Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 326 (2025) 125225

Contents lists available at ScienceDirect

Spectrochimica Acta Part A:
Molecular and Biomolecular Spectroscopy

journal homepage: www.journals.elsevier.com/spectrochimica-acta-part-a-molecular-and-biomolecular-spectroscopy

Semi-quantitative chemometric models for characterization of mixtures of sugars using infrared spectral data

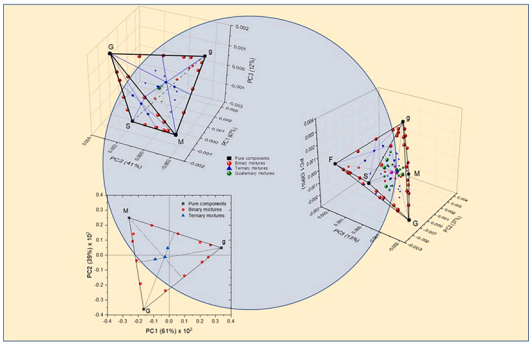
Anna Luiza B. Brito [a,*], Inês F. Cardoso [a], Luís P. Viegas [a], Rui Fausto [a,b]

[a] CQC-IMS, Department of Chemistry, University of Coimbra, 3004-535 Coimbra, Portugal
[b] Faculty of Sciences and Letters, Department of Physics, Istanbul Kultur University, Ataköy Campus, Bakirköy 34156, Istanbul, Turkey

HIGHLIGHTS

- Chemometric models based on IR spectroscopy data were developed for the analysis of mixtures of sugars.
- The compositions of mixtures of up to 5 sugars were estimated successfully at semi-quantitative level.
- Novel method based on the PCA scores along the most significant PCs, used as barycentric compositional coordinates.
- The new method is practical, fast, simple to implement and can be applied even in absence of data for the pure components.
- The method can used for the semi-quantitative analysis of other types of mixtures and applicable to other types of data.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Saccharides
Mixtures
Qualitative/semi-quantitative analyses
ATR-FTIR
PCA scores plots
MCR-ALS

ABSTRACT

Sugars (saccharides) are sweet-tasting carbohydrates that are abundant in foods and play very important roles in living organisms, particularly as sources and stores of energy, and as structural elements in cellular membranes. They are desirable therapeutic targets, as they participate in multiple metabolic processes as fundamental elements. However, the physicochemical characterization of sugars is a challenging task, mostly due to the structural similarity shared by the large diversity of compounds of this family. The need for fast, accurate enough, and cost-effective analytical methods for these substances is of extreme relevance, in particular because of the recently increasing importance of carbohydrates in Medicine and food industry. With this in view, this work focused on the development of chemometric models for semi-quantitative analysis of samples of different types of sugars (glucose, galactose, mannitol, sorbose and fructose) using infrared spectra as data, as an example of application of a novel approach, where the Principal Component Analysis (PCA) score plots are used to estimate the composition (weight-%) of the mixtures of the sugars. In these plots, polygonal geometric shapes emerge in the vectorial space of the most significant principal components, that allow grouping different types of samples on the vertices, edges, faces and interior of the polygons according to the composition of the samples. This approach was applied successfully to mixtures of up to 5 sugars and shown to appropriately extract the compositional information from the hyper-redundant complex spectral data. Thought the method has been applied here to a specific problem, it shall be considered as a general procedure for the semi-quantitative analysis

* Corresponding author.
E-mail address: anna.brito@qui.uc.pt (A.L.B. Brito).

https://doi.org/10.1016/j.saa.2024.125225
Received 18 July 2024; Received in revised form 9 September 2024; Accepted 26 September 2024
Available online 27 September 2024

of other types of mixtures and applicable to other types of data reflecting their composition. In fact, the methodology appears as an efficient tool to solve three main general problems: (i) use hyper-redundant (in variables) data, as spectral information, directly and with minimum pre-treatment, to evaluate semi-quantitatively the composition of mixtures; (ii) do this for systems which produce data that can be considered rather similar; and (iii) do it for a number of substances present in the mixtures that might be greater than that usually considered in chemistry, which in general is limited to 3 components. In addition, this work also demonstrates that, similarly to the developed analysis based on the PCA score plots, the Multivariate Curve Resolution with Alternating Least Squares (MCR-ALS) chemometric method can also be used successfully for the qualitative (when used without any previous knowledge of the components present in the samples) or semi-quantitative (when the pure components spectral profiles are provided as references) analyses of mixtures of (at least) up to 5 distinct sugars.

1. Introduction

Carbohydrates are among the main products of photosynthesis, but their biological importance is not only connected with their participation in energy-related functions. They are also present on the external surface of cell membranes, acting as structural elements that provide rigidity and support to plant cells and form the exoskeletons of arthropods, and work as receptors and signalers [1]. They are also components in the structure of the nucleic acids, forming the backbone of their characteristic helical structures [1]. Carbohydrates play a crucial role in the nutrition of living beings, counting human diet, being found in a wide variety of nourishments, including grains, cereals, legumes, vegetables and fruits [2].

One of the distinct groups of carbohydrates is that of sugars (saccharides), which include mono-, di-, oligo- and polysaccharides [3]. Monosaccharides, such as glucose, fructose and galactose, are the simplest carbohydrates, while disaccharides, such as sucrose, lactose and maltose, are formed by the combination of two monosaccharides. Oligo- and polysaccharides, on the other hand, are larger molecules composed of several monosaccharides (3–10 in the first case, more than 10 in the second) linked together by glycosidic bonds, as it is the case of raffinose, amylose and cellulose [4]. Sugars provide energy (calories), which corresponds to an essential nutritional requirement for living organisms. Nevertheless, excessive sugar consumption can lead to severe diseases, in particular *diabetes mellitus*, and has been one of the main factors contributing to the continued increase of obesity worldwide [5].

The physicochemical characterization of sugars is a challenging task, mostly due to the structural similarity shared by a large diversity of compounds of this family. Even in the case of the simplest monosaccharides, structural differences can be as subtle as stereochemical details resulting, for example, from a different position or orientation of a single OH group, which leads to the existence of isomers, including enantiomers. Indeed, given the simultaneous presence of multiple OH groups in the sugar structures, this type of structural variations can give rise to a large number of different compounds sharing many common structural features. This is also true for more complex polysaccharides, for which one has to consider also the numerous possible ways of combination of their constituting monosaccharide units [5]. Due to the close structural similarity exhibited by sugars within all their diversity, the need for efficient, fast and cheap types of analytical methods for these substances is of extreme relevance, in particular because of the recently increasing importance of carbohydrates in Medicine and food industry [6].

Details regarding current methods for the chemical analysis of carbohydrates based on different techniques can be found in various recent publications [3,6–16].

Infrared (IR) spectroscopy is widely used as one of the primary spectroscopic tools for the structural study as well as qualitative/semi-quantitative and quantitative analyses of many different types of materials, including simple biological molecules and biopolymers. It is a fast, inexpensive and sensitive method, with sampling techniques available that can be tuned to the type of material under investigation. It is considered a valuable technique for research and quality control in the food industry, particularly because it provides insights into the relationship between the structure and properties of substances, while serving also as a compositional analysis tool [6–8].

Due to the extensive information present in infrared spectra, their interpretation frequently requires the use of multivariate statistical methods (or chemometrics), which act as powerful data analysis auxiliary tools, enabling the organization, classification and quantification of the acquired spectroscopic data [17,18]. Among these methods, the Principal Component Analysis (PCA) is a well-established method that allows data reduction, outlier samples' detection, and samples categorization [19–24]. PCA is a linear projection method frequently employed in exploratory non-supervised data analysis that allows reducing data dimensionality by generating new uncorrelated variables (referred to as principal components, PCs) from the original variables set [19–24]. In the practice, PCA projects multivariate data into a new system of coordinates that maximize the variance in the dataset, expressing the results in terms of simple two- or three-dimensional graphs (scores plots). Another chemometric method that has been extensively employed in chemical analysis is the Multivariate Curve Resolution with Alternating Least Squares (MCR-ALS) method [25,26]. This method is effective for monitoring mixtures of compounds, by mathematically decomposing instrumental responses into pure contributions from each component present in the system. Through this approach, it becomes possible to identify and quantify constituents within a sample, particularly when their experimental separation (chemical or physical) is not readily achievable, a frequently occurring scenario when dealing with complex samples or partially unknown chemical processes. As a result, MCR-ALS provides estimations of the concentration profile of each component of a mixture of substances, alongside with its respective instrumental response (*e.g.*, a spectrum). This approach has proven to be very effective across a wide range of applications [26].

The objective of the present article is to introduce a new strategy applicable to solve three main general problems: (i) use hyper-redundant (in variables) data, as infrared spectral information, directly and with minimum pre-treatment, to evaluate semi-quantitatively the composition of mixtures; (ii) do this for systems which produce data that can be considered rather similar, like sugars; and (iii) do it for a number of substances present in the mixtures that might be greater than that usually considered in chemistry, which in general is limited to 3 components. PCA and MCR-ALS chemometric methods are used to perform the data analysis, with PCA-based models being built from which the compositions of the samples can be estimated from the scores plots coordinates associated with the most significant principal components. The followed approach was applied successfully to mixtures of up to 5 sugars, as an example of its application to an important relevant practical type of problem, but the method shall be considered as a general procedure for the semi-quantitative analysis of other types of mixtures and applicable to other types of data reflecting their composition.

