



Review

Application of multivariate data analysis for food quality investigations: An example-based review

Carolien Buvé^a, Wouter Saeys^b, Morten Arendt Rasmussen^{c,d}, Bram Neckebroek^a,
Marc Hendrickx^a, Tara Grauwet^a, Ann Van Loey^{a,*}

^a KU Leuven Department of Microbial and Molecular Systems, Laboratory of Food Technology, Kasteelpark Arenberg 22 Box 2457, 3001 Leuven, Belgium

^b KU Leuven Department of Biosystems, MeBioS division, Kasteelpark Arenberg 30, 3001 Leuven, Belgium

^c University of Copenhagen, Department of Food Science, Faculty of Science, Rolighedsvej 26, 1958 Frederiksberg, Denmark

^d COPSAC, Copenhagen Prospective Studies on Asthma in Childhood, Herlev and Gentofte Hospital, University of Copenhagen, 2100 Copenhagen, Denmark

ARTICLE INFO

Keywords:

Example-based review
Food quality assessment
Multivariate data analysis
Large data sets
Quality changes
Chemometrics
Omics approaches

ABSTRACT

These days, large multivariate data sets are common in the food research area. This is not surprising as food quality, which is important for consumers, and its changes are the result of a complex interplay of multiple compounds and reactions. In order to comprehensively extract information from these data sets, proper data analysis tools should be applied. The application of multivariate data analysis (MVDA) is therefore highly recommended. However, at present the use of MVDA for food quality investigations is not yet fully explored. This paper focusses on a number of MVDA methods (PCA (Principal Component Analysis), PLS (Partial Least Squares Regression), PARAFAC (Parallel Factor Analysis) and ASCA (ANOVA Simultaneous Component Analysis)) useful for food quality investigations. The terminology, main steps and the theoretical basis of each method will be explained. As this is an example-based review, each method was applied on the same experimental data set to give the reader an idea about each selected MVDA method and to make a comparison between the outcomes. Numerous MVDA methods are available in literature. Which method to select depends on the data set and objective. PCA should be the first choice for data exploration of two-dimensional data. For predictive purposes, PLS is the most appropriate method. Given an underlying experimental design, ASCA takes into account both the relation between the different variables and the design factors. In case of a multi-way data set, PARAFAC can be used for data exploration. While these methods have already proven their value in research, there is a need to further explore their potential to investigate the complex interplay of compounds and reactions contributing to food quality. With this work we would like to encourage food scientists with no or limited knowledge of MVDA to get some first insights into the selected methods.

1. Introductory example: are food quality data sets really complex?

The example of strawberry juice volatile changes during storage

To motivate this *example-based review* we start with an example to show the relevance of multivariate data analysis (MVDA) in the area of food science. These data have already been published partially by the

authors of this paper and can be consulted in the work of Buvé et al. (2018) (with permission of Elsevier).

Fruit juices are very popular among consumers *thanks* to their organoleptic properties. Unfortunately, they are prone to quality changes (e.g. colour and flavour changes) during storage. To gain insight into these changes, Buvé et al. (2018) stored pasteurised strawberry juice at 20 °C (32 weeks), 28 °C (32 weeks), 35 °C (12 weeks) and 42 °C (8 weeks) in PET bottles and collected samples 12 times during storage

Abbreviations: ANOVA, Analysis of Variance; ASCA, ANOVA Simultaneous Component Analysis; GC-MS, Gas Chromatography – Mass Spectrometry; LC-TOF-MS, Liquid Chromatography - Time of Flight – Mass Spectrometry; LV(s), Latent Variable(s); MANOVA, Multivariate Analysis of Variance; MVDA, Multivariate data analysis; PARAFAC, Parallel Factor Analysis; PC(s), Principal Component(s); PCA, Principal Component Analysis; PLS, Partial Least Squares Regression; PLS-DA, Partial Least Squares-Discriminant Analysis; RMSECV, Root Mean Squared Error of Cross Validation; SCA, Simultaneous Component Analysis; VID, Variable Identification; VIP, Variable Importance in Projection.

* Corresponding author.

E-mail address: ann.vanloey@kuleuven.be (A. Van Loey).

<https://doi.org/10.1016/j.foodres.2021.110878>

Received 27 July 2021; Received in revised form 29 November 2021; Accepted 4 December 2021

Available online 8 December 2021

0963-9969/© 2021 Elsevier Ltd. All rights reserved.

at each temperature. The designed experiment allowed to study volatile changes at ambient conditions (20 °C) as well as to which extent storage at higher temperatures (28–42 °C) accelerated these changes compared to normal (i.e. ambient) storage conditions. An untargeted GC-MS fingerprinting approach was used to evaluate volatile changes (Grauwet, Vervoort, Colle, Van Loey, & Hendrickx, 2014). This approach allowed to detect a wide range of *a priori* unknown strawberry juice volatiles (± 130) in one run. To give the reader an idea about the complexity of the volatile profile of strawberry juice and the changes during storage at 20 °C, representative chromatograms of a non-stored and stored sample, together with an indication of key volatile compounds, can be found in [Supplement 1](#).

Without a doubt, relevant information linked to quality changes during storage could be obtained when evaluating the behaviour of all 130 volatile compounds individually. However, this would be a time consuming task for a food scientist with a risk of overlooking important changes as well as loss of information which is captured as patterns of correlations between the volatiles. Moreover, when a high number of samples is analysed, the task becomes even more complex. As an example, each of the strawberry juice samples (i.e. stored at a certain temperature for a certain time) was analysed six times independently with the GC-MS fingerprinting method. As a result, a data table with 72 samples (i.e. 12 storage times $\times$ 6 repetitions; the rows in the table) and 130 volatiles (the columns in the table) was obtained for each storage temperature. Moreover, given the specific experimental design, the GC-MS data could be arranged in a three-dimensional array (*storage times $\times$ volatiles $\times$ storage temperatures*), which makes it even more complex as one can look at the evolution as a function of storage time and temperature. Extracting relevant and conclusive findings from such information rich data sets is a pertinent challenge for food scientists, in which the application of multivariate methods, the main topic of this paper, can be a crucial step.

2. Introduction to multivariate data analysis (MVDA) and its potential to study food quality changes

Traditionally, food scientists were focussing on a single food quality attribute (e.g. texture, enzyme activity, colour). Such a univariate approach is the most straightforward strategy for food quality assessment and is definitely of great value. However, this strategy ignores the complexity of a food product and the possible interactions between the different attributes. The complexity of a food product is reflected in the definition of food quality as given by The Knowledge Centre for Food Fraud and Quality. They describe food quality as “a complex and multi-dimensional concept which is influenced by a wide range of situational and contextual factors”. The influencing factors include among others: safety, origin, nutrition, sensorial properties, authenticity and convenience (Knowledge Centre for Food Fraud and Quality, 2021). Food quality changes can thus be described as those changes that are impacting the nutritional or sensorial properties or the authenticity of a food product. Due to the complexity of food products, food scientists are more than ever focussing on the simultaneous analysis of multiple quality attributes (‘multivariate’ and ‘omics’ approaches). This trend has been accelerated by the development of high-throughput analytical equipment (GC-MS, LC-TOF-MS, etc.) producing large data sets. The use of MVDA approaches allows the researcher to obtain an integrated view on food quality and possibly a deeper mechanistic understanding. Moreover, the use of multivariate methods to extract relevant information out of such data sets seems to be a logical step to go along with the analytical progress (Bro, 1997; Cevallos-Cevallos, Reyes-De-Corcuera, Etxeberria, Danyluk, & Rodrick, 2009; Grauwet et al., 2014; Guo, Wu, Massart, Boucon, & de Jong, 2002; Khakimov, Gürdeniz, & Engelsen, 2015; Lenhardt, Zeković, Dramićanin, & Dramićanin, 2015; Mehmood, Sæbø, & Liland, 2020; Roberts & Cozzolino, 2016).

MVDA or chemometrics can be defined as the discipline that uses mathematical and statistical methods to extract information from large

data sets with chemical or biological information. The concept to extract information from chemical systems by means of multivariate statistics was already investigated in the 1960s (Brereton, 2014). Consequently, the use of multivariate methods is by far the most popular in the area of analytical/applied chemistry (Brereton et al., 2017, 2018; Kumar, Bansal, Sarma, & Rawal, 2014). Later on, MVDA has been introduced step-by-step in different research areas including the field of food science and postharvest technology (Nunes, Ortiz Alvarenga, de Souza Sant’Ana, Sousa Santos, & Granato, 2015; Roberts & Cozzolino, 2016; Roussel, Preys, Chauchard, & Lallemand, 2014; Saeys, Nguyen Do Trong, Van Beers, & Nicolai, 2019). However, the application of MVDA in the area of food quality investigations has been explored much less, despite its potential to extract relevant and new information out of large data sets. The major reason can be addressed to the fact that many food scientists find it difficult to understand the statistical and mathematical principles underlying these analyses thereby possessing a hurdle for a wider and general utility (Granato, de Araújo Calado, & Jarvis, 2014; Nunes et al., 2015). Consequently, there is a need for documentation aiming to give food scientists the basic knowledge of the principles as well as specific application driven ideas to show its potential. In this context, it is noteworthy that several authors (e.g. Bevilacqua et al., 2017; Granato et al., 2018; Khakimov et al., 2015; Roberts & Cozzolino, 2016; Skov, Honoré, Jensen, Næs, & Engelsen, 2014) already tried to make MVDA and other statistical/mathematical methods more accessible for food scientists. Rather than focussing on the theory of these methods, this paper aims to show the applicability of MVDA for food quality investigations to encourage food scientists to explore such techniques when dealing with multivariate data. As a result, this paper will focus on the practical aspects of a MVDA and can be used as a guide to become familiar with the concepts. Linked to the practical aspects of MVDA, well-known softwares for MVDA are PLS toolbox and Solo (Eigenvector) and Matlab (The MathWorks inc.).

MVDA covers a wide range of data analysis principles and data types. Linked to the diversity of the challenges originating from the analytical procedures as well as the research questions, these methods can be used for different purposes such as data exploration (Bro & Smilde, 2014), prediction and classification (Abdi, 2003; Saeys et al., 2019; Wold, Eriksson, Trygg, & Kettaneh, 2004) or inference (Jansen et al., 2005). In addition, MVDA is an active field of research in which methods are being developed or improved frequently. However, this review will specifically focus on a particular number of methods being Principal Component Analysis (PCA), Partial Least Squares (PLS) regression, Parallel Factor Analysis (PARAFAC) and ANOVA Simultaneous Component Analysis (ASCA). These data analysis techniques were selected based on the number of publications in which they were used, their applicability to a wide range of common problems and hence their potential for food quality investigations. For details of additional methods (e.g. PCR, iPLS, TUCKER, MCR, SIMCA), the reader is referred to specialised literature (e.g. Amigo, Skov, Bro, Coello, & Maspocho, 2008; Jansen et al., 2008; Nørgaard et al., 2000; Saccenti, Smilde, & Camacho, 2018).

This paper will start with the basic terminology (Section 3) and a brief overview of the main steps (Section 4) of MVDA. In Section 5, the theoretical background of each selected method (PCA, PARAFAC, PLS, ASCA) will be described. The applicability of the selected techniques in the area of food science and technology will be shown by several examples. As this is an example-based review, the selected methods will be directly applied on the study of quality changes of shelf-stable strawberry juice as introduced in Section 1. To the best of our knowledge, a comparative study applying different methods on one food-related data set has never been performed before. Some points of attention of MVDA (Section 6) and the concluding remarks (Section 7) will be presented to finalise the paper.

3. Multivariate data analysis: terminology

3.1. X- and Y-variables

In the most simple case, the starting point of a MVDA is a two-dimensional matrix containing the experimental data (e.g. concentrations, peak areas, absorbance values) as visualised in Fig. 1A for the strawberry juice data set. In the context of MVDA, the experimentally obtained data are called the X-variables. Next to the X-variables, additional information can be available, for example, the class of a sample (e.g. processing type or botanical origin), the concentration of an analyte to be predicted or the storage time. Such information is considered as the Y-variable (Fig. 1A).

3.2. Principal components and latent variables

It is time consuming to look at all measured variables one by one for all samples to get insights into the data even when the experimentally obtained data are arranged in a data table as shown in Fig. 1A. Next to arranging the data in a data table, they can be graphically represented in a J -dimensional coordinate system in which each axis represents one of the X-variables. The experimentally obtained data can be considered as the coordinates of each sample in the J -dimensional coordinate system. However, when J is very large (for example 132 as in Fig. 1A), it is difficult for a human to visualise the data set. To overcome this issue and to assist with the interpretation of the data, a new coordinate system with a new and reduced set of variables is created starting from the original data set when performing MVDA. These new variables in the new coordinate system are called the principal components (PCs) or latent variables (LVs). These PCs/LVs are created using the relationships present in the original data set. This is only possible, when several sources of information follow the same pattern and this can be represented by a single new variable (i.e. the PC or LV). Consequently, MVDA methods aim to create a new set of variables which are lower in number compared to the original number of X-variables. In other words, these methods describe the variation (i.e. variability or differences in the samples) in the original data set (with J variables) as efficiently as possible with a few number of new variables R (with $R \ll J$). In summary, MVDA methods transform a $(I \times J)$ matrix with the information of I samples and J variables (e.g. detected chemical compounds or wavelengths) into a $(I \times R)$ matrix (Fig. 1A). The transformation of the original data set to PCs/LVs can increase its interpretability. Whether the term PC or LV is used depends on the method that is applied (see Section 5).

