CHEMOMETRIC METHODS IN PLANT METABOLOMICS

Keijo Isomaa
Master’s thesis
19.4.2013
University of Helsinki
Laboratory of Analytical Chemistry
This study focuses on chemometric analysis of instrumental data that has been obtained from chemical analysis of plant extracts. Chemometric analysis applies statistical and mathematical tools on chemical data, aiming to find new information or classifying samples in categories defined by the analyst. Chemometric analysis is based on computational pattern recognition and reveals any features that studied samples may have in common.

In the literature part of this study, chemometrics and relevant concepts closely related to it are first explained and four commonly used chemometric methods are introduced, namely principal component analysis, hierarchical cluster analysis, k nearest neighbors and soft independent modeling of class analogy. The text is written with emphasis on being easily understandable without prior knowledge on the subject. After introducing these concepts, the literature concerning metabolomic studies of plant extracts published in the recent ten years are reviewed. This literature commonly employs chemometrics, aiming to discover if two or more varieties of the same plant species have markedly differing metabolomes and whether they can be exploited to automatically recognize these varieties. Additionally, the chemometric approaches often attempt to discover what factors are causing the successful findings. The purpose of the literature survey is to concretely show how chemometrics can achieve these goals, and to learn what the most common ways to treat the analytical data prior to chemometric analysis are.

The experimental part applies chemometric methods to study bean extracts of the Ricinus communis plant, aiming to reveal if seed extracts of a same plant variety can be observed being similar, but clearly different from extracts of other varieties. Such situation could be exploited to develop a method that automatically identifies unknown seeds of the plant. The experimental work consisted of extracting homogenized samples with dilute aqueous acid, analyzing the extracts by three different instrumental techniques (liquid chromatography with ultraviolet light detection, liquid chromatography-mass spectrometry, and proton nuclear magnetic resonance spectroscopy) and finally analyzing the instrumental data by chemometric methods.

Chemometrics research suffers from nonexistent standard operating procedures, since there is no universal way to treat a sample or data derived from it. While the main steps are often same, the details of sample preparation and preprocessing of analytical data vary greatly and can have a significant impact on the outcome. Despite, the data preprocessing is often left partially or completely manifested. The experimental finding was that six varieties of Ricinus communis could be successfully discriminated by both principal component analysis and hierarchical cluster analysis, applied on chromatographic data, while the results for spectroscopic data were not successful. The results encourage continuing the research, but with more emphasis on peak alignment and further experimenting with the preprocessing of the spectroscopic data. Choosing different short segments of the original spectroscopic profile is suggested, to leave out excessive information that is not helpful in discriminating the plant varieties but could obscure the relevant information.
**Kemometriset menetelmät kasvien metabolomiikassa**

**Analyyttinen kemia**

**Pro gradu -tutkielma**

<table>
<thead>
<tr>
<th>Aika</th>
<th>18.4.2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sivunäärä</td>
<td>100</td>
</tr>
</tbody>
</table>


Kirjallisuusosuudessa selviytyy, mikäli kemometrissaan tutkimuksessa käytetään senänteitä, mitä se on ollut olemassa yhtä kaikkeen soveltuvaa näyteenkäsittelytapaa eikä mitataudata. Vaikka pääpiirteet ovat usein samoja, näyteenkäsittelyn ja mitataudan esikäsitellen yksityiskohtaiden vaihtelevat paljon ja niillä voi olla suuri vaikutus tutkimuksen lopputulokseen ja johtopäätöksiin. Tästä huolimatta, datan esikäsitteily jätetään usein osittain tai kokonaan kuvauimattomana. Kokeellisessa työssä havaittiin, että risiniakin kutsui tutkittuja lajitteita saadaan onnistuneesti eronnuttuja omiksi ryhmiksi sekä pääkomponenttianalyysiin, että hierarkkisen klusterianalyysiin avulla, kun analyysi tehtiin kromatografiselle mitataudatalalle, mutta spektroskoopiseelle datalle sama ei onnistunut. Tulokset kannustavat jatkamaan tutkimusta, mutta jatkossa suositellaan käytettävän enemmän aikaa dataan esiintyvien piikien kohdistamiseen. Lisäksi spektroskoopisen datan esikäsitteily tulee kehittää, jotta siinä heikkoa esiintyvää biologista informaatio saadaan kunnolla näkyville. Analyysin voidaan esimerkiksi valita muita lyhyitä pätkiä kokonaissektoria, jotta dataassa enimmäiseen esiintyvyyteen epäoleellista ja mahdollisesti mielenkiintoisen informaation allein peittävää informaatiota saadaan vähennettyä.

**Avainsanat / Keywords**

kemometria, kasvien metabolomiikka, pääkomponenttianalyysi, hierarkkinen klusterianalyysi

**Sähköposti / Where deposited**

Muita tietoja / Övriga uppgifter / Additional information
# TABLE OF CONTENTS

List of abbreviations .................................................................................................................... 3

1. Introduction ............................................................................................................................. 4

2. Pattern recognition .................................................................................................................. 8
   2.1. Unsupervised pattern recognition ................................................................................. 10
      2.1.1. Hierarchical cluster analysis ................................................................................ 10
      2.1.2. Principal component analysis ............................................................................. 13
   2.2. Supervised pattern recognition ..................................................................................... 18
      2.2.1. K nearest neighbours .......................................................................................... 20
      2.2.2. Soft independent modeling of class analogy ....................................................... 22
   2.3. Data preprocessing ......................................................................................................... 23
      2.3.1. Smoothing ........................................................................................................... 23
      2.3.2. Baseline correction .............................................................................................. 24
      2.3.3. Peak alignment .................................................................................................... 25
      2.3.4. Binning ............................................................................................................... 27
      2.3.5. Normalization ..................................................................................................... 28
      2.3.6. Centering ............................................................................................................ 29
      2.3.7. Scaling ............................................................................................................... 30

3. Chemometric approaches in plant metabolomics ............................................................... 31
   3.1. Common sample preparation and data preprocessing procedures .............................. 38

4. Study of differences between cultivars of *Ricinus communis* using chemometric methods ......................................................................................................................... 43
   4.1. Abstract ....................................................................................................................... 43
   4.2. Sample preparation ....................................................................................................... 43
4.3. Methods ........................................................................................................................................... 45
  4.3.1. High performance liquid chromatography ............................................................................. 46
  4.3.2. Liquid chromatography–mass spectrometry ........................................................................... 47
  4.3.3. Nuclear magnetic resonance spectroscopy .............................................................................. 49
  4.3.4. Datasets and preprocessing .................................................................................................... 50
4.4. Results and discussion ...................................................................................................................... 55
  4.4.1. General observations during experimental work ........................................................................ 55
  4.4.2. Results of hierarchical cluster analysis ................................................................................... 56
    4.4.2.1. Data from liquid chromatography with ultraviolet light detection .................................. 57
    4.4.2.2. Data from liquid chromatography–mass spectrometry .................................................. 60
    4.4.2.3. Data from proton nuclear magnetic resonance mass spectroscopy .................................. 63
  4.4.3. Results of principal component analysis .................................................................................. 66
    4.4.3.1. Data from liquid chromatography with ultraviolet light detection .................................. 66
    4.4.3.2. Data from liquid chromatography–mass spectrometry .................................................. 73
    4.4.3.3. Data from proton nuclear magnetic resonance mass spectroscopy .................................. 80
  4.4.4. Evaluation of the results ........................................................................................................... 81
5. Conclusions ........................................................................................................................................... 82
6. Acknowledgements ............................................................................................................................... 83
7. References ............................................................................................................................................. 84
8. Appendices ............................................................................................................................................ 90
LIST OF ABBREVIATIONS

Chemometrics
HCA Hierarchical Cluster Analysis
kNN K Nearest Neighbours
OPLS–DA Orthogonal Partial Least Squares–Discriminant Analysis
PC Principal Component
PCA Principal Component Analysis
PLS Partial Least Squares (synonymous to Projection to Latent Variables)
PLS–DA Partial Least Squares–Discriminant Analysis
RMSECV Root Mean Square Error of Cross-validation
SIMCA Soft Independent Modeling of Class Analogy
UPGMA Unweighted Pair Group Method with Arithmetic mean

Instrumentation
APCI Atmospheric Pressure Chemical Ionization
DAD Diode Array Detector
DART–MS Direct Analysis in Real Time Mass Spectrometry
ESI Electrospray Ionization
FTIR Fourier Transmission Infrared spectroscopy
GC–MS Gas Chromatography–Mass Spectrometry
HPLC High Performance Liquid Chromatography
1H NMR Proton Nuclear Magnetic Resonance Spectroscopy
LC–ITMS Liquid Chromatography–Ion Trap Mass Spectrometer
LC–MS Liquid Chromatography–Mass Spectrometry
LC–UV Liquid Chromatography with Ultraviolet light Detection
MALDI–FTMS Fourier Transmission Mass Spectrometer with Matrix Assisted Laser Desorption Ionization
MALDI–TOF/TOF Tandem Time of Flight Mass Spectrometer with Matrix Assisted Laser Desorption Ionization
Q–TOF Quadrupole–Time of Flight mass spectrometer
RPLC Reversed Phase Liquid Chromatography

Data preprocessing
COW Correlation Optimized Warping
MSPA Multiscale Peak Alignment
PAGA Peak Alignment with Genetic Algorithm
PARS Peak Alignment with Reduced set Mapping
PQN Probabilistic Quotient Normalization
RSPA Recursive Segment-wise Peak Alignment

Other
CWC Chemical Weapons Convention
MWCO Molecular Weight Cut-Off Filter
RCB Ricinus Communis Biomarker
TIC Total Ion Current/Chromatogram
TSP–d₄ Trimethylsilyl propionic-2,2,3,3-d₄ acid
1. INTRODUCTION

Metabolism within the cells of an organism is a product of its genome and habitat, and all metabolites synthesized by the organism form its metabolome.\(^1\) Since the metabolome is different between different organisms and even between individuals of the same organism, it can be exploited to make a distinction between subjects. For example, for a bunch of trees of same species it is possible to tell which of them have grown in a dry climate and which have grown in a humid climate, since the dry climate leaves its mark on the tree’s metabolome.\(^2\) Different approaches to study the metabolome of an organism are targeted analysis, metabolic profiling and metabolic fingerprinting.\(^1,3\) In targeted analysis the structures of the compounds of interest are known, the sample extract often goes through extensive clean-up to reach low limits of detection and quantification, and the compounds of interest are quantified using pure reference chemicals.\(^3\) Metabolic profiling, on the other hand, is measuring the levels of a pre-defined set of metabolites,\(^3\) e.g. organic acids, meaning that true quantitation and therefore reference chemicals are not required. For example, chromatographic peak areas could be measured for the metabolites of interest. One goal of metabolic profiling is to evaluate differences between samples.\(^4\) Metabolic profiling is a common practice used in biomedical sciences to monitor the health state of a patient, by analyzing biological fluids.\(^3\) The third approach – metabolic fingerprinting – does not try to identify or quantify the metabolites in the sample, but the total profile works as a unique fingerprint and is considered as being a snapshot of the metabolism in the cell or tissue.\(^3\) Computational pattern recognition tools are used to compare and classify the fingerprints by their similarity.

Studying the profiles of all or a significant number of cell metabolites is known as metabolomics.\(^3\) Alternatively, such studies are also commonly called metabonomics.\(^5-7\) Although these two terms are commonly used synonymously\(^5,7\), some experts\(^1\) of the field deny the identical meaning of these terms and provide more specific definitions. Metabolomics can be defined to be “comprehensive analysis in which all the metabolites of a biological system are identified and quantified”.\(^1\) Metabonomics, on the other hand, has been defined as “quantitative measurement of the multiparametric metabolic response of biological systems to pathology or genetic modification”.\(^8\) Therefore, by this definition
metabonomics involves disturbing the studied cell or tissue by pathological, chemical,\textsuperscript{9,10} or physical\textsuperscript{11,12} stimulus and studying the resulting changes in the metabolome. Based on this, it is quite easy to see the difference between the two methods, but it is easy to get confused by the literature that commonly use the two terms interchangeably.\textsuperscript{5} Therefore, when “metabolomics” is mentioned in the literature, there’s no way to know which of the two methods is meant, except by reading further to see what kind of experiment is in question. The matter is further complicated by the synonymous use of the term “metabolic profiling”.\textsuperscript{7} According to Fiehn,\textsuperscript{1} metabolic fingerprinting has been erroneously called metabonomics, although the two have completely different goals. As already discussed, metabonomics requires identification and quantitation of the metabolites, while metabolic fingerprinting does not even require identification. Fiehn also claims that metabolic profiling cannot be called metabolomics, unless a comprehensive portion of the metabolome is covered. Whether the metabolome is studied as such (as in metabolomics), or after some kind of stimulation (as in metabonomics), the analytical methods used to study the metabolome are same in both fields. Therefore, as far as the analytical methods are the subject of interest, it may be best to ignore whether the study is labeled “metabolomics” or “metabonomics”. Despite the critical appeals of some experts, most of the scientific community in the field seems to consider “metabolomics” simply as study of metabolome.

Study of a metabolome generates large amounts of data.\textsuperscript{3} Metabolic fingerprinting produces whole spectra or chromatograms that consist of vast amount of intensity values. To analyze the overwhelming amount of data, computational pattern recognition techniques are utilized. Pattern recognition techniques apply mathematical and statistical methods to summarize the data and to extract meaningful information from the bulk data. Analyzing chemical data with mathematical or statistical methods is a field of science known as chemometrics,\textsuperscript{5,13} often also referred to as multivariate statistical analysis,\textsuperscript{14} although the latter term is more general and does not restrict the data to be chemical data. In addition to the fields of metabolomics and metabonomics, chemometrics are commonly applied to quality control purposes. One good example is food industry, where multivariate methods are used \textit{e.g.} for food authentication.\textsuperscript{15} When cheap inauthentic products are labeled as expensive brands, the fraud may be detected by chemometric analysis. For example,
dilution or blending of whiskies can be detected with good sensitivity.\textsuperscript{16} Other applications include monitoring batch processes in chemical industry\textsuperscript{17} and even automatically discovering subtle changes in chromatographic purity of quality control runs.\textsuperscript{18}

Finnish Institute for Verification of the Chemical Weapons Convention, officially abbreviated as VERIFIN, has shown interest to adopt chemometric techniques and establish methods that ultimately aim at cultivar determination of the castor bean plant \textit{Ricinus communis}. Cultivar stands for “cultivated variety” that is reproduced vegetatively to remain identical to the parent plant\textsuperscript{19} and does not appear in the wild.\textsuperscript{20} If the plant is allowed to grow naturally, it is simply called “variety”. Methods for cultivar determination can potentially be utilized in forensics\textsuperscript{21,22} to link castor bean evidence to a particular source, geographic region, or batch.\textsuperscript{23} Around 250 different cultivars for \textit{Ricinus communis} exists,\textsuperscript{21} because the plant has been cultivated to serve different applications, such as to produce maximum amounts of castor oil for numerous industrial applications\textsuperscript{23,21} or to produce more aesthetic phenotypes for ornamental use.\textsuperscript{21, 23} Unfortunately, \textit{Ricinus communis} produces highly toxic ricin protein that is declared as a Schedule 1 chemical by the Chemical Weapons Convention (CWC).\textsuperscript{24} VERIFIN supports the disarmament of chemical weapons by developing identification methods for the scheduled chemical agents,\textsuperscript{25} such as ricin. The chemometric approach to study different \textit{Ricinus communis} cultivars was inspired by Ovenden \textit{et al.}\textsuperscript{21} and Pigott \textit{et al.}\textsuperscript{14} from the Defence Science and Technology Organisation (DSTO) of Australia. They showed that such methods made it possible to differentiate between wild \textit{Ricinus communis} seeds harvested from different geographical locations within Australia.\textsuperscript{21} Additionally, it was shown that seeds of known cultivars and provenance, harvested outside Australia, could be discriminated and random test samples could be successfully classified.\textsuperscript{14} In less technical terms, the data for all cultivars was not too similar, but it was possible to see a member of one cultivar being different from a member of another cultivar, and trained computer could determine both the cultivar and provenance for random test samples without a single mistake.

The aim of this thesis was to explore the literature on chemometrics and its applications, particularly focusing on cultivar discrimination, to pave the way for chemometric analysis of \textit{Ricinus communis} within VERIFIN. The main goal was to learn and gather information
on the requirements, common pitfalls and practical execution of chemometric analysis that could be beneficial for the future research at VERIFIN. Because the future research of VERIFIN aims to study both cultivar discrimination and classification, techniques for both approaches, namely *unsupervised* pattern recognition and *supervised* pattern recognition, are introduced. Two unsupervised techniques that are used in discrimination study were selected for this study, namely *Principal Component Analysis* (PCA) and *Hierarchical Cluster Analysis* (HCA). Similarly, two supervised techniques that are used in classification study were selected, namely *k Nearest Neighbors* (kNN) and *Soft Independent Modeling of Class Analogy* (SIMCA). These techniques were selected, because of being widely used and because the Laboratory of Analytical Chemistry had prior experience on these methods.

Literature on chemometric data analysis aimed at exploring differences within samples and published in the past decade is reviewed. Due to the large amount of publications, the review is not comprehensive, but only gives a general picture. In the experimental part of this study, discrimination between different cultivars of *Ricinus communis* was attempted by analyzing seed extracts with different kinds of instrumental techniques, followed by chemometric analysis. If sufficient discrimination is achieved, it is possible to train a computer to identify the cultivars and develop classification methods for automatic cultivar identification. These classification techniques were not applied, however, due to insufficient number of samples, although this was the original plan. Seed extracts were analyzed by liquid chromatography with UV detection (LC–UV), liquid chromatography–mass spectrometry (LC–MS) and proton nuclear magnetic resonance spectroscopy ($^1$H NMR). Metabolic fingerprinting approaches were applied on LC–UV chromatograms, LC–MS total ion chromatograms and NMR spectra. Additionally, the metabolic profiling approach was attempted, where a table of peak areas extracted from the LC–MS data was subjected to chemometric analysis. The review begins with introduction of the selected pattern recognition techniques. Instrumental techniques are assumed familiar and are not described to keep the paper more compact.
2. PATTERN RECOGNITION

Humans are good at seeing differences between shapes of different objects. For example, a child can tell the difference between a circle and a rectangle at very young age. However, any patterns or trends in a table full of figures is difficult to see and therefore visualization of the numbers by plotting them into a graph is often carried out, resulting in geometric shapes better suiting the human pattern recognition. Such operation is indeed useful, but such visualization is restricted to three dimensions. It is possible to draw a three-dimensional graph, but not a four-dimensional graph. This means that only data consisting of three or less columns, or variables, can be plotted at the same time and so any patterns in higher dimensional data are not easily seen.