2. Experimental methods

2.1. Materials

In this study, 5 sugars were analyzed, which were used as provided

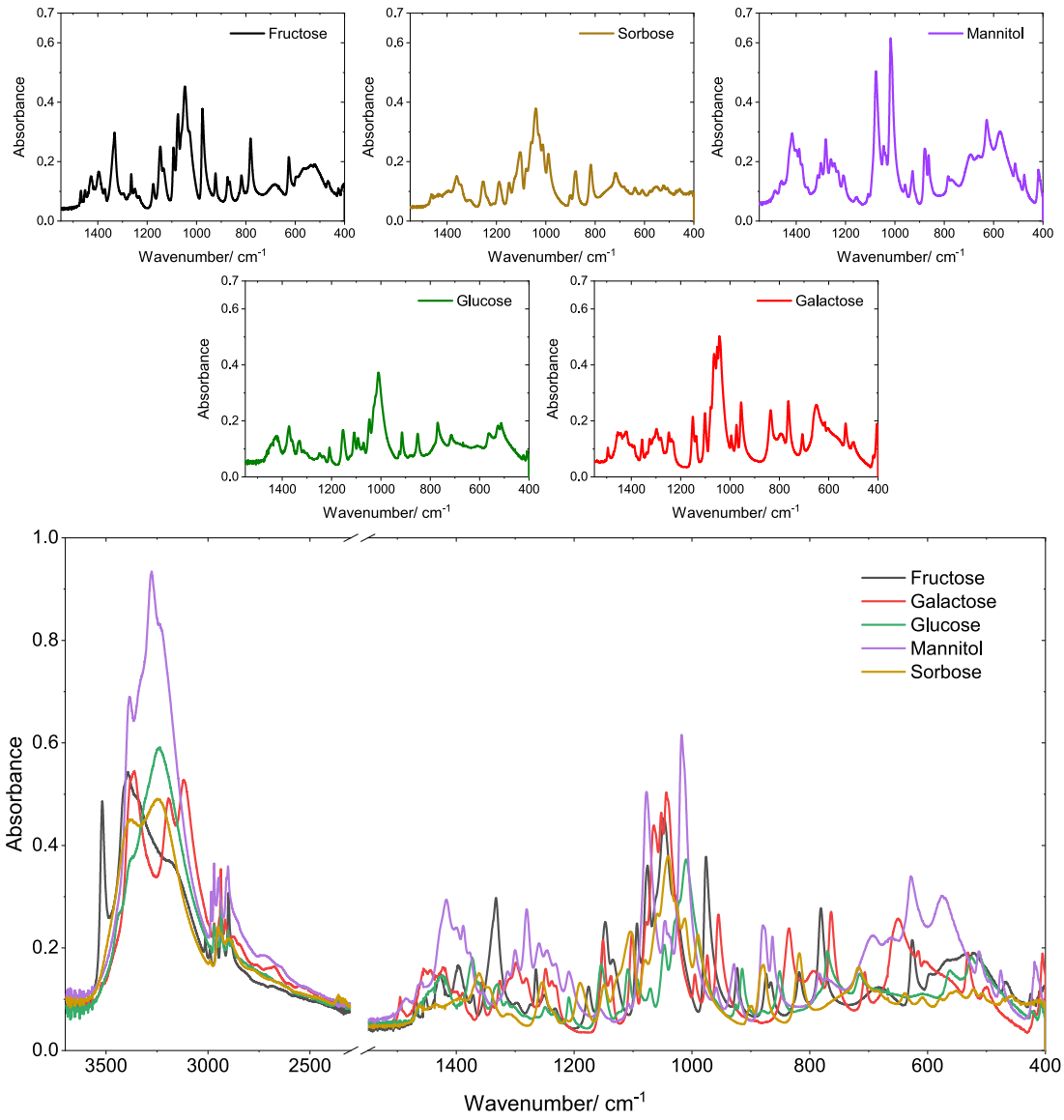


Fig. 1. ATR-FTIR spectra of the pure sugars. The spectra are shown in two views: superimposed in the full measured spectral region (*bottom panel*), and as individual spectra in the 1500–400 cm^{-1} region (*top panels*).

analytical grade: glucose (Merck), fructose (Riedel – de Haën), mannitol (Analar), galactose (Janssen Chimica) and sorbose (BDH Biochemical). Samples corresponding to mixtures of these sugars (with up to 5 components) were prepared by weighting, with the sugars being present in different proportions: binary mixtures in approximate 15/85, 25/75, 50/50 and 75/25 %-weight proportions, ternary mixtures in 40/20/40, 20/20/40 %-weight and equiponderal proportions, quaternary as 25/25/25/25, 20/30/20/30 and 10/20/30/40 %-weight proportions, and quinary mixtures in 20/20/20/20/20, 15/20/30/20/15 and 10/15/20/25/30 %-weight proportions of the constituents. In all cases, the total mass was approximately 100 mg (see Tables S1 and S2 in the Supplementary Material for more details, including the designations given to the different samples).

2.2. Instrumentation

Analytical weighing was conducted using a Kern ALJ 220-5DNM balance (Kern & Sohn, GmbH). Before weighing, the samples were ground in a mortar to enhance homogeneity, and then subjected to vortex mixing (Janke and Kunkel, VF2) at maximum speed for approximately 2 min to ensure thorough mixing of the various components present in the samples.

The FTIR spectra were obtained in a Nicolet iS5 ATR-FTIR spectrometer (Thermo Fisher Scientific), equipped with a Globar source, a KBr beamspliter and a deuterated triglycine sulphate (DTGS) detector, in the attenuated total reflectance (ATR) mode, using an ATR cell with a diamond crystal. The spectra were acquired within the range of 4000–400 cm^{-1}, with a resolution of 1 cm^{-1}, 32 scans, and data spacing of 0.120 cm^{-1}, resulting in a matrix of 29,869 variables (wavenumbers), which was used as input data for the chemometrics analysis. For all samples, spectra were taken in triplicate and averaged.

2.3. Data analyses

The data preprocessing was carried out using Unscrambler X (version 10.5.1) [27] and Origin Pro (version 2021) [28]. The spectra were first subjected to baseline linear correction and area normalized.

In the PCA analysis, the data were mean-centered, and equal weights were assigned to the variables (wavenumbers). The Single Value Decomposition (SVD) algorithm [29] was used, with cross validation (random with 20 segments). PCA-based analytical models were obtained using mixtures of 3 (glucose, galactose and mannitol), 4 (glucose,

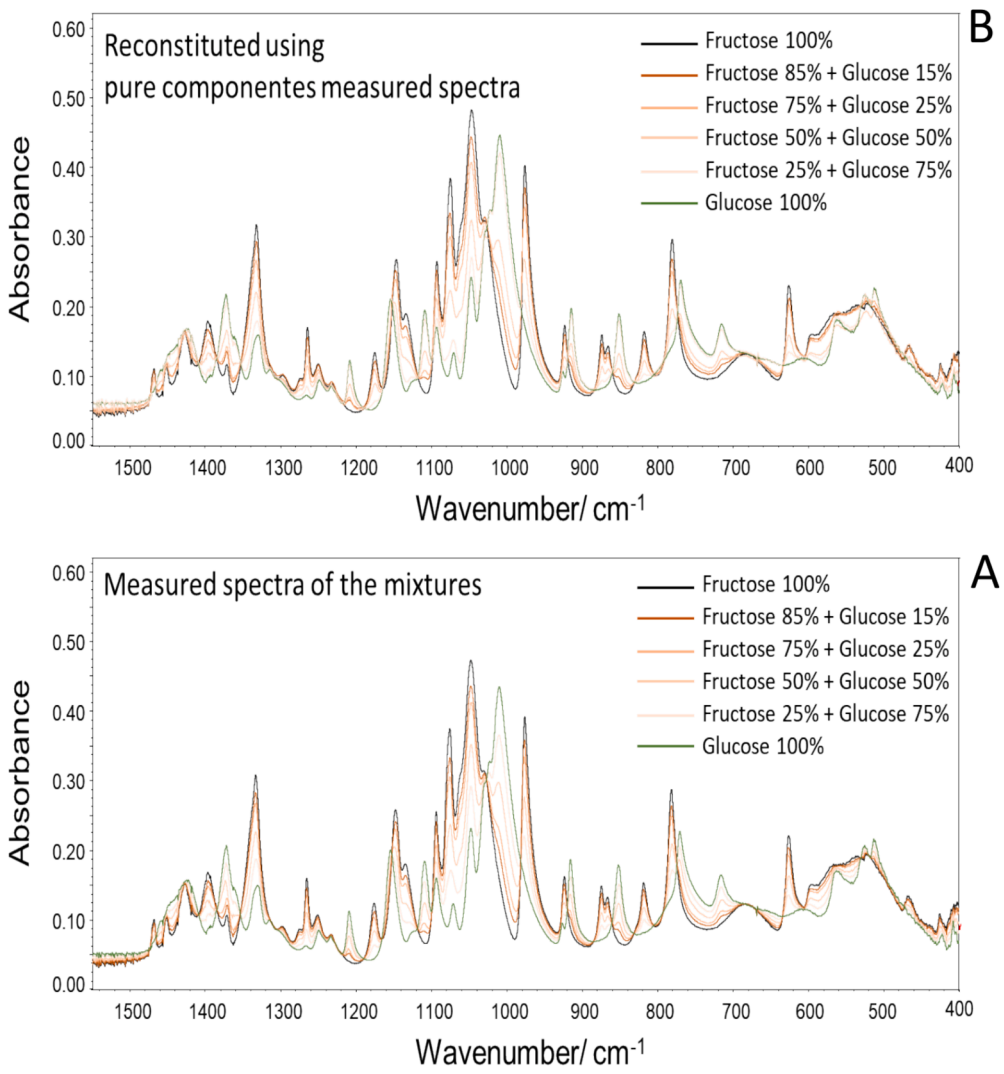


Fig. 2. A: ATR-FTIR spectra of the prepared binary mixtures of fructose and glucose with different component proportions (1500–400 cm^{-1} region). B: Reconstituted spectra of the mixtures, in the same spectral region, constructed using the spectra of the pure components added with weights according to the composition of the samples.

galactose, mannitol and sorbose), and 5 (glucose, galactose, mannitol, sorbose and fructose) sugars. In the development of the models, only the spectra of the pure components and binary mixtures were used, with higher-order mixtures data being subsequently projected onto the space of the relevant PCs of the models. The obtained scores plots were then used for semi-quantitative compositional evaluation.

