Acta Alimentaria, Vol. 44 (1), pp. 4–31 (2015)
DOI: 10.1556/AAlim.44.2015.1.1

TRENDS IN THE APPLICATION OF CHEMOMETRICS TO FOODOMICS STUDIES

B. KHAKIMOV [a,b,1], G. GÜRDENIZ[c,1,] and S.B. ENGELSEN[a]*

[a]Department of Food Science, Faculty of Science, University of Copenhagen, Rolighedsvej 26, Frederiksberg C, 1958 Copenhagen, Denmark
[b]Department of Plant and Environmental Sciences, Faculty of Science, University of Copenhagen, Thorvaldsensvej 40, Frederiksberg C, 1871 Copenhagen, Denmark
[c]Department of Nutrition, Exercise and Sports, Faculty of Science, University of Copenhagen, Rolighedsvej 26, 1958 Frederiksberg C, Denmark

(Received: 4 September 2014; accepted: 30 October 2014)

There is an ever-increasing trend in advanced food analysis and foodomics to use more and more sophisticated analytical platforms that generate large and complex data structures, which in turn require more and more sophisticated data analysis tools for converting data into information. The choice of multivariate chemometric methods is primarily determined by the design of the study, type of the data, and the conclusions sought. In order to validate multivariate models, scientists are required to have basic chemometric knowledge and to be familiar with the variance structure of the investigated data. This review outlines some of the key aspects of applying common chemometric methods used within foodomics and provides selected examples of current applications. The review aims to provide simple insight into various multivariate methods and to illustrate pros and cons of unsupervised and supervised methods. The main analytical platforms used in foodomics are briefly discussed from the application point of view and the utilization of the generated data is illustrated. In addition, advanced data pre-processing tools, prior to multivariate analysis, are explained and relevant tools are demonstrated.

Keywords: chemometrics, food control, foodomics, nutritional metabolomics

1. Introduction

The increasing world population and the continuous climate change result in reduction of agricultural lands for food production and shortage of drinking water. Subsequently this urges modern food science to develop sustainable food production systems and improve nutritional value of food products, while keeping the cost as low as possible. Quality and nutritional value of foods are highly dependent on environment, agricultural practices, production conditions, and consumer preferences, which all may provide different effects for human health. One of the main challenges of the food science is to optimize food production to have minimum environmental footprint, lower production costs, and improving quality and nutritional value. This societal quest has brought a new multidisciplinary area into the food science, namely foodomics (Fig. 1) (CIFUENTES, 2009; CAPOZZI & BORDONI, 2013). Foodomics has been defined as a new discipline that studies food and nutritional domain via cutting-edge analytical technologies and multivariate data analysis (chemometrics).

* To whom correspondence should be addressed.
Phone: +45-20-200-064; e-mail: se@food.ku.dk
[1]: B. Khakimov and G. Gürdeniz contributed equally

0139–3006/$ 20.00 © 2015 *Akadémiai Kiadó, Budapest*

Foodomics studies generally attempt to cover broader range of more holistic research questions, such as for example how crop plants grown under drought conditions will change its chemical composition and how this in turn will affect food production and the health of the consumers. The questions behind most foodomics studies require untargeted analytical methods and the utilization of as much information as possible. Direct and indirect effects of food products on human health are studied within nutritional metabolomics that require untargeted and/or semi-targeted methods of analysis. In contrast, some of the important aspects of the food safety and food control require more targeted methods, such as for example targeted detection of food contaminants, where predefined questions can be answered by measuring relatively few compounds. The sheer size of the generated data in foodomics often becomes too big to be effectively evaluated by conventional univariate approaches and requires multivariate data analysis and pattern recognition methods that utilize all measured variables simultaneously and are able to identify underlying latent factors that carry important biological information.

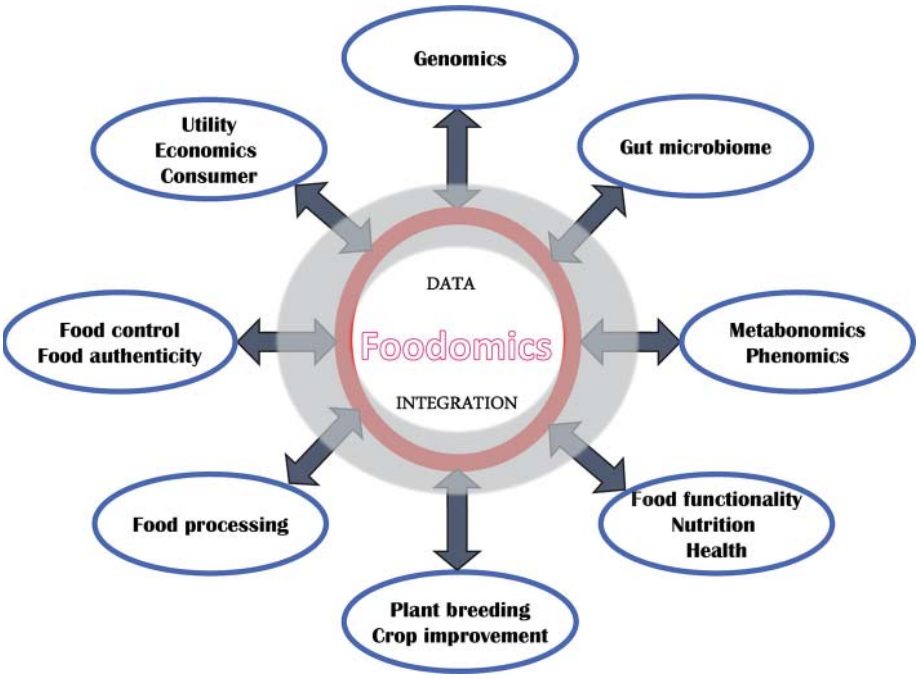

Fig. 1. Foodomics in the centre of the multidisciplinary scientific cyclone with interactions to multiple disciplines and surrounded by a cloud of multivariate chemometric data integration

Analytical platforms, such as Gas Chromatography-Mass Spectrometry (GC-MS), Liquid Chromatography-Mass Spectrometry (LC-MS), Capillary Electrophoresis–Mass Spectrometry (CE-MS), Nuclear Magnetic Resonance spectroscopy (NMR), Infrared Spectroscopy (IR), Near-Infrared Spectroscopy (NIR), and Raman spectroscopy, are frequently employed in various foodomics studies and allow semi-quantitative and quantitative detection of broad range of molecules, which in turn may provide insight into biology and/or food process related changes. Foods are complex heterogeneous systems

including complex mixtures of different molecular families and therefore analytical platforms are required to be unbiased, selective, and sensitive in order to detect the wide range of molecules that might be present at very different concentrations. Such analytical platforms often generate several thousand variables per sample and depending on the analysis mode, the data may have two, three, and more dimensional structures. Analysis of such multivariate data with highly co-linear variables requires appropriate multivariate chemometric methods.

This review focuses on pros and cons of the most commonly applied chemometric methods in foodomics. The most frequently used unsupervised methods, such as Principal Component Analysis (PCA) (Pearson, 1901; Hotelling, 1933), Hierarchical Cluster Analysis (HCA) and supervised methods like Partial Least Squares Regression (PLSR) (Wold, 1975, 1980, Wold et al., 1983), Partial Least Squares Discriminant Analysis (PLS-DA) (Ståhle & Wold, 1987), Extended Canonical Variates Analysis (ECVA) (Nørgaard et al., 2006), and Soft Independent Modelling of Class Analogy (SIMCA) (Wold & Sjöström, 1977), are discussed, and selected applications from literature are highlighted. Moreover, the utility of more advanced chemometric methods, such as ANOVA–Simultaneous Component Analysis (ASCA) (Smilde et al., 2005), sparse PCA (SPCA) (Zou et al., 2006), PARAllel FACtor Analysis (PARAFAC) (Harshman, 1970; Bro, 1997), and PARAllel FACtor Analysis 2 (PARAFAC2) (Harshman, 1972; Bro et al., 1999), will be explained in details and their advantages for exploring complex foodomics data sets are demonstrated. Alongside with short descriptions of the different chemometric methods, few examples of their use within food exploration (Munck et al., 1998), food control (Arvanitoyannis & van Houwelingen-Koukaliaroglou, 2003, Arvanitoyannis & Tzouros, 2005, Berrueta et al., 2007), GMO foods (Ahmed, 2002, Valdés et al., 2013), food adulteration (Arvanitoyannis et al., 2005a, Karoui & De Baerdemaeker, 2007), pesticide detection (Mas et al., 2010), and nutritional metabolomics (Savorani et al., 2013) will be demonstrated.

2. Analytical platforms

Foodomics studies employ various analytical platforms differing by their sensitivity, selectivity, and high-throughput capacity (Table 1), and the molecular composition of the investigated samples significantly vary by their physico-chemical properties and concentrations (Fig. 2). In order to provide a holistic evaluation of the molecular perturbations, comprehensive studies require unbiased analytical platforms covering a broad range of metabolites. Modern analytical platforms can be divided into two categories: 1) separation followed by detection techniques, e.g. GC-MS, LC-MS, CE-MS, LC-NMR, and 2) direct detection techniques, e.g. Fluorescence, Raman, IR, NIR, and NMR spectroscopy.

GC-MS is based on separation of molecules in a gas phase based on their boiling points by applying heat and vaporized molecules fly through the GC column under the steam of carrier gas, e.g. helium or hydrogen. Then, vaporized molecules reach the ionization chamber, where they will be fragmented into several characteristic *m/z* ions and subsequently separated and detected forming two-dimensional data (Fig. 3). Standardized electron ionization (EI) techniques with 70 eV energy electron beam provide unique mass spectra per chemical structure and have formed some of the richest metabolite databases, e.g. NIST and Wiley. However, GC-MS requires molecules to be thermally stable and volatile, thus samples are derivatized, where non-volatile and/or thermally unstable metabolites are chemically altered for increasing their detection (Khakimov et al., 2013). GC-MS is widely utilized in various

food control (Durante et al., 2006), foodomics (Bianchi et al., 2001), and nutritional metabolomics (Zafra-Gomez et al., 2010) applications.

Table 1. Main advantages and drawbacks of analytical platforms GC-MS, LC-MS, CE-MS, NMR, and vibrational spectroscopy

Analytical platform	Advantages	Drawbacks
GC-MS	• High chromatographic resolution • Great sensitivity towards non-polar and volatile metabolites • Rich EI-MS metabolite databases are available • Cheaper than LC-MS and lower running cost (solvent free)	• Requires sample derivatization for detection of polar and non-volatile metabolites • Only small MW metabolites can be detected (MW < 1000 Da) • Usually EI-MS does not provides mass of molecular ions
LC-MS	• Wide range of metabolites can be detected (MW < 60 kDa) • No requirements for metabolites to be volatile • Great sensitivity towards polar metabolites • Larger volume of sample can be injected • Allows metabolite purification	• Mobile phase (pH, polarity, gradient program) dependent sensitivity towards polar metabolites • Ion suppression • Difficult to ionize volatile metabolites • Expensive (especially for high mass accuracy platforms)
CE-MS	• Provides higher resolution of metabolites compared to LC • Allows separation of proteins, nucleic acids, ionic and very polar metabolites that are complicated in LC and GC • Provides highly reproducible profiles when experimental conditions are robust • Allows the analysis of heterogeneous samples	• Small volume of sample introduction (1 µl) limits sensitivity and metabolite purification • Resolution power highly depends on polarity and pH of solvent • Migration times of the same metabolites fluctuate with the changing of the environment temperature • Limitations in electrolyte selections
NMR	• Allows structure elucidation of unknown metabolites • Unbiased and inherently quantitative • Provides higher reproducibility and lower experimental error than MS based methods • Non-destructive with minimal sample preparation • Metabolite coverage is excellent	• Lower sensitivity than MS based methods • Lower selectivity than MS based methods • Signals of metabolites may be overlapped and hamper quantification • High running costs (deuterated solvents and cryogenic gasses) • Measurement speed is medium
NIR IR Raman	• Non-destructive • High reproducibility and measurement speed • Applicable in on-line measurements • Well established and validated methods are available • Metabolite coverage is excellent • No sample preparation • Allows the analysis of solid state samples	• Lower sensitivity compared to NMR and MS • Lower selectivity compared to NMR and MS • Metabolite structural information is limited to the type of functional groups, polarity