The pattern recognition of a computer is the opposite of human pattern recognition: computers are very good at comparing numerical data in the forms of tables. While a chromatogram has geometric shape and is easily comprehensible to a human, the true form of the chromatographic data is a table having two rows and as many columns as there are data points (or more specifically measurements taken from the sample). One of the two rows contains the intensity value (Y-data) for each measurement and the other contains the time from the start of the analysis (X-data).

Data to be used for computational pattern recognition has to be organized into a dataset table so that each sample is lying on a separate row and the dataset has as many columns as there are data points. The data points are separate measurements or variables, and can be anything from weight or temperature values to peak areas or intensity values. If the sample data is a chromatogram, only the relevant row, being the row containing the intensity values, is used (since the row containing the time values is identical to all samples due to equal amount data points and equally long analysis time). When each sample data is forming a separate row in the dataset table, the data is forming a “row space”. The row space can be visualized as long as there are three or less measured variables per each sample. For example, Figure 1a is showing the row space of a dataset consisting of two samples (rows), both having only two measured variables (columns). In the visualization each sample in the dataset is represented by a point in the graph. The visualization is
carried out so that there’s a separate axis for each variable and the variable values are the coordinates of the points. The comparison of the samples is based on the distance between the points (samples) in the row space: the closer the samples are to each other in the row space, the more similar the samples are considered. For example, two samples that have all their variable values identical are located in the same point in the row space (the distance between them is zero). If there was a third measured variable, the data could be visualized as in Figure 1b, where the graph has three axes. If there are more than three variables, e.g. four variables, the data can no longer be visualized for humans, but a computer can still compare the variables using the same routine it has used for the two- and three-dimensional data. There can be thousands of variables to compare, as can be the case with chromatograms or spectra, and the computer will still analyze the data the same way, only taking a little more time now. Therefore, the pattern recognition ability of a computer is not limited to three dimensions, as is the case for man.

\[
R = \begin{bmatrix}
2 & 3 \\
1 & 1
\end{bmatrix}
\]

\[
R = \begin{bmatrix}
2 & 3 & 4 \\
1 & 1 & 3
\end{bmatrix}
\]

**Figure 1.** The row space plot of a matrix with two columns (a) and three columns (b). The figures are adopted from “Chemometrics: a practical guide” by Beebe et al.

In computational pattern recognition techniques, the computer compares the samples in the dataset – a huge table full of numbers – and reports the found patterns in a simple graph suitable for human pattern recognition. When the goal of pattern recognition is just to see if some samples are similar and differ from other samples, no pre-known information on the identities of the samples is used in the analysis and therefore the method is said to be
unsupervised. On the other hand, if the goal is to find out whether the samples belong to any pre-defined class, information on samples that define these pre-defined classes is utilized and the method is then said to be supervised.

2.1. Unsupervised pattern recognition

Unsupervised pattern recognition techniques are often used to see similarities and dissimilarities between samples. It is interesting to see if some samples are clearly different from the rest. Such result would raise questions like what these samples have in common and what makes them special and launch investigations that never would have been carried out without knowing the hidden pattern in the data. Two widely used unsupervised pattern recognition techniques, Hierarchical Cluster Analysis (HCA) and Principal Component Analysis (PCA) are described in the following.

2.1.1. Hierarchical cluster analysis

Hierarchical Cluster analysis (HCA) is a pattern recognition technique that groups objects in the analyzed dataset into clusters solely based on the information in the numeric data. The objects discussed here can be thought to be different measurements or samples, for example. More specifically, the objects (samples) are points in row space, their coordinates being all the measured variables, and the nearness of the samples in the row space is the basis for clustering. In research use, the goal of HCA usually is to discover hidden structures in the data in the form of groupings. The clusters are formed from the objects without any knowledge of what these objects are, and therefore HCA is unsupervised, making it an unsupervised pattern recognition technique. Since no pre-known information describing what the objects are is used in the grouping, all objects in the dataset are considered as equals and if some of them are forming a distinct cluster, there’s something in common with those objects. What the common property is cannot be determined by HCA – only information on what samples/objects form a cluster is gained, and what the objects have in common is for the analyst to determine based on any knowledge he has of the samples. More specifically, information on the samples, such as sample codes or
sampling site information usually is included in the dataset to be analyzed but none of that information is used by the algorithm that calculates the clustering.

What exactly is a cluster is a relevant question. **Figure 2a** shows plotted data for twenty objects and their clustering, depending on the criteria that define a cluster. As visualized by **Figure 2b–d**, various amounts of clusters can be formed from the same set of points. At least two clusters can be seen in the plotted data, but these two clusters can be further divided into sub-clusters, giving a total of four or six clusters, for example. The question is, are these sub-clusters truly meaningful clusters or are those patterns merely formed by chance and just look like clusters? The more distance there is between the clusters and the less distance between objects within a cluster, the better is the clustering. Dozens of different clustering algorithms exists that differ in the way the clusters are defined and how the distances between the clusters are calculated, and changing the algorithm often results in different clustering. Therefore, the type of clustering algorithm needs to be suitable for the application.

![Figure 2](image-url)  
**Figure 2.** Different ways of clustering the same set of points. The figure is adopted from *Introduction to Data Mining* by Tan et al. (Addison-Wesley 2005).

HCA is an **agglomerative** method, meaning that it starts with each object being its own cluster, calculates distances to all other clusters and always merges two clusters based on the method’s linkage rule (e.g. two nearest clusters) to form a larger cluster until there’s only one major cluster left in the end.13,29 “Agglomerative method” indicates starting with many clusters and ending up with a single cluster in the end. Some other forms of cluster analysis are having the opposite approach, starting with a single large cluster and dividing it to smaller clusters, which is known as **partitional** or **divisive** method,29 but the focus here is
solely on HCA and other methods are not discussed here any further. The clustering of a dataset, calculated by the chosen clustering algorithm, is visualized by a dendrogram: a branching diagram illustrated in Figure 3. The samples are listed at the left in the figure and the axis below the dendrogram is describing the distance between two clusters. Samples or clusters merged near the left side of the dendrogram are interpreted as being similar and those merged near the right side as very different. The dendrogram in the figure is interpreted as follows: the distance between samples \(a\) and \(b\) is shortest so these samples are most similar. Together samples \(a\) and \(b\) form cluster \(A\). Distance of sample \(c\) to cluster \(A\) is quite short and therefore sample \(c\) is similar to samples \(a\) and \(b\). Together sample \(c\) and cluster \(A\) form cluster \(B\). Samples \(d\) and \(e\) are significantly more different than samples \(a\) and \(b\), for example, but still somewhat similar and together samples \(d\) and \(e\) form cluster \(C\). Cluster \(C\) is very different from cluster \(B\), but if these clusters were joined – despite their distances being very long – a single cluster \(D\) would be attained. Obviously, cluster \(D\) would be very non-ideal, since there are huge distances between samples inside the cluster and therefore cluster \(D\) is not likely to be relevant. Nevertheless, this is how the agglomerative algorithms work: merging clusters until there’s only one single cluster left. On the other hand, the five samples seem to separate into two distinct clusters \(B\) and \(C\) that could be interesting.

![Dendrogram](image)

**Figure 3.** A simple dendrogram for a small hypothetical dataset, consisting of five samples \(a\)–\(e\). The calculated distance between clusters (\(A\)–\(D\)) or samples (\(a\)–\(e\)) increases from left to right.

Doing hierarchical cluster analysis requires the selection of a distance measure and a linkage method for linking two clusters. The most common distance measure used is the Euclidean distance\(^{30}\) that is the actual geometric distance between two points – something
that can be measured with a ruler, for example. Examples of possible linkage methods are Nearest neighbor, Furthest neighbor, Pair-group average, Centroid, Median and Ward’s method. For example, for the Nearest neighbor linkage rule the distance between two clusters is defined as the distance between the two closest objects of the two clusters, while for the Furthest neighbor rule the distance between the two furthest objects is calculated, respectively. The former tends to perform well for elongated chain-like cluster, while the latter usually excels with round, distinct clusters. Ward’s method differs from the rest of the methods in that instead of looking at the distance between two clusters it selects the two clusters to be merged so that the within-cluster variance of the resulting cluster is minimal and therefore the method is also known as Ward’s minimum variance method. Ward’s method is generally considered to be very efficient and is said to generally produce small clusters and being the most common approach to doing hierarchical cluster analysis. It is not possible to say what linkage method works best, because that depends on the shape of the clusters, but if the goal is to learn as much as possible from the data, using more than one method and comparison of the results is recommended.

2.1.2. Principal component analysis

Principal Component Analysis (PCA) is a way to summarize the original data with a large amount of variables to simpler data with only a few variables. As a result, a simple model that approximates the original data is constructed. Similar to HCA, PCA is often used to find patterns in the data – to see if some samples differ from others, observed as clusters of samples located at a distance from other samples in plots called scores plots.

As discussed earlier, the samples in the dataset form a multidimensional cloud of points in row space. If there were a great number of samples in the dataset, but only three variables were recorded for all samples, the cloud of points (samples) would be three dimensional and could be imagined to look like a swarm of bees. Let’s consider a case where the cloud of points has a symmetrical shape of French bread (long, narrow and flat). The direction where the bread is “pointing” in the row space is not relevant here. If the spread of the points is considered, the spread is smallest in the bread’s “height” direction, since the bread was said to be flat. The spread is slightly larger in the “width” direction, but is by far greatest in the “length” direction of the bread, since the bread is very long compared to its
width or height. In other words, the distance between individual points is farthest in the “length” direction of the bread-shaped cloud of points.

Principal component analysis finds the directions with the largest variance, meaning the largest spread of the points, and fits a line into that direction in such a way that that the minimum squared distance from the line to each point is minimized\textsuperscript{13} and therefore the operation is analogous to fitting a calibration curve to a set of measurement points with the least squares method. Such fitted line is called Principal Component (PC). Fitting the PCs for a hypothetical two-dimensional data is illustrated in Figure 4. In the discussed case of the data cloud in the shape of a French bread, the first principal component would be fitted along the “length” direction of the cloud. The second principal component would be fitted along the direction with next largest spread of points (variance), being the “width” direction of the cloud in this case. Finally, the third and last principal component would be fitted along the “height” direction of the cloud, since the spread of points is small when the bread is flat. This way the cloud of points is given a new coordinate system, the new axes being the principal components PC1, PC2 and PC3. The coordinates of the points with respect to the original axes – one axis for each measured variable – were the numerical values of those variables for the samples. The coordinates of the points in the new coordinate system, whose axes are the principal components, are called \textit{scores}.\textsuperscript{13} For example, the points at the farthest end of the bread along the PC1 axis (the “length” direction of the cloud) are far away from centre of the bread (the origin) and therefore the scores for those points is said to be high. Similarly, the scores of the points near the centre of the bread are said to be low. The example discussing the “French bread”-shaped cloud of points is very artificial and was only introduced here to demonstrate how the principal components are fitted to the cloud of points in PCA. The clouds of points from real scientific data are not symmetrical and have unimaginable shapes for being highly multi-dimensional and therefore the French bread example will no longer be referenced in the following.
Figure 4. Plot of a hypothetical two-dimensional data. The direction with the maximum variance is modeled by PC1 that is constructed by the least squares method. PC2 is modeling the next largest variance of the data, while being orthogonal to PC1 by definition. The figure is adopted from “Chemometrics: a practical guide” by Beebe et al.\textsuperscript{13}

The relative directions of the original variable axes to the principal component axes in space are telling how strongly the variables influence the principal components.\textsuperscript{13} For example, if PC1 is perfectly aligned with one variable axis, it means that PC1 is explaining all the variance coming from that variable. On the other hand, if PC1 is orthogonal (90º) to a variable axis, it means PC1 is explaining none of the variance coming from that variable. When the angle between the PC axis and the variable axis is zero (perfectly aligned) the cosine of that angle equals 1, and when the angle is 90º (orthogonal) the cosine is 0. The cosine values that vary between 0 and 1 are called loadings and describe how strongly a variable is contributing to a specific PC.\textsuperscript{13}

As already mentioned, PCA aims to describe as much of the data as possible with as little principal components as possible, to extract most of the interesting data from the vast amount of initial data that is mostly irrelevant. In other words, a PCA model that consists of only a few principal components is created to approximate and simplify the data. The PCs are fitted so that the first principal component is aligned with the direction with the most variance (the direction where the spread of the points is largest). The second PC is fitted so, that it goes to that direction orthogonal to PC1 that has the next largest variance in the points. The third PC and all the rest of the PCs are always orthogonal to all of the preceding PCs and similarly go to the direction of the largest remaining variance. The PCs are in order of significance: PC1 describes the most variance and is therefore most important. Because there is random noise distributed evenly to the whole row space and the amount of variance explained by the first PC is largest, the signal-to-noise ratio of PC1 is highest of
all PCs, decreasing for each following PC. Eventually, the later PCs have signal-to-noise so low that they mostly describe just noise and are therefore useless. Determining which PCs contain relevant information and which should be discarded as uninformative is up to the analyst.

The most interesting information gained from PCA is the scores and loadings. As already discussed, the distances of samples (points) from a particular PC are called the scores on the respective PC. The scores are holding information on the samples. Samples with similar scores are close to each other in the row space and are therefore somehow similar. The goal of PCA is to summarize and visualize multivariate data and this visualization is given in the form of scores plots. The loadings on the other hand, are holding information on the variables and are a measure of alignment of the particular PC with the variable axes. Variables with high loadings on a particular PC are well described by that PC. The loadings plots make it possible to find out why some samples are being similar and forming clusters in the scores plots, which is a great advantage of PCA over HCA.

The variance of an original variable that is not described by the PCA model is called residual variance, or simply residual. For example, if the first three PCs describe a total of 90% of the variance in the data, the total residual, or variance unexplained by the model, is 10%, and is the sum of the residuals of all variables on all three PCs.

Determining the appropriate number of PCs to use in constructing the PCA model is the main challenge of PCA. Using too many PCs adds too much noise to the model, while using too few PCs leaves essential information out of the model, which both are detrimental to loadings interpretation. No fool-proof way of determining the number relevant PCs exists, but there are several rules of thumb that provide useful suggestions that should be used in combination rather than selecting one to follow, and two concrete ones are presented here. One is plotting the eigenvalues of the PCs or amount of variance explained by the PCs against the principal component number that both give a declining curve, where the point (PC) after which the explained variance does not grow significantly (the steepness of the curve is significantly reduced) is the last relevant PC. Another concrete way to determine the relevant PCs is to cross-validate the model. Cross-validation excludes one or more samples from the model, calculates a new model from the remaining samples
and then tries to estimate the excluded samples with the new model.\textsuperscript{31} The resulting cross-validation residuals are a measure of the error of model reconstruction.\textsuperscript{13,31} Cross-validation will produce a plot where the Root Mean Square Error of Cross-validation (RMSECV), also called reconstruction error,\textsuperscript{31} is plotted against the number of principal components retained in the model and the plot is again a declining curve. The prediction error is based on the residuals: when too few PCs are included in the model, it approximates the data too much, meaning that the residuals for the variables are large (too much variance is left unexplained).\textsuperscript{13} When too many PCs are involved, the prediction error no longer decreases but may even increase due to inclusion of PCs that more or less describe only noise.\textsuperscript{31} A PC whose inclusion increases the prediction error (RMSECV) should not be included in the model. The RMSECV is calculated from the combined residuals for all variables and its value depends on the number of PCs included in the model.\textsuperscript{31}

The presence of outliers has a negative influence on the PCA model, but are sometimes quite easy to spot in the scores plots.\textsuperscript{34} Outliers are samples that are unique in the data that could be due to errors in sample preparation or measurement, for example. Another possibility is that no error had occurred, but the outlier sample is simply different from all the rest. In the latter case, the outlier should either be removed from the data and the model be recalculated, or more samples similar to the outlier sample should be analyzed and added to the model.\textsuperscript{31} An outlier-free scores plot is illustrated in figure Figure 5a, while an outlier sample with very high score on PC1 is found in the scores plot of Figure 5b. The negative effect of the outlier is that PC1 completely focuses on modelling the outlier sample and is not helpful in discriminating the other samples. This is seen in the samples locating at similar distances from the origin in the PC1 direction (x-axis, labelled as $t_1$ in the figure). In contrast, the samples are located at different distances from the origin in the PC2 direction (y-axis, labelled as $t_2$ in the figure), meaning that PC2 is helpful in sample discrimination. PC2 is unaffected by the outlier, because the outlier has almost zero score on PC2 and is therefore practically completely unmodeled by PC2. Strong outliers are found outside the 95\% confidence ellipse in the scores plots. The 95\% confidence limit means that 5 samples out of 100 are expected to be seen outside the confidence ellipse and therefore samples outside it are not necessarily outliers.\textsuperscript{34}
To summarize and compare the two reviewed unsupervised pattern recognition techniques, PCA and HCA are both used for same purpose, but PCA is obviously more complicated. According to the author’s own experience, there are more terms to understand and more things realize and take into account in PCA than in HCA. Therefore doing HCA is a more convenient way to explore new data. However, PCA offers the possibility to observe what properties the clustered samples have in common, which is why it is not advised to choose between the two methods.

2.2. Supervised pattern recognition

The difference between supervised pattern recognition techniques and unsupervised ones is that unsupervised methods have no knowledge on the samples, while supervised methods have been trained to distinguish between different kinds of samples. In other words the objective of supervised methods is to classify samples to different classes. The class could be “gender”, for example, and the method could then try to classify the samples between males and females based on the knowledge it has got from previously analyzed samples. An example of such a study is the successful predicting of the gender of tamarin monkey from its scent mark, analyzed by GC–MS and followed by supervised pattern recognition.

Supervised pattern recognition comprises of three phases: training, validation and finally applying the model on unknown samples. In training phase, sufficiently large set of known samples, called training set, is introduced to the computer. The amount of samples is
adequate when they represent well the different possible variations within the class. Consider, for example, a child who knows nothing about cars being taught to tell if a car driving by is a Mercedes or not. If he was only taught a few most common models seen on the road, he could probably do well most of the time, but every time a rare model drives by he would fail. The question is then, are occasional misclassifications acceptable or should they be avoided at all costs? If the latter is the case, the amount of samples in the training set has to be large, otherwise a few is enough, depending how abundant different variations are within a class, of course. The benefit from increasing the training set size gets smaller as the size increases. The additional benefit comes from the increased ability to classify rare samples, while the more common samples are already well represented by the large training set. Therefore, it could be true for some applications that for a training set already comprised of 500 samples, for example, adding 100 additional samples could have a negligible benefit while analyzing the additional 100 samples could be very laborious.