As shown in Sections 3.1–3.3, samples containing different numbers of sugars appear defined in the scores plots distributed in accordance with their composition along concentration gradients, forming specific geometrical figures. For mixtures up to 3 sugars, the PC2 *vs.* PC1 2D scores plot defines a triangle whose vertices are the pure components, binary mixtures are located along the sides of the triangle, and ternary mixtures within the triangle. For mixtures to up 4 sugars, the positions of the samples in the PC3 *vs.* PC2 *vs.* PC1 3D scores plot define a tetrahedron, where pure components and binary mixtures are located at the apexes and along the edges, respectively, ternary mixtures are located in the faces of the polygon, and quaternary mixtures within the polygonal figure. In the case of the mixtures up to 5 sugars, the PC4 *vs.* PC3 *vs.* PC2 *vs.* PC1 4D scores plot defines a pentachoron [30], the positions of the samples in this geometrical figure following the generalization to 4 dimensions of what is observed for the lower dimensional problems. The obtained geometrical figure for the case of mixtures up to 3 sugars corresponds to the usual compositional triangle for ternary systems [31], while the tetrahedron and pentachoron are the generalization of this geometric figure to 3- and 4-dimensional representations of quaternary and quinary mixtures, respectively. Compositional diagrams of high dimensionality have been very scarcely used hitherto [32–35].

For binary mixtures, the composition was calculated *via* a simple projection to the respective line segment (1-dimensional simplex) bounded by the two pure components that form the mixture. For the compositional evaluation of ternary, quaternary and quinary mixtures a dedicated algorithm has been developed. First, the PCA scores coordinates of the pure components are read, along with the PCA scores coordinates of the unknown mixture. Then, the barycentric coordinates of the unknown mixture with respect to the vertices of the corresponding simplex (PCA scores coordinates of the pure components that form a 2-, 3- or 4-dimensional simplex) are calculated. If the barycentric coordinates are all positive, it follows that the point for the unknown mixture is inside the simplex and the calculated barycentric coordinates directly correspond to the composition of the mixture. However, if at least one of the barycentric coordinates is negative, then it follows that the point for the unknown mixture is outside the simplex and it needs to be projected to the nearest edge or face of the corresponding simplex. This is done in a conditional third step of the algorithm, where the coordinates of a point inside the simplex (our initial guess is the centroid of the simplex) are numerically optimized such as to minimize the distance

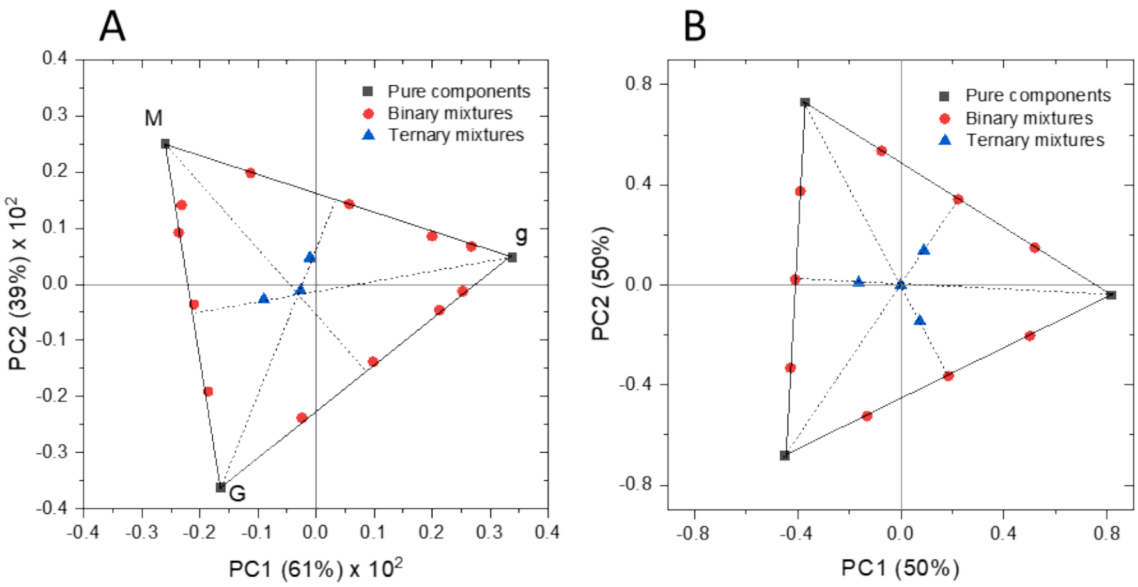


Fig. 3. A: PC2 *vs.* PC1 scores plot obtained using the PCA model developed based on the pure components and binary mixtures of 3 sugars (glucose, **G**, galactose, **g**, and mannitol, **M**). The points corresponding to the ternary mixtures were projected into the plot; B: PC2 *vs.* PC1 scores plot obtained in a PCA calculation where the exact percentual samples' composition data matrix (Table S3) was used.

between the point for the unknown mixture outside the simplex and the point inside the simplex. After the minimization is finished, the barycentric coordinates of the projected point are determined, yielding the composition of the unknown mixture. The barycentric coordinates were calculated resorting to a FORTRAN library called GEOMPACK (available at https://people.sc.fsu.edu/~jburkardt/f_src/geompack/geompack.html) which carries out tasks in computational geometry and is based on the work of Barry Joe [36]. The numerical minimizations were

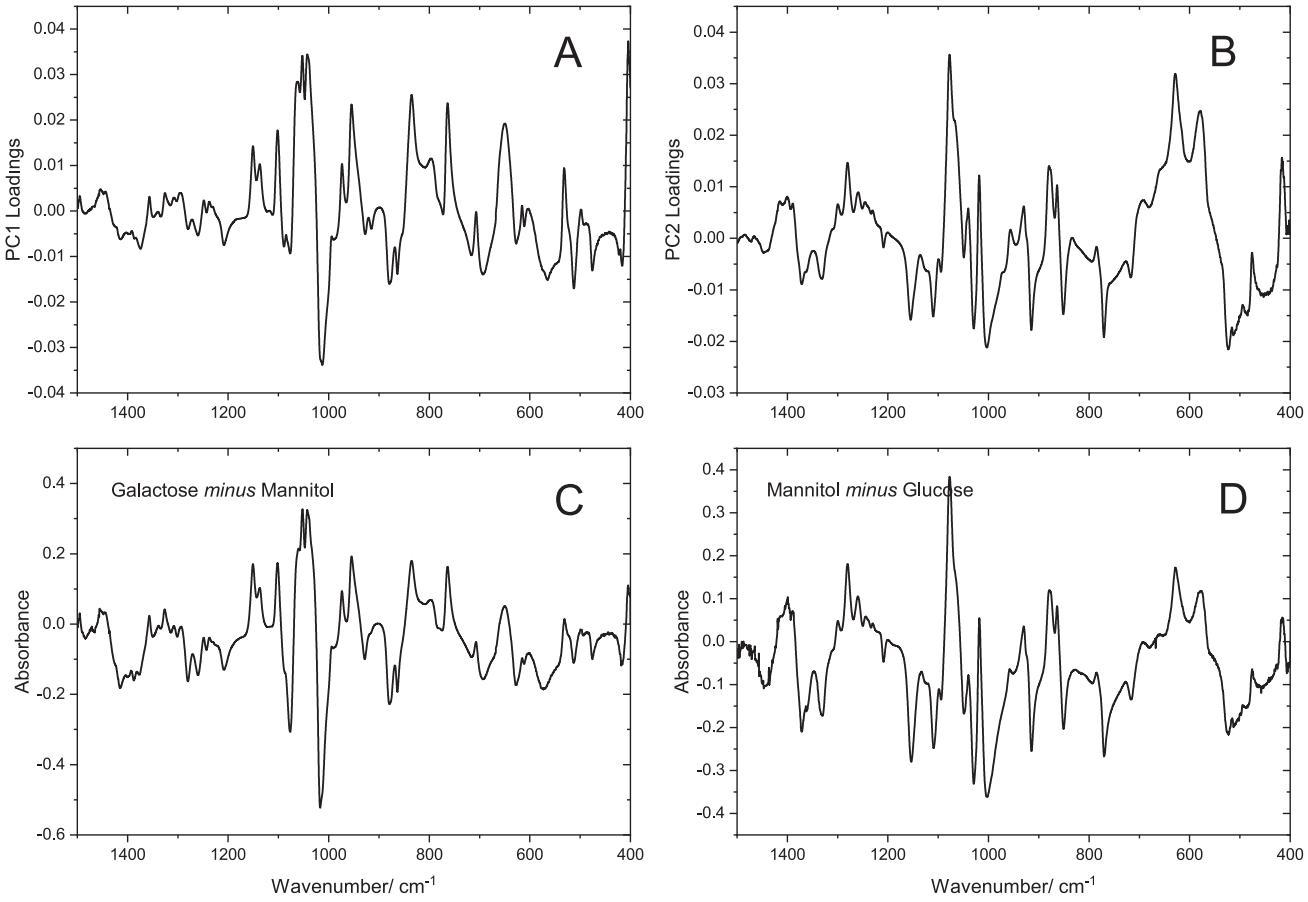


Fig. 4. A: PC1 loadings. B: PC2 loadings. C: Infrared difference spectrum **g** – **M** (galactose *minus* mannitol). D: Infrared difference spectrum **M** – **G** (mannitol *minus* glucose).

carried out with a downhill simplex method from Numerical Recipes [37], with the addition of a penalty function to penalize solutions outside the simplex.

The MCR-ALS analyses were performed under %-amount and spectra non-negativity constraints and subsequently normalized to %-amount of 100 % [25,26].

3. Results and discussion

Fig. 1 shows the ATR-FTIR spectra of the 5 investigated pure sugars, fructose, glucose, galactose, mannitol and sorbose, in the full measured spectral region (4000–400 cm^{-1}). As expected, though the spectra of the different sugars share many common features, they exhibit clearly distinctive global profiles, in particular in the 1500–400 cm^{-1} region.