3.3. Scores and loadings

As described above, the original data table contains the coordinates of each sample in the original J -dimensional coordinate system. Upon the creation of the PCs/LVs, a new set of coordinates is generated in the new coordinate system for each sample. In the context of MVDA, the coordinates of the samples in the new coordinate system are called the scores t_{ip} . The transformation into the new coordinate system also results in the creation of loadings p_{ij} . Loadings quantify the contribution of the original X-variables to the PCs/LVs (Fig. 1A and B) as well as the correlation between the original variables. Variables with a high loading for a particular PC/LV are contributing more to that PC/LV (Brereton, 2017). Scores and loadings can be presented in a score- and loading plot (Fig. 1A and B), respectively. Alternatively, scores and loadings can be represented together in a biplot (see Section 5 for more information).

3.4. Supervised and unsupervised methods

Among the MVDA methods, a distinction can be made between supervised and unsupervised methods. Unsupervised techniques only use the information captured by the X-variables to describe the data. The

information captured by the Y-variable is not included in the analysis (Berrueta, Alonso-Salces, & Héberger, 2007; Bevilacqua et al., 2017; Dettmer, Aronov, & Hammock, 2007). Among the unsupervised techniques, PCA and PARAFAC will be discussed in this review. These methods are used to extract common patterns of variation out of data sets and are usually applied for data exploration (Bevilacqua et al., 2017). For many research problems, the aim is to predict a characteristic of a sample (e.g. class, concentration, storage time). When dealing with large data sets, supervised methods can be used for such cases. Supervised methods take into account the information captured by the X-variables (the independent predictor variables) and Y-variables (the dependent variable). The Y-variable can be continuous (e.g. concentration, storage time) or categorical (e.g. processing type, origin) (Berrueta et al., 2007; Bevilacqua et al., 2017; Dettmer et al., 2007). Among the supervised techniques, PLS regression will be discussed in this paper.

3.5. Pre-processing and unfolding of raw data

Before performing the actual analysis, data are often pre-processed (see also Section 6.1). Some examples of data pre-processing are peak alignment, mean centering, detrending and scaling. The reader is referred to Engel et al. (2013); Gerretzen et al. (2015); Nicolai et al. (2007); Roberts and Cozzolino (2016) and many more for more details about data pre-processing.

So far, we assumed that the original data set could be arranged in a two-dimensional data table or matrix. However, very often food scientists are dealing with a multi-dimensional data set (i.e. having more than two dimensions). This was also the case for the strawberry juice data set, as the data could be arranged in a three-dimensional array (*storage times* $\times$ *volatiles* $\times$ *storage temperatures*). Notwithstanding the fact that such data could be analysed without unfolding, very often the original multi-way data set is unfolded prior to the pre-processing. Unfolding or matricising is the action that rearranges a multi-way array into a matrix (e.g. from a three-dimensional array to a two-dimensional data set). This is often referred to as *slicing* of the array in a particular dimension. These slices (or matrices) are combined next to each other to obtain one large matrix, which can be analysed using well-known multivariate methods (e.g. PCA or PLS). The drawback of unfolding is that part of the information captured by the original data set is lost and that the nature of the model is different from the nature of the original data, which can lead to overfitting (i.e. when the error of prediction increases with increasing number of model variables) (Abdi & Williams, 2010; Abdi, 2010). Moreover, unfolding can cause difficulties to interpret the model output and to make predictions (Kiers, 2000). Lastly, notwithstanding the fact that such data could be analysed without unfolding, very often the original multi-way data set is unfolded prior to the pre-processing.

4. Main steps of a multivariate data analysis

It is not straightforward to propose a general protocol for a multivariate data analysis that is applicable to all data. However, some steps should always be included (Bevilacqua et al., 2017; Roberts & Cozzolino, 2016):

Step 1 - Data pre-processing: removal of unwanted variation in the data linked to sampling and instrumental artefacts in order to focus on the information of interest captured by the data (see Section 6.1) (Engel et al., 2013).

Step 2 - Data exploration: assessing the quality of the pre-processing and detecting correlation patterns and outliers in the samples using unsupervised models. Data exploration can be performed with or without the research question in mind.

Step 3 - Supervised modelling (if applicable): taking into account the information of both the X- and Y-variables. This allows to identify which information captured by the X-variables is relevant to predict the Y-variable.

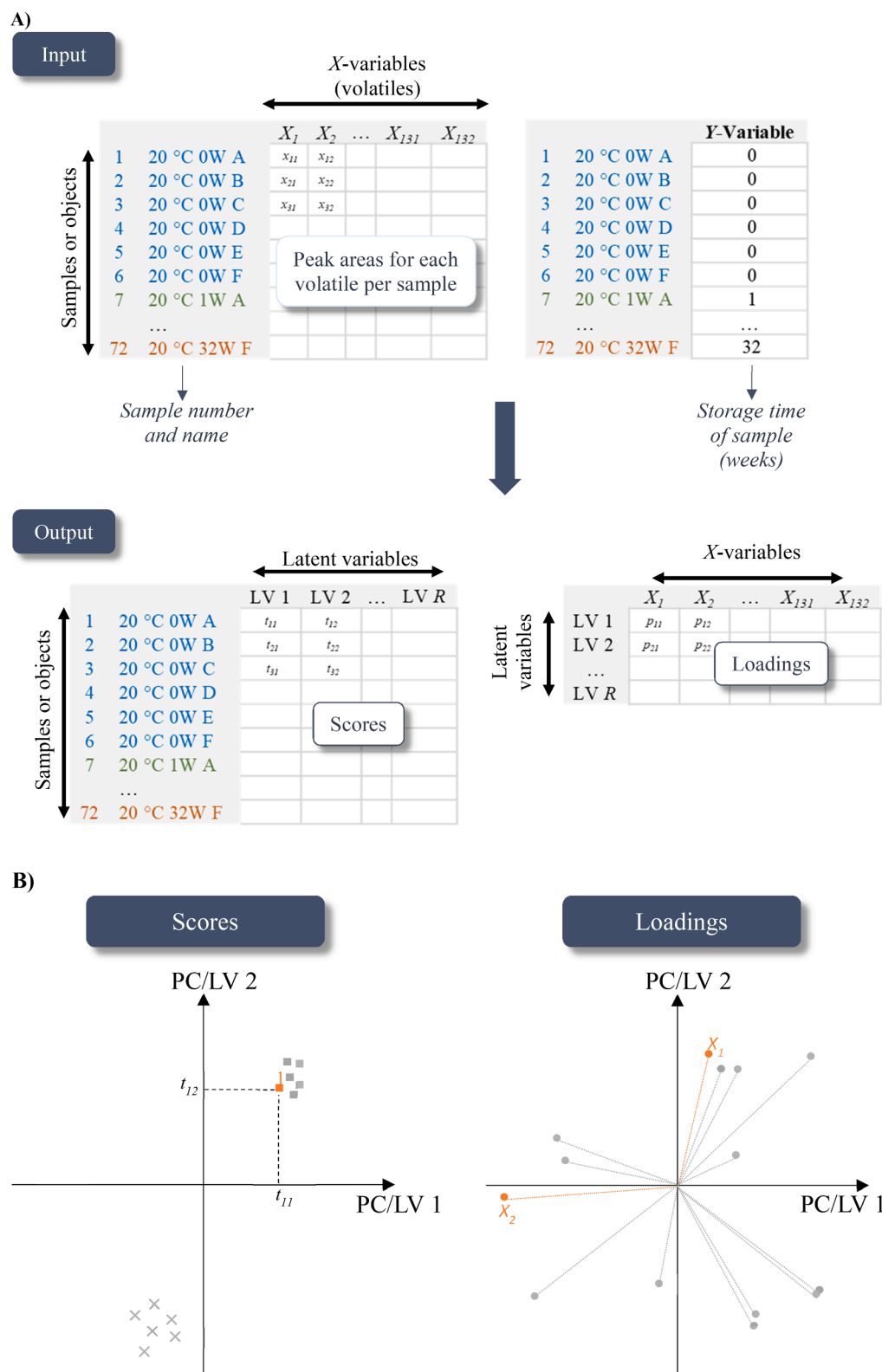


Fig. 1. Schematic representation of the terminology (X- and Y-variables, PCs or LVs, scores, loadings) of a multivariate data analysis (MVDA) using the strawberry juice data set (samples stored at 20 °C for 32 weeks) as an example., A) A two-dimensional data table (*samples* $\times$ *volatiles*) containing information about the X-variables (volatile compounds, $J = 132$ in total) and Y-variable(s) (storage time) was used as the input for the MVDA. Each sample was analysed six times (indicated with the letters A-F). Based on the original data set, a new set of R variables (the latent variables, LVs), was created which are lower in number compared to the original number of X-variables (i.e. $R \ll J$). Associated with the LVs, scores (t_{ir}) and loadings (p_{rj}) are calculated, which are linked to the samples and the X-variables, respectively, B) Score and loading plots to visualise the concept of scores and loadings. The scores (t_{ir}) are the coordinates of the samples in the new coordinate system spanned by the PCs/LVs. Samples represented by the same symbol are considered as analytical replicates. The scores of sample 1 (indicated in orange in the left plot) for PC/LV 1 and 2 are t_{11} and t_{12} , respectively. The loadings quantify the contribution of the original X-variables (indicated as dots in the right plot) to the PCs/LVs. Variable X_1 is largely contributing to PC/LV 2, while it only has a small contribution to PC/LV 1. For variable X_2 the opposite can be deduced.

Step 4 - Model refinement: improvement of the model to increase its performance and interpretability.

Step 5 - Model interpretation: translating the model output into relevant information from a food scientific point of view.

Related to the main steps of a multivariate data analysis, a number of questions should be considered by a food scientist before the start of the analysis in order to make some important decisions. The decisions to be made are related to the type of analysis as well as the pre-processing method(s) to be applied. Next to the points addressed below, it is noteworthy to mention that prior to the actual analysis, food scientists should also think about the experimental design, sample selection and data collection as they are of critical importance for the actual data analysis (Bevilacqua et al., 2017; Roberts & Cozzolino, 2016).

• What is the purpose of the work?

The purpose of the work is strongly linked to the **method** that should be applied (Steps 2 and 3). Classification, prediction, inference in relation to an experimental design, visualisation of trends or outlier detection require different methods (see Section 5).

• What does the structure of the data look like? Can they be organised in a two- or multi-dimensional structure?

The data structure determines the **analyses** that can be performed. Two-dimensional data sets should be analysed using the basic MVDA methods. Multi-dimensional data sets can be analysed using MVDA methods which were developed for such cases or they can be unfolded into two-dimensional data sets and analysed with the basic MVDA methods (see Section 5).

• What is the origin of the data? Are they the result of one analysis (e.g. untargeted GC-MS fingerprinting) or more than one analysis (e.g. combining the output of GC-MS, HPLC data, etc.)?

Depending on the nature of the data, different **pre-processing** methods need to be applied. Data pre-processing allows, next to removal of artefacts, that signals originating from different samples or different types of data can be compared and used in the same application.

As mentioned before, it is impossible to focus on all the aspects of a MVDA in detail in one paper. For this reason, this paper will focus on a selected number of methods which have, based on our opinion, a broad applicability and which can be/are of high relevance in the area of food science.

5. Theoretical background of a selection of multivariate data analysis methods and their application in a case study

As mentioned before, a wide variety of MVDA methods exist in literature. The theory of four methods will be discussed: PCA, PARAFAC, PLS and ASCA. Next to this, a summary of each selected method is given in Table 1. For each method a selection of scientific papers is summarised in Table 2 to show their applicability. As this is a example-based review, the selected methods were applied on the strawberry juice data set.

5.1. Principal component analysis (PCA) for two-way data

5.1.1. Theoretical background

PCA is the most popular MVDA technique to extract information out of large multivariate data sets. It is used for initial data inspection and exploration, to look for overall trends, quality control, outlier detection and overall pattern recognition (Abdi & Williams, 2010; Wold, Esbensen, & Geladi, 1987). PCA, combined with visualisation of scores and loadings, can be applied to almost all data, thereby making it a versatile

Table 1

Overview of selected data analysis techniques: original paper, data structure, important points and examples of applications.

