UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE

New tools for exploratory analysis fusing information from different sources

PhD thesis 2019 | Nicola Cavallini

UNIVERSITÀ DEGLI STUDI
DI MODENA E REGGIO EMILIA

Dottorato di ricerca in
"Models and methods for material and environmental sciences"

Ciclo XXXI

New tools for exploratory analysis fusing information from different sources

in co-tutela con la Københavns Universitet

Candidato: CAVALLINI Nicola

Relatore italiano (Tutor): Prof.ssa Marina Cocchi

Relatore danese (Tutor): Prof. Rasmus Bro

Coordinatore del Corso di Dottorato: Prof.ssa Maria Giovanna Vezzalini

This page intentionally left blank

New tools for exploratory analysis fusing information from different sources

PhD thesis

by

Nicola Cavallini

Department of Food Science

Faculty of Science

University of Copenhagen

under the double degree agreement with

Università degli Studi di Modena e Reggio Emilia

Title

New tools for exploratory analysis fusing information from different sources

Submission date

February 20th, 2019

Final revisions

March 19th, 2019

Defence date

March 29th, 2019

Supervisors

Associate Professor Marina Cocchi
Department of Chemical and Geological Sciences, University of Modena and Reggio Emilia, Italy

Professor Rasmus Bro
Department of Food Science, Faculty of Science, University of Copenhagen, Denmark

Opponents

Professor Beata Walczak
Department of Analytical Chemistry, Institute of Chemistry, University of Silesia, Poland

Associate Professor Klavs Martin Sørensen
Department of Food Science, Faculty of Science, University of Copenhagen, Denmark

Professor Maria Cristina Menziani
Department of Chemical and Geological Sciences, University of Modena and Reggio Emilia, Italy

Cover photo by Nicola Cavallini

Cover's font: Leelawadee
Thesis's font: Cambria (large majority of the text parts)

Table of Contents

« Quality. That's what they'd been talking about all the time. “Man, will you just please, kindly dig it,” he remembered one of them saying, “and hold up on all those wonderful seven-dollar questions? If you got to ask what is it all the time, you'll never get time to know.” Soul. Quality. The same? »

ROBERT M. PIRSIG, *Zen and the art of motorcycle maintenance*

La perfezione è un falso e rende pazzi
E invece questo è il circo di 'sticazzi

ELIO E LE STORIE TESE, *Il circo discutibile*

PINGUINI TATTICI NUCLEARI, *Montanelli (intro) + Sciare*

Preface

This thesis has been submitted to the PhD School of Models and Methods for Material and Environmental Sciences, University of Modena and Reggio Emilia, as well as to the PhD School of the Faculty of Science, University of Copenhagen, in fulfilment of the requirements to obtain the PhD degree.

I did not consciously plan to do a PhD. I have experienced swinging moods about my whole PhD path, especially during writing this thesis, which is one of the results that I reached during the last three years spent working in the field of chemometrics. I also managed to create a network of colleagues, that in many cases I can hardly identify as such: "friends" is a concise but clear definition (not the Facebook meaning of the word).

I hope that you, the reader, will find this manuscript balanced enough between the more technical sections like the novel exploratory tool described and tested in Chapter 3, and the more experimental parts about whisky and beer analysis, especially the "linguistics study" of Chapter 4. I am not ashamed to admit that the latter has been the most enjoyable part to write, and taste.

Nicola Cavallini
Torino, March 2019

List of publications

Published (1)

Fused Adjacency Matrices to enhance information extraction: the beer benchmark

Nicola Cavallini, Francesco Savorani, Rasmus Bro, Marina Cocchi
Analytica Chimica Acta (2019) doi:10.1016/J.ACA.2019.02.023.

Close to submission (3)

Beer's linguistics and chemistry: an investigation

Nicola Cavallini, Francesco Savorani, Rasmus Bro, José Camacho Páez, Marina Cocchi

The whisky space: a chromatographic point of view

Nicola Cavallini, Mikael Agerlin Petersen, Jens Risbo, Rasmus Bro, Marina Cocchi

Advances in Glioma Grading: a metabolomic data fusion approach

Valeria Righi, Nicola Cavallini, Antonella Valentini, Giampiero Pinna, Giacomo Pavesi, Maria Cecilia Rossi, Annette Puzzolante, Adele Mucci, Marina Cocchi

List of abbreviations and notation

Chemometric methods

PCA = Principal Component Analysis

PC = Principal Component

MCR = Multivariate Curve Resolution

PARAFAC = PARAllel FACtor Analysis

SOM = Self-Organizing Kohonen's Map

SMR = Sparse Matrix Regression

OPTICS = Ordering Points To Identify the Clustering Structure

RD = Reachability Distance

RP = Reachability Plot

PA = Procrustes Analysis

SNV = Standard Normal Variate

PCA–GCA = Principal Component Analysis–Generalized Canonical Analysis

t-SNE = t-distributed Stochastic Neighbour Embedding

PMD = Penalized Matrix Decomposition

Analytical methods

Vis = Visible

NIR = Near-Infrared

NMR = Nuclear Magnetic Resonance

GC-MS = Gas Chromatography-Mass Spectrometry

DHS = Dynamic Headspace Sampling

Other

ABV = Alcohol by volume (%)

NAS = Non-Age Statement

NLP = Natural Language Processing

Notation

$\mathbf{X}$ = matrix

$\mathbf{X}^{\mathbf{T}}$ = transpose of a matrix

$\bar{\mathbf{X}}$ = averaged matrix

$\hat{\mathbf{X}}$ = reconstructed data

$\mathbf{v}$ = vector

s = scalar

Abstract

The main focus of this PhD project was on the development and application of new chemometric tools for multivariate exploratory analysis for dealing with data not showing simple groupings or trends, even when projected to spaces of lower dimensionality. Such data may be so complex that common visualizations tools are only shedding limited light on the underlying structures. Starting from these premises, the *Fused Adjacency Matrix* approach was developed as main outcome of the project. The approach was tested on a benchmark dataset of beer samples acquired using three spectroscopic techniques, namely visible, near-infrared (NIR) and nuclear magnetic resonance (NMR) spectroscopies.

Another important part of the PhD project concerned the extension of methods for integration of data sources of very different nature, like numerical and text data, within the food chemistry framework. As a matter of fact, analytical chemistry in synergy with advanced data analysis methods can be profitably used to build new tools to aid consumers to choose and pair foodstuff as well as producers to meet the consumers' expectations and desires. In this perspective, an investigation of the links between the "objective" world of analytical chemical profiling and the "subjective" world of consumers tasting and describing food was carried out, in the context of beer analysis and consumption. By means of text analysis methods, a set of user-generated reviews were processed and converted into a numeric format suitable for data analysis, and then linked by principal component analysis–generalized canonical analysis (PCA–GCA) to the spectral information provided by Visible, NIR and NMR spectroscopies.

Dansk Resumé

Hovedfokus i dette ph.d.-projekt var udvikling og anvendelse af nye kemometriske værktøjer til multivariate eksplorative analyser af data, der ikke umiddelbart viser simple grupperinger eller tendenser, selv når de projiceres til rum af lavere dimensionalitet. Sådanne data kan være så komplekse, at almindelige visualiseringsværktøjer kun kaster begrænset lys på de underliggende strukturer. Ud fra disse præmisser blev der udviklet et sæt metoder kaldet *Fused Adjacency Matrix*. Denne nye tilgang til data-analyse blev testet på et benchmark datasæt af ølprøver erhvervet ved anvendelse af tre spektroskopiske teknikker, nemlig synlige, nær-infrarøde (NIR) og NMR-spektroskopier.

En anden vigtig del af ph.d.-projektet vedrørte udvikling af metoder til integration af datakilder af meget forskellig art, såsom tal og tekstdata. Analytisk kemi kombineret med avancerede dataanalysemetoder kan med fordel anvendes til at opbygge nye værktøjer til eksempelvis at hjælpe forbrugerne med at vælge og parre levnedsmidler og også hjælpe producenter med at imødekomme forbrugernes forventninger og ønsker. I dette perspektiv blev der foretaget en undersøgelse af forbindelserne mellem den "objektive" verden af analytisk kemisk profilering og den "subjektive" verden af forbrugere, der smager og beskriver mad i forbindelse med øl-analyse. Ved hjælp af tekstanalysemetoder blev et sæt brugergenererede anmeldelser behandlet og omdannet til et numerisk format, der var egnet til dataanalyse, og derefter komineret ved hjælp af "principal component analysis–generalized canonical analysis" (PCA–GCA) med spektralinformation fra Vis, NIR og NMR spektroskopier.

Riassunto

L'obiettivo principale del progetto di Dottorato è stato lo sviluppo e l'applicazione di nuovi strumenti chemiometrici per l'analisi esplorativa multivariata atti al trattamento di dati caratterizzati dall'assenza di chiari raggruppamenti o tendenze. Dati di questo tipo possono rivelarsi complessi a tal punto che i normali strumenti di visualizzazione risultano poco efficaci nell'estrarre e descrivere le strutture latenti. Sulla base di queste premesse è stato sviluppato l'approccio *Fused Adjacency Matrix* (matrice di adiacenza fusa), il quale rappresenta il risultato principale del progetto. L'approccio è stato testato mediante un dataset di riferimento ottenuto da misurazioni spettroscopiche (visibile, vicino infrarosso e risonanza magnetica nucleare) su campioni di birra. Inoltre, per comprenderne più a fondo le caratteristiche di funzionamento, l'approccio è stato testato su altri dataset di diversa natura.

Un'altra parte importante del progetto di Dottorato ha riguardato lo sviluppo di modelli per integrare dati da sorgenti di diversissima natura, come ad esempio dati numerici e di testo, nel contesto della scienza degli alimenti. I metodi della chimica analitica possono essere utilizzati sinergicamente con l'analisi dei dati avanzata per la creazione di strumenti utili al consumatore, fornendo assistenza nella scelta degli alimenti, nell'appaiamento dei sapori, e al contempo per aiutare i produttori a soddisfare le esigenze e i desideri dei clienti. In questa prospettiva è stato quindi sviluppato uno studio dei collegamenti tra il mondo "oggettivo" della chimica analitica e quello "soggettivo" del consumo e della valutazione degli alimenti, sulla base dei dati spettroscopici sulla birra (già oggetto dello sviluppo di *Fused Adjacency Matrix*). Una serie di recensioni raccolte online è stata processata mediante metodi di analisi del testo e trasformata in formato numerico, idoneo per essere analizzato con metodi comuni di analisi del dato. Tale dato è stato poi collegato all'informazione spettrale mediante un metodo chiamato analisi delle componenti principali–analisi canonica generalizzata (principal component analysis–generalized canonical analysis, PCA–GCA), dal quale è stata ottenuta informazione utile a collegare le espressioni utilizzate per descrivere la birra in relazione ai composti chimici identificati mediante la spettroscopia.

Chapter 1 | Introduction

1.1. General context

The main focus of this PhD project was on the development and application of new chemometric tools for multivariate exploratory analysis for dealing with very information rich data. Especially data where it is difficult to elucidate the underlying structure in terms of e.g. groupings or trends. Such data may be so complex that common visualizations tools are only shedding limited light on the underlying structures. Starting from these premises, the *Fused Adjacency Matrix* approach was developed as main outcome of the project.

The overall idea of the proposed approach was inspired by the concept of combining so-called "weak sources" of information, which is taken from the supervised classification context, where the combination of multiple weak classifiers is aimed at providing better discriminatory information [1].

The *Fused Adjacency Matrix* approach is based on the fusion of several adjacency matrices (i.e. the weak sources of information) obtained from different distance measures [2] and a neural network method [3]. The approach is also suitable as a mid-level data fusion [4] tool to combine data obtained from different analytical platforms (e.g. spectroscopic fingerprints). The approach was tested on a benchmark dataset of beer samples acquired using three spectroscopic techniques, namely visible, near-infrared and nuclear magnetic resonance spectroscopies. The results of this study are reported in Chapter 3 of this thesis as well as in a recently published paper [5].

Another important part of this PhD project concerned the extension of methods for integration of data sources of very different nature, like numerical and text data, within the food chemistry framework. As a matter of fact, analytical chemistry in synergy with advanced data analysis methods can be profitably used to build new tools to aid consumers to choose and pair foodstuff as well as producers to meet the consumers' expectations and desires. In this perspective, an investigation of the links between the "objective" world of analytical chemical profiling and the "subjective" world of

consumers tasting and describing food was carried out, the results of which are reported in Chapter 4 of this dissertation.

Online reviews about the same beer samples as the aforementioned beer datasets were harvested from a social network dedicated to beer reviewing. By means of text analysis methods [6,7] these user-generated reviews were processed and converted into a numeric format suitable for data analysis. Principal component analysis–generalized canonical analysis (PCA–GCA, [8]) was used to investigate the links between spectral and text data, leading to very interesting results.

1.2. Organization of the thesis

The thesis is structured in five chapters, including the Introduction. A brief summary of all the remaining parts of the thesis:

Chapter 2: the chemometric background supporting the whole thesis is given, as well as a description of the analytical techniques employed for obtaining the data analysed in the other chapters.

Chapter 3: the novel proposed method for exploratory analysis and mid-level data fusion is first introduced and illustrated, and then tested on the datasets described in Chapter 2.

Chapter 4: an investigation study of the connections between the analytical world of spectroscopy with the world of consumer tasting, in the framework of beer analysis is reported; this chapter is the very basis of an article close to submission in the present moment.

Chapter 5: the final remarks and future perspectives conclude the thesis.

References | Chapter 1

[1] Chuanyi Ji, Sheng Ma, Combinations of weak classifiers, IEEE Trans. Neural Networks. 8 (1997) 32–42. doi:10.1109/72.554189.

[2] R. Todeschini, D. Ballabio, V. Consonni, Distances and Other Dissimilarity Measures in Chemometrics, Encycl. Anal. Chem. Appl. Theory Instrum. (2015) 1–34. doi:10.1002/9780470027318.a9438.

[3] F. Marini, R. Bucci, A.L. Magrì, A.D. Magrì, Artificial neural networks in chemometrics: History, examples and perspectives, Microchem. J. 88 (2008) 178–185. doi:10.1016/j.microc.2007.11.008.

[4] E. Borràs, J. Ferré, R. Boqué, M. Mestres, L. Aceña, O. Busto, Data fusion methodologies for food and beverage authentication and quality assessment – A review, Anal. Chim. Acta. 891 (2015) 1–14. doi:10.1016/j.aca.2015.04.042.

[5] N. Cavallini, F. Savorani, R. Bro, M. Cocchi, Fused Adjacency Matrices to enhance information extraction: the beer benchmark, Anal. Chim. Acta. (2019). doi:10.1016/J.ACA.2019.02.023.

[6] R.E. Banchs, Text Mining with MATLAB®, Springer New York, New York, NY, 2013. doi:10.1007/978-1-4614-4151-9.

[7] A. Hotho, A. Nürnberger, G. Paaß, F. Ais, A Brief Survey of Text Mining, 2005. http://www.crisp-dm.org/Process/index.htm.

[8] A.K. Smilde, I. Måge, T. Naes, T. Hankemeier, M.A. Lips, H.A.L. Kiers, E. Acar, R. Bro, Common and distinct components in data fusion, J. Chemom. 31 (2017) e2900. doi:10.1002/cem.2900.

Chapter 2 | Materials and Methods

2.1. Chemometric background

> « *The art of extracting chemically relevant information from data produced in chemical experiments is given the name of '**chemometrics**'* […] »
>
> – Svante Wold [1]

Extracting chemically relevant information, is the heart of this quote from Svante Wold, one of the two widely recognized fathers of Chemometrics. The quote's message should not be a surprise, since the very reason why Chemometrics was born in the first place is that there was a strong need for tools able to extract the relevant information from the increasingly bigger and more complex data that any scientist struggled analysing and fully understanding. As time goes by, with the Digital Revolution first and the advent of the Information Age then, computers became widespread and powerful enough for both testing and refining "old" methods and for developing new ones.

Nowadays we can use very well-established methods to face a large variety of data types and challenges. Depending on the scope, the chemometrician may need to build *unsupervised* models, e.g. when the aim is to explore the data or there is no additional information about the set of measures under examination. On the contrary, *supervised* methods may be required when the need is to model a response, for instance if a property is not directly measurable or measuring it by traditional means is too time-consuming or expensive, and one may desire to use other, cheaper and faster techniques to achieve the same result. In any case, a sound approach to data analysis always involves a step of *unsupervised* data modelling, to both gain knowledge about the quality of the data and to decide which further direction should/can be taken.

It is important to consider that even though data may only seem numbers on spreadsheets, the information they carry, the structures that relate objects and variables upon which the measures were made can be very different, therefore an

approach tailored to the problem is always required. The chemometrician should picture himself/herself at the border between the modelling problem and the physical/chemical system under examination, with the overall driving force of understanding. For example, spectral data such as near infra-red spectra may seem easy to manage, analyse and interpret because of their (sometimes quite boring) look, but with strong band overlaps, background effects, scattering and water influence they can be as complicated as nuclear magnetic resonance spectra, with their often overwhelming abundance of peaks and all the many problems such as peaks shifting and heterogeneous baseline effects.

In such complex cases a different approach may be required. For instance, an interval-based approach aimed at extracting relevant features from the NMR data may greatly improve their interpretability and usability, simplifying the modelling problem from both the points of view of data processing and computational resources.

In this Section, an overview of the chemometric tools used in this thesis for exploratory analysis and feature extraction is given. Please refer to Chapter 3 for the more specific chemometric background of the Fused Adjacency Matrix approach.

2.1.1. Exploratory data analysis

The very first step in data analysis is the quality assessment of the collected data, and their consequent exploration. Data is not information, and a starting point for distinguishing among signals (i.e. information) and sources of, in a broad sense, noise, is found in Exploratory Data Analysis (EDA).

In 1977 John Tukey published a book [2] that is nowadays considered a milestone reference for EDA. Tukey described a framework for data analysis based on statistically relevant visual representations of datasets, aimed at helping the analyst to formulate hypotheses and gain deeper understanding of the phenomena occurring in the data [3]. The data and the information they bring is the focus, not any hypothesis or prior knowledge on the system under examination. On this basis, the aim of EDA is to reveal hidden and unknown information [3]. To this aim, a rather rich set of established exploratory tools is available, and new methods are constantly developed.

Most EDA methods are *unsupervised*, which means that no *a priori* assumptions are imposed when modelling the data: the analyst's intuitions or prior knowledge on the set of samples under examination or the experiment is not *actively* used in the modelling process, but instead can be verified/validated more efficiently when interpreting the modelling results.

The unsupervised approach in EDA is a good way to provide new, sometimes unexpected, information. Unsupervised methods such as Multivariate Curve Resolution (MCR, [4]), Principal Component Analysis (PCA, [5]), some of its variants like Maximum Likelihood PCA [6] or Projection Pursuit PCA [7,8], or linear methods such as Independent Component Analysis (ICA**,** [9,10]) and Multidimensional Scaling (MDS, [3,11]), provide easy, 2D or 3D representations of the groupings and structures present in the data.

Non-linear mapping methods like Kohonen's Self-Organizing Maps (SOMs, [12–14]) are considered complementary to methods like PCA [15] because of their ability to account for non-linear phenomena. SOMs, for instance, is a method that also provides an easy, low-dimensional representation of the data structures but it does so by training a neural network. Groups, or clusters of objects are also provided by the clustering methods [16], whose final aim is to detect and represent how the objects group and how distant each cluster and/or sample are. All these approaches are commonly applied for inspecting structures in the samples space, but PCA and MCR, for instance, also provide easy-to-access information about the variables' influence on these structures.

From the analyst's point of view, the fact that these methods supply information about the natural groupings present in the data – provided that proper care is taken during the data pre-processing and modelling – is of great importance: deeper understanding of the phenomena occurring in the data can be attained, when "unbiased", *unsupervised* sources of information are combined with the analyst's knowledge of the larger picture.

2.1.1.1. Principal Component Analysis (PCA)

Principal Component Analysis (PCA, [5,17]) is probably the most famous multivariate tool for exploratory data analysis. It is an unsupervised decomposition method that operates a projection of the data from the original high-dimensional space to a space of lower dimensionality, which is defined by a set of new variables, called Principal Components (PCs).

The PCs are derived by linear combination of the original variables, and they account for the largest sources of variability. PCs are *nested*, meaning that, starting from the first PC describing the direction of maximum variance in the data, each following component in turn will have the highest possible variance, under the constraint of orthogonality with respect to all the preceding components. Each PC is composed of a vector of scores **t** and a vector of loadings **p**. The PCA decomposition can be represented by Equation 2.1:

$$\mathbf{X} = \sum_{f=1}^{F} \mathbf{t}_f \cdot \mathbf{p}_f^{\mathrm{T}} + \mathbf{E} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} = \hat{\mathbf{X}} + \mathbf{E} \quad (2.1)$$

In Equation 2.1 the original data **X** are modelled using *F* PCs, therefore the $\hat{\mathbf{X}}$ matrix represents the modelled part of the original data, while the **E** matrix corresponds to the unmodeled part, or the residuals. Given that set of PCs corresponds to the axes of the low-dimensional space the data are projected to, the values of the score vector $\mathbf{t}_f$ represent the sample's coordinates on the *f*th PC or axis, and the loadings vector $\mathbf{p}_f$ represents the contribution weights of the original variables to that PC. Another way of representing the PCA model is depicted in Figure 2.1, which highlights how the product $\mathbf{T}\mathbf{P}^{\mathrm{T}}$ represents the modelled part of the original data, as opposed to the residual matrix **E**.

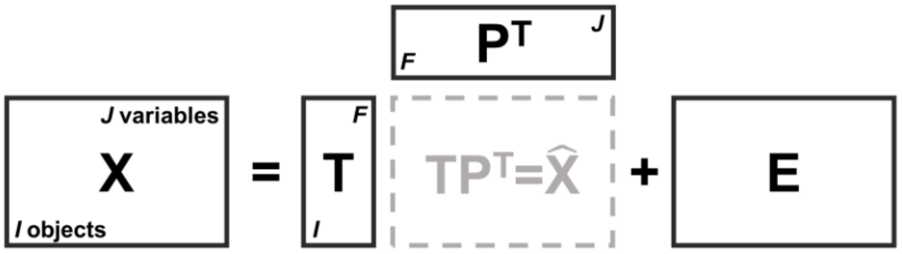


Figure 2.1. Graphical representation of a PCA model.

Ideally, in a good PCA model all the structured variability is included in $\hat{\mathbf{X}}$, the modelled part, while all the random and uninformative variability of the original data is left in the residuals matrix **E**.

Data described by a PCA model can be inspected by looking at the score plots, the loadings plots and the residuals plots. Groups of similar samples and their distribution can be inspected with the score plots, which are a 2D or 3D representation of the modelled data and are obtained by plotting one score vector **t** against another. Likewise, the loadings plots allow inspecting groups and relations among the original variables, whose values on the different PCs (the values contained in the loading vectors **p**) also describe how the original variables influence each PC. For the same pair of inspected PCs, directions on the scores and loadings plots are coincident.

The fact that the original variables are combined into a few new ones makes PCA also a good "compression" or features extraction method, which is extensively used for handling large datasets and reducing computational time needed for modelling.

2.1.1.2. Multivariate Curve Resolution (MCR)

Multivariate Curve Resolution (MCR, [4,18]) is a decomposition method that can extract pure contributions from overlapped signals. From a mathematical point of view, MCR is related to PCA, but its components are not forced to be orthogonal: for this reason, to reduce rotational ambiguity it is necessary to constraint the components.

$$\mathbf{X} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E} = \hat{\mathbf{X}} + \mathbf{E} \tag{2.2}$$

Even if the decomposition equations of PCA (Eq. 2.1) and MCR (Eq. 2.2) look the same, the way MCR works is quite different. Instead of describing the largest sources of variability (i.e. maximizing the variance explained by each component), MCR aims to obtaining pure signals and their relative concentration in each sample. For this reason, PCA's scores matrix **T** translates into the pure concentrations matrix **C** and PCA's loadings matrix **P** translates into the pure resolved spectra matrix **S** of Equation 2.2, also represented in Figure 2.2.

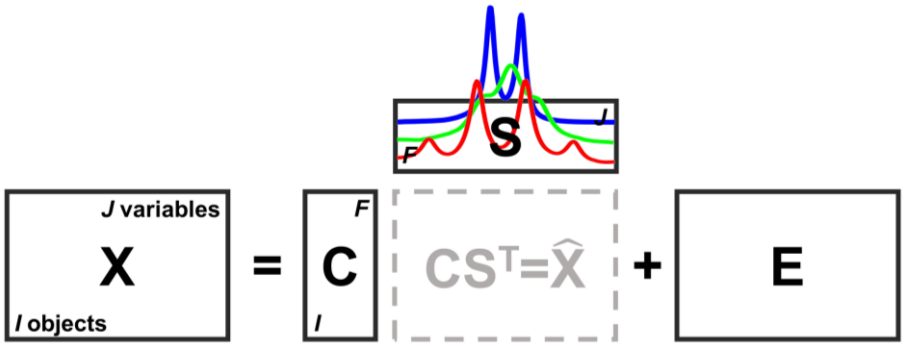


Figure 2.2. Graphical representation of an MCR model.

MCR is based on Beer's Law [18], since it treats the input data as a mixture of pure signals mixed in different ratios, the concentrations: resolving such a mixture implies obtaining the pure contributions (the resolved spectra) and their "mixing ratios", the concentrations.

Given that pure signals are obtained, MCR also works as a method for filtering the data: undesired sources of variability such as noise or background effects can be efficiently removed, and end up in the residual matrix **E**.

MCR is particularly suitable for resolving overlapped signals, especially on selected intervals, where a few signals are present and can be more easily extracted. Since each extracted component is characterized by a pure (spectral) profile, it becomes very easy to match it with a known chemical compound or at least interpret it in terms of the latent phenomena generating the resolved signal. For this reason, MCR is a powerful method for obtaining an easy interpretation of data, while achieving strong compression at the same time.

2.1.1.3. PARAllel FACtor Analysis 2 (PARAFAC2)

PARAllel FACtor analysis (PARAFAC, [19,20]) is a multi-way decomposition method that can be considered as a generalisation of PCA. If PCA can handle 2D matrices, or better, 2-way data, PARAFAC can handle *n*-way data. The most common application of PARAFAC is with three-way data, and an example of the PARAFAC decomposition of a three-way dataset $\underline{\mathbf{X}}$ ($I \times J \times K$) with F components (or factors) is given by Equation 2.3:

(2.3) $$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk}$$

where x_{ijk} are the elements of the three-way array **X**, and a_{if}, b_{jf} and c_{kf} are the elements of the three loadings matrices **A**(I×F), **B**(J×F) and **C**(K×F). Similarly to PCA, a three-way PARAFAC model can be represented as shown in Figure 2.3. In its common use, the first mode is associated with the samples and the remaining two modes are associated with the measured variables.

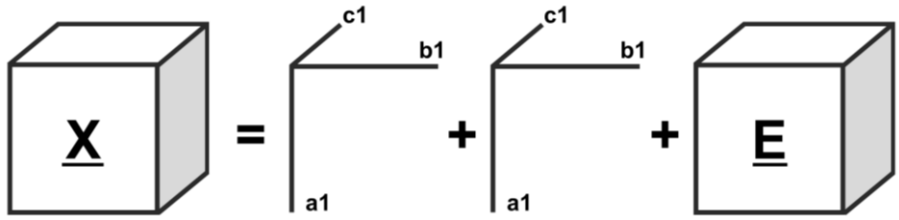


Figure 2.3. Graphical representation of a PARAFAC model.

A good example of three-way data comes from Gas Chromatography-Mass Spectrometry (GC-MS). In this case, the chromatogram's retention time is associated with the second mode, and the mass spectrum is associated with the third mode. This kind of data, however, are better modelled using a more general version of PARAFAC, called PARAFAC2.

PARAFAC2 [21,22] operates much the same as the trilinear decomposition of PARAFAC(1), with the only difference that the shape and the length of the elution profile (keeping the previous example) are not assumed to be the same in each sample [23]: this means that in the second mode one elution profile for each sample is obtained. In this way it is possible to manage shifts in the retention time direction (second mode), which may arise due to experimental factors like column ageing, changes in temperature or variation in the mobile phase flux.