After the computer has been taught to classify unknown samples with the knowledge within the training set, the resulting classification model needs to be validated. The validation is done using another set of known samples, called validation set or test set, and the purpose is to evaluate whether a correct classification was just a lucky guess or if the classification is reliable. Validation yields data on the ratio of correct classifications and also whether the computer was e.g. 50 % or 90 % sure of a classification that proved correct. The validation set also needs to be of sufficient size to represent well all possible kinds of samples and it must not contain any samples that have been used to train the computer. Otherwise the classification seems to work a lot better than it actually does, as would a student who had obtained the questions prior to taking an exam also perform better than expected, for example. It is also possible to validate the model by cross-validation, instead of using a validation set. The cross-validation approach is recommended when samples are hard to get, expensive or limited. The rationale behind this recommendation is that if the amount of samples is low, not many samples can be spent for validation, which does not result in trustworthy assessment of model fit but samples that could be used to train the model are wasted instead. More information on the cross-validation approach is given in the paper by Hawkins et al.
When the model is found to be sufficiently reliable, it is ready to be used for classifying unknown samples. In an ideal case, the computer can now tell what class an unknown sample belongs to. For example, the computer could determine which olive tree variety a sample represents, by analyzing the data acquired from a leaf sample.\(^2\) If the classes under study are observed to be well separated with unsupervised methods, \textit{i.e.} forming tight clusters distant from other clusters, the classification with supervised methods is also likely to be reliable. When the classes form clusters that are located near each other, the supervised methods may also find it hard to tell which class a sample belongs to.

Two widely used unsupervised pattern recognition techniques, \textit{k Nearest Neighbours} (kNN) and \textit{Soft Independent Modeling of Class Analogy} (SIMCA) are described in the following.

\textbf{2.2.1. K nearest neighbours}

Like other supervised pattern recognition techniques, \textit{k Nearest Neighbours} (kNN) is used for classifying unknown samples to predefined categories.\(^3\) For kNN the classification principle is very simple. The classification is based on the distances between samples in the row space. As already stated, there are many different distance measures of which Euclidean distance, being simply the geometric distance between samples, is often used\(^*\). The distance of an unknown sample is calculated to \textit{k} nearest known samples that have been used to train the computer. If the value of \textit{k} is chosen to be “1”, the unknown sample is classified as being the same as the class of a sample nearest to it in row space. Such choice works well if the classes are well separated, but if two or more classes overlap the closest sample may well belong to a different class than the sample to be classified really is and using a bigger value for \textit{k} could yield more correct classification. In that case, distance to multiple samples is calculated and the unknown sample is assigned to the class with most nearby samples. The optimal value for \textit{k} can be determined by cross-validation,\(^3\) such as \textit{leave-one-out} cross-validation. The leave-one-out algorithm leaves one sample out of the

\(^*\) Other distance measures, such as \textit{Mahalanobis distance}, can be used for emphasizing some variables over others, for example.\(^3\)
training set and predicts the class of that sample by comparing its distance to the samples in the training set. The operation is repeated so that every sample has been left out and its class been predicted once and the whole thing is repeated using different values for \( k \). From the cross-validation results the \( k \) value with least misclassifications is then selected.\(^{13}\) After the value for \( k \) has been chosen, the model is validated to see if the classification results are trustworthy. Even if all samples were correctly classified, further evaluation should be done. Firstly, it is useful to check how many of the \( k \) nearest neighbours belong to the assigned class. For example, if \( k = 5 \) and all the five nearest neighbours belong to the same class, it is likely that the classification is reliable. In contrast, if only three of the five nearest neighbours belong to same class and two are something else, the probability of misclassification is higher; in this case, the confidence of correct classification would only be 60%.

Even if all the five nearest samples belonged to same class and the classification confidence would therefore be 100%, additional validation is required. This is because even if the five closest samples were all of the same class, they could still be very distant, implying that the sample to be classified would not belong to any of the classes included in the training set. Nevertheless, the sample would be classified according to the closest class and would therefore be misclassified. Another kind of validation that takes the distance to the nearest class into account should be used in addition. For example, if the unknown sample was \( X \) and the nearest class was \( A \), the average distance between samples of class \( A \) in the training set and perhaps their standard deviation could be calculated. Then, the distance of \( X \) from the nearest class \( A \) sample could be calculated and if the distance was significantly longer than the average distance between class \( A \) samples, the classification of \( X \) as being of class \( A \) could be considered unreliable. One way to quantify how well sample \( X \) to fits to class \( A \) is to use Equation 1.\(^{13}\)

\[
G = \frac{d_X - \bar{d}_A}{\sigma_A}
\]

Equation 1,

where \( d_X \) is the distance from the unknown sample \( X \) to the nearest class \( A \) sample, \( \bar{d}_A \) is the average distance between known samples of class \( A \) in the training set, \( \sigma_A \) is the standard deviation of the distances between the samples of class \( A \) and \( G \) is the “goodness”
value for the classification. Such goodness value would be small for good classification and large for bad classification.\textsuperscript{13} An appropriate value for the highest acceptable goodness value is determined experimentally.

\textbf{2.2.2. Soft independent modeling of class analogy}

\textit{Soft independent modeling of class analogy} (SIMCA) is a supervised pattern recognition technique that is based on PCA,\textsuperscript{13,31} which makes it a more complex technique than kNN. In contrast to kNN, in which only one model is constructed from the whole training set data, SIMCA requires a separate model to be constructed for each class.\textsuperscript{13} Therefore, if the training set includes samples from 10 different classes, 10 PCA models will need to be constructed that together form the SIMCA model. Therefore, enough samples to sufficiently represent each class is required for each PCA model. Additionally, the appropriate amount of PCs to retain in each PCA model needs to be determined with the methods previously described for PCA. As was already discussed, selecting either too few or too many PCs to keep is detrimental to the model and such choice is needed for every class in the training set. Needless to say, classifying unknown samples by using SIMCA is a lot more laborious and complicated than by using kNN.

When an unknown sample is classified using SIMCA, the sample is separately compared to each class to determine if the sample fits the PCA model of any class.\textsuperscript{13} It is therefore possible that the sample fits to more than one PCA model and the sample is then classified as belonging to multiple classes. For example, both “monkey” and “snake” can be classified as “animal”, but only monkey can be classified as “mammal”. In contrast to SIMCA, kNN always classifies a sample as belonging to one class – the one that is closest to the unknown sample no matter how distant that class is in row space.\textsuperscript{2,13} Therefore, if the computer was only trained to classify monkeys and snakes and it was then asked to classify an elephant, kNN could easily classify it as a “monkey”, while SIMCA would say the sample is neither. This is why it does matter what kind of supervised method is selected for classification and the right choice depends on the nature of the classes involved. If the classes are mutually exclusive, kNN works well, but if not, SIMCA should be used for classifying the samples instead of kNN.
2.3. Data preprocessing

Data is often ‘cleaned’ prior to chemometric analysis. Such ‘cleaning’ is usually called \textit{signal preprocessing} or \textit{data preprocessing}. Signal preprocessing is a very important part of chemometric analysis,\textsuperscript{13,34} but care should be taken not to create wrong knowledge or remove important information in the process.\textsuperscript{13,38} The goal of signal preprocessing is to remove measurement noise and only leave the relevant biological information for analysis.\textsuperscript{13,38} If done right, data preprocessing can make analysis easier, more robust and more accurate.\textsuperscript{7} Sometimes irrelevant disturbing variation from the data acquisition can completely hide the relevant variation, making data preprocessing mandatory.\textsuperscript{13} Therefore, if no discrimination between samples is observed although expected, some discrimination could be seen after appropriate data preprocessing. Examples of different data preprocessing operations are \textit{smoothing}, \textit{baseline correction}, \textit{peak alignment}, \textit{binning}, \textit{normalization}, \textit{transformation} such as logarithmic transformation, \textit{centering} and \textit{scaling}. Van den Berg calls these operations that are used for cleaning the data “signal \textit{preprocessing}” or “data \textit{preprocessing}” and call operations that only prepare the data to better suit the chemometric analysis “data \textit{pretreatment}”.\textsuperscript{39} According to this, transformations, centering and scaling would rather be called data pretreatment operations than data preprocessing. However, synonymous use of all these terms seems to be a lot more common in the literature, meaning any operation done to the data. The meaning of different data preprocessing operations is explained in the following.

2.3.1. Smoothing

Spectra or chromatograms can be smoothed to reduce \textit{random high-frequency noise} from the signal and so improve the signal-to-noise ratio.\textsuperscript{13} One commonly used smoothing algorithm is a running polynomial smoother known as Savitzky-Golay algorithm.\textsuperscript{13} Like many other smoothing algorithms, Savitzky-Golay algorithm requires definition of a smoothing window, but selection of a proper polynomial order is also required. The polynomial order defines what kind of polynomial is fitted to the data window. Choosing appropriate values for the window size and polynomial order is very important and they have to be determined by trial and error. As the window width is increased, more noise is smoothed, but too wide a window leads to removal of sharp peaks and distortion of the
remaining peaks.\textsuperscript{13} For a comparison of different smoothing algorithms, the book by Beebe \textit{et al.}\textsuperscript{13} is recommended.

\subsection*{2.3.2. Baseline correction}

In addition to random noise, a spectrum or chromatogram may also contain systematic baseline patterns that are not due to the chemical signal of the sample, but rather caused by other sources.\textsuperscript{13} The baseline patterns are mostly due to \textit{chemical noise} and are matrix dependent.\textsuperscript{38} If the chemical noise is characteristic to certain class of compounds, the baseline could become a significant discriminating factor in chemometric analysis,\textsuperscript{38} although the discrimination should be originating from biological information in the signals of interest. Another possible consequence of chemical noise is signals of interest being obscured by baseline.\textsuperscript{38}

One example of chemical noise is a linearly increasing baseline signal that is caused by a change in eluent composition when gradient elution program is used in chromatography. Another example is a contaminant in the mobile phase, which would raise the baseline signal by some constant amount from the value that would be observed with a clean mobile phase. To clean the signal from those contributions not coming from the compounds of interest and so remove irrelevant variance, baseline correction can be done.

One simple way to do baseline correction is to select one or preferably more points that only represent baseline, calculate their average intensity and then simply subtract this value from all variables.\textsuperscript{13} This would only apply to constant baseline. If the baseline was a linear slope, then a line would be estimated from one or more points that only contain baseline and the line would be subtracted from the signal.

Another way to eliminate baseline from a spectrum or chromatogram is \textit{differentiation}.\textsuperscript{13} Differentiation is a purely mathematical way to remove baseline and may result to the processed signal no longer being similarly interpretable that it originally was. Equation 2 shows a hypothetical signal, a function where $y$ can be thought as being the intensity and $x$ being the time. The first term, $Y(x)$ represents the true signal and the rest of the terms represent the baseline. A constant part of the baseline signal is represented by the term $a$, a linear part is represented by the term $bx$ and so on.
\[ y(x) = Y(x) + a + bx + cx^2 + dx^3 \ldots \]  
\text{Equation 2}

If the Equation 2 is differentiated, according to the common differentiation rules the constant term \( a \) in the equation becomes zero and so a constant contribution of a contaminant compound to the signal would be completely eliminated from the signal. The differentiated signal, \( \text{i.e.} \) the derivative, is represented by Equation 3.

\[ y'(x) = Y'(x) + b + 2cx + 3dx^2 \ldots \]  
\text{Equation 3}

A possible linear increase in the baseline – caused by a change in eluent composition, for example – that originally was described by the term \( bx \) has now been transformed to \( b \). If the Equation 3 is now differentiated, the term \( b \) becomes zero and thus another irrelevant part of the signal has been eliminated. Although two kinds of irrelevant variation have now been eliminated by taking the second derivative of the original signal, the preprocessed signal now looks very different from the original and can’t be recognized to represent a chromatogram anymore. In conclusion, the signal is cleaned from non-informative systematic baseline features with the expense of interpretability.

Differentiation of noisy signal results in poor signal-to-noise ratio of the resulting signal.\textsuperscript{13} Therefore a smoothing operation should be done prior to differentiation and choice of a proper window is again necessary. It is recommended that smoothing and baseline correction is done using system-specific software applications.\textsuperscript{34}

\textbf{2.3.3. Peak alignment}

As described earlier, the datasets that are created for analysis by pattern recognition techniques are large tables full of figures. Each sample fills one row and the amount of rows in the table is equal to the amount of samples in the dataset. A simple dataset could consist of three samples that could be studied persons, for example. Three variables for each person could be studied, their height, weight and age for example. Height could be stored in the first column of the dataset table, weight in the second and age in the third column. Computational pattern recognition techniques compare the values within a column, the heights of the persons, for example. If the analyst accidentally stored one or more age values to the first column that is supposed to contain only height values, the computer
would compare height values to age values and meaningless results would be calculated since the values are not comparable as they represent different variables. The purpose of this example is to point out that same variables should always be compared. Such mistake by the analyst is really not a significant concern, but there are analytical issues that could lead to such comparison of wrong variables.

Both in chromatography and NMR spectroscopy, shifts in peak positions are possible. In liquid chromatography, retention time drifts may result from variation in temperature, pressure or mobile phase (composition or flow rate), changes in stationary phase, or sample matrix effects. The retention time issues are similar in gas chromatography. In NMR spectrometry, the shifts have been reported to be due to variations in the background matrix and instrument instabilities. Often the variation in peak positions in biological samples is due to variation in pH and intermolecular interactions, both in liquid chromatography and NMR spectrometry, and adjusting the pH of the samples may solve the problem. The data can be binned to decrease the effect of varying peak position to the variables. The meaning of binning is explained later.

In the ideal case, peaks in chromatograms or spectra are perfectly aligned so that their apex intensity values are contained within the same column of the dataset table. In reality, peak apexes are not perfectly aligned and may be found at slightly different positions due to chromatographic drift or other reason. Poor alignment introduces error, since the compared variables, e.g. variables supposed to be peak apexes, could for some samples be points quite far away from the true apex and can lead to erroneous comparison results. This can be illustrated by measuring the heights of different persons, for example, but taking the measurements from the top of the head for some people and from the shoulder level for others. Since the measurements are not comparable, the results become erroneous and the person truly tallest of the group may not be measured as one, in case his height was measured from the shoulder level. In the worst case scenario, the peak has shifted so much that there is plain baseline where the apex is expected to be, which could be interpreted as a missing peak in the sample, even though the peak is present in the sample. It is also good to realize that intensity increases and decreases at faster rate for sharp and narrow peaks than it does for broad peaks. If the intensity value of the 10th data point after a peak apex is
considered, its value has decreased from the apex value more for a sharp peak than it has for a broad peak, implying that proper alignment may be more critical for sharp peaks than for broad peaks.

There is a great number of peak alignment algorithms reported. Just a few examples are Correlation Optimized Warping (COW), peak alignment using beam search, genetic algorithm (PAGA), and reduced set mapping (PARS), multiscale peak alignment (MSPA), and recursive segment-wise peak alignment (RSPA). COW has been extensively used and found effective for chromatographic data, but is less suitable for NMR spectra, although it has been utilized nevertheless. All peak alignment algorithms require careful optimization of some parameters, such as “slack size” and “segment length” for COW. Helpful instructions on parameter optimization with the COW algorithm is given by Skov et al. Of different preprocessing steps, peak alignment is especially time-consuming and there are large differences in aligning speed of different algorithms. The icoshift algorithm (interval-correlation-shifting) is a promising and relatively new peak alignment algorithm shown to work well for both chromatographic and NMR spectroscopic data. It is a downloadable open source peak alignment algorithm for Matlab and is reported to be several orders of magnitude faster than COW and having better aligning performance, which makes it an attractive choice for peak alignment algorithm. For more information on different alignment algorithms and alignment in general, papers by Gong et al., Savorani et al., Zhang et al., and Peters et al. provide some descriptions and references.

### 2.3.4. Binning

The goal of binning, also called bucketing, is to reduce the number of variables in the dataset to improve the statistical analysis or make it easier. In standard binning the spectra or chromatograms are divided to equal sized fragments (bins). If bin width of ten data points is selected, the intensity values of ten consecutive data points are summed, averaged or their largest value is extracted, and this value is set as the intensity of that particular bin. Selecting a bin width of ten data points will result in a spectrum or chromatogram that has one tenth the amount of data points the original had, and therefore some resolution is lost. A typical bin width for NMR spectra is 0.04 ppm, which reduces a 65k NMR spectrum...
to 300 bins. In other words, 65 500 variables are reduced to only 300. Some more advanced binning algorithms exist, such as adaptive-intelligent binning and Gaussian binning,\textsuperscript{54} but are not covered here.

### 2.3.5. Normalization

Normalization is a row operation that is executed to make all samples directly comparable with each other.\textsuperscript{44} The reason for the samples to be incomparable could result, for example, from seeds chosen for analysis having differing masses. The samples made from bigger seeds could be more concentrated than those made from smaller ones. Other reason could be a dilution error occurred in the sample preparation or varying injection volume, for example.

Normalization can be done by dividing each row (sample) by a sample specific normalization factor, for example the mass of the seed that was used to make the sample. This would be a simple matrix operation, easily done with software such as Matlab. Normalization can also be done using mathematical methods, such as integral normalization (also called area normalization or normalization to constant sum\textsuperscript{7}), vector normalization or maximum normalization. Of these, integral normalization is most used and has become more or less a standard in the field of metabolomics, at least with NMR spectroscopy.\textsuperscript{7} Integral normalization assumes that the concentration of a sample is proportional to the total intensity of its spectrum, total intensity being the sum of the intensities of all variables. The differences in the metabolomes are expected to be seen in the variations of intensities of some selected few variables rather than in the variation of the total intensity.\textsuperscript{7} In integral normalization, the values of all variables in a sample are summed together and every variable is divided by the sum. In other words, the new value for each variable is its fraction of the total intensity. Since more concentrated samples are expected to have larger summed intensities, they are divided by larger values and the larger concentrations are thereby compensated. It is important to know that integral normalization fails when there are dramatic differences in the intensities of only a few peaks between samples.\textsuperscript{44}
In vector normalization all variables are divided by the length of the sample vector, and in maximum normalization by the variable with the largest value, giving the largest peak a value of 100 %. In a comparative study, vector normalization has been found to perform worse than integral normalization and being highly sensitive to changes of a single peak. In addition to these three simple normalization methods, there are a large number of advanced normalization methods, of which many have been found to outperform the classic integral normalization. One of these is Probabilistic Quotient Normalization (PQN), that has been tested along with integral and vector normalization methods on real datasets from $^1$H NMR analyses as well as on simulated datasets. The simulated datasets simulated a gradual increase in the overall concentration or a gradual increase in some variables or both. The quotient normalization showed superior capability to normalize the spectra and was concluded to be by far more robust and accurate than the widely used integral normalization, and vector normalization. The two classic normalization methods only performed equally well in normalization of a dataset where only unspecific variation (the relative intensities of all variables remained the same) was present, simulating a situation where the samples are different dilutions of a same stock solution. In contrast to the classic normalization methods, PQN requires a reference spectrum that is similar to the spectra to be normalized. For more information on different advanced normalization methods, a paper by Kohl et al. is recommended.