Method	Original paper	Data structure	Important points	Application in food science and examples of the selected data analysis techniques in the area of food science
PCA	Hotelling, 1933	2 dimensional	- Unsupervised method - Data exploration: obtain first insights into data, look for trends and clustering	Widely applied (Belmonte-Sánchez, Romero-González, Martínez Vidal, Arrebola, & Garrido French, 2020; Kang et al., 2020; López-Maestresalas et al., 2016)
PARAFAC	Harshman, 1970	Multi-dimensional	- Unsupervised method - Decomposition of multi-way data - Data exploration	Mostly applied on fluorescence spectroscopy data (Callejón et al., 2012; Elcoroaristizabal et al., 2016; Lenhardt et al., 2017; Rubio, Sanllorente, Sarabia, & Ortiz, 2019)
PLS-(DA)	Wold, 1980	2 dimensional	- Supervised method - Prediction or classification	Widely applied, often on NIR spectroscopy data (Aernouts, Polshin, Saeys, & Lammertyn, 2011; Nascimento, Carvalho, Júnior, Pereira, & Teixeira, 2016; Wafula et al., 2020; Wibowo, Grauwet, Kebede, Hendrickx, & Van Loey, 2015)
ASCA	Smilde et al., 2005	2 dimensional	- ANOVA combined with PCA - Inference - Address variation in the data to a specific experimental factor	Only limitedly applied in food science (Bevilacqua, Bucci, Materazzi, & Marini, 2013; Strani, Grassi, Casiraghi, Alamprese, & Marini, 2019).

technique for data exploration. Consequently, PCA is often used as a first step to obtain insight into large data sets (Abdi & Williams, 2010; Jolliffe, 2002; van den Berg, Hoefsloot, Westerhuis, Smilde, & van der Werf, 2006; Wold et al., 1987).

PCA is by far the most popular MVDA method used by food scientists. PCA is classified as an unsupervised method, meaning that only the information of the X -variables is used (see Section 3). The starting point for PCA is a two-dimensional data set or matrix X . This is a $(I \times J)$ matrix containing the information of I samples and J variables (the X -variables, e.g. detected chemical compounds, wavelengths). PCA linearly transforms the matrix X into a new and smaller set of R variables, the PCs, by which the dimensionality of the original data set is reduced as $R \leq \min(J, I)$. The PCs are created in such a way that they capture as much as possible of the variation in the original data set: the first PC captures most of the variance followed by the second PC and so on (Jolliffe, 2002; Wold et al., 1987). As introduced in Section 3, each PC is characterised by scores and loadings. The PCA scores are the coordinates of the samples in the new coordinate system and the PCA loadings quantify the contribution of the original variables to the PCs (Abdi & Williams, 2010; Jolliffe, 2002; Wold et al., 1987). A detailed mathematical description of PCA can be found in Supplement 2.

The outcome of a PCA model can be visualised in individual scatter plots (e.g. score and loading plots of PC 1 versus PC 2). However, the

Table 2

Selection of scientific papers published in the last 10 years to show the applicability of the selected data analysis techniques in the area of food science. Examples in which the quality of food products was investigated were selected. The studied quality attributes are indicated in *italic*.

Food product	Research objective	Result	Reference
PCA			
Rum	Classification of golden rum based on different quality factors (<i>fermentation barrel, raw material, distillation method and aging</i>) using NMR data.	PCA was used for data exploration, to show trends and grouping in the rum samples relying on the selected factors.	(Belmonte-Sánchez et al., 2020)
Sea cucumber	Identification of <i>origin</i> of Chinese sea cucumbers using stable isotope ratios and composition	PCA was used for data exploration. No clear difference between sea cucumbers with different origin based on PCA model.	(Kang et al., 2020)
Mozzarella cheese	Discriminate between real and fake mozzarella cheese based on <i>fatty acid profile and phytosterol content</i> .	PCA enabled to differentiate between real and fake mozzarella by using a score plot based on PC 1 and PC 2.	(Kim et al., 2014)
Potato	Check the potential of PCA to group potato samples based on <i>blackspot damage</i> .	PC 4 caused separation between the damaged (blackspot) and non-damaged potato samples.	(López-Maestresalas et al., 2016)
PARAFAC			
Sparkling wine	Study <i>browning</i> in sparkling wines	PARAFAC decomposition of fluorescence spectroscopy data (excitation emission matrices can be used tool to study wine quality in terms of browning).	(Elcoroaristizabal et al., 2016)
Strawberry jam	Determination of <i>food colorant cochineal</i> in the presence of carmoisine in strawberry jams using fluorescence spectroscopy	A four-component PARAFAC model of the fluorescence data allowed to identify and quantify cochineal in the presence of carmoisine (a fluorescent quencher).	(Rubio et al., 2019)
Sherry vinegar	Classify Sherry vinegars based on their <i>ageing time</i> using fluorescence spectroscopy data.	PARAFAC allowed to extract relevant information about the vinegars (fluorophores). PLS-DA starting from the PARAFAC components allowed to classify Sherry vinegar based on their ageing time.	(Callejón et al., 2012)
Cereal flour	Characterise different cereal flour samples to be able to link it to their <i>botanical origin</i> using fluorescence spectroscopy.	Four fluorophores were present in the cereal flour based on the outcome of a four component PARAFAC model. Two fluorophores corresponded to amino acids and the two others were most probably low molecular weight molecules. PLS-DA was used to classify the different flours starting from the output of PARAFAC.	(Lenhardt et al., 2017)
PLS regression			
Milk	Prediction of <i>fat, crude protein, lactose and urea</i> in raw and homogenised milk using mid IR data	The success of prediction of the PLS model was depending on the applied sampling technique and homogenisation.	(Aernouts et al., 2011)
Orange juice	Study the evolution of the <i>volatile fraction</i> of pasteurised orange juice during storage	PLS has been successfully applied to study the evolution of the volatile fraction. Calculated VID coefficients to select the volatiles changing the most.	(Wibowo et al., 2015)
Peaches	Predict the <i>soluble solid content and firmness</i> of peaches using NIR spectroscopy data	NIR spectroscopy in combination with PLS regression has the potential to predict the soluble solid content and firmness of peaches. However, the PLS model should be optimised to minimise prediction errors.	(Nascimento et al., 2016)
Common beans	Predict the <i>cooking time</i> of aged common beans using NIR spectroscopy	The cooking times of aged common beans could be predicted using NIR data and PLS regression.	(Wafula et al., 2020)
Chickpeas	Study the evolution of the <i>volatile fraction</i> of sterilised chickpeas during storage	The application of PLS in combination with the calculation of VID coefficients allowed to gain insight into the volatile changes of sterilised chickpeas.	(Noordraven, Buvé, Chen, Hendrickx, & Van Loey, 2021)
ASCA			
Milk	Study impact of <i>process factors</i> (temperature, time, fat content and pH) on milk renneting with FT-NIR.	ASCA showed that the four main effects and second order interactions have a significant impact on the milk renneting process.	(Strani et al., 2019)
Dried egg-pasta	Study the effect of the <i>manufacturing process</i> (drying time and temperature), <i>egg content</i> and their interactions on dried egg-pasta quality using NIR spectroscopy	The results of ASCA showed that drying time, drying temperature, egg content and the interaction (egg content × temperature) had a significant effect on dried egg-pasta quality.	(Bevilacqua et al., 2013)

results of a PCA model are often represented in a two-dimensional biplot. This is a combined graphical representation of the samples and variables in a two-dimensional space using the scores and loadings. A PCA biplot gives an indication about trends or groupings of samples in the original data set and allows to detect the variables which are indicative for the structuring/grouping among the samples (Jolliffe, 2002). Examples of biplots are given in Fig. 2.

5.1.2. PCA to study strawberry juice volatile changes during storage

As an example, a PCA was performed to investigate the impact of storage at 20 °C on the volatile fraction of a shelf-stable strawberry juice. More details on the strawberry juice data set can be found in Supplement 3. Prior to modelling, mean-centering and scaling using the standard deviation were applied as pre-processing. During mean-centering and scaling, the mean of the variables (i.e. volatile compounds) was subtracted and each variable was divided by its standard deviation, respectively. As a result, the new mean of each variable is equal to zero and its standard deviation is equal to one. Mean-centering was applied to assure that the first PC was not pointing wrongly to the center of the data cloud, which will also affect the position of the following PCs (Jolliffe, 2002). Scaling was applied because the peak areas (i.e.

measured values) of the different volatiles were considerably different. In other words, they had different ranges.. As PCA is unsupervised, only the X-variables were modelled. It was not straightforward to select the most appropriate number of PCs in the model, because the outcomes of the different model selection methods (see Section 6.2 for details) were contradictory. For example, based on the eigenvalues (>1), more than 20 PCs should be selected, while the plot of log(eigenvalues) showed a clear 'jump' for the model with 3 PCs. From PC 6 on, the contribution to the variance was 2.5% or less. For these reasons, a model with 5 PCs was selected starting from the pre-processed two-way data set (*storage times* × *volatile peak areas*) to evaluate volatile changes as a function of shelf-life. This model only explained 49.0% of total variance present in the original data set, with the first two PCs explaining 37.2% of the variance. It is important to notice for this example that only a small fraction of the variance present in the original data set was explained by the model and that from PC 6 on the explained variance was rather low (see Section 6.2 for more information about model variable selection). The effect of interest (i.e. storage time) on the changes in volatiles is, however, so dominant that a clear trend could be observed in the modelled data. Readers should keep in mind that this will not always be the case as all selected PCs (or LVs) can explain some interesting effects in the data set.

In contrast, it is also possible that no clear trends are observed in the modelled data (for example when the studied samples are very similar, when they are not affected by the applied processing, etc.). This is of course case specific and will not be further considered in detail. However, the researcher can in this case take a look at its research hypotheses, check the raw data (e.g. making some plots) and ask the question whether MVDA is the most appropriate data analysis technique.

Part of the output of the selected PCA model was visualised in a biplot of the first two PCs (Fig. 2). From this biplot, it can be observed that PCA allows to detect outliers in the replicates (e.g. for the storage time of 20 weeks, ●). Secondly, there is a separation between the samples with the different storage times (represented by the different symbols). This indicates that storage time affects the volatile fraction of strawberry juice during storage. Thirdly, the non-stored (0 weeks) samples are located on the left side of the biplot (i.e. compared to the centre of the biplot), while the stored samples are located more to the right. However, the symbols representing the stored samples are not ordered chronologically along PC 1. In other words, PC 1 only is not enough to adequately describe the effect of storage time on the volatile composition. PC 2 results in the grouping of some samples (projection on the vertical axis). However, the separation seems to be not linked to storage time. When looking at PC 3–5, no extra information related to the storage time could be deduced (results not shown).

In summary, PCA resulted in grouping of different samples based on the storage time, but it does not provide a clear view of the effect of storage time on the volatiles. This is most probably related to the fact that PCA is an unsupervised technique in which the information about the Y-variable (storage time) is not explicitly included (in contrast to PLS, see Section 5.2). Nevertheless, the trends in the data set linked to storage time will become visible because the data set which was used as the input for the PCA included the abundance of the different volatile compounds (in terms of peak areas) at the different storage times.

Lastly, it should be noted that the total variance explained by the

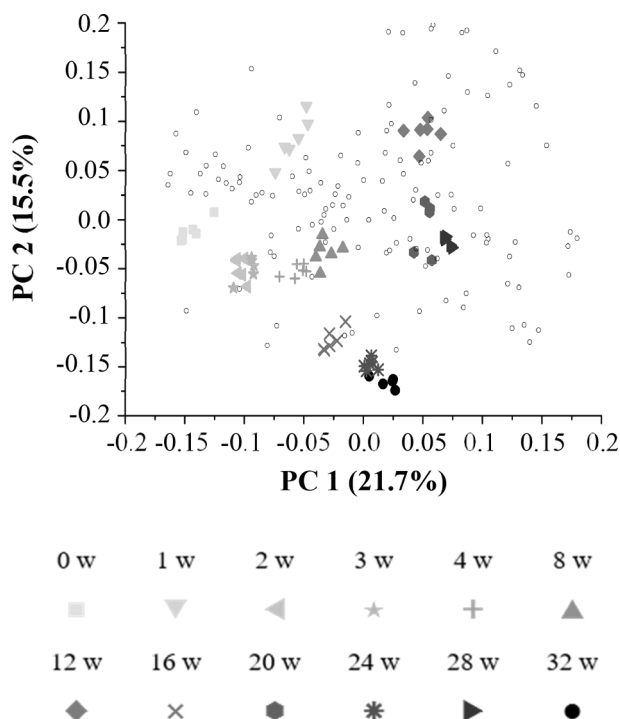


Fig. 2. PCA biplot describing the change in volatile fraction of pasteurised strawberry juices during storage at 20 °C in a standard PET bottle. The percentages of variance explained by the two principal components (PCs) are indicated on the respective axes. The strawberry juice samples (i.e. storage times) are displayed by the different symbols (w = week(s)) and the volatile compounds by the open circles.

model is rather low. Similar trends were observed when comparing the outcome of the PCA models for the other temperatures (28–42 °C).

5.2. Parallel factor analysis (PARAFAC) for multi-way data

5.2.1. Theoretical background

PARAFAC can be seen as a generalisation of PCA (Bro, 1997). Similar to PCA, PARAFAC is an unsupervised technique, which can be used for data exploration. Like PCA, PARAFAC aims to describe the original data in a few latent components/factors which are characterised with loadings for all dimensions (including for the samples) (Smilde, Bro, & Geladi, 2004). Due to the similarities between PCA and PARAFAC, a short mathematical description is included as Supplement 4. The theory and mathematical background of PARAFAC have already been discussed extensively by Bro (1997), Harshman (1970) and Smilde, Bro, and Geladi (2004).

