. The orthogonal projection therefore removed the baseline.

Some information that did not appear before is now apparent (Fig. 4b). The three peaks at 2309 nm, 2348 nm and 2400 nm show a difference in fat



(a) (b)

Figure 4: Loadings on PC1 of the term $\mathbf{X}_{species}$: (a) Without pretreatment (b) by using the projection orthogonally to the 5 first components of $\mathbf{W}$.

content [7]. The peaks at 1728 nm and 1763 nm reveal a strong presence of fatty acids [28]. It's known that Arabica beans has higher lipid content than Robusta ones [28]. The two species can be differentiated by the lipid content. A less important peak at 1212 nm also shows a difference in caffeine and carbohydrates. A negative plateau between 2070 nm and 2238 nm shows a difference in proteins and chlorogenic acids content.

4.5. Impact of the number of spectra in $\mathbf{W}$

4.5.1. Number required for $\mathbf{W}$

In this study, $\mathbf{W}$ contains 24 spectra. This number is more than enough to describe the repeatability error. The impact of the number of observations on the structure of the subspace of $\mathbf{W}$ is studied using the RV coefficient between the matrices $\mathbf{J}_n$ and $\mathbf{J}_{n-1}$ with n ranging from 4 to 24 spectra. Bootstrap is performed to check that the results are not due to the random selection of observations available.

The graph 5 shows the evolution of the mean and standard deviation of this coefficient according to n. Mean values are higher than 0.90 for all the

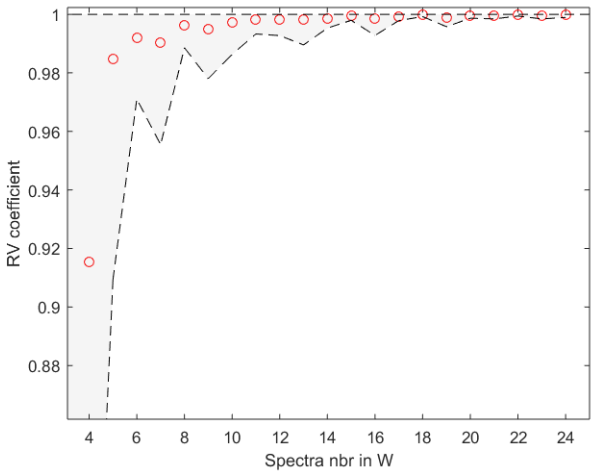


Figure 5: Evolution of the mean (red circle) and the standard deviation (grey area) of the RV coefficient between $\mathbf{J}_n$ and $\mathbf{J}_{n-1}$ according to the number n of spectra contained in $\mathbf{W}$

values of n. A plateau is visible with values close to 1 when n is greater than 10. The standard deviation displayed is about 0.18 when $n < 6$. When n increases ($n > 11$), standard deviation gradually decreases to values below 0.01. For a given n, a high value of the standard deviation means that the RV coefficient varies according to the n-th spectrum added in $\mathbf{W}$. The opposite is true, with a low standard deviation, the RV coefficient random varies slightly.

For $n < 8$, RV coefficients have low values meaning that the repeatability error is not correctly described. Additional observations must then be added. The increase of RV coefficient values with n is expected. Indeed, adding an observation has less impact in a dataset containing many observations. The plateau located from $n = 10$ shows that the structure of subspace $\mathbf{W}$ is stabilizing. From 10 observations, adding spectra in $\mathbf{W}$ does not provide

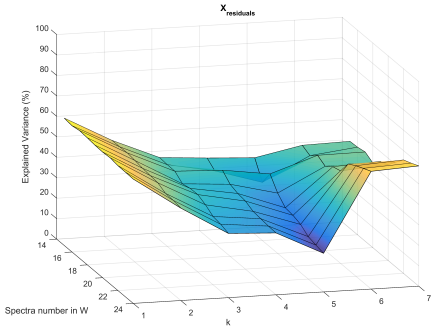


Figure 6: Explained variance of residuals according to number of projections and spectra number

more information on the repeatability error.

4.5.2. Impact on the selection of k

In this study, repeated measures are used to construct the $\mathbf{X}_s$ matrix. This matrix is used to describe the repeatability error. The impact of the number of spectra used for $\mathbf{X}_s$ matrix and the number of components used for orthogonal projection on the explained variance of residuals is illustrated in Fig. 6. The variance decreases up to the 5th component for a spectra number ranging from 15 to 24 in $\mathbf{X}_s$. The part of variance explained by the residuals is 37.31% and is 18.88% respectively for a number of 15 and 24 spectra in $\mathbf{X}_s$. The fewer spectra in $\mathbf{X}_s$, the greater residual variance. As a result, the repeatability error is then less described.