2.3.6. Centering

Mean-centering is a column operation. Average value (mean) is calculated for each column from all the values in that column and the mean is subtracted from each value in the column. The result is that those values that used to be smaller than the average value now become negative figures and those that were larger than the mean retain their positive sign. The new mean of the values is now zero and exceptionally large values show as large deviation from zero and are more readily observed now. Centering effectively moves the focus to the fluctuating part of the data. Mean-centering prior to PCA is very important and should always be used, as the PCs cannot be fitted effectively otherwise. This is illustrated in Figure 6.
2.3.7. Scaling

As is the case with many other analysis techniques, in chromatography and NMR spectroscopy the response signal is related to the concentration of a compound in question. Compounds present in high concentrations have high intensities while those in low concentrations may be barely visible in the chromatogram or spectrum. However, from a biological point of view, the compounds present in high concentrations are not necessarily the most relevant ones in the metabolome.\textsuperscript{34,39} For example, the signal from a relevant metabolite may be very small compared to irrelevant biological compounds, such as ATP (adenosine triphosphate), present in high amounts. The differences in concentrations of the compounds in metabolomic data can be 5000-fold.\textsuperscript{39} Because PCA is looking for the highest variation in the data and large peaks are more likely to show high variance,\textsuperscript{34} the largest peaks get most of its attention and minor peaks have small role in the model. That is, peaks with large intensities get the highest weights in the model and the model focuses in describing them. While in some applications it may be best to look for the biggest differences in the variables, in the field of metabolomics the metabolites of interest are present in minor amounts and the weight of the model should often be transferred to those.

Scaling is a routine data pretreatment method used to give all the variables in the data similar weight.\textsuperscript{13} It is a column operation, which means that every variable (column) is divided (or multiplied) by a scaling factor, which is different for each variable. Many different kind of scaling methods exist, such as autoscaling, pareto scaling, range scaling, vast scaling and level scaling.\textsuperscript{39} By far the most used one is autoscaling, also known as unit
variance scaling. Autoscaling includes mean-centering, followed by scaling by standard deviation of the variable (dividing by standard deviation results in the variables having a variance of 1, hence the name “unit variance scaling”). The problem with autoscaling is that because small variation is given more weight, the measurement error is also magnified and noise is observed more pronounced. This can lead to interpretation difficulties, since the loadings plots often appear very noisy and little or no information is gained from the source of the variation that is responsible for the discrimination of the samples, seen as clustering. So in practice it may well be the case, that obvious clustering of samples is observed, but the reasons causing the clustering cannot be pointed out. While the magnification of noise is an unwelcome side effect of autoscaling, it may not be a problem if the preceding data preprocessing removes the noise efficiently.

If autoscaling results in uninterpretable loadings plot, pareto scaling can be a good alternative. Pareto scaling is very similar to autoscaling, but the scaling factor is the square root of the standard deviation rather than the standard deviation itself. This way, variables with low intensities still getting more weight (although not as much as with autoscaling) and noise is also less pronounced. For more information on different scaling methods, an excellent open access research article by van den Berg et al. is recommended.

3. CHEMOMETRIC APPROACHES IN PLANT METABOLOMICS

A common application of chemometrics involves studying the differences between different kinds of samples and determining what variables are causing the differences. This will be referred to as discrimination, as has been done so far. Different kinds of samples can be considered to belong to different sample classes. Such research is done by utilizing unsupervised pattern recognition. The study can be taken one step further and a computer can be trained to recognize the different classes and to sort samples according to which class they belong to. This is referred to as classification, and is done utilizing supervised pattern recognition. A few studies employing pattern recognition are introduced in some detail to serve as examples of what kind of results can be obtained with these techniques, applied to different kinds of instrumental data. All studies focus on metabolic profiling or
metabolic fingerprinting of different plant species. A more general survey that aims to find information on typical data preprocessing and sample preprocessing procedures follows the introductory examples.

Kim and Park have studied differences between two cactus species with unsupervised pattern recognition (PCA). Their aim was to establish a LC–MS method for metabolic profiling and compare the flavonoids found in the two species. Chemical discrimination of the two species could be utilized in quality control of dietary supplement foods manufactured from cactus extracts. The dataset for metabolic profiling consisted of peak areas for 15 marker peaks (glycoside derivatives of different flavonoids) analyzed from 34 samples (17 samples of both species) and was analyzed by PCA using Xlstat 6.0 software. The results from PCA are shown in Figure 7. Metabolic profiling allowed the two species to be discriminated on PC1 (Figure 7a), which was found to be due to different isorhamnetin glycosides (peaks 6–8 and 15) and kaempherol glycosides (peaks 12 and 13) as seen in Figure 7b. The kaempferol compounds also caused a significant variation within the cactus species marked with round symbols in the scores plot. The experiment was done using Agilent HP110 liquid chromatograph combined with Bruker HCT 3000 mass spectrometer operated in negative mode with heated (365 °C) electrospray ionization. The scanned mass range was m/z 100–1000, the column was Xterra MSC-18 with dimensions of 3.0 x 150 mm and particle size of 3.5 µm. The eluents were 1% acetic acid in water (A) and 1% acetic acid in acetonitrile (B), and the gradient elution program lasted 50 min. The simple sample preparation consisted of homogenizing 2 g of stem material in methanol and extraction for 12 h at room temperature, and finally filtering the sample through 0.2 mm Millipore filter.
Figure 7. Discrimination of two cactus species in a scores plot obtained with PCA by Kim and Park from a table of peak areas.\textsuperscript{55} The figure (a) shows scores on PC1 plotted against scores on PC2. The two species marked by different symbols are well discriminated in the PC1 dimension. The loadings in the right plot (b) show that the kaempferol glycosides (peaks 12 and 13) contribute to negative scores on PC1 and consequently make the samples appear left in the scores plot, while isorhamnetin glycosides (peaks 6–8 and 15) do the opposite and contribute to positive scores on PC1, moving the samples right in the scores plot.

Xiang et al.\textsuperscript{56} have studied the discrimination of three *Curcuma* species based on common compounds found in their essential oils by GC–MS analysis followed by unsupervised (PCA) and supervised (PLS–DA) pattern recognition. Medicinal products can be manufactured from the cactus species and therefore, as was the case with Kim and Park,\textsuperscript{55} the ultimate goal was to establish a method for herbal product quality control. There were five different kinds of samples in the study: firstly, for each species some herbs had been cultivated in their original cultivating areas in China (authentic herbs). Secondly, five samples had been cultivated elsewhere and, thirdly, 14 samples were made of commercial herbs. The dataset for metabolic profiling consisted of 75 manually integrated peak areas that were common to all samples, had a signal-to-noise ratio higher than 10, and that had been normalized to total area. Unsupervised pattern recognition (PCA) was used to find compounds that discriminate between the three different species. Supervised pattern recognition (PLS–DA) was then applied to train the computer to recognize the different species using the information gained from the unsupervised analysis (training set), the classification ability was then tested with known samples (test set) and ultimately, applied to classify the commercial samples. The PCA and PLS–DA were done with Matlab 6.5 software. Discrimination of all three studied *Curcuma* species proved possible. Scores and
loadings plots from PCA are shown in Figure 8. In the scores plot, three distinct clusters are seen, labeled as I–III, that each consists of only one authentic species (CURCW, CURCK or CURP). Interestingly, the five herbs not cultivated in their original cultivating areas are mostly close, but outside the cluster boundaries (95% confidence limit), showing that the cultivation area does affect the metabolome and can be observed. The species represented by the commercial samples can be deduced in the scores plots by looking at which cluster they reside in. The loadings plot shows that the reason for the Curcuma species III being so different from the other two species is different epicurzerenone levels that are strongly modeled by PC2. On PC1, the main factor for discrimination between species I and II is different germacrone levels, although curdione and curzerenone levels are significant also.

Figure 8. Discrimination of three Curcuma species (labeled as I, II and III) achieved by Xiang et al. from a table of common peak areas. The scores on PC1 vs. PC2 are shown in the left plot and loadings on PC1 vs. PC2 in the right plot. For the three species, samples cultivated in their original cultivation areas are labeled as CURCW, CURCK and CURCP, samples cultivated elsewhere as CURI and commercial samples as CURC.

Two thirds of the authentic samples were randomly selected and used for building the training set for supervised classification. The remaining one third, were used for testing the classification performance. Supervised PLS–DA successfully predicted the class of 12 out of 14 test samples (86%). For the samples that were used for train the model, 28 out of 29 authentic samples (97%) were correctly classified. The former result obtained with samples not used for building the model gives more realistic picture of the classification performance, while the latter way always gives over-optimistic results. Overall, 40 out of
43 samples (93 %) were correctly classified. Of the commercial herb samples, 12 out of 14 (86 %) was correctly classified. Of the 75 compounds included in the original dataset, four were found to serve as marker compounds that could be used to discriminate between species (see loadings in Figure 8) and repeating the supervised analysis with a dataset containing the peak areas of only these four compounds gave identical classification results. Leaving out two marker compounds caused the classification performance to drop from 97 % to 69 % when applied to the training set and from 86 % to 64 % when applied to the test set.

The analysis was done with Agilent 6890 gas chromatograph coupled with Agilent 5965B mass spectrometer using electron impact ionization. Helium was used as carrier gas with a flow rate of 1.0 ml/min and splitting ratio of 20:1. Analytical column was HP-5MS capillary column with dimensions of 30m x 0.25mm, coated with 0.25 µm film of 5% phenyl methyl siloxane. The temperature program lasted 55 min. The sample preparation started with boiling, drying (at 35 °C) to constant weight, cutting, mincing, and sieving of the rhizome samples. The essential oil was extracted by steam distillation, dried with NaSO₄ and finally dissolved in ethyl acetate for GC analysis.

The Defence Science and Technology Organisation (DSTO) of Australia has been studying the metabolome of *Ricinus communis* with sophisticated techniques. Firstly, by utilizing liquid chromatography coupled to ion trap (LC–ITMS) and quadrupole–time-of-flight (Q–TOF) mass spectrometers, in addition to fourier transform mass spectrometry (MALDI–FTMS) and tandem time-of-flight mass spectrometry, both with matrix assisted laser desorption ionization (MALDI–TOF/TOF), they have identified and sequenced three peptides, namely *Ricinus communis* biomarkers (RCB) 1–3 from the metabolomes. The peptides vary in concentration in the eight studied cultivars and could potentially be helpful to differentiate between different cultivars or provenance. Secondly, they showed that at least some wild Australian varieties of *Ricinus communis* can be differentiated by PCA. The varieties were analyzed by LC–UV, LC–ITMS and ¹H NMR spectroscopy. Some of the results from PCA of LC–UV data are shown in Figure 9.
The figure illustrates the discrimination of wild Australian *Ricinus communis* varieties, collected from different regions of Australia. Some samples (e.g. from Warrnambool, Westgate, Coopers Plains) are different from the rest, but the within-region variance is significant, possibly because of cross-pollinating in the wild.

Most recently, DSTO used $^1$H NMR spectroscopy and Orthogonal Partial Least Squares Discriminant Analysis (OPLS–DA) to study the metabolic fingerprints of eight imported cultivars.$^{14}$ They found that all cultivars as well as the different provenances could be completely discriminated (Figure 10) on the first three latent variables (corresponding to principal components in PCA). The discrimination was found to be due to sucrose, ricinine, both $N$-demethyl and $O$-demethyl ricinine, and phenylalanine (see loadings in Figure 11). The data was acquired with Bruker Avance-500 NMR spectrometer (500 MHz) using *noesypresat* solvent suppression pulse sequence. Acquisition time was 3.17 s and relaxation time was set to 5 s. Sample preparation included crushing of the seed, oil-removal with acetone wash, extraction of the seed mash with 2 % acetic acid, filtering with a 30 kDa molecular cut-off (MWCO) filter, freeze-drying and resuspension in D$_2$O. The data preprocessing included the following steps: the spectra were manually phased and baseline corrected, and binned into bins of δ 0.005 in width over the chemical shift range δ 0.2–10.00 from which the signals of HDO (δ 4.68–5.00) and acetic acid (δ 1.82–2.28) had been removed. Each spectrum was normalized to the peak area of the reference chemical (3-(trimethylsilyl)-2,2,3,3,-$d_4$-propionic acid, TSP-$d_4$) and logarithmically transformed. The 56 x 1649 data matrix was then normalized using probabilistic quotient normalization and,
finally, pareto scaled. Multivariate statistical analysis by OPLS–DA was done with both PLS Toolbox (for Matlab) and SIMCA P12+ software.

![Diagram](image)

**Figure 10.** The scores on LV1 vs. LV2 (a) and LV1 vs. LV3 (b) from the metabolic fingerprint acquired with $^1$H NMR and analyzed with OPLS–DA by Pigott et al. The latent variables are marked by “t” in figures. The first three latent variables allow complete discrimination of cultivars and provenance.

OPLS is a modification of PLS that was created to improve model interpretation. OPLS–DA is an improved version of PLS–DA, which is a pattern recognition method similar to PCA, but uses class information to maximize the class separation. Similar to PCA, PLS produces scores plots and loadings plots, but principal components are replaced by latent variables (LV) that are corresponding axes in PLS. Since PCA models the largest variation in the data, but the variance that causes the class separation does not necessarily lie in the same dimensions with the largest variation, PCA may fail to show sufficient discrimination. In such case, supervised methods like PLS–DA or OPLS–DA, which use class information to find the variance that best separates the classes, can be used.
Figure 11. Loadings on LV1 (a) and LV3 (b), marked by pq[1] and pq[3], that gave the best discrimination in the study by Pigott et al. Sucrose has a great impact on LV1 and LV3 and ricinene on LV3.

3.1. Common sample preparation and data preprocessing procedures

In an ideal metabolomic sample preparation the metabolome remains intact from the beginning of the sampling to the end of the analysis, and in an ideal metabolomic analysis the whole metabolome is measured. In reality the metabolome is slowly changing and sample preparation and instrumental techniques favor some compounds over others. For example, still after the samples are collected enzymatic activity remains, modifying the metabolome. It is also known that wounding a plant causes the plant to respond to the stimulus, which leaves its mark to the metabolome and so sampling itself may contribute to the metabolome. Extracting samples with a polar solvent discriminates non-polar metabolites. Common means to suppress enzymatic activity are acidic treatment, immediate freezing in liquid nitrogen and freeze clamping. Acidic treatment is problematic due to decomposition of some compounds. Grinding under liquid nitrogen followed by freeze-drying suppresses enzymatic activity since enzymes need water to work. If the powdered sample is stored, it should be done in –80 °C or kept in a desiccator to prevent it
from absorbing moisture from air. It should be noted that the even the freezing approaches have their problems, as explained in the paper by Fiehn.1

The literature concerning plant metabolomics and chemometric methods, primarily PCA and HCA, was explored, covering 37 publications.2, 4, 9-12, 14, 21, 55, 56, 59-85 The aim was to summarize common trends and procedures in sample and data preprocessing. The covered plant species include olive, cactus, curcuma, cannabis, pegaga, Catharanthus roseus, birch, Arabopsis thaliana, Castor bean plant, tomato, holly, dandelion, catuaba, hop, melon, cherry, green tea, Magnoliae Flos, Mung bean, ephedra, Quillaja saponaria, grape, potato, duckweed, tobacco, and balloon pea.

Every third study reported grinding the samples after being frozen by liquid nitrogen and in almost all of those cases the samples were then freeze-dried prior to extraction. Ground samples were often extracted with two immiscible solvents which were vortexed and both phases were then analyzed separately to get more comprehensive and complementary data. One third of all studies controlled the pH of the dissolved sample with a buffer and often further adjusted the pH. Samples for GC–MS required two derivatization steps, namely silylation and methoxylation. While the sample preparation for GC–MS analyses was most laborious, it was simplest for analysis by Direct Analysis in Real Time Mass Spectrometry (DART–MS) where the tissues were directly analyzed with no sample processing. Analysis of olive oils by LC–MS with atmospheric pressure chemical ionization (APCI) was also very straight-forward, as the oil was simply diluted 3000-fold with methanol.

As is commonly stated in the literature, the most popular instrumental technique seems to be NMR spectrometry, involved in half of the covered studies.4, 10, 12, 21, 61-64, 71-76, 79-84 Utilization of LC–MS and GC–MS seemed quite equally common as both techniques were reported in 21 %11, 55, 60, 67, 68, 75, 76, 85 and 16 %9, 65, 73, 76-78 of the covered studies, respectively. Application of mass spectrometry as a stand-alone technique was also relatively common, applied in ca. 10 % of the studies66, 69, 70, 76 Other techniques less applied in metabolomics, such as Fourier Transmission infrared spectroscopy (FTIR),59 were also noticed. Of all chemometric techniques, PCA was applied in 86 % of the covered publications, while 30 % of them also utilized HCA. Many times, HCA was applied on data that had been first reduced by PCA.63, 71, 74, 83
Surprisingly, the widely available supervised methods kNN and SIMCA,\(^{13}\) used for classification, turned out to be less popular than initially expected. Only two studies\(^ {2, 72}\) used SIMCA and only one\(^ {2}\) applied kNN. Of the supervised methods, Partial Least Squares Discriminant Analysis (PLS–DA) was by far the method of choice, applied in 32% of all covered studies.\(^ {10, 12, 59, 66, 67, 71, 73, 76, 80-83}\) Additionally, two studies (5%) utilized orthogonal PLS–DA (OPLS–DA),\(^ {14, 85}\) which has improved model interpretation over the original.\(^ {57}\) However, no confident conclusions on the relative popularity of the chemometric methods can be made, since the sample consisting of the 37 covered publications are possibly biased by the key words used for searching the literature.