The main advantage of PARAFAC when dealing with multi-way data, is the fact that it can be applied on the data set as such without unfolding. Unfolding of a data set can result in a loss of information, because the nature of the original data set does not correspond anymore with the nature of the applied model (Bro, 1997). Based on the work of Bro (1997) and Smilde et al. (2004), a comparison of PARAFAC with unfolding combined with PCA shows that (i) unfolding has to deal with issues related to overfitting of the multi-way data and leads to models which are less robust and less interpretable; (ii) a PCA model will always fit the data better than a PARAFAC model; (iii) multi-way models are less sensitive to noise, (iv) when a PARAFAC model is able to describe the data well, the PCA model will include noise and (v) the unique information captured in the original multi-way data set is preserved when applying PARAFAC (Bro, 1997; Smilde et al., 2004). It should be noted that until now PARAFAC was almost exclusively applied on fluorescence spectroscopy data due to its ability to model physical phenomena directly (i.e. to split the fluorescence signal into signals of pure constituents) (see Table 2 for examples). For this reason, it can be questioned whether these statements are valid for all types of multi-way data. Therefore, we recommend to compare the outcome of a PARAFAC model and the results of PCA modelling applied on the unfolded data carefully in order to choose the most appropriate method for data exploration.

5.2.2. PARAFAC to study strawberry juice volatile changes during storage

The obtained three-way data set (storage time $\times$ peak areas $\times$ storage temperature) was analysed with PARAFAC to explore the data and to gain a first insight into the effect of storage time and temperature on the strawberry juice volatile fraction. As a pre-processing step, the data were multi-way scaled on mode 2 (i.e. the dimension of the volatile compounds) with the standard deviation as a weighting factor. Multi-way scaling is similar to the scaling introduced in Section 5.1 (PCA), but only the mode (or dimension) in which the scaling is done should be selected.

A two-component PARAFAC model best described the evolution of the volatile fraction during storage. Model selection was based on the percentage of variance explained (72.5%) and the degree of similarity after performing a split-half analysis (85.4%) (Bro & Kiers, 2003). The results of the PARAFAC model are graphically represented in Fig. 3. The results for component 1 of the PARAFAC model show that the loadings along the sample mode (mode 1) (Fig. 3A) decreased as a function of storage time. The loadings linked to the volatiles (mode 2) (Fig. 3C) are both positive and negative, indicating that the detected strawberry juice volatiles are behaving differently during storage. The loadings of mode 3 (storage temperature) (Fig. 3B) are decreasing with increasing temperature. The results of the second PARAFAC component are shown in Fig. 3D–F. The loadings for mode 1 are increasing as a function of storage time, being the opposite trend compared to component 1. For mode 2, both positive and negative loadings are observed. It should be noted that for mode 2 there are more negative loadings for component 2

compared to component 1. The loadings for mode 3 show no clear trend as a function of temperature.

Based on the observed trends in the two PARAFAC components, it can be hypothesised that PARAFAC component 1 is related to the loss of fresh strawberry juice characteristics. Therefore, volatiles with positive loading in mode 2 for this PARAFAC component tend to decrease during storage and *vice versa*. In Fig. 3C, the chemical classes of the volatiles are indicated. A comparison of the volatile changes during storage with literature data (Buvé et al., 2018; Golaszewski, Sims, O'Keefe, Braddock, & Littell, 1998; Siegmund, Derler, & Pfannhauser, 2001) confirms that the compounds with positive loadings are degraded during storage. The decreasing trend of the loadings for mode 3 (temperature) for the first PARAFAC component indicate a faster change at higher storage temperatures of volatiles related to fresh strawberry juice characteristics. It is hypothesised that the second PARAFAC component describes the formation of volatiles induced by storage. For some volatiles, an opposite trend in the loadings is observed compared to PARAFAC component one, which confirms this hypothesis. Due to the non-consistent trend in the loadings for mode 3, it is difficult to draw conclusions about the effect of temperature on PARAFAC component 2. In summary, both storage time and temperature seems to have a clear impact on the change in volatile fraction of strawberry juice during storage. However, interpreting the outcome of the PARAFAC model is not straightforward and conclusions should be drawn with caution. When interpreting the results of PARAFAC, it should be remembered that a linear relationship between the storage times, volatiles and storage temperatures is assumed in a PARAFAC model. Based on the outcome of the second PARAFAC components, it can be seen that this assumption is not completely correct as the loadings of 28 °C are higher than the ones of 35 °C. The effect of temperature can most probably be described by the Arrhenius equation which has an exponential nature (Buvé et al., 2018; van Boekel, 2008). Further optimisation of the model and a better understanding of the different components of the model would allow to gain more insight into the volatile changes and the possibilities of PARAFAC to study food quality changes during storage.

5.3. Partial Least Squares (PLS) regression for two-way data

5.3.1. Theoretical background

PLS regression is categorised as a supervised method and aims to predict a characteristic of a sample (e.g. concentration of a compound). To be able to predict this characteristic, the X-variables (analytical data) and Y-variables (characteristic of interest) are used. PLS regression combines the principles of PCA and multiple linear regression to predict the dependent (and continuous) Y-variable based on the independent X-variables. PLS decomposes the matrices X and Y (with the X- and Y-variables, respectively) simultaneously to create a common set of R new LVs. The LVs are linear combinations of the original X-variables and are constructed such that they maximally capture the covariance between X and Y (Abdi, 2010). In a second step, Y is regressed on the R LVs to be able to predict the Y-variable (e.g. concentration, storage time, cooking time, etc.) (Abdi, 2010; Smilde, Bro, & Geladi, 2004; Wold, Sjöström, & Eriksson, 2001). Supplement 5 gives a mathematical description of PLS.

Very often, the goal is to predict more than one Y-variable (e.g. concentration of different compounds). In this case, two approaches can be followed. Firstly, a standard PLS regression (PLS-1 regression) can be performed to predict each Y-variable individually. By this, a new set of LVs is defined for each Y-variable. Secondly, one can apply a PLS-2 regression in which one common set of LVs is defined for the prediction of all Y-variables simultaneously. The best approach will depend on the data set and the research goal. PLS-1 will result in more precise models and has higher predictive power in case there are non-linear relationships in the data. In contrast, if the set of Y-variables is correlated and if they contain noise, PLS-2 is the best option to model the data and exclude noise (Camo Software, 2006; Pedro & Ferreira, 2007). Examples of the use of PLS-2 regression can be found in the work of

Gendrin, Roggo, and Collet (2007) and Pedro and Ferreira (2007).

Next to the possibility of having multiple continuous Y-variables, it is also possible to have a categorical Y-variable (e.g. classes related to the origin of the sample or the processing type). In this case, an extended version of PLS can be applied, Partial Least Squares-Discriminant Analysis (PLS-DA). PLS-DA is a classification technique which aims to separate samples belonging to different classes/groups as much as possible. Like in PLS regression, this is performed predicting the class (Y-variable) based on the X-variables (measured data). When executing a PLS-DA, a dummy matrix is defined in which the number of columns is equal to the number of classes and the number of rows corresponds to the total number of samples. If a sample belongs to a class, the dummy variable in the corresponding column is equal to one. In contrast, the dummy variable in a column is zero if it does not belong to that class. This dummy matrix is used as the Y-matrix in a PLS-DA regression (Barker & Rayens, 2003; Bevilacqua et al., 2017; Brereton & Lloyd, 2014; Gromski et al., 2015). Examples of PLS-DA can be found in the work of Cynkar, Damberg, Smith, and Cozzolino (2010) and Kebede et al. (2014).

PLS regression is a widely used MVDA tool. An advantage of PLS in comparison with PCA is that information of both the X- and Y-variables is used to describe the data. This allows to gain more insight into the data and to make predictions (Jolliffe, 2002; Maitra & Yan, 2008). According to Khakimov et al. (2015), supervised regression methods (like PLS) are frequently used to predict characteristics of food products which are not easy to measure and of which the analyses are expensive. PLS regression is able to handle variables which are correlated and noisy. In addition, PLS regression is capable of predicting different Y-variables at the same time even if they are correlated (Abdi, 2010; Smilde et al., 2004; Wold et al., 2001). When performing PLS-DA, it should be remembered that the number of classes/groups and number of samples per class has a large effect on the PLS-DA model. The model output of PLS-DA will be better when the data set has a limited number of classes all having a large number of samples (Granato et al., 2018).

5.3.2. PLS to study strawberry juice volatile changes during storage

The data of the samples stored at 20 °C (X- and Y-variables) were modelled with PLS regression to gain insight into the effect of storage time on the strawberry volatiles (published by Buvé et al., 2018). Before modelling, the data were mean-centered and scaled using their standard deviation (see Section 5.1). A PLS regression model with three LVs was selected to describe the evolution during storage. The RMSECV of the selected model was 12.3 weeks (see Section 6.2) and did not reduce significantly if more LVs were included. The first two LVs explained a high percentage of Y-variance (96.5%). Using the X- and Y-variables for modelling allowed to more adequately describe the variation in the volatile components related to storage time in comparison with the PCA model (49.0% of explained variance, Section 5.1).

When examining the biplot of LV 1 and LV 2 (Fig. 4), it can be seen that the non-stored pasteurised strawberry juice samples (0 weeks) are located on the left side of the biplot and that the samples which were stored for 32 weeks (i.e. the longest time) are located on the right compared to the centre of the biplot. All other storage times are located in between in a chronological order along the LV 1 axis. Consequently, LV 1 is describing the largest part of the effect of storage time (92.0 %). The clear effect of storage time is also represented by the long and almost horizontally positioned Y-vector on the biplot. It should be noted that the length and position of the Y-vector is highly depending on the data set (and the selected model). In the example of strawberry juice, the effect of storage time (Y-variable) was represented by LV 1 and therefore the Y-vector is positioned almost horizontally. In case the Y-vector would have been short, it would mean that the continuous Y-variable would have a rather limited effect on the variation in the data. In addition, the Y-vector can point to other directions as well (e.g. the lower quadrant on the right or vertically). If the Y-vector would have been long and positioned almost horizontally, but pointing to the lower

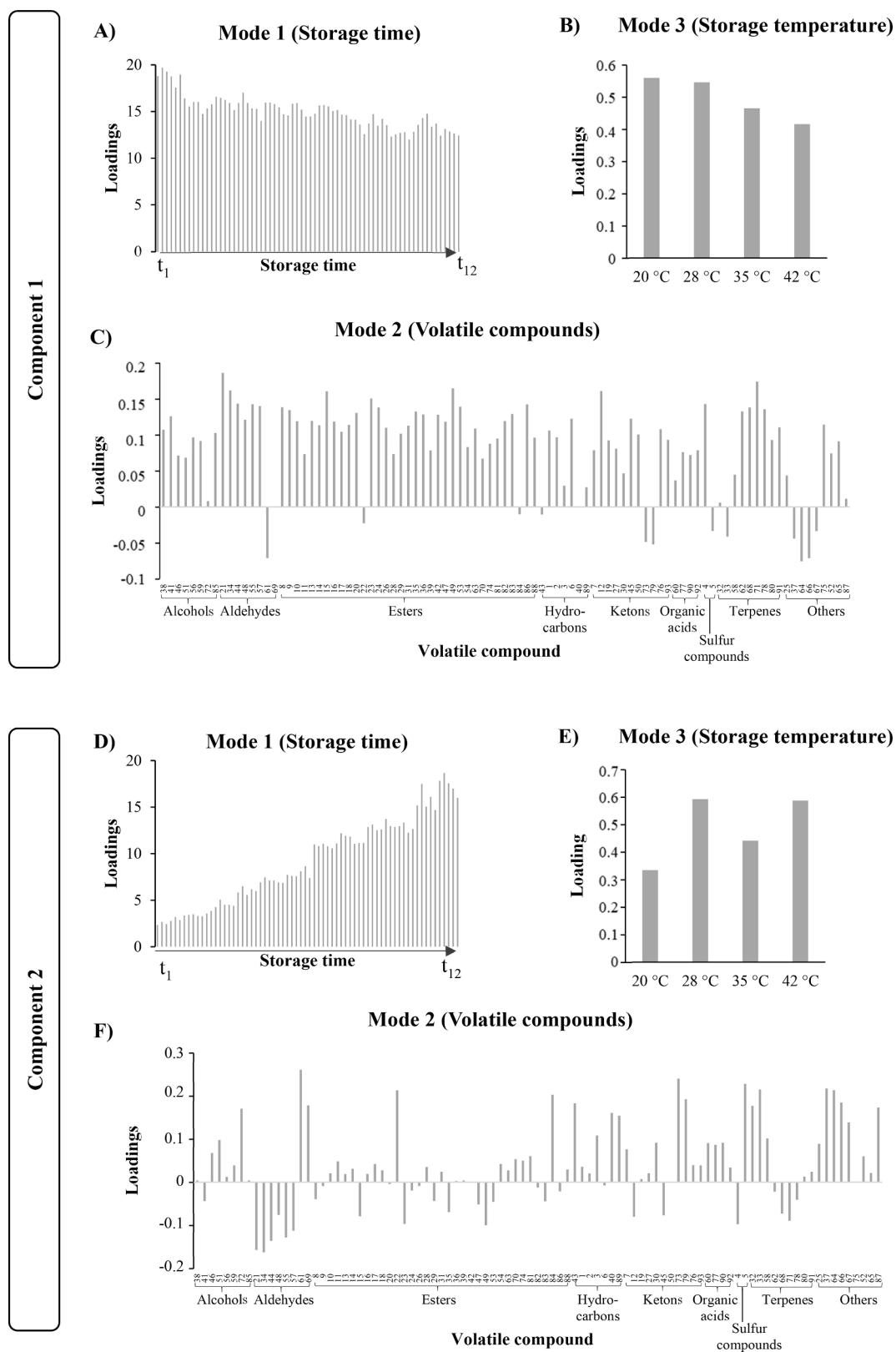


Fig. 3. Output of a two-component PARAFAC model for the three modes (storage time, volatiles and storage temperature). (A-C) Loadings of the first PARAFAC component and (D-F) loadings of the second PARAFAC component. The volatile compounds are ranked based on their chemical class which was obtained using NIST spectral database.

quadrant on the right of the reference point (i.e. centre of the biplot) the same conclusions would have been drawn as is the case now. However, in the case the Y-vector would be long, but positioned almost vertically, the effect of storage time would be explained by the LV which is represented by the vertical axis in the biplot (e.g. LV 2).