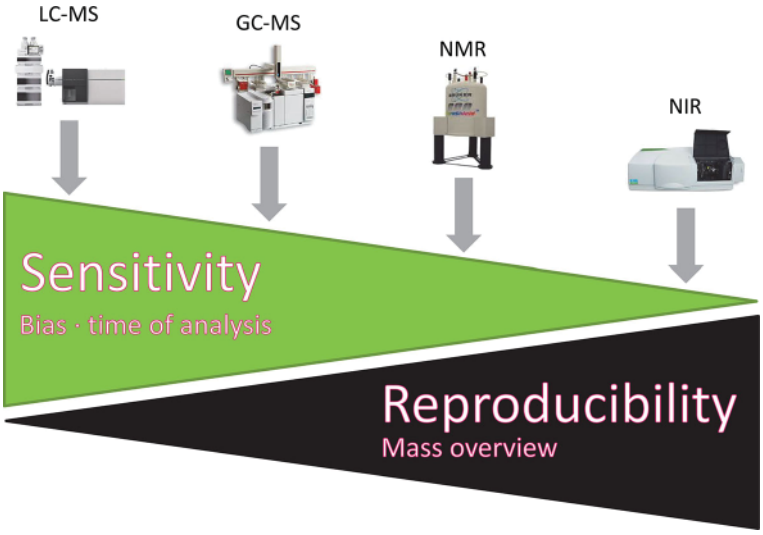

Fig. 2. Analytical platforms frequently applied in foodomics

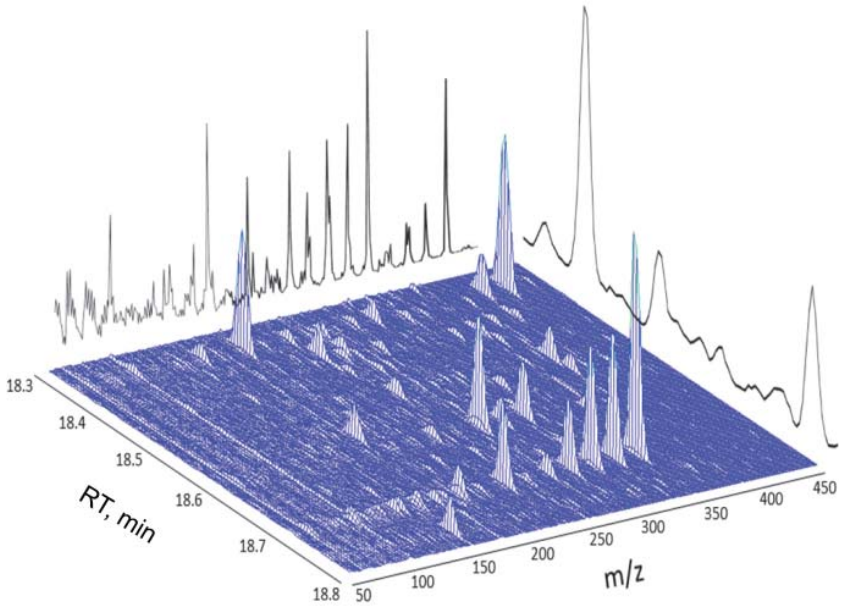

Fig. 3. A segment of the raw GC-MS data showing the three dimensional nature of the data (intensity as a function of elution time and mass)

LC-MS is a technique that allows mass spectrometric detection of metabolites that are separated in liquid phase due to their different mobile (solvent)-stationary (column) phase partitioning coefficients. Development of electrospray ionization (ESI) method has largely broadened the applications of LC-MS by offering an efficient ionization of non-volatile/polar

metabolites that cannot be ionized by the EI and minimized problems related to the LC solvent interferences.

CE-MS is one of the most versatile analytical techniques commonly applied in foodomics studies (Kolchet al., 2005; Mischak et al., 2009; Castro-Puyana et al., 2012). In CE-MS the molecules travel through a capillary tube by using an electric field and electrolytic solution and separation is based on migration rate, which depends on the molecular charge, the molecular size, and the electroosmotic flow of the solution. Resolution, sensitivity, and reproducibility of CE-MS depend on the nature of the capillary tube coating, stability of pH and temperature (Erny et al., 2006, Kasicka, 2012).

In a non-targeted set-up, the above-mentioned hyphenated platforms allow detection of up to 800 small biomolecules/metabolites of the investigated sample mixtures, and identification of up to 200 molecules at level 1, 2, and 3 according to the Metabolomics Standards Initiatives (Sumner et al., 2007). However, due to the complexity of the sample preparation and analysis protocols, the levels of experimental errors are usually higher than in the direct detection platforms.

Direct detection analytical platforms are used for measuring various physico-chemical properties of food samples as one whole system (non-destructive), but they are normally less specific to individual molecules. However, these methods provide effective quantitative analysis of the bulk primary metabolites, such as total starch content, fat content, protein content, dietary fibres, and sugars, and facilitate rapid (high throughput) and non-destructive evaluation of the samples. Vibrational spectroscopic techniques, such as IR and NIR, are considered as a fingerprint of the underlying molecular structures and measure the energies of the fundamental vibrations, such as stretching, scissoring, wagging, rocking, and twisting of functional groups in the molecules, and last but not least, the absorption is directly proportional to concentration (Lambert-Beer's law). Near infrared spectroscopy is analogous to IR spectroscopy (fundamental vibrations), but due to the shorter wavelengths and thus higher energy photons, it measures the energy absorbance due to molecular overtone and combination vibrations. As the "symphony" of overtones and combination tones usually gives rise to very complex and holistic spectra, multivariate regression methods are required to uncover the information about the overall physico-chemical state of the sample. NIR technology has a long tradition as a powerful tool for the rapid proxy evaluation of foods in food control and in process analytical technology applications (van den Berg et al., 2013).

NMR is one of the mostly used non-destructive and unbiased analysis method that provides medium rapid, highly reproducible, and quantitative detection of the broad range of metabolites that possess ^{1}H atoms or any other atom with NMR active nuclei, such as ^{31}P, ^{15}N, and ^{13}C. NMR fingerprints possess both qualitative and quantitative metabolomic information that easily can be turned into biological information with the help of multivariate data analysis (Engelsen et al., 2013). NMR has lower sensitivity compared to hyphenated platforms and it allows the detection of the most abundant metabolites (e.g. first 50 metabolites) of sample mixture, while low concentration metabolites remain undetected.

3. Data pre-processing prior to chemometric analysis

Chemometric data analysis methods are sensitive to the noise and other non-sample related variations that might hide the true biological information and/or lead to misinterpretations. In foodomics, the level of non-sample related variations depends on complexity of protocols,

and data acquired using direct detection platforms possess less artificial variation than hyphenated platforms. In order to reduce such variations, complex biological samples are spiked with an internal standard and/or control samples are run at frequent intervals throughout the analysis. Data normalization to these standard samples significantly reduces the experimental and instrumental variation to a certain extent. However, several other variations, e.g. retention time shift, chemical shift inconsistence, baseline drift, and occurrence of artificial peaks, might significantly hamper extraction of data. Thus pre-processing of raw foodomics data is often required prior to the application of chemometrics. Most often foodomics data are pre-processed to remove the baseline, reduce noise and align peaks. This section provides an overview to the different pre-processing tools used with GC-MS, LC-MS, NMR and NIR data.

GC-MS. Non-specialist analysts need user-friendly methods for quantification and identification of resolved or partially resolved GC-MS peaks from the complex raw data. This is for a large part possible by using commercial chromatographic data processing software, such as ChemStation and Mass Profiler (Agilent), DataAnalysis (Bruker), and ChromaTOF (LECO). However, it becomes complex and laborious when dealing with profiles with several overlapped and/or closely eluted peaks that make reliable and high-throughput quantification and identification difficult. Although automatic quantification of resolved and overlapped peaks is possible using such software, it may not be reliable due to the inconsistencies in retention times of peaks, changes of peak shape, and omission of peaks having a low s/n ratio. One of the most utilized GC-MS data processing software, Automated Mass Spectral Deconvolution and Identification System (AMDIS) (Stein, 1999), allows automatic deconvolution of mass spectra from complex profiles and compares with NIST database. However, the technique requires validation of deconvoluted mass spectra, since the level of false positive results is high. Recently a similar method to AMDIS, GAVIN was developed as a freely available software (Behrends et al., 2011). Another approach is based on multivariate curve resolution (MCR) (Lawton & Sylvestre, 1971; de Juan & Tauler, 2006) that decomposes three-way GC-MS data and allow spectral deconvolution and estimation of metabolite concentrations via modelling (Hantao et al., 2012).

Comprehensive GC-MS foodomics requires high-throughput and reliable pre-processing methods, which will allow automated and/or semi-automated deconvolution of mass spectra, baseline correction, RT alignment, metabolite quantification, and identification (Amigo et al., 2010b, Khakimov et al., 2014). While AMDIS and MCR can handle only one sample at a time, new technology based on multi-way decomposition modelling, PARAFAC2 performs the same task in a more efficient and robust manner by extracting the same features across all samples. However, this approach also has some disadvantages related to its use by non-specialist users. PARAFAC2 based chromatographic data processing requires division of the data (e.g. Elution time × Mass spectra × Samples) into smaller (less complex) intervals in elution time dimension in order to reduce model complexicity and improve model validation (Bro et al., 1999; Amigo et al., 2010a; Khakimov et al., 2012). An example of processing raw GC-MS data interval using PARAFAC2 is illustrated in Fig. 4. PARAFAC2 will be discussed in more detail in section 5.4.4.

LC-MS. Pre-processing of LC-MS data is complex and a number of good conducting habits and *tricks of the trade* is given in Skov and Engelsen (2013). The LC-MS data pre-processing involves conversion of the complex raw data into a simple metabolite table by

extracting informative characteristics of each detected ion in order to proceed with data analysis methods for interpretation. The informative characteristics include *m/z*, retention time of the ion (usually referred as a 'feature'), and intensity measurements (height or area). In LC-MS based foodomics studies, thousands of peaks can be detected and usually the aim is to determine which of these features are responsible for distinguishing between two or more sample groups. Many software tools are available for pre-processing LC-MS data, either commercial from instrument vendors like MarkerLynx from Waters and Metabolic Profiler Pro from Agilent or freely available, such as XCMS (Smith et al., 2006), MetAlign (Lommen, 2009), and MZmine (Pluskal et al., 2010). The common task of these tools involves two critical steps: (1) feature detection, which aims to detect and integrate all true peaks within a chromatographic run for each sample, and (2) feature alignment, which intends to match the features representing the same ion in multiple samples. All these tools require selection of several parameters that has to be optimized based on instrumental conditions, and comparative studies have shown that the same data pre-processed with different softwares can lead to detection of only 20–40% common features (Tautenhahn et al., 2008; Gürdeniz et al., 2012). This mismatch can be explained by large number of peaks with varying peak shapes and differences in the implemented peak detection method and corresponding parameter settings for each software. So far, there has been no study to clearly demonstrate that any of the software performs better than the others and thus many factors are to be considered for the choice of software platform, such as programming skills, provided GUIs for visualization, and the computer power for pre-processing large number of samples. Some of the practical properties of XCMS, MZmine, and MarkerLynx are listed in Table 2. The pros and cons of several LC-MS data pre-processing tools have been reviewed by Castillo and co-workers (2011).