Contrary to PCA, which has solution "rotational freedom", PARAFAC and PARAFAC2 lead to unique solutions. To reach an unambiguous solution to the multilinear problem, three conditions must be satisfied: 1) data must be trilinear; 2) data must show random, not too intense noise; 3) a good estimation of the chemical rank of the system (i.e. the number of independent chemical components) must be provided.

2.1.1.4. Co-clustering with non-negative matrix factorization

Coclustering[1] is a group of methods for exploratory analysis that has its roots in the Seventies [24] but has gained popularity with the growing trend of measuring bigger and bigger amounts of data and, at the same time the need of focusing on finding few relevant variables. The discovery of biomarkers [25,26] and the *omics* fields represent some of the many promising applications for coclustering methods.

Contrary to other exploratory methods such as PCA and traditional clustering, the aim of coclustering is to find and select areas of interest in the data: instead of modelling everything the data matrix is simultaneously clustered in its objects and variables. In this way it is possible to obtain sets of objects sharing a particular behaviour in relation to a select number of variables.

Even though this method is related to clustering in the sense that information about groups is obtained, one of the main differences is that the same object or variable can simultaneously be in different clusters. This feature is referred to as *overlapping coclustering*, in contrast to non-overlapping coclustering or traditional clustering, where each object or variable is assigned to at most one cluster.

One of the mathematical formulations of coclustering consists of performing the decomposition of the data matrix **X** using a bilinear model. The method used in this thesis is described by Bro *et al.* in [27] and is based on Sparse Matrix Regression (SMR, [28]). This algorithm can be considered a soft or fuzzy coclustering algorithm, since the samples' and variables' assignments to a cluster are not binary but can have any value between zero and one. The SMR algorithm operates a bilinear decomposition of the data matrix **X** by minimizing the loss function of Equation 2.4:

$$\|\mathbf{X} - \mathbf{A}\mathbf{B}^{\mathrm{T}}\|_F^2 + \lambda \sum_{i,k} |\mathbf{A}_{ik}| + \lambda \sum_{j,k} |\mathbf{B}_{jk}| \qquad (2.4)$$

where the columns of A ($I \times K$) can be referred to as *scores* and the columns of B ($J \times K$) are the *loadings*, while K corresponds to the number of extracted factors, or coclusters;

[1] An interesting and clear YouTube video about coclustering can be found at: https://www.youtube.com/watch?v=mnDC6hWWbwY (accessed: 08/01/2019)

the parameter λ represents the penalty factor imposed to make the scores and loadings sparse.

In SMR each cocluster is represented by a rank-1 component of the decomposition: in other words, imposing sparsity allows selecting the suitable rows and columns belonging to each cocluster, making all the remaining coefficients correspondent to objects and variables not belonging to the coclusters exactly zero.

Interpreting a coclustering model is very straightforward, since for each group of objects a set of variables is provided. Inspecting the coclusters one by one allows to identify their specific features and since coclustering's solutions are approximately nested [27], coclusters of low-dimensional models will also be present in models with more components. For this reason, it is good practice to inspect many models to choose a reasonable number of coclusters.

Limitations to this method come when the data are very non-quadratic or if they are not discrete. Spectral (continuous) data are difficult to process, probably because the strong natural correlation among those kind of variables makes it difficult to impose sparseness in a meaningful way. A way to bypass this limitation can be to turn the elution profiles into resolved features.

2.1.1.5. Data visualization techniques

This section is devoted to explaining the background theory of the clustering structure-revealing OPTICS algorithm and its use in combination with heatmaps. An example of this combination is given in Figure 2.5.

2.1.1.5.1. Ordering Points To Identify the Clustering Structure (OPTICS)

OPTICS [29–31] is a density-based clustering method aimed at revealing the data clustering structure. This method consists of an iterative procedure that only needs an initial input parameter, namely *k*, which is the minimal number of objects forming a cluster. Daszykowski and Walczak [31] suggested a rule of thumb for choosing the *k* value:

(2.5) $k = \text{integer}\left(\frac{m}{25}\right)$ where m is the number of samples.

However, based on the author's experience working with OPTICS, it often happens that values lower than the ones computed with Equation 6.1 provide better results than sticking to the value obtained by the rule. It is advisable to slightly change this parameter and assess different outputs to obtain a better insight into the structure of the data.

OPTICS is based on the concept of Reachability Distance (RD), an abstract similarity measure [31]. RD is basically a Euclidean distance that describes how distant/similar is an object from the one processed at the preceding step. The graphical output of OPTICS is called Reachability Plot (RP), and it is obtained by plotting the RDs as vertical bars arranged along the x-axis according to the processing sequence.

At each iteration, the OPTICS algorithm selects one object and compares it with all the objects that have not been processed yet. This is done by computing all the pairwise Euclidean distances between the selected object and the ones to be processed. Then, the next object to be processed is selected among the k-nearest neighbours: the distance at which this next object is found becomes its RD, which is stored unchanged until the end of the procedure. The final output is therefore a vector of RD values, which can be plotted as bars in the RP.

A cluster is generally formed by objects that happen to be very close to each other, so it can be expected that these objects would have, on average, a similar number of neighbours at similar distances, i.e. they would have similar neighbourhoods: these short distances among neighbours also result in similar RD values.

Generally, when an entire cluster has been processed, then the next object would likely belong to another cluster. If the inter-cluster distance is larger than the intra-cluster variability, then the next RD value in the processing sequence is going to be larger than the values preceding it, which are related to previous cluster. This "jump" from one cluster to another is graphically recognizable in the RP because it corresponds to a very high bar, standing out among the preceding and following positions. Clusters therefore appear as hollows created by groups of samples sharing similarly low RDs, separated by high bars representing the jump to another cluster.

It is important to consider that the RP does not explicitly cluster the objects [31], but it rather allows deducing the number of clusters in the data. The value of *k* operates as a "smoothing parameter", allowing to obtain deeper or shallower hollows in the RP, therefore highlighting a finer or coarser structure in the data.

2.1.1.5.2. Cluster Heatmaps

The heatmap is a widely used tool all across the fields of science and its earliest appearances can be dated back to the end of the 19th century [32]. It consists of a rectangular tiling, whose tiles are shaded on a colour scale representing the value corresponding to the tile. A two-dimensional data matrix can be easily visually displayed by means of a heatmap, potentially providing clear insight into the data.

However, the origin of data matrices is generally subject to a large variety of factors which may affect the way the samples and variables are organized in the matrix. Regular experimental practice, for instance, envisages experiment randomization when analysing a set of samples, to avoid confusion between time effects (instrumental drifts, laboratory conditions) and a very ordered sequence of experiments. Therefore, just visualizing the data as a heatmap usually does not provide much information about the data structure.

For this reason, the cluster heatmaps were developed over the years and are now used in field like bioinformatics [33] and sensomics [34,35]. Their history and main uses up to 2009 is nicely reported by Wilkinson *et al.* [32]. A cluster heatmap is a clever visualization tool which combines a heatmap and one or more clustering methods to permute (in other words, to reorder) both the rows and the columns of the heatmap. Clustering is then operated once in the samples' direction and once in the variables' direction. The clustering outputs, such as the common clustering tree, are then appended on the sides of the reordered heatmap. The result is a powerful combined representation in which similar samples as well as similar variables are grouped together, allowing to interpret the patterns in the data by directly linking groups of samples to groups of variables. An example with the honey data from Marini *et al.* [36] and OPTICS as a clustering method [30] is given in Figure 2.4.

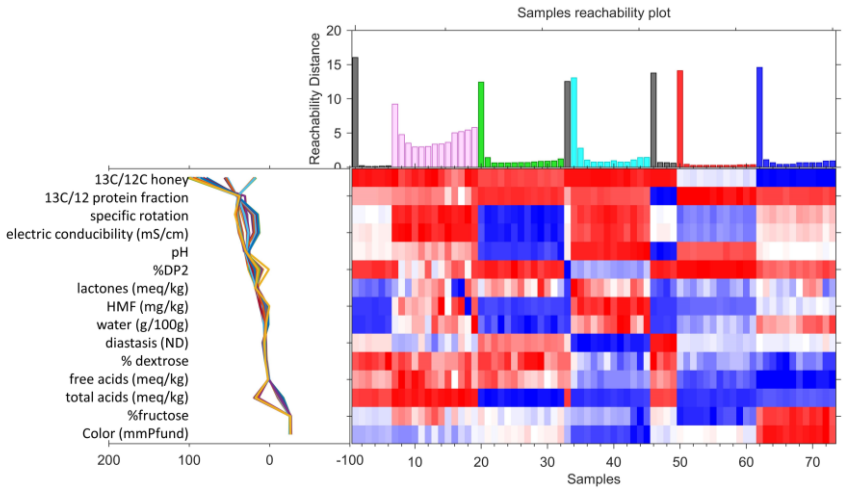


Figure 2.4. Cluster heatmap of the honey data from Marini *et al.* [36].

Different clustering methods can be used for reordering both the columns and the rows, but if the original variables are continuous (e.g. chromatograms, spectra) then reordering is not advisable: in such cases the variables are already "ordered" in an interpretable way, according to their spectral or chromatographic structure.

Another aspect that must be considered with this visualization tool is the distribution of the values of the represented matrix. Extreme values may compromise the interpretability by hiding smaller but potentially meaningful structures. This may not be the case of data such as NIR or Visible spectra, but discrete data may suffer from this effect. A way for dealing with this problem is to normalize the data between zero and one, as it was done in Figure 2.4.

The colour map also plays a role in obtaining a clear and interpretable representation of the data. Linear and diverging colour maps usually make pattern interpretation in data straightforward, as opposed to rainbow colour maps. Linear colour maps follow a linear variation of lightness, either monotonically increasing or decreasing, which is very suitable for general purpose data display [37] and continuous data such as NIR or visible spectra. On the contrary, diverging colour maps are aimed at displaying the data as compared to a well-defined reference value, emphasizing whether a data value lies above or below the reference. The most common diverging map is the red-white-blue map, also used in the large majority of the colored plots of this thesis. Because of

its emphasis on the difference between high and low values, the diverging maps are also suitable for highlighting the extremes.

2.1.2. Features extraction

Features extraction goes together with the concept of *data compression*, since it is aimed at selecting what is important/relevant in the data: features are obtained, and noise/unwanted variability is largely filtered out. The resulting reduced data are generally easier to interpret, as working with fewer variables makes it simpler to inspect and manage the data. Moreover, it is often possible to give meaningful names to the extracted features, like in the case of relative concentrations from MCR or the peak areas from PARAFAC, by identifying and assigning the extracted pure profiles.

These two examples are the most relevant methods for this thesis, especially their application within an interval-based approach [38]. Interval-based approaches or "*i*-chemometric" methods, as Savorani *et al.* [38] also call them, allow extracting large amounts of information and, at the same time, taking care of the local characteristics of the data. Different spectral regions, chromatographic time intervals or just portions of the data may contain very different chemical information, may have different scale or dynamics, or may have different density or noise: with an interval-based approach it is possible to tailor the feature extraction process to the region of interest.

A nice and clear example of the whole process of extracting relevant information from complex data by means of PARAFAC2 interval-modelling can be found in this article [39] by Bevilacqua *et al.*, which focuses on the contributions of chemometrics to the foodomics field.

Finally, another fundamental application of features extraction concerns data-fusion methods [40–42]. In so-called *mid-level* data fusion approaches, features extracted from different data blocks are combined into a new dataset, thus reducing the number of variables, but also exploiting the different – hopefully complementary – pieces of information, often obtained from different analytical techniques. More details about data fusion approaches and theory can be found in Section 3.4.

2.1.3. Procrustes Analysis (PA)

When dealing with different sources of information, e.g. if more data blocks of measures made on the same samples set are available, it can be useful to compare these sources to assess whether they carry the same information or not. For example, a set of objects may be described by two distinct sets of PC scores, obtained from two different analytical sources: Procrustes Analysis (PA, [43,44]) is a method that can be used for the purpose of comparing the two PC sets.

The aim of PA is to obtain the closest match between the two PC spaces by applying operations such as scaling, rotation, reflection and translation. If the two PC sets are named **Z** and **Y**, the mathematical operations operated by PA on **Y** to match it to **Z** are described by Equation 2.5 [44]:

(2.6) $$\mathbf{Z} = a\mathbf{YR} + \mathbf{1_m b} + \mathbf{E} = \hat{\mathbf{Z}} + \mathbf{E}$$

where a is a scalar constant for scaling, **R** is a rotation/reflection matrix, **b** is a translation vector and $\mathbf{1_m}$ indicates a vector of ones. The **E** matrix is the residuals matrix, and all operations in PA are optimized towards the minimization of the sum of squared residuals: in other words, the set of operations leading to the closest match between **Z** and **Y** is sought.

The similarity of the two spaces is expressed using a dissimilarity parameter d (Equation 2.6, [44]), ranging from zero (perfect alignment) to one (no similarity):

(2.7) $$d = \frac{\sum_i \sum_j \left(z_{ij} - \hat{z}_{ij}\right)^2}{\sum_i \sum_j \left(z_{ij} - \bar{z}_{ij}\right)^2}$$

2.2. Analytical Techniques

An overview of the analytical techniques applied in the thesis is given in this section, the principles and data characteristics of each technique are discussed.

2.2.1. Nuclear Magnetic Resonance (NMR) spectroscopy

Nuclear Magnetic Resonance (NMR) spectroscopy is an analytical technique widely used in the field of metabolomics [45,46] and Food science [47]. Due to the large

amount of information that an NMR spectrum can yield, NMR spectroscopy is also a very appealing fingerprinting technique. For instance, some examples from the field of *beeromics* [48] include studying the chemical composition [49–53], performing product authentication [54,55] or the using NMR spectroscopy for quality control purposes [48,56,57]. A very detailed dissertation about the use of NMR spectroscopy within the foodomics approach can be found in a paper from 2014 by Laghi *et al.* [58].

NMR spectroscopy only requires little specimen preparation and since the analysis does not induce any physical or chemical change in the specimen, it is a non-destructive technique. However, even if the specimen's integrity is preserved, some purification steps would be required to recover its original composition.

2.2.1.1. Principles and spectral characteristics

The NMR signal is produced by excitation of the nuclei within the sample using radio waves: when a radio frequency pulse is applied, the nuclei start resonating with it. The frequency at which each nucleus resonates strongly depends on its chemical environment: for this reason, two physically identical nuclei which occupy different positions within the structure of a molecule will generate different signals. Such an influence from neighbouring atoms and/or molecular structures makes the resonance frequency highly characteristic to individual functional groups, therefore allowing to match each signal to the position of the nucleus emitting it within the structure of a molecule.

Only the nuclei which possess a magnetic spin momentum can resonate and be detected in NMR spectroscopy. The most common isotopes are ^{1}H and ^{13}C. Considering the different natural abundances of carbon isotopes[2], the occurrence of ^{13}C is only 1.07%, while ^{12}C represents some 98.93%. The very abundant ^{12}C isotope would be the ideal choice for NMR spectroscopy, but unfortunately it is an NMR-silent nucleus. In addition to hydrogen and carbon, many other nuclei can be used in NMR spectroscopy, if they have a nuclear spin magnetic moment.

[2] https://www.webelements.com/carbon/isotopes.html (accessed: 08/01/2019)

During an NMR experiment, only a specific isotope is stimulated. All the nuclei of that isotope will simultaneously resonate at their own frequency and the signal that is collected by the instrument, the "free induction decay" (FID), is composed by all these contributions. The FID therefore contains all the information generated by the stimulated nuclei. The NMR spectrum is obtained by deconvolution of the FID using the Fourier transform. This conversion from the time domain of the FID to the frequency domain of the spectrum is fundamental to gain access to the collected information, as it allows to obtain the individual contributions to the signal, the NMR peaks.

The peaks' position is referred to as *chemical shift* δ and it is conventionally reported using the dimensionless unit of parts per million (ppm). Raw frequencies are not used because the axis scale would be dependent on the magnetic field strength of the instrument, making interpretation and comparisons among spectra more difficult. Therefore, all spectra are generally referenced to a selected signal, from which the chemical shift δ of a random signal n is thus derived[3]:

$$\delta_n \equiv \frac{\nu_n - \nu_R}{\nu_R} \cdot 10^6 \qquad (2.8)$$

Reference compounds should be as chemically inert as possible, show a well-resolved singlet, and have chemical shifts independent of external variables (i.e. temperature or ionic strength, which are the largest sources of shifting effects). Tetramethylsilane (TMS) and 3-(trimethylsilyl)propanoic acid (TSP) are two of the most commonly used reference compounds for ^{1}H NMR analysis.

2.2.1.2. The NMR spectrometer

Without an intense magnetic field, NMR spectroscopy would not be possible. For this reason, a big magnet is at the core of any NMR spectrometer. The required intense magnetic field is nowadays provided by superconducting solenoid magnets made of a niobium-tin alloy. To exploit the magnet's superconductivity properties very low temperatures are needed. The magnet is therefore fully immersed in a bath of liquid

[3] https://goldbook.iupac.org/html/C/C01036.html (accessed: 08/01/2019)

helium (4 K); the helium tank is then surrounded by a thermal jacket filled with liquid nitrogen (77 K) which, together with an external high-vacuum jacket, acts as a thermal buffer between the core of the system and the surrounding environment.

To collect any signal, the sample needs to be inserted in the system, and placed at the center of the solenoid, where the magnetic field is most intense. An insertion system carries the NMR tube through a cavity (usually form the top of the instrument) to the center of the magnet, where a zone at room temperature is guaranteed. This part of the instrument is provided with the "probe", a device consisting of a series of coils and other systems that represents the "eyes" of the instrument. The probe is both an emitter and a receiver as it is dedicated to generating the radio signals for stimulating the nuclei and to collecting the FID originating from the nuclei. The probe is also equipped with fine tuning systems, which are fundamental for collecting good signals: the magnetic fields that are present within the small area occupied by the sample must be as homogeneous as possible, so that any equivalent nucleus can resonate in the same way.

Field homogeneity is not the only factor that may affect the quality of the collected signal. Temperature fluctuations and vibrations may also cause problems: to keep those aspects under control, external control systems allow to directly operate on parameters such as temperature and field homogeneity, while possible vibrations are reduced using damping systems. The whole instrument is controlled by a dedicated computer, with which it is usually possible to operate the autosampler system, if equipped. Autosampler systems allow to carry out automated analysis sequences, resulting in time savings and human error reduction.

2.2.1.3. Post-acquisition signal processing

Right after collection, the NMR spectra are usually Fourier transformed and inspected. To improve the quality of the original data, some common post-acquisition processing tools such as phase correction, baseline correction and zero-filling are generally employed [59,60]. Once these operations are completed, the spectra are ready for export and further data analysis.

2.2.2. Visible/Near-infrared (Vis/NIR) spectroscopy

Visible (Vis) and Near-infrared (NIR) spectroscopies are fast, non-destructive techniques, and modern instruments allow to record large numbers of reproducible spectra in a short time, with almost no sample preparation required. NIR spectroscopy is especially suitable for very practical applications where a quick response or a pre-screening is required.

Interaction between electromagnetic radiation and matter can occur in many ways. Figure 2.5 represents the electromagnetic spectrum and the conventional regions that are identified and used for technical applications: from signal and information transmission (radio waves), to everyday uses such as food heating and cooking (microwaves) or all the many applications of X-ray in the medical field.

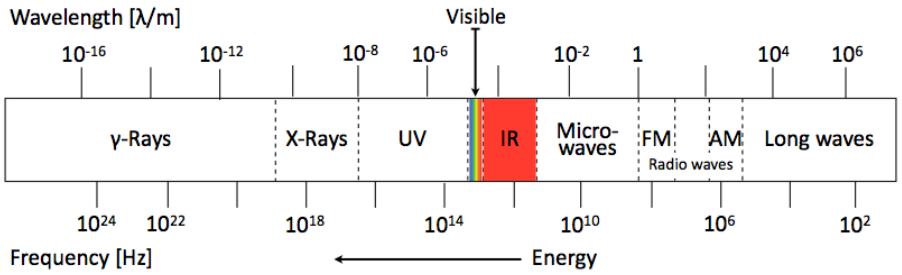


Figure 2.5. the electromagnetic spectrum.

Molecular spectroscopies are based on the interaction between molecules and the central part of the electromagnetic spectrum. Different phenomena arise depending on the considered interval:

- **molecular rotations**: collective motions of the molecules stimulated by microwave and far-infrared (FIR) radiation;
- **molecular vibrations**: relative motions of the atomic nuclei involved in any molecular bond, stimulated by infrared radiation;
- **electronic transitions** to excited states which occur when visible (Vis) and ultraviolet (UV) radiation is involved.

2.2.2.1. Visible spectroscopy

Visible radiation is usually considered together with ultraviolet (UV) radiation, since in both cases the absorption of light from these two intervals results in electronic transitions of bonding electrons.

Molecules responsible for colour are called "chromophores" which means "carrier of colour". These molecules contain π-electrons or non-bonding electrons and can absorb energy in the form of ultraviolet or visible light to excite these electrons to higher energy states. Molecules containing structures such as extended conjugated systems (corresponding to moieties with increased electronic density), double bonds or metal complexes are generally also chromophores. Biological compounds like carotenoids, chlorophylls and anthocyanins are examples of chromophores, as well as melanoidins, brownish heterogeneous polymers resulting from the combination of sugars and amino acids via the Maillard reaction in foodstuff [61].

2.2.2.2. Near-Infrared (NIR) spectroscopy

Taking as a reference the Visible interval (a very human-based point of view), the infra-red (IR) interval is found at lower energy, as the prefix "infra" ("below") suggests. Three sub-regions are generally identified within the IR interval: near-infrared (NIR), middle infra-red (MIR) and far infra-red (FIR). The NIR and MIR intervals are at the basis of two of the most used spectroscopic techniques in analytical chemistry, and even if the physical mechanisms that characterize them are basically the same, they are usually treated and used separately.

MIR radiation directly stimulates the *vibrational modes* of molecules, while in the case of NIR radiation more than one vibrational mode may be stimulated, due to the higher energies involved. Signals originating from integer multiples of normal vibrational modes are called *overtones*, while signals resulting from combinations of integer multiples of normal vibrational modes are referred to as *combination bands*.

If the fundamental vibrational modes are detected by MIR spectroscopy, only the overtones and combination bands can be detected by NIR spectroscopy. Compared to the MIR interval, signals in the NIR interval are much less intense, because their

probability of excitation is much lower than the one associated with the normal vibrational modes. A graphical representation of this difference in intensity is given in Figure 2.6.

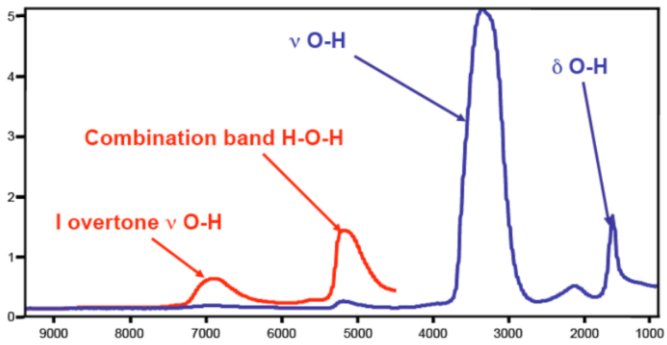


Figure 2.6. Intensity difference between MIR (intense peaks in blue) and NIR (enhanced in red) signals in the water spectrum. (from Paolo Belloni, Brucker)

2.2.2.3. NIR instruments

NIR spectroscopy has nowadays many industrial applications from process control using optical fibres to portable-miniaturized instruments, often dedicated to specific purposes for in-field and in-situ measurements [62,63].

NIR instruments can generally operate in three different modes: transmittance, reflectance and transflectance. Transmittance is, in a sense, the most straightforward mode: a light beam is directed through the specimen (usually a liquid) and is recorded on the other side: transmitted light will result poorer of certain wavelengths, as a result of the interaction between matter and light (i.e. absorption), and the recorded pattern is the transmittance (or absorbance) spectrum. The reflectance mode is used with solid samples and is used in a large variety of industrial applications. In the transflectance mode the light beam goes through the specimen, gets reflected by a surface placed on the other side of the cuvette and passes through the specimen again, finally reaching the detector.

The NIR spectra recorded and analysed in this thesis were obtained with an instrument operating in transflectance mode, for a detailed description of the experimental conditions for the beer datasets, please refer to Section 2.3.2.1.2.

2.2.3. Gas Chromatography-Mass Spectrometry (GC-MS)

Gas Chromatography-Mass Spectrometry (GC-MS) is a hyphenated technique [64] which combines the resolution power of chromatography with the high selectivity and sensitivity provided by the mass spectrometer. For this reason, it can separate and analyse complex mixtures, characterizing all the components both qualitatively and quantitatively.

GC-MS has a long history and one of its most recent and fruitful uses is in the field of metabolomics [65]. Like NMR spectroscopy it has been used in many applications in food science. Its strong points have made it a successful and reliable technique because, in general:

1. it is a widespread instrument, which can be found and afforded by many laboratories;
2. it is cheap and easy to develop and apply methods;
3. it is a robust instrument, with high reproducibility (provided that adequate and regular maintenance is performed);
4. due to its "long" history, many rich reference libraries are available.

However, to balance these strong points, GC-MS also has some drawbacks. Most instruments have quite low resolution and high-resolution instruments (such as the Quadrupole-Time of Flight, the Quadrupole-Ion Trap or the Orbitrap) can be very expensive. Moreover, some sample preparation steps are required, to make the analytes volatile. To this aim, large or non-volatile molecules are difficult, if not generally impossible to analyse.

The principles of GC-MS are briefly described in the next section, for more detailed dissertation reference [66] can be useful.

2.2.3.1. Principles and data characteristics

Separation of the sample's chemical components is achieved by running the sample mixture through a capillary column, using an inert carrier gas, also called "mobile phase". Common carrier gases are helium, hydrogen and nitrogen. The chromatographic column is located in an oven, whose temperature can be precisely

varied, usually following an increasing thermal ramp: the ramp's slope, its starting and ending temperatures depend on the type of experiment.

Analytes in a mixture get separated because of their different affinity with the stationary phase, a thin layer covering the inner surface of the column. Depending on the analytes' polarity, different columns may be required. The overall idea is that analytes with little affinity with the stationary phase will travel faster through the column than those which, on the contrary, may interact more, getting slowed down. The "order of release" from the chromatographer generates the chromatographic profile, which is directly related to the amount of analytes detected at the end of the column.

Right after being released from the chromatographer, the analytes enter the mass spectrometer, where ionization and fragmentation take place. The mass spectrometer acts as a detector and as an analytical instrument, providing both the signal that generates the chromatogram and the information about the chemical structures (mass spectrum) detected at each point of the chromatogram.

Different ionization techniques are available, but electron impact ionization is the one of interest in the context of this thesis. Electron impact shows good sensitivity and produces unique patterns of fragmentation, which are both desired features for running many reproducible experiments. Electron impact happens at the interface between the chromatographer and the mass spectrometer and it is a "hard" ionization method, since highly energetic electrons are used to produce ions. A large amount of energy is transferred to the molecule, which will dissipate this excess of energy by breaking up and producing many fragments. These fragments are detected by their mass-to-charge ratio (m/z).

Instruments equipped with a quadrupole detector can work in two modes, scan or SIM (Selected Ion Monitoring), which generally correspond to two different analytical approaches. The scan mode is usually employed for *untargeted* analyses, therefore when the mass profile of the molecule is unknown or when there is no target molecule. In this mode, a m/z range is defined, and it is scanned by changing the electric potentials of the quadrupole's bars. The scan mode works very well for collecting large amounts of rich data for untargeted analysis, but at the cost of sensibility: compared to

focusing on few m/z values, a whole range has to be covered at each scan therefore, for this reason, during the time spent recording each m/z value a low amount of signal will be recorded.

The SIM mode, on the contrary, allows to focus on known molecules, which can be set as targets to follow by recording their characteristic fragmentation ions. This also allows to focus on specific classes of known molecules, which may be related because of similar fragmentation patterns, e.g. they all lose a peculiar functional group – i.e. a specific ion – which can be traced and quantified. Obviously, all the m/z values that are not recorded are lost, but the boost in sensibility is of great value for *targeted* analyses.