Due to the clear effect of storage time (represented by LV 1), the biplot allows to gain insights into the relation between the strawberry juice volatiles and storage time. The location of an X-variable (volatile) in a biplot gives an indication about its importance to describe the changes related to the Y-variable. For this purpose, correlation circles are added to a PLS biplot. These circles correspond to the correlation coefficient between the original peak areas and PLS scores. In Fig. 4, the 70 and 100% correlation circles were added. If a volatile is projected far from the centre of the biplot and between the inner (70%) and outer (100%) circles, the volatile is important for describing the observed trend. The concentration of these volatiles changed consistently and considerably during the whole time frame of storage. In addition, for present case study, if a compound is located on the right side of the biplot (i.e. compared to the centre of the biplot which is the reference point), it is positively correlated with storage time. And hence, the concentration of this volatile increased during storage. Volatiles which are located on the left side compared to the centre of the biplot are negatively correlated with time and decreased during storage. Similar trends were observed for the other storage temperatures (28, 35 and 42 °C) (results not shown).

5.4. ANOVA-Simultaneous component analysis (ASCA) for designed experiments

5.4.1. Theoretical background

Very often the experiment that is performed by a food scientist has an underlying design. Examples of the design factors of an experiment can be time, type of treatment, dose of a drug, storage temperature, etc. MVDA techniques such as PCA are not able to take into account the underlying structure or design factors of an experiment. Consequently, the contribution of these design factors to the variation in the data set is not always clear in the model output and these methods might fail to extract information related to the design factors (Jansen et al., 2005). Analysis of variance (ANOVA) is commonly used to study the impact of design factors on one dependent variable. However, as today's research focusses on multiple response variables rather than one, ANOVA is not the best choice, because it cannot take into account the relation between the different variables of interest. In this case, MANOVA (multivariate ANOVA) is a valuable alternative. However, MANOVA fails to properly analyse multivariate data when the number of variables is higher than the number of measurements (Jansen et al., 2005; Smilde et al., 2005; Zwanenburg, Hoefsloot, Westerhuis, Jansen, & Smilde, 2011). To provide an alternative to analyse multivariate data from an experimental design, Smilde et al. (2005) proposed ANOVA Simultaneous Component Analysis (ASCA).

ASCA combines the features of ANOVA and PCA to analyse multivariate data and eliminates the disadvantages of both methods. In ASCA, ANOVA is used to decompose the multivariate data matrix into different effect matrices related to the different design factors and a residual matrix containing the unexplained variation. In a next step, a Simultaneous Compound Analysis (SCA) is performed. SCA can be considered as a PCA when dealing with multiple X-matrices, here being the different effect matrices (one per design variable). The different effect matrices are analysed simultaneously with PCA to extract information about the different design factors (e.g. time, type of treatment) based on the X-variables. By this, the relationship between the different variables in a multivariate data set is modelled (Bevilacqua et al., 2017; Jansen et al., 2005; Timmerman, Hoefsloot, Smilde, & Ceulemans, 2015; Zwanenburg et al., 2011). A concise mathematical description of ASCA can be found in Supplement 6.

The most important advantage of ASCA is that it takes into account

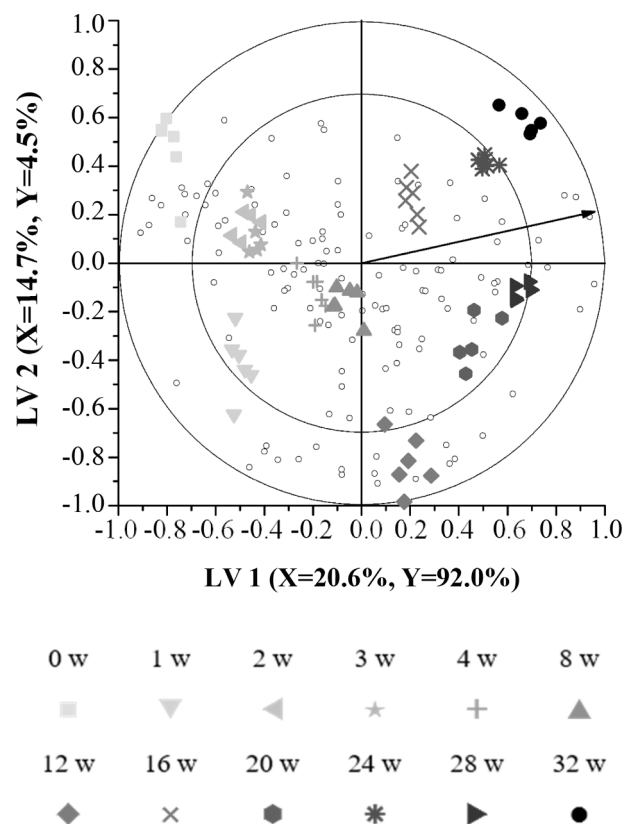


Fig. 4. PLS biplot describing the change in volatile fraction of pasteurised strawberry juices during storage at 20 °C in a standard PET bottle. The percentages of X- and Y-variance explained by the two latent variables are indicated on the respective axes. The strawberry juice samples (i.e. storage times) are shown by the different symbols (w = week(s)) and the headspace compounds by the open circles. Correlation circles of 100% (outer circle) and 70% (inner circle) are giving an indication about the importance of each volatile in describing the change during storage. The Y-vector represents the correlation loadings of storage time (Y-variable) (Buvé et al., 2018, with permission from Elsevier).

both the relation between the different variables and the design of the experiment, which is not the case when ANOVA or PCA are applied separately (Jansen et al., 2005; Zwanenburg et al., 2011). One of the drawbacks of ASCA is the possibility to overestimate the effect of a certain design factor (Zwanenburg et al., 2011).

5.4.2. ASCA to study strawberry juice volatile changes during storage

The performed storage experiment had two design factors: storage time and temperature. First, an ANOVA decomposition was performed to separate the variation in the data into different factor effects (time and temperature). It was assumed that: (i) all non-stored strawberry juice samples (0 weeks of storage) were the same, consequently, no effect of temperature before the start of storage was observed and (ii) that temperature accelerates the effect of storage time. Because of this the matrix X with the volatile compounds for the different storage times and temperatures could be described with the ANOVA model represented in Eq. (1). An average volatile concentration at the start of storage was defined as the reference effect ($\mu(t_0)$). The effect of time at 20 °C was defined as an additive term ($X_{t20} \text{ } ^\circ\text{C}$), while the effect of time at higher temperatures were defined as additive to the effect at 20 °C ($X_{t28} \text{ } ^\circ\text{C}-20 \text{ } ^\circ\text{C}$, $X_{t35} \text{ } ^\circ\text{C}-20 \text{ } ^\circ\text{C}$ and $X_{t42} \text{ } ^\circ\text{C}-20 \text{ } ^\circ\text{C}$). E is the matrix with the residuals.

$$X = \mu(t_0) + X_{t20} \text{ } ^\circ\text{C} + X_{t28} \text{ } ^\circ\text{C}-20 \text{ } ^\circ\text{C} + X_{t35} \text{ } ^\circ\text{C}-20 \text{ } ^\circ\text{C} + X_{t42} \text{ } ^\circ\text{C}-20 \text{ } ^\circ\text{C} + E \quad (1)$$

After ANOVA decomposition, the four effect matrices were analysed

with PCA, with mean-centering and scaling using their standard deviation as selected pre-processing methods (see Section 5.1). To describe the effect of storage time at 20 °C, a model with 9 PCs was built which explained 93.5% of the total variance. PCA models with 9, 6 and 7 PCs were selected to describe the effect of storage at 28, 35 and 42 °C compared to 20 °C, respectively. These models explained 96.6%, 94.4% and 95.9% of the variance in the data. The same evaluation criteria (explained variance, eigenvalues) as described in Section 5.1 for PCA were applied for the selection of PCs in the ASCA model. For all the effect factors, a high number of PCs was selected. As an example, score and

loading plots of the PCA submodel $X_{t_{20} \text{ } ^\circ\text{C}}$ are shown in Fig. 5A-D and a PC 1-PC 2 score plot to show the effect of storage at 35 °C compared to 20 °C (Fig. 5E).

In the submodel $X_{t_{20} \text{ } ^\circ\text{C}}$, PC 1 (28.4%) and PC 2 (25.1%) explained together 53.5% of the variance. The score plot of PC 1 (Fig. 5A) showed that the scores of the different samples are increasing as a function of storage time for all storage temperatures. The score plot of PC 2 (Fig. 5B) shows the opposite trend, a decrease of the scores as a function of storage time for all temperatures. In other words, the volatile fraction is clearly evolving during storage at all temperatures. The latter can also be

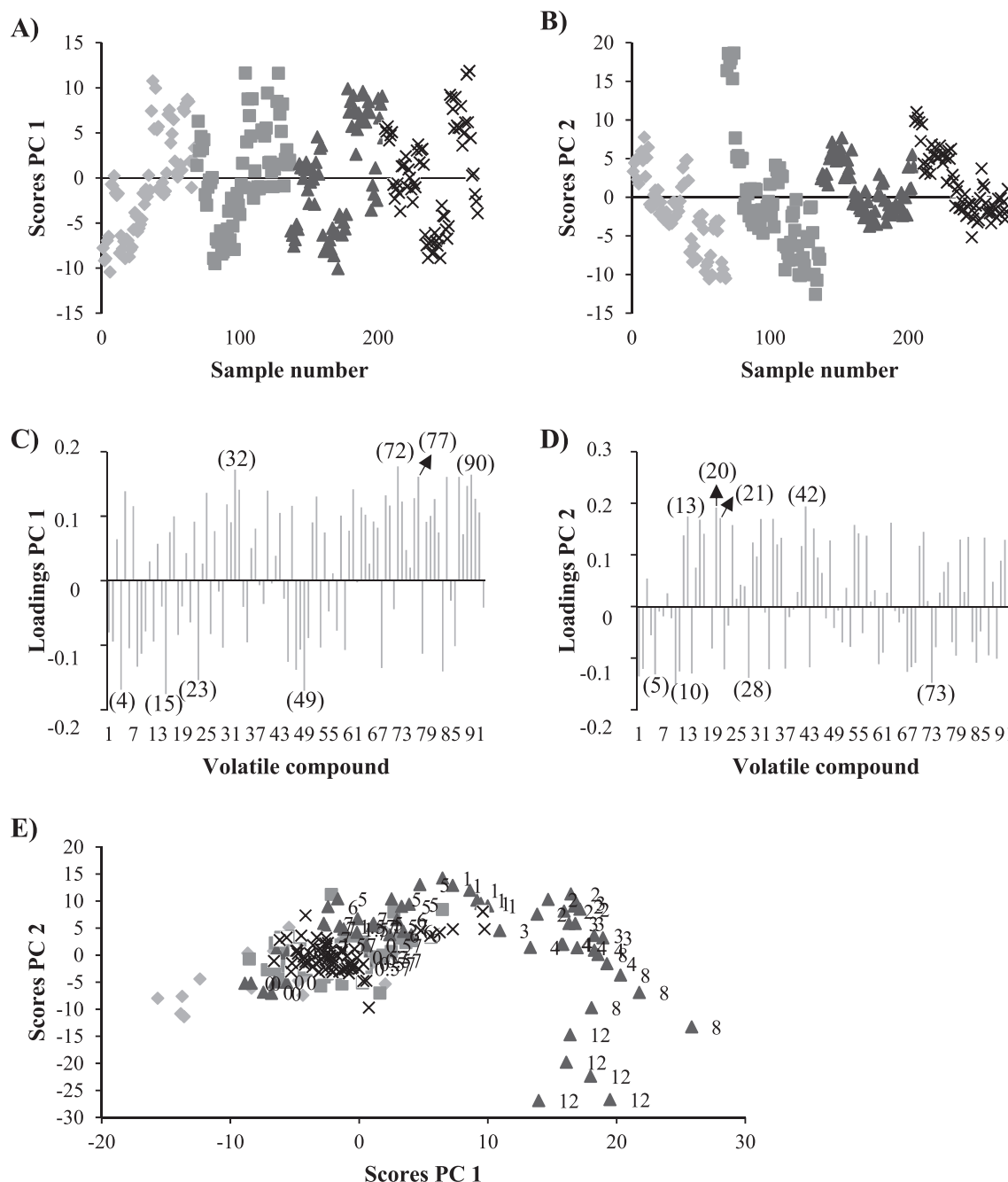


Fig. 5. (A-B) Score plots of PC 1 and PC 2 describing the common effect of storage time at 20 °C, calculated over all temperatures (◆ 20 °C, ■ 28 °C, ▲ 35 °C, × 42 °C) and ordered per storage temperature. An increasing sample number corresponds to an increase in storage time. (C-D) Loading plots of PC 1 and PC 2 describing which volatiles are most affected by storage at 20 °C. Volatile compounds with the highest and lowest loadings were identified as (4) Carbon disulfide, (5) Dimethyl sulfide, (10) Isopropyl acetate, (13) Methyl butanoate, (15) Isobutyl acetate, (20) Butyl acetate, (21) Hexanal, (23) 2-methylbutyl acetate, (28) Ethyl butanoate, (32) Limonene, (42) Hexyl acetate, (49) 2-hexenyl acetate, (72) 1-octanol, (73) Unidentified, (77) butanoic acid, (90) Hexanoic acid. (E) PC 1-PC 2 score plot showing the additional effect of storage at 35 °C compared to 20 °C. Samples of all storage temperatures are shown in the plot. For the samples stored at 35 °C (▲), the storage times are indicated in weeks.