Fig. 2 shows the spectra for the series of binary mixtures fructose/glucose, presented here as an example of the full set of spectra obtained for the prepared mixtures of sugars. As exemplified in Fig. 2, the spectra of the mixtures (Fig. 2A) are well reproduced by the sum of the spectra of the individual constituents in the corresponding proportions (Fig. 2B), and exhibit well-defined isosbestic points, indicating that the presence of one sugar in a mixture does not affect the spectral properties of the other component(s). These data also demonstrates that the spectral data for the studied systems exhibit linear response regarding composition, which is the type of systems for which the analytical methodology described in the next sections has been validated.

3.1. Samples with up to 3 different sugars

A set of samples with different composition were prepared using 3 selected sugars: glucose, galactose and mannitol. The sugars were combined in binary and ternary mixtures of different proportions (see Tables S1 and S2). The infrared spectra of the different samples were collected, and the spectral data (1500–400 cm^{-1} region) were subjected to PCA after pre-processing as described in Section 2.3.

Fig. 3A presents the obtained 2D PCA scores plot built using the infrared spectra (1500–400 cm^{-1}) of the prepared mixtures of the three selected sugars, together with those of the pure components. The axes correspond to the 2 most significant principal components, PC2 *vs.* PC1, which together account for 100 % of the total variance present in the spectra (with 61 % being explained by PC1, and 39 % by PC2).

Fig. 3A clearly shows that the samples of the binary mixtures appear distributed in the PC2 *vs.* PC1 scores plot in accordance with their composition along the edges of a triangle, whose vertices correspond to the samples of the pure sugars. The PCA model developed based on the infrared spectra of the pure components and their binary mixtures was then applied to project in the PC2 *vs.* PC1 scores plot the data obtained for the ternary mixtures. As expected, the equimolar ternary mixture is located nearly at the center of the triangle.

For comparison, the PC2 *vs.* PC1 scores plot of a PCA calculation where the exact percentual samples' composition data matrix was used (Table S3) is shown in Fig. 3B. The obtained triangle corresponds to the classical equilateral triangular figure used to express the composition of mixtures of up to 3 components with constant sum [31,32]. It is then clear that the PCA carried out using the infrared spectral data of the studied mixtures of sugars successfully recovered the compositional differences of the samples, extracting from the complex spectral data the appropriate information to generate the triangular polygon in the scores plot, while reducing the dimensionality of the data to only 2 relevant variables, the first two principal components, which define the directions in the space of the dataset that maximize the variance. It is important to note that the data used to make the plot shown in Fig. 3B have no redundancies (*i.e.*, the number of variables are strictly that required to determine the system), so that the algorithm assigns 50 % of variance to each one of the PCs, and the triangle is strictly equilateral. On the other hand, the spectroscopic data of the sugars are hyper-redundant. Application of the PCA algorithm thus leads to an augmented variance explanation for PC1 and results in a non-equilateral triangle, as it could be anticipated.

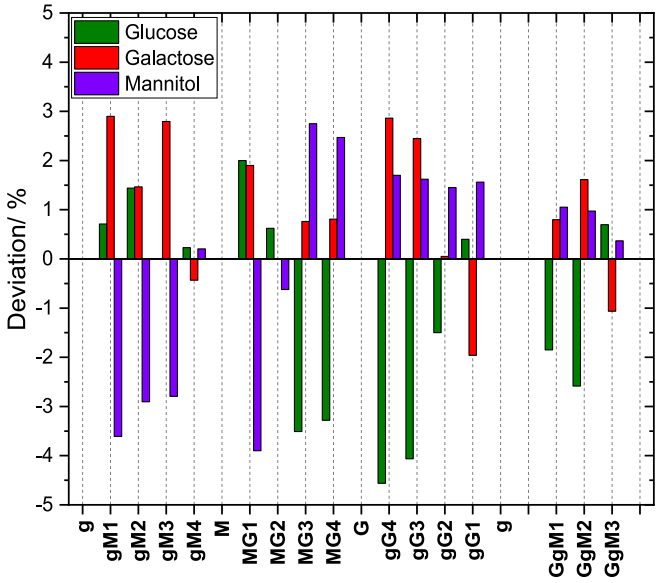


Fig. 5. Deviations between the predicted and real %-weight compositions for the mixtures of up to 3 sugars (glucose, **G**; galactose, **g**; mannitol, **M**).

The PC1 and PC2 loadings are shown in Fig. 4 (A and B, respectively). As expected considering the scores plot shown in Fig. 3A, the loadings of PC1 correlate well with the **g** – **M** (galactose *minus* mannitol) infrared difference spectrum (Fig. 4C), and the loadings of PC2 with the **M** – **G** (mannitol *minus* glucose) infrared difference spectrum.

From the positions of the points corresponding to a given sample in the PC2 *vs.* PC1 scores plot (Fig. 3A), expressed in barycentric coordinates with respect to the vertices of the triangle defined by the positions in the plot of the pure components, the compositions of the different samples in %-weight of the three sugars were obtained. Fig. 5 shows the deviations between the predicted and real compositions. The deviations, including those for the ternary mixtures (that were not used to develop the PCA model, but projected in the scores plot using the model developed based only on the pure components and their binary mixtures, as described in Section 2.3) are within the −4.6 and 2.9 % range, the average errors for glucose, galactose and mannitol being all below 2 % (1.83, 1.46 and 1.87 %, respectively), and the average total error being 1.72 %. These errors demonstrate that the developed model has a predicted ability close to quantitative.

From the practical point of view, the strategy outlined above has the limitation of requiring the knowledge about the pure components.[1] This limitation can, however, be overcome by performing a preliminary MCR-ALS analysis for identification of the sugars that are present in the samples. As shown in Fig. 6A, the application of the MCR-ALS method in a "blind" way (*i.e.*, without having any information on the spectra of the pure components), can be successfully used for identification of the components of the mixtures by comparing the obtained components spectral profiles with spectra from a suitable spectral database. On the other hand, it is unable to predict appropriately the samples compositions in terms of weight percentage of the components, with deviations relative to the real values in the range −27.1 and 21.8 % (Fig. 6B), and an average total error of 10.7 %. These errors might be useful for a qualitative analysis, but are not acceptable for a semi-quantitative one.

[1] In fact, this knowledge may not be required if the positions of the points of the samples under analysis in the scores plot allow the determination by simple geometrical considerations of the position of the (absent) sample(s) of the pure component(s).

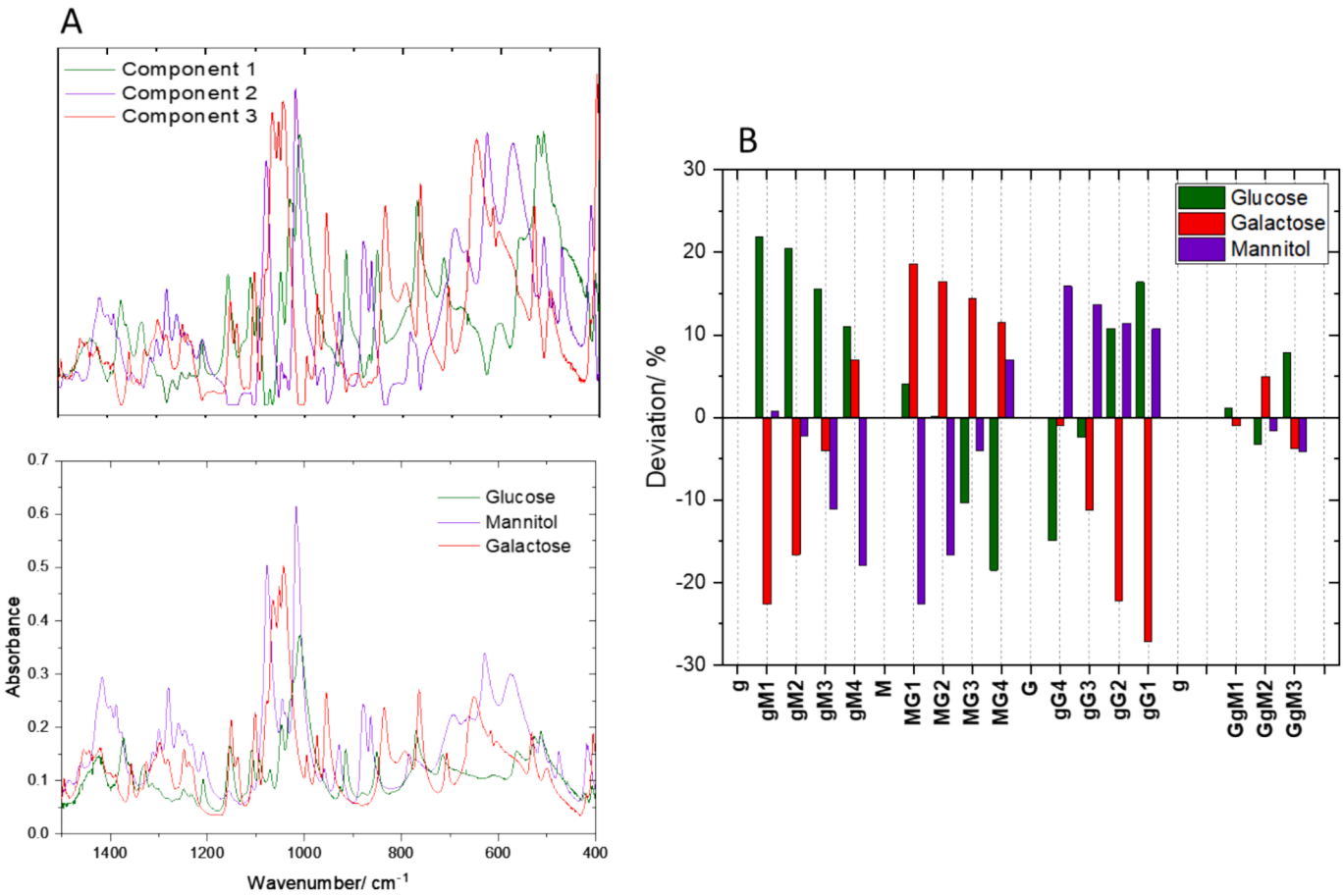


Fig. 6. A: "Blind" MCR-ALS predicted components' spectral profiles for the mixtures of up to 3 sugars (*top panel*), compared to the spectra of glucose (**G**), galactose (**g**), and mannitol (**M**) (*lower panel*). B: deviations between the "blind" MCR-ALS predicted and real %-weight compositions.