Most of the NMR-based studies (90%) explored the data by PCA.\(^ {4, 10, 12, 21, 61-64, 71, 72, 74-76, 79, 81-84}\) 15% of the studies applied HCA,\(^ {62, 63, 71, 74, 83, 84}\) but always employed PCA too. Half of those using HCA applied it on data that had been first reduced by PCA.\(^ {63, 71, 74}\) Where mentioned, the linkage algorithm applied in HCA was Unweighted Pair Group Method with Arithmetic Mean (UPGMA),\(^ {63}\) Ward’s method\(^ {62}\) or incremental linkage,\(^ {74}\) and the distance measure was Euclidean distance.\(^ {62, 63}\) Supervised data analysis was primarily done by PLS–DA,\(^ {10, 12, 71, 73, 76, 80-83}\) whereas other supervised techniques, OPLS–DA and SIMCA, were each reported in only one study.\(^ {14, 72}\) The most popular software for data analysis was by far SIMCA P, used by 65% of the NMR-based studies.\(^ {10, 12, 14, 61, 62, 64, 71-73, 76, 79, 82, 83}\) Other software used were Matlab (20%),\(^ {4, 12, 63, 84}\) PLS Toolbox (15%),\(^ {14, 21, 80}\) Pirouette\(^ {74}\), XCMS\(^ {75}\), Win-Das\(^ {81}\) and SPSS\(^ {83}\) (5% each).

The applied data preprocessing is often left unmentioned, despite the fact that it may have a large impact on the result. The reason for this could be considering it redundant or not realizing its significance. None of the papers mentioned smoothing the data and only a few reported a peak alignment operation. Two of these only reported the peak alignment software (SpecAlign),\(^ {63, 84}\) while the other two named the actual alignment algorithm (COW,\(^ {73}\) RSPA\(^ {12}\)). The NMR spectra are routinely baseline corrected, primarily by the software provided with the instrument.\(^ {4, 10, 12, 14, 21, 62, 71, 73, 81-84}\) In a couple of cases this was also done by other software (ADVASP Lite\(^ {63}\), PLS Toolbox\(^ {80}\)). Binning was found to be commonly used to reduce the data, used in 75% of the NMR-based studies.\(^ {10, 14, 21, 61, 62, 64, 71-73, 75, 79-82, 84}\) A bin-width of 0.04 ppm was found to be most common (used by every third
which is also stated in the literature. Other applied bin widths were 0.002, 0.005, 0.01 and 0.02 ppm. The most common data normalization approach was normalizing to the peak area (or -height) of a reference standard (55% of NMR-based studies). Normalization to the sum of signals (integral normalization) was another common method (25%). In two cases an advanced normalization method was used, namely Probabilistic Quotient Normalization. The most common way to scale the data was autoscaling (45%), pareto scaling being second (25%). Only one paper mentioned doing logarithmic transformations.

In contrast to the popular NMR spectrometry, only two metabolomic studies utilized LC–UV instrumentation. One had the chromatographic data binned to bins of 20 s, whereas the other exploited the whole chromatographic fingerprint. Both studies had the data normalized (method unspecified) and autoscaled. No further information on data preprocessing was given.

LC–MS analyses were predominantly done using electrospray ionization (ESI) as only one out of nine studies employed atmospheric pressure chemical ionization (APCI). Based on the covered literature, metabolomic studies employing LC–MS seem to prefer the metabolic profiling approach over the fingerprinting approach, since in 7 out 8 studies the datasets were tables of peak areas and only one study made use of the entire chromatographic pattern. Again, PCA was the unsupervised method of choice and HCA was merely applied to data that had been first reduced by PCA. The linkage algorithms used in HCA were Complete linkage and Ward’s method with Euclidean distance as the distance measure. Integral normalization was commonly used to normalize the data (half of all cases) and autoscaling (50%) was slightly preferred over pareto scaling (38%). although some papers did not mention the scaling method or the normalizing method (50%). The only study utilizing the chromatographic fingerprint ended up using pareto scaling to scale the signals. This was also the only paper that reported smoothing the data. Use of peak alignment was only reported in one study that referred to MetAlign software.
As was the case with LC–MS-based metabolomic studies, the datasets for chemometric analysis of GC–MS data are usually tables of integrated peak areas or peak heights (in all cases) and primarily studied with PCA (83 %). Two studies employed HCA and the one mentioning the methods reported using Ward’s method with Euclidean distances. In contrast to studies using other instrumental techniques, those using GC–MS did not report integral normalization, but normalizing to the peak area of a reference standard (half of the GC-based studies) and/or to the weighted sample mass instead (33 %). Peak alignment procedures were mentioned in only one study. Use of both autoscaling and pareto scaling were reported, although in most cases (66 %) scaling was not discussed.
4. STUDY OF DIFFERENCES BETWEEN CULTIVARS OF RICINUS COMMUNIS USING CHEMOMETRIC METHODS

4.1. Abstract

The aim of this study was to find out if different cultivars of *Ricinus communis* could be discriminated by their metabolome, analyzed by hierarchical cluster analysis (HCA) and principal component analysis (PCA). The metabolomic data was acquired with liquid chromatography with ultraviolet light detection (LC–UV), liquid chromatography–mass spectrometry (LC–MS) and proton nuclear magnetic resonance (\(^1\)H NMR) spectroscopy. Discrimination between cultivars proved to be possible, at least with the chromatographic techniques, although some uncertainty remained whether the discrimination was due to biological variation or perhaps partially to misaligned peaks in the data. In addition to chemometric results, other experimental findings (such as the presence of potential biomarkers) are reported as well, for being potentially interesting to VERIFIN. The presence of *Ricinus communis* peptide biomarkers in different cultivars were found to be inconsistent with what has been reported previously.\(^{22}\)

4.2. Sample preparation

Six different *Ricinus communis* cultivars were selected to be studied and six seeds of similar size, mass and appearance were selected from each cultivar for analysis and the seed weights were recorded. In total, 36 seeds were prepared to make 36 samples. Two fractions were taken from each sample, one for analysis by liquid chromatographic techniques and the other was further prepared for \(^1\)H NMR analysis.

The sample preparation consisted of peeling and grinding of the seed, removal of disturbing ricin oil by acetone extraction, drying the residue under air, dissolving the residue into 2 % deuterated acetic acid and filtration of the acetic acid solution through a molecular weight cut-off (MWCO) filter in a centrifuge. The sample preparation was done by the experienced M.Sc. Marja-Leena Rapinoja in a containment hood within the facilities of VERIFIN, due
to presence of the highly toxic ricin. The filtrate, free of all compounds with molar mass more than 30 kDa, was collected as the sample to be analyzed.

Prior to grinding the seed was peeled and the peel discarded. The ground seed mass was transferred to a Falcon tube and 20 ml of cold (+4 °C) acetone was added to extract the ricin oil. The mixture was then stirred overnight at +4 °C. Next, the mixture was centrifuged at 3750 rpm for 10 min and the acetone solution was decanted off. The residue was washed with 10 ml of acetone and the mixture was again centrifuged and the solvent decanted off. After acetone treatment, the mass had turned from stiff to light and powder-like. Now, acetone was once again added, but this time only 1 ml and only to help distribute the mass evenly to the walls of the tube to enable faster and more complete drying, in contrast to all the mass sitting on the bottom of the tube in a pile. The tubes containing the fine residues were left uncapped to dry overnight in the containment hood with the ventilation turned off. Next, the dry, fine and white powder was extracted with 6 ml of cold 2 % acetic acid-\textsuperscript{d}4 and vortexed for 1 minute. The tubes were left to stand in an ice bath for 1 hour, after which they were centrifuged (again 3750 rpm for 10 min) and the supernatant was decanted into smaller, 15-ml Falcon tubes. The volume was adjusted to the 6 ml mark of the Falcon tube with 2 % acetic acid-\textsuperscript{d}4 solution and the capped tube was turned upside down and back a few times to homogenize the solution. Because the capacity of the MWCO filter was only 0.5 ml, several filtrations were necessary to obtain enough filtrate. To obtain ca. 2 ml of ready sample, four MWCO cartridges were injected with the acidic solution, 500 µl in each cartridge. The cartridges were put into 2-ml Eppendorf tubes, and centrifuged at 8000 rpm for 2 min. Roughly 150 µl remained on the filter after centrifugation, meaning that approximately 350 µl of the sample was obtained in the Eppendorf tube for every filtration. Some variation was observed in the amount of sample remaining on the filter.

The resulting filtrate from two of the four Eppendorf tubes was combined and transferred to a 2-ml vial and analyzed directly by both chromatographic techniques, but in making samples for NMR analysis, further sample preparation was needed, including the addition of the trimethylsilyl propionic-2,2,3,3-\textsuperscript{d}4 acid (TSP-\textsuperscript{d}4) as a chemical shift reference and
deuterated water to be used for locking to the deuterium signal, and is described in the following.

Filtrates from the remaining two Eppendorf tubes were combined to one, making a combined volume on the order of 0.7 ml and leaving the other Eppendorf tube empty. From the combined solution, 500 µl was pipetted into the empty Eppendorf tube, using a Finnpipette. Next, 100 µl of TSP-$d_4$ in D$_2$O was added to the sample, making the final sample volume 600 µl. The solution was stirred with the tip of the Finnpipette and some solution from the bottom of the tube was transferred multiple times to the surface of the solution to mix and homogenize the solution. The finished solution was carefully transferred into an NMR tube (standard quality). Bubbling or foaming (possible unwanted behavior of proteins and peptides) of the sample was avoided all the time by careful movements of the pipette.

Fractions for chromatographic analysis are indicated by “S02” and fractions for NMR analysis by “S04” in the sample code.

**4.3. Methods**

The preprocessing of the datasets constructed of data from different instrumental analyses was minimal. Many preprocessing operations require in-depth understanding as well as trial and error$^{13}$ to determine valid and/or optimum parameters and there was insufficient time to optimize such parameters for some preprocessing operations properly. Therefore no smoothing, baseline correction (except for NMR data) or advanced peak alignment operations were done. All datasets were normalized with masses of the seeds selected for analysis to avoid the possibility of the samples to simply cluster according to the weight of the seeds. For example, all seeds weighed close to 0.5 g, except ‘Carmencita pink’ seeds weighed only ca. 0.3 g, and it was thought that this significant mass difference could have led to ‘Carmencita pink’ appearing exceptional. Additionally, all datasets were integral normalized to compensate possible errors in concentration, such as dilution errors occurred in sample preparation or variation in injection volume *etc.* Finally, the datasets were either autoscaled, pareto scaled or merely mean-centered, depending on the dataset. All data
preprocessing and the chemometric study was conducted with a multivariate analysis package *PLS Toolbox* (Eigenvector Research Inc.), freely downloadable add-on for *Matlab* software (MathWorks Inc.). The datasets are described in more detail in the following.

4.3.1. High performance liquid chromatography

The HPLC instrument was an Agilent 1200 HPLC consisting of a solvent degasser, a binary pump, an autosampler, a fraction collecting unit (Bruker Biospin BPSU-36), a column oven and a diode array detector (DAD). One goal was to learn about the performance of C18 columns in this application, even though the retention for polar compounds was expected to be quite poor (C18 columns are known to be reliable and the application requires robustness from the analysis). Therefore, a Waters XBridge C18 column with the dimensions of 150 x 4.6 mm and particle size of 5 µm was initially tried, but was later replaced by a Waters Atlantis T3 C18 column with the same dimensions to improve the retention of polar compounds. Since the extraction solvent was aqueous acetic acid (2 %, deuterated), lots of polar compounds were expected to be present.

0.1 % aqueous HCOOH was used as eluent A and 0.1 % HCOOH in methanol as eluent B. Methanol was HPLC grade and purchased from Sigma-Aldrich, while the water used was UHQ water obtained from an Elgastat UHQ water purification system. A flow rate of 1 ml/min and injection volume of 10 µl were used. The diode array detector was recording a wavelength range of 190–950 nm, but the wavelength of 254 nm was selected to be used in the data analysis. This wavelength was adopted from Ovenden *et al.*, who have studied similarly pretreated castor bean extracts in the past using a C18 column, and who were the inspiration for this study.

Gradient elution was used to separate the compounds of the castor bean extract and was optimized to separate as much compounds as possible. The gradient started with an isocratic period with 5 % B lasting 5 min, after which the ratio of the organic eluent was raised from 5 to 100 % in 14 minutes (6.8 %/min). Next, the column was cleaned for 5 minutes with 100 % B, after which the original conditions were restored in 1 min and equilibrated for 5 min. The purpose of the initial isocratic part was to give early eluting
polar compounds more time to separate and proved to be beneficial, but was perhaps unnecessarily long for that purpose.

A quality control sample, being a mixture of phenols (NIP-MIX-4, Bruker BioSpin, Germany), was run prior to each sample batch to discover possible unusual behavior of the instrument by monitoring the peak intensities and peak widths.

4.3.2. Liquid chromatography–mass spectrometry

The LC–MS instrument was a commercial ensemble manufactured by Thermo Scientific, consisting of a solvent degasser, a Finnigan Surveyor LC Pump Plus quaternary pump, a Finnigan Surveyor Autosampler Plus autosampler and a Finnigan LXQ iontrap mass spectrometer with an electrospray ion source, operated in positive mode. The analytical column was the same used for LC–UV analyses, namely Waters Atlantis T3 C18 with dimensions of 150 x 4.6 mm and particle size of 5 µm. The gradient program was not optimized for LC–MS, but the conditions were kept as similar to the LC–UV conditions as possible, to make the chromatographic profiles and retention times similar. For the same reason a flow rate of 1 ml/min was used, but for ESI it was split to 1:20 to introduce a flow rate of 50 µl/min to the mass spectrometer, a value considered optimal for the instrument. Eluents were same that were used with the LC–UV system and the same injection volume of 10 ml was injected using “Partial Loop Injection” as the injection method.

All samples were analyzed using a mass range of $m/z$ 70–2000. The signal was optimized for a peptide mixture used also as a chromatographic and mass spectrometric quality control sample before each sample batch. The optimized ion source parameters are listed in Table 1. The peptide mixture consisted of four peptides, namely glycine-tyrosine (Gly-Tyr, $m/z$ 239.1), methionine enkephalin ($m/z$ 574.2), leucine enkephalin acetate ($m/z$ 556.3) and angiotensin II ($m/z$ 523.8$^{2+}$). A peptide mixture was chosen for optimization and as the quality control, because peptide biomarkers for *Ricinus communis*, present in different amounts in different cultivars$^{22}$ and possibly useful for discrimination between them, were expected to be seen in the castor bean extracts.
Table 1. Ion source parameters used for ionization of quality control and castor bean samples.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sheath Gas Flow Rate (arb)</td>
<td>14</td>
</tr>
<tr>
<td>Auxiliary Gas Flow Rate (arb)</td>
<td>15</td>
</tr>
<tr>
<td>Sweep Gas Flow Rate (arb)</td>
<td>0</td>
</tr>
<tr>
<td>Spray Voltage (kV)</td>
<td>5.00</td>
</tr>
<tr>
<td>Capillary Temperature (°C)</td>
<td>275.00</td>
</tr>
<tr>
<td>Capillary Voltage (V)</td>
<td>35.00</td>
</tr>
<tr>
<td>Tube Lens (V)</td>
<td>85.00</td>
</tr>
</tbody>
</table>

Two datasets were made from the LC–MS data. First, the total ion current (TIC) chromatograms were saved as text files using the export function of the Xcalibur software. Additionally, five peaks were extracted, smoothed (“Gaussian” fit with a window of 15 points) and integrated and the peak areas were saved as another dataset. The goal with the extracted ion dataset was to see if the peak areas or their relative ratios could be used to discriminate between cultivars, as is proposed in the literature\(^1\). The five picked peaks were ricine (m/z 165), three known peptide biomarkers, called *Ricinus communis* biomarkers 1–3 (RCB-1–3) by Ovenden *et al*\(^2\), and an unknown compound suspected to be demethylated analog of ricinine (see the different ricinine structures in Figure 12). This assumption was based on its mass (m/z 151), elution before ricinine (as the substitution of methyl with a proton at either the oxygen or nitrogen in the ricinine molecule would lead to formation of a more polar group that would elute earlier in RPLC), and appearing only in castor bean samples and being common in them. Additionally, the unknown compound showed similar fragmentation to ricinine. Although the product ion mass spectra of ricinine and the suspected demethyl ricinine were very uninformative for structure determination, they both seem to exclusively lose a fragment of m/z 27, possibly being hydrogen cyanide, with normalized collision energy of 23 % (see Appendix 1 for product ion spectra). For the peptide biomarkers, the masses used for peak picking – two for each peptide – were m/z 690.3\(^3\) and 1034.5\(^2\) for RCB-1, 661.0\(^3\) and 990.9\(^2\) for RCB-2, and 655.0\(^3\) and 982.0\(^2\) for RCB-3, as reported by Ovenden *et al*\(^2\). The more intensive ion, being the triply charged

\(^1\) The capillary temperature was a remnant from another application. While its value was not found to have any effect on the peptide signals during optimization, it was later found that it had a dramatic effect on ricinine signal and a collapse is observed when the temperature reaches 145 °C. This was observed when ricinine signal was optimized for a product ion scan experiment. The value has possibly been detrimental to the analysis.
ion for each RCB, was used for integration. An example of the peak picking is seen in Appendix 2.

Figure 12. The structure of ricinine (1) and two possible structures for demethylated ricinine (2 and 3).

Electrospray ionization (ESI) operated in positive mode was a natural choice for the ionization method, since the extract was aqueous and acidic, containing predominantly polar compounds and peptide biomarkers were expected to be present in the samples. An alternative ionization technique available was Atmospheric Pressure Chemical Ionization (APCI), which generally can be used for compounds having a mass-to-charge ratio up to ca. 1500 amu. However, because ionization in APCI occurs in gas phase and large polar molecules such as proteins and larger peptides are nonvolatile, they are not expected to be transferred to the gas phase and thus, ionize. In contrast to APCI, in ESI the ions are formed already in the liquid phase and the ions are transferred to the gas phase by a strong electric field.

4.3.3. Nuclear magnetic resonance spectroscopy

The NMR experiments were carried out measuring the proton signal with a Bruker Avance III 500 NMR spectrometer operating at a frequency of 500 MHz. Solvent signals were suppressed using “Purge NMR” method by A.J. Simpson et al. and trimethylsilyl propionic-2,2,3,3-$d_4$ acid (TSP-$d_4$) was used as the chemical shift reference (0 ppm). Protons were excited with a 90° pulse 8.3 µs in width and the proton signal was recorded for 2.7 s. The relaxation delay was 5 s. A spectrum was 6010 Hz in width and consisted of 65536 data points, acquired with 256 scans. The recorded spectra were phased and baseline corrected and saved in JCAMP-DX file format using Bruker TopSpin 3.0 software.