derived from the loading plots in Fig. 5C-D. More specifically, due to the trends in the scores of both PCs, a positive loading for PC 1 is related to an increase in concentration as a function of storage time and *vice versa*. In contrast, a positive loading for PC 2 represents a decrease during storage, while a negative loading represents an increase during storage. Volatile compounds which were affected the most by storage should have the most positive or negative loadings. Based on the different score and loading plots for all PCs (data not shown), it was noticed only for PC 1 and 2 the effect of storage time is clear. Next to this, the effect of time is not exclusively explained by one of these PCs. This was also the case for the PARAFAC model where the two components described the effect of storage time (Section 5.2 and Fig. 3).

To describe the accelerated effect of storage time at higher temperatures, three additional factors were included in Eq. (1). As an example, the results of the submodel $t_{35^{\circ}\text{C}-20^{\circ}\text{C}}$ are shown in the PC 1-PC 2 score plot in Fig. 5E. These PCs explained 60.7% of the variance (40.4% by PC 1 and 20.3% by PC 2). From Fig. 5E, it can be concluded that there was an additional effect of storage at 35 °C on the volatile fraction compared to ambient conditions, because the 35 °C stored samples were separated from the other storage temperatures. In addition, a clear effect of storage time could be observed in the score plot. Further research to correctly implement and optimise the model and interpret the data is recommended and will be a goal for future work.

5.5. Comparison of the output of different techniques to study volatile changes of strawberry juice during storage

Based on the results of the different methods, it can be seen that some common trends could be detected in the output of the different models (e.g. clear change of the volatile fraction during storage, effect of storage temperature). All applied methods showed that there was a clear effect of storage time on the volatile composition. PLS better described the effect of time than PCA thanks to the inclusion of the Y-variable to model the data. PARAFAC and ASCA also showed that there was a significant effect of storage temperature on volatile changes. Moreover, PARAFAC and ASCA showed that there was a different effect of time for each storage temperature. The latter shows the advantage of analysing the whole (multi-way) data set.

6. Points of attention when performing a multivariate data analysis

Each of the above described methods has certain advantages and disadvantages which can be a reason to apply it or not. The points addressed below should be kept in mind regardless of the method that has been selected.

6.1. Data pre-processing

Adequate data pre-processing prior to the actual data analysis is of high importance as it will have a large impact on the data analysis itself and the final model output. Data pre-processing removes undesired and irrelevant random or systematic variation (e.g. instrumental artefacts, inhomogeneity of biological sample) in the data set. Data pre-processing can be seen as a transformation of the raw data into 'cleaned' data to make them more suitable for the actual data analysis (Engel et al., 2013; Nicolaï et al., 2007; Nunes et al., 2015; Pierce, Kehimkar, Marney, Hoggard, & Synovec, 2012; Roberts & Cozzolino, 2016; Saeys et al., 2019). Several pre-processing methods can be found in literature. Some examples of pre-processing techniques are mean-centering, standardisation, scaling, baseline correction, alignment (Engel et al., 2013; Khakimov et al., 2015; Roberts & Cozzolino, 2016; Skov et al., 2014). Which pre-processing method to select depends on several factors, such as the characteristics of the experimental data, the purpose of the analysis, etc. Unfortunately, there are no unequivocal guidelines which data pre-processing method to select in which case. In addition, it should

be noted that it is not straightforward to select to appropriate pre-processing method to reach your objective (Engel et al., 2013). The reader is advised to check literature (Bro, 1997; Engel et al., 2013; Khakimov et al., 2015; Nicolaï et al., 2007; Roberts & Cozzolino, 2016; Saeys et al., 2019; Skov et al., 2014; Timmerman et al., 2015; Wold et al., 1987; 2001) to look for the details of the existing pre-processing methods which are compatible with the method that was selected in order to look for the most appropriate method. Pre-processing is beyond the scope of this paper, but PCA is a versatile technique to evaluate and compare the consequences of different pre-processing techniques and settings.

6.2. Selection of model variables/factors

A food scientist should properly select the number of PCs/LVs/factors of the model. Selection of too many PCs/LVs/factors will result in overfitting and introduces noise in the model. In contrast, when the number of PCs/LVs/factors is too small, important information can be lost and the predictive power can be low (if applicable). Different criteria to select the most appropriate number of new variables were found in literature. Jolliffe (2002) and Bro and Smilde (2014) gave several suggestions to decide how many PCs should be selected to adequately describe the data with PCA: (1) a percentage of total variance that should be explained by the model can be selected beforehand, (2) a PC should be included in the model if it explains more than a certain percentage of variance (Kaiser's rule), (3) PCs with eigenvalues larger than 1 should be selected (the eigenvalues are linked to the singular value decomposition, see Supplement 2), (4) make a scree plot of the eigenvalues (or their logarithm) as a function of the number of PCs and look for an 'elbow' in the plot (i.e. an indication of a change in the slope and importance of the PCs in the model), (5) check the model with the lowest Root Mean Square Error of Cross Validation (RMSECV) (only valid when cross-validation was performed; during cross-validation, part of the data set is used to build the model and the other part is used to test the performance of the model). For PLS regression, the same rules can be applied, next to an inspection of the residuals (Abdi & Williams, 2010; Luis et al., 2004; Wold et al., 2001). Selection of the appropriate number of PARAFAC components can be based on a split-half analysis, the application of cross-validation, a visual check of the residuals, a check of the percentage of explained variance, etc. (Smilde et al., 2004). It is important to notice that variable/factor selection is not an easy task and very often the methods described above do not result in an unequivocal conclusion. Consequently, it is advised to not rely on one method, but to critically revise the different outcomes to select the most appropriate number of PCs/LVs/factors. In addition, for all methods, in the best case, the selected PCs/LVs/factors reflect real chemistry and are interpretable.

6.3. Selection of characteristic sample variables

Researchers are often interested in the attributes or compounds (X-variables) which are most characteristic/important to describe the observed trends in the data. For instance, to know which attributes are causing the grouping of samples or to select the compounds which are changing the most during storage.

Depending on the method, several strategies for sample variable selection can be applied. Irrespective of the method, one can look at the loadings as these are related to the different attributes which were analysed. If the PC/LV/factor describing the effect of interest is identified, the loadings related to this PC/LV/factor can be ranked. The variables with the highest loadings in absolute value are contributing the most to this PC/LV/factor and are the most important to describe the trend in the data set (Guo et al., 2002; Idborg, Edlund, & Jacobsson, 2004). It should be noted that ranking of loadings is only relevant in case all variables have the same variance. This can be achieved by proper pre-processing (e.g. scaling of the variables using the standard deviation as a

weighting factor) (Jolliffe, 2002). The use of loadings to select the most important variables is interesting, but it is not straightforward which cut-off value to select to distinguish between more and less important variables.

Many sample variable selection methods are available to be used for PLS regression: Variable Importance in Projection (VIP) scores, Variable Identification (VID) coefficients, selectivity ratio, interval PLS, sparse PLS, etc. (Farrés, Platikanov, Tsakovski, & Tauler, 2015; Lê Cao, Rossouw, Robert-Granié, & Besse, 2008; Nørgaard et al., 2000; Ooms, 1996; Wold, Johansson, & Cocchi, 1993). Selected volatile compounds using the VID criterion based on a PLS model for the case study presented in Section 5.3 can be found in the work of Buvé et al. (2018).

In general, it should be noted that sample variable selection methods for PARAFAC and ASCA are not (well) documented in literature. Yet, it should be investigated whether the selection techniques for PCA can be extended to PARAFAC. To the best of our knowledge, no methods have been explicitly proposed in literature for ASCA. However, as ASCA is linked to PCA, we are convinced that the methods described for PCA can be modified to be applicable for ASCA. Examples of possible important volatile markers based on the loadings of an ASCA model are included in Fig. 5.

6.4. Other points of attention

The application of MVDA can open a new world for a food scientist. Besides the points discussed above, there are more things to be aware of when applying MVDA as already stressed by Roberts and Cozzolino (2016). First of all, the quality of the experimental data is of utmost importance as a model can only be as good as the underlying data. Related to this, all steps that can have an effect on the generated data (processing steps, sampling, etc.) should be taken into account. In addition, the total number of samples and replicates included in the analysis will have a large effect on the model output. The number of replications is depending on the uncertainty and the reproducibility of the measurement. In the strawberry juice example, six independent replications were used. Unfortunately, there are no general rules about the number of replicates in a multivariate analysis. As a general rule, it is advised to have as many replications as possible and especially if the sampling and measurement uncertainty is high (Roberts & Cozzolino, 2016).

Furthermore, it is important to identify outliers in the data set. An outlier can be defined as an observation which is behaving differently compared to the others. Outliers could be related to the experimental work (problems during sampling or extraction, problems with the equipment, etc.), errors during copying or writing of the data, etc. If an outlier is present in the data set, it will have a large effect on the variance of the variables and thereby the model output. Unsupervised methods (like PCA) can be used for this purpose. The reader is referred to the work of Rousseeuw, Debruyne, Engelen, and Hubert (2006) to learn more about outlier detection for MVDA. In addition, more robust MVDA methods can be used such as ROBPCA (Hubert, Rousseeuw, & Vanden Branden, 2005), RSIMPLS (Hubert & Branden, 2003) or MCW-PLS (Fonseca Diaz, De Ketelaere, Aernouts, & Saeys, 2020).

When building a model for predictive purposes (e.g. PLS), it is important to include both model calibration and validation. During model calibration, one tries to build a model which allows to make accurate predictions. For this reason, the calibration data set should contain a large number of data. In addition, model validation checks how well the model is performing in predicting new samples (Roberts & Cozzolino, 2016; Saeys et al., 2019).

It should be noted that all selected methods linearly transform the data to reduce the dimensionality of the original data set. In case non-linear relationships are expected, the data should be pre-processed properly (e.g. log transformation) to solve this problem (Abdi & Williams, 2010; Jolliffe, 2002; van den Berg et al., 2006; Wold et al., 1987) before moving to non-linear methods (Saeys et al., 2019).

7. Concluding remarks

The overall quality of a food product is the result of an interplay of multiple attributes. More than ever, food scientists are using advanced and multivariate instruments to study food quality characteristics and changes. The use of these approaches demands the application of proper data analysis techniques to extract relevant and conclusive information from these large data sets. In this context, MVDA or chemometrics should be the first choice. This review paper gives an overview of a number of tools to analyse multivariate data with a specific focus on food quality aspects. The most suitable technique depends on the nature of the data set and the objective of the experiment. In brief, a researcher first needs to identify the dimensions of the data set. If a two-dimensional (multivariate) data set is obtained, basic MVDA can be applied. For multi-way data sets, one can use techniques developed for such data sets or the data can be unfolded to a series of two-way data sets and analysed with MVDA. In a next step, the researcher needs to decide whether unsupervised (data exploration) or supervised (prediction) methods will be used and what type of data pre-processing is recommended. After modelling, the outcome should be critically evaluated. In addition, evaluation of the output of different methods can result in a complementary understanding of the data.

PCA is still the most common method to analyse multivariate data. Although PCA is useful for food quality investigations (e.g. to detect trends and outliers), other methods may be more suitable as a second step depending on the aim. For predictive and classification purposes, supervised methods like PLS regression (for continuous Y-variables) or PLS-DA (for categorical Y-variables) are preferred. Multi-way data analysis techniques such as PARAFAC (unsupervised) may evoke the feeling to be more complex compared to MVDA techniques such as PCA or PLS. However, when dealing with multi-way data they can result in new insights. If the aim is to assign the variation in a data set to the design factors, the application of ASCA can be very useful.

Different MVDA methods have already been applied in food research. It is interesting to notice that researchers often interpret the output and graphs of a multivariate analysis without having knowledge on the statistics or mathematics behind the technique. In addition, food scientists are not always aware of the critical points of a method or the different steps the data went through. Hence, this shows the need for interdisciplinary collaborations between food scientists and statisticians in order to use the data analysis methods properly and interpret the outcome correctly.