In both cases, the data that are produced consist of a three-way array: each sample can be imagined as a collection of mass spectra, ordered according to the time at which they were recorded. The time mode corresponds to the moment when a particular analyte or group of analytes was released into the mass spectrometer. Ions generated by fragmentation are recorded when they hit the charged bars of the detector: a small amount of electric current is associated to each ion hitting the detector, and by measuring the impact position and charge, the mass spectrum is generated.

The amount of detected analyte is usually represented using a chromatogram, which corresponds to the total intensity recorded over time. Chromatograms in this case can be referred to as "total ion current" profile or TIC.

2.2.3.2. Data pre-processing

Chromatographic three-way data are usually rather rich in information, and the main challenge to the analyst is the process of unravelling it and making it available.

A very simple approach is to work with TIC profiles. This are 1D-signal and provided that peaks are well resolved they may be assigned on the basis of refence databases. However, the large majority of the information carried by the data is ignored, since the whole mass spectrum dimension is removed when TICs are generated. Co-eluting compounds may not be noticed, and even if an internal standard may be used, quantification of peaks containing more than one chemical component may easily result wrong. Moreover, the same peak may result shifted from one sample to another,

because of small variations in the chromatographic conditions, column aging or matrix effects.

The mass spectra yield the key for resolving these situations, since it provides the information about the fragmentation patterns of all the molecules present at that specific point in time (i.e. corresponding to a chromatographic peak). Data processing therefore needs the power of methods like PARAFAC for understanding how many components may be hidden within a single peak.

The GC-MS whisky data described and analysed in this thesis were processed using the PARADISe software [67], a PARAFAC2-based deconvolution and identification system. For a more detailed description of how it was used with the whisky data, please refer to Section 2.3.3.2.1.

PARAFAC2 (described in Section 2.1.1.3) can model and extract different components from a peak made of more than one chemical co-eluting components, handling at the same time possible shifts. The result is that each extracted component can be quantified relatively but also it can be identified by means of its resolved mass spectrum – i.e. its fingerprint.

Such an approach requires some time for processing the data, but it results very efficient: the whole picture is collected at once to be mathematically deconvoluted; the resolved signals can later be matched with digital libraries. Any further analysis will be done on resolved components, each one with a chemical name (certain or tentative). This also represents a good starting point for further refinement e.g. by focusing on specific compounds, their recognition and quantification, which can be made and tracked with the use of standards and an instrument operating in the SIM mode.

2.3. Experimental section and datasets

The datasets discussed and analysed in Chapter 3 and 4 are described in this section. The approach taken towards the whole experimental process was aimed at ensuring the quality of the data from the laboratory to the results. A summary description of the datasets is given in Table 2.1.

Table 2.1. Overview of the datasets.

dataset	origin	dimensions	variables
Beer	Visible	100 × 600	wavelengths
	NIR	100 × 3600	wavelengths
	NMR	100 × 61	features
Whisky	GC-MS	54 × 194	features
Globular #1	simulated	300 × 600	/
Globular #2	simulated	900 × 250	/
Circles	simulated	1000 × 250	/
t4.8k	simulated	2000 × 2	/

2.3.1. Barley's children: beer and whisky

The production processes of beer and whisky are so similar that these two beverages can be imagined not only as siblings, but also as homozygous twins: almost indistinguishable when they are young but destined to grow up very different.

Everything starts with a mother, which in our case is a source of sugars, barley. Barley is tricked into sprouting, but then it gets roasted and dried, in a process called "malting". Malting allows sugars and enzymes to become available for fermentation. Malt is then soaked with water to produce a malty-tea called "wort". This is the first time in the twin's life they start to differentiate: beer likes hops and spices, but whisky is pickier, and avoids them.

Then yeasts come in as puberty suddenly does and change everything. Fermentation yields ethanol and a myriad of other aroma and flavour compounds, and the character of the two siblings starts to get defined. However, the most drastic change happens next: whisky undergoes double (or even triple! [68]) distillation, while beer ripens for a while and quite soon it gets bottled, ready to start exploring the world. Whisky needs more time, as it prefers to stay home at least three years before even considering leaving the family nest [69].

During this very standard life-development, a lot of deviations may occur, resulting in very different beer and whisky products: even if both products have proud and long traditions, experiments for brewing new flavours has been on the rise for some years now. New beer styles, spices and malts are mixed nowadays, and even cross-overs between whisky and beer meet the market[4].

In this context, this part of the thesis work was aimed at the characterization, using different spectroscopic and chromatographic techniques, of one-hundred samples of beer and fifty-four whisky samples. Spectroscopies such as Visible, NIR and NMR were employed for studying the beer [56,70], while in the whisky case dynamic headspace GC-MS [71] was chosen. Chemometrics tools were used for exploring the data and for extracting relevant chemical features. Specifically, interval-based [38] methods were employed for the integration of the NMR peaks and for the mathematical deconvolution of the GC-MS chromatographic peaks.

2.3.2. Beer datasets

Beer has been the object of several studies, mostly focused on specific beer types or local products, aimed either at analysing its composition [51,56,72,73] or at controlling the brewing process [74,75]. To these aims very different analytical techniques have been applied: NMR spectroscopy [50–53,56,73], liquid- and gas-chromatography (LC-MS [50,76,77] and GC-MS [78,79]), vibrational (NIR and IR, [56,72,74,80]) and UV-Visible [70] spectroscopies.

[4] Elisabeth Sherman, *5 Things to Know About Beer Barrel-Aged Whiskey*, Food & Wine (2018), https://www.foodandwine.com/news/beer-barrel-aged-whiskey-jameson-caskmates (accessed: 08/01/2019)

The beer dataset consists of three data blocks obtained from different spectroscopic techniques, namely Visible, NIR and NMR, the latter as interval-resolved data. Related to the same set of samples, consumer-generated ratings and comments were obtained from the RateBeer[5] website. Text analysis techniques were applied for processing and extracting features from the comments, with the aim of linking them to the chemical information represented by the spectroscopic data blocks. Chapter 4 of this thesis is devoted to this study.

2.3.2.1. Experimental

This experimental section is devoted to describing product sampling, sample preparation and spectra acquisition, and the subsequent data preprocessing.

2.3.2.1.1. Sampling and sample preparation

One hundred beer products were purchased from local stores. Only beers rather pale in colour and with very low turbidity (i.e. no clearly visible particles suspended in the liquid), differing by brand, location of production, percentage of alcohol by volume (ABV), colour hue and beer style were selected.

A collection of 2 mL eppendorfs was prepared directly from the original commercial containers (cans or glass bottles): three eppendorfs for each beer sample were prepared and kept frozen at –20°C.

Since all the specimens were clear (i.e. no suspended particles), filtration was not required. The degassing procedure is highly recommended by literature studies [51,56,73] and it is aimed at reducing measurement interferences due to bubble formation which may cause strong interferences in both the Vis/NIR and the NMR measurements. Therefore, after thawing, only degassing by ultrasonication was performed. The initial steps of thawing and degassing were common across all the different spectroscopic techniques, and were performed as follows:

1) 10 minutes thawing in water bath at room temperature;
2) 20 minutes of ultrasonic bath in water at room temperature.

[5] https://www.ratebeer.com/ (accessed: 08/01/2019)

2.3.2.1.2. Vis/NIR data acquisition and preprocessing

The Vis and NIR spectra were acquired together using a NIRS FOSS DS2500 spectrometer. The range 400–2500 nm was recorded with 0.5 nm resolution. A cup with a round quartz window was equipped with a 0.2 mm-gap golden reflector to operate in transflectance mode. Each spectrum was obtained by taking the average over 16 scans acquired at different positions of the quartz window.

No additional steps to the preparation procedure described in the previous paragraph were necessary prior to recording the Vis/NIR spectra. The specimens were prepared in batches of twenty-five samples and then stored inside a thermally insulated styrofoam box, equipped with ice chips and a lid. This setup was made to keep the specimens in stable conditions while running the experiments.

For each sample three replicates were acquired, and the order of acquisition was randomized both with respect to samples and replicates. A control sample for each batch was also prepared under the same conditions as the other specimens: a pack of six canned beers was purchased from a local store and kept in a fridge at 4°C; right before preparing each batch, one eppendorf was filled with fresh beer and then processed together with the other samples. This allowed checking for time drifts among different batches, since they were analysed at different points in time.

Similarity among replicates was assessed by performing a PCA on the data centered with respect to the replicates (i.e. subtracting from each sample the average of its replicates). The first principal component explained 88.33% of the total variance, and the spectra far exceeding the scores confidence limits were identified as anomalous. Six outliers were identified and by inspecting the raw spectra it was found that all of them were affected by scattering effects. After removing those six outliers, each sample had at least two replicates. A new dataset consisting of one hundred spectra was finally obtained by taking the average over each set of replicates.

The Standard Normal Variate (SNV) correction was separately performed on the Vis [81] and the NIR [82] datasets. Mean centering was finally applied prior to data analysis.

2.3.2.1.3. ^{1}H-NMR data acquisition

After thawing and degassing, the specimens were kept at 5°C. Preparation of the NMR tubes was executed in batches of twelve samples, which were collected from the fridge and placed within a thermally insulated styrofoam box equipped with a ground of ice chips and closed with a lid. The newly prepared tubes were placed into the autosampler rack, which was also stored within the thermal box.

All the specimens were prepared to contain 10% D_2O, 0,02% of sodium-3-(trimethylsilyl)propionate-d4 (TSP-d4) as a chemical shift reference [49–53,56,73,83] and 20% phosphate buffer (pH = 3.55). The required volume for the NMR tubes was 600 µL, and it was obtained by mixing: 420 µL of beer specimen, 60 µL of D_2O and 120 µL of phosphate buffer in H_2O. Duarte *et al.* [83] studied the composition of ale and lager beers, reporting pH values in the 3.7–4.4 range. The phosphate buffer (pH = 3.55) was added with the aim of obtaining more homogeneous pH values, also reducing the signals' horizontal shifts due to different protonation forms of compounds such as organic acids [52,53]. All the ^{1}H-NMR profiles were acquired in random order with respect to samples and replicates on a Bruker Avance III 600 spectrometer (Bruker Biospin Gmbh, Rheinstetten, Germany) operating at Larmor frequency of 600.13 MHz for protons, equipped with a double tuned cryoprobe (TCl) set for 5 mm sample tubes and a cooled autosampler (SampleJet, at 5°C).

Spectra were acquired from all the beer specimens using TOPSPIN 2.1 (Bruker Biospin Gmbh, Rheinstetten, Germany), with the NOESYGPPR1D sequence [53,73]. Presaturation of the water signal (4.77 ppm, [49–53,56,57,73,83]) was employed, while the ethanol signals were not suppressed [52,53,73]. All the experiments were performed at 298 K with a fixed receiver gain. Each FID was collected using a total of 64 scans plus 4 dummy scans.

Zero-filling to 64k points and 0.3 Hz Lorentzian line broadening were applied to the FIDs prior to Fourier transformation. Depending on the results of the automatic baseline and phase correction (assessed by a trained NMR user) some of the spectra were manually corrected using the TOPSPIN processing tools. For all the spectra, the ppm scale was referenced to the TSP peak at 0.00 ppm. The recorded spectral window was 20.5 ppm.

2.3.2.1.4. ^{1}H-NMR features extraction: peak integration via MCR

NMR data carry different information in different spectral regions. As a consequence, NMR spectra are usually roughly split into three regions [38,83]: the aliphatic/organic acids region (0–3 ppm), the carbohydrates region (3–5 ppm) and the aromatic region (6–9 ppm). These regions mainly differ because of involved metabolites/molecules, baseline noise, and signal's average intensity [38]. For instance, as it is shown in Figure 2.7, the carbohydrates region of the beer spectra contains on average signals much more intense than the rest of the spectrum. By using an interval-based approach [38] it is possible to efficiently handle those differences and to obtain meaningful chemical features from each region.

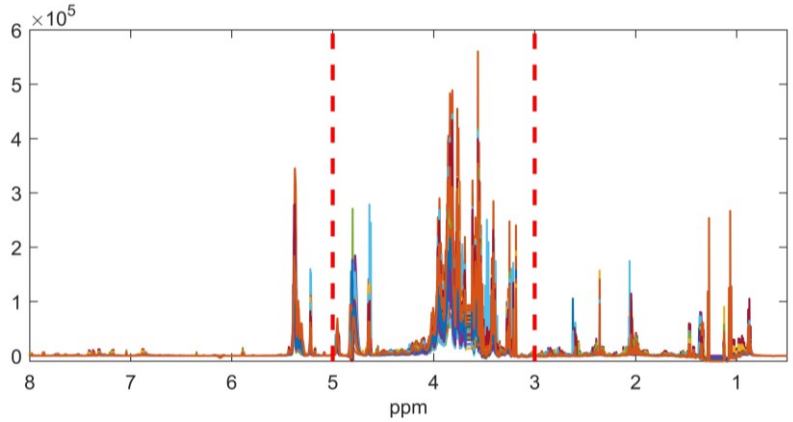


Figure 2.7. The beer NMR spectra.

The NMR spectra were imported on MATLAB and inspected to promptly spot any low-quality sample. Then, without aligning the whole spectrum, an interval-by-interval processing procedure was performed:

1) alignment by means of *i*coshift [84,85];
2) peak deconvolution and integration by means of MCR;
3) peak assignment (comparisons with literature and digital libraries).

Working on single intervals from their alignment to the assignment of the resolved signals allows to focus and to obtain meaningful chemical features, together with a deep insight in the nature of the data. One MCR model was built for each custom-defined interval, using non-negativity constraint (more specifically, the fast non-

negativity-constrained least squares algorithm [86]) on both profiles and concentrations. The initial estimates were selected using the SIMPLISMA [87] algorithm, even though in some special cases a set of simulated spectra were used. The pure profiles matrix was normalized to the Euclidean norm and the max number of iterations was set to 500.

Those models for which converge was not achieved, but clearly resolved one or more plausible components, were inspected in greater detail. It was generally found that the reason why the convergence criterion was not met lied in one or more baseline-like or just meaningless components which were difficult to model: since the profiles of the plausible components were basically constant when the 500^{th} iteration was reached, it was decided to keep them.

For each model, the components representing chemical information were retained, whereas components describing baseline variations or noise were excluded. Once all the intervals had been processed a new "features dataset" was composed using sixty-one resolved components: the relative concentrations of each component provided by MCR were merged to create the new dataset, including those for which it was not found a chemical label either in the literature or in a reference library: in those cases, the unassigned resolved signals showed clear peak-like characteristics, and were therefore included in the new dataset. Twenty-one of these features were tentatively assigned with the procedures described in the next Section.

2.3.2.1.5. ^{1}H-NMR peak assignment

One of the main goals of features extraction is to make the data analysis simpler, by reducing the number of variables and removing at the same time meaningless variables and noise. Feature extraction by peak deconvolution allows taking one step further towards clarity and interpretability: each extracted feature is ideally a signal, or a group of signals directly related to a specific molecule. Assigning a chemical name to these new variables makes the data analysis and the consequent model interpretation much easier.

The assignments given to the resolved peaks were made starting from the signals' shape and position. By comparison with assignments from literature it was possible to give a name to most of the resolved profiles. Another source of information was the reference library from the Chenomx NMR Suite[6], a software dedicated to the interpretation of NMR spectra.

For the purposes of our research, a detailed and extensive assignment was not required. However, an interesting guide for thorough identification of metabolites in NMR-based metabolomics was published by Dona *et al.* [88] in 2016.

2.3.3. Whisky dataset

Whisky, or whisk*e*y outside of Scotland, is a very popular distilled drink obtained from fermented grain mash. Whisky has been studied for a long time and a hint about that can be found in *The Lancet* first (1905, [89]) and in *Nature* right after (1906, [90]) where the question "What is whisk(e)y?" was asked. Funny enough, this dilemma came *after* another article from *The Lancet* titled "The chemistry of whisky-and-soda" (1903, [91]), suggesting that bartenders were probably some steps ahead of scientist at that time.

The chemical composition of whisky has been studied at different levels, from general [92–94] to quite specific chemical classes of compounds [95–98]. Aroma and flavour profiling is another important point of view in whisky analysis [99–105], which is directly linked to the sensory analysis approach [106–108]. Moreover, whisky has always been very popular, and many whisky products are considered luxury goods. For these reasons, whisky is particularly at risk of potential adulteration [109]. Authenticity is therefore one of the main issues regarding whisky analysis [110–117].

This dataset originates from a project that involves both the University of Copenhagen and the University of Modena and Reggio Emilia and has not been published before. The overall aim of the project was to explore the "whisky space" to gain information about similarities and differences in aroma and flavour of several products, from a chemical point of view. The idea is to extract the volatile compounds under conditions

[6] https://www.chenomx.com/ (accessed: 08/01/2019)

resembling the human oral cavity and separate the extracted mixture through Gas Chromatography coupled with Mass Spectroscopy (GC-MS).

2.3.3.1. Experimental

In this section, an overview of the variety among the selected whisky products is given, followed by the description of the procedures and conditions used for collection the volatile compounds and analysing them via GC-MS spectrometry.

2.3.3.1.1. Samples collection

A set of fifty-four whisky small bottles ("miniatures", size from 2 to 5 mL) was purchased from online retailers. The sampling process was aimed at covering the most important known whisky features. The selected products differ by type, distillery, country of origin, mash bill, cask type, age, alcohol content, peat-drying, colour and price. More detailed information is reported in Table 2.2.

The dataset's composition is naturally dominated by Scottish whisky, or *Scotch*, since Scotland historically represents the origins of modern whiskies. However, Scotch whiskies are traditionally associated to their region of provenience, with which peculiar features are associated. Figure 2.8 depicts the "whisky regions" of Scotland.

Figure 2.8. Whisky regions of Scotland[7], and a portion of Ireland on the bottom left corner.

[7] *Scotch regions* by Briangotts - Own work. Licensed under the Creative Commons Attribution-Share Alike 3.0 Unported: https://commons.wikimedia.org/wiki/File:Scotch_regions.svg (accessed: 08/01/2019)

Table 2.2. The whisky dataset, features, counts and related studies from literature.

feature		related studies
type	single malt (33) blended (16) Tennessee whisky (3) bourbon (2)	[118]
distillery (30 producers)	Amrut, Antiquary, Ardbeg, Auchentoshan, Ballantine's, Ben Nevis, Douglas Laing & Co., BenRiach, Benromach, Buffalo Trace, Chivas, Fary Lochan, Glenfarclas, Glenfiddich, Glenlivet, GlenDronach, Jameson, Jack Daniel's, Jura, Johnnie Walker, Kilbeggan, Loch Lomond, Nikka Taketsuru, Penderyn, Tomatin, Tullamore D.E.W., The Wild Geese, Whyte & Mackay, Evan Williams, Thylandia	[119]
country of origin	Ireland (6), USA (5), Japan (3), India (2), Denmark (2), Wales (1), Scotland (35) - Speyside (13) - Highland (7) - Lowland (4) - Islands (3) - Islay (2) - Glasgow (1)	[120,121]
mash bill (or **malt type)**	barley (38) rye (3) corn (1) mixed (11)	
cask type	ex-bourbon ex-sherry new charred American oak casks for special purposes combinations of casks	[122,123]
age	range 4–25 years 21 samples with non-Age Statement (NAS[8])	[120,123–126]
alcohol content	range 40–50% one "whisky liquor" with 22%;	
peat-drying	peated (13) non-peated (41)	[127,128]
colour	natural colour (20) added colouring (31)	
price level[9]	low (6) medium-low (17) medium (20) medium-high (4) high (1) no rating available for 5 samples	

[8] https://scotchaddict.com/nas-no-age-statement-whisky.html (accessed: 08/01/2019)

[9] https://distiller.com/ (accessed: 08/01/2019)

2.3.3.1.2. Sample randomization and analysis schemes

All samples were analysed in triplicates and each set of replicates was prepared and analysed separately. The randomization scheme described in Figure 2.9 was applied to each set of replicates. Due to the limited capacity of the autosampler, it was decided to divide the 54 samples of each set of replicates in three groups (GR1-2-3), also to avoid overusing the instrument.

Three GC-MS runs of 18 samples were scheduled (GR1-2-3), and for each run a pair of identical control samples (CTRL) was prepared from a cheap whisky: to check for instrumental drifts the control samples were placed at the beginning and at the end of the processing sequence.

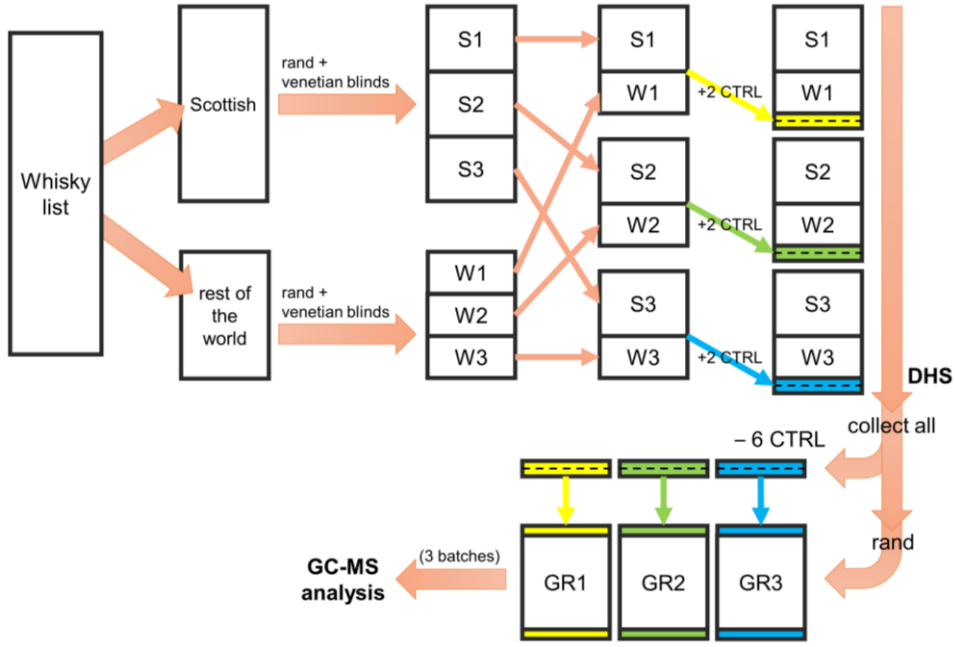


Figure 2.9. Randomization scheme for the whisky analysis. Due to the large majority of Scottish products, it was decided to randomize (rand) them separately. To add one additional mixing step and obtain the three groups for DHS, a venetian blinds scheme was applied to the randomized Scottish list and the randomized rest of the world (W) list. After DHS was carried out, all the Tenax-TA traps were collected, randomized and then divided into three new groups (GR1-2-3), before undergoing the GC-MS analysis.

2.3.3.1.3. Dynamic Headspace Sampling (DHS)

Volatile compounds were collected using a dynamic headspace sampling (DHS) system. For each sample, 5 mL of whisky were placed in a 100 mL gas washing flask equipped with a purge head. A trap containing Tenax-TA (250 mg, mesh size 60/80; Buchem BV, Apeldoorn, The Netherlands) was attached to the purge head. The flask containing the sample was immersed in a laboratory water bath and held at 37 °C (temperature resembling the physical conditions of the human mouth). To collect the volatiles, the sample was purged with nitrogen (100 mL/min) for 20 minutes and under magnetic stirring (200 rpm). The traps were then dry purged with nitrogen (100 mL/min) for 15 min to remove excess water trapped during the sampling procedure. Finally, the Tenax-TA traps were sealed and kept at 5 ℃ before GC-MS analysis.

2.3.3.1.4. Gas Chromatography-Mass Spectrometry (GC-MS)

The collected volatiles were thermally desorbed from the Tenax-TA traps using an automatic thermal desorption unit (ATD 400; Perkin Elmer, Waltham, MA). Primary desorption was carried out at 250 ℃ (15 min) to a cold trap (30 mg Tenax TA, 5 ℃), with a hydrogen flow of 50 mL/min. Volatiles were desorbed from the cold trap by heating to 300 ℃ for 4 min (secondary desorption), using a split ratio of 1:10. The volatiles were then transferred through a heated transfer-line (225 ℃) to a gas chromatograph-mass spectrometer (GC-MS, 7890A GC-system interfaced with a 5975C VL MSD with triple-axis detector from Agilent Technologies, Palo Alto, CA) equipped with a J&W Scientific DB-Wax column of 30 m length and 0.25 mm internal diameter, with 0.50 μm film thickness. The column pressure was held constant at 2.3 psi, using helium as carrier gas (1 mL/min). The column temperature was kept at 30℃ for 10 min, increased at 8 ℃/min to 240℃, and kept isothermal for 5 min. The mass selective detector was in electron impact mode (70 eV). Mass spectra were obtained at a mass/charge (m/z) range between 15 and 300. GC-MS data processing was carried out under Matlab environment and is described in Section 2.3.3.2.

2.3.3.2. Features extraction and data preprocessing

Chromatograms originating from GC-MS instrument are three-way data. This is because for each datapoint along the retention time mode, there is one corresponding mass spectrum. These data contain therefore information on both where in time a compound is (its retention time) and what its fingerprint is (the mass spectrum).

As explained in Section 2.1.1.3, this abundance of information can be unravelled by means of PARAFAC modelling. Pure resolved components are obtained, so the complex starting data can be compressed into fewer variables which represent the relative concentrations in the samples of each resolved component. By organizing these extracted concentrations (the features) in a 2D matrix and by matching the mass spectra to library and literature references, this new features data matrix will have columns directly related to chemical compounds. Such a data matrix is easy to process and interpret.

This section is devoted to explaining in more detail the features extraction steps and the peak assignment procedure taken for the analysis of the Whisky dataset.

2.3.3.2.1. Mathematical chromatography: features extraction via PARAFAC2

Features extraction was performed on the GC-MS data using the PARADISe software [67], a PARAFAC2-based deconvolution and identification system. The purpose of this software is to allow extracting features from the very rich GC-MS data, by providing an easy graphical user interface through which the user can inspect the chromatograms and define intervals small at will, ideally containing as few as possible peaks. Each interval is then processed individually by building many PARAFAC2 models of increasing dimensionality. All the models can be later inspected through another interface, through which it is possible to select the optimal model for each interval, but also the nicest extracted chemical components. Once the components have been selected it is possible to compare their mass spectra with digital libraries, using the

built-in function that communicates with the NIST MS software[10] and produces a report of the tentative assignments.

After aligning the whole dataset using the Dynamic Time Warping (DTW) and Correlation Optimized Warping (COW) algorithms [129,130], the data were processed with PARADISe by separately modelling small user-defined intervals. For each interval, eight PARAFAC2 models of increasing dimensionality (i.e. from one to eight components) were built, and the better resolved chemical components were selected. A total of 194 resolved components were obtained, and the assignments provided by the NIST software were carefully checked and compared with literature sources (Section 2.3.3.2.2).

2.3.3.2.2. Peak assignment

Peak assignment was done using the built-in "Make report" function of PARADISe that compares the mass spectra of the resolved components with digital libraries, via the NIST software. More specifically, the digital libraries[11] used in this work were the Main EI MS Database (electron impact mass spectroscopy database, "mainlib", 212.961 spectra) and the Replicate spectra Database ("replib", 30.932 spectra).

The output report consists of an Excel file with two worksheets, one containing the areas of all the resolved chemical components for each sample, and the other one containing the first five matches to the compounds libraries for each resolved component. Matching probabilities are also provided, so that if an assignment looks suspicious, one of the other options may be used.

All the resolved components were carefully inspected and compared with literature sources, both for validating the labels provided by the software and for defining chemical classes that could be subsequently used for interpreting the results from data analysis.

[10] https://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:start (accessed: 30/01/2019)

[11] https://www.nist.gov/sites/default/files/documents/srd/NIST1a11Ver2-0Man.pdf (accessed: 30/01/2019)

2.3.3.2.3. Data preprocessing

Once the resolved areas were obtained, a PCA model was built on the autoscaled data, for spotting weird samples and/or variables. All replicates and control samples were included and, by taking advantage of their meta-information, possible time drifts related to the headspace sampling batch, the GC-MS analysis group or to the replicate were checked. No time drifts were found. Instead, four faulty samples were identified: those samples have very low intensities, which became very clear when they were plotted against their correspondent replicates which, on the contrary, had much higher and comparable intensities. A possible explanation for this deficiency could be a temporary malfunctioning of the nitrogen pump during the purging phase of the headspace sampling procedure.

