Hawbeer Jamal Ahmed

PERFORMANCE CHARACTERIZATION OF MINIATURIZED NEAR-INFRARED (NIR) SPECTROMETERS FOR THE CLASSIFICATION OF SWEET AND BITTER ALMONDS

MASTER'S DEGREE THESIS

supervised by Dr Jordi Riu, Dr Ricard Boqué, and Dr Barbara Giussani

MASTER'S DEGREE IN NANOSCIENCE, MATERIALS, AND PROCESSES: CHEMICAL TECHNOLOGY AT THE FRONTIER

UNIVERSITAT ROVIRA I VIRGILI

Tarragona

2022

Performance characterization of miniaturized near-infrared (NIR) spectrometers for the classification of sweet and bitter almonds

Hawbeer Jamal Ahmed
Master Program in Nanoscience, Materials, and Processes: Chemical Technology at the Frontier
2021-2022
e-mail: hawbeerjamal.ahmed@estudiants.urv.cat

Supervisors: Dr Jordi Riu[1], Dr Ricard Boqué[1], and Dr Barbara Giussani[2]
[1]*Department of Analytical Chemistry and Organic Chemistry. Universitat Rovira i Virgili*
c/ Marcel·lí Domingo 1, 43007 Tarragona, Spain.
[2]*Dipartimento di Scienza e Alta Tecnologia, Università degli Studi dell'Insubria, Via Valleggio, 9, 22100 Como, Italy*

Abstract: Near-infrared (NIR) spectroscopy is a well-established analytical technique that has been used in many applications over the years. Due to the advancements in the semiconductor industry, NIR instruments have evolved from benchtop instruments to miniaturized portable devices. The miniaturized NIR instruments have gained more interest in recent years because of the fast and robust measurements they provide with almost no sample pretreatments. However, due to the very different configurations of these instruments, they still lack a proper optimization of the measurement conditions, which is crucial for obtaining reliable results. In this work, different sources of variability have been studied that may affect the performance of two portable low-cost NIR devices (SCiO and NeoSpectra). Measurement error covariance and correlation matrices were calculated and then visually inspected to evaluate the sources and structures of the errors associated with both devices and to find the optimal preprocessing technique that may result in the improvement of the models built with both devices. This strategy has been applied to the measurement of sweet and bitter almonds, which were measured with both devices as in-shell and shelled forms since their classification is of utmost importance for the almond industry. The results showed that bitter almonds can be classified from sweet almonds using both instruments after selecting optimal preprocessing methods obtained through inspection of covariance and correlation matrices. SCiO measurements provided better classification models with sensitivities and specificities of ones in all cases. NeoSpectra measurements showed some variations related to the analytical sessions. The chosen strategy provides new insight into the performance characterization of the fast-growing miniaturized NIR instruments.

1. Introduction

Near-infrared (NIR) spectroscopy is a type of vibrational spectroscopy that detects changes in the vibrations of molecules in response to electromagnetic radiation. Vibrational spectroscopy has some advantages over other analytical techniques because it can monitor materials with high specificity and selectivity without the need for well-trained technicians and extensive sample treatments [1,2]. Because it is robust and non-destructive, NIR spectroscopy has become a successful analysis tool in many fields, such as the pharmaceutical, agri-food, and polymer industry [3].

Usually, chemometrics tools are employed during data analysis to extract information from NIR spectra because of the complexity of the raw signal, which is of multivariate nature and therefore needs advanced techniques to correctly process the information. One of the main challenges in using NIR techniques is that a NIR spectrum consists of a number of bands emerging from overtones and combination modes that substantially overlap with each other, making it difficult to analyze the spectra. Another major problem is that NIR spectroscopy usually deals with real samples that produce low signal-to-noise (SN) ratio and baseline variations. As a result, useful information can be obtained by combining NIR spectral data with chemometrics [4,5].

Recent advancements in the semiconductor industry have led to the miniaturization of classical laboratory instruments into portable hand-held devices. One such improvement is in the field of spectroscopy. The downsizing of the spectrometers is, among others, due to incorporating novel technological solutions based on MEMS (micro-electro-mechanical systems) and MOEMS (micro-opto-electro-mechanical systems) which are techniques that can be used to fabricate electronic and optical components of devices. The fabricated miniaturized spectrometers have lower costs compared to benchtop instruments, which allows their use by a broader range of consumers. These spectrometers follow the objectives of green analytical chemistry, providing rapid, non-destructive, and on-field analysis with almost no sample pretreatments and minimal use of reagents. These on-field analyses with miniaturized spectrometers are of utmost importance in the industry because using conventional benchtop instruments requires transferring the sample to a lab, and the sample may undergo alteration. The real-time monitoring in the production facilities ensures quality and safety matters and allows rapid intervention when a problem is detected. Moreover, it allows fast quality

control checks by regulatory bodies in the markets and factories [6–8].

The aforementioned reduction in the dimension and cost of the spectrometers may result in lower performances compared with bulky instruments. The benchtop NIR instruments are mature devices that have been well studied and characterized in the past 20 years. They have uniform performances and rely on the same types of instrumentation such as the light source. The spectral wavelengths of the NIR region range from 800 to 2500 nm (12,500 to 4000 cm^{-1}), and benchtop devices usually cover the entire region. On the other hand, most portable NIR spectrometers cover a portion of the NIR region and have lower spectral resolutions. They rely on different and new solutions due to the engineering difficulties of miniaturization aspects. Such diverse technologies cause non-uniform performance and require more device-specific optimizations[3,9,10].

There has been an increasing effort to characterize these fast-growing miniaturized NIR devices in recent years. However, most of the works found in the literature are carried out to assess the classification or identification abilities of these devices applied to different fields and their performance comparison with benchtop devices[11–17]. Most of the instruments are designed to be used with solid samples with a few companies that are developing devices and accessories for other types of samples. Since the field of miniaturized NIR spectrometers is very new and evolving very fast, optimal analytical procedures are not developed yet for most of the instruments and applications. The results obtained from these studies are not directly reproducible nor transferable to other measurements because a complete characterization of these newly developed spectrometers still needs to be explored, although there are some attempts to transfer the calibration data from benchtop instruments to miniaturized devices[18,19]. Most of the studies so far do not investigate the sources of variability in miniaturized devices that affect their performance, and without understanding the underlying error structures it is more difficult to develop optimal strategies for new measurements and applications. Thus, these miniaturized spectrometers require a thorough systematic evaluation of the underlying sources of variability to get the best performance characterization that can produce reliable models.

This master's thesis aims to optimize the performance and characterize the errors associated with two low-cost miniaturized NIR instruments: SCiO (Consumer Physics, Herzliya, Israel) and NeoSpectra Micro Development Kit (Si-Ware Systems, Cairo, Egypt). Each of the two NIR spectrometers covers a different region of the NIR spectra so that the data are complementary and do not overlap in any part of the spectra. The instrumentations of both devices are quite different, allowing for a better characterization of their intrinsic characteristics. The study identifies underlying error types and structures through the analysis of multivariate measurement errors, which is a well-established statistical way to characterize error structures by building error covariance and correlation matrices from the obtained spectra. The effectiveness of multivariate measurement error and its incorporation during data analysis for other analytical instruments has been studied [20–23] but its application on miniaturized NIR spectrometers is very limited and missing in the literature. Recently *Gorla et al.* [24] tried to reveal the error sources in one miniaturized instrument and used the information to determine different properties of forage samples. They compared the results with a benchtop instrument and concluded that by evaluating error structures, optimal analytical strategies can be developed. In this current work, the best preprocessing is identified through the inspection of error covariance and correlation graphs. Finally, to test the effectiveness of the proposed method, the performances of both instruments are evaluated to classify batches of bitter and sweet almonds by using the information from multivariate measurement errors in the construction of the classification models. One of the most important aspects of the almond industry is the discrimination of bitter almonds from sweet ones since it affects their commercialization and usage in a variety of foods. Apart from the unpleasant taste, bitter almonds have serious health risks because they contain toxic compounds such as amygdalin that cause poisoning and accidental death. Since miniaturized NIR spectrometers have already proved their applicability for this kind of almond classification[25,26], different batches of bitter and sweet almonds were used for this study using the information from multivariate measurement errors.

2. Materials and Methods

2.1. Instrumentation

All measurements were performed using two NIR spectrometers; SCiO (Consumer Physics, Herzliya, Israel) and NeoSpectra Micro Development Kit (Si-Ware Systems, Cairo, Egypt). The working principle and the instrumental solutions are totally different in both devices. SCiO is a pocket-size NIR spectrometer that has a weight of 35 g and its dimensions are 67.7 × 40.2 × 18.8 mm. It acquires spectra in the 740 - 1070 nm wavelength range with interpolated spectra of 1 nm spacing; although it is not declared by the manufacturer, there are some claims that the actual resolution is considerably lower[27,28]. It can perform measurements both in contact and distance mode and is usually used for solid samples. The distance between the sample and the device should be less than 10 mm. SCiO should be connected to a smartphone via Bluetooth, and the spectra are recorded using 'The Lab' app available for Android and iOS systems. The app does not allow setting any measurement parameters, and the scan time is less than 5 seconds. The spectra can then be downloaded from the private area of a cloud web address (thelab.consumerphysics.com). The device should be calibrated each time it is turned on using a calibration standard on the back cover of the device. The device can be used in two operation modes; connected to a power supply or running on a battery. It is equipped with a Li-ion battery and can be recharged via a Universal Serial Bus (USB) cable. Regarding technological solutions, SCiO has a light source based on light-emitting diodes (LEDs), making the

device more cost-effective and decreasing power consumption. SCiO contains a silicon detector based on complementary metal-oxide-semiconductor (CMOS) in the form of a 4 × 3 photodiode array with optical filters over the individual pixels. No initial warm-up is required due to the high thermal stability of the light source.

The NeoSpectra device has a weight of 17 g and dimensions of 32 × 32 × 22 mm. The wavelength range is from 1350 to 2558 nm with a spectral resolution of 16 nm. Only contact measurements can be performed. The device must be connected to a computer via a USB cable, and the spectra are collected by a software (SpectroMOST Micro). It allows configuring some parameters such as the scan time, run mode (single or continuous), display mode (reflectance or absorbance), and data interpolation in each spectrum collected. A reflection standard such as Spectralon (99% reflectance) is required when the software is started or when the scan time is changed. The instrumental design of NeoSpectra consists of a light source made of three halogen tungsten lamps that require an initial warm-up before performing the measurements to stabilize the light intensity. The wavelength selector is a Michelson interferometer made by the MEMS technique and has a single InGaAs photodetector[15,16].

2.2. Samples and materials

The samples included different varieties of almonds from the Ribera d'Ebre region of Tarragona, Spain. The almonds were used to test the effect of the experimental and instrumental errors of the devices and their classification ability to distinguish bitter and sweet almonds. All almonds were from the same harvesting season with two different times of collection. The first group was collected in February 2022 and the second group was collected in June 2022. The different collection times aimed to observe the classification ability over time for the same types of almonds. Both in-shell and shelled forms of the almonds were measured and analyzed. A total of 247 in-shell almonds were analyzed of which 130 were sweet almonds, and 117 were bitter almonds. The sweet almonds were from two different varieties; 81 were from *Llargueta* and 49 were from *Comuna* almonds which are almonds mainly used in the industry. The bitter almonds belonged to different non-identified varieties of the region. The shells were then cracked manually to measure the almond kernels. Only whole kernels were used for the measurements, thus broken or empty almonds were dismissed, making a total of 216 shelled samples of which 120 were sweet and 96 were bitter almonds. Two different types of replicates were used in the measurements: *replacement replicate* which is changing the position of the sample each time when it is measured to account for variations related to the sample position, and *instrumental replicate* which is taking the spectra without moving the sample to account for variations related to the instrument[21]. Three replacement replicates and five instrumental replicates were measured per each sample making a total of 15 replicates to account for the heterogeneity of the samples and instrumental variations. The mean spectra of all replicates for each almond were calculated and used during the data analysis.