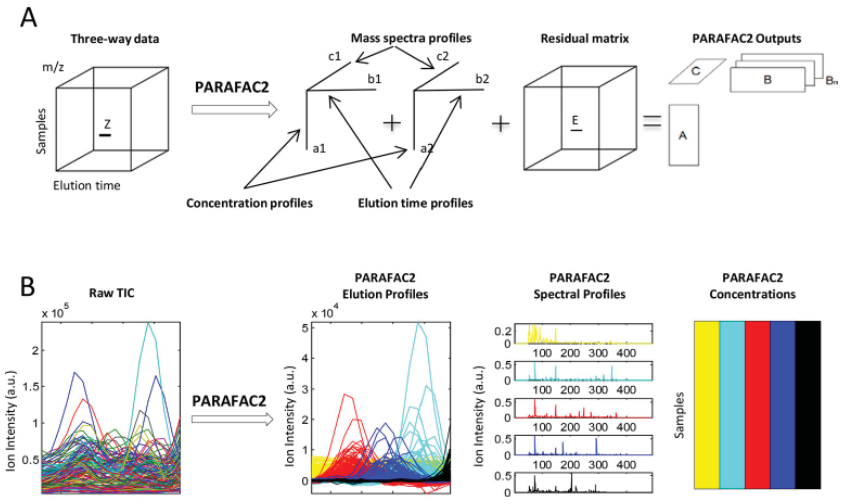

Fig. 4. (A) PARAFAC2 decomposition of the three-way GC-MS data, (B) example of the PARAFAC2 based processing of the raw GC-MS data interval

Table 2. Properties of commonly used LC-MS data pre-processing tools, MZmine, XCMS, and MarkerLynx

	MZmine	XCMS	MarkerLynx
Availability	Free	Free	Commercial
User interface	• GUI* • No requirement of programming skills	• R software command line • Some programming skills are required	• GUI* • No requirement of programming skills
Memory usage	• Adjustable to maximum available memory in the PC. • Less efficient than XCMS, e.g. 16 GB RAM = maximum ~2000 samples	• Adjustable to maximum available memory in the PC, e.g. 16 GB RAM = maximum ~5000 samples	• Fixed, e.g. maximum ~1000 samples
CPU usage	• Adjustable to maximum available CPU in the PC	• Adjustable to maximum available CPU in the PC	• Fixed
Identification tools	• Basic identification tools • Automated advanced tool CAMERA is partially incorporated	• Automated advanced identification tool CAMERA	• Basic identification tools
Coverage of pre-processing pipeline	All steps	Final feature table includes isotopic peaks	All steps
Visualization of the results	Yes	Yes	No

NMR. In NMR spectroscopy, the chemical shifts derive from free induction decay by utilizing a Fourier transformation, and the first step of pre-processing is phase and baseline correction, which is often performed by the software tools provided by machine-vendor software. Due to pH differences, overall dilution of samples and relative concentrations of some metabolites, the chemical shift of the same analyte signal may vary across different samples. In order to correct for these variations, a simple and common approach, spectral binning, has been widely applied. The main disadvantage of binning is loss of spectral resolution, and in order to avoid this problem, more sophisticated alignment tools have been proposed, utilizing varying procedures to determine optimum values for alignment, such as genetic algorithms (Forshed et al., 2003), partial linear fit (Vogels et al., 1996), and correlations (Veselkov et al., 2009; Savorani et al., 2010b). Correlation based alignment methods, such as recursive segment-wise peak alignment (Veselkov et al., 2009) and interval correlated shifting (icoshift) (Savorani et al., 2010b) use the efficient fast Fourier transform engine to optimize the algorithms to be able to handle large data sets in real time. The same methods can be used to align chromatographic data (Tomasi et al., 2011), but in chromatography, there has been a tradition of using the more flexible and meta-parameter demanding Correlation Optimized Warping (COW) method (Nielsen et al., 1998, Tomasi et al., 2004).

Vibrational spectroscopy. Vibrational spectroscopic data, such as IR spectra and especially NIR spectra, are influenced by undesired scatter effects caused by the sample morphology and the particle size of biological samples. The scatter effects are observed both as baseline shifts, multiplicative effects, and non-linearities that hinder the extraction of

relevant information using chemometric bilinear modelling (e.g. PCA and PLS). Bilinearity requires that the signal of each identical compound is stored in a column among all samples. Pre-processing of NIR spectral data aims to eliminate undesirable scatter effects prior to data analysis (Rinnan et al., 2009). Two of the most common pre-processing methods are Multiplicative Scatter Correction (MSC) (Geladi et al., 1985) and Standard Normal Variate (SNV) (Barnes et al., 1989). The MSC eliminates irrelevant variation due to scatter effects in two steps: (1) regression of a sample against a reference spectrum, (2) correction of the sample spectrum using intercept and slope of this fit. The SNV performs a correction of each sample by the standard deviation of all variables for that sample. In the course of SNV, each sample is processed individually, independently from entire set unlike MSC, where a reference spectrum is required.

4. Chemometric analysis methods

Unsupervised and supervised multivariate data analysis methods play a key role in exploration of multivariate data sets, where *K* number of variables are measured for *N* number of samples. Usually, the main aim from the analysis of such a data matrix, e.g. $\mathbf{X}$ *(N×K),* is to explore the underlying patterns in the dataset and extract variables that provide quantitative (e.g. prediction of melamine concentration in milk) and/or qualitative (e.g. discrimination of food products according to their origin) information. In foodomics studies, it is normal to include 100–500 samples, and if these samples were to be measured by the classical quality control (QC) analysis set-up like NIR spectroscopy as is common in the classical quality control (QC) analysis set-up, typically 1000 spectral variables are recorded. Such data sets of the order of 100×1000 can, with advantage, be dealt with using chemometrics, which efficiently can model collinear data sets with many more variables than samples. However, nowadays it is not uncommon that the analytical platforms may record more than 100 000 variables from each sample, which pushes the chemometric tools to the limit, as this increase will also increase the chances of spurious correlations and include more interferences (Skov et al., 2014). Indeed, in omics based profiling experiments, *K* is much larger than *N* and highly multi-collinear. Multivariate methods make use of all these variables simultaneously and deal with the relationship amongst the variables. In the following sections, the commonly applied unsupervised and supervised multivariate methods will be discussed.

4.1. Data pre-treatment methods

Normalization, scaling, and centering are usually necessary for efficient application of multivariate data analysis methods. Initial normalization can be applied for potential systematic error arising from sample preparation and/or in cases where instrumental issues can bring out unwanted variations between samples that may hinder the extraction of relevant variation.

The unwanted variation may appear in two forms: 1. Overall concentration variations between samples, i.e. the signal increase in all analytes of one sample compared to another sample. In this case, a scaling factor based normalization method can be used for correction of between–sample variations. Scaling factor based normalization is performed by division of each analyte in a sample by a factor, such as unit norm, total area, and total sum of intensities, calculated for that sample. 2. Analyte specific fluctuations between samples, i.e. the signal increases for one analyte, while decreasing for another analyte. This may be an

issue in GC-MS and LC-MS applications, which can ideally be corrected by utilization of isotope labelled internal standards. It is suggested where each internal standard is added to each sample in identical concentrations. However, a fully labelled reference metabolome is not feasible. Thus, multiple internal standards, each representing metabolites from chemically related groups, can be used for correction of systematic errors on the metabolite level (Bijlsma et al., 2006; Sysi-Ahoet al., 2007).

Particularly in GC-MS and LC-MS based studies, the analyte levels can differ with orders of magnitude, yet this may not correspond to the biological question addressed. For instance, two metabolites with signals of 5000 and 50 are usually of equal importance. However, PCA tends to gravitate upon the larger variation that is provided by larger peaks. Thus, scaling is necessary prior to PCA or PLSR, to put metabolites on similar or equal basis. Centering adjusts for differences in the offset between high and low abundant metabolites, while mean centering forces the corrected (centered) metabolite concentrations to fluctuate around zero mean. In most cases, centering is applied in combination to scaling. Autoscaling and pareto scaling are the most commonly employed scaling strategies in metabolomics. Autoscaling, which is the combination of unit scaling and mean centering, uses standard deviation as the scaling factor. After unit scaling of the data, all metabolites have standard deviation of one so that they have equal chance to influence the model. The main disadvantage of autoscaling is that it also inflates the noise, particularly for NMR and NIR profiles, which in turn may hinder the extraction of relevant patterns. Pareto scaling utilizes square root of standard deviation as scaling factor. As a result, it reduces large scale differences between metabolites, but still they are close to the original measurements. Although some studies in LC-MS based metabolomics use pareto scaling of the data, in most cases autoscaling seems to be a better choice, unless there is a specific interest or situation (e.g. very noisy data). The reason is that the magnitude of metabolite concentration differences are not representative of the biological relevance, which can only be provided by autoscaling (van den Berg et al., 2006).

4.2. Unsupervised exploration and classification

Unsupervised data analysis methods do not utilize any prior information on sample characteristics (quantity or category) in the modelling. In food control, foodomics and nutritional metabolomics, the most commonly used unsupervised data analysis methods are PCA and HCA.

4.2.1. Principal Component Analysis (PCA). PCA is the mostly utilized unsupervised method, which aims to extract the dominant patterns in a data matrix consisting of a large number of interrelated variables in terms of lower dimensional variables called principal components. Principal components represent linear combinations of original variables. The components are approximated as orthogonal directions in original variable space with the aim of capturing maximum variance. PCA can be formulated as

$$\mathbf{X}=\mathbf{T}\times\mathbf{P}'+\mathbf{E} \quad (1)$$

where **T** is the score matrix ($n\times i$-components), **P** is the loadings matrix ($k\times i$-components), and **E** is the residuals. The sample patterns are commonly visualized by a scatter plot of scores, e.g., $\boldsymbol{t}_1$ *vs.* $\boldsymbol{t}_2$ for the first two components, and the corresponding variable patterns are represented by a loadings plot e.g., $\boldsymbol{p}_1$ *vs* .$\boldsymbol{p}_2$.

PCA can with advantage be applied to most multivariate data sets (Engelsen et al., 2013) for outlier detection, data exploration, sample classification, and replicate analysis. PCA has been in use in food analysis (Willson & Freeman, 1970; MacNaughton et al., 1972) and spectroscopy/spectrometry (Rasmussen et al., 1978) since the 1970s and is nowadays commonly used in various foodomics studies, such as food authenticity and food traceability. Typical examples could be the determination of the geographical origin of wine (Schlesier et al., 2009), classification of fish oil products (Aursand et al., 2007), and the authenticity of grape cultivars based on their antioxidant compounds detected using HPLC (Berente et al., 2000). In addition, PCA is frequently used to detect food adulteration and explore effects of food on human health. For example, Oliveira and co-workers (2009) applied PCA on GC-MS aroma profiles of coffee samples to study adulteration of roasted coffee with roasted barley and Savorani and co-workers (2013) used PCA to explore contrasts between diets in the human metabolome.

In particular, PCA is the powerful tool for exploring complex data sets, such as NIR spectroscopic data, and such a combination have successfully been applied in various applications within foodomics (Cozzolino, 2014). Botros and co-workers (2013) attempted to develop an untargeted adulterant detection method in milk powders based on PCA and NIR spectroscopy, and the same techniques have previously been applied for the identification of mutant endosperm genes from NIR spectra of genetically modified barley cultivars with exceptionally high beta-glucan content (Munck et al., 2004).

4.2.2. Hierarchical Cluster Analysis (HCA). The other main unsupervised method, HCA (Cattell, 1943), is commonly used to study similarities and differences present amongst the investigated samples. HCA is based on either an agglomerative approach, that considers each sample as a separate cluster and then gradually merges it with other similar samples to form clusters, or a divisive approach that assumes that all samples constitute a single cluster and recursively splits the samples moving down the hierarchy. In HCA, similarities between samples can be estimated using metric systems, such as Euclidean distance, Manhattan distance, or Mahalanobis distance. The results of HCA are normally presented in dendrograms. The HCA method has been applied to complex data sets, such as LC-MS for studying a relationship between metabolites and plant's resistance against insects (Kuzina et al., 2009) and GC-MS data, where covariance of primary and secondary metabolites of seven rice cultivars have been explored (Kim et al., 2013). In addition, HCA is frequently used on spectroscopic data, such as Raman spectroscopy for classification of citrus fruit (Feng et al., 2013) and FT-IR based molecular structural analysis of various feed and food mixture samples (Abeysekara et al., 2013).

4.3. Partial Least Squares Regression (PLSR)

PLSR is a linear regression based method for relating a set of predictor variables, **X**, with one or more response variables, **Y**. As mentioned previously, PCA aims to find a subspace that explains the maximum amount of variation in **X** (*N* samples and *K* features). PLSR, on the other hand, tries to find a smaller dimensional subspace that describes the **X** well, but at the same time the coordinates of this new subspace are good predictors of the response variable **Y**. Similar to PCA, the components are orthogonal. For PLSR, **X** matrix is

decomposed by using eqn (1), providing **T** and **P**. Decomposition of **Y**, on the other hand can be formulated as:

$$\mathbf{Y}=\mathbf{U}\times\mathbf{Q}'+\mathbf{F} \quad (2)$$

such that **U** and **Q** are scores and loading matrices, respectively. In PLSR, the calculated **T** and **P** values differ as the criteria for PLSR is not only to describe **X** but also to provide a relation between **X** and **Y**. Thus, **P** and **Q** are calculated in such a way that the covariance between **T** and **P** is maximum. The PLSR scores and loadings can be interpreted as in PCA.