After removing the faulty samples, the data were row-normalized to unit area. Finally, the average over all the replicates was taken, and a dataset of 54 unique samples was obtained, and used for further data analysis. The data were always autoscaled prior to modelling, except for the coclustering (Section 2.1.1.4.) analysis, where in order to use the non-negativity constraint, the data were scaled to unit variance bay not mean centered. This preprocessing is also called *unit variance scaling*, as described in [131].

2.3.4. Simulated datasets

The simulated datasets were chosen and designed for testing different sets of parameters of the Fused Adjacency Matrix approach. Variations to these parameters are described in more detail in Section 3.6.1.

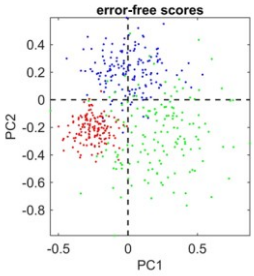
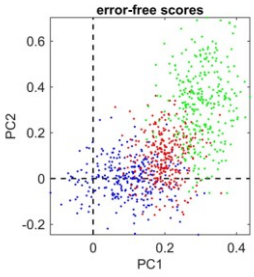
Four simulated datasets were used: two of them consist of globular clusters with different positions and overlaps (Figures 2.10 and 2.11); the *circles* dataset consists of a non-linear situation with a central dense cluster surrounded by a larger, more dispersed set of samples and a third globular-like cluster in a non-central position (Figure 2.12); in the fourth dataset six distinct clusters are present, five of which have regular-polygonal shapes while the last one is more “curvy” (Figure 2.14).

2.3.4.1. Globular clusters datasets (#1 and #2)

The globular datasets represent the simplest situations for testing the Fused Adjacency Matrix approach. Three globular clusters for each dataset were generated, and their position and overlap were designed to get a rather confused situation, in which, without the class information it would be moderately difficult to distinguish the clusters.

First, a simple error-free set of scores was generated in a two-dimensional space and normally distributed random noise (homoscedastic noise) was added to it. A set of loadings was generated by combining different gaussian curves, to obtain a NIR-like look. Homoscedastic and heteroscedastic noise was added to it. Finally, a multivariate dataset was obtained by multiplying the scores and the transposed loadings.

Table 2.3. Characteristics of the globular datasets.

	clusters			noise	
	cluster 1	**cluster 2**	**cluster 3**	**scores**	**loadings**
globular dataset #1 (450 × 600)					
	n = 150 $\sigma_{var1,2}$ = 0.15	n = 150 $\sigma_{var1,2}$ = 0.10	n = 150 $\sigma_{var1,2}$ = 0.25	σ = 0.3	m = 600 σ_{Homo} = 0.05 σ_{Het} =0.05
globular dataset #2 (900 × 250)					
	n = 300 σ_{var1} = 0.08 σ_{var2} = 0.08	n = 300 σ_{var1} = 0.05 σ_{var2} = 0.1	n = 300 σ_{var1} = 0.05 σ_{var2} = 0.15	σ = 0.15	m = 250 σ_{Homo} = 0.02 σ_{Het} = 0.015

Three clusters for each dataset were built following an approach similar to the one described by Wentzell *et al.* [132]. The objects belonging to each cluster were randomly distributed around the desired centres, according to a symmetric bivariate normal distribution. Both datasets have three clusters, but with different position and distribution properties. Figures 2.10 and 2.11 report the visual descriptions of the two datasets.

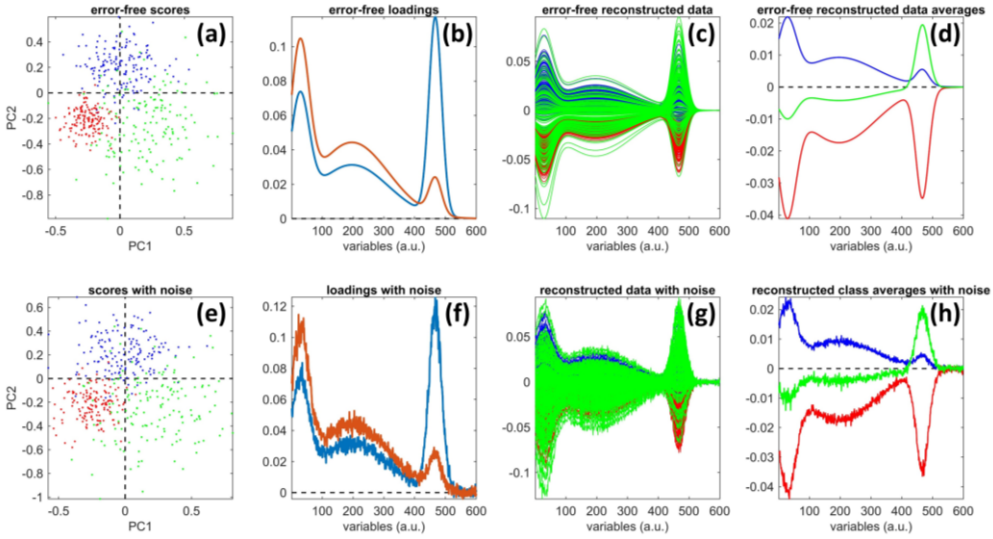


Figure 2.10. The globular dataset #1: error-free scores (a) and with noise (e); error-free loadings (b) and with noise (f); error-free reconstructed data (c) and with noise (g); class averages from reconstructed data, error-free (d) and with noise (h).

The globular dataset #1 consists of 450 samples, equally divided into three clusters of 150 objects each. The clusters' centres in the starting two-dimensional dataset, as depicted in Figure 2.10, are located at the vertices of an equilateral triangle, centered at the origin and with sides of unit length [132]. As reported in Table 2.3, the three clusters have different standard deviations and different noise levels were added to the scores and loadings.

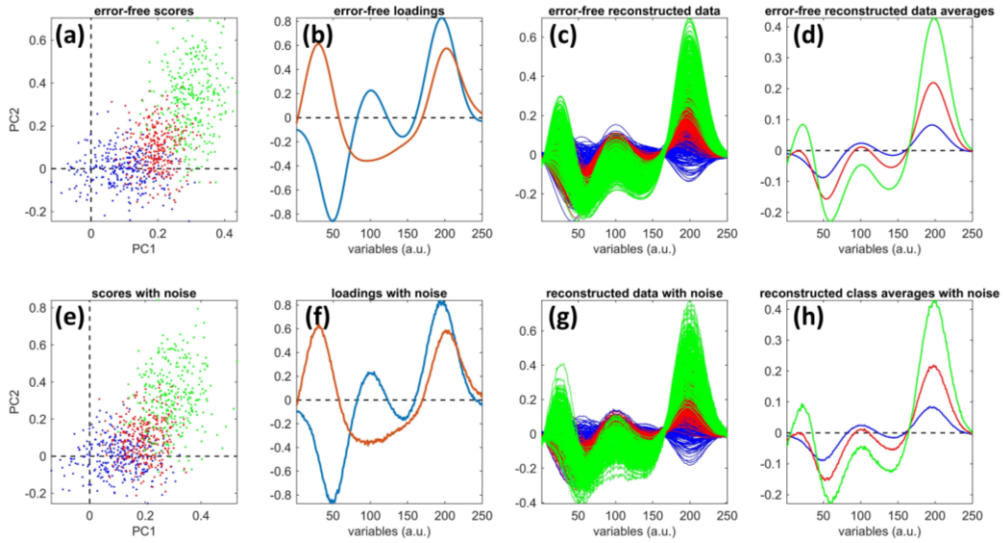


Figure 2.11. The globular dataset #2: error-free scores (a) and with noise (e); error-free loadings (b) and with noise (f); error-free reconstructed data (c) and with noise (g); class averages from reconstructed data, error-free (d) and with noise (h).

The globular dataset #2 consists of 900 samples, equally divided into three clusters of 300 objects each. The clusters' centres in the starting two-dimensional dataset, as depicted in Figure 2.11, are located so that it is difficult to distinguish the groups along the first simulated PC. As reported in Table 2.3, the three clusters have different standard deviations and different noise levels were added to the scores and loadings.

2.3.4.2. Circles dataset

In the *circles* dataset a peculiar non-linear situation is simulated. A central dense cluster (cluster 2, in red in Figure 2.12) is surrounded by a larger, more dispersed set of samples (cluster 1, in blue in Figure 2.12). A third cluster (cluster 3, in green in Figure 2.12) is partially surrounded by the cluster in blue, but even if its shape resembles the globular clusters (Section 2.3.6.1), it was not generated using a normal bivariate distribution.

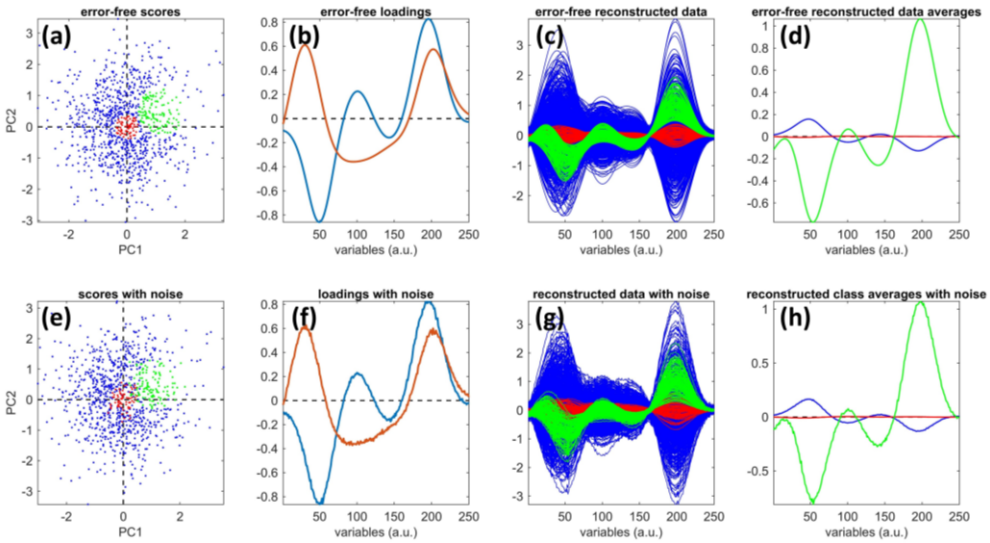


Figure 2.12. The *circles* dataset: error-free scores (a) and with noise (e); error-free loadings (b) and with noise (f); error-free reconstructed data (c) and with noise (g); class averages from reconstructed data, error-free (d) and with noise (h).

The *circles* dataset has 1000 samples divided into three clusters: cluster 1 with 799 samples, cluster 2 with 74 samples and cluster 3 with 127 samples. The same set of loadings as the globular dataset #2 was used and a reconstructed dataset with dimensions 1000 × 250 was used.

2.3.4.3. t4.8k dataset

This dataset was chosen for testing how the Fused Adjacency Matrix approach would perform with clusters with very well-defined geometric shapes (Figure 2.13). Density-based methods like OPTICS and DBSCAN usually are the optimal choice for this type of situation and can be taken as a benchmark.

The dataset was used by Karypis *et al.* [133] and can be downloaded from the "Clustering basic benchmark" [134] webpage of the School of Computing of the University of Eastern Finland: http://cs.joensuu.fi/sipu/datasets/.

The original size of the dataset was 8000 × 2, but due to computational and time constraints it was decided to reduce it by randomly selecting 2000 samples. The dimensions of the dataset used for testing was therefore 2000 × 2.

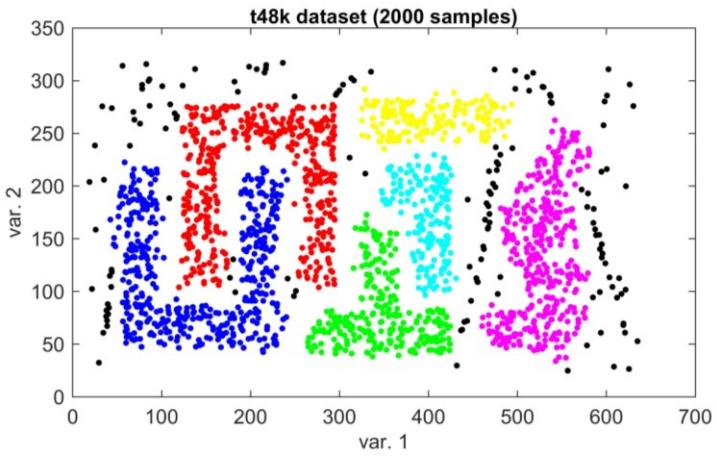


Figure 2.13. The t4.8k dataset. The six clusters are depicted with different colours and the samples not belonging to any cluster are depicted in black.

2.4. Software

All the data analyses described and reported in this thesis were carried out under MATLAB environment (2016a/2017b, Mathworks, MA, USA).

PCA analysis was performed using the PLS Toolbox 8.6 (Eigenvector Research Inc. WA, USA).

NMR spectral alignment was performed using *i*coshift [84,85], and it can be downloaded from:

http://www.models.life.ku.dk/icoshift (accessed: 08/01/2019)

NMR features extraction was performed by means of the MCR-ALS GUI by Joaquim Jaumot, Anna de Juan and Romà Tauler [135]. The MATLAB package can be found at:

https://mcrals.wordpress.com/ (accessed: 08/01/2019)

GC-MS alignment was operated using the Dynamic Time Warping (DTW) and Correlation Optimized Warping (COW) algorithms [129,130] created by Giorgio Tomasi, Thomas Skov and Frans van den Berg; it can be downloaded from:

http://www.models.life.ku.dk/DTW_COW (accessed: 08/01/2019)

GC-MS features extraction was performed using the PARADISe software [67], which can be downloaded from:

http://www.models.life.ku.dk/paradise (accessed: 08/01/2019)

The **OPTICS algorithm** was written by Michal Daszykowski and it can be found at:

http://chemometria.us.edu.pl/download/OPTICS.M (accessed: 08/01/2019)

The MATLAB function for performing **co-clustering with non-negative matrix factorization** can be downloaded from:

http://www.models.life.ku.dk/cocluster (accessed: 08/01/2019)

Kohonen's Self-Organizing Maps were computed using a homemade routine by Federico Marini (Università La Sapienza, Roma).

The **Fused Adjacency Matrices** were computed using in-house written MATLAB routines, which can be downloaded at:

http://www.models.life.ku.dk/algorithms (accessed: 08/01/2019)

The **simulated globular (#1 and #2) and circles datasets** were generated using in-house MATLAB routines, based on the simulation approach used by Wentzell *et al.* [132].

The **t4.8k dataset** can be downloaded from the "Clustering basic benchmark" webpage of the School of Computing of the University of Eastern Finland:

http://cs.joensuu.fi/sipu/datasets/ (accessed: 08/01/2019)

References | Chapter 2

[1] S. Wold, Chemometrics; what do we mean with it, and what do we want from it?, Chemom. Intell. Lab. Syst. 30 (1995) 109–115. doi:10.1016/0169-7439(95)00042-9.

[2] J.W. Tukey, Exploratory data analysis, (1977). https://www.popline.org/node/499313 (accessed January 8, 2019).

[3] M. Li Vigni, C. Durante, M. Cocchi, Exploratory Data Analysis, in: Data Handl. Sci. Technol., Elsevier, 2013: pp. 55–126. doi:10.1016/B978-0-444-59528-7.00003-X.

[4] S.C. Rutan, A. de Juan, R. Tauler, Introduction to Multivariate Curve Resolution, in: Compr. Chemom., Elsevier, 2009: pp. 249–259. doi:10.1016/B978-044452701-1.00046-6.

[5] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods. 6 (2014) 2812–2831. doi:10.1039/C3AY41907J.

[6] P.D. Wentzell, Other Topics in Soft-Modeling: Maximum Likelihood-Based Soft-Modeling Methods, in: Compr. Chemom., Elsevier, 2009: pp. 507–558. doi:10.1016/B978-044452701-1.00057-0.

[7] J.H. Friedman, J.W. Tukey, A Projection Pursuit Algorithm for Exploratory Data Analysis, IEEE Trans. Comput. C-23 (1974) 881–890. doi:10.1109/T-C.1974.224051.

[8] J.F.Q. Pereira, C.S. Silva, A. Braz, M.F. Pimentel, R.S. Honorato, C. Pasquini, P.D. Wentzell, Projection pursuit and PCA associated with near and middle infrared hyperspectral images to investigate forensic cases of fraudulent documents, Microchem. J. 130 (2017) 412–419. doi:10.1016/j.microc.2016.10.024.

[9] F. Westad, M. Kermit, Independent Component Analysis, in: Compr. Chemom., Elsevier, 2009: pp. 227–248. doi:10.1016/B978-044452701-1.00045-4.

[10] A. Hyvärinen, Independent component analysis: recent advances., Philos. Trans. A. Math. Phys. Eng. Sci. 371 (2013) 20110534. doi:10.1098/rsta.2011.0534.

[11] M.C. Hout, M.H. Papesh, S.D. Goldinger, Multidimensional scaling, Wiley Interdiscip. Rev. Cogn. Sci. 4 (2013) 93–103. doi:10.1002/wcs.1203.

[12] F. Marini, R. Bucci, A.L. Magrì, A.D. Magrì, Artificial neural networks in chemometrics: History, examples and perspectives, Microchem. J. 88 (2008) 178–185. doi:10.1016/j.microc.2007.11.008.

[13] T. Kohonen, Essentials of the self-organizing map, Neural Networks. 37 (2013) 52–65. doi:10.1016/J.NEUNET.2012.09.018.

[14] T. Kohonen, Self-organizing maps, Springer, 2001. doi:10.1007/978-3-642-56927-2.

[15] K. Varmuza, P. Filzmoser, Introduction to multivariate statistical analysis in chemometrics, CRC Press, 2009.

[16] I. Lee, J. Yang, Common Clustering Algorithms, in: Compr. Chemom., Elsevier, 2009: pp. 577–618. doi:10.1016/B978-044452701-1.00064-8.

[17] I.T. Jolliffe, Principal component analysis, 2nd ed., Springer, 2002.

[18] A. de Juan, R. Tauler, Multivariate Curve Resolution (MCR) from 2000: Progress in Concepts and Applications, Crit. Rev. Anal. Chem. 36 (2006) 163–176. doi:10.1080/10408340600970005.

[19] R. Bro, PARAFAC. Tutorial and applications, Chemom. Intell. Lab. Syst. 38 (1997) 149–171. doi:10.1016/S0169-7439(97)00032-4.

[20] A.K. Smilde, R. Bro, P. Geladi, Multi-way analysis with applications in the chemical sciences, J. Wiley, 2004. https://www.wiley.com/en-us/Multi+way+Analysis%3A+Applications+in+the+Chemical+Sciences-p-9780471986911 (accessed January 11, 2019).

[21] H.A.L. Kiers, J.M.F. ten Berge, R. Bro, PARAFAC2—Part I. A direct fitting algorithm for the PARAFAC2 model, J. Chemom. 13 (1999) 275–294. doi:10.1002/(SICI)1099-128X(199905/08)13:3/4<275::AID-CEM543>3.0.CO;2-B.

[22] R. Bro, C.A. Andersson, H.A.L. Kiers, PARAFAC2—Part II. Modeling chromatographic data with retention time shifts, J. Chemom. 13 (1999) 295–309. doi:10.1002/(SICI)1099-128X(199905/08)13:3/4<295::AID-CEM547>3.0.CO;2-Y.

[23] J.M. Amigo, T. Skov, R. Bro, J. Coello, S. Maspoch, Solving GC-MS problems with PARAFAC2, TrAC Trends Anal. Chem. 27 (2008) 714–725. doi:10.1016/J.TRAC.2008.05.011.

[24] J.A. Hartigan, Direct Clustering of a Data Matrix, J. Am. Stat. Assoc. 67 (1972) 123–129. doi:10.1080/01621459.1972.10481214.

[25] D. Hanisch, A. Zien, R. Zimmer, T. Lengauer, Co-clustering of biological networks and gene expression data, Bioinformatics. 18 (2002) S145–S154. doi:10.1093/bioinformatics/18.suppl_1.S145.

[26] S.C. Madeira, A.L. Oliveira, Biclustering algorithms for biological data analysis: a survey, IEEE/ACM Trans. Comput. Biol. Bioinforma. 1 (2004) 24–45. doi:10.1109/TCBB.2004.2.

[27] R. Bro, E.E. Papalexakis, E. Acar, N.D. Sidiropoulos, Coclustering-a useful tool for chemometrics, J. Chemom. (2012). doi:10.1002/cem.1424.

[28] E.E. Papalexakis, N.D. Sidiropoulos, M.N. Garofalakis, Reviewer Profiling Using Sparse Matrix Regression, in: 2010 IEEE Int. Conf. Data Min. Work., IEEE, 2010: pp. 1214–1219. doi:10.1109/ICDMW.2010.87.

[29] M. Ankerst, M.M. Breunig, H.-P. Kriegel, J. Sander, OPTICS: Ordering Points To Identify the Clustering Structure, in: Proc. 1999 ACM SIGMOD Int. Conf. Manag. Data - SIGMOD '99, ACM Press, New York, New York, USA, 1999: pp. 49–60. doi:10.1145/304182.304187.

[30] M. Daszykowski, B. Walczak, D.L. Massart, Looking for Natural Patterns in Analytical Data. 2. Tracing Local Density with OPTICS, J. Chem. Inf. Model. 42 (2002) 500–507. doi:10.1021/CI010384S.

[31] M. Daszykowski, B. Walczak, Density-Based Clustering Methods, in: Compr. Chemom., 2009: pp. 635–654. doi:10.1016/B978-044452701-1.00067-3.

[32] L. Wilkinson, M. Friendly, The History of the Cluster Heat Map, Am. Stat. 63 (2009) 179–184. doi:10.1198/tas.2009.0033.

[33] R.M. Podowski, B. Miller, W.W. Wasserman, Visualization of complementary systems biology data with parallel heatmaps, IBM J. Res. Dev. 50 (2006) 575–581. doi:10.1147/rd.506.0575.

[34] D. Intelmann, G. Haseleu, A. Dunkel, A. Lagemann, A. Stephan, T. Hofmann, Comprehensive Sensomics Analysis of Hop-Derived Bitter Compounds during Storage of Beer, J. Agric. Food Chem. 59 (2011) 1939–1953. doi:10.1021/jf104392y.

[35] I. Stanimirova, C. Boucon, B. Walczak, Relating gas chromatographic profiles to sensory measurements describing the end products of the Maillard reaction, Talanta. 83 (2011) 1239–1246. doi:10.1016/J.TALANTA.2010.09.018.

[36] F. Marini, A.L. Magrì, F. Balestrieri, F. Fabretti, D. Marini, Supervised pattern recognition applied to the discrimination of the floral origin of six types of Italian honey samples, in: Anal. Chim. Acta, 2004. doi:10.1016/j.aca.2004.01.013.

[37] P. Kovesi, Good Colour Maps: How to Design Them, (2015). doi:citeulike-article-id:14289773.

[38] F. Savorani, M.A. Rasmussen, Å. Rinnan, S.B. Engelsen, Interval-Based Chemometric Methods in NMR Foodomics, in: Cyril Ruckebusch (Ed.), Data Handl. Sci. Technol., Elsevier, 2013: pp. 449–486. doi:10.1016/B978-0-444-59528-7.00012-0.

[39] M. Bevilacqua, R. Bro, F. Marini, Å. Rinnan, M.A. Rasmussen, T. Skov, Recent chemometrics advances for foodomics, TrAC - Trends Anal. Chem. (2017). doi:10.1016/j.trac.2017.08.011.

[40] M. Silvestri, A. Elia, D. Bertelli, E. Salvatore, C. Durante, M. Li Vigni, A. Marchetti, M. Cocchi, A mid level data fusion strategy for the Varietal Classification of Lambrusco PDO wines, Chemom. Intell. Lab. Syst. 137 (2014) 181–189. doi:10.1016/j.chemolab.2014.06.012.

[41] E. Borràs, J. Ferré, R. Boqué, M. Mestres, L. Aceña, O. Busto, Data fusion methodologies for food and beverage authentication and quality assessment – A review, Anal. Chim. Acta. 891 (2015) 1–14. doi:10.1016/j.aca.2015.04.042.

[42] A. Biancolillo, R. Bucci, A.L. Magrì, A.D. Magrì, F. Marini, Data-fusion for multiplatform characterization of an italian craft beer aimed at its authentication, Anal. Chim. Acta. 820 (2014) 23–31. doi:10.1016/j.aca.2014.02.024.

[43] J.M. Andrade, M.P. Gómez-Carracedo, W. Krzanowski, M. Kubista, Procrustes rotation in analytical chemistry, a tutorial, Chemom. Intell. Lab. Syst. 72 (2004) 123–132. doi:10.1016/J.CHEMOLAB.2004.01.007.

[44] P.D. Wentzell, S. Hou, C.S. Silva, C.C. Wicks, M.F. Pimentel, Procrustes rotation as a diagnostic tool for projection pursuit analysis, Anal. Chim. Acta. 877 (2015) 51–63. doi:10.1016/j.aca.2015.03.006.

[45] A. Tomassini, G. Capuani, M. Delfini, A. Miccheli, NMR-Based Metabolomics in Food Quality Control, Data Handl. Sci. Technol. (2013). doi:10.1016/B978-0-444-59528-7.00011-9.

[46] Alessandra Sussulini, ed., Metabolomics: From Fundamentals to Clinical Applications, Springer, 2017. doi:10.1007/978-3-319-47656-8.

[47] A. Trimigno, F.C. Marincola, N. Dellarosa, G. Picone, L. Laghi, Definition of food quality by NMR-based foodomics, Curr. Opin. Food Sci. 4 (2015) 99–104. doi:10.1016/j.cofs.2015.06.008.

[48] C.A. Hughey, C.M. McMinn, J. Phung, Beeromics: from quality control to identification of differentially expressed compounds in beer, Metabolomics. 12 (2016) 11. doi:10.1007/s11306-015-0885-5.

[49] A.M. Gil, I.F. Duarte, M. Godejohann, U. Braumann, M. Maraschin, M. Spraul, Characterization of the aromatic composition of some liquid foods by nuclear magnetic resonance spectrometry and liquid chromatography with nuclear magnetic resonance and mass spectrometric detection, Anal. Chim. Acta. 488 (2003) 35–51. doi:10.1016/S0003-2670(03)00579-8.

[50] I.F. Duarte, M. Godejohann, U. Braumann, M. Spraul, A.M. Gil, Application of NMR Spectroscopy and LC-NMR/MS to the Identification of Carbohydrates in Beer, J. Agric. Food Chem. 51 (2003) 4847–4852. doi:10.1021/JF030097J.

[51] C. Almeida, I.F. Duarte, A. Barros, J. Rodrigues, M. Spraul, A.M. Gil, Composition of Beer by 1H NMR Spectroscopy: Effects of Brewing Site and Date of Production, J. Agric. Food Chem. 54 (2006) 700–706. doi:10.1021/JF0526947.

[52] L.I. Nord, P. Vaag, J.Ø. Duus, Quantification of Organic and Amino Acids in Beer by 1H NMR Spectroscopy, Anal. Chem. 76 (2004) 4790–4798. doi:10.1021/ac0496852.

[53] J.A. Rodrigues, G.L. Erny, A.S. Barros, V.I. Esteves, T. Brandão, A.A. Ferreira, E. Cabrita, A.M. Gil, Quantification of organic acids in beer by nuclear magnetic resonance (NMR)-based methods, Anal. Chim. Acta. 674 (2010) 166–175. doi:10.1016/j.aca.2010.06.029.

[54] L. Mannina, F. Marini, R. Antiochia, S. Cesa, A. Magrì, D. Capitani, A.P. Sobolev, Tracing the origin of beer samples by NMR and chemometrics: Trappist beers as a case study, Electrophoresis. 37 (2016) 2710–2719. doi:10.1002/elps.201600082.