2.3. Statistical Data Analysis

MATLAB R2021b (Mathworks Inc., Natick, MA, USA) and PLS Toolbox version 9.0 for MATLAB (Eigenvector Inc, Manson, WA, USA) installed on a PC with Windows operating system were used for data analysis. Data were organized with in-house routines building an **X** matrix containing the samples in the rows and the wavelengths (nm) in the columns.

Error Covariance Matrices (ECMs) were calculated using in-house routines with MATLAB while Principal Component Analysis (PCA) and Partial Least Squares Discriminant Analysis (PLS-DA) were performed with the PLS Toolbox. Different spectral pre-processing methods were tested: multiplicative scatter correction (MSC), standard normal variate (SNV), detrend, and first and second Savitzky–Golay derivatives with a different number of smoothing points (from 7 to 15 points). After spectral pre-processing, data were finally mean-centered in all calculations.

To validate the multivariate models, both cross-validation (i.e. using the same samples to build and to validate the model) and external validation (i.e. using a set of samples to build the model and a different set of samples to validate the model) were used. For cross-validation, the Venetian blinds method (with 10 data splits) was used. For the external validation models, about 2/3 of the data were used in the calibration sets, and about 1/3 of the data were used in the test sets. The samples were randomly split for both calibration and test sets by taking relative samples from each variety to make a representative model of all types. The separation of sets was done after acquiring the spectra of all almonds and before starting the calculations. The best models were chosen based on fewer latent variables (LVs), higher sensitivity (samples belonging to a class correctly assigned to that class) and higher specificity (samples not belonging to the class correctly not assigned to that class), and lower root mean square error of calibration (RMSEC) cross-validation (RMSECV) and prediction (RMSEP).

2.3.1. Error Covariance Matrix (ECM)

ECMs are a common way to characterize multivariate measurement errors by describing the relationships between measurement errors across the channels/wavelengths. It is a symmetric matrix in which the diagonal elements contain the variance in the measurement error at each channel and the off-diagonal elements contain the covariance of the errors between pairs of channels. ECMs are typically represented graphically and the conclusions are taken from these graphs). The ECM is a useful tool to find the structure associated with the measured errors (e.g. proportional errors, constant errors ...) so that the choice of the optimal data preprocessing may be easily derived. When there is not enough prior knowledge about the error types and structures in a measurement (as is the case in

this work with miniaturized instruments), then the ECMs are calculated from the experimental estimation method, which is based on the analysis of the replicates. To calculate the covariance matrix using this method, the (approximate) real value of a sample is estimated from the mean of the spectral replicates of that sample ($\bar{\mathbf{X}}$). Then, a residual matrix is calculated by subtracting the estimation of the real value ($\bar{\mathbf{X}}$) from each replicate spectrum (**X**). Finally, the error covariance matrix (**Σ**) is calculated as the covariance between the residuals as shown in Eq (1):

$$\mathbf{\Sigma_i} = \frac{1}{(n-1)} \sum_{k=1}^{r} (\mathbf{X}_k - \bar{\mathbf{X}})^T (\mathbf{X}_k - \bar{\mathbf{X}}) \qquad \textbf{(1)}$$

where n is the number of replicates of the i^{th} sample, $\mathbf{X}_k$ is the measured spectrum of the k^{th} replicate, $\bar{\mathbf{X}}$ is the mean spectrum of n replicates, $\mathbf{\Sigma}_i$ is the covariance matrix of the i^{th} sample.

The covariance matrix depends on the magnitudes of the variances, and this makes their visual interpretation difficult when there are few channels with high variances (e.g. certain wavelengths with much lower precision than the rest of the wavelengths). These few high variances can hide the variations among other channels. For this reason, error correlation matrices were also calculated in this work by scaling the variances and thus removing the effects of magnitudes of variations. The error correlation matrix, with all the values scaled between +1 and -1, contains the correlation coefficients of the covariance matrix and is calculated as shown in Eq (2).

$$\mathbf{\Sigma_{corr}} = \mathbf{\Sigma_{cov}} ./ \sqrt{diag(\mathbf{\Sigma_{cov}}) diag(\mathbf{\Sigma_{cov}})^{\mathrm{T}}} \qquad \textbf{(2)}$$

Unless the number of replications for a sample is very high, the error covariance matrix has a high degree of uncertainty. For this reason, it is important to have a sufficient number of replicates or otherwise to pool the error covariance over different subsets of samples by taking the mean of all covariance matrices (**Σ**pooled). The pooling solution is generally preferred with NIR spectra because the measurement data do not change very much for the same types of samples[23].

2.3.2. Principal Component Analysis (PCA)

PCA is the most widely used statistical exploratory method for multivariate data. PCA reduces the dimensions of original variables in large data sets into smaller sets of variables known as principal components (PCs) by keeping the most important variations of the data set. The key information retained by the PCs is uncorrelated with each other, that is PCs are orthogonal among them. The highest important information is preserved in the first PC and progressively downwards. Eq (3) shows the general formula of PCA where an **X** matrix is linearly decomposed into a score matrix, **T** (explaining the similarities/differences between samples), and a loading matrix, **P** (explaining the weight of each variable retained in the PC). The residuals matrix **E** (contains part of the data that is not explained by the selected PCs)[29].

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \qquad \textbf{(3)}$$

2.3.3. Partial Least Squares Discriminant Analysis (PLS-DA)

PLS-DA is a supervised method used to classify samples based on specific properties assigned as different classes. In PLS-DA, a PLS regression model is calculated that links the independent variables (**X** matrix of NIR spectra in our case) to a vector **y** containing the assigned classes as integer numbers. For example, 1 was assigned to indicate sweet almonds, and 0 was assigned for bitter almonds. An unknown sample is classified using the PLS model's projected value. This number, which is a real number rather than an integer, ought to ideally be near to the values used to define the class (here either 0 or 1). A cut-off value between 0 and 1 is established by the model so that an unknown sample is assigned to class 1 if the prediction is larger than the cut-off value, or assigned to class 0 if it is lower than the cut-off value. For the construction of the models, the right number of latent variables (LVs) must be chosen to prevent underfitting or overfitting the models. The LVs (equivalents to PCs in PCA) are linear combinations of the initially selected variables that maximize the discrimination among the groups[30].

3. Results and Discussion

3.1. Optimization of the Instrumental Setup

In the measurements performed with SCiO, the device was fully charged and background acquisitions were acquired before starting each measurement session. The spectra were obtained by directly pointing the SCiO device onto the almonds at a fixed distance of around 0.5 cm. For the NeoSpectra measurements, background acquisitions were performed at the beginning of the measurements and every hour thereafter because NeoSpectra needs frequent background resets due to the heating of optical components[24]. An initial warm-up of 20 minutes was performed before each measurement session and the scanning time was set to 2 seconds in all measurements with NeoSpectra. The almonds were put directly on the NeoSpectra window. All measurements with both devices were performed at room temperature under ambient light as shown in Fig. 1. The spectra were acquired in reflectance mode with both devices. The average raw spectra for both in-shell and shelled almonds with SCiO and NeoSpectra are shown in Fig. 2a and Fig. 2b respectively.

a) b)

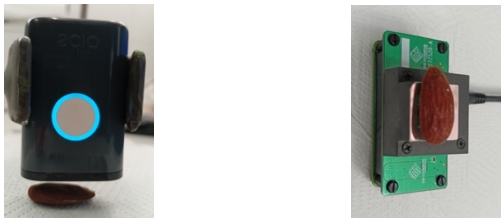
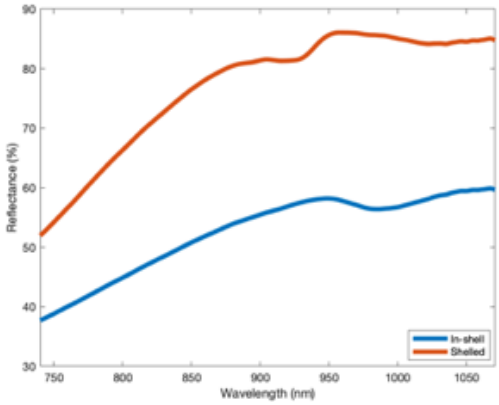


Fig. 1. Almond measurements performed with a) SCiO, b) NeoSpectra.

a)



b)

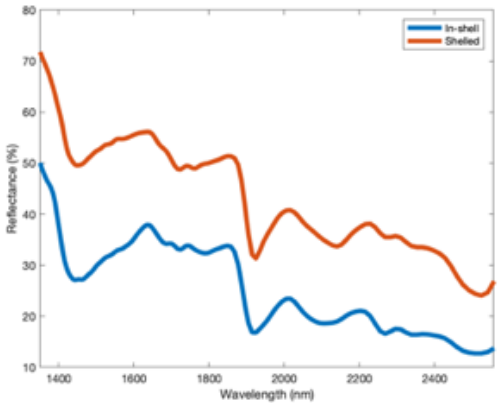


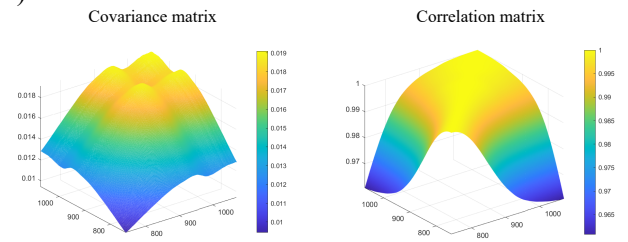
Fig. 2. Average raw spectra of in-shell and shelled almonds measured with a) SCiO, b) NeoSpectra.