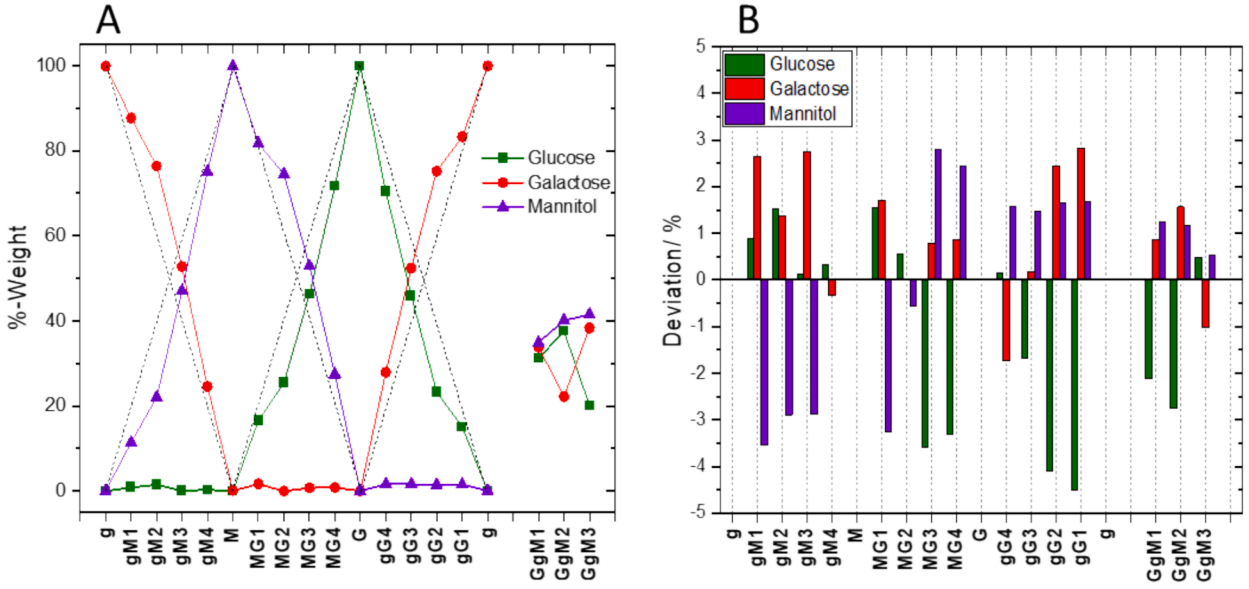


Fig. 7. A: MCR-ALS predicted %-weight compositions for the mixtures of up to 3 sugars (glucose, **G**; galactose, **g**; mannitol, **M**). B: deviations between the MCR-ALS predicted and real %-weight compositions.

MCR-ALS analysis of the data with the spectra of the pure components provided as references, on the other hand, shows almost quantitative composition predicted ability, which is similar to that obtained using the PCA scores. The results are given in Fig. 7. The deviations of the MCR-ALS predicted %-weight values in relation to the real ones vary between −4.5 and 2.8 %, the average errors for glucose, galactose and mannitol being 1.85, 1.41 and 1.85 %, respectively. The average total error is 1.70 % (1.85, 1.41 and 1.85 %, for glucose, galactose and

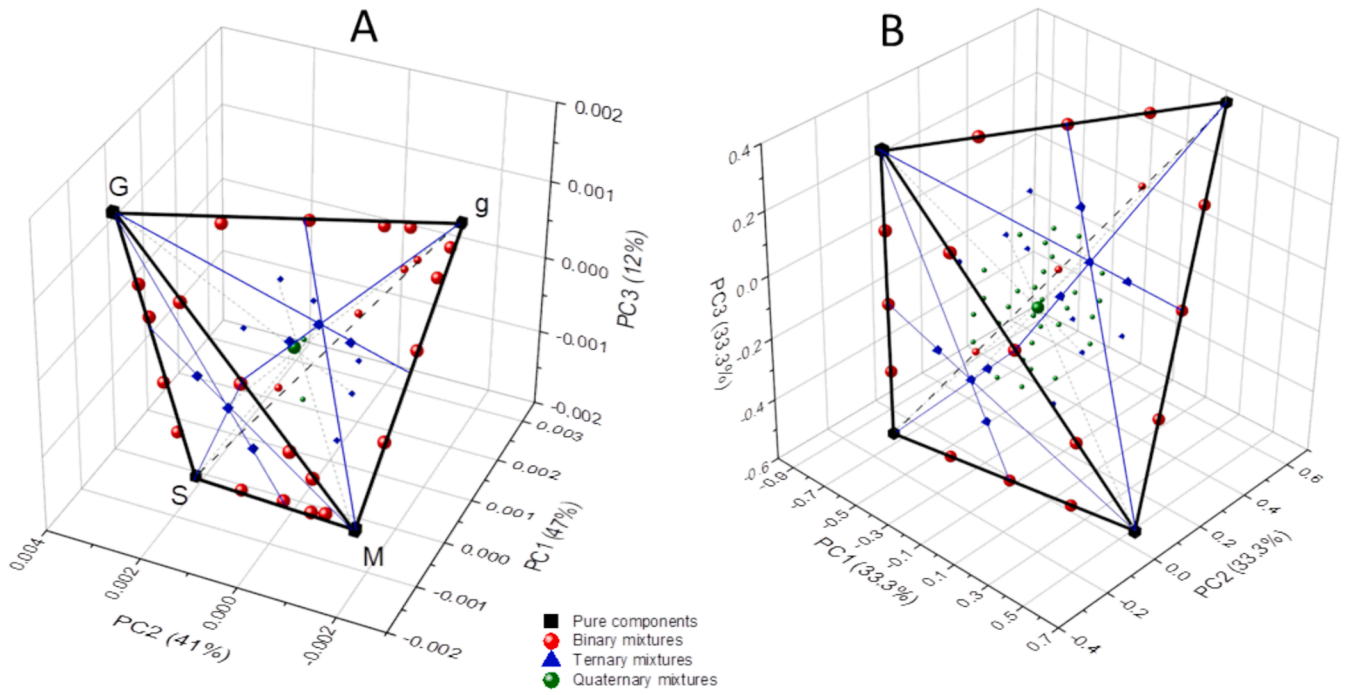


Fig. 8. A: 3D scores plot (PC3 *vs.* PC2 *vs.* PC1) obtained using the PCA model developed based on the pure components and binary mixtures of 4 sugars (glucose, **G**, galactose, **g**, mannitol, **M**, sorbose, **S**). The points corresponding to the ternary and quaternary mixtures were projected into the plot; B: PC3 *vs.* PC2 *vs.* PC1 scores plot obtained in a PCA calculation where the exact percentual samples' composition data matrix (Table S4) was used.

mannitol, respectively). All these values are identical to those obtained from the PCA scores.

3.2. Samples with up to 4 different sugars

For the investigation on samples with up to 4 different sugars, the same set of samples used in the analysis of mixtures of up to 3 sugars was used, complemented by samples having also sorbose in their composition (see Tables S1 and S2). For the PCA-based analysis, following a procedure identical to that used for the analysis of mixtures of up to 3 components, the infrared spectra (1500–400 cm^{-1} region) of the different binary mixtures of the four sugars were used to develop the PCA model, together with the spectra of the pure components. The spectral data for the ternary and the quaternary mixtures were then projected onto the PC3 *vs.* PC2 *vs.* PC1 3D scores plot of the developed model. The obtained results are shown in Fig. 8A, and they can be compared with the results obtained using an exact composition data matrix for 4 components (Table S4) that is shown in Fig. 8B.

As expected, the scores shown in Fig. 8A form a tetrahedral polygon, where the vertices correspond to the samples of the pure sugars, the binary mixtures appear distributed along the edges of the polygon according to their composition, the ternary mixtures in the faces of the tetrahedron, and the quaternary mixtures within the polygon, with the equiponderal quaternary mixture located in the center of the geometrical figure.

The results show that the PCA undertaken using the spectral data of

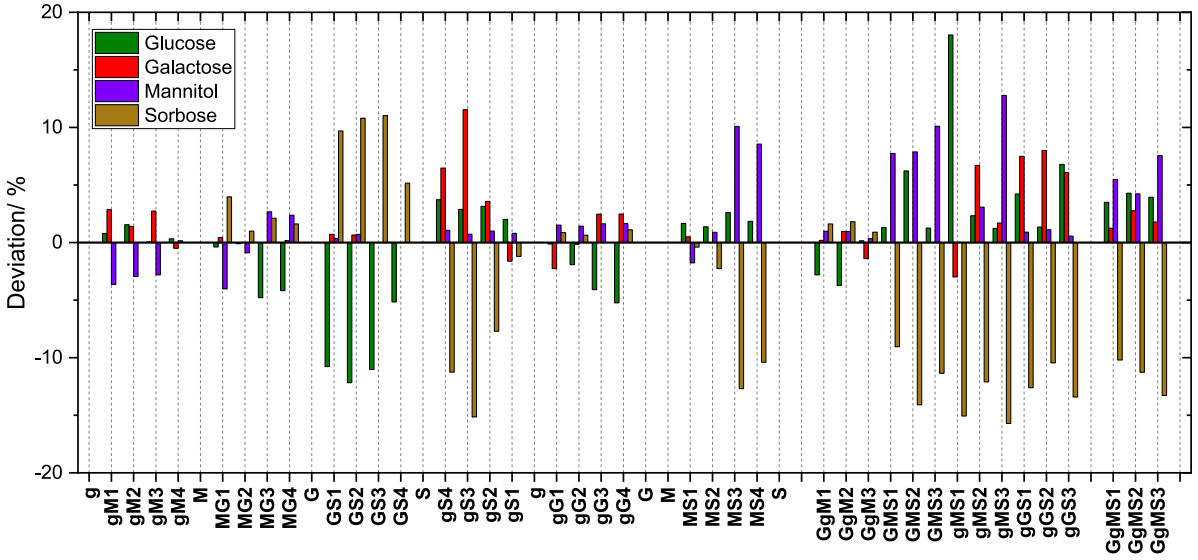


Fig. 9. Deviations between the predicted and real %-weight compositions for the mixtures of up to 4 sugars (glucose, **G**; galactose, **g**; mannitol, **M**; sorbose, **S**).