Finally, the intensity data from the JCAMP-DX files was imported to Matlab, where specific spectral regions were extracted from the full spectra. These regions, namely
aromatic region (7.6–9.4 ppm), main saccharide signal region (4.3–5.5 ppm), anomeric signal region (another saccharide dataset, 6.2–6.7 ppm) and amino acid signal region (1.9–3.2 ppm). Additionally, one large dataset including all the spectral areas that were considered to be free of solvent peaks, other disturbing peaks or long segments of baseline, was created from spectral regions of 1.9–3.2, 3.6–5.9 and 6.2–9.9 ppm. These were saved as five separate datasets that were subjected to chemometric analyses.

4.3.4. Datasets and preprocessing
The LC–UV dataset analyzed by different chemometric methods is visualized in Figure 13. The only operations done to the raw data shown in the figure is normalization by seed mass and manual alignment of the chromatograms. To be exact, the alignment is truly an alignment of the ricinine peaks and it was assumed that if the ricinine peak in one chromatogram had a retention time 0.1 min shorter than in another chromatogram, the rest of the peaks eluted that much earlier also. This optimistic assumption may not be correct, but the effect was assumed to be minor, as the retention times were quite reproducible, varying between 11.55 and 11.64 min (average and median both 11.61 min) for ricinine. Following the alignment operation, unnecessary baseline at the beginning and the end of the chromatograms was cut off. All operations were performed by self-written Matlab scripts.

Different data preprocessing operations were tried on the data, and best the combination for the LC–UV data was found to be integral normalization followed by autoscaling. This combination not only worked best for PCA, but for HCA, as well. Not surprisingly, the same data preprocessing worked well for LC–MS data too, since the data produced by both chromatographic techniques, being chromatograms with relatively few (resolved) peaks, is very similar. However, replacing autoscaling with pareto scaling (very similar to autoscaling), including mean-centering, was found to give even better clustering in PCA. Exclusion of the highly intensive segment (data point region 0–500) of unresolved peaks in the beginning of the TIC was also found to significantly improve clustering. After the exclusion the remaining signals are quite similar in intensity, which probably explains the lesser need of scaling. The complete LC–MS TIC dataset is shown in Figure 14 before (a) and after (b) the exclusion of the segment with the unresolved peaks. Prior to any other
preprocessing, both LC–MS datasets had been normalized by seed mass and the chromatograms of the TIC dataset had been manually aligned using the ricinine peak, as was the case with the LC–UV dataset. The extracted ion dataset was integral normalized and mean-centered for PCA and autoscaled for HCA, and is shown in Figure 15 prior to this preprocessing.

![Figure 13](image)

**Figure 13.** Visualization of the dataset consisting of LC–UV chromatograms of 36 samples. The only data preprocessing steps done to the visualized data are normalization by seed mass and a simple alignment of the 36 chromatograms.

![Figure 14](image)

**Figure 14.** Complete LC–MS TIC dataset before (a) and after (b) exclusion of the first 500 data points. The only data preprocessing steps done to the visualized data are normalization by seed mass and a simple alignment of the 36 chromatograms.
The extracted ion dataset looks very different from all the other datasets, because it is neither a table of chromatograms nor a table of spectra, but a table of peak areas with only five variables in contrast to thousands of variables (intensity values) in spectra and chromatograms.

The different $^1$H NMR datasets analyzed are shown in Figure 16. While there was no significant difference in clustering between autoscaled and merely mean-centered datasets for the individual regions, autoscaling did seem to benefit the dataset that contained all the individual regions together. The reason for this was probably the individual regions being many orders of magnitude different in intensity. Figure 16e shows that if all regions are contained in one dataset, the saccharide signals (Figure 16d) dominate the spectrum. Such difference in orders of magnitude is removed by autoscaling. For the individual regions, integral normalization followed by mean-centering was selected as data preprocessing because of better interpretability of the loadings plots for mean-centered data. In contrast, for the combined regions dataset, mean-centering was replaced with autoscaling and same preprocessing was found to slightly improve clustering in HCA.
Contrary to the initial plan, the supervised pattern recognition techniques, namely kNN and SIMCA, were not applied because of insufficient amount of samples per cultivar. Both of
these techniques require that the computer is trained with enough samples to be able classify any samples. If all samples were used to train the computer, no samples would be left to be classified. If the same samples that were used to train the computer are given to be classified, they fit too well to the training data and the classification is unrealistically successful and gives optimistic picture of the classification model. More samples, which shall not be used to train the computer, would be needed to put the classification model to test. If some samples were simply taken away from the training set – already consisting of only six samples per cultivar – to be used as a test set, the training would become even more insufficient and it is not useful to create an unreliable model.
4.4. Results and discussion

4.4.1. General observations during experimental work

In the initial tests it was soon found out that the MWCO cartridge used for filtering the castor bean extracts contaminated the samples with glycerol. The glycerol peaks seen in the NMR spectrum were intensive and likely to be detrimental to chemometric analysis, overlapping other peaks in the sample. Washing the MWCO cartridge three times with purified water greatly reduced the glycerol signals to an acceptable level, but the signals were visible even after six consecutive washes. Washing with 2 % acetic acid or 0.1 M NaOH did not provide a better washing result and therefore washing with pure water is recommended. Fortunately, the glycerol was not a problem with the chromatographic methods, since it does not have a chromophore to show up in a UV chromatogram, nor is it basic enough to ionize in the conditions used with ESI.

Different injection volumes (10, 15, 20 and 25 µl) were also tested in the initial tests, to see how they affect the peak intensities and shape. 20 and 25 µl turned out to be too much, observed as fronting peaks in the LC–MS chromatogram. What comes to the mass range that was intentionally set as wide as \( m/z \) 70–2000 to not leave any peptides outside the mass window, it could narrowed to ca. \( m/z \) 1300, since there was a little data beyond this value and starting from \( m/z \) 1500 there was increased background noise.

The reproducibility for the LC–UV instrument was very good and the peak intensities for the four QC compounds had relative standard deviations less than 1 % for 21 consecutive quality control runs. The (relative) standard deviation for QC samples run prior to each castor bean sample batch was 1–2% for peak intensities and 0.02–0.06 min for retention times. The good reproducibility is largely explained by the QC method, which is run in isocratic conditions (80 % B for 10 min). However, the reproducibility with the gradient used to run the castor bean samples was still good, since the standard deviation for the retention time of the risinine peak was 0.02 min within a batch. The quality control samples for LC–MS were run with the same gradient used to analyze castor bean samples. The reproducibility test with six consecutive runs gave (relative) standard deviations of 2–4 % for the peak intensities and 0.01–0.02 min for retention times. Between batches these
values were 0.02–0.03 min for retention times and 5–20 % for peak intensities. Therefore, there were some relatively high deviations in peak intensities for LC–MS. The QC data is found in Appendix 3.

The chromatograms or spectra of all castor bean samples within each instrumental method looked very similar. The amount of peaks was generally more or less similar as well as the intensity of the peaks, although there were some peaks whose presence and intensity varied greatly even within the same cultivar. The most interesting example is the *Ricinus communis* peptide biomarker RCB-3 observed with LC–MS. Ovenden *et al.* have previously studied the peptide biomarkers of ricin for several different cultivars, four of which were also involved in this study, namely ‘Gibsonii’, ‘Zanzibariensis’, ‘Impala’ and ‘Carmencita’. They found that RCB-3 was only observed in ‘Carmencita’ extracts, in contrast to a result of this study that the biomarker was found in three of these cultivars, ‘Gibsonii’ being the only one not containing it. It was also found in the two other cultivars, ‘Carmencita pink’ and ‘Carmensita bright red’ involved in this study. The three peptide biomarkers involved in this study were RCB-1, RCB-2 and RCB-3 and the idea was to pick the peaks characteristic to these peptides from the chromatograms and analyze a table of peak areas with chemometric methods to see if the peak areas can be used to discriminate between cultivars. The presence of peptide biomarkers and their peak intensities were more or less consistent for all cultivars, ‘Zanzibariensis’ making a distinct exception. In fact, both the presence of peaks and their intensities were very inconsistent for this cultivar. Two samples (Y023bS02 and Y023dS02) were found to contain biomarkers RCB-1 and RCB-2, but not RCB-3. For the next sample (Y023aS02) the situation was completely inverse, since it did not contain either RCB-1 or RCB-2, but did contain RCB-3. The rest of the samples (Y023cS02, Y023eS02 and Y023fS02) contained all three biomarkers and for Y023fS02 they were significantly more intensive.

4.4.2. Results of hierarchical cluster analysis

Seven different linkage methods (Ward’s method, nearest neighbor, furthest neighbor, average paired distance, centroid, median and K-means) were applied to get more accurate view of the data and avoid random error due to arbitrary choice of linkage method (the
different linkage algorithms may give slightly different clustering and determining which one is most “correct” is problematic).

4.4.2.1. Data from liquid chromatography with ultraviolet light detection

Of the seven obtained dendrograms the one resulting from Ward’s method is shown (Figure 17) for being easiest to read and showing the finest clustering. Four of the linkage methods: Ward’s method, furthest neighbor, average paired distance and K-means, all agree that there are six clusters that exclusively consist of members of a single cultivar. This result was expected to be seen as long as the sample preparation, data acquisition method or data preprocessing would not fail, since samples of six known cultivars of *Ricinus communis* were analyzed. Nearest neighbor algorithm found that one sample (Y023aS02) did not cluster with the rest of the ‘Zanzibariensis’ samples and same was observed for one ‘Carmencita bright red’ sample (Y021aS02), but otherwise the outcome was similar to the four other methods. Ward’s method and furthest neighbor method both found that the six clusters could be further joined to form two major clusters, one consisting of the ‘Carmencita’ (Y024), ‘Carmencita pink’ (Y020) and ‘Carmencita bright red’ (Y021) cultivars and the other consisting of the rest of the cultivars, similarly. The former cluster makes sense, since the cluster is made up of cultivars of the same family. However, the distances between clusters inside the two major clusters are similar in both and therefore claiming that the cultivars of the ‘Carmencita’ family are very similar would not be well supported by this data.
The dendrogram obtained with Ward’s method from the integral normalized and autoscaled LC–UV dataset. The cultivars are coded as follows: ‘Gibsonii’ (Y019), ‘Carmencita pink’ (Y020), ‘Carmencita bright red’ (Y021), ‘Impala’ (Y022), ‘Zanzibariensis’ (Y023) and ‘Carmencita’ (Y024).

The result of the centroid and median linkage algorithms differed markedly from the other methods (Figure 18). These two linkage methods came to the same conclusion that three of the cultivars (Y019, Y022 and Y024) form their own clusters pure of samples of other cultivars, but the rest form mixed clusters. Also, sample Y021aS02 was found to be a distinct outlier inside its cultivar. It is no surprise that centroid and median linkage methods gave similar results since they work very similarly, while differently from the other linkage methods. Since five of the seven linkage methods gave very similar results, while the two based on centroids of the clusters gave different and more arbitrary results it is concluded that the data truly is clustered so that each cultivar forms a separate cluster, but the samples within the clusters are not evenly distributed around the “center” of the cluster, i.e. the centroid of the cluster is not in the middle of the cloud of points (samples) in the row space. This is probably due to the small amount of samples per cultivar (only six) in the dataset and having a greater amount of samples would probably lead to the centroid better matching the center of the cloud of points and therefore allow more successful clustering.
with the centroid and median algorithms. The clustering results obtained with different linkage algorithms are summarized in Table 2.

![Dendrogram of Data with Preprocessing: Normalize (1-Norm, Area = 1) + Autoscale](image)

**Figure 18.** The dendrogram obtained with the centroid method from the integral normalized and autoscaled LC–UV dataset. The cultivars are coded as follows: ‘Gibsonii’ (Y019), ‘Carmencita pink’ (Y020), ‘Carmencita bright red’ (Y021), ‘Impala’ (Y022), ‘Zanzibariensis’ (Y023) and ‘Carmencita’ (Y024).

**Table 2.** Summary of the clustering results for integral normalized and autoscaled LC–UV data showing the number of clusters that only consist of samples of same cultivar. Additionally, the cultivars that form such pure clusters are listed. In the best scenario, six pure clusters are observed and the results of different linkage algorithms are consistent.

<table>
<thead>
<tr>
<th>Ward’s method</th>
<th>Nearest neighbour</th>
<th>Furthest neighbour</th>
<th>Average paired distance</th>
<th>K means</th>
<th>Centroid</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 All</td>
<td>Y019, Y020, Y023, Y024</td>
<td>6 All</td>
<td>6 All</td>
<td>6 All</td>
<td>3 Y019, Y022, Y024</td>
<td>2 Y019, Y024</td>
</tr>
</tbody>
</table>

59
4.4.2.2. Data from liquid chromatography–mass spectrometry

For the LC–MS extracted ion dataset all linkage algorithms found that ‘Gibsonii’ certainly forms a clear cluster free of samples from other cultivars. Additionally, they all found 4–5 samples of ‘Carmencita’ and three samples of ‘Zanzibariensis’ (Y023cS02, Y023eS02 and Y023fS02) to cluster. Otherwise no interesting clustering was observed. Dendrogram obtained with Ward’s method is shown in Figure 19. The clustering results obtained with different linkage algorithms are summarized in Table 3.

![Dendrogram obtained with Ward’s method](image)

**Figure 19.** The dendrogram obtained with Ward’s method from autoscaled LC–MS extracted ion dataset. The cultivars are coded as follows: ‘Gibsonii’ (Y019), ‘Carmencita pink’ (Y020), ‘Carmencita bright red’ (Y021), ‘Impala’ (Y022), ‘Zanzibariensis’ (Y023) and ‘Carmencita’ (Y024).

For the LC–MS TIC dataset clustering was found to be similar to the clustering of LC–UV data. Again, Ward’s method provided the finest dendrogram, shown in Figure 20. Ward’s method and average paired distance algorithm found that all cultivars form a cluster, pure from other cultivars. K-means came to otherwise same conclusion, but it found ‘Impala’ sample Y022bS02 to be an outlier and not cluster with the rest ‘Impala’ samples. Nearest neighbor otherwise agreed with K-means, but it also found ‘Carmencita pink’ sample Y022cS02 not to cluster with the rest of its kind (Figure 21). Furthest neighbor and median
algorithms found that only ‘Gibsonii’, ‘Zanzibariensis’ and ‘Carmencita’ cluster ideally. Finally, centroid found that only ‘Gibsonii’ and ‘Zanzibariensis’ form pure clusters.

Figure 20. The dendrogram obtained with Ward’s method from the integral normalized, pareto scaled and mean-centered LC–MS TIC dataset. The cultivars are coded as follows: ‘Gibsonii’ (Y019), ‘Carmencita pink’ (Y020), ‘Carmencita bright red’ (Y021), ‘Impala’ (Y022), ‘Zanzibariensis’ (Y023) and ‘Carmencita’ (Y024).

Table 3. Summary of the clustering results for integral normalized, pareto scaled and mean-centered LC–MS TIC dataset and integral normalized and autoscaled extracted ion dataset showing the number of clusters that only consist of samples of same cultivar. Additionally, the cultivars that form such pure clusters are listed. In the best scenario, six pure clusters are observed and the results of different linkage algorithms are consistent.

<table>
<thead>
<tr>
<th>Ward’s method</th>
<th>Nearest neighbour</th>
<th>Furthest neighbour</th>
<th>Average paired distance</th>
<th>K means</th>
<th>Centroid</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIC dataset</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 All</td>
<td>Y019, Y021, Y023, Y024</td>
<td>Y019, Y023, Y024</td>
<td>6 Y019, Y020, Y021, Y023, Y024</td>
<td>5 Y019, Y020, Y021, Y023, Y024</td>
<td>2 Y019, Y020, Y023</td>
<td>3 Y019, Y020, Y023</td>
</tr>
<tr>
<td>Extracted ion dataset</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Y019</td>
<td>1 Y019</td>
<td>1 Y019</td>
<td>1 Y019</td>
<td>1 Y019</td>
<td>1 Y019</td>
<td>1 Y019</td>
</tr>
</tbody>
</table>
Figure 21. The dendrogram obtained with the nearest neighbor algorithm from the integral normalized, pareto scaled and mean-centered LC–MS TIC dataset. The cultivars are coded as follows: ‘Gibsonii’ (Y019), ‘Carmencita pink’ (Y020), ‘Carmencita bright red’ (Y021), ‘Impala’ (Y022), ‘Zanzibariensis’ (Y023) and ‘Carmencita’ (Y024).
4.4.2.3. Data from proton nuclear magnetic resonance spectroscopy

For the data of the region containing the aromatic proton signals, five out of seven clustering algorithms found that ‘Carmencita’ samples form a cluster free from other cultivars (the two inconsistent ones being furthest neighbour and median algorithms, see Appendix 4 for a dendrogram). Additionally, Ward’s method and average paired distance method found ‘Zanzibariensis’ samples to cluster, as well as ‘Carmencita Bright red’ with exception of sample Y021cS04. The uniqueness of Y021cS04 was also stated by furthest neighbour algorithm.

For the saccharide region data, four algorithms (Ward’s method, furthest neighbour, average paired and $k$ mean) found ‘Gibsonii’ to form a pure cluster (see Appendix 4 for dendrograms). $K$ means also found ‘Impala’ to cluster – something that all the rest of the algorithms more or less agree with. Three algorithms (Ward’s, furthest neighbour and $K$-means) found ‘Carmencita’ to cluster with the exception of sample Y024eS04. To get further information whether saccharides can be used to discriminate between cultivars, another and more simple region containing only the signals from anomeric protons was analyzed. None of the clustering algorithms found any good clustering in this dataset, although ‘Impala’ seemed to cluster to some extent.

For the amino acid region, Ward’s method found all cultivars to form a separate cluster, with the exception of the ‘Impala’ cluster that was missing two samples (Y022dS04 and Y022fS04) that formed a seventh cluster (see Figure 22 for the dendrogram). However, the other algorithms did not show much support for this finding (see Appendix 4 for a contrasting dendrogram by furthest neighbour), and the only cultivar agreed to form a pure cluster was found to be ‘Carmencita’ (only median algorithm failed to find this). Besides Ward’s method, average paired distance was the only one finding the clustering of ‘Gibsonii’ and ‘Carmencita Bright red’. Furthest neighbour is the only algorithm agreeing with Ward’s method that ‘Zanzibariensis’ forms a pure cluster, although they all indicate its clustering to some extent.

The final NMR dataset analyzed was a large dataset including most parts of the whole NMR spectra, excluding known solvent peaks and long segments of baseline. The resulting dendrogram is shown in Figure 23. Ward’s method, furthest neighbour and average paired
distance algorithms found that ‘Gibsonii’, ‘Carmencita bright red’ and ‘Carmencita’ clustered nicely. All three also found that ‘Impala’ samples formed a pure cluster, but the sample Y022cS04 being an outlier and not included in the cluster. Additionally, ‘Zanzibariensis’ was found to cluster, but the samples Y023bS04 and Y023dS04 were outside the cluster forming a small cluster of their own. Compared to the amino acid data, which showed best discrimination from the individual regions, not as many cultivars can be discriminated, but the clustering result is more consistent between the different algorithms. All clustering results are summarized in Table 4.