3.2. Multivariate Statistical Analysis

3.2.1. Error Covariance Matrix (ECM)

The error covariance and correlation graphs were visually inspected to understand the error types and structures in both devices. Fig. 3 shows ECMs of SCiO for both in-shell and shelled almonds. Both forms of almonds had very similar error structures. The errors were heteroscedastic which is a type of error that has a non-uniform variance across the channels on the diagonal of the covariance matrix. This can be seen from the figure in which diagonal elements have different values. The errors were also highly correlated which means that there is a relationship among the errors for different variables since most of the correlation matrix was close to 1, which exhibits a high correlation among the errors. A constant offset noise was observed which is a type of correlated noise that shifts the entire signal by a constant value. This can be seen from the fact that covariance values were non-zero across all channels. This offset noise might be due to the repositioning of the almonds between replacement replicate scans. Additionally, a multiplicative noise can also be observed from the fact that the error covariance is proportional to the spectral signal (e.g., the peaks between 900-1000 nm) by comparing covariance matrix from Fig. 3 with the spectral signal from Fig. 2. This multiplicative noise could be related to variations in the light source. Constant offset and multiplicative noise are typical characteristics of NIR spectra[23].

a)



b)

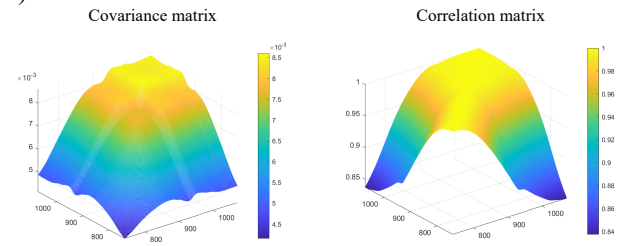


Fig. 3. Measurement error covariance and correlation matrices calculated from spectra acquired with SCiO. Case of a) in-shell almonds, b) shelled almonds

Regarding the measurements with NeoSpectra (Fig. 4), the error structures were different for in-shell and shelled almonds as can be seen from the covariance matrix which was not the case in SCiO measurements. In the case of in-shell almonds, errors were homoscedastic and correlated in the spectral region of 2000-2500nm. The shelled almonds showed more heteroscedastic errors across all regions. The noise proportional to the spectral signal was also present. Errors were higher in magnitude than in SCiO, which was already predicted from the noisier spectra given by NeoSpectra, especially in the range of 1400-1800 nm (the raw individual reflectance spectra of shelled almonds for SCiO and NeoSpectra are given in Supplementary Information Fig. S1). The higher errors may be related to the heating up of the optical components of the NeoSpectra and to some inherent design and manufacturing of the instrument. As can be seen from the ECM graphs of NeoSpectra, the error magnitudes are higher in shelled almonds compared with in-shell almonds. This higher error in the case of shelled almonds may affect their classification as will be discussed in the following sections. It was wondered if the higher errors were related to the size of almond kernels since they were smaller than the acquisition window of NeoSpectra. It was

assumed that the surrounding light may interfere with the measurements and result in higher errors. To evaluate this assumption, a different measurement was performed in which 9 almonds were measured under ambient light and then measured again in a dark environment by putting the NeoSpectra device in a box preventing the surrounding light from reaching the device. No obvious differences were observed in the ECM graphs (Supplementary Information Fig. S2) when the measurements were performed in ambient light or in the box, suggesting that the interferences from surrounding light have no effect on the measurements even though the size of shelled almonds were smaller than the acquisition window. The results showed that the errors were related to some other reasons and not to the surrounding interference.

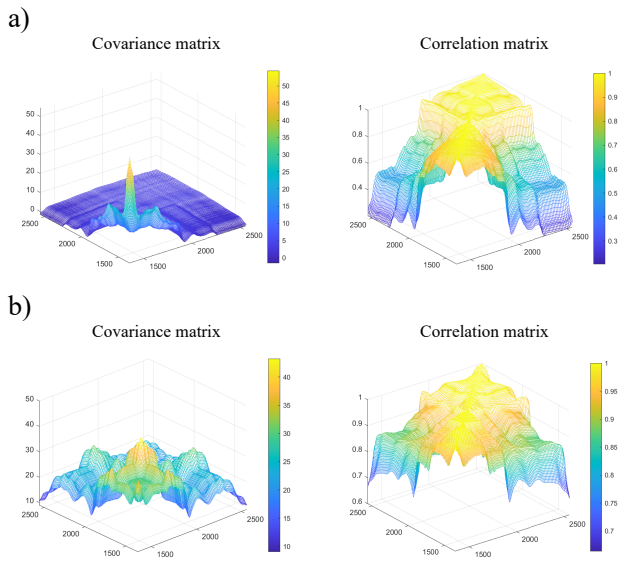


Fig. 4. Measurement error covariance and correlation matrices calculated from spectra acquired with NeoSpectra. Calculations of a) in-shell almonds, b) shelled almonds.

In order to evaluate the effect of instrumental and replacement replicates on the measurement errors, ECMs were calculated for a subset of almonds which contained 9 samples. The covariance and correlation matrices were calculated based on three replacement replicates and then based on five instrumental replicates separately. The results of this experiment for SCiO and NeoSpectra are given in Supplementary Information Fig. S3 and Fig. S4 respectively. As was expected and also concluded from previous works[24], the instrumental variations had a smaller effect on the measurement errors than the sample variations due to the heterogeneity of the sample. The comparison can be made through the magnitude of the covariance errors which is lower for instrumental replicates compared with replacement replicates. It is worth mentioning that in the measurements with NeoSpectra, the differences between replacement replicates and instrumental replicates in covariance matrices were smaller than in SCiO, suggesting higher variations in the NeoSpectra instrument. It may be required when designing an experiment with NeoSpectra to include a proper number of instrumental replicates to account for these variations.

Different preprocessing methods were applied while evaluating the plots to see their effect on the noise structures. One of the limitations of performing PCA for exploratory analysis is that it assumes all the noises present within a signal to be independent and identically distributed (known as *iid)*[23]. In this study, the aim of evaluating the preprocessing methods through ECM graphs was to find a method that best represents the *iid* assumption of PCA. It was observed that preprocessing methods that resulted in smoother surfaces and less correlated (more randomly distributed) errors were the optimal methods to apply for the classification of almonds. Table 1 shows different preprocessing methods applied for SCiO in shelled form. As it can be seen from the table, a combination of 2nd-degree derivative combined with multiplicative scatter correction (MSC) was the best selection because these graphs have fewer errors and have more random distribution compared to the raw data or other preprocessing. With NeoSpectra for shelled almonds, the optimal preprocessing was found to be smoothing combined with 2nd-degree derivative. The smoothing was usually required to correct the noisier signal acquired by NeoSpectra. The results for NeoSpectra are given in Supplementary Information Table S1. The different preprocessing was applied only to shelled almonds because the in-shell almonds were classified easily without using preprocessing as will be discussed later in the classification section. It is important to note the scale of the error covariance matrices while evaluating the graphs since it depends on the values of the errors as the smaller ones usually mean better preprocessing methods. After calculating the measurement errors through ECMs, and selecting the optimal preprocessing technique, questions arose on the validity of such a method to types of samples other than almonds. To test the effectiveness of this approach using ECMs to choose the best preprocessing method regardless of the sample type, several different solid samples were measured with both instruments. The optimal preprocessing selection was also performed through the ECM calculations. The selected preprocessing resulted in the best separation among the different classes of solid samples, which can be observed on the PCA scores plot. These results confirmed the validity of this approach. The details of the samples and measurements of the later experiment are found in the Supplementary Information from Fig. S5 to Fig. S8. From these findings, an interesting conclusion was obtained, which is a robust way that we found for choosing the best preprocessing method to be included in the prediction models of the measurements.

3.2.2. Principal Component Analysis (PCA)

PCA analysis was used to confirm the error types and structures identified through ECMs in the analysis of almonds. To evaluate the contribution from uncorrelated errors, a comparison between the cumulative percent variance of the error covariance matrix and the cumulative percent variance of the residual matrix from Eq (1) can be

used. If the differences are high, it shows a significant contribution from independent noise and smaller differences indicate high correlation noise[23]. The results (Table 2 and Table 3) showed that in both devices the contribution from independent errors was very small relative to the correlated errors except with NeoSpectra in-shell measurements which show a higher contribution from independent noises.

3.2.3. *Classification of Almonds*

By incorporating the information obtained from ECM analysis, classification models were built to distinguish sweet and bitter almonds. The PLS-DA models were built for each measurement session separately and the performance was evaluated. Then a global model was built by putting one measurement session in the calibration set and another measurement session in the prediction set to evaluate the performance of each device based on the analytical session. In all models, the in-shell almonds were better classified than the shelled almonds. It is not clear that the better results are related to any property of the shells. To determine if the radiation passes through the shell and reaches the almond kernels, a different measurement was performed. A subset of 21 bitter and sweet almonds was measured with the shell, then the almonds were carefully cracked and the kernels were taken out. The measurements were repeated only on the empty shells resembling the almonds. It was observed that the ECM and the models were very similar for both measurements, suggesting that the radiation does not pass through the shell and that the classification models for in-shell almonds were based only on the profile of the shell and not the kernels. The ECM graphs and the PCA scores plots are given in Supplementary Information from Fig. S9 to Fig. S12. Further investigation into the profile of the shells is required to determine the reasons that affect their classification.

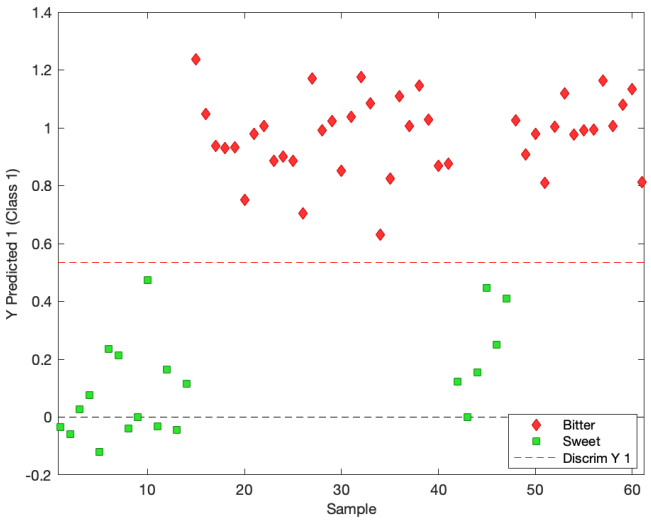


Fig. 5. PLS-DA discrimination plot for the prediction of bitter and sweet almonds measured with SCiO and based on one measurement session.

Regarding the models built with separate measurement sessions, in SCiO the in-shell almonds required no preprocessing because the models performed very well with 3 LVs and sensitivities and specificities of one, while in the shelled almonds it required the optimal preprocessing determined from ECMs which was 2nd-degree derivative combined with MSC (Fig. 5). With NeoSpectra, the in-shell almond did not require preprocessing to be applied since with 3 LVs a good classification was obtained. The shelled almonds required using an optimal preprocessing, which was smoothing combined with 2nd-degree derivative and the LVs increased to 5 to separate both types of almonds (Fig. 6). SCiO performed better with fewer LVs and lower RMSECV and RMSEP values for shelled almonds. The results were similar for both measurement sessions. The details of models based on separate measurement sessions are given in Table 4.

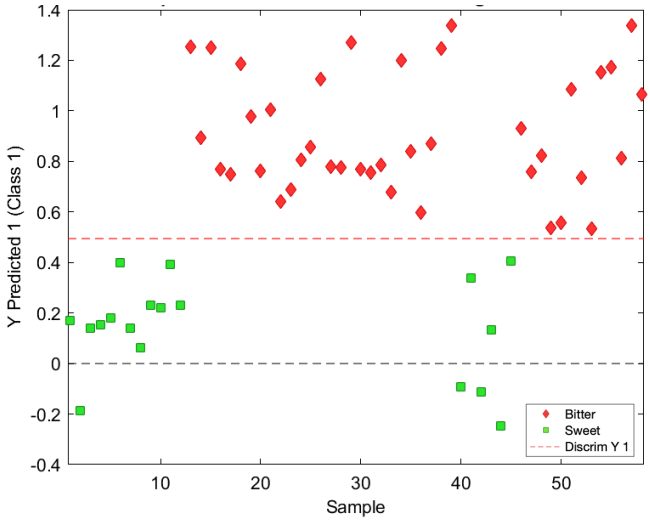


Fig. 6. PLS-DA discrimination plot for the prediction of bitter and sweet shelled almonds measured with NeoSpectra and based on one measurement session.