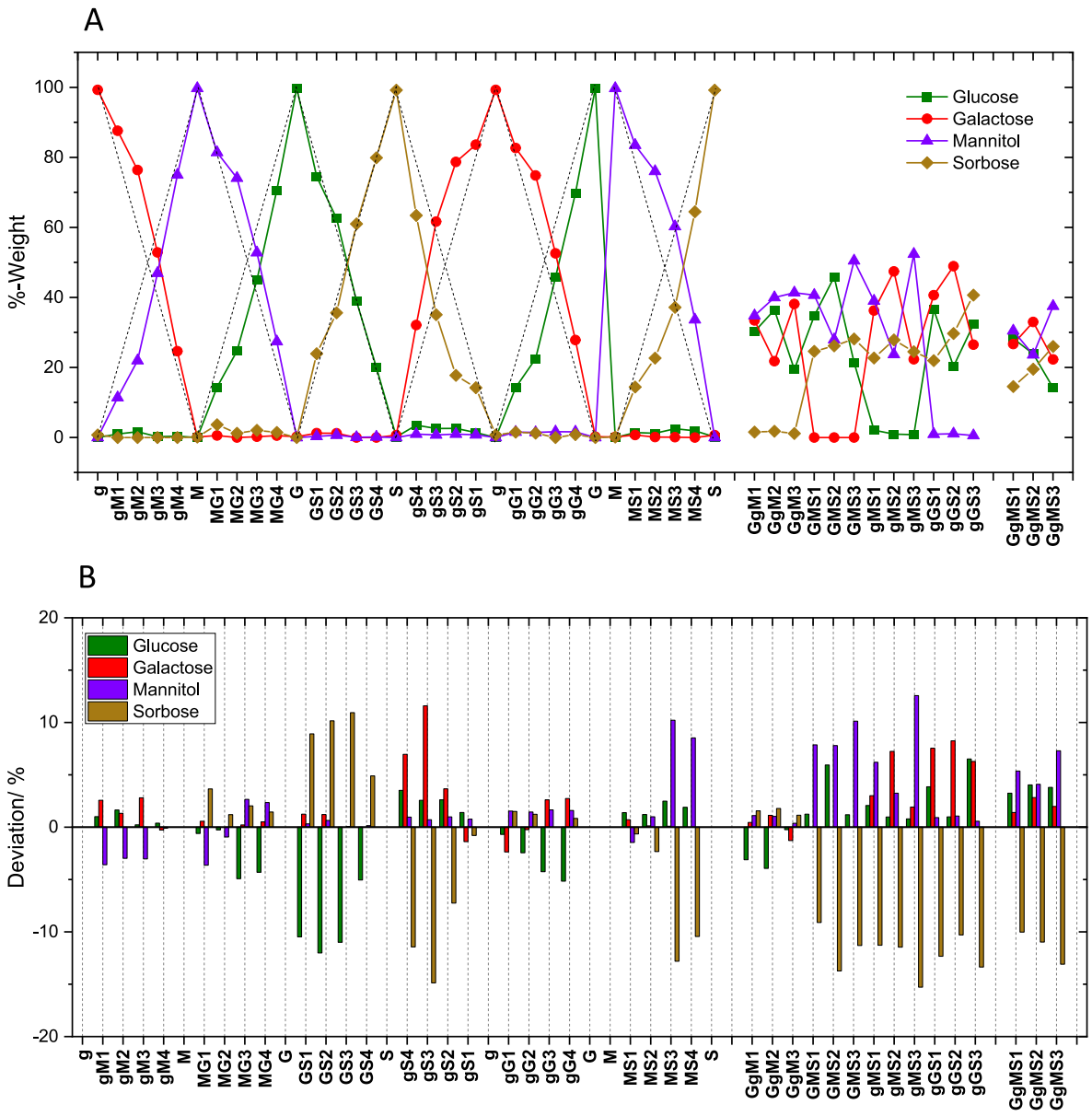


Fig. 10. A: MCR-ALS predicted %-weight compositions for the mixtures of up to 4 sugars (glucose, **G**; galactose, **g**; mannitol, **M**; sorbose, **S**). B: deviations between the MCR-ALS predicted and real %-weight compositions.

the mixtures of the 4 sugars, like for the case of 3 components, describe properly the variability in the composition of the samples, while successfully reducing the highly-dimensional data to the 3 relevant composition-dependent variables. Also as noticed above for the case of mixtures of 3 components, introduction of hyper-redundancy in the data leads to both unequal variance explanation by the 3 PCs (PC1: 47 %, PC2: 41 %; PC3: 12 %) and distortion of the tetrahedron, both properties being strictly obeyed when the variables non-redundant exact composition data matrix is used (see Fig. 8B).

The PC1, PC2 and PC3 loadings are shown in Fig. S1 (Supplementary Material), where it can be seen that they correspond mostly to the **g** – **M** (galactose *minus* mannitol), **G** – **M** (glucose *minus* mannitol), and **G** – **S** (glucose *minus* sorbose) infrared difference spectra, respectively.

From the PC3 *vs.* PC2 *vs.* PC1 scores (Fig. 8A), used as barycentric coordinates with respect to the vertices of the tetrahedron defined by the positions in the scores plot of the pure components, the compositions of the samples of the mixtures of the four sugars (in %-weight) were estimated.

The deviations between the predicted and real compositions of the samples are shown in Fig. 9, where it can be seen that the deviations are within the −15.1 and 18.0 % interval. Average errors for glucose, galactose, mannitol and sorbose are 3.67, 2.10, 2.96 and 6.72 %, respectively, the average total error being 3.86 %. These errors demonstrate that the developed model has a very good semi-quantitative prediction ability (see Fig. 10).

When applied to the spectral data, "blind" MCR-ALS yielded four components, that compare well with the four sugars, as shown in Fig. S2. When the spectra of the individual sugars are provided as data, the MCR-ALS method recovered the samples compositions at semi-quantitative level, with deviations in the predicted %-weight values in relation to the real ones varying between –15.3 and 12.6 %, the average errors for glucose, galactose, mannitol and sorbose being 3.16, 2.18, 3.10 and 6.15 %, respectively, and the average total error 3.75 %. As for the case of mixtures of up to 3 sugars, these values are similar to those obtained from the PCA scores.

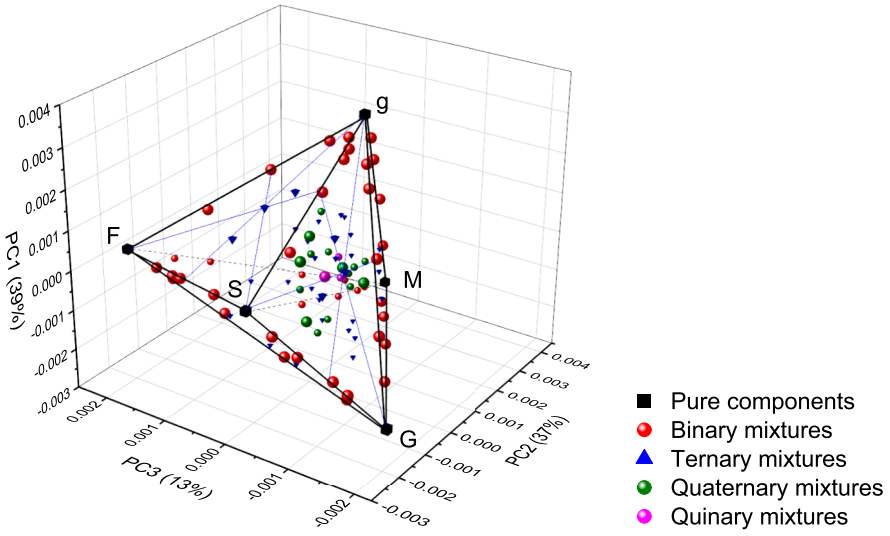


Fig. 11. 3D scores plot (PC3 *vs.* PC2 *vs.* PC1) obtained using the PCA model developed based on the pure components and binary mixtures of 5 sugars (glucose, **G**, galactose, **g**, mannitol, **M**, sorbose, **S**; fructose, **F**). The points corresponding to the ternary, quaternary and quinary mixtures were projected.