PLSR is one of the most powerful regression methods that exist, which can deal with highly collinear data, but it is important to note that it is prone to over-fitting, i.e. good calibration fitting, but with no predictive ability. Thus, determination of correct model complexity, in other words correct number of components, is critical and model must be validated before interpretation. In PLSR one of the most common validation methods is cross-validation, the samples are divided into training and validation sets. The training set is used to develop models with different number of components (i.e. from 1 to *n*). These models are evaluated based on their performance for correctly predicting the training set and then, the number of components providing the lowest value for Root Mean Square Error of Cross-Validation (RMSECV) is selected. However, assessment of the performance of the final model by RMSECV of the training sample set may lead to over-optimistic validation results. The model is optimized for the samples that are left out, so, those do not assess the validity of the final model. For a proper validation, the total data should be divided into training, validation, and test sets. Then the optimized model using the training and validation sample sets is used to evaluate a final model performance using the virgin test samples (Brereton, 2006).

Supervised regression analysis is probably the most commonly used method for rapid prediction one or more valuable features of food products that are costly and difficult to measure. The first applications of PLSR in food analysis (Martens et al., 1983; Frank & Kowalski, 1984) and spectroscopy/spectrometry (Lindberg et al., 1983, 1985) was published in the 1980s, and PLSR continues to play a most important role for analyzing foodomics data. PLSR is commonly used in food control, and new applications are emerging in nutritional metabolomics, such as for example to provide high throughput methods for measuring the postprandial levels of chylomicrons in the blood (Savorani et al., 2010a). PLSR has also been used to predict dioxin content in fish meal samples using GC-FID profiles (Bassompierre et al., 2007), for the prediction of caffeine and chlorogenic acid content in instant coffee mixtures from IR and NIR spectra (Fabián et al., 1994), and for the prediction of adulteration level of caprine and ovine milk with bovine milk from pyrolysis mass spectrometry data (Goodacre, 1997). Recently, PLSR was applied to develop a NIR based method for rapid prediction of sugar content of sugarcane plants (Taira et al., 2013) and for prediction of the adulteration level of butter with margarine using Raman spectroscopy (Uysal et al., 2013).

4.3.1. PLSR with variable selection. PLSR is able to deal with large number of variables, yet, in many situations it is desired to reduce the number of variables in order to improve the model predictions and/or to obtain better interpretation. However, the vast amount of variables compared to the relative few objects may often generate spurious correlations when Y-based variable selection methods, such as "forward selection", are applied (i.e. find the variable that best predicts the response variable; find the second variable that best improves the prediction; etc.). A large diversity of variable selection methods exist (Andersen & Bro,

2010), but commonly applied variable selection methods are based on the model parameters, describing relevance of each variable, such as regression coefficients, variable importance of projections (VIPs), and selectivity ratio. Regression coefficient represents the importance of a given variable for modelling dependent **Y,** whereas VIP summarises its importance for both independent variable **X** and dependent variable **Y** (Wold et al., 2001). Selectivity ratio is calculated for a variable as the ratio between explained and residual variance on the target projected component, which is a single latent variable explaining the covariance of the **X** variables with the **Y** (Rajalahti et al., 2009). In order to develop a highly predictive PLSR model with as low RMSECV as possible, the variables with insignificant regression coefficients, VIPs, or selectivity ratios are excluded, and a final model usually includes only the selected variables with high predictive power. However, such models require an appropriate validation against an intact test set samples, which have been involved neither in the variable selection, nor in the regression procedures.

In this context, two other variable selection methods based on PLSR should be mentioned: interval Partial Least Squares (*i*PLS) (Nørgaard et al., 2000) and recursive weighted Partial Least Squares (*r*PLS) (Rinnan et al., 2014). *i*PLS is commonly used for spectroscopic data, where adjacent variables are highly correlated. It involves the division of spectral data into a number of smaller intervals where PLSR is calculated for each interval. Subsequently, the intervals that are better for predictions than compared to the whole spectrum are selected using the RMSECV as evaluation criteria.

The combination of the *i*PLS with spectroscopic data has found widespread use in foodomics due to the improved interpretability and better performance (Larsen et al., 2006; Kristensen et al., 2010; Di Anibal et al., 2011). A few selected examples could be the prediction of the glucose content in various sport drinks by micro-Raman spectroscopy as an on-line quantification tool (Delfino et al., 2011), the prediction of crystalline lactose content in whey permeate powder by NIR spectroscopy (Nørgaard et al., 2005). The extension of *i*PLS to discriminant analysis, namely *i*PLS-DA, was used for discrimination of red wines adulterated with anthocyanins from black rice, using NMR data (Ferrari et al., 2011).

*r*PLS iteratively uses the regression coefficients to magnify important variables and thus down-weight less important variables. Recursive weighted PLS is based on an iterative process of repeated PLSR models, in which the current regression coefficients are used as cumulative weights on **X**. The *r*PLS model has the excellent property that it will converge to a limited number of variables (equal to the number of PC's), but it will exhibit optimal performance before normally including co-linear neighbouring variables. The *r*PLS method has great potential in foodomics studies, but so far the use of *r*PLS for exploitation of foodomics data has not been documented in literature.

4.4. Discrimination analysis

4.4.1. Partial Least Squares – Discriminant Analysis (PLS-DA). In foodomics, PLSR has been extensively applied in discrimination problems, where class labels (e.g. case *vs.* control, exposed *vs.* unexposed, pure *vs.* adulterated) are used as **Y** vector. PLS-DA (Ståhle & Wold, 1987) is a classical PLS regression used to discriminate samples, considering two-class case, the **Y** variable is set to have zero and one entries for each class (dummy matrix), respectively. PLS-DA aims to improve the separation between the two groups by using the class information.

The orthogonal PLS-DA (OPLS-DA) has been developed as an extension of PLS-DA and it is extensively used in metabolomics (Trygg & Wold, 2002). In OPLS-DA, the **Y**

unrelated (orthogonal) variation is removed from **X**. In this way, OPLS-DA attempts to describe the classification information in one component, which may provide advantages in terms of interpretation. However, the prediction power of PLS-DA and OPLS-DA is identical (Kemsley & Tapp, 2009).

Over-fitting is a potential danger in foodomics data as it usually contains a large number of irrelevant variables. Yet many studies have appeared to present PLS-DA scores and loadings plots from models without any indication of validation diagnostic statistics. To point out this issue, Westerhuis and co-workers (2008) performed PLS-DA on NMR spectra of 23 healthy volunteers, which were arbitrarily divided into two classes. Although PLS-DA scores plot showed a clear separation, cross-validation of revealed Q^2 values of –0.18 (no classification) illustrated the importance of validation (Westerhuis et al., 2008). In addition to the proposed validation routine described for PLSR, double cross-validation has been suggested for reducing over-optimism in PLS-DA cross-validation (Anderssen et al., 2006; Smit et al., 2007). In this case, training, validation, and test sets were selected randomly a high number of times. PLSR variable selection methods also apply for PLS-DA, but the danger of over-fitting increases. PLS-DA model performance has been evaluated using different diagnostic statistics, such as RMSECV, number of misclassifications (NMC), the Area Under the Receiver Operating Characteristic (AUROC), Q^2 and Discriminant Q^2 (DQ^2). Szymanska et al. (2012) demonstrated that for the evaluation of two group discriminmation problem, NMC and AUROC are efficient and reliable diagnostic statistics compared to DQ^2 and Q^2. It is important to note that although the importance of validation and validation tools are described under the PLS-DA section, it is applicable to other supervised methods described in the subsequent sections (ECVA and SIMCA).

In order to demonstrate the over-fitting issues of PLS-DA for datasets, where small number of samples were represented with large number of variables, an example dataset comprising 3703 variables (measured by UPLC-QTOF) representing of 41 urine samples, 23 and 18 samples after coffee and water intake, respectively, has been used. Figure 5 shows an example of PLS-DA based misclassification percentages based on calibration set, two different kinds of CV and independent test set validation on the data with original and randomly permuted classes. As shown in Fig. 5, the calibration set does not provide meaningful information, as the dataset with both original and random class labels had zero misclassifications. This example shows that PLS-DA without a proper validation provides a wrong classification of the sample groups. The leave-one-out CV also showed an over-optimistic result indicating just above 20% of misclassification error for random class dataset and zero error for the original class dataset. However, the random subset CV, when one fourth of the samples were selected as CV set with 10 iterations, also displayed a relatively over-optimistic classification performance and showed just above 30% of misclassification error for the random class dataset. These results urge to apply independent test set validation for the reliable classification, since with large number of variables and relatively few samples PLS-DA is always capable of finding dimensions that may separate sample groups.

PLS-DA is one of the most powerful classification tools and it is normally the reference method in supervised classification studies within food adulteration, food authenticity, food traceability, and food effects on human health. In nutritional metabolomics PLS-DA is the favourite classification method due to its sensitivity to reveal hidden patterns related to diets effects that are difficult to be revealed by unsupervised techniques (Andersen et al., 2014). Other recent studies include the authentication of geographical origin of palm oils based on GC-MS profiles of triacylglycerols (Ruiz-Samblás et al., 2013) and differentiation of NIR

spectra of cow milk samples based on farm altitude, different feeding regimes, and different breeding systems (Valenti et al., 2013). The combination of NIR spectroscopy and PLS-DA classification have also been proposed to augment the control of GMO foods by allowing high-throughput screening and discrimination of transgenic potato line (LeETR2) from its parental non-transgenic lines (Xie et al., 2007).

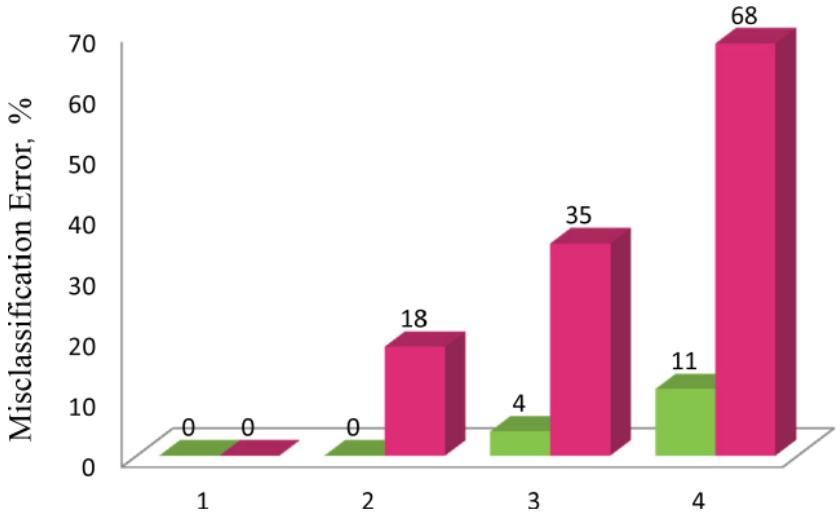

Fig. 5. Misclassification errors in PLS-DA models with different validation methods in the examples of the UPLC-QTOF data (41 samples and 3703 variables) from a metabolomics intervention study. PLS-DA misclassification percentages based on (1) calibration set, (2) leave-one-out cross validation (CV), (3) random subset CV (1/4 of the data is selected as CV set, with 10 iterations), and (4) independent test set validation.
■: original class; ■: random class

4.4.2. SIMCA. SIMCA is a classification technique based on disjoint PCA models for each class within data (Wold & Sjöström, 1977). The method assumes that it is possible to capture information reflecting similarities of individuals within each class, by class specific PCA models using training samples from each class. Then unknown samples are compared to the class models, and assigned to classes according to their analogy to training samples. Originally, the number of components required for PCA models to describe the training samples in each class is determined using cross-validation. The distance of a sample to the class model is calculated utilizing the orthogonal distance of the sample to the model space and the distance of the sample to the scores space. The SIMCA classification rule is then based on the comparison of the squared distance, with the class residual variance, by means of an F-test. The unknown samples are assigned to a class if the test is passed. The sample may also be assigned to several or none of the classes. More recently, the distance is determined by means of some other methods, such as Hotelling's T2 and Mahalanobis distances, that were also used for outlier detection (Hawkins et al., 1983). SIMCA is useful in situations when variance is relevant to separate classes. However, in cases where the between-class variation is smaller than the within-class variation, the classes will merge, therefore, SIMCA provides poor classification results (Galtier et al., 2011).