Finally, the global models were built for each device by putting one measurement session (186 almonds) in calibration and the other measurement session in prediction (61 almonds). The aim was to evaluate the performance of each device for predicting results from different measurement sessions. The results (Table 5) showed that the classification models were good and similar to previous models for in-shell almonds with both devices.

For shelled almonds, SCiO gave good results and produced similar results with separate measurement sessions (Fig. 7). NeoSpectra did not produce good models and most of the almonds were not correctly classified (Fig. 8). It was concluded that NeoSpectra provides different results depending on the analytical session that may depend on the instrumental variations. This type of bias was not observed for SCiO as the results were similar to previous models.

To observe if there is a shift from one measurement session to the next, a PCA analysis was performed on shelled almonds measured with NeoSpectra. The scores plot (Supplementary Information Fig. S13) showed a clear shift on the third PC after the smoothing and 2nd-

degree derivative were applied. This confirms the instrumental variations that may affect the device's performance. A viable solution might be the recalibration of the models with new data sets each time a new measurement is performed. It is important to note that this variation did not affect the classification of in-shell almonds with NeoSpectra since their classification may depend on other properties of the shell such as texture and colour and not the amygdalin content as was previously discussed. Further study is required to find the exact sources of this variation with NeoSpectra.

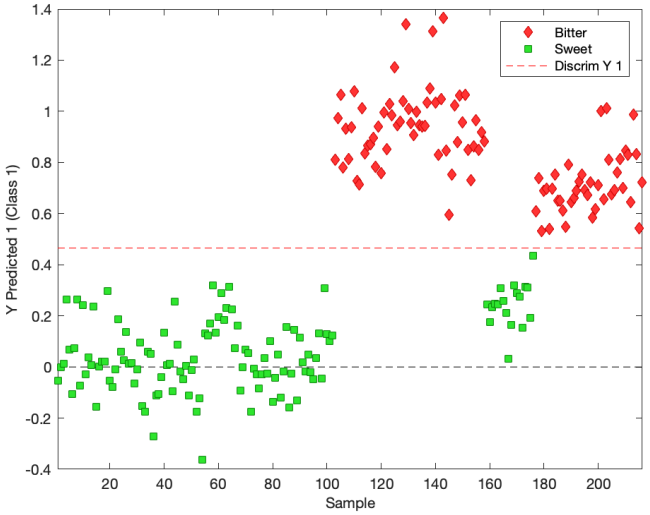


Fig. 7. PLS-DA discrimination plot for the prediction of bitter and sweet almonds measured with SCiO based on the global models.

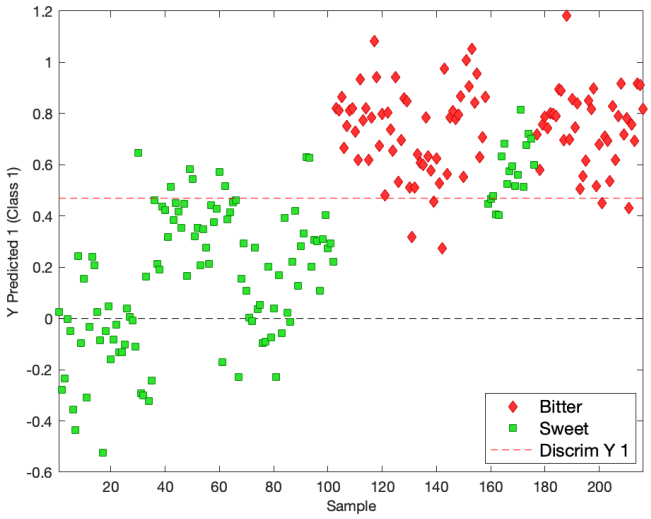


Fig. 8. PLS-DA discrimination plot for the prediction of bitter and sweet almonds measured with NeoSpectra based on the global models.

Table 1. Error covariance and correlation matrices calculated from raw spectra and using different preprocessing methods with SCiO instrument.

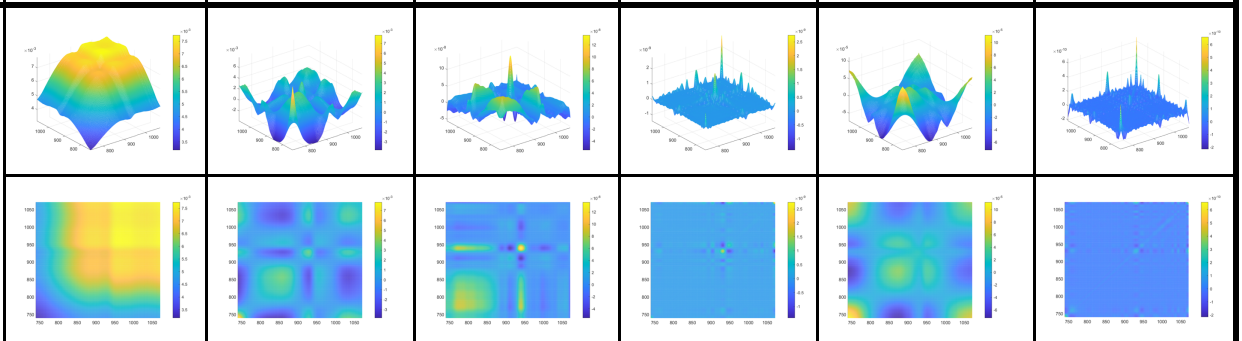
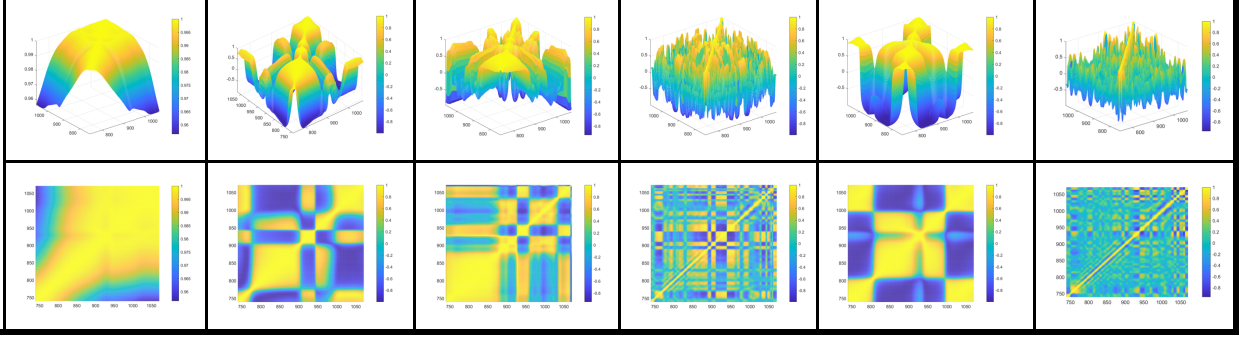
Preprocessing	Raw	MSC	1st-Derivative	2nd-Derivative	Detrend	2nd-Derivative +MSC
Covariance Matrix						
Correlation Matrix						

Table 2. Cumulative percent variance captured from the residual matrix and the error covariance matrix by PCA models performed on SCiO measurements.

	Principal Component Number	*% Cumulative variance of residuals matrix*	*% Cumulative variance of error covariance matrix*
In-shell	1	99.68	99.97
	2	99.98	100.00
Shelled	1	99.35	99.97
	2	99.96	100.00

Table 3. Cumulative percent variance captured from the residual matrix and the error covariance matrix by PCA models performed on NeoSpectra measurements.

	Principal Component Number	% Cumulative variance of residuals matrix	% Cumulative variance of error covariance matrix
In-shell	1	89.80	96.95
	2	95.08	99.45
	3	96.61	99.89
Shelled	1	97.01	99.02
	2	98.98	99.96
	3	99.55	99.99

Table 4. PLSD-DA classification results of almonds based on separate analytical sessions performed with SCiO and NeoSpectra.

		SCiO	NeoSpectra
In-shell	LVs	3	3
	Sensitivity	1	1
	Specificity	1	1
	RMSEC	0.16	0.19
	RMSECV	0.17	0.20
	RMSEP	0.18	0.18
Shelled	LVs	3	5
	Sensitivity	1	1
	Specificity	1	1
	RMSEC	0.13	0.23
	RMSECV	0.15	0.28
	RMSEP	0.12	0.27

Table 5. PLSD-DA classification results of almonds based on global models performed with SCiO and NeoSpectra.

		SCiO	NeoSpectra
In-shell	LVs	3	3
	Sensitivity	1	1
	Specificity	1	0.98
	RMSEC	0.14	0.19
	RMSEP	0.18	0.23
Shelled	LVs	3	3
	Sensitivity	1	0.95
	Specificity	1	0.22
	RMSEC	0.14	0.59
	RMSEP	0.29	0.41

4. Conclusions

Miniaturized instruments are evolving every day and reliable results can only be obtained through structured characterizations of the sources of variability that may influence their performance. In this work, two low-cost miniaturized NIR spectrometers have been evaluated through multivariate measurement errors using covariance and correlation matrices as a method to understand the sources of variability that affect the models and to select the optimal preprocessing methods that could be included in the classification models during the data analysis stage to distinguish sweet and bitter almonds. The effect of instrumental and replacement replicates on the data sets was also studied through the error covariance matrices. In both devices, the instrumental variations had a smaller effect on the measurements compared with the variations due to the repositioning of the samples. However, in NeoSpectra the differences between both types of replicates were comparable suggesting a greater contribution from instrumental variations.

PLS-DA successfully classified the almonds after applying the optimal preprocessing determined through the visual inspection of covariance and correlation matrices. Both SCiO and NeoSpectra can be used to classify almonds by bitterness with and without the shells with SCiO performing better in all models. SCiO is also able to predict new measurements from old data sets. NeoSpectra showed some variations along the analytical sessions and may require recalibration with new data sets to produce reliable results.

ACKNOWLEDGMENTS

I would like to express my gratitude to Dr. Jordi Riu, Dr. Ricard Boqué, and Dr. Barbara Giussani for their scientific guidance, valuable feedback, and constant encouragement throughout the master project.

REFERENCES

(1) Rodriguez-Saona, L.; Aykas, D. P.; Borba, K. R.; Urtubia, A. *Current Opinion in Food Science* **2020**, *31*, 136–150. https://doi.org/10.1016/j.cofs.2020.04.008.

(2) Agelet, L. E.; Hurburgh, C. R. *Critical Reviews in Analytical Chemistry* **2010**, *40* (4), 246–260. https://doi.org/10.1080/10408347.2010.515468.

(3) Beć, K.; Grabska, J.; Huck, C. *Chemistry - A European Journal* **2020**. https://doi.org/10.1002/chem.202002838.

(4) Ozaki, Y.; Šašić, S.; Jiang, J. H. *Journal of Near Infrared Spectroscopy* **2001**, *9* (2), 63–95.

(5) Ozaki, Y.; Morita, S.; Du, Y. Spectral Analysis. In *Near-Infrared Spectroscopy in Food Science and Technology*; John Wiley & Sons, Ltd, **2006**; pp 47–72. https://doi.org/10.1002/9780470047705.ch3.