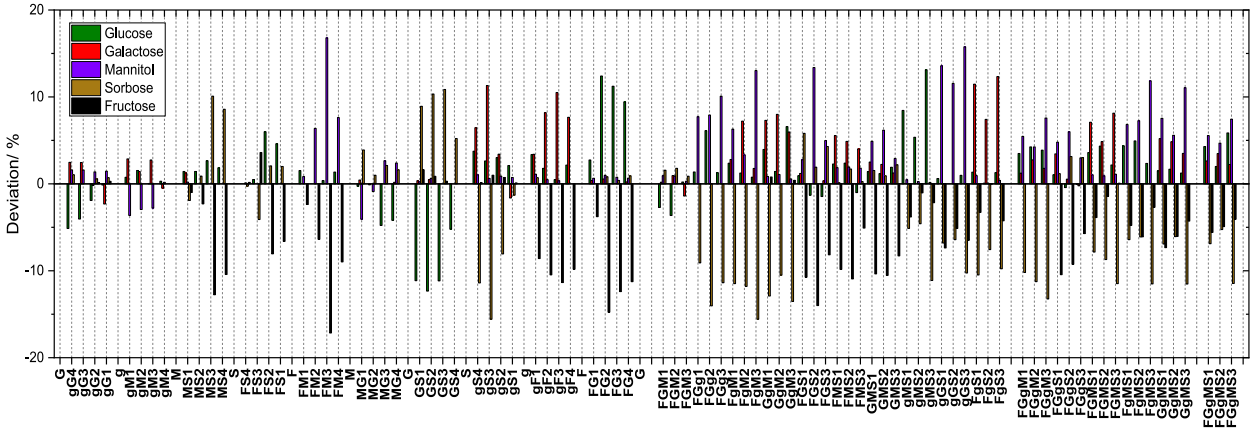


Fig. 12. Deviations between the predicted and real %-weight compositions for the mixtures of up to 5 sugars (glucose, **G**; galactose, **g**; mannitol, **M**; sorbose, **S**; fructose, **F**).

3.3. Samples with up to 5 sugars

In this section, the generalization of the approaches presented in the two previous sections to mixtures of up to 5 sugars is presented. The use data was expanded to include also mixtures containing fructose (see Tables S1 and S2).

As stated in Section 2.3, the composition of mixtures of up to 5 components can be represented by using a pentachoron (4D simplex), the equivalent geometrical figure of the tetrahedron in a 4D-space. The composition of a given mixture is expressed by the barycentric co-ordinates in the pentachoron whose vertices correspond to the pure components. Following the same approach used above to deal with mixtures of up to 3 and up to 4 sugars, the PC1, PC2, PC3 and PC4 scores resulting from a PCA calculation made based on the infrared spectra of the five pure sugars and their binary mixtures were made to correspond to the barycentric coordinates yielding the estimations of the %-weight composition of the mixtures (including the ternary, quaternary and quinary mixtures, which were subsequently projected in the PC4 *vs.* PC3 *vs.* PC2 *vs.* PC1 4D-space of the developed PCA model).

The 4D scores plot cannot be graphically depicted, but the data can still be visualized through different 3D projections [30,38], which result from using different combinations of the 4 relevant PCs. One example of these 3D graphs is shown in Fig. 11 (PC3 *vs.* PC2 *vs.* PC1), where it is visible that, as for the lower dimensionality problems considered in the previous sections, the positions of the points corresponding to the different types of mixtures in the geometric figure can be directly related with the compositions of the mixtures.

The percentages of variance explained by PC1, PC2, PC3 and PC4 amount to 39, 37, 13 and 10 %, respectively, *i.e.*, fully explain the variance in the samples and showing that the algorithm is robust in dealing with hyper-redundancy in the data.

The deviations between the predicted and real compositions of the samples are shown in Fig. 12, where it can be seen that they stay within the −17.2 to 16.8 % range. The average errors for glucose, galactose, mannitol, sorbose and fructose are 2.60, 1.87, 2.75, 4.10 and 3.01 %, respectively, and the average total error 2.87 %, demonstrating that the developed model has, also in this case, a very good semi-quantitative prediction ability.

The spectral profile of the five components predicated by the "blind" MCR-ALS are shown in Fig. S3, showing a good concordance with the spectra of the pure sugars, while the composition predictions of this method when the spectra of the individual sugars are given as references are displayed in Fig. 13, together with the corresponding deviations relative to the real values. Also in this case, the MCR-ALS method was

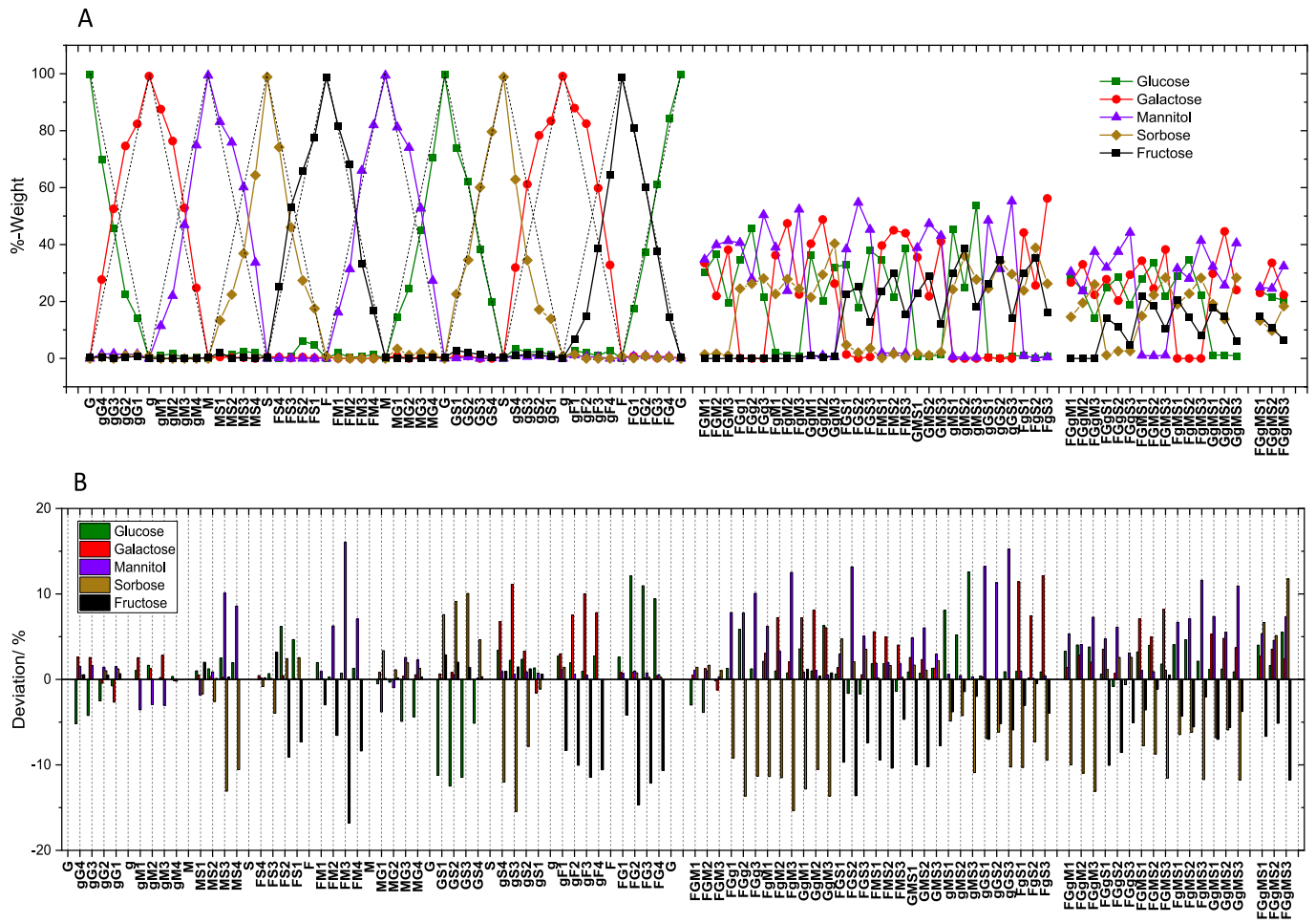


Fig. 13. A: MCR-ALS predicted %-weight compositions for the mixtures of up to 5 sugars (glucose, **G**; galactose, **g**; mannitol, **M**; sorbose, **S**; fructose, **F**). B: deviations between the MCR-ALS predicted and real %-weight compositions.

able to predict the samples compositions at semi-quantitative level and with errors similar to those obtained from the PCA scores. The deviations vary between −16.8 and 16.1 %, the average errors for glucose, galactose, mannitol, sorbose and fructose are 2.47, 2.04, 2.75, 4.15 and 3.18 %, respectively, and the average total error is 2.94 %.

4. Conclusion

In this study, chemometric models based on infrared spectral data were developed that are able to predict, at the semi-quantitative level, the composition (%-weight) of mixtures of up to 5 sugars. Besides application of the MCR-ALS method, which was shown to only be able to attain a semi-quantitative level when the spectral profiles of the pure components are provided as reference (though being able to attain a qualitative analytic level in the absence of these), a new method based on the scores of the most representative PCs was developed. In this method, the scores are used as barycentric coordinates relative to the simplex of the proper order (triangle, tetrahedron, tetrachoron, for 3, 4 and 5 maximum sugars in the mixtures, respectively) defined by the positions in the score plots of the points corresponding to the pure sugars, thus allowing direct determination of the compositions. The developed method has been shown to be able to appropriately extract the compositional information from the complex spectral data and the algorithm to be robust in dealing with hyper-redundancy in the data. In relation to the MCR-ALS method, the PCA-based approach has the advantage of not necessarily needing previous information on the pure components, if the positions of the points of the samples under analysis in the relevant scores plot allow the determination by simple geometrical considerations of the position of the (absent) sample(s) of the pure component(s). The method is also able to efficiently handle interferents, since their effects will be described by variance expressed in higher-order principal components. Moreover, though the ability of the PCA scores-based method to represent data for graphical visual direct analysis is, obviously conditioned by the impossibility to represent the data in more than 3 dimensions, its ability to represent the data for any dimension is excellent in numeric terms, providing simple orthogonal coordinates for each sample in the proper vectorial space.

The approach developed in this investigation was applied successfully to the specific problem of determining the composition of mixtures of up to 5 sugars, as a suitable example of application to a relevant system. However, it should be seen as a general procedure for the semi-quantitative analysis of other types of mixtures and applicable to other types of data reflecting linearly their composition. In addition, considering the observed robustness of the method upon increasing dimensionality from 3 to 5 components, it might be expected that it can also be used successfully for higher dimensions.