Like PCA, SIMCA can be used in various foodomics studies, such as food adulteration, food authenticity, food traceability, and food effects on human health. However, in contrast to PCA the class information is used actively. Recently, SIMCA was applied to detect adulteration of hazelnut paste with almond paste or chickpea flour based on NIR spectroscopy (López et al., 2014), authentication of geographical origin of honey (Latorre et al., 2013), and for the discrimination of fresh and frozen beef burger products from beef offal adulteration using IR spectroscopy (Zhao et al., 2014).

4.4.3. Extended canonical variates analysis (ECVA). ECVA (Nørgaard et al., 2006) has been developed as a modification of Canonical Variates Analysis (CVA). CVA (Campbell & Atchley, 1981) aims to estimate directions in space that maximize the differences between the groups according to well-defined optimization criterion, which is finding a direction that maximizes difference between projected mean values of each group relative to projected variance within groups.

Data generated in foodomics experiments are multi-collinear and CVA cannot deal with multi-collinear data. ECVA is based on the standard CVA, yet it reformulates eigenvector problem in CVA as a regression problem, which can be solved by using PLSR. In this way, canonical variates are calculated in the original high-dimensional space making it possible to deal with such data. Application of linear discriminant analysis (LDA) to the canonical variates allows the discriminative directions to be estimated directly in the original multidimensional space (Nørgaard et al., 2006). The number of canonical directions is always one less than the number of classes in the dataset. ECVA provides plots easily interpretable, similar to PLS-DA loadings and scores, canonical weights, and variates explain the patterns related to variables and observations, respectively, in each canonical direction. Additionally, ECVA has been shown to be an efficient classification method for more than two class problems (Nørgaard et al., 2006).

Application of supervised classification method ECVA in foodomics studies is scarce, despite its potential for solving classification problems involving several classes. The method was successfully applied in nutritional metabolomics to study effects of onion intake, in which ECVA was used to distinguish between the urine NMR metabolic profiles of rats with normal feed and rats with an onion diet (Winning et al., 2009) and to the discrimination of Rioja wineries, which are even geographically very close (terroir) (Lopez-Rituerto et al., 2012).

4.5. Advanced data analysis methods

4.5.1. ANOVA Simultaneous Component Analysis (ASCA). Experimental design structures that investigate food systems as a function of more than one underlying design factors, such as different treatments, time course, doses of an active compound, and different processing, are commonly used in foodomics studies. In order to utilize the design information in the multivariate data analysis, ASCA (Smilde et al., 2005) has been developed. ASCA allows for investigation of the individual design factor contributions and their interactions. ASCA utilizes advantages of both ANOVA in terms of partitioning the sources of variance and PCA for explaining maximum variance. Let us assume a data matrix, **X** with I observations and J variables, consists of balanced experimental design with two factors, A and B, which both have two levels (e.g. control *vs.* treatment, high *vs.* low dose, two different varieties or time points, etc.). The ANOVA model can be represented as shown in Figure 6. The first term in

this model, the 'overall' mean matrix, $\mathbf{X}_{mean}$, represents the variation of the variables profile from zero. It is calculated as the column means of the original data, **X.** Initially, $\mathbf{X}_{mean}$ is removed from **X** and then contributions of factor matrices, $\mathbf{X}_A$ and $\mathbf{X}_B$, are determined by taking the mean of each factor for each level. Later the factor matrices are also subtracted from $\mathbf{X}–\mathbf{X}_{mean}$ and then $\mathbf{X}_{AxB}$ interaction matrix as the average of columns of remaining residuals for combined levels of factors A and B. Finally, the effect of interaction terms is removed and the errors related to individual samples are collected in the residual matrix, $\mathbf{X}_{individual}$.

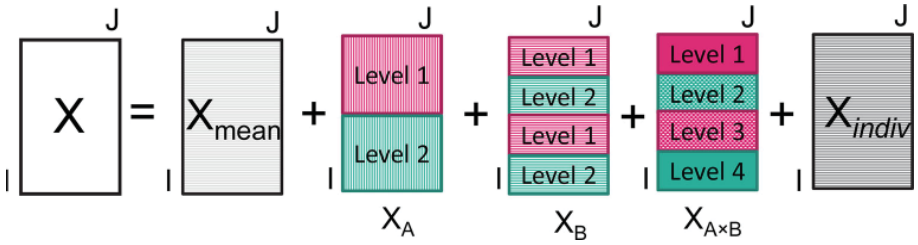

Fig. 6. Balanced ANOVA–simultaneous component analysis (ASCA) methodology

It is important to point out that the matrices provided by ANOVA decomposition are orthogonal, or in other words are independent from each other. Therefore, the effect of each of the term, i.e. the contribution of the factors and their interactions to the total variability, is estimated by means of the sum of squares (SS) of the corresponding matrix. Furthermore, each isolated matrix can be investigated independently by simultaneous component analysis, which is analogous to PCA, which can be formulized as:

$$\mathbf{X}_A=\mathbf{T}_A\mathbf{P}^T_A+\mathbf{E}_A, \quad \mathbf{X}_B=\mathbf{T}_B\mathbf{P}^T_B+\mathbf{E}_B, \quad \mathbf{X}_{A\times B}=\mathbf{T}_{A\times B}\mathbf{P}^T_{A\times B}+\mathbf{E}_{A\times B} \tag{3}$$

This leads to a reduction of *J* variables to the number of principle components, which satisfies one of the main aims of ASCA – namely the dimension reduction. Here, **T** and **P** are the scores and loadings, respectively, for each decomposed matrix. The information that is not described by any of the ASCA sub-models is represented by total residual matrix **E** matrix defined as the sum of the sub-model residual matrices $\mathbf{E}_A+\mathbf{E}_B+\mathbf{E}_{A\times B}$.

Nevertheless, the application of ASCA is limited to balanced design structures, so that there will be equal number of samples per each factors and their levels. In order to deal with unbalanced designs, a modified version of ASCA has been suggested (Stanimirova et al., 2011; Rago et al., 2013).

Some foodomics studies use balanced experimental designs and may possess underlying factors, such as time resolved measurements, varying doses of a specific diet and/or drug in intervention studies, or by varying growth temperature in plant metabolomic studies. In such studies, separation of effects of individual factors and interactions between factors are essential. Due to the possibilities of including the experimental design into the data analysis, the ASCA method is becoming a powerful tool in nutritional metabolomics. One of the first applications of ACSA was made to study the effect of grape/wine extracts on the NMR based human urine metabolome by dividing the effects into between and within the human subjects

(van Velzen et al., 2008). Another early application was made to study the three different factors, apple dose (0, 5, and 10 g), carcinogen induction, and fasting/fed states, in the LC-MS nutritional metabolomics study of rat plasma samples (Rago et al., 2013). The latter study demonstrated that ASCA was able to reduce the complexity, such that, the unique variation reflecting apple intake was isolated, which simplified subsequent data analysis.

4.5.2. Multi-block analysis or data fusion. Foodomics studies are normally concerned with multifactorial problems and it makes good sense to explore and measure the same samples on several complementary analytical platforms (Skov et al., 2014). Measurements at multiple platforms, such as NMR, LC-MS and GC-MS, provide a larger number of regulated metabolites and thus increase the chance for obtaining a better mechanistic understanding related to a specific diet or growth condition. However, the extraction of relevant information from such a complex dataset is challenging. Rather than analyzing each data block individually, simultaneous analysis of blocks of data using multi-block data analysis tools may provide better understanding of the investigated systems of interest. Multi-block methods attempt to extract the relevant information between and within blocks in terms of components or latent variables.

As an example, considering the same set of samples is analyzed in an IR and mass spectrometer, a multi-block component model can give insight into both uniquely and commonly represented information in the IR spectra and the mass spectra.

Similar to single block data analysis methods, multi-block techniques can be categorized into supervised and unsupervised approaches. Unsupervised multi-block methods include various extensions of PCA, such as SUM PCA (Smilde et al., 2003), consensus PCA (Wold et al., 1996), and hierarchical PCA (Wold et al., 1987). Corresponding methods perform data integration in two steps approach providing (1) scores and loadings for each block, and (2) consensus information derived from combination of all blocks. When choosing the multi-block method, it is important to clarify the objective of the study. The first issue to consider is whether the primary interest is to study the variation between the blocks or the variation within each block of data is also of interest. Another point to consider is fairness, that is, whether each block contributes equally (Smilde et al., 2003).

In many cases the goal of multi-block analysis or data fusion is to extract patterns not only common to all sources, but also specific to each source. For instance, in metabolomics, where the samples are measured by two different platforms, LC-MS and NMR, the interest may be to know whether the metabolites explaining a biological phenomenon are provided by one of the platforms (LC-MS or NMR) or there is a synergy between the variables provided for both platforms. Therefore, several methods aiming at accurate calculation of common and distinctive components have been developed, and their application potential in -omics based studies has been demonstrated (Acar et al., 2012, Van Deun et al., 2012).

Supervised multi-block models, which may be regarded as an extension to PLS regression, aim to relate multiple blocks of data provided by different sources with a response variable Y. Actually, the data from each analytical platform can be regarded as an interval – to follow the *i*PLS scheme –, the art is how to weight the individual analytical blocks. Some of the simple methods are regression based solutions of previously mentioned unsupervised methods, such as hierarchical PLS (Westerhuis et al., 1998), multi-block redundancy (Bougeard et al., 2011), and a consensus orthogonal PLS (O-PLS) (Bylesjo et al., 2007).

Thus far multi-block and data-fusion applications in foodomics have been limited, but coupled matrix factorization was applied on NMR and LC-MS profiles of rat plasma and

illustrated the usefulness of the method in a metabolomics application, where potential markers for apple intake are identified through coupled analysis of LC-MS and NMR data (Acar et al., 2012).

4.5.3. Sparse PCA (SPCA). In PCA, PCs are linear combinations of all variables (nonzero loadings), thus interpretation is often very difficult due to high number of irrelevant variables. Particularly for large and complex data sets, such as in foodomics (many variables for relatively few samples), it becomes very difficult to identify and select groups of important variables from many irrelevant ones. In order to overcome this issue, sparse PCA (SPCA) has been developed (Zou et al., 2006). SPCA aims to produce modified PCs with sparse loadings by imposing penalties on the model parameters (in this case loadings vectors), the less influential variables are forced to have zero influence on the model. Several methods have been proposed for estimating SPCA, utilizing either the regression error property or the maximum variance property of principal components (Witten et al., 2009). In the context of maximizing variance, SPCA can be formulated as a penalized optimization problem with the main objective being a minimization problem similar to PCA but with L_1 norm penalties imposed on the loadings:

$$argmin(\|\mathbf{X}-\mathbf{TP}^{\mathbf{T}}\|_F^2) \tag{4}$$

$$\text{subject to } \|\boldsymbol{p}_i\|_1^1 \leq c \text{ and } \|\boldsymbol{p}_i\|_2^2 = 1, \text{ for } i = 1,\ldots,k$$

where X**X**(n×k), is the data matrix, $\|\boldsymbol{p}_i\|_1^1$ is the sum of absolute values ($L_{\neg 1}$ norm) of the columns of loading matrix **P**, and **T** is the score matrix. The tuning parameter c is a positive penalty parameter bounding the sum of absolute values of the normalized loading vector ($\|\boldsymbol{pi}\|1 \leq c$). Thus, it leads to some loadings being exactly zero. If c is chosen large enough, it will lead to the unconstrained PCA solution. A meaningful sparse solution can be found when c is chosen between 0 and the sparsity level, producing unconstrained solution (Rasmussen & Bro, 2012). Unlike PCA, SPCA does not impose orthogonality constraint between components. SPCA components are correlated and the SPCA loadings are not orthogonal, which in some cases improves the extraction of relevant biological information from large metabolomics data as it forces less effective metabolites to have zero loadings, thereby variable selection becomes more efficient. As an example, in a LC-MS based metabolomics study, in which the metabolites reflect the time since last meal, the information has been more efficiently extracted using SPCA compared to PCA (Gürdeniz et al., 2013). It has been noted that the main obstacle in SPCA is the selection of sparsity level (meta parameter) that adjusts the number of variables with zero loadings. The evaluation is based on visual inspection of scores and loadings and might be rather time demanding. Sparsity penalty has also been implemented in unsupervised multi-block models applied on nutritional metabolomics to ease the selection of relevant metabolites (Acar et al., 2012).