(6) Giussani, B.; Gorla, G.; Riu, J. *Critical Reviews in Analytical Chemistry* **2022**, *0* (0), 1–33.

https://doi.org/10.1080/10408347.2022.2047607.
(7) Huang, J.; Wen, Q.; Nie, Q.; Chang, F.; Zhou, Y.; Wen, Z. *Micromachines (Basel)* **2018**, *9* (10). https://doi.org/10.3390/mi9100478.
(8) Schuler, L. P.; Milne, J. S.; Dell, J. M.; Faraone, L. *Journal of Physics D: Applied Physics* **2009**, *42* (13), 133001. https://doi.org/10.1088/0022-3727/42/13/133001.
(9) Pasquini, C. *Analytica Chimica Acta* **2018**, *1026*, 8–36. https://doi.org/10.1016/j.aca.2018.04.004.
(10) Beć, K. B.; Grabska, J.; Siesler, H. W.; Huck, C. W. *NIR news* **2020**, *31* (3–4), 28–35. https://doi.org/10.1177/0960336020916815.
(11) Kirchler, C. G.; Pezzei, C. K.; Beć, K. B.; Mayr, S.; Ishigaki, M.; Ozaki, Y.; Huck, C. W. *Analyst* **2017**, *142* (3), 455–464. https://doi.org/10.1039/C6AN02439D.
(12) Wiedemair, V.; Mair, D.; Held, C.; Huck, C. *Talanta* **2019**, *205*, 120115. https://doi.org/10.1016/j.talanta.2019.120115.
(13) Hoffmann, U.; Pfeifer, F.; Hsuing, C.; Siesler, H. *Applied Spectroscopy* **2016**, *70*, 852–860. https://doi.org/10.1177/0003702816638284.
(14) Yan, H.; Siesler, H. W. *J Pharm Biomed Anal* **2018**, *160*, 179—186. https://doi.org/10.1016/j.jpba.2018.07.048.
(15) Giussani, B.; Escalante-Quiceno, A. T.; Boqué, R.; Riu, J. *Foods* **2021**, *10* (11). https://doi.org/10.3390/foods10112856.
(16) Riu, J.; Gorla, G.; Chakif, D.; Boqué, R.; Giussani, B. *Foods* **2020**, *9* (8). https://doi.org/10.3390/foods9081090.
(17) Yan, H.; Siesler, H. W. *Applied Spectroscopy* **2018**, *72* (9), 1362–1370. https://doi.org/10.1177/0003702818777260.
(18) Pierna, J. A. F.; Vermeulen, P.; Lecler, B.; Baeten, V.; Dardenne, P. *Applied Spectroscopy* **2010**, *64* (6), 644–648. https://doi.org/10.1366/000370210791414353.
(19) Zamora-Rojas, E.; Pérez-Marín, D.; de Pedro-Sanz, E.; Guerrero-Ginel, J. E.; Garrido-Varo, A. *Chemometrics and Intelligent Laboratory Systems* **2012**, *114*, 30–35. https://doi.org/10.1016/j.chemolab.2012.02.001.
(20) Matinrad, F.; Kompany-Zareh, M.; Omidikia, N.; Dadashi, M. *Analytica Chimica Acta* **2020**, *1129*, 98–107. https://doi.org/10.1016/j.aca.2020.06.066.
(21) Wentzell, P. *J Braz Chem Soc* **2013**, *25*. https://doi.org/10.5935/0103-5053.20130293.
(22) Wentzell, P.; Wicks, C.; Braga, J.; Soares, L.; Pastore, T.; Coradin, V.; Davrieux, F. *Canadian Journal of Chemistry* **2018**, *96*. https://doi.org/10.1139/cjc-2017-0730.
(23) Leger, M. N.; Vega-Montoto, L.; Wentzell, P. D. *Chemometrics and Intelligent Laboratory Systems* **2005**, *77* (1), 181–205. https://doi.org/10.1016/j.chemolab.2004.09.017.
(24) Gorla, G.; Taiana, A.; Boqué, R.; Bani, P.; Gachiuta, O.; Giussani, B. *Analytica Chimica Acta* **2022**, *1211*, 339900. https://doi.org/10.1016/j.aca.2022.339900.
(25) Vega-Castellote, M.; Pérez-Marín, D.; Torres, I.; Moreno-Rojas, J.-M.; Sánchez, M.-T. *Journal of Food Engineering* **2021**, *294*, 110406. https://doi.org/10.1016/j.jfoodeng.2020.110406.
(26) Torres, I.; Sánchez, M.-T.; Vega-Castellote, M.; Pérez-Marín, D. *Foods* **2021**, *10* (6). https://doi.org/10.3390/foods10061221.
(27) Wiedemair, V.; Langore, D.; Garsleitner, R.; Dillinger, K.; Huck, C*Molecules* **2019**, *24* (3). https://doi.org/10.3390/molecules24030428.
(28) Huck, C. W. New Trend in Instrumentation of NIR Spectroscopy—Miniaturization. In *Near-Infrared Spectroscopy: Theory, Spectral Analysis, Instrumentation, and Applications*; Ozaki, Y., Huck, C., Tsuchikawa, S., Engelsen, S. B., Eds.; Springer Singapore: Singapore, **2021**; pp 193–210. https://doi.org/10.1007/978-981-15-8648-4_8.
(29) Bro, R.; Smilde, A. K. *Anal. Methods* **2014**, *6* (9), 2812–2831. https://doi.org/10.1039/C3AY41907J.
(30) Borràs, E.; Amigo, J. M.; van den Berg, F.; Boqué, R.; Busto, O. *Food Chemistry* **2014**, *153*, 15–19. https://doi.org/10.1016/j.foodchem.2013.12.032.

Performance characterization of miniaturized near-infrared (NIR) spectrometers for the classification of sweet and bitter almonds

Supplementary Information

Hawbeer Jamal Ahmed
Master Program in Nanoscience, Materials, and Processes: Chemical Technology at the Frontier
2021-2022
e-mail: hawbeerjamal.ahmed@estudiants.urv.cat

Supervisors: Dr Jordi Riu[1], Dr Ricard Boqué[1], and Dr Barbara Giussani[2]
[1]*Department of Analytical Chemistry and Organic Chemistry. Universitat Rovira i Virgili*
c/ Marcel·lí Domingo 1, 43007 Tarragona, Spain.
[2]*Dipartimento di Scienza e Alta Tecnologia, Università degli Studi dell'Insubria, Via Valleggio, 9, 22100 Como, Italy*

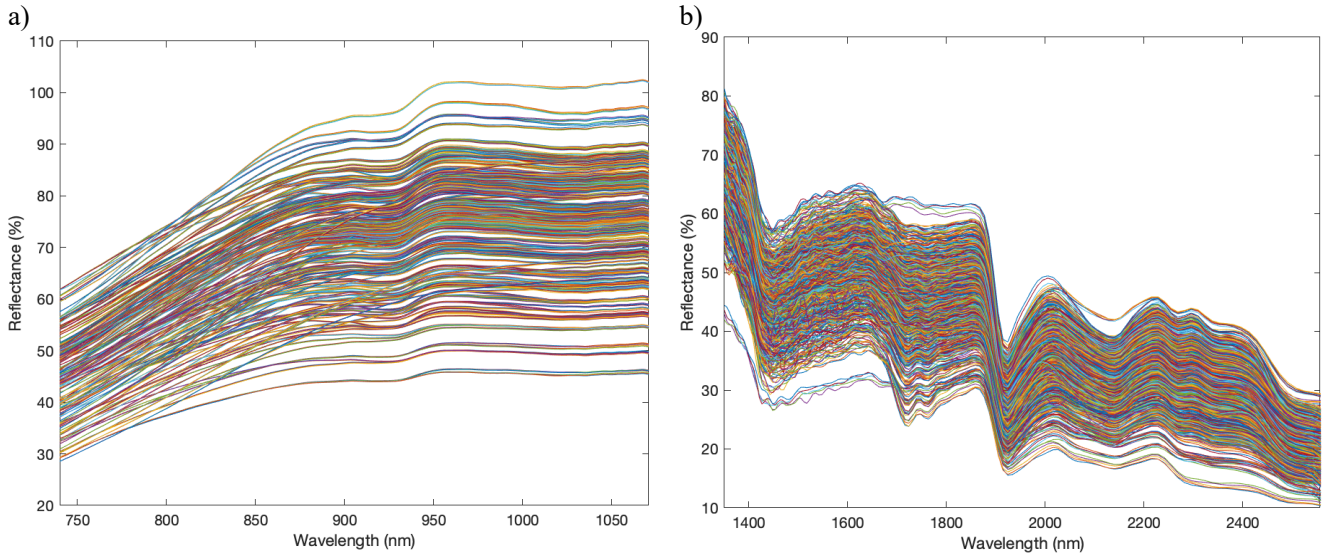


Fig. S1. Raw spectra of 58 shelled almonds acquired with a) SCiO, b) NeoSpectra. The NeoSpectra measurements appear to be noisier than SCiO, especially in the range of 1400-1800 nm.

a)

b)

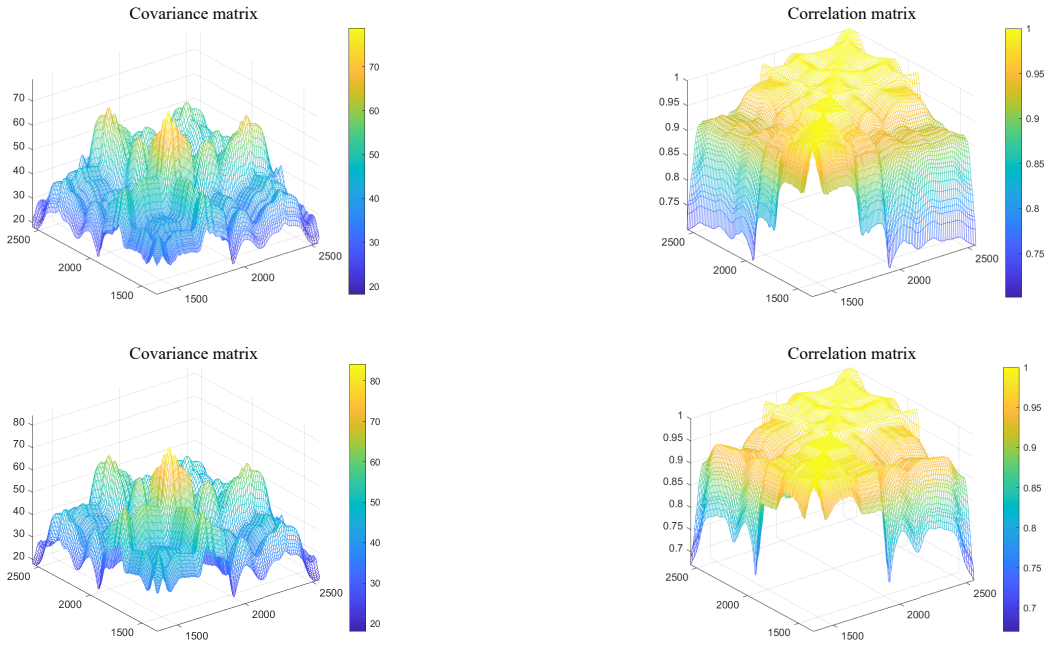


Fig. S2. Error covariance and correlation matrices calculated from test measurements performed on 9 shelled almonds measured with NeoSpectra; a) under ambient light, b) in a dark box. Results showed no obvious differences between ECMs, suggesting that the interferences from surrounding light have no effect on the measurements.

a)

b)