These methods can be extended to nearly all multivariate and multi-way problems in food science. Further exploring and optimising the use of MVDA for food quality investigations should be encouraged. In addition, the application of these methods should go beyond the interpretation of the graphs. Having basic knowledge about the mathematical and statistical background of the method as well as reflecting on the characteristics of the data set should be kept in mind. Progress in other research areas (analytical chemistry, spectroscopy, metabolomics, medical imaging, geology, cancer diagnostics) can be a starting point to further implement these advanced data analysis techniques in the area of food science.

8. Funding sources

This work was supported by the Research Foundation Flanders (FWO) [Project G.OA76.15N]. Bram Neckebroek has a doctoral scholarship of the Research Foundation Flanders (FWO) [grant 1S14717N].

CRedit authorship contribution statement

Carolien Buvé: Conceptualization, Investigation, Formal analysis, Writing – original draft. **Wouter Saeys:** Formal analysis, Writing – review & editing. **Morten Arendt Rasmussen:** Writing – review & editing. **Bram Neckebroek:** Investigation, Formal analysis,

Conceptualization. **Marc Hendrickx**: Conceptualization. **Tara Grauwet**: Conceptualization, Supervision, Writing – review & editing. **Ann Van Loey**: Conceptualization, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.foodres.2021.110878>.

References

- Abdi, H., 2003. Partial Least Squares (PLS) Regression. In: *Encyclopedia for research methods for the social sciences* (pp. 792–795). 10.4135/9781412950589.n690.
- Abdi, H. (2010). Partial least squares regression. Projection on latent structures regression (PLS Regression). *WIREs Computational Statistics*, 2(1), 97–106. <https://doi.org/10.1002/wics.051>
- Abdi, H., & Williams, L. (2010). Principal component analysis. *Computational Statistics*, 2 (4), 433–459. <https://doi.org/10.1002/wics.101>
- Aernouts, B., Polshin, E., Saeys, W., & Lammertyn, J. (2011). Mid-infrared spectrometry of milk for dairy metabolomics: A comparison of two sampling techniques and effect of homogenization. *Analytica Chimica Acta*, 705(1–2), 88–97. <https://doi.org/10.1016/j.aca.2011.04.018>
- Amigo, J. M., Skov, T., Bro, R., Coello, J., & Maspocho, S. (2008). Solving GC-MS problems with PARAFAC2. *TrAC - Trends in Analytical Chemistry*, 27(8), 714–725. <https://doi.org/10.1016/j.trac.2008.05.011>
- Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics*, 17(3), 166–173. [https://doi.org/10.1002/\(ISSN\)1099-128X10.1002/cem.v17:310.1002/cem.785](https://doi.org/10.1002/(ISSN)1099-128X10.1002/cem.v17:310.1002/cem.785)
- Belmonte-Sánchez, J. R., Romero-González, R., Martínez Vidal, J. L., Arrebola, F. J., & Garrido Frenich, A. (2020). 1H NMR and multi-technique data fusion as metabolomic tool for the classification of golden rums by multivariate statistical analysis. *Food Chemistry*, 317, 126363. <https://doi.org/10.1016/j.foodchem.2020.126363>
- Berrueta, L. A., Alonso-Salces, R. M., & Héberger, K. (2007). Supervised pattern recognition in food analysis. *Journal of Chromatography A*, 1158(1–2), 196–214. <https://doi.org/10.1016/j.chroma.2007.05.024>
- Bevilacqua, M., Bro, R., Marini, F., Rinnan, Å., Rasmussen, M. A., & Skov, T. (2017). Recent chemometrics advances for foodomics. *TrAC - Trends in Analytical Chemistry*, 96, 42–51. <https://doi.org/10.1016/j.trac.2017.08.011>
- Bevilacqua, M., Buccì, R., Materazzi, S., & Marini, F. (2013). Application of near infrared (NIR) spectroscopy coupled to chemometrics for dried egg-pasta characterization and egg content quantification. *Food Chemistry*, 140(4), 726–734. <https://doi.org/10.1016/j.foodchem.2012.11.018>
- Brereton, R. G. (2014). A short history of chemometrics: A personal view. *Journal of Chemometrics*, 28(10), 749–760. <https://doi.org/10.1002/cem.2633>
- Brereton, R. G. (2017). Visualizing matrices. *Journal of Chemometrics*, 31(3), 1–3. <https://doi.org/10.1002/cem.2834>
- Brereton, R. G., & Lloyd, G. R. (2014). Partial least squares discriminant analysis: Taking the magic away. *Journal of Chemometrics*, 28(4), 213–225. <https://doi.org/10.1002/cem.2609>
- Brereton, R. G., Jansen, J., Lopes, J., Marini, F., Pomerantsev, A., Rodionova, O., ... Tauler, R. (2017). Chemometrics in analytical chemistry—part I: History, experimental design and data analysis tools. *Analytical and Bioanalytical Chemistry*, 409(25), 5891–5899. <https://doi.org/10.1007/s00216-017-0517-1>
- Brereton, R. G., Jansen, J., Lopes, J., Marini, F., Pomerantsev, A., Rodionova, O., ... Tauler, R. (2018). Chemometrics in analytical chemistry—part II: Modeling, validation, and applications. *Analytical and Bioanalytical Chemistry*, 410(26), 6691–6704. <https://doi.org/10.1007/s00216-018-1283-4>
- Bro, R. (1997). PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, 38(2), 149–171. [https://doi.org/10.1016/S0169-7439\(97\)00032-4](https://doi.org/10.1016/S0169-7439(97)00032-4)
- Bro, R., & Kiers, H. A. L. (2003). A new efficient method for determining the number of components in PARAFAC models. *Journal of Chemometrics*, 17(5), 274–286. [https://doi.org/10.1002/\(ISSN\)1099-128X10.1002/cem.v17:510.1002/cem.801](https://doi.org/10.1002/(ISSN)1099-128X10.1002/cem.v17:510.1002/cem.801)
- Bro, R., & Smilde, A. K. (2014). Principal component analysis. *Analytical Methods*, 6(9), 2812–2831. <https://doi.org/10.1039/c3ay41907j>
- Buvé, C., Neckebroek, B., Haenen, A., Kebede, B., Hendrickx, M., Grauwet, T., & Van Loey, A. (2018). Combining untargeted, targeted and sensory data to investigate the impact of storage on food volatiles: A case study on strawberry juice. *Food Research International*, 113, 382–391. <https://doi.org/10.1016/j.foodres.2018.07.022>
- Callejón, R. M., Amigo, J. M., Pairo, E., Garmón, S., Ocaña, J. A., & Morales, M. L. (2012). Classification of Sherry vinegars by combining multidimensional fluorescence, parafac and different classification approaches. *Talanta*, 88, 456–462. <https://doi.org/10.1016/j.talanta.2011.11.014>
- Camo Software, A. (2006). The Unscrambler Methods. [http://www.camo.com/downloads/U9.6/pdf/manual/The Unscrambler Methods.pdf](http://www.camo.com/downloads/U9.6/pdf/manual/The%20Unscrambler%20Methods.pdf) Accessed on 14 April 2020.
- Cevallos-Cevallos, J. M., Reyes-De-Corcuera, J. I., Etxeberria, E., Danyluk, M. D., & Rodrick, G. E. (2009). Metabolomic analysis in food science: A review. *Trends in Food Science & Technology*, 20(11–12), 557–566. <https://doi.org/10.1016/j.tifs.2009.07.002>
- Cynkar, W., Damberg, R., Smith, P., & Cozzolino, D. (2010). Classification of Tempranillo wines according to geographic origin: Combination of mass spectrometry based electronic nose and chemometrics. *Analytica Chimica Acta*, 660 (1–2), 227–231. <https://doi.org/10.1016/j.aca.2009.09.030>
- Detmer, K., Aronov, P. A., & Hammock, B. D. (2007). Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, 26(1), 51–78. <https://doi.org/10.1002/mas.v26:110.1002/mas.20108>
- Elcoroaristizabal, S., Callejón, R. M., Amigo, J. M., Ocaña-González, J. A., Morales, M. L., & Ubeda, C. (2016). Fluorescence excitation-emission matrix spectroscopy as a tool for determining quality of sparkling wines. *Food Chemistry*, 206, 284–290. <https://doi.org/10.1016/j.foodchem.2016.03.037>
- Engel, J., Gerretzen, J., Szymańska, E., Jansen, J. J., Downey, G., Blanchet, L., & Buydens, L. M. C. (2013). Breaking with trends in pre-processing? TrAC - Trends in Analytical Chemistry, 50, 96–106. <https://doi.org/10.1016/j.trac.2013.04.015>
- Farrés, M., Platikanov, S., Tsakovski, S., & Tauler, R. (2015). Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation. *Journal of Chemometrics*, 29(10), 528–536. <https://doi.org/10.1002/cem.v29:1010.1002/cem.2736>
- Fonseca Diaz, V., De Ketelaere, B., Aernouts, B., & Saeys, W. (2020). Robustness control in bilinear modeling based on maximum correntropy. *Journal of Chemometrics*, 34(4), 1–16. <https://doi.org/10.1002/cem.3215>
- Gendrin, C., Roggo, Y., & Collet, C. (2007). Content uniformity of pharmaceutical solid dosage forms by near infrared hyperspectral imaging: A feasibility study. *Talanta*, 73 (4), 733–741. <https://doi.org/10.1016/j.talanta.2007.04.054>
- Gerretzen, J., Szymańska, E., Jansen, J. J., Bart, J., van Manen, H.-J., van den Heuvel, E. R., & Buydens, L. M. C. (2015). Simple and Effective Way for Data Preprocessing Selection Based on Design of Experiments. *Analytical Chemistry*, 87 (24), 12096–12103. <https://doi.org/10.1021/acs.analchem.5b0283210.1021/acs.analchem.5b02832.s001>
- Golaszewski, R., Sims, C. A., O'Keefe, S. F., Braddock, R. J., & Littell, R. C. (1998). Sensory attributes and volatile components of stored strawberry juice. *Journal of Food Science*, 63(4), 734–738. <https://doi.org/10.1111/j.1365-2621.1998.tb15825.x>
- Granato, D., de Araújo Calado, V. O. M., & Jarvis, B. (2014). Observations on the use of statistical methods in Food Science and Technology. *Food Research International*, 55, 137–149. <https://doi.org/10.1016/j.foodres.2013.10.024>
- Granato, D., Putnik, P., Kovačević, D. B., Santos, J. S., Calado, V., Rocha, R. S., ... Pomerantsev, A. (2018). Trends in chemometrics: Food authentication, microbiology, and effects of processing. *Comprehensive Reviews in Food Science and Food Safety*, 17(3), 663–677. <https://doi.org/10.1111/crf3.2018.17.issue-310.1111/1541-4337.12341>
- Grauwet, T., Vervoort, L., Colle, I., Van Loey, A., & Hendrickx, M. (2014). From fingerprinting to kinetics in evaluating food quality changes. *Trends in Biotechnology*, 32(3), 125–131. <https://doi.org/10.1016/j.tibtech.2014.01.002>
- Gromski, P. S., Muhamadali, H., Ellis, D. I., Xu, Y., Correa, E., Turner, M. L., & Goodacre, R. (2015). A tutorial review: Metabolomics and partial least squares-discriminant analysis - a marriage of convenience or a shotgun wedding. *Analytica Chimica Acta*, 879, 10–23. <https://doi.org/10.1016/j.aca.2015.02.012>
- Guo, Q., Wu, W., Massart, D., Boucon, C., & de Jong, S. (2002). Feature selection in principal component analysis of analytical data. *Chemometrics and Intelligent Laboratory Systems*, 61(1–2), 123–132. [https://doi.org/10.1016/S0169-7439\(01\)00203-9](https://doi.org/10.1016/S0169-7439(01)00203-9)
- Harshman, R. A. (1970). Foundations of the PARAFAC procedure: Models and conditions for an “experimental” multimodal factor analysis. *UCLA Working Papers in Phonetics*, 16(10), 1–84.
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(7), 498–520.
- Hubert, M., & Branden, K. V. (2003). Robust methods for partial least squares regression. *Journal of Chemometrics*, 17(10), 537–549. [https://doi.org/10.1002/\(ISSN\)1099-128X10.1002/cem.v17:1010.1002/cem.822](https://doi.org/10.1002/(ISSN)1099-128X10.1002/cem.v17:1010.1002/cem.822)
- Hubert, M., Rousseeuw, P. J., & Vanden Branden, K. (2005). ROBPCA: A new approach to robust principal component analysis. *Technometrics*, 47(1), 64–79. <https://doi.org/10.1198/004017004000000563>
- Idborg, H., Edlund, P.-O., & Jacobsson, S. P. (2004). Multivariate approaches for efficient detection of potential metabolites from liquid chromatography/mass spectrometry data. *Rapid Communications in Mass Spectrometry*, 18(9), 944–954. [https://doi.org/10.1002/\(ISSN\)1097-023110.1002/rcm.v18:910.1002/rcm.1432](https://doi.org/10.1002/(ISSN)1097-023110.1002/rcm.v18:910.1002/rcm.1432)
- Jansen, J. J., Bro, R., Hoefsloot, H. C. J., van den Berg, F. W. J., Westerhuis, J. A., & Smilde, A. K. (2008). PARAFASCA: ASCA combined with PARAFAC for the analysis of metabolic fingerprinting data. *Journal of Chemometrics*, 22(2), 114–121. [https://doi.org/10.1002/\(ISSN\)1099-128X10.1002/cem.v22:210.1002/cem.1105](https://doi.org/10.1002/(ISSN)1099-128X10.1002/cem.v22:210.1002/cem.1105)
- Jansen, J. J., Hoefsloot, H. C. J., van der Greef, J., Timmerman, M. E., Westerhuis, J. A., & Smilde, A. K. (2005). ASCA: Analysis of multivariate data obtained from an experimental design. *Journal of Chemometrics*, 19(9), 469–481. [https://doi.org/10.1002/\(ISSN\)1099-128X10.1002/cem.v19:910.1002/cem.952](https://doi.org/10.1002/(ISSN)1099-128X10.1002/cem.v19:910.1002/cem.952)
- Jolliffe, I. T. (2002). Principal component analysis, second ed. In *Principle Component Analysis*. 10.2307/1270093.
- Kang, X., Zhao, Y., Shang, D., Zhai, Y., Ning, J., Ding, H., & Sheng, X. (2020). Identification of the geographical origins of sea cucumbers in China: The application