Table 4. Summary of the clustering results for the different integral normalized and autoscaled 1H NMR datasets, showing the number of clusters that only consist of samples of same cultivar. Additionally, the cultivars that form such pure clusters are listed. In the best scenario, six pure clusters are observed and the results of different linkage algorithms are consistent.

<table>
<thead>
<tr>
<th>Ward’s method</th>
<th>Nearest neighbour</th>
<th>Furthest neighbour</th>
<th>Average paired distance</th>
<th>K means</th>
<th>Centroid</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aromatic region</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Y023, Y024</td>
<td>Y024</td>
<td>0</td>
<td>1</td>
<td>Y024</td>
<td>Y024</td>
</tr>
<tr>
<td>Nearest neighbour</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y023, Y024</td>
<td>Y024</td>
<td>-</td>
<td>Y023, Y024</td>
<td>Y024</td>
<td>Y024</td>
<td>-</td>
</tr>
<tr>
<td>Furthest neighbour</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y023, Y024</td>
<td>Y024</td>
<td>-</td>
<td>Y023, Y024</td>
<td>Y024</td>
<td>Y024</td>
<td>-</td>
</tr>
<tr>
<td>Average paired distance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y023, Y024</td>
<td>Y024</td>
<td>-</td>
<td>Y023, Y024</td>
<td>Y024</td>
<td>Y024</td>
<td>-</td>
</tr>
<tr>
<td>K means</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Y019</td>
<td>Y019</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Centroid</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y019, Y019</td>
<td>Y019</td>
<td>Y019</td>
<td>Y019</td>
<td>Y019</td>
<td>Y019</td>
<td>-</td>
</tr>
<tr>
<td>Median</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y019, Y022</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saccharide region</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>Y019</td>
<td>Y019</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Anomeric region of saccharides</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>Y019, Y020, Y021, Y023, Y024</td>
<td>Y024</td>
<td>Y023, Y024</td>
<td>Y019, Y021, Y024</td>
<td>Y024</td>
<td>-</td>
</tr>
<tr>
<td>Amino acid region</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Y019, Y020, Y021, Y023, Y024</td>
<td>Y024</td>
<td>Y023, Y024</td>
<td>Y019, Y021, Y024</td>
<td>Y024</td>
<td>-</td>
</tr>
<tr>
<td>Large dataset including most signals</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Y019, Y020, Y021, Y024</td>
<td>Y019, Y023</td>
<td>Y019, Y021, Y024</td>
<td>Y019, Y021, Y024</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Figure 22. The dendrogram obtained with Ward’s method from the integral normalized and autoscaled ¹H NMR dataset representing the region of proton signals from amino acids. The cultivars are coded as follows: ‘Gibsonii’ (Y019), ‘Carmencita pink’ (Y020), ‘Carmencita bright red’ (Y021), ‘Impala’ (Y022), ‘Zanzibariensis’ (Y023) and ‘Carmencita’ (Y024).

Figure 23. Dendrogram of the largest NMR dataset, including most parts of the whole NMR spectra, but excluding solvent peaks and long segments of baseline. Obtained with Ward’s method from integral normalized and autoscaled data. The cultivars are coded as follows: ‘Gibsonii’ (Y019), ‘Carmencita pink’ (Y020), ‘Carmencita bright red’ (Y021), ‘Impala’ (Y022), ‘Zanzibariensis’ (Y023) and ‘Carmencita’ (Y024).
4.4.3. Results of principal component analysis

4.4.3.1. Data from liquid chromatography with ultraviolet light detection

The integral normalized and autoscaled dataset was analyzed by PCA and the resulting eigenvalues, the amount of variance captured (%) for a given PC and the cumulative variance captured (%) – numbers all indicating the amount of information contained in the principal components – are shown in Table 5 for the first 20 principal components.

<table>
<thead>
<tr>
<th>Principal Component</th>
<th>Eigenvalue of Cov(X)</th>
<th>% Variance captured by this PC</th>
<th>% Variance captured total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.06e+003</td>
<td>25.80</td>
<td>25.80</td>
</tr>
<tr>
<td>2</td>
<td>7.91e+002</td>
<td>19.29</td>
<td>45.09</td>
</tr>
<tr>
<td>3</td>
<td>5.16e+002</td>
<td>12.58</td>
<td>57.67</td>
</tr>
<tr>
<td>4</td>
<td>2.78e+002</td>
<td>6.77</td>
<td>64.44</td>
</tr>
<tr>
<td>5</td>
<td>2.18e+002</td>
<td>5.33</td>
<td>69.77</td>
</tr>
<tr>
<td>6</td>
<td>1.77e+002</td>
<td>4.33</td>
<td>74.10</td>
</tr>
<tr>
<td>7</td>
<td>1.32e+002</td>
<td>3.22</td>
<td>77.31</td>
</tr>
<tr>
<td>8</td>
<td>1.06e+002</td>
<td>2.59</td>
<td>79.90</td>
</tr>
<tr>
<td>9</td>
<td>1.02e+002</td>
<td>2.48</td>
<td>82.38</td>
</tr>
<tr>
<td>10</td>
<td>9.53e+001</td>
<td>2.32</td>
<td>84.70</td>
</tr>
<tr>
<td>11</td>
<td>8.23e+001</td>
<td>2.01</td>
<td>86.70</td>
</tr>
<tr>
<td>12</td>
<td>6.75e+001</td>
<td>1.65</td>
<td>88.35</td>
</tr>
<tr>
<td>13</td>
<td>5.77e+001</td>
<td>1.41</td>
<td>89.76</td>
</tr>
<tr>
<td>14</td>
<td>4.66e+001</td>
<td>1.14</td>
<td>90.89</td>
</tr>
<tr>
<td>15</td>
<td>4.08e+001</td>
<td>0.99</td>
<td>91.89</td>
</tr>
<tr>
<td>16</td>
<td>3.77e+001</td>
<td>0.92</td>
<td>92.81</td>
</tr>
<tr>
<td>17</td>
<td>3.54e+001</td>
<td>0.86</td>
<td>93.67</td>
</tr>
<tr>
<td>18</td>
<td>3.04e+001</td>
<td>0.74</td>
<td>94.41</td>
</tr>
<tr>
<td>19</td>
<td>2.91e+001</td>
<td>0.71</td>
<td>95.12</td>
</tr>
<tr>
<td>20</td>
<td>2.69e+001</td>
<td>0.66</td>
<td>95.78</td>
</tr>
</tbody>
</table>

The eigenvalues for the principal components in Table 5 are also graphically visualized in Figure 24 – a graph, which is one possible way to determine the appropriate number of PCs to keep in the model.\textsuperscript{31} For example, it is seen in the figure that the first four principal components contain by far the most information and starting from PC5 only little information is added to the model by inclusion of an additional PC. As seen in Table 5, the first four PCs explain a total of 64.44 % of the variance, which is a fair amount. The relatively low amount of variance explained by the first PCs were considered to be due to magnified noise level – a side effect of autoscaling discussed more a little later.
Figure 24. The eigenvalues of the first 20 principal components. This figure is one of the several ways to determine the amount of PCs to keep in the model. Since the first four principal components contain the most information and adding more PCs add only little information to the model, taking only the first four to the model would be a rational choice.

Another method used to determine the amount of PCs to keep was cross-validation and “Leave-one-out” algorithm was selected. The cross-validation results are shown in Figure 25, in which the reconstruction error, namely the Root Mean Square Error of Cross-Validation (RMSECV), is plotted against the principal component number. It can be seen from the figure that the reconstruction error decreases as more PCs are taken into the model, until seven PCs are included. Inclusion of PC8 and PC9 actually increases the reconstruction error, suggesting that they are describing a significant amount of noise. From the cross-validation results it was assumed that the first six or seven PCs contain more or less meaningful information that can be used to look for similarities and dissimilarities between cultivars and all the rest PCs described mainly noise. Based on Figure 24 and Figure 25, the first six PCs were taken into the model and the rest were discarded. Compared to the model containing only the first four PCs, this model describes a little more variation (74.10 % compared to 64.44 %, see Table 5) and has somewhat smaller reconstruction error.
The scores for PCs 1–6 are shown in Figure 26. It is obvious in the figure that the separation of different classes representing the cultivars is excellent. The Figure 26a shows that PC1 alone separates the six classes into three groups: first with highly positive scores including ‘Carmencita Pink’ and ‘Carmencita’, second with average scores near zero, including ‘Gibsonii’ and ‘Carmencita Bright Red’, and third with highly negative scores including ‘Impala’ and ‘Zanzibariensis’. While ‘Impala’ and ‘Zanzibariensis’ do not separate from each other on PC1 (Figure 26a), they separate completely on PC2 (Figure 26b) and same is true for ‘Gibsonii’ and ‘Carmencita Bright Red’, as well as for ‘Carmencita Pink’ and ‘Carmencita’ too. Therefore, only two PCs are needed to see a clear separation between all the cultivars. Still, PC3 gives further information on the cultivars, showing that ‘Gibsonii’ is in some way completely different from the rest of the cultivars (Figure 26c). PC4 also shows some capability of telling difference between the cultivars (Figure 26d), in contrast to PCs 5 and 6 (Figure 26e and f) that seem to find differences between the samples representing the same cultivar, possibly because of difference between samples inside the same class or simply because of noise. It may be no coincidence that the first three PCs that explained by far the most variance (Figure 24) are best at discriminating between classes.
The scores on PC1 (a), PC2 (b), PC3 (c), PC4 (d), PC5 (e) and PC6 (f) calculated for an integral normalized and autoscaled LC–UV dataset. The PCs 1–4 primarily separate the classes/cultivars from each other, but PCs 5 and 6 also separate individual samples out of a cluster formation. The classes are ‘Gibsonii’ (+), ‘Carmencita Pink’ (■), ‘Carmencita Bright Red’ (●), ‘Impala’ (○), ‘Zanzibariensis’ (▲) and ‘Carmencita’ (▼).

Figure 26. The scores on PC1 (a), PC2 (b), PC3 (c), PC4 (d), PC5 (e) and PC6 (f) calculated for an integral normalized and autoscaled LC–UV dataset. The PCs 1–4 primarily separate the classes/cultivars from each other, but PCs 5 and 6 also separate individual samples out of a cluster formation. The classes are ‘Gibsonii’ (+), ‘Carmencita Pink’ (■), ‘Carmencita Bright Red’ (●), ‘Impala’ (○), ‘Zanzibariensis’ (▲) and ‘Carmencita’ (▼).

The combined discriminative ability of PC1 and PC2, describing a total of 45% of the variance, proved to be excellent and is shown in Figure 27. At least for such a small amount samples (six per cultivar), the separation of the clusters looks almost ideal. In a less ideal case, where two or more clusters overlapped, it could still be possible to well discriminate between the classes, since only two dimensions are observed in Figure 27, where all the samples lie in a plane. It is not seen in the figure if some of the classes are located above or below the other classes in space. Therefore it can be useful to view the samples from a different angle, as is done in Figure 28. In the figure it is seen that ‘Gibsonii’ cluster, located in the middle of the plot in Figure 27, floats far above the other classes in a third dimension, defined by PC3. If this still wasn’t enough, as many dimensions can be used to examine the data as there are principal components included in the model. Because six PCs are included in this model, the data resides in a six-dimensional space. It is likely, however, that using to PC5 or PC6 to view the data is not helpful, because they seem to rather scatter the classes than to cluster them (Figure 26e and f).

69
Figure 27. Scores on PC 1 plotted against scores on PC 2. These two PCs have the best separation performance and their combined separation ability is excellent. All six cultivars form easily recognizable clusters with reasonably low distance between the samples in a same cluster. Together the first two PCs explain 45 % of the variance. The classes are ‘Gibsonii’ ( ), ‘Carmencita Pink’ ( ), ‘Carmencita Bright Red’ ( ), ‘Impala’ ( ), ‘Zanzibariensis’ ( ) and ‘Carmencita’ ( ).

Figure 28. Scores on PC 1 plotted against scores on PC 3. On PC 3, ‘Gibsonii’ separates from the rest. A total of 38 % variance is explained by the PCs. The classes are ‘Gibsonii’ ( ), ‘Carmencita Pink’ ( ), ‘Carmencita Bright Red’ ( ), ‘Impala’ ( ), ‘Zanzibariensis’ ( ) and ‘Carmencita’ ( ).
What variables or peaks are causing the scores in Figure 26 can be determined by examining the loadings plots. The more a variable is responsible for separating a sample or class away from the average (the zero level) on a PC, the higher loading that variable has on that PC (either negative or positive). A high loading is observed as a large deviation from the zero level (average) in a loadings plot. The loadings plots for the two most informative PCs, PC1 and PC2, are shown in Figure 29. Unfortunately, while autoscaling as a data pretreatment method works excellently to highlight the clustering for this data, it does have an unwanted side effect of producing loadings plots very difficult to interpret, as is the case in Figure 29. This effect is also acknowledged in the literature.\textsuperscript{34,39} The reason for the noisy plots is the scaling operation: since autoscaling has a goal of giving more weight for small signals in the data, also random noise signals are given more weight, making them appear larger in the plots, while some originally large signals containing genuine information are given less weight, making them appear smaller in the plots. This “explosion of noise” may not be a problem, if the noise is properly removed from the data prior to autoscaling,\textsuperscript{34} but such noise removal had not been done here.

![Figure 29](image)

Figure 29. The loadings on PC1 (a) and PC2 (b). The loadings plots look noisy and are not easily interpretable, as a high-looking loading on a variable could be just a result of noise. There are certainly more peaks in the plots than there are in the chromatograms.

The Hotelling’s $T^2$ values and Q residuals that describe the variance explained by the model and the variance left unmodeled, respectively,\textsuperscript{31} are plotted against the variables in Figure 30. In addition to the loadings plots, these two plots are also difficult to interpret. By looking at the plots, it looks like the variance is more or less evenly distributed across all the variables, although some peaks can be recognized. However, telling which ranges of variables really are peaks and which are not, is not very easy.
Figure 30. The Hotelling’s $T^2$ values (a) and $Q$ residuals (b) for all variables in the dataset. The first plot (a) is showing in which variables the variance described by the model is located. The second plot (b), on the other hand, is indicating which variables contain the variance not explained by the model.
4.4.3.2. Data from liquid chromatography–mass spectrometry

For the integral normalized, pareto scaled and mean-centered TIC dataset analyzed by PCA, the appropriate amount of PCs to keep was again determined using the same methods used with LC–UV data. The first 4–5 PCs contained most of the systematic variance, but the inclusion of six PCs seemed to significantly decrease the cross-validation residuals, indicating that the information in the sixth PC would be useful (see Appendix 5 for the plots leading to this conclusion). Therefore, six PCs describing a total of 82.31 % variance were kept. The scores on all PCs are shown in Figure 31. PC2, explaining 18.19 % of the variance, was found to best discriminate between cultivars, while PC1 finds ‘Zanzibariensis’ exceptional from other cultivars. PC3, on the other hand, is least useful. The combined discrimination ability of PC1 and PC2 is shown in Figure 32.

Figure 31. The scores on PC1 (a), PC2 (b), PC3 (c), PC4 (d), PC5 (e) and PC6 (f) calculated from an integral normalized, pareto scaled and mean-centered LC–MS TIC dataset. The classes are ‘Gibsonii’ (+), ‘Carmenita Pink’ (■), ‘Carmenita Bright Red’ (♦), ‘Impala’ (◇), ‘Zanzibariensis’ (▲) and ‘Carmenita’ (▼).
Figure 32. Scores on PC1 plotted against scores on PC2, showing good clustering. Most of the clusters (excluding ‘Impala’) are even better separated, than for the LC–UV data. The explained variance is 49\%.

The variance explained and not explained by the model is seen in Figure 33, where it is seen that the variance left unmodeled is mainly noise. The peptide biomarkers RCB-1–3, seen as three peaks in the data point range 1400–1600 in Figure 33a, seem to have a fair influence on the model. The peaks or variables responsible for the discrimination can be tracked down by examining in the loadings plots in Figure 34.

Figure 33. The variance explained by the model (a) and left unmodeled (b) for the integral normalized, pareto scaled and mean-centered TIC dataset.
In the Figure 34a it is seen that four peaks have the highest influence on PC1: one having apex at variable number 1000, two peaks at variable range 1100–1300 and one having apex around variable number 1500. These four peaks can be easily recognized from the original chromatographic profile in Figure 14b. The first-mentioned peak is the only one having highly negative loadings, while the rest have highly positive loadings. Since ‘Zanzibariensis’ samples are having highly negative scores on PC1 (Figure 31a), they are immediately expected to have the intensity of the negatively loading peak being significantly higher that the intensities of the positively loading peaks. This expectation is confirmed by the Figure 35e showing the plotted intensities of the peaks in question for ‘Zanzibariensis’ (Y023). The high intensities of the two peaks at the data point range 1100–1300 are explaining the relatively higher scores for ‘Carmencita Bright Red’ (Y021) and ‘Impala’ (Y022) (compare Figure 31a to Figure 35e–f), but the low intensity of the peak at variable number 1000 (Figure 35h) seems also important, especially in explaining the high scores for ‘Carmencita’ (Y024).
The loadings plot for PC2 (Figure 35b), which was best at discriminating between cultivars, looked confusing. It was clear that two peaks within data point range 1100–1300 had the most influence on PC2, but the same two peaks seemed to have both positive and negative loadings at the same time: the first half of the peak having negative loading and the remaining half having positive loadings. This was considered to be due to poor aligning of these peaks. Worse misalignments had been observed, but in this particular case it had a large effect on the outcome. The suspected misalignment is more readily seen in the mean-centered data (Figure 36b) than in the unpreprocessed data (Figure 36a). Unfortunately, it could therefore be possible, that the fine discrimination observed is not due to biological
variation, but instead due to artificial variation resulting from instrumental analysis, for example. It should be quite easy to find out if misalignment is causing the discrimination, by carefully aligning the peaks and then re-analyzing the data. The reason for the misalignment could be anything from differing pH of the samples or eluents used in the runs to the two compounds responsible for the peaks being different compounds (perhaps slightly differing analogs of the same compound), or just pure coincidence resulting from chromatographic drift.