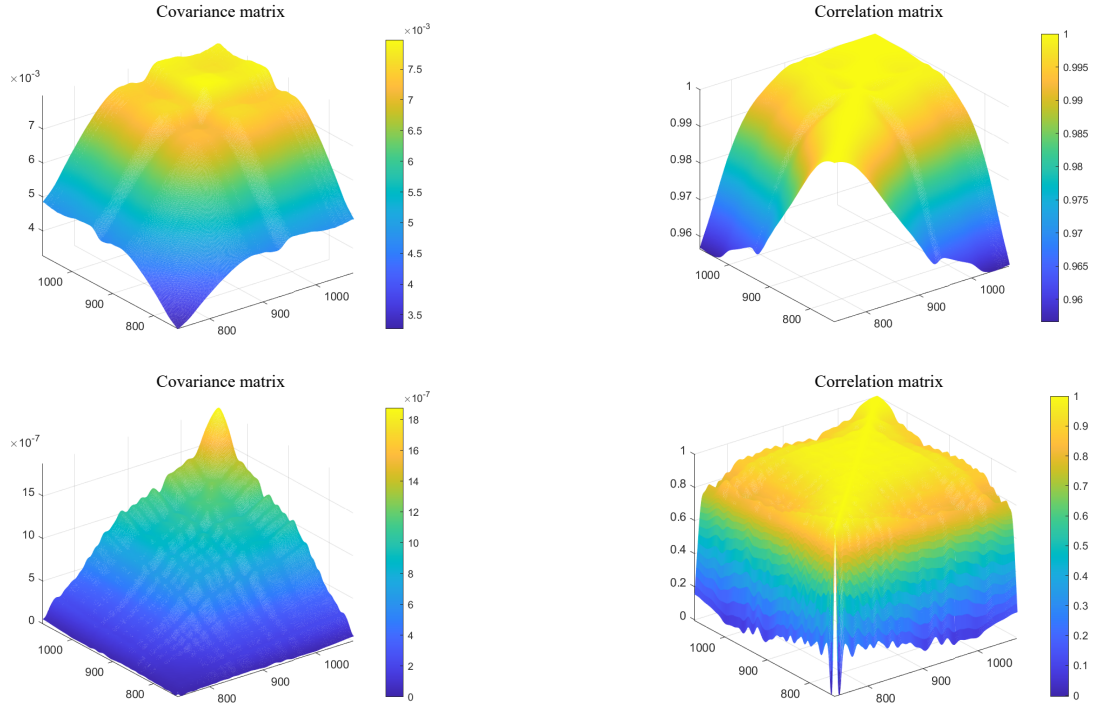
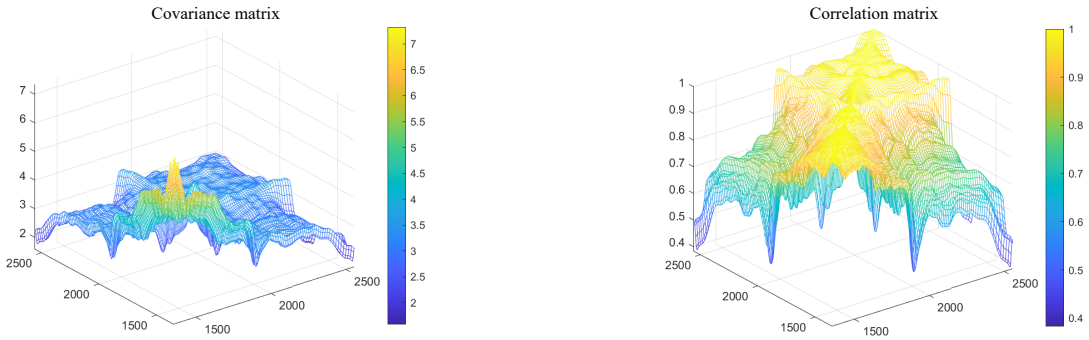


Fig. S3. Error covariance and correlation matrices calculated from test measurements performed on 9 shelled almonds measured with SCiO by using; a) replacement replicates, b) instrumental replicates. As can be seen from the magnitude of the covariance matrix, the instrumental variations had a much smaller effect on the measurement errors than the sample variations.

a)



b)

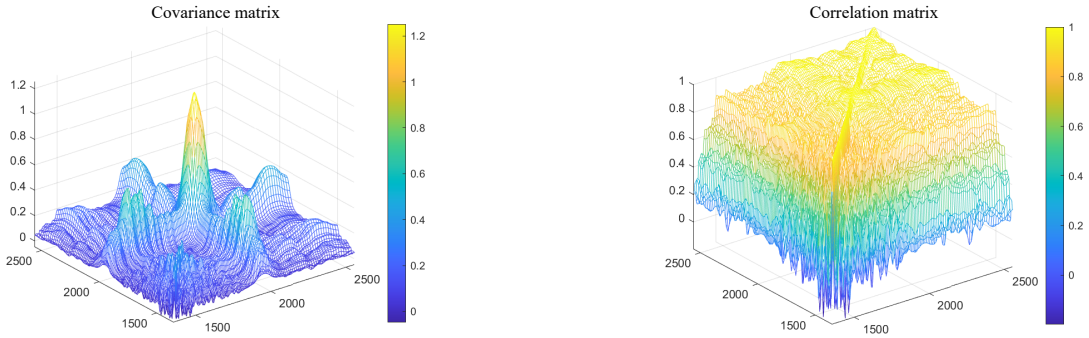
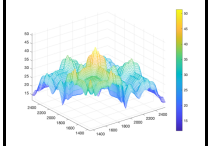
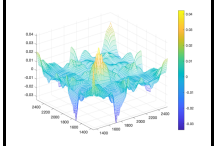
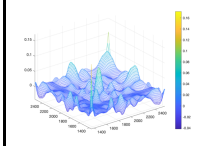
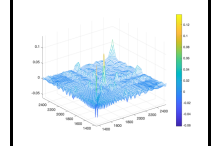
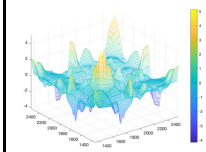
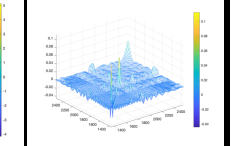
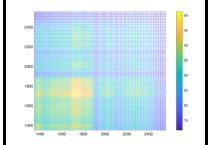
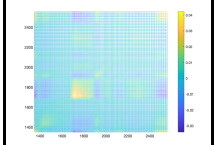
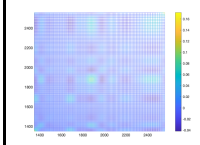
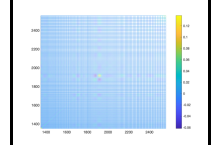
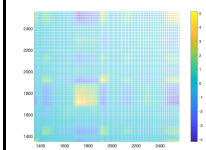
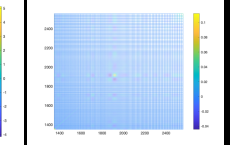
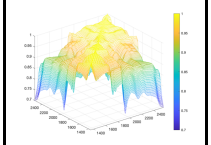
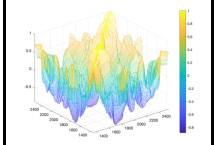
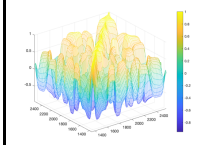
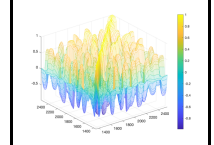
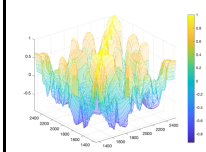
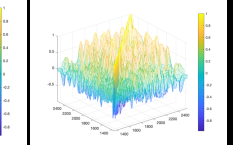
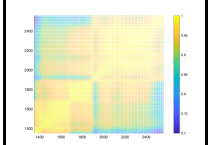
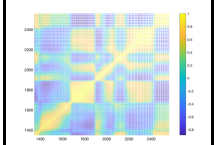
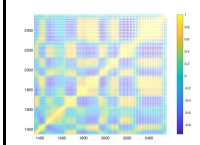
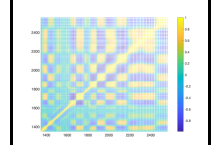
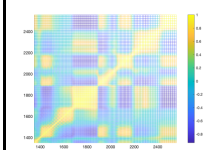
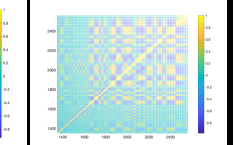


Fig. S4. Error covariance and correlation matrices calculated from test measurements performed on 9 shelled almonds measured with NeoSpectra by using; a) replacement replicates, b) instrumental replicates. As can be seen from the magnitude of the covariance matrix, the instrumental variations had a comparable effect on the measurement errors to the sample variations.

Table S1. Error covariance and correlation matrices calculated for raw spectra and different preprocessing methods with the NeoSpectra instrument. Smoothing combined with 2nd-degree derivative proved to be the optimal preprocessing.

Preprocessing	Raw	SNV	1st-Derivative	2nd-Derivative	Detrend	Smoothing+2nd-Derivative
Covariance Matrix						
						
Correlation Matrix						
						

To test the effectiveness of the preprocessing selection through ECM calculations, eight solid samples with different physical and chemical characteristics were measured with both instruments. They were purchased from local markets in Como, Italy. The samples were: brown powder sugar, white powder sugar, sugar cube, white homogeneous pill, white nonhomogeneous pill, white rice, brown rice, and colored pill. The spectra were acquired with 15 replacement replicates for each sample. The optimal preprocessing selection was also performed through the visual inspection of ECM graphs.

With SCiO, the 1st-degree derivative combined with SNV gave the best separation. The results of the raw signal compared with the best preprocessing can be seen in Fig. S5. The scores plot of the PCA analysis for both raw and pre-processed data is given in Fig. S6. With NeoSpectra, smoothing combined with the 2nd-degree derivative gave the best separation. The results of the raw signal compared with the best preprocessing can be seen in Fig. S7. The scores plot of the PCA analysis for both raw and preprocessed data is given in Fig. S8.