4.5.4. PARAllelFACtor Analysis 2 (PARAFAC2). PARAFAC2 is an extension of PARAFAC (Harshman, 1970, Bro, 1997) that is able to model more complex three-way data sets with disturbed trilinear structure. Both PARAFAC and PARAFAC2 can be considered as the generalization of the principal component analysis (PCA) to higher order data arrays. In contrast to PCA, PARAFAC2 does not suffer from rotational problems and is able to model three-way data sets by decomposing it into a smaller number of components that will be represented by scores and loadings (Fig. 4A). In order to condense the three-way data in such

a way, PARAFAC2 applies some constraints that restrict the degree of freedom and provide simpler and more robust models. Therefore, PARAFAC2 may not be able to model all the variation that a PCA can capture, which is often the case when dealing with complex data sets with higher order of variations. Any three-way data can be unfolded to two-way matrix (in any mode) and modelled by PCA, which may result in explanation of the substantial part of the data variation, however, interpretation of such a model is challenging. Instead, PARAFAC2 offers several advantages for exploration of three-way arrays. Firstly, the obtained solutions are unique, which means that its scores and loadings directly represent the modes of the investigated data array. Secondly, models are robust and easily interpretable, however, model validation might be challenging if the data are complex. However, recent studies performed for improvement of PARAFAC and PARAFAC2 model validation (Kamstrup-Nielsen et al., 2013) have described more reliable and easier way of deciding the number of components of the PARAFAC models.

The main difference between PARAFAC and PARAFAC2 is that PARAFAC2 is less restrictive to the trilinear structure of the data and it is able to cope with data shifts in some extent. For example, in chromatography, retention time shifted peaks of the same metabolites over the different samples can still be modelled as the same metabolite, because PARAFAC2 uses not only the retention time dimension but also the mass spectral information (since mass spectra of these shifted peaks will be identical if they are derived from the same metabolite). All these features of PARAFAC2, e.g. uniqueness, shift and noise handling, and easier interpretation, make the method very useful for processing raw metabolomic three-way data sets derived from hyphenated platforms, such as GC-MS (Amigo et al., 2010a), LC-MS (Khakimov et al., 2012), and LC-DAD (Marini et al., 2011). By PARAFAC2 processing of such hyphenated metabolomic data, it is possible to extract most if not all the quantitative and qualitative information (Fig. 4B). The PARAFAC2 model of the three-way GC-MS data defined by elution times × mass spectra × samples provide three following outputs: (1) PARAFAC2 elution time profiles that represent the elution profiles of the resolved peaks, (2) PARAFAC2 mass spectral profiles that correspond to the actual mass spectra of the resolved peaks that can be used for metabolite identification from libraries, and (3) PARAFAC2 concentration profiles, which represent the areas of the resolved peaks.

The method provides processing raw metabolomic data and allows extraction of vast amount of information in a high-throughput manner. PARAFAC2 allows processing of all samples simultaneously, deconvolution of mass spectra of hundreds of metabolites per sample, and separation of pure analyte peaks by eliminating chromatographic baseline and alignment of retention time shifts. However, the method has some drawbacks primarily related to its availability and use by non-specialists. Despite its comprehensiveness, the method can provide fruitful results, when the data is less complex. Therefore, PARAFAC2 based processing of raw chromatographic data is mainly performed in baseline separated intervals, where the data is divided into smaller intervals in retention time dimension.

5. Current challenges and perspectives

Foodomics is a multidisciplinary research field that investigates foods, food processing, and foods effects on human health and wellbeing. In order to extract useful and reliable information from foodomics studies, they must have an appropriate design depending on the purpose of research. While the success of foodomics studies depends on the obtained data, the quality

and amount of this data are determined by the applied analytical platforms. State-of-the-art analytical instruments allow simultaneous detection of up to several thousands of analytes from the investigated sample matrix. Despite today's advances in analytical platforms, it is still not possible to detect a whole metabolome, and even more challenging to obtain reproducible quantitative data. In fact, there is always a compromise between the amount of the data and their quantitative/qualitative quality.

As the data generated by analytical platforms applied in foodomics are becoming more and more megavariate, appropriate validation of the results is becoming more and more important (Szymanska et al., 2012). Unfortunately, the validation of multi- and megavariate foodomics models is not always straightforward, which makes many new findings, especially made by the PLS-DA type of classification tools, questionable.

Foodomics studies normally concern multifactorial problems, wherefore more and more studies employ two or more analytical platforms, which urge researchers to develop new chemometrics tools that handle such large data sets and assist to extract relevant (meaningful) biological information. For example, development of new multi-block methods is one of the first steps made towards analysis of megavariate metabolomics data acquired on various systems simultaneously. This, in fact, has a high potential in near future to combine various data sets obtained from metabolomic, proteomic, transcriptomic, and genomic studies and perform in-depth data exploration to analyze the variance and co-variances present along different omics data sets. In foodomics, not only the human metabolome is in play, but also the food microbiota co-metabolome and the food metabolome (genotypes). Obtaining dynamical and coherent foodomics data from all these three streams of causality tracks is a tremendous task, which is far from being resolved yet. Nevertheless, the foodomics approach is here to stay and will be an integrated part of future discussions on bioactive substances, their bioavailability, and part of the documentation for obtaining health claims. The high-throughput analytical platforms and the related advanced multivariate chemometric methods will also be integral parts of future global health screenings and for development of stratified nutrition.

References

Abeysekara, S., Damiran, D. & Yu, P.Q. (2013): Univariate and multivariate molecular spectral analyses of lipid related molecular structural components in relation to nutrient profile in feed and food mixtures. *Spectrochim. Acta A., 102*, 432–442.

Acar, E., Gürdeniz, G., Rasmussen, M.A., Rago, D., Dragsted, L.O. & Bro, R. (2012): Coupled matrix factorization with sparse factors to identify potential biomarkers in metabolomics. *Proceedings of the 2012 IEEE International Conference on Data Mining Workshops*.

Ahmed, F.E. (2002): Detection of genetically modified organisms in foods. *Trends Biotechnol., 20*, 215–223.

Amigo, J.M., Popielarz, M.J., Callejon, R.M., Morales, M.L., Troncoso, A.M., Petersen, M.A. & Toldam-Andersen, T.B. (2010a): Comprehensive analysis of chromatographic data by using PARAFAC2 and principal components analysis. *J. Chromatogr. A, 1217*, 4422–4429.

Amigo, J.M., Skov, T. & Bro, R. (2010b): ChroMATHography: Solving chromatographic issues with mathematical models and intuitive graphics. *Chem. Rev., 110*, 4582–4605.

Andersen, C.M. & Bro, R. (2010): Variable selection in regression-a tutorial. *J. Chemometr., 24*, 728–737.

Andersen, M.B., Rinnan, A., Manach, C., Poulsen, S.K., Pujos-Guillot, E., Larsen, T.M., Astrup, A. & Dragsted, L.O. (2014): Untargeted metabolomics as a screening tool for estimating compliance to a dietary pattern. *J. Proteome Res.*, doi:10.1021/pr400964s.

Anderssen, E., Dyrstad, K., Westad, F. & Martens, H. (2006):Reducing over-optimism in variable selection by cross-model validation. *Chemometr. Intell. Lab., 84*, 69–74.

Arvanitoyannis, A.S. & Tzouros, N.E. (2005): Implementation of quality control methods in conjunction with chemometrics toward authentication of dairy products. *Crit. Rev. Food Sci. Nutr., 45*, 231–249.

Arvanitoyannis, I.S., Chalhoub, C., Gotsiou, P., Lydakis-Simantiris, N. & Kefalas, P. (2005a): Novel quality control methods in conjunction with chemometrics (multivariate analysis) for detecting honey authenticity. *Crit. Rev. Food Sci. Nutr., 45*, 193–203.

Arvanitoyannis, I.S., Tsitsika, E.V. & Panagiotaki, P. (2005b): Implementation of quality control methods (physico-chemical, microbiological, and sensory) in conjunction with multivariate analysis towards fish authenticity. *Int. J. Food Sci. Tech., 40*, 237–263.

Arvanitoyannis, I.S. & van Houwelingen-Koukaliaroglou, M. (2003): Implementation of chemometrics for quality control and authentication of meat and meat products. *Crit. Rev. Food Sci. Nutr., 43*, 173–218.

Aursand, M., Standal, I.B. & Axelson, D.E. (2007): High-resolution (13)C nuclear magnetic resonance spectroscopy pattern recognition of fish oil capsules. *J. Agr. Food Chem., 55*, 38–47.

Barnes, R.J., Dhanoa, M.S. & Lister, S.J. (1989): Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc., 43*, 772–777.

Bassompierre, M., Tomasi, G., Munck, L., Bro, R. & Engelsen, S.B. (2007): Dioxin screening in fish product by pattern recognition of biomarkers. *Chemosphere*, *67*, 28–35.

Behrends, V., Tredwell, G.D. & Bundy, J.G. (2011): A software complement to AMDIS for processing GC-MS metabolomic data. *Anal. Biochem., 415*, 206–208.

Berente, B., De la Calle García, D., Reichenbächer, M. & Danzer, K. (2000): Method development for the determination of anthocyanins in red wines by high-performance liquid chromatography and classification of German red wines by means of multivariate statistical methods. *J. Chromatogr. A, 871*, 95–103.

Berrueta, L.A., Alonso-Salces, R.M. & Heberger, K. (2007): Supervised pattern recognition in food analysis. *J. Chromatogr. A, 1158*, 196–214.

Bianchi, G., Giansante, L., Shaw, A. & Kell, D.B. (2001): Chemometric criteria for the characterisation of Italian Protected Denomination of Origin (DOP) olive oils from their metabolic profiles. *Eur. J. Lipid Sci. Tech., 103*, 141–150.

Bijlsma, S., Bobeldijk, L., Verheij, E.R., Ramaker, R., Kochhar, S., Macdonald, I.A., van Ommen, B. & Smilde, A.K. (2006): Large-scale human metabolomics studies: A strategy for data (pre-) processing and validation. *Anal. Chem., 78,* 567–674.

Botros, L.L., Jablonski, J., Chang, C., Bergana, M.M., Wehling, P., Harnly, J.M., Downey, G., Harrington, P., Potts, A.R. & Moore, J.C. (2013): Exploring authentic skim and nonfat dry milk powder variance for the development of nontargeted adulterant detection methods using near-infrared spectroscopy and chemometrics. *J. Agr. Food Chem., 61*, 9810–9818.

Bougeard, S., Qannari, E., Lupo, C. & Hanafi, M. (2011): From multiblock partial least squares to multiblock redundancy analysis. A continuum approach. *Informatica, 22*, 11–26.

Brereton, R.G. (2006): Consequences of sample size, variable selection, and model validation and optimisation, for predicting classification ability from analytical data. *Trac-Trend. Anal. Chem., 25*, 1103–1111.

Bro, R. (1997): PARAFAC. Tutorial and applications. *Chemometr. Intell. Lab., 38*, 149–171.

Bro, R., Andersson, C.A. & Kiers, H.A.L. (1999): PARAFAC2 - Part II. Modeling chromatographic data with retention time shifts. *J. Chemometr., 13*, 295–309.

Bylesjo, M., Eriksson, D., Kusano, M., Moritz, T. & Trygg, J. (2007): Data integration in plant biology: the O2PLS method for combined modeling of transcript and metabolite data. *Plant J., 52*, 1181–1191.

Campbell, N.A. & Atchley, W.R. (1981): The geometry of Canonical Variate Analysis. *Syst. Zool., 30*, 268–280.

Capozzi, F. & Bordoni, A. (2013): Foodomics: a new comprehensive approach to food and nutrition. *Genes Nutr., 8*, 1–4.