[55] T. Kuballa, T.S. Brunner, T. Thongpanchang, S.G. Walch, D.W. Lachenmeier, Application of NMR for authentication of honey, beer and spices, Curr. Opin. Food Sci. 19 (2018) 57–62. doi:10.1016/J.COFS.2018.01.007.

[56] I.F. Duarte, A. Barros, C. Almeida, M. Spraul, A.M. Gil, Multivariate Analysis of NMR and FTIR Data as a Potential Tool for the Quality Control of Beer, J. Agric. Food Chem. 52 (2004) 1031–1038. doi:10.1021/jf030659z.

[57] D.W. Lachenmeier, W. Frank, E. Humpfer, H. Schäfer, S. Keller, M. Mörtter, M. Spraul, Quality control of beer using high-resolution nuclear magnetic resonance spectroscopy and multivariate analysis, Eur. Food Res. Technol. 220 (2005) 215–221. doi:10.1007/s00217-004-1070-7.

[58] L. Laghi, G. Picone, F. Capozzi, Nuclear magnetic resonance for foodomics beyond food analysis, TrAC Trends Anal. Chem. 59 (2014) 93–102. doi:10.1016/J.TRAC.2014.04.009.

[59] D. Benaki, E. Mikros, NMR-Based Metabolic Profiling Procedures for Biofluids and Cell and Tissue Extracts, in: Humana Press, New York, NY, 2018: pp. 117–131. doi:10.1007/978-1-4939-7643-0_8.

[60] A.-H. Emwas, E. Saccenti, X. Gao, R.T. McKay, V.A.P.M. dos Santos, R. Roy, D.S. Wishart, Recommended strategies for spectral processing and post-processing of 1D 1H-NMR data of biofluids with a particular focus on urine, Metabolomics. 14 (2018) 31. doi:10.1007/s11306-018-1321-4.

[61] S.I.F.. Martins, W.M.. Jongen, M.A.J.. van Boekel, A review of Maillard reaction in food and implications to kinetic modelling, Trends Food Sci. Technol. 11 (2000) 364–373. doi:10.1016/S0924-2244(01)00022-X.

[62] V. Bellon-Maurel, A. McBratney, Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils – Critical review and research perspectives, Soil Biol. Biochem. 43 (2011) 1398–1410. doi:10.1016/J.SOILBIO.2011.02.019.

[63] J.B. Reeves, The potential of mid- and near-infrared diffuse reflectance spectroscopy for determining major- and trace-element concentrations in soils from a geochemical survey of North America, Appl. Geochemistry. 24 (2009) 1472–1481. doi:10.1016/J.APGEOCHEM.2009.04.017.

[64] J. Hajslova, K. Mastovska, T. Cajka, Mass Spectrometry and Hyphenated Instruments in Food Analysis, in: Handb. Food Anal. Instruments, CRC Press, 2008. doi:10.1201/9781420045673.ch10.

[65] A. Garcia, C. Barbas, Gas Chromatography-Mass Spectrometry (GC-MS)-Based Metabolomics, in: Humana Press, 2011: pp. 191–204. doi:10.1007/978-1-61737-985-7_11.

[66] H.-J. Hübschmann, Handbook of GC-MS: Fundamentals and Applications., Wiley, 2015. https://www.wiley.com/en-us/Handbook+of+GC+MS%3A+Fundamentals+and+Applications%2C+3rd+Edition-p-9783527334742 (accessed January 30, 2019).

[67] L.G. Johnsen, P.B. Skou, B. Khakimov, R. Bro, Gas chromatography – mass spectrometry data processing made easy, J. Chromatogr. A. 1503 (2017) 57–64. doi:10.1016/j.chroma.2017.04.052.

[68] P. Valaer, Scotch Whisky, Ind. Eng. Chem. 32 (1940) 935–943. doi:10.1021/ie50367a016.

[69] A.J. Buglass, M. Mckay, C.G. Lee, Whiskeys, in: Handb. Alcohol. Beverages, John Wiley & Sons, Ltd, Chichester, UK, 2010: pp. 515–534. doi:10.1002/9780470976524.ch19.

[70] O. Klein, A. Roth, F. Dornuf, O. Schöller, W. Mäntele, The Good Vibrations of Beer. The Use of Infrared and UV/Vis Spectroscopy and Chemometry for the Quantitative Analysis of Beverages, Zeitschrift Für Naturforsch. B. 67 (2012) 1005–1015. doi:10.5560/znb.2012-0166.

[71] K. Mac Namara, F. Mcguigan, A. Hoffmann, Automated Static and Dynamic Headspace Analysis with GC-MS for Determination of Abundant and Trace Flavour Compounds in Alcoholic Beverages Containing Dry Extract, 2010.

[72] D.W. Lachenmeier, Rapid quality control of spirit drinks and beer using multivariate data analysis of Fourier transform infrared spectra, Food Chem. 101 (2007) 825–832. doi:10.1016/j.foodchem.2005.12.032.

[73] J.A. Rodrigues, A.S. Barros, B. Carvalho, T. Brandão, A.M. Gil, Probing beer aging chemistry by nuclear magnetic resonance and multivariate analysis, Anal. Chim. Acta. 702 (2011) 178–187. doi:10.1016/j.aca.2011.06.042.

[74] S. Grassi, J.M. Amigo, C.B. Lyndgaard, R. Foschino, E. Casiraghi, Assessment of the sugars and ethanol development in beer fermentation with FT-IR and multivariate curve resolution models, Food Res. Int. 62 (2014) 602–608. doi:10.1016/J.FOODRES.2014.03.058.

[75] V. Giovenzana, R. Beghi, R. Guidetti, Rapid evaluation of craft beer quality during fermentation process by vis/NIR spectroscopy, J. Food Eng. 142 (2014) 80–86. doi:10.1016/J.JFOODENG.2014.06.017.

[76] O. Oladokun, A. Tarrega, S. James, K. Smart, J. Hort, D. Cook, The impact of hop bitter acid and

polyphenol profiles on the perceived bitterness of beer, Food Chem. 205 (2016) 212–220. doi:10.1016/j.foodchem.2016.03.023.

[77] C. Andrés-Iglesias, C.A. Blanco, J. Blanco, O. Montero, Mass spectrometry-based metabolomics approach to determine differential metabolites between regular and non-alcohol beers, Food Chem. 157 (2014) 205–212. doi:10.1016/j.foodchem.2014.01.123.

[78] S. Rossi, V. Sileoni, G. Perretti, O. Marconi, Characterization of the volatile profiles of beer using headspace solid-phase microextraction and gas chromatography-mass spectrometry, J. Sci. Food Agric. 94 (2014) 919–928. doi:10.1002/jsfa.6336.

[79] E. Bravi, O. Marconi, V. Sileoni, G. Perretti, Determination of free fatty acids in beer, Food Chem. 215 (2017) 341–346. doi:10.1016/J.FOODCHEM.2016.07.153.

[80] S. Engelhard, H.-G. Löhmannsröben, F. Schael, Quantifying Ethanol Content of Beer Using Interpretive Near-Infrared Spectroscopy, Appl. Spectrosc. 58 (2004) 1205–1209. doi:10.1366/0003702042336000.

[81] L. Vera, L. Aceña, J. Guasch, R. Boqué, M. Mestres, O. Busto, Discrimination and sensory description of beers through data fusion, Talanta. (2011). doi:10.1016/j.talanta.2011.09.052.

[82] D.-W. Sun, Å. Rinnan, L. Nørgaard, F. van den Berg, J. Thygesen, R. Bro, S.B. Engelsen, Data Pre-processing, in: Da-Wen Sun (Ed.), Infrared Spectrosc. Food Qual. Anal. Control, Elsevier, 2009: pp. 29–50. doi:10.1016/B978-0-12-374136-3.00002-X.

[83] I. Duarte, A. Barros, P.S. Belton, R. Righelato, M. Spraul, E. Humpfer, A.M. Gil, High-Resolution Nuclear Magnetic Resonance Spectroscopy and Multivariate Analysis for the Characterization of Beer, J. Agric. Food Chem. 50 (2002) 2475–2481. doi:10.1021/jf011345j.

[84] F. Savorani, G. Tomasi, S.B. Engelsen, icoshift: A versatile tool for the rapid alignment of 1D NMR spectra, J. Magn. Reson. 202 (2010) 190–202. doi:10.1016/j.jmr.2009.11.012.

[85] F. Savorani, G. Tomasi, S.B. Engelsen, Alignment of 1D NMR Data using the iCoshift Tool: A Tutorial, in: J. van Duynhoven, P.S. Belton, Webb. G.A., H. van As (Eds.), Magn. Reson. Food Sci. Food Thought, Royal Society of Chemistry, 2013: pp. 14–24. doi:10.1039/9781849737531-00014.

[86] R. Bro, S. De Jong, A fast non-negativity-constrained least squares algorithm, J. Chemom. 11 (1997) 393–401. doi:10.1002/(SICI)1099-128X(199709/10)11:5<393::AID-CEM483>3.0.CO;2-L.

[87] W. Windig, Two-Way Data Analysis: Detection of Purest Variables, in: Compr. Chemom., 2010. doi:10.1016/B978-044452701-1.00048-X.

[88] A.C. Dona, M. Kyriakides, F. Scott, E.A. Shephard, D. Varshavi, K. Veselkov, J.R. Everett, A guide to the identification of metabolites in NMR-based metabonomics/metabolomics experiments, Comput. Struct. Biotechnol. J. (2016). doi:10.1016/j.csbj.2016.02.005.

[89] What is whisky?, Lancet. 166 (1905) 1490. doi:10.1016/S0140-6736(00)68466-0.

[90] What is Whiskey?, Nature. 73 (1906) 441–442. doi:10.1038/073441b0.

[91] The chemistry of whisky-and-soda, Lancet. 162 (1903) 483. doi:10.1016/S0140-6736(00)67396-8.

[92] F.R. Jack, Whiskies: composition, sensory properties and sensory analysis, in: Alcohol. Beverages, Elsevier, 2012: pp. 379–392. doi:10.1533/9780857095176.3.379.

[93] J.R. Piggott, Whisky, Whiskey and Bourbon: Composition and Analysis of Whisky, in: Encycl. Food Heal., Elsevier, 2016: pp. 514–518. doi:10.1016/B978-0-12-384947-2.00752-2.

[94] W. Kew, I. Goodall, D. Clarke, D. Uhrín, Chemical Diversity and Complexity of Scotch Whisky as Revealed by High-Resolution Mass Spectrometry, J. Am. Soc. Mass Spectrom. 28 (2017) 200–213. doi:10.1007/s13361-016-1513-y.

[95] E. Campo, J. Cacho, V. Ferreira, Solid phase extraction, multidimensional gas chromatography mass spectrometry determination of four novel aroma powerful ethyl esters: Assessment of their occurrence and importance in wine and other alcoholic beverages, J. Chromatogr. A. 1140 (2007) 180–188. doi:10.1016/J.CHROMA.2006.11.036.

[96] M. Fujieda, T. Tanaka, Y. Suwa, S. Koshimizu, I. Kouno, Isolation and Structure of Whiskey Polyphenols Produced by Oxidation of Oak Wood Ellagitannins, J. Agric. Food Chem. 56 (2008) 7305–7310. doi:10.1021/jf8012713.

[97] Y. Nie, E. Kleine-Benne, Determining Phenolic Compounds in Whisky using Direct Large Volume Injection and Stir Bar Sorptive Extraction, 2012.

[98] B. White, M.R. Smyth, C.E. Lunte, Determination of phenolic acids in a range of Irish whiskies, including single pot stills and aged single malts, using capillary electrophoresis with field amplified sample stacking, Anal. Methods. 9 (2017) 1248–1252. doi:10.1039/C6AY03299K.

[99] K. MacNamara, N. Burke, N. Conway, Aroma profiling of distilled spirits using vacuum fractional distillation with specific phase ration capillary GC analysis and selective detection, 1989.

[100] K. Macnamara, Investigation of Medium Volatile Sulfur Compounds in Whiskey, 1992.

[101] J.R. Piggott, J.M. Conner, A. Paterson, Flavour development in whisky maturation, Dev. Food Sci. 37 (1995) 1731–1751. doi:10.1016/S0167-4501(06)80261-X.

[102] J.M. Conner, A. Paterson, J.R. Piggott, Release of distillate flavour compounds in Scotch malt whisky, J. Sci. Food Agric. 79 (1999) 1015–1020. doi:10.1002/(SICI)1097-0010(19990515)79:7<1015::AID-JSFA321>3.0.CO;2-R.

[103] J. Conner, K. Reid, G. Richardson, SPME Analysis of Flavor Components in the Headspace of Scotch Whiskey and Their Subsequent Correlation with Sensory Perception, in: 2001: pp. 113–122. doi:10.1021/bk-2001-0782.ch010.

[104] K. MacNamara, Flavour components of whiskey, 2002. http://scholar.sun.ac.za.

[105] L. Poisson, P. Schieberle, Characterization of the Most Odor-Active Compounds in an American Bourbon Whisky by Application of the Aroma Extract Dilution Analysis, J. Agric. Food Chem. 56 (2008) 5813–5819. doi:10.1021/jf800382m.

[106] K.-Y.M. Lee, A. Paterson, J.R. Piggott, G.D. Richardson, Sensory discrimination of blended Scotch whiskies of different product categories, Food Qual. Prefer. 12 (2001) 109–117. doi:10.1016/S0950-3293(00)00037-9.

[107] K.-Y.M. Lee, A. Paterson, J.R. Piggott, G.D. Richardson, Origins of Flavour in Whiskies and a Revised Flavour Wheel: a Review, J. Inst. Brew. 107 (2001) 287–313. doi:10.1002/j.2050-0416.2001.tb00099.x.

[108] F.R. Jack, G.M. Steele, Modelling the sensory characteristics of Scotch whisky using neural networks—a novel tool for generic protection, Food Qual. Prefer. 13 (2002) 163–172. doi:10.1016/S0950-3293(02)00012-5.

[109] P. Wiśniewska, T. Dymerski, W. Wardencki, J. Namieśnik, Chemical composition analysis and authentication of whisky, J. Sci. Food Agric. 95 (2015) 2159–2166. doi:10.1002/jsfa.6960.

[110] R.I. Aylott, W.M. MacKenzie, Analytical Strategies to Confirm the Generic Authenticity of Scotch Whisky, J. Inst. Brew. 116 (2010) 215–229. doi:10.1002/j.2050-0416.2010.tb00424.x.

[111] W.M. MacKenzie, R.I. Aylott, Analytical strategies to confirm Scotch whisky authenticity., Analyst. 129 (2004) 607. doi:10.1039/b403068k.

[112] W. Meier-Augenstein, H.F. Kemp, S.M.L. Hardie, Detection of counterfeit scotch whisky by 2H and 18O stable isotope analysis, Food Chem. 133 (2012) 1070–1074. doi:10.1016/J.FOODCHEM.2012.01.084.

[113] F.W. Lima, C.M. Silva, R. Guimarães, An actual case of examination of counterfeited whisky, J. Radioanal. Chem. 15 (1973) 157–164. doi:10.1007/BF02516567.

[114] T. Kuballa, T. Hausler, A.O. Okaru, M. Neufeld, K.O. Abuga, I.O. Kibwage, J. Rehm, B. Luy, S.G. Walch, D.W. Lachenmeier, Detection of counterfeit brand spirits using 1H NMR fingerprints in comparison to sensory analysis, Food Chem. 245 (2018) 112–118. doi:10.1016/J.FOODCHEM.2017.10.065.

[115] A.R. Martins, M. Talhavini, M.L. Vieira, J.J. Zacca, J.W.B. Braga, Discrimination of whisky brands and counterfeit identification by UV–Vis spectroscopy and multivariate data analysis, Food Chem. 229 (2017) 142–151. doi:10.1016/J.FOODCHEM.2017.02.024.

[116] P. Wiśniewska, R. Boqué, E. Borràs, O. Busto, W. Wardencki, J. Namieśnik, T. Dymerski, Authentication of whisky due to its botanical origin and way of production by instrumental analysis and multivariate classification methods, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 173 (2017) 849–853. doi:10.1016/J.SAA.2016.10.042.

[117] F. Tosato, R.M. Correia, B.G. Oliveira, A.M. Fontes, H.S. França, W.K.T. Coltro, P.R. Filgueiras, W. Romão, Paper spray ionization mass spectrometry allied to chemometric tools for quantification of whisky adulteration with additions of sugarcane spirit, Anal. Methods. 10 (2018) 1952–1960. doi:10.1039/C8AY00071A.

[118] † Domingo González-Arjona, † Germán López-Pérez, ‡ and Víctor González-Gallero, ‡ A. Gustavo González*, Supervised Pattern Recognition Procedures for Discrimination of Whiskeys from Gas Chromatography/Mass Spectrometry Congener Analysis, (2006). doi:10.1021/JF0517389.

[119] A. Herranz, P. de la Serna, C. Barro, P.J. Martin, M.D. Cabezudo, Application of the statistical multivariate analysis to the differentiation of whiskies of different brands, Food Chem. 31 (1989) 73–81. doi:10.1016/0308-8146(89)90152-0.

[120] K. Sujka, P. Koczoń, The application of FT-IR spectroscopy in discrimination of differently originated and aged whisky, Eur. Food Res. Technol. (2018) 1–7. doi:10.1007/s00217-018-3113-5.

[121] A.G. Mignani, L. Ciaccheri, B. Gordillo, A.A. Mencaglia, M.L. González-Miret, F.J. Heredia, B. Culshaw, Identifying the production region of single-malt Scotch whiskies using optical spectroscopy and

pattern recognition techniques, Sensors Actuators B Chem. 171–172 (2012) 458–462. doi:10.1016/J.SNB.2012.05.011.

[122] J. Clyne, J.M. Conner, A. Paterson, J.R. Piggott, The effect of cask charring on Scotch whisky maturation, Int. J. Food Sci. Technol. 28 (2007) 69–81. doi:10.1111/j.1365-2621.1993.tb01252.x.

[123] K. Macnamara, D. Dabrowska, M. Baden, N. Helle, Advances in the Ageing Chemistry of Distilled Spirits Matured in Oak Barrels, n.d. www.chromatographyonline.com.

[124] H. Aoshima, S. Hossain, H. Koda, Y. Kiso, Why Is an Aged Whiskey Highly Valued?, Curr. Nutr. Food Sci. 5 (2009) 204–208. doi:10.2174/157340109789007090.

[125] K. Macnamara, J. Van Wyk, O.P.H. Augustyn, A. Rapp4, Flavour Components of Whiskey. II. Ageing Changes in the High-Volatility Fraction, South African J. Enol. Vitic. (2001). doi:http://dx.doi.org/10.21548/22-2-2196.

[126] K. Macnamara, J. Van Wyk, P. Brunerie, O.P.H. Augustyn, A. Rapp, Flavour Components of Whiskey. III. Ageing Changes in the Low-Volatility Fraction, 2001. (n.d.).

[127] B. Harrison, J. Ellis, D. Broadhurst, K. Reid, R. Goodacre, F.G. Priest, Differentiation of Peats Used in the Preparation of Malt for Scotch Whisky Production Using Fourier Transform Infrared Spectroscopy, J. Inst. Brew. 112 (2006) 333–339. doi:10.1002/j.2050-0416.2006.tb00739.x.

[128] B.M. Harrison, F.G. Priest, Composition of Peats Used in the Preparation of Malt for Scotch Whisky Production-Influence of Geographical Source and Extraction Depth, J. Agric. Food Chem. 57 (2009) 2385–2391. doi:10.1021/jf803556y.

[129] G. Tomasi, F. van den Berg, C. Andersson, Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data, J. Chemom. 18 (2004) 231–241. doi:10.1002/cem.859.

[130] T. Skov, F. van den Berg, G. Tomasi, R. Bro, Automated alignment of chromatographic data, J. Chemom. 20 (2006) 484–497. doi:10.1002/cem.1031.

[131] L. Eriksson, E. Johansson, N. Kettaneh-Wold, S. Wold, Multi- and Megavariate Data Analysis: Part I – Basic Principles and Applications, Umetrics Acedemy, 2013.

[132] P.D. Wentzell, S. Hou, Exploratory data analysis with noisy measurements, J. Chemom. (2012). doi:10.1002/cem.2428.

[133] G. Karypis, Eui-Hong Han, V. Kumar, Chameleon: hierarchical clustering using dynamic modeling, Computer (Long. Beach. Calif). 32 (1999) 68–75. doi:10.1109/2.781637.

[134] P. Fränti, S. Sieranoja, K-means properties on six clustering benchmark datasets, Appl. Intell. 48 (2018) 4743–4759. doi:10.1007/s10489-018-1238-7.

[135] J. Jaumot, A. de Juan, R. Tauler, MCR-ALS GUI 2.0: New features and applications, Chemom. Intell. Lab. Syst. 140 (2015) 1–12. doi:10.1016/j.chemolab.2014.10.003.

Chapter 3 | Fused Adjacency Matrix approach

3.1. Introduction

Despite the large variety of exploratory multivariate data analysis methods available nowadays [1], there are still cases in which it is difficult to obtain satisfactory results regarding groupings and the latent phenomena buried in the data. Highly complex data may not show simple groupings and/or trends even when projected to a space of lower dimensionality, as they may be so complex that common visualizations tools are only shedding limited light on the underlying structures.

The Fused Adjacency Matrix approach was developed starting from these premises. The overall idea of the approach is to combine multiple "weak sources" of information which, when combined, will provide better discriminatory information [2,3]. The approach is based on the combination of several adjacency matrices, which represent the weak sources of information and contain a "coded version" of the sample-to-sample distance information.

Even if the proposed approach is intended as an unsupervised exploratory tool, it can also be used as a method for mid-level data fusion: if more blocks of data are available [4], the information can be extracted and encoded into as many fused adjacency matrices ($\mathbf{AM_X}$ in Figure 3.3) as the number of data blocks, and then combined into a single final fused adjacency matrix. These two steps – extraction and final data fusion – are marked at the bottom of Figure 3.3.

This Chapter is organized as follows: an overview about the two distance measures employed by the approach is given in Section 3.2, while the theory behind the Self-Organizing Maps is described in Section 3.3; Section 3.4 is devoted to describing the framework of data fusion methodologies; a detailed description of the Fused Adjacency Matrix approach is given in Section 3.5, while a graphical description of the approach is given in Figure 3.3; finally, in Section 3.6 the approach is applied to some dataset as an exploratory tool and in Section 3.7 its application as a mid-level data fusion method is reported using the Beer datasets as a benchmark.

3.2. Distance measures

This section is devoted to the theory behind Euclidean and Mahalanobis distances, the two similarity/dissimilarity measures used in the Fused Adjacency Matrix approach. These are the two most commonly used distance measures [5] and also for this reason it was decided to use them in the approach, in the perspective of working with relatively simple tools, which had to be combined in a more complex system.

Within the framework of dissimilarity measures, distances *D* are functions which satisfy three axioms (Equations 3.1a-b-c), and both the Euclidean and Mahalanobis satisfy all of them:

(3.1a) $D_{xy} \geq 0$ non-negativity (negative distances cannot exist)

(3.1b) $D_{xx} = 0$ reflexivity (an object cannot be dissimilar from itself)

(3.1c) $D_{xy} = D_{yx}$ symmetry (the distance between a pair of objects has no direction)

The Euclidean distance (D^{Euc}) is by far the most common distance measure, and it corresponds to the shortest path joining two points [6]. It is *unbounded*, which means that it has no upper limit, so its range is from zero to infinite. It is easy to compute and interpret. If we consider two *n*-dimensional objects **x** and **y** belonging to a $m \times n$ dataset, the D^{Euc} between them can be computed by Equation 3.2 [6]:

$$(3.2) \qquad D_{xy}^{Euc} = \sqrt{\sum_{j=1}^{m}\left(x_j - y_j\right)^2} = \sqrt{(\mathbf{x}-\mathbf{y})^T \cdot (\mathbf{x}-\mathbf{y})}$$

where x_j and y_j indicate the values for the *j*th variable of the objects **x** and **y**. The second part of the equation expresses the vector form.

The Mahalanobis distance (D^{Mah}) works in a similar way to D^{Euc}, but it also considers the information on the whole structure of the dataset, so that the possible correlation among variables is also taken into account and underweighted, reducing the amount of redundant information [5,6]. This is done by means of the data covariance matrix **S**, as it is shown in Equation 3.3 [6]:

$$(3.3) \qquad D_{xy}^{Mah} = \sqrt{(\mathbf{x}-\mathbf{y})^{\mathrm{T}} \cdot \mathbf{S}^{-1} \cdot (\mathbf{x}-\mathbf{y})}$$

By comparing Equations 3.2 and 3.3, it becomes clear how D^{Euc} and D^{Mah} D are related: if the covariance matrix **S** is replaced by the identity matrix **I**, the D^{Mah} reduces to the D^{Euc}. The D^{Mah} can therefore be considered as a generalization of the D^{Euc} [6]. Moreover, when working in the principal component space, it is possible to compute the D^{Mah} as a D^{Euc} in scores space after autoscaling[1]. The D^{Mah} computed on the PC scores operates with a diagonal **S** matrix, because the PCs are orthogonal by definition, i.e. uncorrelated. Autoscaling the scores converts the covariance matrix **S** to the identity matrix **I**, making the D^{Mah} equation equivalent to the D^{Euc} equation.

3.3. Kohonen's Self-Organizing Maps (SOM)

A way for handling possible non-linearities and more complex structures in the samples' space is the use of the Kohonen's Self-Organizing Maps (SOMs, [7,8]). The SOM method is a type of artificial neural network that is particularly suitable for modelling non-linear boundaries between samples belonging to different groups.

Its aim is to obtain a low-dimensional representation of the high-dimensional input space. The high-dimensional space is mapped using a set of representative coordinates, which are distributed unevenly over the space, based on data structure and sample density. For this reason, this is a non-linear mapping procedure. These coordinates are called nodes (or neurons) and are organized on a "top-map", typically a two-dimensional grid whose geometry may vary (Figure 3.1). Each node is characterized by a weight vector, which has the same dimensions of the original space.

SOM mapping preserves the topology, and this means that distances and proximity relations between samples in the original space are preserved [7]. As a result of this, all the nodes that are at the same topological distance from a given node define a "neighbourhood". Depending on the selected neighbourhood shape, each node may have 4, 6 or 8 nearest neighbours, which correspond to, respectively, square-, hexagonal- and rectangular-shaped neighbourhoods. A representation of the different neighbourhood shapes and levels is given in Figure 3.1.

[1] *Autoscale* = the mean value is subtracted from each column which is then divided by its standard deviation; the result is that each column has mean = 0 and standard deviation = 1.

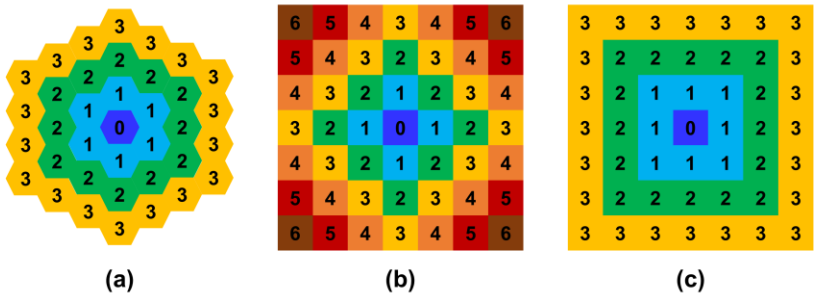


Figure 3.1. Neighbourhood types – hexagonal (a), square (b) and rectangular (c); the numbers indicate the topological distances from the central node (zero). (adapted from Marini *et al.* [9])

The SOM algorithm proceeds by unsupervised competitive learning. "Unsupervised" because no desired outcome is imposed *a priori*, therefore the network must adapt itself (hence the name "self-organizing" maps) according to the data structure, iteratively accommodating all the input samples on the grid. "Competitive" because a *winner-takes-all* approach is implemented: at each iteration, each input is presented to the network, compared to all the nodes and finally it is assigned to the most similar one [10]. When the winning node is selected, its weight vector gets updated based on the difference between its old weight values and the weight vector of the input. This correction is also applied to the neighbouring nodes based on a function that considers their topological distance and the learning rate parameter.