The p-values for each factor are presented in Fig. 7 according to the number of spectra used for $\mathbf{X}_s$ matrix and the number of components used for orthogonal projection. For the species factor (Fig. 7a), reducing the number of spectra from 24 spectra to 15 does not have an impact on the significance even after projection because species is the predominant factor

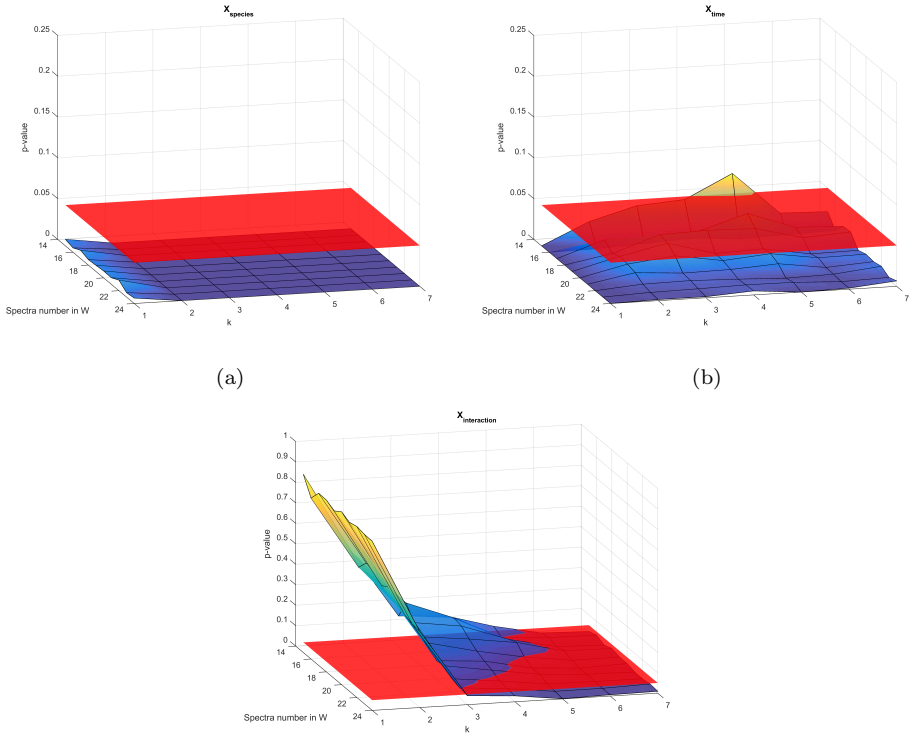


Figure 7: Impact of the spectra number in $\mathbf{X}_s$ and of the number k of components removed on p-values obtained for (a) Species, (b) Roasting time, (c) Interaction; The surface in red represents the value of 0.05 for the p-value.

in terms of portion of variance explained. For the roasting time (Fig. 7b) and the interaction terms (Fig. 7c), the number of spectra in $\mathbf{X}_s$ and the number of components used for orthogonal projection affect the results of the permutation tests. As presented in Fig. 7b, the roasting time term is not significant after a projection orthogonal to the first five components of $\mathbf{W}$ when $\mathbf{X}_s$ contains less than or equal to 15 spectra. For the interaction term (Fig. 7c), significance may change depending on the number of spectra used to build $\mathbf{X}_s$. Permutation tests may be strongly influenced when reducing the number of spectra in $\mathbf{X}_s$. Here, to study the interaction term and the roasting time term, it is necessary to have at least 20 spectra in $\mathbf{X}_s$. As a general, the number of observations must be sufficient to reduce the repeatability error. The more components to be removed, the higher the number of spectra required to be included in $\mathbf{X}_s$.

In the previous study [18], the Standard Normal Variate transform (SNV) [35] was used to reduce additive effects. This pretreatment is common in spectroscopy. It corrects the constant baselines over the whole dataset but SNV cannot correct non-constant (sloping or curving) baseline effects [36, 9].

The advantage of the proposed method is to reduce the repeatability error with more accuracy. Based on the analysis of repeated measures, the components of the repeatability error contain non-constant (sloping or curving) baselines. Indeed, they reflect diffusion and absorption phenomena of both physical and chemical nature.

5. Conclusion

In this study, a method to reduce the repeatability error (REP-ASCA) was tested on spectral data containing large repeatability error. Indeed, the interpretation of the factors through ASCA on such raw data (without correction) could mislead to wrong conclusions. The loadings of factor components contain then only the information related to the repeatability error. It was also illustrated how the significance of the factors can be influenced by the nature of the repeatability error.