Castillo, S., Gopalacharyulu, P., Yetukuri, L. & Oresic, M. (2011): Algorithms and tools for the preprocessing of LC-MS metabolomics data. *Chemometr. Intell. Lab., 108*, 23–32.

Castro-Puyana, M., Garcia-Canas, V., Simo, C. & Cifuentes, A. (2012): Recent advances in the application of capillary electromigration methods for food analysis and Foodomics. *Electrophoresis, 33*, 147–167.

Cattell, R.B. (1943): The description of personality: Basic traits resolved into clusters. *J. Abnorm. Soc. Psych., 38*, 476–506.

Cifuentes, A. (2009): Food analysis and Foodomics Foreword. *J. Chromatogr. A, 1216*, 7109.

Cozzolino, D. (2014): An overview of the use of infrared spectroscopy and chemometrics in authenticity and traceability of cereals. *Food Res. Int., 60*, 262–265.

de Juan, A. & Tauler, R. (2006): Multivariate Curve Resolution (MCR) from 2000: Progress in Concepts and Applications. *Crit. Rev. Anal. Chem., 36*, 163–176.

Delfino, I., Camerlingo, C., Portaccio, M., Ventura, B.D., Mita, L., Mita, D.G. & Lepore, M. (2011): Visible micro-Raman spectroscopy for determining glucose content in beverage industry. *Food Chem., 127*, 735–742.

Di Anibal, C.V., Callao, M.P. & Ruisanchez, I. (2011): H-1 NMR variable selection approaches for classification. A case study: The determination of adulterated foodstuffs. *Talanta, 86*, 316–323.

Durante, C., Cocchi, M., Grandi, M., Marchetti, A. & Bro, R. (2006): Application of N-PLS to gas chromatographic and sensory data of traditional balsamic vinegars of modena. *Chemometr. Intell. Lab., 83*, 54–65.

Engelsen, S.B., Savorani, F. & Rasmussen, M.A. (2013): Chemometric exploration of quantitative NMR data. *eMagRes2*, 267–278.

Erny, G.L., Elvira, C., San Roman, J. & Cifuentes, A. (2006): Capillary electrophoresis using copolymers of different composition as physical coatings: A comparative study. *Electrophoresis, 27*, 1041–1049.

Fabián, Z., Izvekov, V., Salgó, A. & Örsi, F. (1994): Near-infrared reflectance and Fourier transform infrared analysis of instant coffee mixtures. *Anal. Proc., 31*, 261–263.

Feng, X.W., Zhang, Q.H. & Zhu, Z L. (2013): Rapid classification of citrus fruits based on raman spectroscopy and pattern recognition techniques. *Food Sci. Technol. Res., 19*, 1077–1084.

Ferrari, E., Foca, G., Vignali, M., Tassi, L. & Ulrici, A. (2011): Adulteration of the anthocyanin content of red wines: Perspectives for authentication by Fourier Transform-Near InfraRed and 1H NMR spectroscopies. *Anal. Chim. Acta*, *701*, 139–151.

Forshed, J., Schuppe-Koistinen, I. & Jacobsson, S.P. (2003): Peak alignment of NMR signals by means of a genetic algorithm. *Anal. Chim. Acta, 487*, 189–199.

Frank, I.E. & Kowalski, B.R. (1984): Predictions of wine quality and geographic origin from chemical measurements by partial least-squares regression modeling. *Anal. Chim. Acta, 162,* 241–251.

Galtier, O., Abbas, O., Le Dreau, Y., Rebufa, C., Kister, J., Artaud, J. & Dupuy, N. (2011): Comparison of PLS1-DA, PLS2-DA and SIMCA for classification by origin of crude petroleum oils by MIR and virgin olive oils by NIR for different spectral regions. *Vibr. Spectrosc., 55*, 132–140.

Geladi, P., Macdougall, D. & Martens, H. (1985): Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Appl. Spectrosc., 39*, 491–500.

Goodacre, R. (1997): Use of pyrolysis mass spectrometry with supervised learning for the assessment of the adulteration of milk of different species. *Appl. Spectrosc., 51*, 1144–1153.

Gürdeniz, G., Hansen, L., Rasmussen, M.A., Acar, E., Olsen, A., Christensen, J., Barri, T., Tjonneland, A. & Dragsted, L.O. (2013): Patterns of time since last meal revealed by sparse PCA in an observational LC-MS based metabolomics study. *Metabolomics, 9*, 1073–1081.

Gürdeniz, G., Kristensen, M., Skov, R. & Dragsted, L.O. (2012): The effect of LC-MS data preprocessing methods on the selection of plasma biomarkers in fed vs. fasted rats. *Metabolites, 2*, 77–99.

Hantao, L.W., Aleme, H.G., Pedroso, M.P., Sabin, G.P., Poppi, R.J. & Augusto, F. (2012): Multivariate curve resolution combined with gas chromatography to enhance analytical separation in complex samples: A review. *Anal. Chim. Acta, 731*, 11–23.

Harshman, R.A. (1970): Foundations of the PARAFAC procedure: Model and conditions for an "explanatory" multi-mode factor analysis. *UCLA Working Papers in Phonetics, 16*.

Harshman, R.A. (1972): PARAFAC2: mathematical and technical notes. *UCLA Working Papers in Phonetics, 22*.

Hawkins, D.M., Bradu, D., Kass, G.V. & Galpin, J.S. (1983): Outlier detection using elemental sets in regression. *S. Afr. Stat. J., 17*, 184.

Hotelling, H. (1933): Analysis of a complex of statistical variables into principal components. *J. Edduc. Psychiol., 24*, 417–441.

Kamstrup-Nielsen, M.H., Johnsen, L.G. & Bro, R. (2013): Core consistency diagnostic in PARAFAC2. *J. Chemometr., 27*, 99–105.

Karoui, R. & De Baerdemaeker, J. (2007): A review of the analytical methods coupled with chemometric tools for the determination of the quality and identity of dairy products. *Food Chem., 102*, 621–640.

Kasicka, V. (2012): Recent developments in CE and CEC of peptides (2009–2011): *Electrophoresis, 33*, 48–73.

Kemsley, E.K. & Tapp, H.S. (2009): OPLS filtered data can be obtained directly from non-orthogonalized PLS1. *J. Chemometr., 23*, 263–264.

Khakimov, B., Motawia, M.S., Bak, S. & Engelsen, S.B. (2013): The use of trimethylsilyl cyanide derivatization for robust and broad-spectrum high-throughput gas chromatography-mass spectrometry based metabolomics. *Analy. Bioanal. Chem., 405*, 9193–9205.

Khakimov, B., Amigo, J.M., Bak, S. & Engelsen, S.B. (2012): Plant metabolomics: Resolution and quantification of elusive peaks in liquid chromatography-mass spectrometry profiles of complex plant extracts using multi-way decomposition methods. *J. Chromatogr. A, 1266*, 84–94.

Khakimov, B., Bak, S. & Engelsen, S.B. (2014): High-throughput cereal metabolomics: Current analytical technologies, challenges and perspectives. *J. Cereal Sci., 59*, 393–418.

Kiers, H.A.L., Ten Berge, J.M.F. & Bro, R. (1999): PARAFAC2 - Part I. A direct fitting algorithm for the PARAFAC2 model. *J. Chemometr., 13*, 275–294.

Kim, J.K., Park, S.Y., Lim, S.H., Yeo, Y., Cho, H.S. & Ha, S.H. (2013): Comparative metabolic profiling of pigmented rice (*Oryza sativa* L.) cultivars reveals primary metabolites are correlated with secondary metabolites. *J. Cereal Sci., 57*, 14–20.

Kolch, W., Neususs, C., Peizing, M. & Mischak, H. (2005): Capillary electrophoresis - Mass spectrometry as a powerful tool in clinical diagnosis and biomarker discovery. *Mass Spectrom. Rev., 24*, 959–977.

Kristensen, M., Savorani, F., Ravn-Haren, G., Poulsen, M., Markowski, J., Larsen, F.H., Dragsted, L.O. & Engelsen, S.B. (2010): NMR and interval PLS as reliable methods for determination of cholesterol in rodent lipoprotein fractions. *Metabolomics, 6*, 129–136.

Kuzina, V., Ekstrom, C.T., Andersen, S.B., Nielsen, J.K., Olsen, C.E. & Bak, S. (2009): Identification of defense compounds in *Barbarea vulgaris* against the herbivore *Phyllotreta nemorum* by an ecometabolomic approach. *Plant Physiol., 151*, 1977–1990.

Larsen, F.H., van den Berg, F. & Engelsen, S.B. (2006): An exploratory chemometric study of H-1 NMR spectra of table wines. *J. Chemometr., 20*, 198–208.

Latorre, C.H., Crecente, R.M.P., Martin, S.G. & Garcia, J.B. (2013): A fast chemometric procedure based on NIR data for authentication of honey with protected geographical indication. *Food Chem., 141*, 3559–3565.

Lawton, W.H. & Sylvestre, E.A. (1971): Self modeling curve resolution. *Technometrics, 13*, 617.

Lindberg. W., Persson, J.A. & Wold, S. (1983): Partial least-squares methods for spectrofluorimetric analysis of mixtures of humic-acid and ligninsulfonate. *Anal. Chem., 55,* 643–648.

Lindberg, W., Ohman, J., Wold, S. & Martens, H. (1985): Determination of the proteins in mixtures of meat, soymeal and rind from their chromatographic amino-acid patternby the partial least-squares method. *Anal. Chim. Acta, 171,* 1–11.

Lommen, A. (2009): MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. *Anal. Chem., 81*, 3079–3086.

López, M.I., Trullols, E., Callao, M.P. & Ruisánchez, I. (2014): Multivariate screening in food adulteration: Untargeted versus targeted modelling. *Food Chem., 147*, 177–181.

Lopez-Rituerto, E., Savorani, F., Avenoza, A., Busto, J. H., Peregrina, J.M. & Engelsen, S.B. (2012): Investigations of La Rioja Terroir for wine production using H-1 NMR metabolomics. *J. Agr. Food Chem.*, *60*, 3452–3461.

MacNaughton, D., Rogers, L.B. &Wernimont, G. (1972): Principal-component analysis applied to chromatographic data. *Anal. Chem., 44,* 1421–1427.

Marini, F., D'Aloise, A., Bucci, R., Buiarelli, F., Magri, A.L. & Magri, A.D. (2011): Fast analysis of 4 phenolic acids in olive oil by HPLC-DAD and chemometrics. *Chemometr. Intell. Lab., 106*, 142–149.

Martens, M., Martens, H. & Wold, S. (1983): Preference of cauliflower related to sensory descriptive variables by partial least-squares (PLS) regression. *J. Sci. Food Agr., 34,*715–724.

Mas, S., de Juan, A., Tauler, R., Olivieri, A.C. & Escandar, G.M. (2010): Application of chemometric methods to environmental analysis of organic pollutants: A review. *Talanta, 80*, 1052–1067.

Mischak, H., Coon, J.J., Novak, J., Weissinger, E.M., Schanstra, J.P. & Dominiczak, A.F. (2009): Capillary electrophoresis-mass spectrometry as a powerful tool in biomarker discovery and clinical diagnosis: An update of recent developments. *Mass Spectrom. Rev., 28*, 703–724.

Munck, L., Moller, B., Jacobsen, S. & Sondergaard, I. (2004): Near infrared spectra indicate specific mutant endosperm genes and reveal a new mechanism for substituting starch with (1→3,1→4)-beta-glucan in barley. *J Cereal Sci., 40*, 213–222.

Munck, L., Norgaard, L., Engelsen, S.B., Bro, R. & Andersson, C.A. (1998): Chemometrics in food science – a demonstration of the feasibility of a highly exploratory, inductive evaluation strategy of fundamental scientific significance. *Chemometr. Intell. Lab., 44*, 31–60.

Nielsen, N.P.V., Carstensen, J.M. & Smedsgaard, J. (1998): Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *J. Chromatogr. A., 805*, 17–35.

Nørgaard, L., Hahn, M.T., Knudsen, L.B., Farhat, I.A. & Engelsen, S.B. (2005): Multivariate near-infrared and Raman spectroscopic quantifications of the crystallinity of lactose in whey permeate powder. *Int. Dairy J., 15*, 1261–1270.

Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J.P., Munck, L. & Engelsen, S B. (2000): Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. *Appl. Spectrosc., 54*, 413–419.