The top-map can be used as an exploratory tool for the identification of clusters [7], since it allows to assess similarity between samples in a simple and direct way, by comparing their position on the top-map.

3.4. Data fusion techniques

Data fusion methods are strategies for combining different sources of complementary information, for instance data blocks obtained from the analysis of the same set of samples by means of different analytical techniques. Data fusion strategies are generally grouped into three families, low-, mid- and high-level methods [4,10,11], as represented in Figure 3.2.

In low-level data fusion, two or more data blocks are simply joined together by augmentation in the variables' or samples' direction. In the case of many data blocks, the resulting matrix will have as many columns as the sum of the number of columns of all the data blocks. It is generally required little preprocessing of the resulting augmented matrix prior to modelling, mainly a scale correction to make each source comparable.

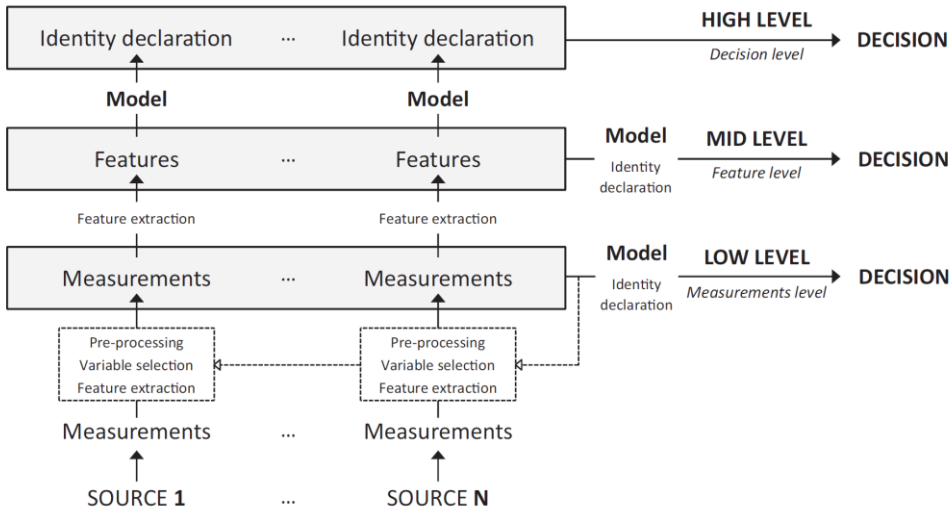


Figure 3.2. Data fusion strategies. (from Borràs *et al.* [4])

Mid-level data fusion [10,12] is accomplished by combining relevant features separately extracted from each data block into a new data matrix: these extracted features can be for instance PCA scores, resolved resonance or chromatographic peaks, or scores from partial-least squares discriminant analysis [12]. In general, block-scaling and autoscaling are performed on the resulting fused data matrix prior to any further modelling.

High-level data fusion concerns the combination of the outputs of different supervised models, and for this reason it is also referred to as "decision level fusion".

3.5. The Fused Adjacency Matrix approach

The Fused Adjacency Matrix approach is based on the concept of combining different *weak sources* of information [2,3,13,14] as it is done, for instance, in the classification context by the Random Forest algorithm [13]. In Random Forest the results of several *weak classifiers* are merged by counting how many times a sample was assigned to one of the defined categories; then the sample is assigned to the category to which it was more often assigned. Another strategy also used in the supervised context is to combine the results obtained from an ensemble of different classification methods [15,16] which, individually, may not be good enough. Several fusion rules to combine the different classifiers/classifier outcomes were proposed [15–17] and more recently, a combination strategy for non-optimized classifiers based on defining a window of tuning parameters values for each classifier was proposed by Brownfield *et al.* [18].

In our unsupervised case, the distance information is converted into several *adjacency matrices*, which represent the weak sources of information. Adjacency matrices (AMs) are squared binary symmetric matrices ($m \times m$) in which a one is present when the *adjacency condition* is fulfilled by the pair of samples under examination, while a zero is present when this condition is not fulfilled. In other words, these matrices carry the information about the pairwise relations between the objects: a relation exists if a pair of objects is close enough (i.e. they are "adjacent") as compared to, for instance, a distance threshold (i.e. the adjacency condition). An example with a hard threshold t as the adjacency condition is given in Equation 3.4:

$$\mathbf{AM} = \begin{cases} a_{i,j} = 1 \Rightarrow d(i,j) \leq t \\ a_{i,j} = 0 \Rightarrow d(i,j) > t \end{cases} \tag{3.4}$$

where $d(i,j)$ represents the distance between the ith and the jth objects.

In the Fused Adjacency Matrix approach, several AMs are generated and merged step-by-step using a sum rule [15]: the output adjacency matrix $\mathbf{AM_X}$, which is still squared and symmetric, is eventually obtained as a result. In $\mathbf{AM_X}$ those pairs of samples that were consistently found adjacent will be characterized by high values, while those pairs of samples which were consistently found far apart will have low values or, even better,

values close to zero. This is the overall idea of the proposed approach, and Figure 3.3 provides a graphical representation of the whole approach.

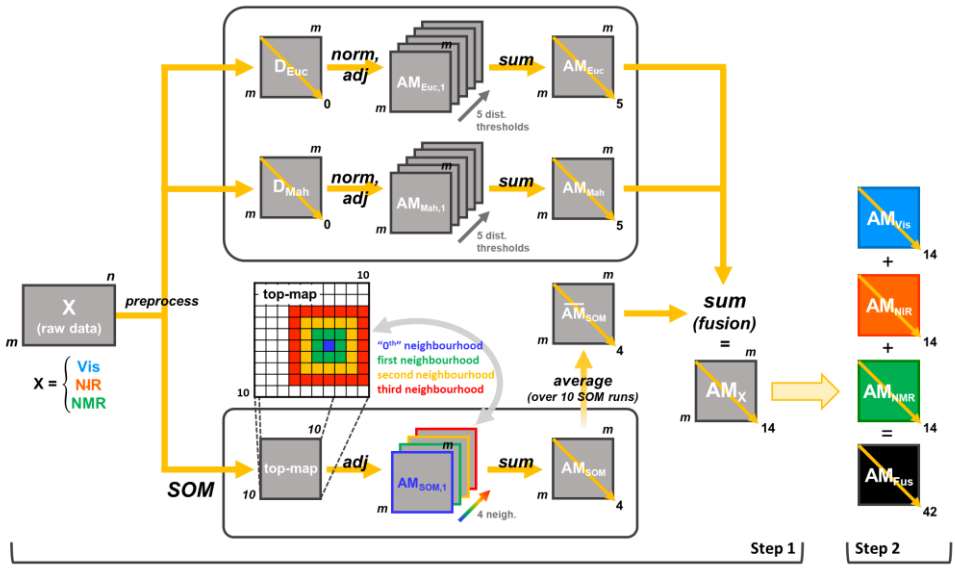


Figure 3.3. Graphical representation of the Fused Adjacency Matrix approach; in the top box, the adjacency matrices are obtained from Euclidean and Mahalanobis distances, while in the lower box they are obtained using SOM.

For a given data block **X**, the corresponding output is the matrix $\mathbf{AM_X}$, which is obtained from the combination of the AMs obtained from the Euclidean distance (5 AMs), the Mahalanobis distance (5 AMs) and SOM (4 AMs), as explained by Equations 3.5–3.8:

$$\mathbf{AM_{Euc}} = \sum_{t=1}^{5} \mathbf{AM_{Euc},}_{t} \tag{3.5}$$

$$\mathbf{AM_{Mah}} = \sum_{t=1}^{5} \mathbf{AM_{Mah},}_{t} \tag{3.6}$$

$$\mathbf{AM_{SOM}} = \sum_{g=1}^{4} \mathbf{AM_{SOM},}_{g} \tag{3.7}$$

$$\mathbf{AM_X} = \mathbf{AM_{Euc}} + \mathbf{AM_{Mah}} + \mathbf{AM_{SOM}} \tag{3.8}$$

where t represents the five distance thresholds used for Euclidean and Mahalanobis distances, while g corresponds to the four neighbourhoods used with SOM's top-maps.

The output matrix $\mathbf{AM_X}$ obtained from Equation 3.8 is symmetric, squared and its diagonal values are all equal to 14, like the total number of generated AMs. Its values are within the range 0–14, and because of the way the AMs from SOM are obtained, they may not assume even values, as one would expect from a sum of binary matrices. This fact will be better elucidated in the next paragraphs.

Due to the number of implemented thresholds (Euc = 5, Mah = 5 and SOM = 4), the contribution of each distance measure to the $\mathbf{AM_X}$ is comparable. However, the use of a weighted sum can be advised in the more general case.

3.5.1. Use of Euclidean and Mahalanobis distances in the approach

In an early version of the approach, the Euclidean and Mahalanobis distance matrices ($\mathbf{D_{Euc}}$ and $\mathbf{D_{Mah}}$ in Figure 3.3) were both normalized between zero and one, and the same window [18] of five threshold values (0.05 / 0.1 / 0.2 / 0.3 / 0.4) was applied to them. With this system however, extreme outliers can be very influential on the generation of the AMs, because of the normalization step: Figure 3.4a shows how strongly extreme outliers can "compress" the distribution of distance values using the Euclidean distance and a "toy" dataset of honey samples [19]. The more extreme an outlier, the stronger this "compression effect" becomes.

Since the distance information was coded into the AMs through fixed hard thresholds, the result was that all the samples would result equally close, and any similarity or difference among the samples that would characterize the dataset would be ironed out.

On the other hand, it was also necessary to avoid the opposite problem. The distances' distribution may also shift towards the right, meaning that, on average, the samples are located further apart from each other. This may be not so influential from the point of view of group and data structure, but with the use of fixed hard thresholds, this may easily lead to a lacking sampling of the distribution, with the consequence of having all the samples equally distant, and very few of them being adjacent.

To avoid both these potentially disrupting effects, it was decided to use the median and the minimum values of the distances' distribution to define the range in which the thresholds would be defined and used. A "moving set" of thresholds was implemented

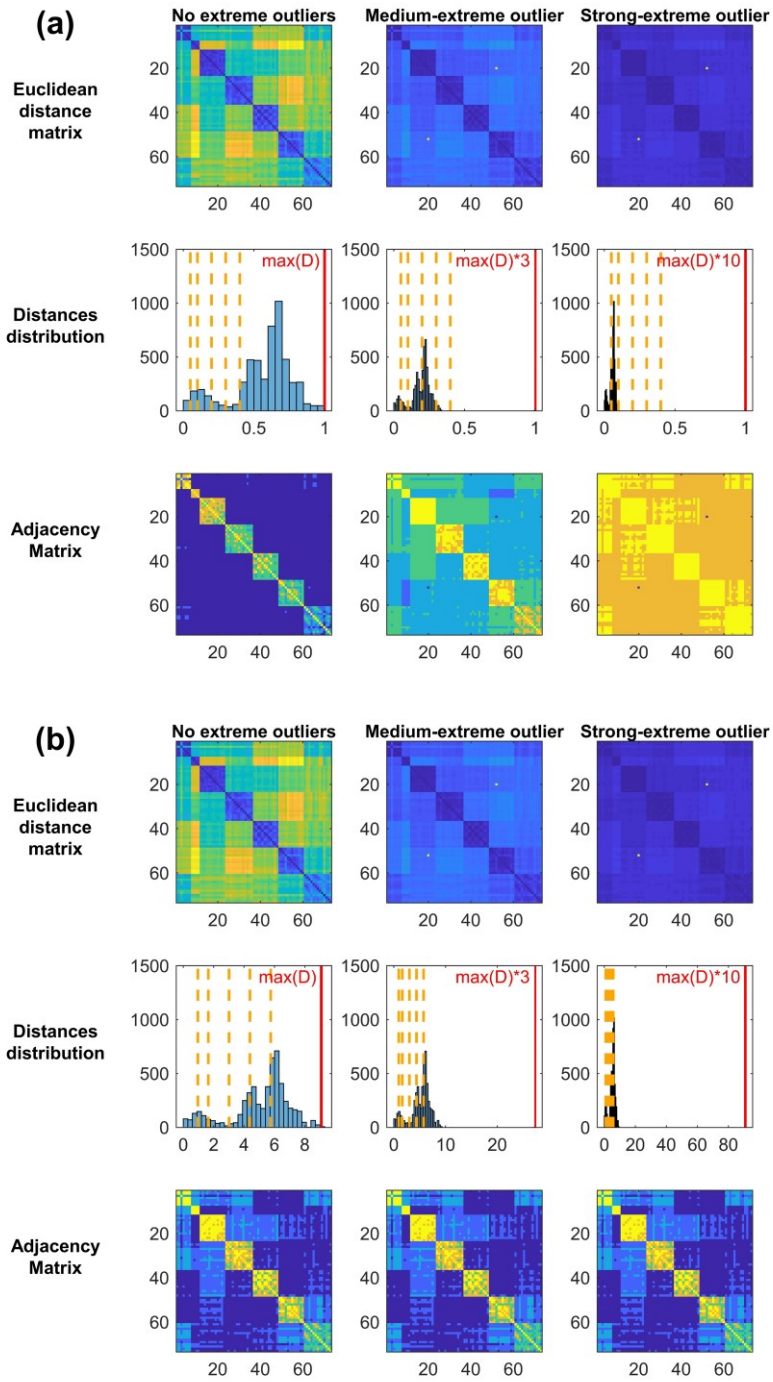


Figure 3.4. (a) The "compression effect" due to the presence of increasingly extreme outliers; (b) the moving thresholds can avoid the disruptive effect of extreme outliers. (the honey dataset, from Marini *et al.* [19])

as shown in Figure 3.4b. The normalization step became therefore useless and was removed, together with the potential influence of extreme outliers.

The threshold values are then allowed to change depending on the input data, while the use of different distance measures allows to get a more complete representation of the data.

3.5.2. Use of SOM in the approach

SOM does not provide a distance matrix, but instead a grid of nodes (the top-map), on which the samples are arranged. In this case, the adjacency condition to be checked is whether the two samples under examination are found at a topological distance equal to or lower than g, a parameter that defines the considered topological distance. Rectangular topological neighbourhoods [9] were defined, and the adjacency condition was checked along four levels (g = 0, 1, 2, 3), with the "zero*th* level" corresponding to a single node (i.e. topological distance is zero).

Since different SOM runs generally produce slightly different outputs, the average over ten runs was taken to make the resulting adjacency matrix $\mathbf{AM_{SOM}}$ more robust.

To work with a basic but general set of parameters, a simple two-dimensional 10 × 10 squared grid of nodes was used for SOM modelling, as suggested by Kohonen [8]. Regarding the grid's dimensions, as reported by Simon *et al.* [20], using too small grids may increase the chance of class conflicts, while using too big grids may cause the samples to end too much spread out, hampering the recognition of clusters.

The network was trained for at least 5000 epochs, with rectangular neighbourhoods and a gaussian function for modulating the distance based-learning. The initial neighbourhood size was made coincident with the grid's size, as suggested by Marini *et al.* [7]. The number of training epochs was set quite high because, from the point of view of modelling, the map needs enough time to arrange itself, and the only rule-of-thumb for this parameter is that the number should not be too low. Setting such a high value increases the likelihood that the map will reach an optimal organization, but also allows to model possible smaller differences among samples to a higher degree. Moreover, compression was applied prior to SOM modelling, therefore the required computational

time was substantially reduced, compared to the same calculation using the uncompressed input data.

This set of parameters was used for modelling the real-case datasets (Beer, Whisky and Olive oil datasets), while it was decided to test the algorithm using different, adaptive parameters, together with an additional hexagonal grid. In Section 3.6.1 it is described how the parameters were chosen.

It is worth considering that in SOM modelling the fact that a sample is assigned to a specific node does not imply that the selected node represents a good approximation for describing the sample. Since the aim of the approach is to obtain a "simpler", coded version of the distance information, this uncertainty is probably negligible. For instance, in the Euclidean and Mahalanobis distances cases hard thresholds are used, therefore the coded information about a pair of samples at a distance slightly shorter than a threshold (where a "1" is assigned) would result completely different from a pair of similar samples but at a distance slightly longer than the threshold (where a "0" is assigned). In the same way, the fact that a node could be a better or worse approximation depending on the sample represents a much smaller uncertainty than the assignment to a different (even if close) node.

3.5.3. Use of the approach as a mid-level data fusion method

When more **X** data blocks are available (like in the benchmark case presented in Section 3.7.1), the resulting $\mathbf{AM_X}$ matrices can be combined using, again, a sum rule [18] (Equation 3.9). The resulting matrix is the Fused Adjacency Matrix $\mathbf{AM_{Fus}}$, depicted in black in Figure 3.3.

(3.9) $$\mathbf{AM_{Fus}} = \sum_X \mathbf{AM_X}$$

3.5.4. Preprocessing of symmetric squared matrices

The preprocessing of symmetric squared matrices such as distance and adjacency matrices or the Fused Adjacency Matrix should be done preserving their symmetric

structure. The *double centering* preprocessing described by Vitale *et al.* [21] operates on an adjacency matrix **AM** as follows:

$$\mathbf{AM_{dc}} = \mathbf{AM} - \overline{\mathbf{AM}}_{\mathbf{columns}} - \overline{\mathbf{AM}}_{\mathbf{rows}} + \overline{\mathbf{AM}} \quad (3.10)$$

which corresponds to removing both the column mean $\overline{\mathbf{AM}}_{\mathbf{columns}}$ and the row mean $\overline{\mathbf{AM}}_{\mathbf{rows}}$ (which are the same when the matrix is squared and symmetric), and finally adding back the overall mean $\overline{\mathbf{AM}}$. The diagonal elements of this matrix are no more all equal, but they correspond to the row (or column) sum.

3.6. Exploratory applications

This section is organized in two parts: first the test results obtained with four simulated datasets are reported, followed by the exploratory application of the Fused Adjacency Matrix approach on the whisky dataset.

The simulated datasets were used for testing the influence of different sets of thresholds on the final Fused Adjacency Matrix, as well as the contribution of an additional hexagonal SOM grid. Parameters such as maximum connectivity and the number of non-zero elements from the difference of subsequent AMs were used for choosing the optimal numbers of thresholds, with the aid of visual inspection of the AMs corresponding to each tested threshold value.

However, the optimal threshold number choice can be still considered work in progress.

3.6.1. Simulated datasets

As described in Section 2.3.5, four simulated datasets were designed and used to test the performances of the Fused Adjacency Matrix approach with different sets of parameters. More precisely, some parameters of SOM were adapted to the datasets' characteristics, and an additional SOM grid was used.

The grids' dimensions SOM_x and SOM_y were automatically set using the closer integer to the square root of the number of samples m, as a rule of thumb (Eq. 3.11). As described in Section 3.5.2, the initial neighbourhood size was made coincident with the grid's size.

(3.11) $$SOM_x = SOM_y = \text{int}\left(\sqrt{m}\right)$$

Just like the distance contributions rely on two different types of distance metric, an additional grid for SOM was chosen. Since SOM modelling is the most time-consuming step of the whole approach, it was decided to leave out the quadratic grid and focus on the hexagonal grid, whose shape is more different compared to the rectangular one. The number of g neighbourhoods was automatically set by linking it to the grid's dimensions, as described by Equation 3.12:

(3.12) $$g = 0, 1, 2, \ldots, G-1 \quad \text{with } G = \text{int}\left(\frac{SOM_x}{2.5}\right)$$

The number of SOM replicated runs was left unchanged to 10, mainly for computational time constraints. The automatically determined SOM parameters used for modelling the simulated datasets are reported in Table 3.1.

Table 3.1. SOM parameters for the simulated datasets

	grid's dimensions ($SOM_x \times SOM_y$)	***G*–1**
Globular #1	21 × 21	8
Globular #2	30 × 30	12
Circles	32 × 32	13
t4.8k	45 × 45	18

To test how different numbers of thresholds affect the structure recovery in the Fused Adjacency Matrix approach, the intermediate steps of the procedure were inspected. The procedure for this assessment envisages focusing on each distance measure/SOM grid shape individually: one-to-one comparisons are made with the Fused Adjacency Matrix, starting with the Euclidean distance, then the Mahalanobis distance and then the two grids used is SOM (rectangular and hexagonal). From each comparison the best number of thresholds is obtained, and the set of optimized numbers of thresholds is finally used for computing a new Fused Adjacency Matrix, which is then inspected.

3.6.1.1. Globular clusters dataset #1

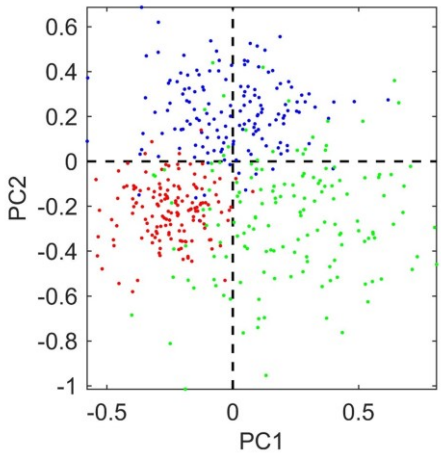


Figure 3.5. The original distribution of the globular dataset #1.

The globular dataset #1 (Figure 3.5) consists of three slightly overlapped clusters, whose structure can be deducted from Figure 3.6b-d. In this Figure, all rows and columns of the Fused Adjacency Matrix (b) and the Euclidean distance matrix (d) were reordered according to the known classes, so the three clusters are rather clearly separated. By inspecting the same matrices reordered according to the OPTICS sequence obtained from the Fused Adjacency Matrix, it can be seen that the original structure is partially recovered. The Fused Adjacency Matrix can recover the structure in more detail than the distance matrix, which appears more blurred. The three clusters can be better seen in the Fused case (Figure 3.6b) than in the Euclidean case (Figure 3.6d), in which the third cluster appears much less defined than the other two.

Five AMs were computed from the Euclidean distance matrix and are represented in Figure 3.7, in which columns and rows were reordered according to the known classes, to allow inspecting if and how the original structure of the data was recovered.

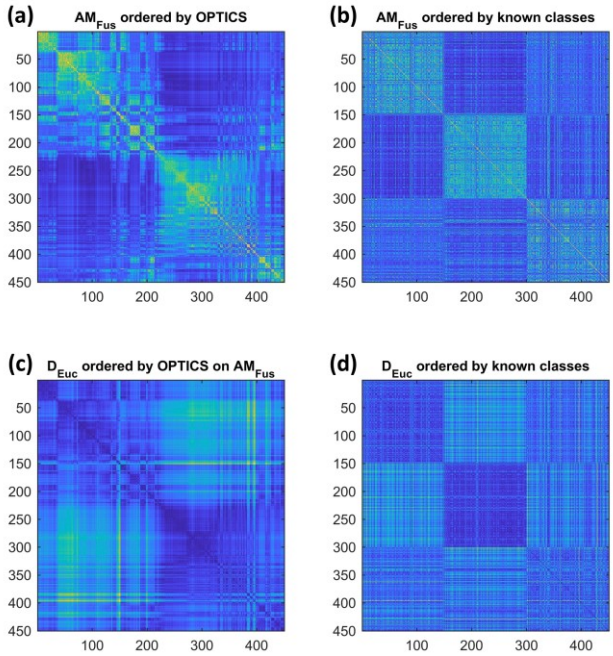


Figure 3.6. Comparison between the Fused adjacency matrix (a-b, AM_{Fus}) and the Euclidean distance matrix (c-d, D_{Euc}) computed from the globular dataset #1.

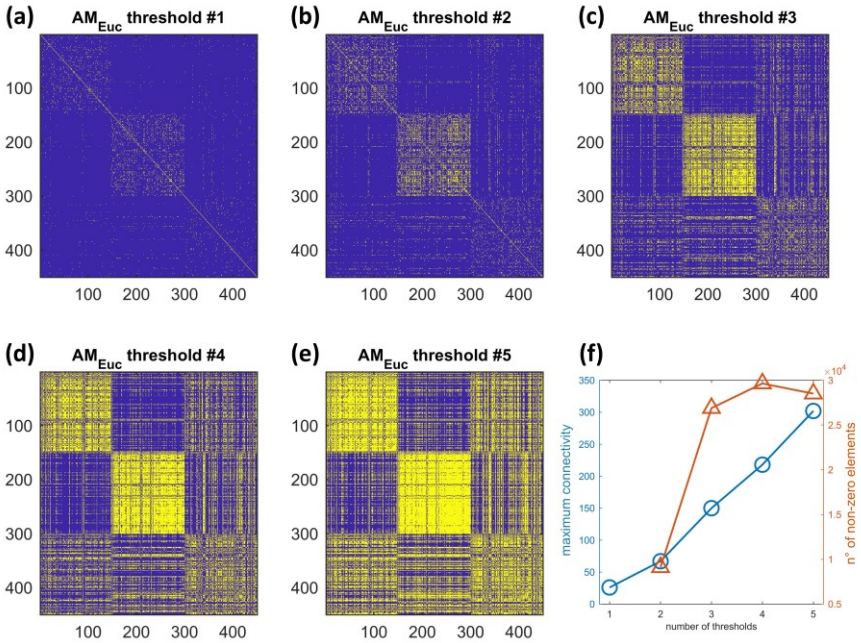


Figure 3.7. Five AMs (a-e) computed from the Euclidean distance matrix, each one corresponding to a threshold value; in (f) maximum connectivity per single sample (○) and sum of non-zero elements in differences of subsequent AM (Δ).

Each AM corresponds to a threshold value belonging to the moving set described in Section 3.5.1. By inspecting the evolution towards increasingly higher threshold values it is possible to deduct the point where the original structure was recovered, which is also the point where the clustering information is expected to start degrading. To this aim, the plot reported in Figure 3.7f is also of help. Depending on the threshold number, the evolution of the two parameters reported in the plot can be used for assessing the optimal threshold value:

- the **maximum connectivity for a single sample** (in blue) at each threshold value is obtained by summing along the rows (or columns, since all these matrices are symmetrical) the corresponding AM, and picking the maximum value; this parameter provides information about the maximum number of connections that any sample of the dataset can have at a given threshold value;
- the **number of non-zero elements computed from the difference between pairs of subsequent AMs** (i.e. corresponding to consecutive threshold values) can provide an indication of whether any changes had occurred moving from one threshold value to the next; this quantity is expected to grow and then stabilize into a plateau, since once any element of the adjacency matrix changes from zero to one (a connection between the corresponding couple of samples is found) its value would remain unchanged for any successive threshold value.

Three seems to be the optimal number of thresholds, given that this value corresponds to the largest change in number of non-zero elements and that the maximum connectivity value is approximately 150, which corresponds to the number of samples belonging to each cluster. By inspecting the AMs, it is worth noting that even though the fourth AM (Figure 3.7.d) highlights in a clearer way the first two clusters, the information about the last one becomes very confused, while the third AM (Figure 3.7c), still keeps it distant from the other two groups. The chosen threshold value is therefore 3.

Moving on to the Mahalanobis distance, the situation reported in Figure 3.8 closely resembles the results obtained in the Euclidean case (Figure 3.6), even if the Mahalanobis distance matrix seems more structured than the Euclidean case (Figure 3.8a). Once again, the third cluster is the most confused with the distance matrix, as it

can be seen in Figure 3.8d, while the Fused approach better highlights the overall clustering structure (Figure 3.8b).

Next, to choose the optimal number of thresholds the information provided by Figure 3.9 can be used. Also, in this case the optimal number of thresholds seems to be three, based on the maximum connectivity value (again approximately equal to the number of expected connected samples within a cluster) and on the change in the number of non-zero elements. The visual inspection of the AMs suggests that either three or four thresholds can be good options. For the sake of simplicity, three thresholds will be used.

Then, the results from the use of the two SOM grids must be evaluated. The results obtained from the rectangular grid are reported in Figures 3.10 and 3.11, while those obtained from the hexagonal grid are reported in Figures 3.12 and 3.13. Since SOM does not provide a distance matrix but instead a top-map, the SOM adjacency matrix was included in Figures 3.10 and 3.12.

It is interesting to notice how similar AM_{SOM} and AM_{Fus} are since this strong resemblance may reveal that SOM had a big influence on the final result. When reordered according to the known classes, both grids seem to be able to recover the clustering structure (Figures 3.10d and 3.12d). Ordering by OPTICS also leads to highlighting the data structure, even if only two clusters are easily recognizable.