![Figure 36. The LC–MS TIC dataset before (a) and after (b) preprocessing (integral normalization, pareto scaling and mean-centering).](image)

PCs 3 and 4 are interesting, in that they largely describe the variance of the three ricin peptide biomarkers peaks RCB-1–3 in the data point range 1400–1600 (Figure 34c–d). Unfortunately, RCB-1 and RCB-2 are not chromatographically well resolved, the difference in their retention being only 0.3 min that results in only one broad peak showing in the TIC, unless the intensities of these two peaks are significantly different. On the other hand RCB-3 is a little better separated and elutes 0.5 min later than the later-eluting one (RCB-2) of the two other compounds. All three biomarkers can be seen in the loadings plot of PC3, RCB-2 in the middle having the highest loading (it is also the most intensive of these three in all the chromatograms) and RCB-3 being the last and slightly the best separated of the three. Interestingly, PCs 3 and 4 separated two ‘Zanzibariensis’ samples (Y023bS02 and Y023dS02) far away from the others.

The presence of outliers in the LC–MS TIC model can be ruled out by looking at Figure 37a, showing that all the variance within the model is inside the 95% confidence limit (under the dash line) and all samples have more or less similar influence on the model. Also, the variance left unmodeled is similar for all samples (Figure 37b).
Figure 37. The amount of variance within the model coming from each sample (a). No sample is responsible of causing such a great amount of variance that it would end up on the other side of the (blue) dash line in the upper end of the left plot and be considered as an outlier. Additionally, the unmodeled variance coming from different samples is shown (b).

For the integral normalized and mean-centered extracted ion dataset, two PCs were found to describe 97% of the variance, shown in Figure 38. While most of the cultivars are clustered, but not well separated, the model found the most important features in the data. Firstly, the model sees that ‘Gibsonii’ is the only cultivar completely lacking RCB-3. Secondly, ‘Zanzibariensis’ samples were very inconsistent with the RCB peak areas, as has been already discussed. The model also recognizes that there is something in common with ‘Zanzibariensis’ samples Y023bS02 and Y023dS02, agreeing with the results of the TIC dataset (see scores on PC3 and PC4 in Figure 31c and d). This common denominator is known to be the lack of RCB-3 and it is the reason why these samples cluster with ‘Gibsonii’ samples (Figure 38), also missing RCB-3.

PC1 described mostly RCB-2 and ricinine, but also RCB-2 to some extent (the figures are found in Appendix 6, if needed). PC2, on the other hand, described mainly RCB-1 and ricinine as well as RCB-3 a little. The suspected demethyl ricinine had no role in the model, probably due to its low intensity compared to the other peaks. Also RCB-3 was mostly unmodeled.
Figure 38. Scores on PC1 plotted against scores on PC2 for integral normalized and mean-centered extracted ion dataset. The plot describes 97% of the variance in data. The classes are ‘Gibsonii’ (+), ‘Carmencita Pink’ (■), ‘Carmencita Bright Red’ ( ◆), ‘Impala’ ( ◇), ‘Zanzibariensis’ ( ▲) and ‘Carmencita’ ( ▼).
4.4.3.3. Data from proton nuclear magnetic resonance spectroscopy

The best discrimination achieved with different NMR datasets is summarized in Figure 39.

![Figure 39](image_url)  
**Figure 39.** Plotted scores on the two most important PCs for different datasets, namely aromatic region (a), amino acid region (b), saccharide region (c), anomeric region (d), combined regions (e) and (f). The data for all plots have been integral normalized, plots a–e are mean-centered and the plot f is the autoscaled version of plot e.

The scores plots from PCA in Figure 39 show that some discrimination between cultivars is achieved. Some clustering is observed, although the separation is modest. The necessity of variable scaling was considered and tested, but didn’t seem to improve discrimination. This is in contrast to chromatographic data, where discrimination of the cultivars was significantly improved by auto or pareto scaling. However, autoscaling did seem to benefit the discrimination when the large combined dataset was in question. This makes sense, since the peak intensities of the other regions are something very different from what they are in the amino acid region (see Figure 16), and combination of these datasets leads to negligible modelling of the those peaks that are significantly less intensive, unless the data is scaled. The loadings plots of the data sets can be seen in Appendix 7, if desired.
4.4.4. Evaluation of the results

Both HCA and PCA found good clustering of samples from the LC–UV data, very evident from the dendrograms and scores plots. Both methods suggested that discrimination between cultivars was very good. Both methods also agreed that only ‘Gibsonii’ forms well separated cluster in the LC–MS extracted ion data. As for LC–UV data, both methods found good clustering for LC–MS TIC data. Although generally HCA found that the LC–MS TIC data forms good clustering, some algorithms did not find ‘Impala’ to form a six-membered cluster pure from other classes, as was the case with ‘Carmencita pink’. According to the scores plot of Figure 32 from PCA, ‘Impala’ has quite long within-cluster distances and form a long chain-like cluster that is not very well separated from ‘Carmencita pink’, thus being consistent with HCA results. The clustering of ‘Carmencita pink’ in the scores plot is also clearly inferior to the other cultivars, excluding ‘Impala’. For NMR data, neither of the methods found as clear clustering as they found from chromatographic data. However, HCA seemed to find better clustering for the individual signal regions than PCA, probably due to autoscaling, as the individual region data for PCA were not autoscaled, but mean-centered. On the other hand, the combined region dataset was autoscaled for PCA and a little better clustering is seen there (Figure 39f). From this dataset most clustering algorithms of HCA found ‘Gibsonii’, ‘Carmencita bright red’ and ‘Carmencita’ to form separate clusters, and the same cultivars can be seen to form the tightest clusters in the scores plot of (Figure 39f). It is concluded that results of HCA and PCA are generally consistent.

Although the chromatographic techniques gave consistent results, chemometric analysis on NMR data failed to find well-separated clusters. The reason for this is unknown, but one difference between the NMR datasets and chromatographic datasets are the amount of peaks in the spectrum/chromatogram. The NMR spectra are “forests” of peaks, whereas the chromatograms have quite a few (resolved) peaks. The moderate (although not robust) discrimination observed in some NMR datasets with certain data preprocessing approaches suggests that there is information in the data that was simply obscured by irrelevant peaks. The issue could therefore also lie in the data preprocessing. While binning was not found beneficial in the early tests of this study, it could be given another try, since binning is very
common in the field with NMR spectroscopy. In the worst case scenario, the class separation for the chromatographic techniques would be artificially produced and was not of biological origin, resulting from misaligned peaks, for example. Some peak misalignments were observed for LC–UV and LC–MS TIC datasets, both giving good clustering. For LC–UV this was observed in loadings plots of a pareto scaled dataset that was not reported here, because of superior performance of the autoscaled dataset. However, such a slight misalignment resulting in such a fine clustering that was observed seems very unlikely. Nevertheless, further study is needed to find out whether misaligned peaks have any significant influence (good or bad) on the outcome and for this purpose advanced peak alignment methods, such as the *icoshift* algorithm, should be used. *Icoshift* is a promising downloadable open source peak alignment algorithm for Matlab, suitable for both chromatographic and NMR spectroscopic data.

5. CONCLUSIONS

Employing chemometrics to different applications has increased during the past decade and seems to only be restrained by the required expertise. There are problems in the field of chemometrics, such as confusion caused by inconsistent use of terms, *e.g.* “metabolomics” and “metabonomics”, unreliable conclusions due to insufficient model validation in significant portion of the published papers, and missing standard operating procedures.

In the experimental work of this study, discrimination between cultivars of *Ricinus communis*, based on chemometric analysis of chromatographic data was found possible, with the previously described sample preparation and data preprocessing. Despite the minimal preprocessing of the datasets, good clustering was obtained and results from HCA and PCA results were found to be consistent. However, at this point it cannot be ruled out that at least some of the discrimination between cultivars is due to artificial systematic variation from insufficiently aligned peaks. While it is unlikely that misaligned peaks would result in such a good discrimination as was observed, its influence is worth investigating by repeating the chemometric analysis of the data with emphasis on peak alignment.
The failure to find satisfactory class separation with the NMR spectroscopic approach is likely to be due to interesting information being obscured by irrelevant information, rather than bad instrumental data or inexistent pattern in the data. The NMR spectra were significantly more complex than the chromatograms and the weak appearance of the potential pattern could be due to some peaks varying significantly also within the samples of same cultivar and could possibly be bypassed by narrowing or changing the studied spectral region. As a conclusion, the issue is likely to lie in the data mining approach.

Based on the good class separation observed in the chromatographic data, it seems plausible that computer could be trained to classify seeds of unknown *Ricinus communis* varieties and therefore automatically detect the variety. This conclusion is based on unsupervised pattern recognition results of six cultivars that separated in six classes. It remains to be investigated if all classes remain well separated when more cultivars are added to the study, but the current promising results strongly encourage continuing the research.

6. ACKNOWLEDGEMENTS

I want to thank Professor Paula Vanninen for giving me the interesting topic for my thesis and for compensation during my experimental work. I also want to thank M. Sc. Marja-Leena Rapinoja for all her time performing the sample preparation and for her idea to adopt the previously elsewhere reported chemometric approach to study *Ricinus communis* also at VERIFIN. I want to thank Ph. D. Harri Koskela for his help with the Matlab software and for supervising my work, as well as to thank Ph. D. José Ruiz-Jiménez for his help with the chemometric techniques and for co-supervising my work. Finally, I want to thank my dearest fiancée Laura Anttila for enduring all the long days I’ve spent with this work.
7. REFERENCES


84


8. APPENDICES

Appendix 1. Product ion scan data for a sample (Y021eS02) observed to be abundant in an unknown compound suspected to be demethyl ricinine (a) and the product ion data of ricinine acquired from a reference standard (b). For both data there are three chromatograms: one TIC and two extracted ion chromatograms (EIC). The upmost chromatogram is the TIC, the second is EIC of ricinine (m/z 165) and the third is EIC of the suspected demethyl ricinine (m/z 151). Along with the chromatograms, a mass spectrum for each extracted peak is shown (if found). From the castor bean extract both ricinine and its suspected analog are found and their product ion spectra look very similar: only visible cleavage is m/z 27, possibly hydrogen cyanide. The normalized collision energy used was 23 % and was determined by optimizing the collision energy for the ricinine standard to get the maximum TIC signal.
Appendix 1 (continued). The suspected ricinine analog was not found in the ricinine standard (b).
Appendix 2. An example of peak picking for a castor bean sample (Y021fS02). The five picked peaks in the picture are ricinine (m/z 165), suspected demethyl ricinine (m/z 151), RCB-1 (m/z 690), RCB-2 (m/z 661) and RCB-3 (m/z 655).
Appendix 3. Quality control data for chromatographic methods. The reproducibility was determined for both instruments: LC–UV (a) and LC–MS (b). For LC–UV more runs were made simply because one run was short in duration (10 min). The quality control data for QC experiments run prior to the castor bean samples analyzed with LC–UV (c) and LC–MS (d) are also provided.

a) Reproducibility test with 21 consecutive runs for LC–UV.

<table>
<thead>
<tr>
<th>Run #</th>
<th>Retention time [min]</th>
<th>Intensity [mAu]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Peak 1</td>
<td>Peak 2</td>
</tr>
<tr>
<td>1</td>
<td>2,39</td>
<td>2,63</td>
</tr>
<tr>
<td>2</td>
<td>2,39</td>
<td>2,63</td>
</tr>
<tr>
<td>3</td>
<td>2,40</td>
<td>2,63</td>
</tr>
<tr>
<td>4</td>
<td>2,39</td>
<td>2,63</td>
</tr>
<tr>
<td>5</td>
<td>2,40</td>
<td>2,63</td>
</tr>
<tr>
<td>6</td>
<td>2,39</td>
<td>2,63</td>
</tr>
<tr>
<td>7</td>
<td>2,40</td>
<td>2,63</td>
</tr>
<tr>
<td>8</td>
<td>2,39</td>
<td>2,63</td>
</tr>
<tr>
<td>9</td>
<td>2,40</td>
<td>2,63</td>
</tr>
<tr>
<td>10</td>
<td>2,39</td>
<td>2,63</td>
</tr>
<tr>
<td>11</td>
<td>2,39</td>
<td>2,63</td>
</tr>
<tr>
<td>12</td>
<td>2,39</td>
<td>2,62</td>
</tr>
<tr>
<td>13</td>
<td>2,39</td>
<td>2,63</td>
</tr>
<tr>
<td>14</td>
<td>2,39</td>
<td>2,62</td>
</tr>
<tr>
<td>15</td>
<td>2,39</td>
<td>2,63</td>
</tr>
<tr>
<td>16</td>
<td>2,39</td>
<td>2,63</td>
</tr>
<tr>
<td>17</td>
<td>2,39</td>
<td>2,62</td>
</tr>
<tr>
<td>18</td>
<td>2,39</td>
<td>2,63</td>
</tr>
<tr>
<td>19</td>
<td>2,39</td>
<td>2,63</td>
</tr>
<tr>
<td>20</td>
<td>2,39</td>
<td>2,63</td>
</tr>
<tr>
<td>21</td>
<td>2,40</td>
<td>2,63</td>
</tr>
</tbody>
</table>

Average: 2,39 2,63 3,01 3,57
Standard deviation: 0,00 0,00 0,00 0,01
RSD: 0,2 % 0,1 % 0,1 % 0,1 %

The four quality control compounds for LC–UV was a commercial LC–NMR quality control mixture “NIP-MIX-4” (Bruker BioSpin, Germany), whose contents are not revealed by Bruker BioSpin.

b) Reproducibility test with six consecutive runs for LC–MS.

<table>
<thead>
<tr>
<th>Run #</th>
<th>RT (min)</th>
<th>Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Peak 1</td>
<td>Peak 2</td>
</tr>
<tr>
<td>1</td>
<td>4.54</td>
<td>5.60</td>
</tr>
<tr>
<td>2</td>
<td>4.55</td>
<td>5.59</td>
</tr>
<tr>
<td>3</td>
<td>4.57</td>
<td>5.58</td>
</tr>
<tr>
<td>4</td>
<td>4.56</td>
<td>5.64</td>
</tr>
<tr>
<td>5</td>
<td>4.54</td>
<td>5.62</td>
</tr>
<tr>
<td>6</td>
<td>4.53</td>
<td>5.59</td>
</tr>
</tbody>
</table>

Average: 4.55 5.60 6.20 5.96
Standard deviation: 0,01 0,02 0,02 0,02
RSD (%): 0,3% 0,4% 0,4% 0,3%

The four quality control peptides for LC–MS were glycine-tyrosine (m/z 239.1), methionine enkephalin (m/z 574.2), leucine enkephalin acetate (m/z 556.3) and angiotensin II (m/z 523.8²).
Appendix 3 (continued).

The table below shows the quality control runs for LC–UV and LC–MS batches. These runs were performed prior to analyzing castor bean samples on different days.

### c) Quality control runs for LC–UV batches

Run prior to analyzing castor bean samples on different days.

<table>
<thead>
<tr>
<th>Batch #</th>
<th>RT (min)</th>
<th>Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Peak 1</td>
<td>Peak 2</td>
</tr>
<tr>
<td>1</td>
<td>2.30</td>
<td>2.53</td>
</tr>
<tr>
<td>2</td>
<td>2.33</td>
<td>2.58</td>
</tr>
<tr>
<td>3</td>
<td>2.30</td>
<td>2.54</td>
</tr>
<tr>
<td>4</td>
<td>2.31</td>
<td>2.55</td>
</tr>
<tr>
<td>5</td>
<td>2.34</td>
<td>2.59</td>
</tr>
</tbody>
</table>

**Average:**

- RT (min): 2.32, 2.56, 2.95
- Intensity: 3.53

**Standard deviation:**

- RT (min): 0.02, 0.03, 0.04, 0.06
- Intensity: 16, 19, 22, 27

**RSD (%):**

- RT (min): 0.8 %, 1.0 %, 1.4 %, 1.8 %
- Intensity: 1.1 %, 1.3 %, 1.6 %, 2.1 %

### d) Quality control runs for LC–MS batches

Run prior to analyzing castor bean samples on different days.

<table>
<thead>
<tr>
<th>Batch #</th>
<th>RT (min)</th>
<th>Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Peak 1</td>
<td>Peak 2</td>
</tr>
<tr>
<td>1</td>
<td>4.57</td>
<td>5.59</td>
</tr>
<tr>
<td>2</td>
<td>4.62</td>
<td>5.64</td>
</tr>
<tr>
<td>3</td>
<td>4.63</td>
<td>5.59</td>
</tr>
<tr>
<td>4</td>
<td>4.56</td>
<td>5.58</td>
</tr>
<tr>
<td>5</td>
<td>4.62</td>
<td>5.59</td>
</tr>
</tbody>
</table>

**Average:**

- RT (min): 4.60, 5.60, 6.19, 5.97
- Intensity: 1,40E+04, 1,05E+06, 5,09E+05, 3,20E+05

**Standard deviation:**

- RT (min): 0.03, 0.02, 0.03, 0.02
- Intensity: 7E+02, 2,1E+05, 5,2E+04, 2,5E+04

**RSD (%):**

- RT (min): %, %, %, %
- Intensity: 5.3 %, 20.1 %, 10.2 %, 7.7 %
Appendix 4. Examples of dendrograms obtained from different NMR datasets. The cultivars are coded as follows: ‘Gibsonii’ (Y019), ‘Carmencita pink’ (Y020), ‘Carmencita bright red’ (Y021), ‘Impala’ (Y022), ‘Zanzibariensis’ (Y023) and ‘Carmencita’ (Y024).

a) Dendrogram of the aromatic region dataset analyzed using Ward’s method.

b) Dendrogram of the main saccharide region dataset analyzed using Ward’s method.
Appendix 4 (continued).

c) Dendrogram of the region containing anomeric protons of saccharides. Dataset was analyzed using Ward’s method.

d) Dendrogram of the amino acid region dataset analyzed using the furthest neighbor algorithm.
Appendix 5. Determining the appropriate amount of PCs to include in the model calculated from integral normalized, pareto scaled and mean-centered TIC dataset.

The amount of information contained by the PCs (a) and the reconstruction error (cross-validation residual) of the model as more PCs are taken into the model (b). The former plot shows the sixth PC containing very little additional information, but according to the latter plot the residuals decrease significantly (error is decreased).
Appendix 6. Data for the extracted ion dataset. Loadings on PC1 (a), loadings on PC2 (b), the variance described by the model (c) and the unmodeled residual variance (d). The variables “1” and “2” are derived from peak areas of ricinine and demethyl ricinine, respectively. The variables “3”, “4” and “5” are the corresponding values for RCB-1–3.
Appendix 7. Loadings plots for the different $^1$H NMR datasets (first two PCs only). All data has been integral normalized and mean-centered, except for the "combined region" whose data has been integral normalized and autoscaled.
Appendix 7 (continued).

AMINOACID REGION

COMBINED REGION