- of stable isotope ratios and compositions of C, N, O and H. *Food Control*, 111, 107036. <https://doi.org/10.1016/j.foodcont.2019.107036>
- Kebede, B. T., Grauwet, T., Mutsokoti, L., Palmers, S., Vervoort, L., Hendrickx, M., & Van Loey, A. (2014). Comparing the impact of high pressure high temperature and thermal sterilization on the volatile fingerprint of onion, potato, pumpkin and red beet. *Food Research International*, 56, 218–225. <https://doi.org/10.1016/j.foodres.2013.12.034>
- Khakimov, B., Gürdeniz, G., & Engelsens, S. B. (2015). Trends in the application of chemometrics to foodomics studies. *Acta Alimentaria*, 44(1), 4–31. <https://doi.org/10.1556/AAlim.44.2015.1.1>
- Kiers, H. A. L. (2000). Towards a standardized notation and terminology in multiway analysis. *Journal of Chemometrics*, 14(3), 105–122. [https://doi.org/10.1002/\(ISSN\)1099-128X10.1002/1099-128X\(200005/06\)14:3<105::AID-CEM582>3.0.CO;2-I](https://doi.org/10.1002/(ISSN)1099-128X10.1002/1099-128X(200005/06)14:3<105::AID-CEM582>3.0.CO;2-I)
- Kim, N. S., Lee, J. H., Han, K. M., Kim, J. W., Cho, S., & Kim, J. (2014). Discrimination of commercial cheeses from fatty acid profiles and phytosterol contents obtained by GC and PCA. *Food Chemistry*, 143, 40–47. <https://doi.org/10.1016/j.foodchem.2013.07.083>
- Knowledge Centre for Food Fraud and Quality (2021). *Food Quality*. Retrieved from http://s://knowledge4policy.ec.europa.eu/food-fraud-quality/topic/food-quality_en. Accessed on October 4, 2021.
- Kumar, N., Bansal, A., Sarma, G. S., & Rawal, R. K. (2014). Chemometrics tools used in analytical chemistry: An overview. *Talanta*, 123, 186–199. <https://doi.org/10.1016/j.talanta.2014.02.003>
- Lê Cao, K. A., Rossouw, D., Robert-Granié, C., & Besse, P. (2008). A sparse PLS for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*, 7(1). <https://doi.org/10.2202/1544-6115.1390>
- Lenhardt, L., Zeković, I., Dramićanin, T., Miličević, B., Burojević, J., & Dramićanin, M. D. (2017). Characterization of cereal flours by fluorescence spectroscopy coupled with PARAFAC. *Food Chemistry*, 229, 165–171. <https://doi.org/10.1016/j.foodchem.2017.02.070>
- López-Maestresalas, A., Keresztes, J. C., Goodarzi, M., Arazuri, S., Jarén, C., & Saeys, W. (2016). Non-destructive detection of blackspot in potatoes by Vis-NIR and SWIR hyperspectral imaging. *Food Control*, 70, 229–241. <https://doi.org/10.1016/j.foodcont.2016.06.001>
- Luis, M. L., Fraga, J. M. G., Jiménez, A. I., Jiménez, F., Hernández, O., & Arias, J. J. (2004). Application of PLS regression to fluorimetric data for the determination of furosemide and triamterene in pharmaceutical preparations and triamterene in urine. *Talanta*, 62(2), 307–316. <https://doi.org/10.1016/j.talanta.2003.07.010>
- Maitra, S., & Yan, J. (2008). Principle component analysis and partial least squares: Two dimension reduction techniques for regression. *Casualty Actuarial Society*, 79–90.
- Mehmood, T., Sæbø, S., & Liland, K. H. (2020). Comparison of variable selection methods in partial least squares regression. *Journal of Chemometrics*, 34(6), 1–14.
- Nascimento, P. A. M., Carvalho, L. C. de, Júnior, L. C. C., Pereira, F. M. V., & Teixeira, G. H. de A. (2016). Robust PLS models for soluble solids content and firmness determination in low chilling peach using near-infrared spectroscopy (NIR). *Postharvest Biology and Technology*, 111, 345–351. <https://doi.org/10.1016/j.postharvbio.2015.08.006>
- Nicolai, B. M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K. I., & Lammertyn, J. (2007). Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. *Postharvest Biology and Technology*, 46(2), 99–118. <https://doi.org/10.1016/j.postharvbio.2007.06.024>
- Noordraven, L. E. C., Buvé, C., Chen, C., Hendrickx, M. E., & Van Loey, A. M. (2021). Impact of Processing and Storage Conditions on the Volatile Profile of Whole Chickpeas (*Cicer arietinum* L.). *ACS Food Sci. Technol.* (accepted manuscript). 10.1021/acsfdsctech.1c00108.
- Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J. P., Munck, L., & Engelsens, S. B. (2000). Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy. *Applied Spectroscopy*, 54(3), 413–419. <https://doi.org/10.1366/0003702001949500>
- Nunes, C. A., Ortiz Alvarenga, V., de Souza Sant'Ana, A., Sousa Santos, J., & Granato, D. (2015). The use of statistical software in food science and technology: Advantages, limitations and misuses. *Food Research International*, 75, 270–280. <https://doi.org/10.1016/j.foodres.2015.06.011>
- Ooms, K. (1996). *Identification of potentially causal regressors in PLS models*. PhD thesis. Belgium: KU Leuven.
- Pedro, A. M. K., & Ferreira, M. M. C. (2007). Simultaneously calibrating solids, sugars and acidity of tomato products using PLS2 and NIR spectroscopy. *Analytica Chimica Acta*, 595(1–2), 221–227. <https://doi.org/10.1016/j.aca.2007.03.036>
- Pierce, K. M., Kehimkar, B., Marney, L. C., Hoggard, J. C., & Synovec, R. E. (2012). Review of chemometric analysis techniques for comprehensive two dimensional separations data. *Journal of Chromatography A*, 1255, 3–11. <https://doi.org/10.1016/j.chroma.2012.05.050>
- Roberts, J. J., & Cozzolino, D. (2016). An Overview on the application of chemometrics in food science and technology—An approach to quantitative data analysis. *Food Analytical Methods*, 9(12), 3258–3267. <https://doi.org/10.1007/s12161-016-0574-7>
- Rousseeuw, P. J., Debruyne, M., Engelen, S., & Hubert, M. (2006). Robustness and outlier detection in chemometrics. *Critical Reviews in Analytical Chemistry*, 36(3–4), 221–242. <https://doi.org/10.1080/10408340600969403>
- Roussel, S., Preys, S., Chauchard, F., & Lallemand, J. (2014). Chapter 2 Multivariate Data Analysis (Chemometrics). In C. P. O. O'Donnell (Ed.), *Process Analytical Technology for the Food Industry* (pp. 7–17). 10.1007/978-1-4939-0311-5.
- Rubio, L., Sanllorente, S., Sarabia, L. A., & Ortiz, M. C. (2019). Fluorescence determination of cochineal in strawberry jam in the presence of carmoisine as a quencher by means of four-way PARAFAC decomposition. *Food Chemistry*, 290, 178–186. <https://doi.org/10.1016/j.foodchem.2019.03.113>
- Saccenti, E., Smilde, A. K., & Camacho, J. (2018). Group-wise ANOVA simultaneous component analysis for designed omics experiments. *Metabolomics*, 14(6). <https://doi.org/10.1007/s11306-018-1369-1>
- Saeys, W., Nguyen Do Trong, N., Van Beers, R., & Nicolai, B. M. (2019). Multivariate calibration of spectroscopic sensors for postharvest quality evaluation: A review. *Postharvest Biology and Technology*, 158, 110981. <https://doi.org/10.1016/j.postharvbio.2019.110981>
- Siegmund, B., Derler, K., & Pfannhauser, W. (2001). Changes in the aroma of a strawberry drink during storage. *Journal of Agricultural and Food Chemistry*, 49(7), 3244–3252. <https://doi.org/10.1021/jf010116u>
- Skov, T., Honoré, A. H., Jensen, H. M., Næs, T., & Engelsens, S. B. (2014). Chemometrics in foodomics: Handling data structures from multiple analytical platforms. *TrAC - Trends in Analytical Chemistry*, 60, 71–79. <https://doi.org/10.1016/j.trac.2014.05.004>
- Smilde, A. K., Jansen, J. J., Hoefsloot, H. C. J., Lamers, R. J. A. N., van der Greef, J., & Timmerman, M. E. (2005). ANOVA-simultaneous component analysis (ASCA): A new tool for analyzing designed metabolomics data. *Bioinformatics*, 21(13), 3043–3048. <https://doi.org/10.1093/bioinformatics/bti476>
- Smilde, A., Bro, R., & Geladi, P. (2004). *Multi-way analysis* (first edition; A. Smilde, R. Bro, & P. Geladi, Eds.). Wiley & Sons, Inc.
- Strani, L., Grassi, S., Casiraghi, E., Alamprese, C., & Marini, F. (2019). Milk Renneting: Study of process factor influences by FT-NIR spectroscopy and chemometrics. *Food and Bioprocess Technology*, 12(6), 954–963. <https://doi.org/10.1007/s11947-019-02266-2>
- Timmerman, M. E., Hoefsloot, H. C. J., Smilde, A. K., & Ceulemans, E. (2015). Scaling in ANOVA-simultaneous component analysis. *Metabolomics*, 11(5), 1265–1276. <https://doi.org/10.1007/s11306-015-0785-8>
- van Boekel, M. A. J. S. (2008). Kinetic modeling of food quality: A critical review. *Comprehensive Reviews in Food Science and Food Safety*, 7(1), 144–158.
- van den Berg, R. A., Hoefsloot, H. C. J., Westerhuis, J. A., Smilde, A. K., & van der Werf, M. J. (2006). Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics*, 7, 1–15. <https://doi.org/10.1186/1471-2164-7-142>
- Wafula, E. N., Wainaina, I. N., Buvé, C., Nguyen, N.-D.-T., Kinyanjui, P. K., Saeys, W., ... Hendrickx, M. (2020). Application of Near-infrared Spectroscopy to Predict the Cooking Times of Aged Common Beans (*Phaseolus vulgaris* L.). *Journal of Food Engineering*, 284, 110056. <https://doi.org/10.1016/j.jfoodeng.2020.110056>
- Wibowo, S., Grauwet, T., Kebede, B. T., Hendrickx, M., & Van Loey, A. (2015). Study of chemical changes in pasteurised orange juice during shelf-life: A fingerprinting-kinetics evaluation of the volatile fraction. *Food Research International*, 75, 295–304. <https://doi.org/10.1016/j.foodres.2015.06.020>
- Wold, H. (1980). Model construction and evaluation when theoretical knowledge is scarce. In J. B. Kmenta, & J. Ramsey (Eds.), *Evaluation of Econometric Models* (pp. 47–74). Academic Press.
- Wold, S., Eriksson, L., Trygg, J., & Kettaneh, N. (2004). The PLS method – partial least squares projections to latent structures – and its applications in industrial RDP (research, development, and production). *Compstat*. <https://doi.org/10.1109/JMEMS.2011.2159097>
- Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1–3), 37–52.
- Wold, S., Johansson, E., & Cocchi, M. (1993). PLS: Partial Least Squares Projections to Latent Structures. In H. Kubinyo (Ed.), *3D QSAR in Drug Design: Theory, Methods and Applications*. (3rd edition, pp. 523–550).
- Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2), 109–130. [https://doi.org/10.1016/S0169-7439\(01\)00155-1](https://doi.org/10.1016/S0169-7439(01)00155-1)
- Zwanenburg, G., Hoefsloot, H. C. J., Westerhuis, J. A., Jansen, J. J., & Smilde, A. K. (2011). ANOVA-principal component analysis and ANOVA-simultaneous component analysis: A comparison. *Journal of Chemometrics*, 25(10), 561–567. <https://doi.org/10.1002/cem.v25.1010.1002/cem.1400>