The optimal number of neighbourhood levels can be estimated with the same procedure employed for the distance cases. By inspecting Figures 3.11 and 3.13 it can be deduced that the optimal numbers of neighbourhoods for the rectangular and hexagonal grid are, respectively, 6 and 5.

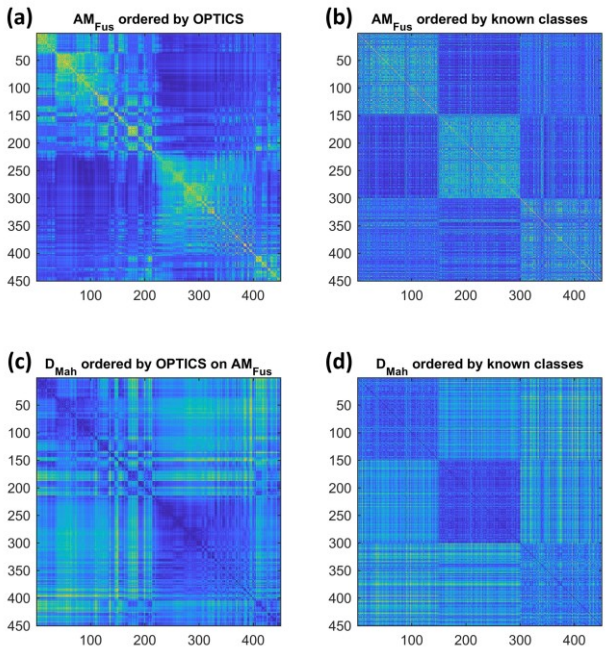


Figure 3.8. Comparison between the Fused adjacency matrix (a-b, AM_{Fus}) and the Mahalanobis distance matrix (c-d, D_{Mah}) computed from the globular dataset #1.

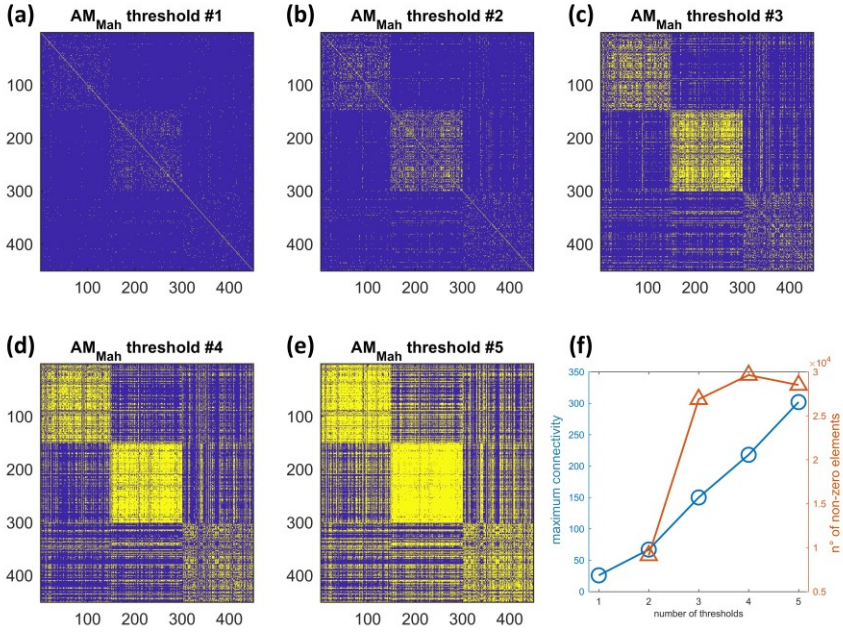


Figure 3.9. Five AMs (a-e) computed from the Mahalanobis distance matrix, each one corresponding to a threshold value; in (f) maximum connectivity per single sample (○) and sum of non-zero elements in differences of subsequent AM (Δ).

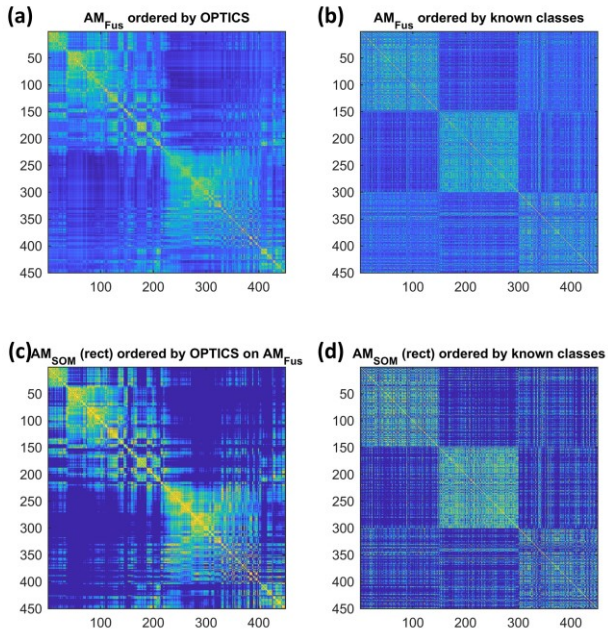


Figure 3.10. Comparison between the Fused adjacency matrix (a-b, AM_{Fus}) and the SOM adjacency matrix (c-d, AM_{SOM}) computed from the globular dataset #1 using a rectangular SOM grid.

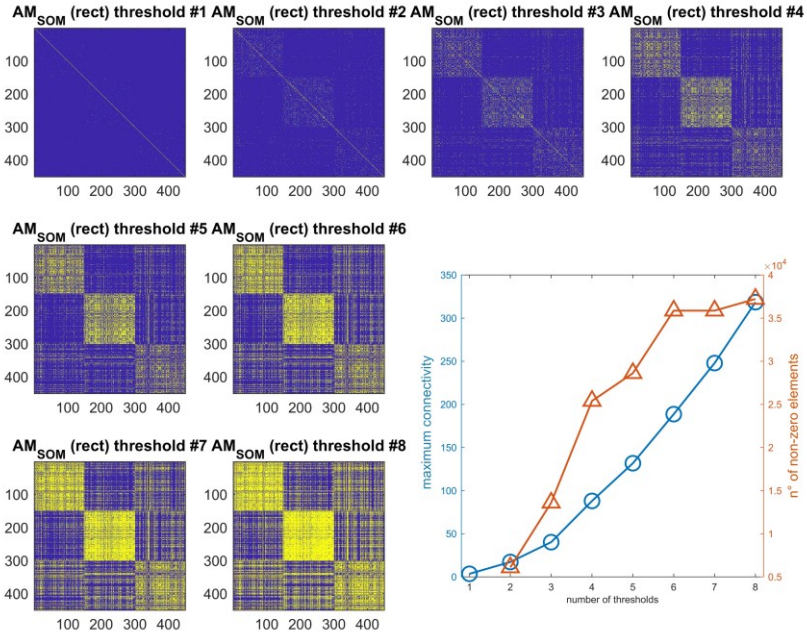


Figure 3.11. Eight AMs computed from the SOM rectangular top-map, each one corresponding to a neighbourhood level; in the bottom-right position, maximum connectivity per single sample (○) and sum of non-zero elements in differences of subsequent AM (Δ).

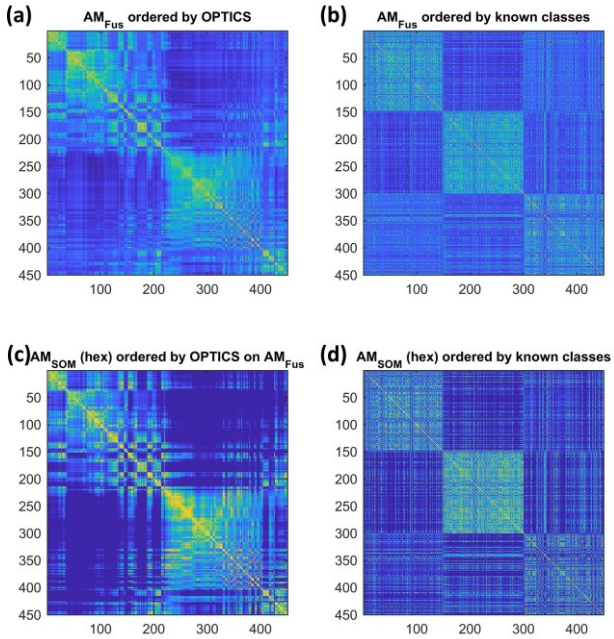


Figure 3.12. Comparison between the Fused adjacency matrix (a-b, AM_{Fus}) and the SOM adjacency matrix (c-d, AM_{SOM}) computed from the globular dataset #1 using a hexagonal SOM grid.

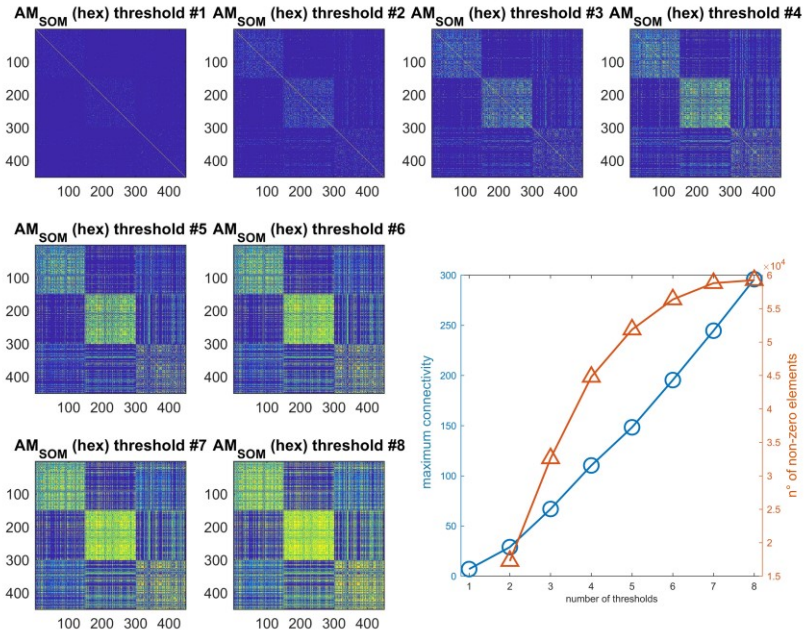


Figure 3.13. Eight AMs computed from the SOM hexagonal top-map, each one corresponding to a neighbourhood level; in the bottom-right position, maximum connectivity per single sample (○) and sum of non-zero elements in differences of subsequent AM (Δ).

The new Fused Adjacency Matrix of the globular dataset #1, was obtained using the manually optimized threshold values/neighbourhoods summarized in Table 3.2.

Table 3.2. Set of optimized parameters for the globular dataset #1.

	Euclidean	Mahalanobis	SOM rectangular	SOM hexagonal
globular #1	3	3	6	5

At the center of Figure 3.14 (either a or b), there is a void: only some samples belonging to Class 2 are present. The samples' distribution seems to be stretched away from the origin of the plot towards three directions of increased samples density. This can also be noticed in the colourless version of the score plot (Figure 3.14a).

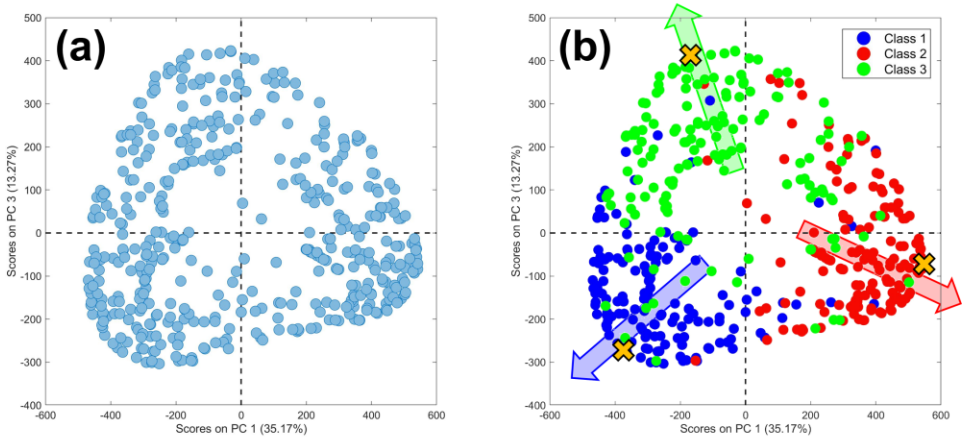


Figure 3.14. PCA score plot (PC1-PC3) of the optimized Fused Adjacency Matrix of the globular dataset #1.

These three directions probably represent the centres of the clusters, and this is confirmed by the fact that the samples closest to their own cluster's center (computed as the mean position of each individual class) almost perfectly lie in the cluster's direction. Those samples are identified by orange crosses on Figure 3.14b.

These results seem therefore to suggest that in the Fused Adjacency Matrix the cluster's centres tend to move further apart, together with the cluster's bulk, which becomes more grouped (i.e. zones of increased density are obtained). It is worth noticing that PC1 and PC3 are the best PCs in relation to the cluster's distinction. PC2 plotted against either PC1 or PC3 results somehow confused, but when it is plotted in a 3D score plot

together with these PCs (Figure 3.15), the situation described in Figure 3.14 becomes even clearer. Finally, Figure 3.16 shows how OPTICS is able to almost perfectly distinguish between class 2 (in red) and class 1 (in blue), but class 3 (in green) is the most fragmented.

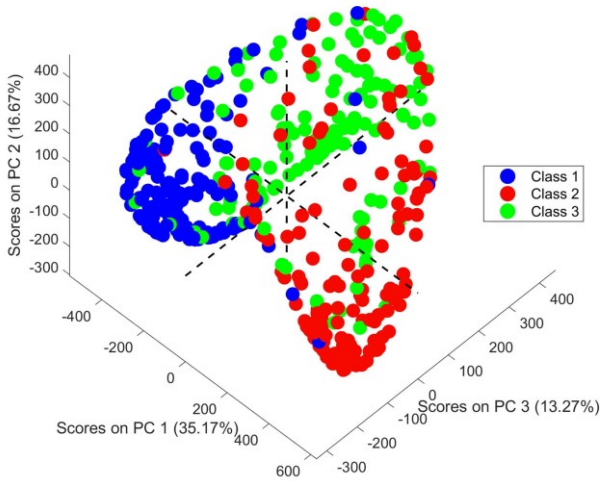


Figure 3.15. PCA score plot (PC1-PC2-PC3) of the optimized Fused Adjacency Matrix of the globular dataset #1.

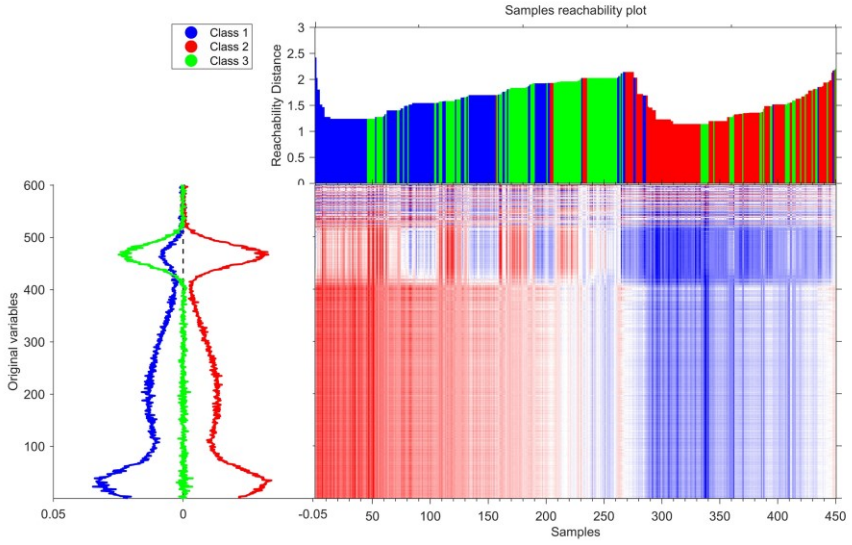


Figure 3.16. Reachability plot and reordered heatmap obtained from the optimized Fused adjacency matrix of the globular dataset #1. On the left, the class-average signals.

3.6.1.2. Globular clusters dataset #2

The globular dataset #2 consists of three overlapped clusters (Figure 3.17) of which the central one (class 2, in red) is by design very mixed with the other two. By the same approach as globular dataset #1, the optimal set of threshold values/neighbourhoods for the globular dataset #2 was chosen.

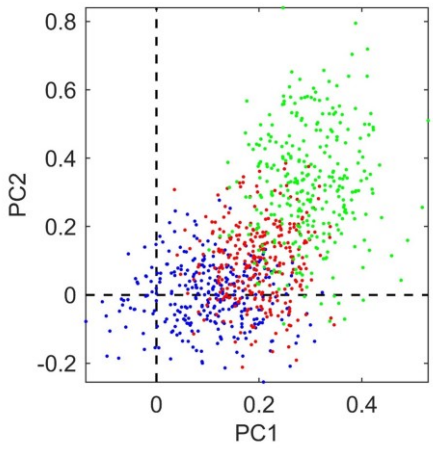


Figure 3.17. The original distribution of the globular dataset #2.

Starting from the Euclidean distance matrix (Figure 3.18), it is possible to notice that reordering by OPTICS obtained from the Fused adjacency matrix allows the identification of one cluster, while the remaining two still result rather overlapped. When reordering according to known classes is performed (Figure 3.18b-d), the clustering structure becomes clearer and the second class results the less dense and, as expected, the most overlapped with the other two, and also. Three Euclidean thresholds seems to be a good compromise between a maximum connectivity value of about 300 and a good change in non-zero elements (Figure 3.19f). Moreover, in the structure highlighted by the AM of the fourth threshold (Figure 3.19d) the overlap between class 1 and class 2 strongly increases, therefore choosing three thresholds seems the most reasonable option.

The Mahalanobis distance looks similar to the Euclidean case, but more structure seems to be recovered by the distance matrix (Figure 3.20c). Again, class 2 is the most overlapped, and the cluster information becomes confused when the fourth threshold

is reached (Figure 3.19d). The maximum connectivity value and the number of non-zero elements also point to choosing three thresholds for the Mahalanobis distance.

Since the globular dataset #2 consists of 900 samples, the chosen SOM grid's dimensions were larger than globular dataset #1, also meaning that more threshold values were used. Figure 3.23 shows the twelve automatically defined thresholds. In this case, either seven or eight thresholds seem good options, even if seven is probably a better choice, since the clustering structure highlighted by the corresponding AM (#7) reveals that class 2 is less connected to the elements of the other classes, if compared to the next AM (#8). The maximum connectivity value is closer to 250 than 300, but this is probably still a good compromise, considered that even if the samples are expected to be connected to about 300 other samples, preferably belonging to the same cluster, in a strongly overlapped structure this value can be set a bit lower.

The same considerations can be made with the hexagonal grid, for which six thresholds seem a good compromise, corresponding to a good AM structure (Figure 3.25, AM threshold #6) and to a connectivity value lower than 300.

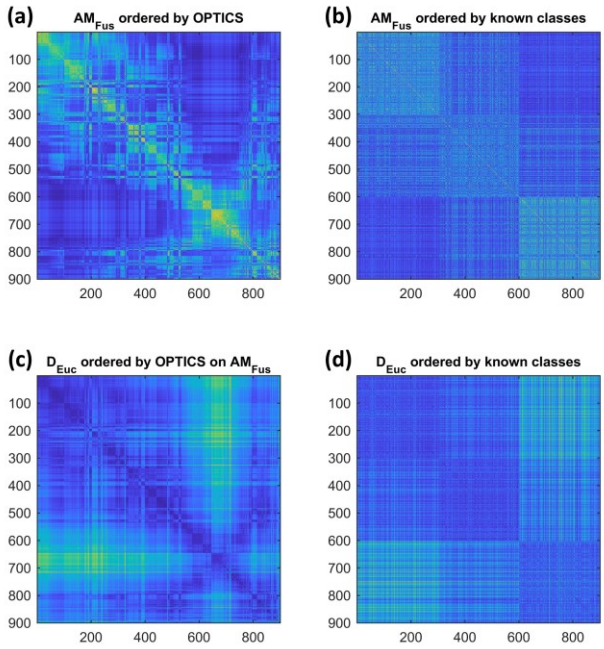


Figure 3.18. Comparison between the Fused adjacency matrix (a-b, AM_{Fus}) and the Euclidean distance matrix (c-d, D_{Euc}) computed from the globular dataset #2.

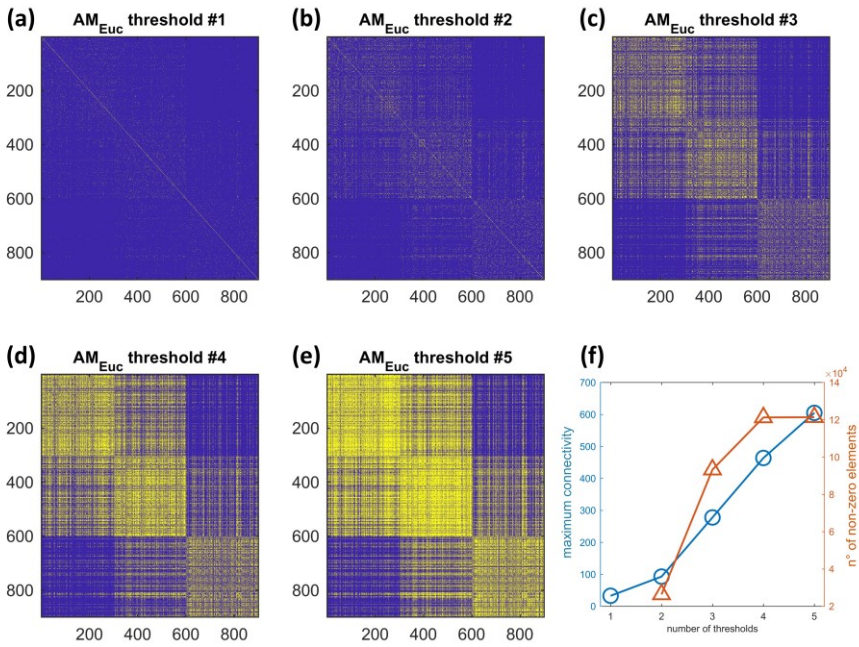


Figure 3.19. Five AMs (a-e) computed from the Euclidean distance matrix, each one corresponding to a threshold value; in (f) maximum connectivity per single sample (○) and sum of non-zero elements in differences of subsequent AM (Δ).

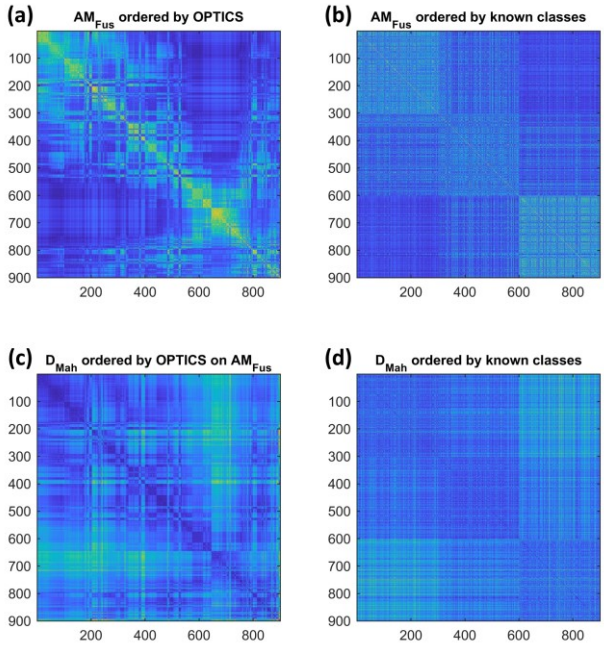


Figure 3.20. Comparison between the Fused adjacency matrix (a-b, AM_{Fus}) and the Mahalanobis distance matrix (c-d, D_{Mah}) computed from the globular dataset #2.

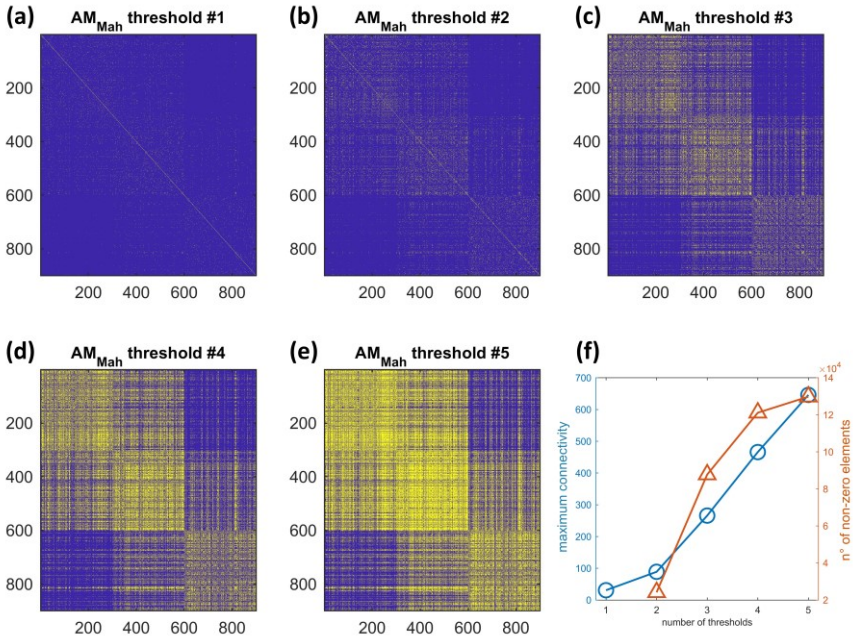


Figure 3.21. Five AMs (a-e) computed from the Mahalanobis distance matrix, each one corresponding to a threshold value; in (f) maximum connectivity per single sample (○) and sum of non-zero elements in differences of subsequent AM (Δ).

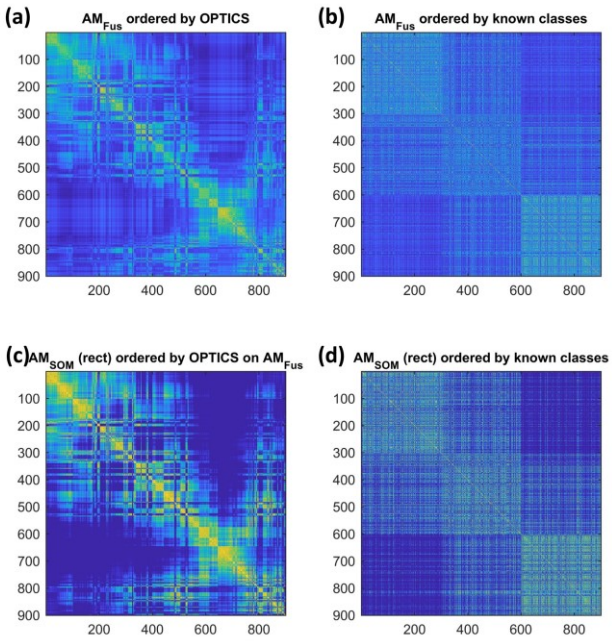


Figure 3.22. Comparison between the Fused adjacency matrix (a-b, AM_{Fus}) and the SOM adjacency matrix (c-d, AM_{SOM}) computed from the globular dataset #2 using a rectangular SOM grid.