Considering that in the cases studied here some differences in the predictive ability regarding the different substances are apparent (for example, sorbose, mannitol and fructose %-weight predicted values show a general trend to be somewhat larger than those obtained for glucose and galactose), it can also be concluded that the precise nature of the data is relevant to determine the performance of the methods, a high degree of similarity between the profiles of the components appearing as a potential limitation for the method, at least when the

problem is of high dimensionality.

For the specific problem addressed in the present article, it shall be highlighted that the PCA/IR-based developed strategy appears of great practical interest. The method requires no sample preparation (or minimum preparation), is fast and cost-effective (in particular when compared to traditional chemical or chromatographic methods), the chemometric analysis is very simple (PCA is the most used unsupervised chemometric method and is available in many software packages), and infrared spectroscopy is a very sensitive technique to structure and composition. Moreover, the fact that the method is robust in relation to the consideration of hyper-redundancy and able to extract from redundant data the requires compositional information in an unsupervised way is of particular relevance, since it allows the user not to be concerned in any way of *a priori* identifying the relevant variables to perform the analysis.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: ANNA LUIZA BIZERRA DE BRITO, RUI FAUSTO MARTINS RIBEIRO DA SILVA LOURENCO, LUIS PEDRO DA FRANCA E SILVA CALEIRAS VIEGAS has patent Chemometrics Protocol for Semi-Quantitative Analysis of Mixtures Using Spectral Data or Other Composition Dependent Hyper-Redundant Data pending to 119690. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

The CQC-IMS is supported by FCT through projects UI0313B/QUI/2020, UI0313P/QUI/2020 and LA/P/0056/2020 (national funds). The authors also acknowledge the Laboratory for Advanced Computing at University of Coimbra (LCA-UC; https://www.uc.pt/lca) for providing computing resources, and Coimbra LaserLab (CLL) for experimental facilities.

Authors contributions

Anna Luiza B. Brito: experimental planning, experiments, spectroscopic data analysis, writing of the preliminary version of the manuscript; Inês Cardoso: experiments; Luís Pedro Viegas: data analysis, programing; Rui Fausto: conceptualization, funding acquisition, data analysis, writing of the preliminary version of the manuscript. All authors contributed to the final written version of the manuscript.

Appendix A. Supplementary material

Table S1, with the composition (%-weight) of the prepared sugar mixtures; Table S2, with the list of the samples used in the different experiments; Tables S3 and S4, with the simple data matrix with exact samples' compositions for mixtures of 3 and 4 components, respectively; Fig. S1, with the PC1 and PC2 loadings for the PCA model developed for mixtures to up 4 sugars, and IR difference spectra correlated with the loadings' profiles; Figs. S2 and S3, with the "blind" MCR-ALS predicted components' profiles for the mixtures of up to 4 sugars and up to 5 sugars, respectively. Supplementary data to this article can be found online at https://doi.org/10.1016/j.saa.2024.125225.

References

[1] V.H. Pomin, P.A.S. Mourão, Carboidratos, Ciência Hoje 39 (2006) 1.
[2] J.E. Holesh, S. Aslam, A. Martin, Physiology, carbohydrates, in: StatPearls [Internet], StatPearls Publishing, Treasure Island (FL), 2023, Available from: https://www.ncbi.nlm.nih.gov/books/NBK459280/.
[3] Z. Farooq, IsA. Successful sugar identification with ATR-FTIR, Agro Food Ind. Hi-Tech 25 (2014) 36.
[4] G. Livesey, Health potential of polyols as sugar replacers, with emphasis on low glycaemic properties, Nutr. Res. Rev. 16 (2003) 163.
[5] X. Qi, R.F. Tester, Fructose, galactose and glucose - in health and disease, Clin. Nutr. ESPEN 33 (2019) 18.
[6] M. Kurzyna-Szklarek, J. Cybulska, A. Zdunek, Analysis of the chemical composition of natural carbohydrates – an overview of methods, Food Chem. 394 (2022) 133466.
[7] M. Kačuráková, R.H. Wilson, Developments in mid-infrared FT-IR spectroscopy of selected carbohydrates, Carbohyd. Pol. 44 (2001) 291.
[8] L. Leopold, N. Leopold, H. Diehl, C. Socaciu, Quantification of carbohydrates in fruit juices using FTIR spectroscopy and multivariate analysis, Spectroscopy 26 (2011) 93.
[9] M.F. Mrozek, D. Zhang, D. Ben-Amotz, Oligosaccharide identification and mixture quantification using Raman spectroscopy and chemometric analysis, Carbohydr. Res. 339 (2004) 141.
[10] E. Wiercigroch, E. Szafraniec, K. Czamara, M.Z. Pacia, K. Majzner, K. Kochan, A. Kaczor, M. Baranska, K. Malek, Raman, and infrared spectroscopy of carbohydrates: a review, Spectrochim. Acta A Mol. Biomol. Spectrosc. 185 (2017) 317.
[11] M. Meyer, L. Montero, S.W. Meckelmann, O.J. Schmitz, Comparative study for analysis of carbohydrates in biological samples, Anal. Bioanal. Chem. 414 (2022) 2117.
[12] J.N. BeMiller, Carbohydrate analysis. Food Analysis. Food Analysis, Springer, Boston, MA, 2010.
[13] M.F. Chaplin, J.F. Kennedy (Eds.), Carbohydrate Analysis: A Practical Approach, second edition, Oxford University Press, London, New York, 1994.
[14] Z. El Rassi, Carbohydrate Analysis by Modern Chromatography and Electrophoresis, second edition, Elsevier Health Sciences, Amsterdam, The Netherlands, 2002.
[15] C. Heiss, P. Azadi, Carbohydrate structure analysis: methods and applications, SLAS Technol.: Transl. Life Sci. Innov. 25 (2020) 305.
[16] F.J. Moreno, M.L. Sanz (Eds.), Food Oligosaccharides: Production, Analysis and Bioactivity, John Wiley & Sons, Ltd., Chicago, IL, USA, 2014.
[17] H. Martens, T. Naes, Multivariate Calibration, Wiley, New York, NY, USA, 1989.
[18] A.C. Olivieri, Introduction to Multivariate Calibration: A Practical Approach, second edition, Springer Intl. Publishers, Cham, Switzerland, 2024.
[19] K. Pearson, On lines and planes of closest fit to systems of points in space, Phil. Magazine. 2 (1901) 559.
[20] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (1933) 417.
[21] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (1933) 498.
[22] H. Hotelling, Relations between two sets of variates, Biometrika 28 (1936) 321.
[23] A. Takamura, K. Watanabe, T. Akutsu, T. Ozawa, Soft and robust identification of body fluid using Fourier Transform infrared spectroscopy and chemometric strategies for forensic analysis, Sci. Rep. 8 (2018) 8459.
[24] G.O. Ildiz, S. Bayari, N. Yorguner, R. Fausto, Blood serum–infrared spectra-based chemometric models for auxiliary diagnosis of autism spectrum disorder, in: A. S. El-Baz, J.S. Suri (Eds.), Neural Engineering Techniques for Autism Spectrum Disorder, Academic Press, 2021, pp. 185–213. Chapter 10.
[25] A. de Juan, J. Jaumot, R. Tauler, Multivariate Curve Resolution (MCR). Solving the mixture analysis problem, Anal. Methods 6 (2014) 4964.
[26] P.H. Marco, P. Valderrama, G.L. Alexandrino, R.J. Poppi, T. Romá, Multivariate Curve Resolution with Alternating Least Squares: Description, operation and applications, Quím. Nova 37 (2014) 1525.
[27] The Unscrambler™ Version 10.5.1, CAMO Software Inc., CAMO A/S, Trondheim, Norway, 2018.
[28] Origin Pro (version 2021), OriginLab Corporation, Northampton, MA, USA.
[29] J. Dongarra, M. Gates, A. Haidar, J. Kurzak, P. Luszczek, S. Tomov, I. Yamazaki, The singular value decomposition: anatomy of optimizing an algorithm for extreme scale, Soc. Industr. Appl. Math. Rev. 60 (2018) 808.
[30] A. Kageyanna, A visualization method of four-dimensional polytopes by oval display of parallel hyperplane, J. Visualiz. 19 (2016) 417.
[31] D. West, Ternary Equilibrium Diagrams, second edition, Springer, New York, 2013.
[32] N. Wei, The 3-dimensional phase diagram in quaternary systems of polymers and solvents, J. Appl. Pol. Sci. 28 (1983) 2755.
[33] T.M. Konyakhina, J. Wu, J.D. Mastroianni, F.A. Heberle, G.W. Feigenson, Phase diagram of a 4-component lipid mixture: DSPC/DOPC/POPC/Chol, Biochim. Biophys. Acta: Biomembr. 1828 (2013) 2204.
[34] T. Shimura, A.I.S. Kemp, Tetrahedral plot diagram: a geometrical solution for quaternary systems, J. Am. Miner. 100 (2015) 2545.
[35] L. Banko, O.A. Krysiak, B. Xiao, T. Löffler, A. Savan, J.K. Pedersen, J. Rossmeisl, W. Schuhmann, A. Ludwig, Unravelling composition–activity–stability trends in high entropy alloy electrocatalysts by using a data-guided combinatorial synthesis strategy and computational modeling, Adv. Ener. Mater. 12 (2022) 2103312.

[36] B. Joe, GEOMPACK - a Software package for the generation of meshes using geometric algorithms, Adv. Eng. Softw. 13 (1991) 325.
[37] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in Fortran 77: The Art of Scientific Computing, second edition, Cambridge University Press, New York, USA, 1992.
[38] A.J. Hanson, Geometry for N-dimensional graphics, in: P. Heckbert (Ed.), "Graphics Gems IV", Academic Press, Cambridge, USA, 1994, Chapter II.6, pp. 149.170.