The proposed method reduces these effects with the help of an additional set. The description of the repeatability error is then made through the study of repeated measurement introduced in the additional set. This description highlights the spectral regions related to the repeatability error. While repeatability error description is carried out in the column-space, reducing this error is realized in the row-space.

The choice of the number of components to be removed is done by studying the impact of projections orthogonally to these components on the various factors of the analysis of variance. Moreover, this choice can be improved by the inspection of the loadings of the main components of the repeatability error. While removing the error due to repeatability, the results obtained with REP-ASCA allow to highlight the spectral regions due to the factors of interest.

REP-ASCA was tested on a dataset containing a repeatability error mainly caused by samples with highly variable physical specificities and associated measurement errors (particle size, angle of view, etc). It would be interesting to apply REP-ASCA to another dataset containing other kinds of

repeatability errors. In addition to reducing repeatability error, REP-ASCA could remove other undesired effects (temporal effect, specific chemical content) by incorporating other components related to these effects. Futhermore, the REP-ASCA methodology could correct any type of unwanted effects, provided that the spectral subspace generated by these effects can be identified. A lot of methods can be used to carry out this identification [37].

On the practical side, adding an additional set to reduce repeatability error may considerably reduce the number of repetitions per modality in the set for the analysis of variance. Using this method, acquisition protocols should include observations to describe these errors independently of the observations for analysis of variance. This is required to make the measures for the analysis of variance more robust. However, a sufficient number of observations should be kept to study the significance of the factors.

References

[1] F. E. Grubbs, On estimating precision of measuring instruments and product variability, Journal of the American Statistical Association 43 (1948) 243–264.

[2] J. N. Miller, J. C. Miller, Statistics and Chemometrics for Analytical Chemistry, Pearson Education, 2005. Google-Books-ID: Efx77dxOC3sC.

[3] R. A. Fisher, The design of experiments, Oliver And Boyd; Edinburgh; London, 1937.

[4] G. W. Oehlert, A first course in design and analysis of experiments, W.H. Freeman, New York, 2000.

[5] R. A. Fisher, Statistical methods for research workers, Genesis Publishing Pvt Ltd, 2006.

[6] H. Scheffé, The analysis of variance, The analysis of variance, Wiley, Oxford, England, 1959.

[7] P. Williams, K. Norris, Near-infrared technology in the agricultural and food industries., American Association of Cereal Chemists, Inc., 1987.

[8] J. Engel, J. Gerretzen, E. Szymańska, J. J. Jansen, G. Downey, L. Blanchet, L. M. Buydens, Breaking with trends in pre-processing?, TrAC Trends in Analytical Chemistry 50 (2013) 96–106.

[9] Å. Rinnan, F. v. d. Berg, S. B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, TrAC Trends in Analytical Chemistry 28 (2009) 1201–1222.

[10] B. M. Bolker, M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, J.-S. S. White, Generalized linear mixed models: a practical guide for ecology and evolution, Trends in Ecology & Evolution 24 (2009) 127–135.

[11] S. L. Zeger, K.-Y. Liang, Longitudinal data analysis for discrete and continuous outcomes, Biometrics (1986) 121–130.

[12] M. Anderson, C. T. Braak, Permutation tests for multi-factorial analysis of variance, Journal of statistical computation and simulation 73 (2003) 85–113.

[13] R. G. Brereton, J. Jansen, J. Lopes, F. Marini, A. Pomerantsev, O. Rodionova, J. M. Roger, B. Walczak, R. Tauler, Chemometrics in analytical chemistry—part I: history, experimental design and data analysis tools, Analytical and Bioanalytical Chemistry 409 (2017) 5891–5899.

[14] A. K. Smilde, J. J. Jansen, H. C. J. Hoefsloot, R.-J. A. N. Lamers, J. van der Greef, M. E. Timmerman, ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data, Bioinformatics 21 (2005) 3043–3048.

[15] R. Climaco-Pinto, A. Barros, N. Locquet, L. Schmidtke, D. Rutledge, Improving the detection of significant factors using ANOVA-PCA by selective reduction of residual variability, Analytica Chimica Acta 653 (2009) 131–142.

[16] R. G. Shaw, T. Mitchell-Olds, Anova for Unbalanced Data: An Overview, Ecology 74 (1993) 1638–1645.

[17] A. K. Smilde, H. C. Hoefsloot, J. A. Westerhuis, The geometry of ASCA, Journal of Chemometrics 22 (2008) 464–471.