Nørgaard, L., Bro, R., Westad, F. & Engelsen, S.B. (2006): A modification of canonical variates analysis to handle highly collinear multivariate data. *J. Chemometr., 20*, 425–435.

Oliveira, R.C.S., Oliveira, L.S., Franca, A.S. & Augusti, R. (2009): Evaluation of the potential of SPME-GC-MS and chemometrics to detect adulteration of ground roasted coffee with roasted barley. *J. Food Compos. Anal., 22*, 257–261.

Pearson, K. (1901): On lines and planes of closest fit to systems of points in space. *Philos. Mag., 2,* 7–12.

Pluskal, T., Castillo, S., Villar-Briones, A. & Oresic, M. (2010): MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *Bmc Bioinformatics,* 11, 395.

Rago, D., Mette, K., Gürdeniz, G., Marini, F., Poulsen, M. & Dragsted, L.O. (2013): A LC-MS metabolomics approach to investigate the effect of raw apple intake in the rat plasma metabolome. *Metabolomics, 9*, 1202–1215.

Rajalahti, T., Arneberg, R., Berven, F.S., Myhr, K.M., Ulvik, R.J. & Kvalheim, O.M. (2009): Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. *Chemometr. Intell. Lab., 95*, 35–48.

Rasmussen, M.A. & Bro, R. (2012): A tutorial on the Lasso approach to sparse modeling. *Chemometr. Intell. Lab., 119*, 21–31.

Rasmussen, G.T., Lowry, S.R. & Ritter, G.L. (1978): Principal component analysis of the infrared spectra of mixtures. *Anal. Chim. Acta-Comp., 2,* 213–221.

Rinnan, Å., van den Berg, F. & Engelsen, S.B. (2009): Review of the most common pre-processing techniques for near-infrared spectra. *Trac-Trend. Anal. Chem., 28*, 1201–1222.

Rinnan, Å., Andersson, M., Ridder, C. & Engelsen, S.B. (2014): Recursive weighted partial least squares (rPLS): an efficient variable selection method using PLS. *J. Chemometr., 28*, 439–447.

Ruiz-Samblás, C., Arrebola-Pascual, C., Tres, A., van Ruth, S. & Cuadros-Rodriguez, L. (2013): Authentication of geographical origin of palm oil by chromatographic fingerprinting of triacylglycerols and partial least square-discriminant analysis. *Talanta, 116*, 788–793.

Savorani, F., Kristensen, M., Larsen, F.H., Astrup, A. & Engelsen, S.B. (2010a): High throughput prediction of chylomicron triglycerides in human plasma by nuclear magnetic resonance and chemometrics. *Nutr. Metab., 7*, 43.

Savorani, F., Rasmussen, M.A., Mikkelsen, M.S. & Engelsen, S.B. (2013): A primer to nutritional metabolomics by NMR spectroscopy and chemometrics. *Food Res. Int., DOI: 10.1016/j.foodres.2012.12.025*.

Savorani, F., Tomasi, G. & Engelsen, S.B. (2010b): icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *J. Magn. Reson., 202*, 190–202.

Schlesier, K., Fauhl-Hassek, C., Forina, M., Cotea, V., Kocsi, E., Schoula, R., van Jaarsveld, F. & Wittkowski, R. (2009): Characterisation and determination of the geographical origin of wines. Part I: overview. *Eur. Food Res. Technol., 230*, 1–13.

Skov, T. & Engelsen, S.B. (2013): Chemometrics, mass spectra, and foodomics. -in: Cifuentes, A. (Ed.): *Foodomics: Advanced mass spectrometry in modern food science and nutrition*. John Wiley & Sons, Inc., Hoboken, N.J., pp. 507–538.

Skov, T., Honoré, H.A., Jensen, H.M., Næs, T. & Engelsen, S.B. (2014): Chemometrics goes into foodomics. *TRAC Trend. Anal. Chem.*, doi: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. *BMCBioinformatics, 21*, 3043–3048.

Smilde, A.K., Westerhuis, J.A. & de Jong, S. (2003): A framework for sequential multiblock component methods. *J. Chemometr., 17*, 323–337.

Smit, S., van Breemen, M.J., Hoefsloot, H.C.J., Smilde, A.K., Aerts, J.M.F.G. & de Koster, C.G. (2007): Assessing the statistical validity of proteomics based biomarkers. *Anal. Chim. Acta, 592*, 210–217.

Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R. & Siuzdak, G. (2006): XCMS: Processing mass spectrometry data for metabolite profiling using Nonlinear peak alignment, matching, and identification. *Anal. Chem., 78*, 779–787.

Ståhle, L. & Wold, S. (1987): Partial least squares analysis with cross-validation for the two-class problem a Monte Carlo study. *J. Chemometr., 1*, 185–196.

Sysi-Aho, M., Katajamaa, M., Yetukuri, L. & Oresic, M. (2007): Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC Bioinformatics, 8,* 93.

Stanimirova, I., Michalik, K., Drzazga, Z., Trzeciak, H., Wentzell, P.D. & Walczak, B. (2011): Interpretation of analysis of variance models using principal component analysis to assess the effect of a maternal anticancer treatment on the mineralization of rat bones. *Anal. Chim. Acta, 689*, 1–7.

Stein, S.E. (1999): An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *J. Am. Soc. Mass Spectr., 10*, 770–781.

Sumner, L., Amberg, A., Barrett, D., Beale, M., Beger, R., Daykin, C., Fan, T., Fiehn, O., Goodacre, R., Griffin, J., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A., Lindon, J., Marriott, P., Nicholls, A., Reily, M., Thaden, J. & Viant, M. (2007): Proposed minimum reporting standards for chemical analysis. *Metabolomics, 3*, 211–221.

Szymansnka, E., Saccenti, E., Smilde, A.K. & Westerhuis, J.A. (2012): Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics, 8*, S3–S16.

Taira, E., Ueno, M., Saengprachatanarug, K. & Kawamitsua, Y. (2013): Direct sugar content analysis for whole stalk sugarcane using a portable near infrared instrument. *J. Near Infrared Spec., 21*, 281–287.

Tautenhahn, R., Bottcher, C. & Neumann, S. (2008): Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics, 9*, 504.

Tomasi, G., van den Berg, F. & Andersson, C. (2004): Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *J. Chemometr., 18*, 231–241.

Tomasi, G., Savorani, F. & Engelsen, S.B. (2011): icoshift: An effective tool for the alignment of chromatographic data. *J. Chromatogr. A, 1218*, 7832–7840.

Trygg, J. & Wold, S. (2002): Orthogonal projections to latent structures (O-PLS). *J. Chemometr., 16*, 119–128.

Uysal, R.S., Boyaci, I.H., Genis, H.E. & Tamer, U. (2013): Determination of butter adulteration with margarine using Raman spectroscopy. *Food Chem., 141*, 4397–4403.

Valdés, A., Simó, C., Ibánez, C. & García-Canas, V. (2013): Foodomics strategies for the analysis of transgenic foods. *TRAC Trend. Anal. Chem., 52*, 2–15.

Valenti, B., Martin, B., Andueza, D., Leroux, C., Labonne, C., Lahalle, F., Larroque, H., Brunschwig, P., Lecomte, C., Brochard, M. & Ferlay, A. (2013): Infrared spectroscopic methods for the discrimination of cows' milk according to the feeding system, cow breed and altitude of the dairy farm. *Int. Dairy J., 32*, 26–32.

van den Berg, F., Lyndgaard, C.B., Sorensen, K.M. & Engelsen, S.B. (2013): Process Analytical Technology in the food industry. *Trends Food Sci. Tech., 31*, 27–35.

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,* 142.

Van Deun, K., Van Mechelen, I., Thorrez, L., Schouteden, M., De Moor, B., van der Werf, M.T.J., De Lathauwer, L., Smilde, A.K. & Kiers, H.A.L. (2012): DISCO-SCA and properly applied GSVD as swinging methods to find common and distinctive processes. *Plos One, 7*, 5.

van Velzen, E.J.J., Westerhuis, J.A., van Duynhoven, J.P.M., van Dorsten, F.A., Hoefsloot, H.C.J., Jacobs, D.M., Smit, S., Draijer, R., Kroner, C.I. & Smilde, A.K. (2008): Multilevel data analysis of a crossover designed human nutritional intervention study. *J. Proteome Res., 7*, 4483–4491.

Veselkov, K.A., Lindon, J.C., Ebbels, T.M.D., Crockford, D., Volynkin, V.V., Holmes, E., Davies, D.B. & Nicholson, J.K. (2009): Recursive segment-wise peak alignmentof biological H-1 NMR spectra for improved metabolic biomarker recovery. *Anal. Chem., 81*, 56–66.

Vogels, J.T. W.E., Tas, A.C., Venekamp, J. & VanderGreef, J. (1996): Partial linear fit: A new NMR spectroscopy preprocessing tool for pattern recognition applications. *J. Chemometr., 10*, 425–438.

Westerhuis, J.A., Kourti, T. & MacGregor, J.F. (1998): Analysis of multiblock and hierarchical PCA and PLS models. *J. Chemometr., 12*, 301–321.

Westerhuis, J.A., Hoefsloot, H.C.J., Smit, S., Vis, D.J., Smilde, A.K., van Velzen, E.J.J., van Duijnhoven, J.P.M. & van Dorsten, F.A. (2008): Assessment of PLSDA cross validation. *Metabolomics, 4*, 81–89.

Willson, K.C. & Freeman, G.H. (1970): Use of principal component analysis on data from chemical analysis of tea leaves. *Exp. Agr., 6 (4),* 319–325.

Winning, H., Roldan-Marin, E., Dragsted, L.O., Viereck, N., Poulsen, M., Sanchez-Moreno, C., Cano, M.P. & Engelsen, S.B. (2009): An exploratory NMR nutri-metabonomic investigation reveals dimethyl sulfone as a dietary biomarker for onion intake. *Analyst, 134*, 2344–2351.

Witten, D.M., Tibshirani, R. & Hastie, T. (2009): A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics, 10*, 515–534.

Wold, H. (1975): Quantitative sociology: Intentional perspective on mathematical and statistical modeling. -in: Blalock, H.M, Aganbegian, A, Borodkin, F.M., Boudon, R., Capecchi, V. (Eds). *Quantitative sociology: Intentional perspective on mathematical and statistical modeling*. Academic Press, New York, pp. 307–357.

Wold, H. (1980): Model construction and evaluation when theoretical knowledge is scarce. Theory and application of partial least squares. -in: Kmenta, J. & Ramsey, J.B. (Eds) *Evaluation of econometric models*. Academic Press, New York, pp. 47–74.

Wold. S., Martens, H. & Wold, H. (1983): The multivariate calibration-problem in chemistry solved by the PLS method. *Lect. Notes Math., 973,* 286–293.

Wold, S., Esbensen, K. & Geladi, P. (1987): Principal Component Analysis. *Chemometr. Intell. Lab., 2*, 37–52.

Wold, S., Kettaneh, N. & Tjessem, K. (1996): Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection. *J. Chemometr., 10*, 463–482.

Wold, S., Sjöström, M. & Eriksson, L. (2001): PLS-regression: a basic tool of chemometrics. *Chemometr. Intell. Lab., 58*, 109–130.

Wold, S. & Sjöström, M. (1977): *Chemometrics: Theory and application,* 52th ed., American Chemical Society, pp. 243–282.

Xie, L.J., Ying, Y.B., Ying, T.J., Yu, H.Y. & Fu, X.P. (2007): Discrimination of transgenic tomatoes based on visible/near-infrared spectra. *Anal. Chim. Acta, 584*, 379–384.

Zafra-Gomez, A., Luzon-Toro, B., Jimenez-Diaz, I., Ballesteros, O. & Navalon, A. (2010): Quantification of phenolic antioxidants in rat cerebrospinal fluid by GC-MS after oral administration of compounds. *J. Pharmaceut. Biomed., 53*, 103–108.

Zhao, M., Downey, G. & O'Donnell, C.P. (2014): Detection of adulteration in fresh and frozen beefburger products by beef offal using mid-infrared ATR spectroscopy and multivariate data analysis. *Meat Sci., 96*, 1003–1011.

Zou, H., Hastie, T. & Tibshirani, R. (2006): Sparse principal component analysis. *J. Comput. Graph. Stat., 15*, 265–286.