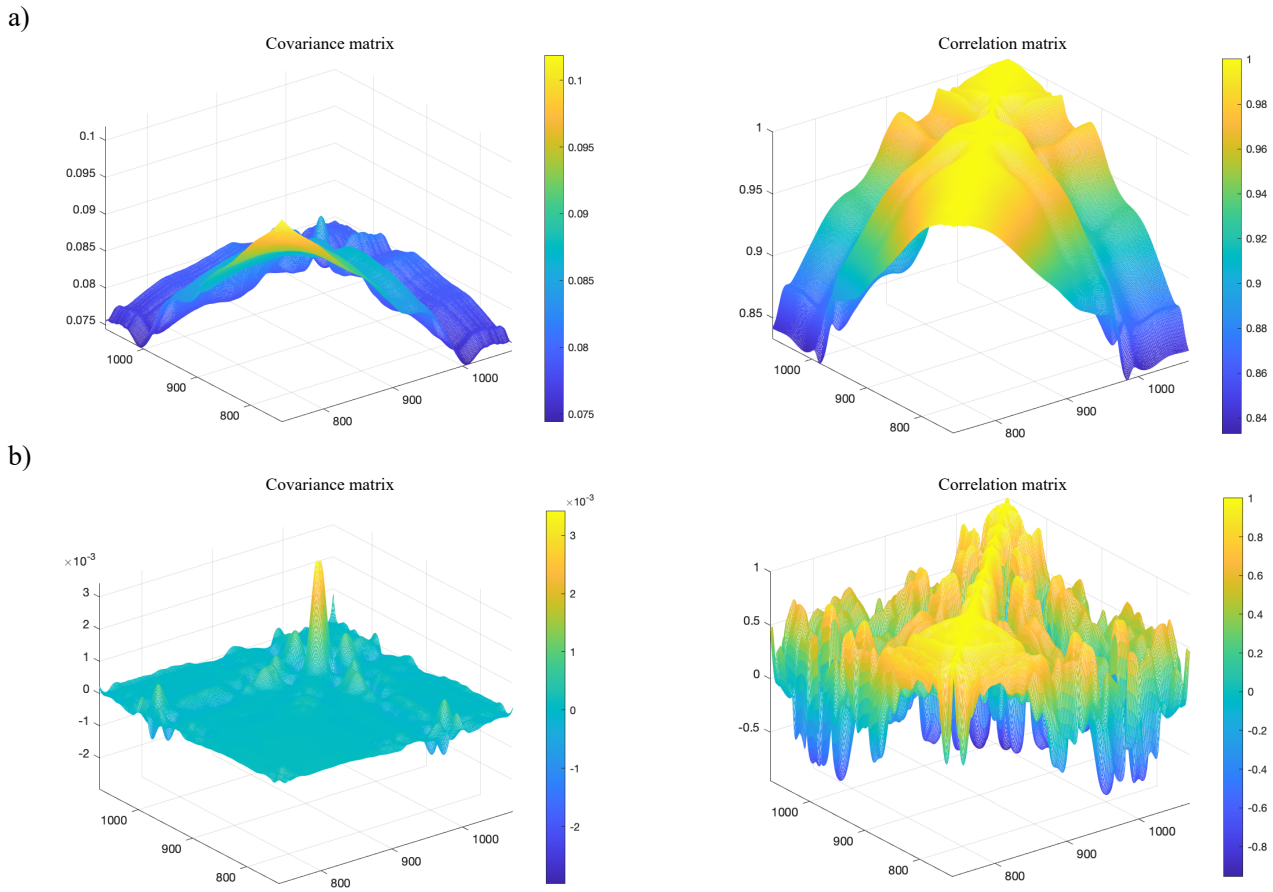


Fig. S5. Error covariance and correlation matrices calculated from test measurements performed on the solid samples measured with SCiO by using; a) raw spectra, b) 1st-degree derivative combined with SNV.

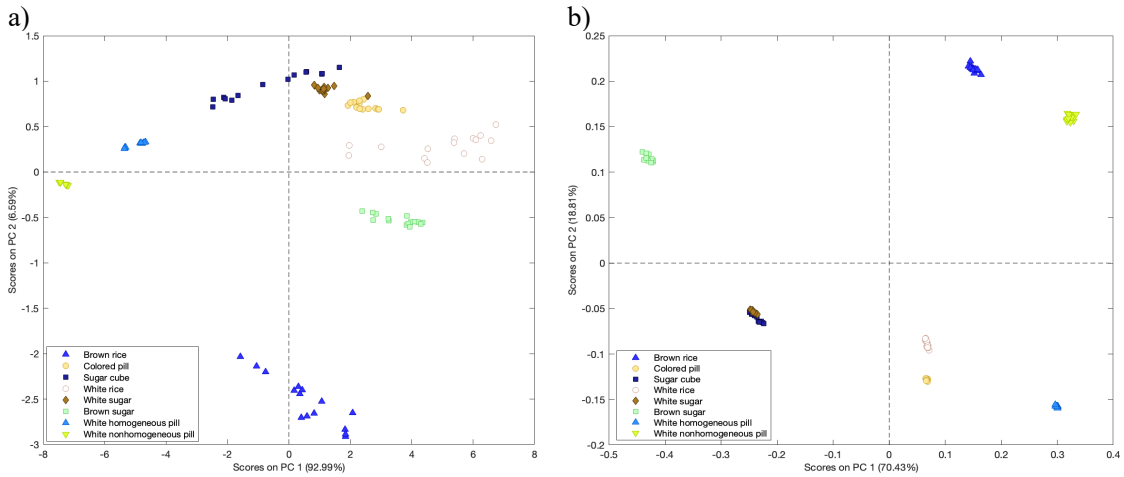


Fig. S6. Scores plots calculated from test measurements performed on the solid samples measured with SCiO by using; a) raw spectra, b) 1st-degree derivative combined with SNV. After applying the optimal preprocessing, a good clustering based on the types of samples was obtained.

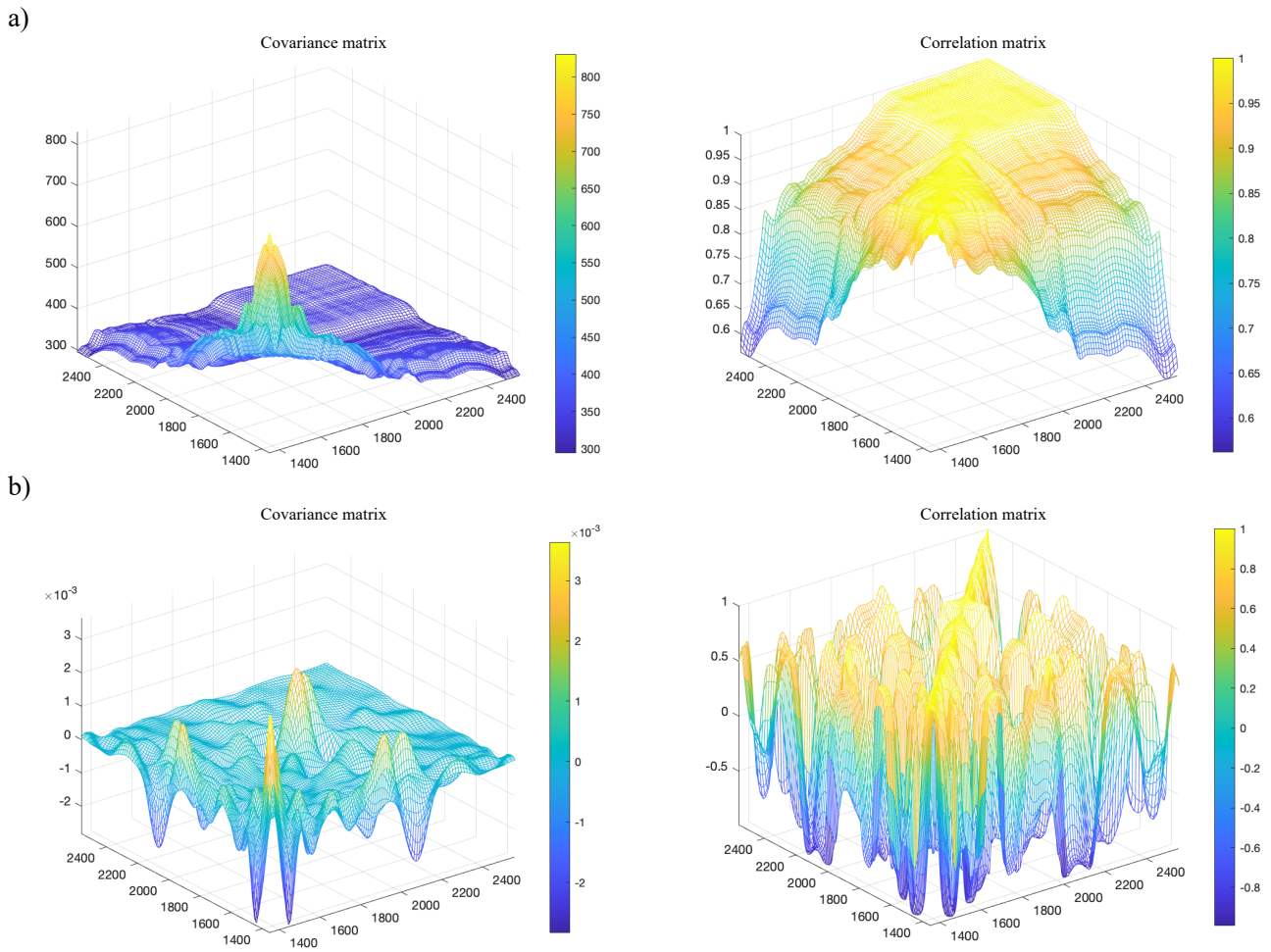


Fig. S7. Error covariance and correlation matrices calculated from test measurements performed on the solid samples measured with SCiO by using; a) raw spectra, b) smoothing combined with 2nd-degree derivative.

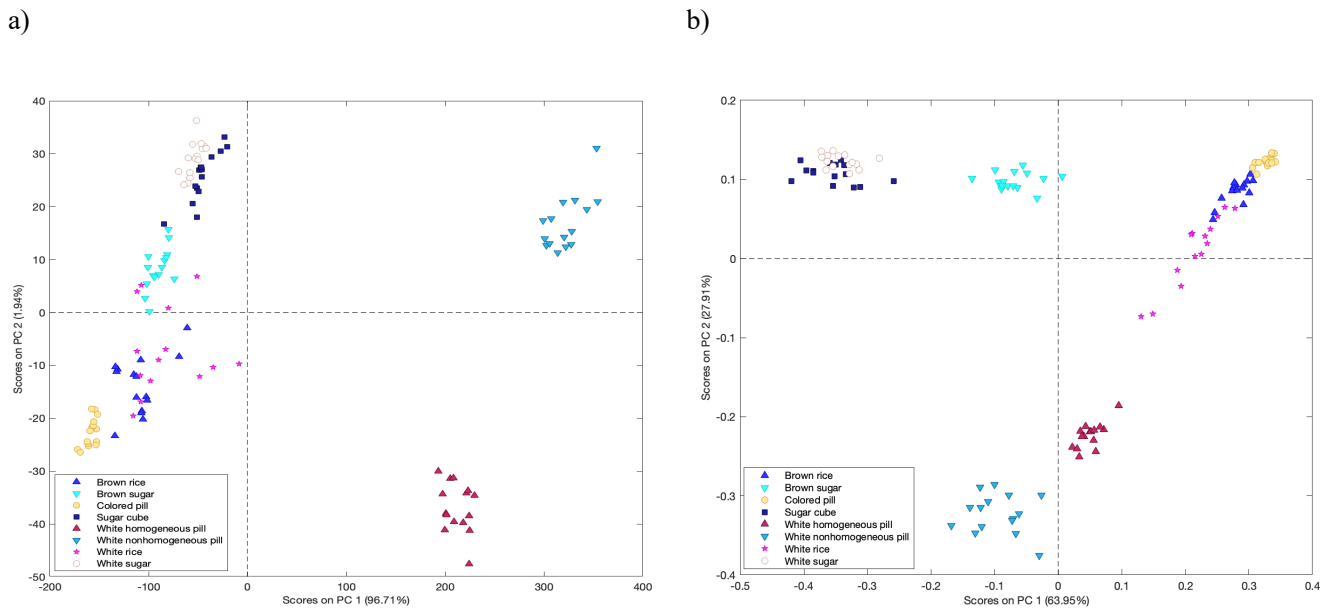


Fig. S8. Scores plots calculated from test measurements performed on the solid samples measured with NeoSpectra by using; a) raw spectra, b) smoothing combined with 2nd-degree derivative. After applying the optimal preprocessing, a better clustering based on the types of samples was obtained.