[18] S. De Luca, M. De Filippis, R. Bucci, A. D. Magrì, A. L. Magrì, F. Marini, Characterization of the effects of different roasting conditions on coffee samples of different geographical origins by HPLC-DAD, NIR and chemometrics, Microchemical Journal 129 (2016) 348–361.

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

[20] M. J. Anderson, Permutation tests for univariate or multivariate analysis of variance and regression, Canadian Journal of Fisheries and Aquatic Sciences 58 (2001) 626–639.

[21] H. A. Kiers, SCA: A Program for Simultaneous Components Analysis of Variables Measured in Two Or More Populations: user's Manual, iec ProGamma, 1990.

[22] K. Van Deun, A. K. Smilde, M. J. van der Werf, H. A. Kiers, I. Van Mechelen, A structured overview of simultaneous component based data integration, Bmc Bioinformatics 10 (2009) 246.

[23] F. Marini, D. de Beer, E. Joubert, B. Walczak, Analysis of variance of designed chromatographic data sets: The analysis of variance-target projection approach, Journal of Chromatography A 1405 (2015) 94–102.

[24] ISO-5725-2 1994(en), International Organization for Standardization, 1994, Accuracy (trueness and precision) of measurement methods and results— Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method, Standard, ????

[25] Y. Escoufier, Le Traitement des Variables Vectorielles, Biometrics 29 (1973) 751.

[26] H. Abdi, RV coefficient and congruence coefficient, Encyclopedia of measurement and statistics 849 (2007) 853.

[27] L. Alessandrini, S. Romani, G. Pinnavaia, M. D. Rosa, Near infrared spectroscopy: An analytical tool to predict coffee roasting degree, Analytica Chimica Acta 625 (2008) 95–102.

[28] I. Esteban-Díez, J. González-Sáiz, C. Pizarro, Prediction of sensory properties of espresso from roasted coffee samples by near-infrared spectroscopy, Analytica Chimica Acta 525 (2004) 171–182.

[29] J. Ribeiro, M. Ferreira, T. Salva, Chemometric models for the quantitative descriptive sensory analysis of Arabica coffee beverages using near infrared spectroscopy, Talanta 83 (2011) 1352–1358.

[30] K. A. Bakeev (Ed.), Process analytical technology: spectroscopic tools and implemented strategies for the chemical and pharmaceutical industries, Wiley, Chichester, West Sussex, 2nd ed edition, 2010. OCLC: ocn473478595.

[31] D. L. Massart, B. G. M. Vandeginste, L. M. C. Buydens, S. de Jong, P. J. Lewi, J. Smeyers-Verbeke, C. K. Mann, Handbook of Chemometrics and Qualimetrics: Part A, Applied Spectroscopy 52 (1998) 302A.

[32] J. Workman, A. W. Springsteen (Eds.), Applied spectroscopy: a compact reference for practitioners, Academic Press, San Diego, 1998.

[33] W. A. Luck, Structure of water and aqueous solutions, Verlag Chemie, 1974.

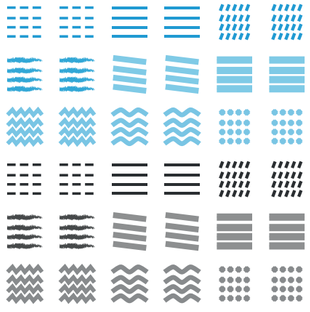
[34] A. A. Gowen, R. Tsenkova, C. Esquerre, G. Downey, C. P. O'Donnell, Use of near Infrared Hyperspectral Imaging to Identify Water Matrix Co-Ordinates in Mushrooms (*Agaricus Bisporus*) Subjected to Mechanical Vibration, Journal of Near Infrared Spectroscopy 17 (2009) 363–371.

[35] R. J. Barnes, M. S. Dhanoa, S. J. Lister, Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra, Applied Spectroscopy 43 (1989) 772–777.

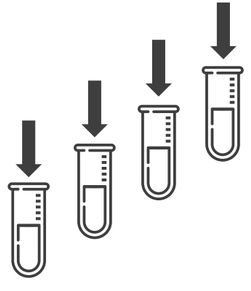
[36] R. P. Cogdill, C. R. Hurburgh, G. R. Rippke, S. J. Bajic, R. W. Jones, J. F. McClelland, T. C. Jensen, J. Liu, others, Single-kernel maize analysis by near-infrared hyperspectral imaging, Transactions of the ASAE 47 (2004) 311.

[37] J.-M. Roger, J.-C. Boulet, A review of orthogonal projections for calibration, Journal of Chemometrics (2018) e3045.

EXPERIMENTAL DESIGN



REPEATED MEASURES



$P_\perp$

ASCA

PCA

LOADINGS