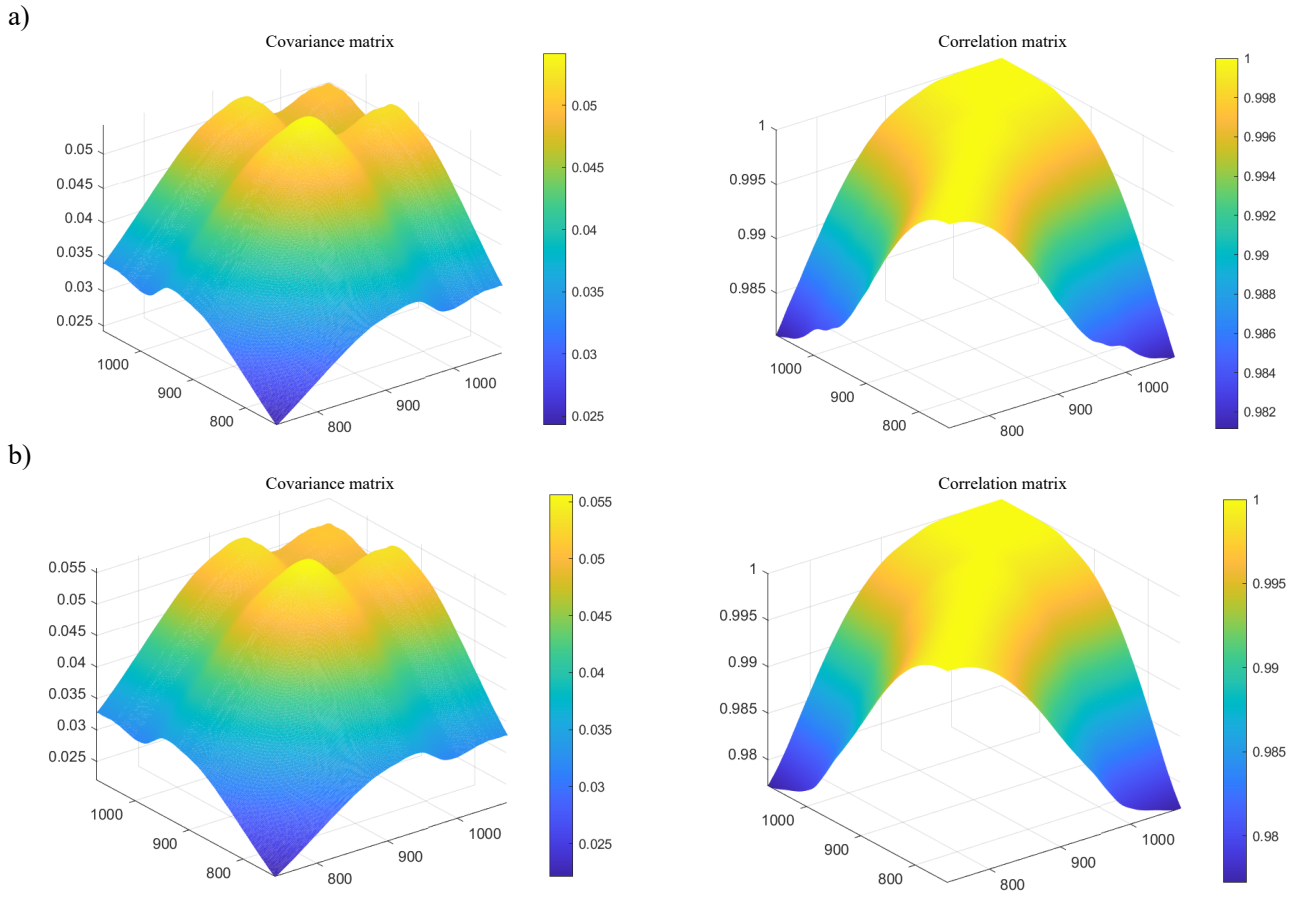


Fig. S9. Error covariance and correlation matrices calculated from test measurements performed on 21 in-shell almonds measured with SCiO by using; a) shell with the kernel, b) shell without the kernel. The ECM is very similar for both measurements, suggesting that the radiation does not pass through the shell.

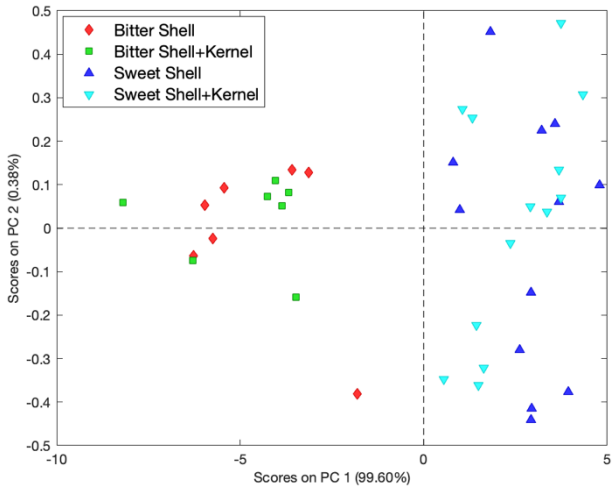


Fig. S10. Scores plot of PCA analysis from test measurements performed on 21 in-shell almonds measured with SCiO. No separation can be seen between the shell and the shell with the kernel suggesting that the radiation does not pass through the shell.

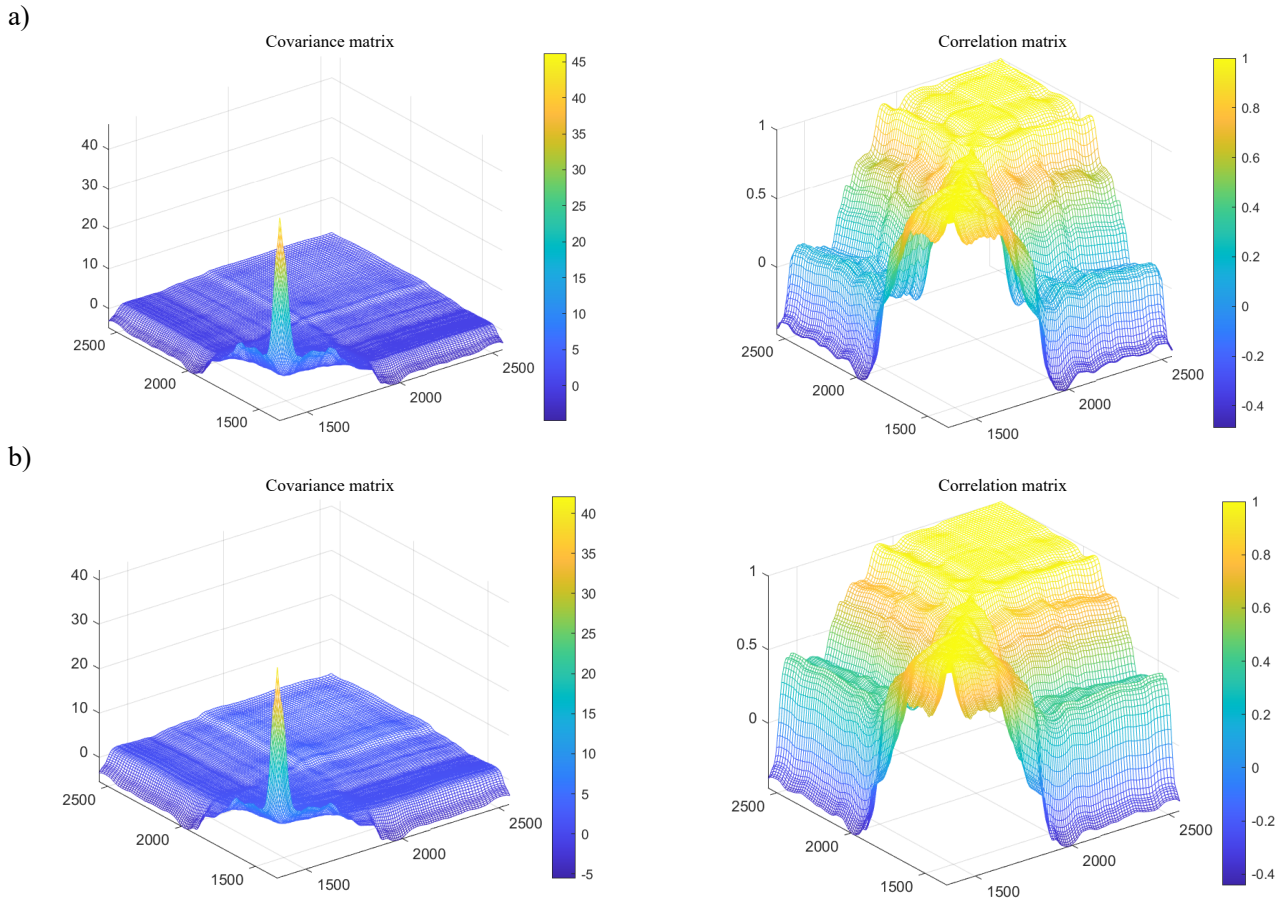


Fig. S11. Error covariance and correlation matrices calculated from test measurements performed on 21 in-shell almonds measured with NeoSpectra by using; a) shell with the kernel, b) shell without the kernel. The ECM is very similar for both measurements, suggesting that the radiation does not pass through the shell.

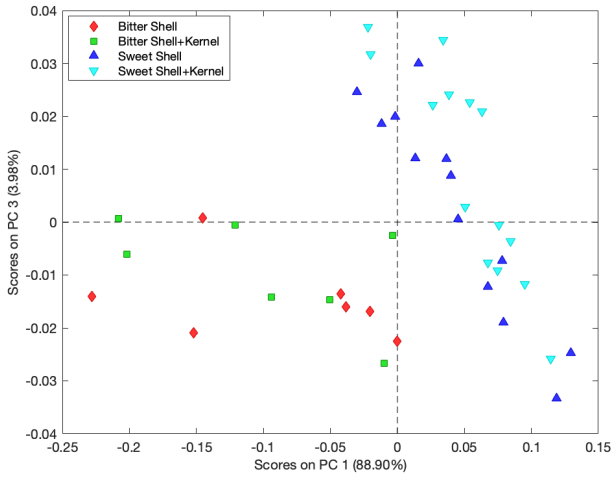


Fig. S12. Scores plot of PCA analysis from test measurements performed on 21 in-shell almonds measured with NeoSpectra. No separation can be seen between the shell and the shell with the kernel suggesting that the radiation does not pass through the shell.

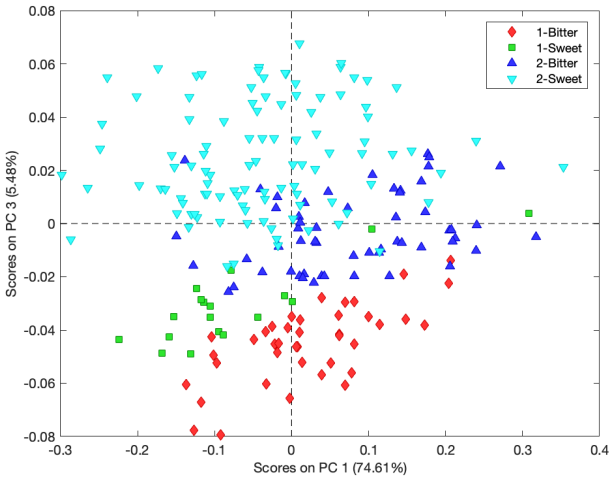


Fig.S13. Scores plot from two measurement sessions performed on shelled almonds measured with NeoSpectra. A clear shift on the third PC can be seen between the two measurement sessions. This confirms the instrumental variations that may affect the performance of NeoSpectra.