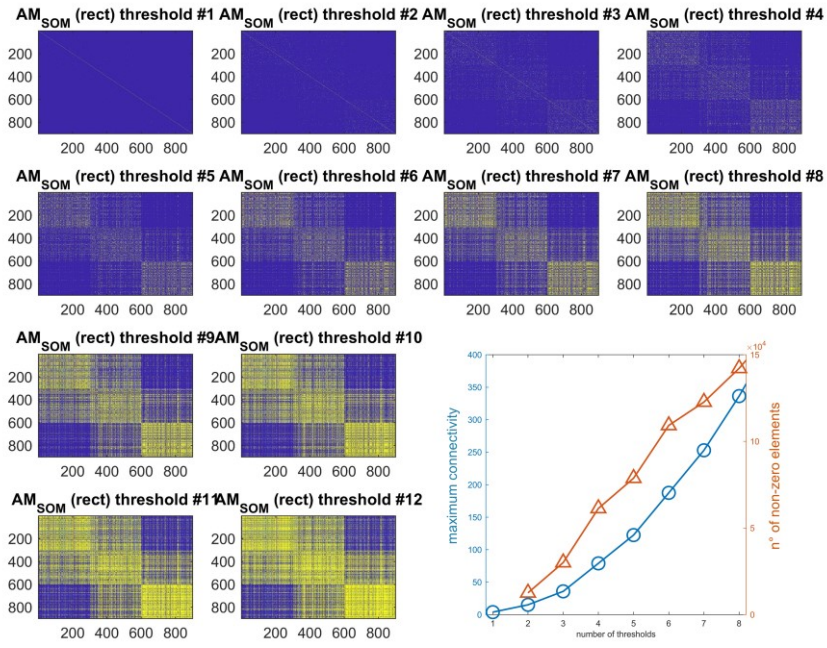


Figure 3.23. Eight AMs computed from the SOM rectangular top-map, each one corresponding to a neighbourhood level; in the bottom-right position, maximum connectivity per single sample (○) and sum of non-zero elements in differences of subsequent AM (Δ).

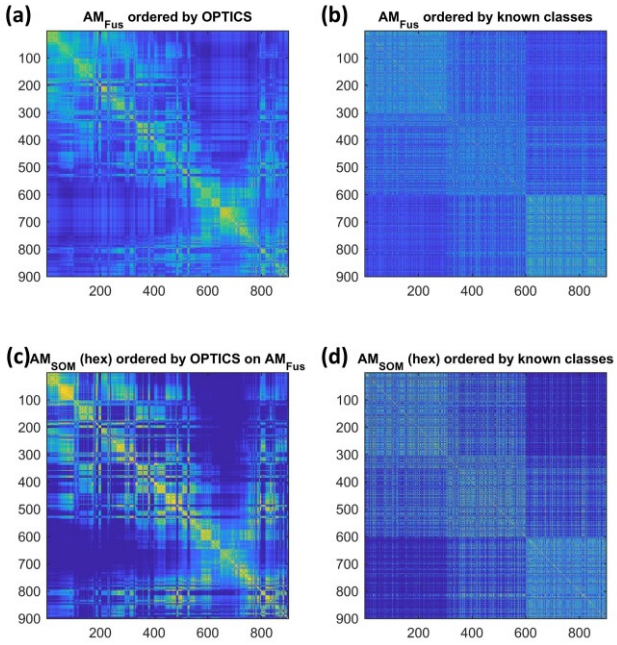


Figure 3.24. Comparison between the Fused adjacency matrix (a-b, AM_{Fus}) and the SOM adjacency matrix (c-d, AM_{SOM}) computed from the globular dataset #2 using a hexagonal SOM grid.

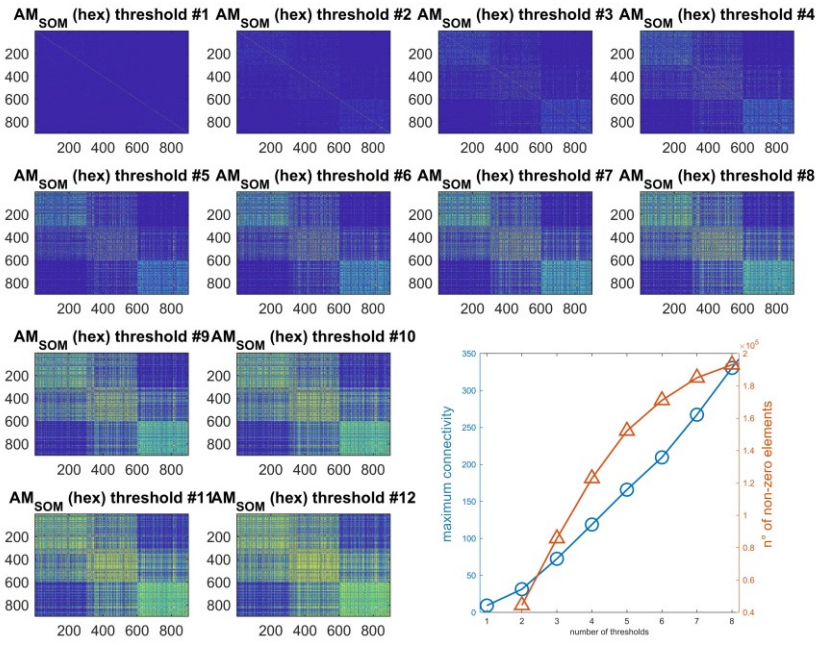


Figure 3.25. Twelve AMs computed from the SOM hexagonal top-map, each one corresponding to a neighbourhood level; in the bottom-right position, maximum connectivity per single sample (○) and sum of non-zero elements in differences of subsequent AM (Δ).

The new Fused Adjacency Matrix of the globular dataset #2, was obtained using the manually optimized threshold/neighbourhood values summarized in Table 3.3.

Table 3.3. Set of optimized parameters for the globular dataset #2.

	Euclidean	Mahalanobis	SOM rectangular	SOM hexagonal
globular #2	3	3	7	6

The new optimized Fused adjacency matrix was inspected by PCA and OPTICS (Figure 3.26 and 3.27, respectively). The distribution highlighted by PC1 and PC2 looks similar to the one obtained with the globular dataset #1 (Figure 3.15) The samples' distribution in the PCA, but it is, as expected, more confused. Class 2, the central and most overlapped class, is not grouped, while class 1 and 3 tend to group into dense "lobes", still following the cluster centres (represented by orange crosses in Figure 3.26). The center of class 2 has moved as well, but the cluster has remained rather disperse.

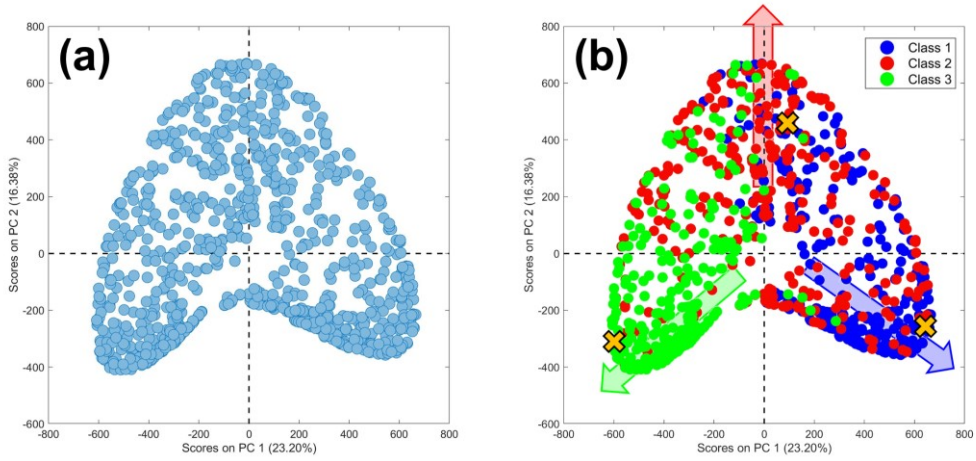


Figure 3.26. PCA score plot (PC1-PC2) of the optimized Fused Adjacency Matrix of the globular dataset #2.

The results obtained by OPTICS on the optimized Fused adjacency matrix are shown in Figure 3.27: as anticipated by the PCA score plot, class 2 (in red) does not get grouped, while two zones of increased sample density are found for classes 1 and 3.

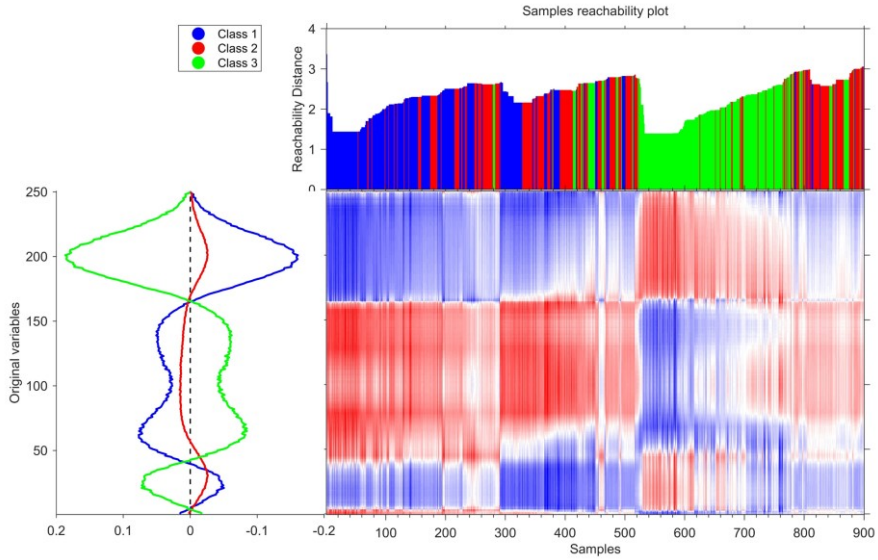


Figure 3.27. Reachability plot and reordered heatmap obtained from the optimized Fused adjacency matrix of the globular dataset #2. On the left, the class-average signals.

3.6.1.3. Circles dataset

The circles dataset consists of three unbalanced clusters (Figure 3.28). The optimal set of threshold values/neighbourhoods was chosen by the same approach as globular datasets #1 and #2.

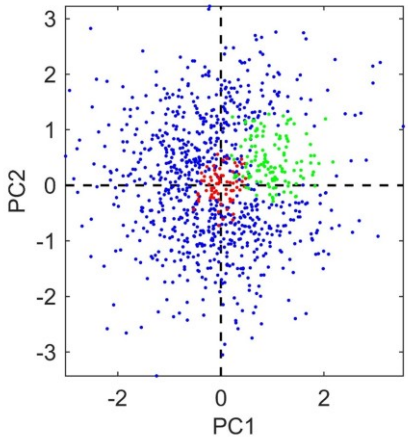


Figure 3.28. The original distribution of the circles dataset.

Reordering by OPTICS on the Fused Adjacency Matrix seems to highlight substructures that are weakly related to the three designed clusters. The best match is with class 3, which corresponds to the square within the interval 300–400 of Figure 3.30a. The last, weak group of samples at the end of Figures 3.29a and 3.29c can be linked to class 2.

Concerning the choice of the number of thresholds to be used in the Euclidean case, three seems the most reasonable option: from the corresponding AM (Figure 3.30c) it can be noticed how the small class 2 cluster has become very linked, and the structure of class 3 is recognized as well. With the fourth threshold value class 2 becomes much more connected to the other two, without a substantial gain with regard to class 1, whose structure is weakly recovered. Once again, the Mahalanobis distance (Figures 3.31 and 3.32) resembles the Euclidean case, and in this case as well three thresholds seem to be the best compromise.

Since the circles dataset consists of 1000 samples, the automatically determined dimensions of the SOM grid led to defining twelve threshold values/neighbourhoods:

Figure 3.34 shows the AMs corresponding to the twelve thresholds. For the same reasons as the Euclidean and Mahalanobis distances, the best option seems to be threshold #6, as the structure highlighted by the AM is strong enough for class 2 and class 3, even if it is still weak for class 1. Concerning the hexagonal SOM grid, the sixth threshold value also seems to be the best option. Maximum connectivity is lower than 250 (Figure 3.36), which is probably still rather high in relation to the unbalanced numbers of samples belonging to the three clusters, but the structure highlighted by the corresponding AM (Figure 3.36) shows a good balance between the gain in structure of class 1 and the not too high connections among the different clusters.

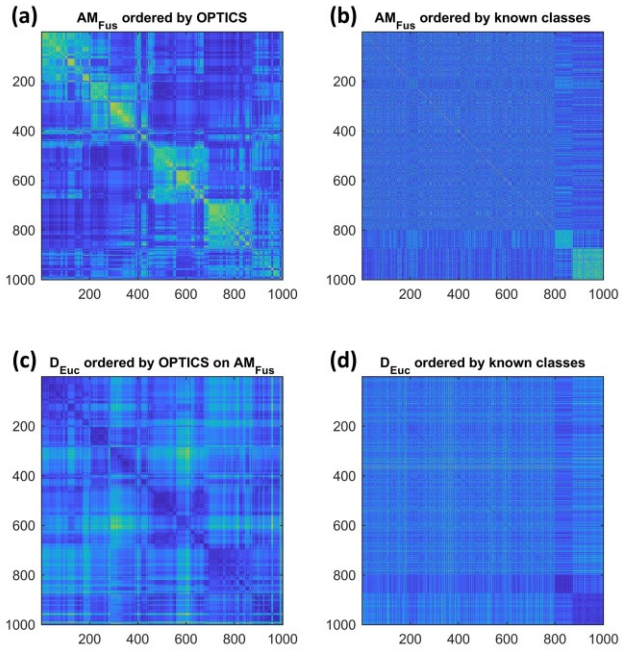


Figure 3.29. Comparison between the Fused adjacency matrix (a-b, AM_{Fus}) and the Euclidean distance matrix (c-d, D_{Euc}) computed from the circles dataset.

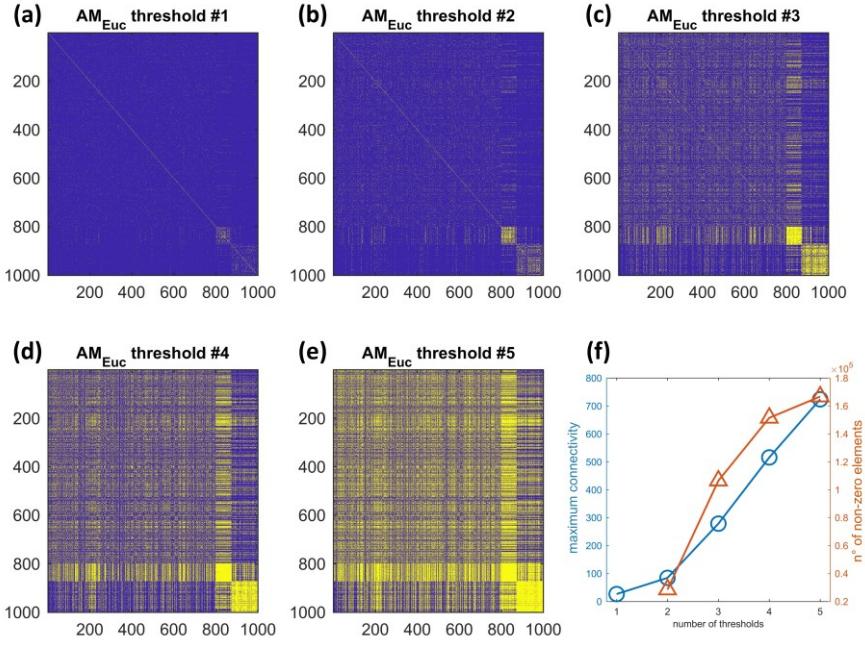


Figure 3.30. Five AMs (a-e) computed from the Euclidean distance matrix, each one corresponding to a threshold value; in (f) maximum connectivity per single sample (○) and sum of non-zero elements in differences of subsequent AM (Δ).

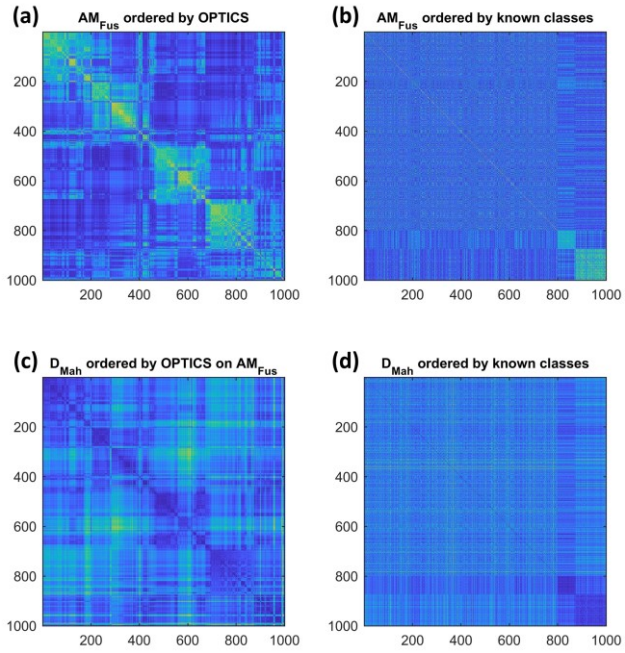


Figure 3.31. Comparison between the Fused adjacency matrix (a-b, AM_{Fus}) and the Mahalanobis distance matrix (c-d, D_{Mah}) computed from the circles dataset.

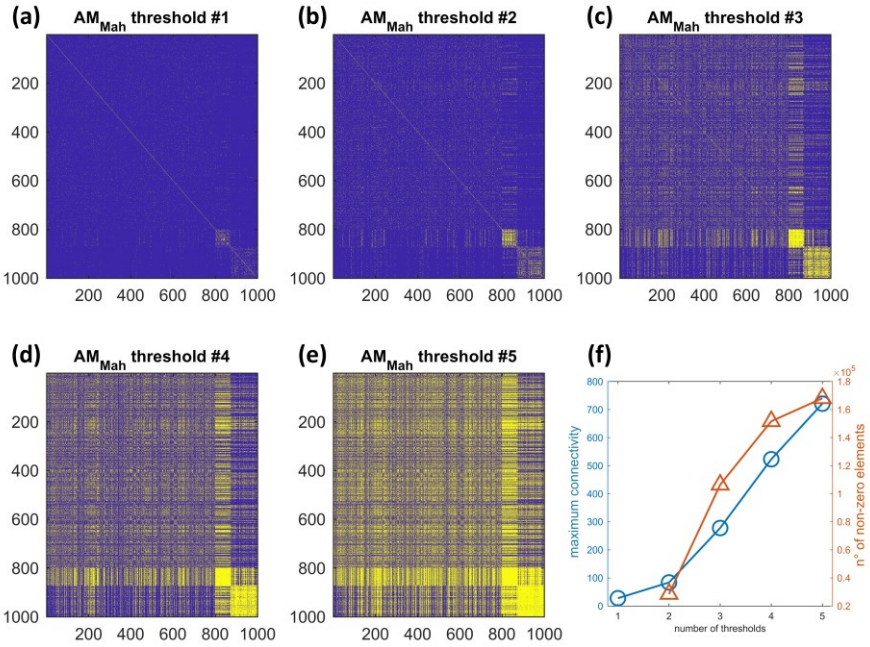


Figure 3.32. Five AMs (a-e) computed from the Mahalanobis distance matrix, each one corresponding to a threshold value; in (f) maximum connectivity per single sample (○) and sum of non-zero elements in differences of subsequent AM (Δ).

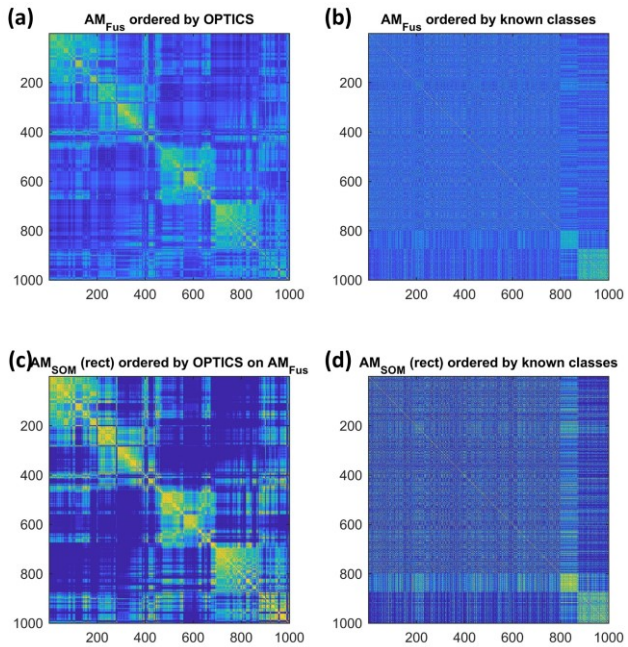


Figure 3.33. Comparison between the Fused adjacency matrix (a-b, AM_{Fus}) and the SOM adjacency matrix (c-d, AM_{SOM}) computed from the circles using a rectangular SOM grid.

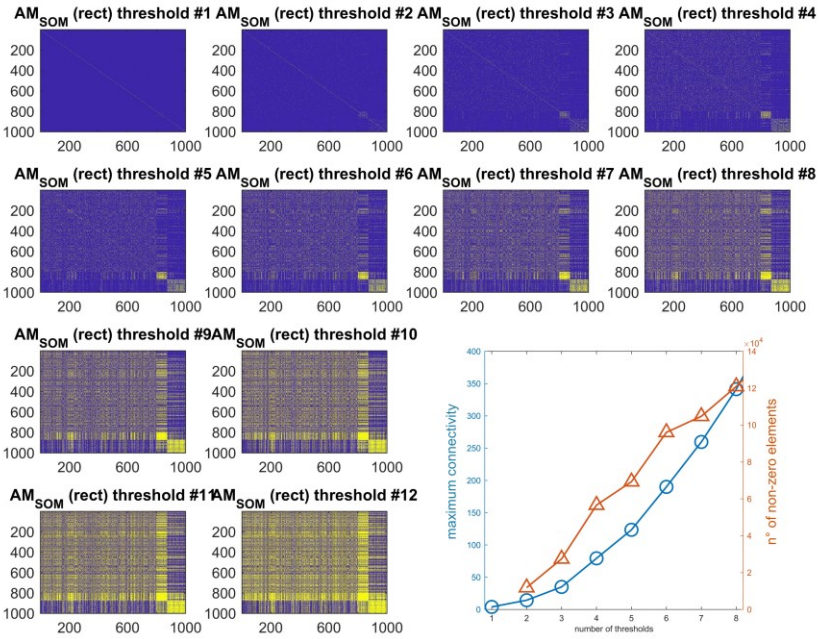


Figure 3.34. Twelve AMs computed from the SOM rectangular top-map, each one corresponding to a neighbourhood level; in the bottom-right position, maximum connectivity per single sample (○) and sum of non-zero elements in differences of subsequent AM (Δ).

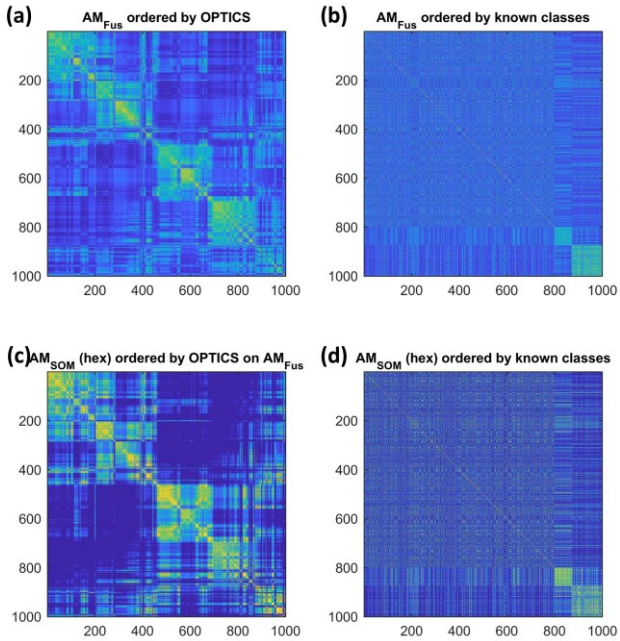


Figure 3.35. Comparison between the Fused adjacency matrix (a-b, AM_{Fus}) and the SOM adjacency matrix (c-d, AM_{SOM}) computed from the circles using a hexagonal SOM grid.

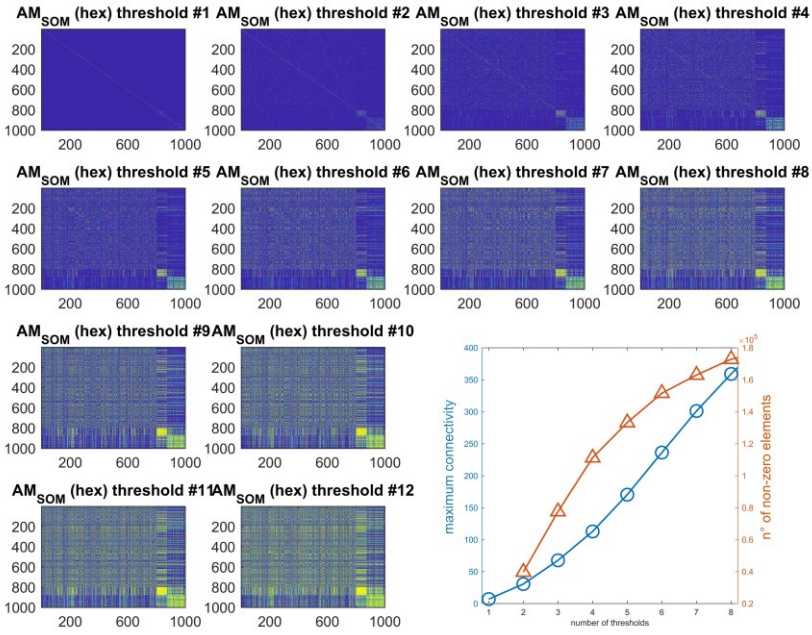


Figure 3.36. Twelve AMs computed from the SOM hexagonal top-map, each one corresponding to a neighbourhood level; in the bottom-right position, maximum connectivity per single sample (○) and sum of non-zero elements in differences of subsequent AM (Δ).

The new Fused Adjacency Matrix of the circles dataset, was obtained using the manually optimized threshold/neighbourhood values summarized in Table 3.4.

Table 3.4. Set of optimized parameters for the circles dataset.

	Euclidean	Mahalanobis	SOM rectangular	SOM hexagonal
circles	3	3	6	6

No direct information about the clusters could be detected by PCA on the new optimized Fused adjacency matrix. The best PC combinations showing classes 2 and 3 are reported in Figure 3.37. Figure 3.37a is referred to the combination of PC2, PC8 and PC10 which provides a rather clear visualization of class 3 (in green), which is found in a lobe of the samples' distribution. Class 2 (which was located at the origin of the original scores distribution, as represented in Figure 3.28) is also the less numerous cluster. The PC3-PC4 score plot (Figure 3.37b) reveals that the cluster's position, with its center represented by an orange cross.

However, these considerations can be made only with a priori knowledge – i.e. the classes are known – therefore with similarly structured datasets the Fused adjacency matrix approach may not yield optimal results.

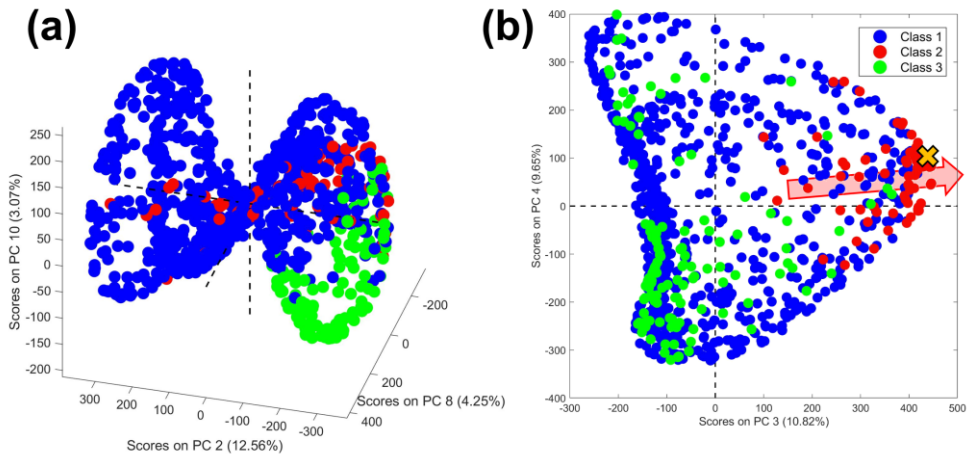


Figure 3.37. PCA score plots of the optimized Fused Adjacency Matrix of the circles dataset. Score plot (a) shows the combination of PC2, PC8 and PC10 that better highlights class 3, while score plot (b) shows the combination of PC3 and PC4 that better highlights class 2.

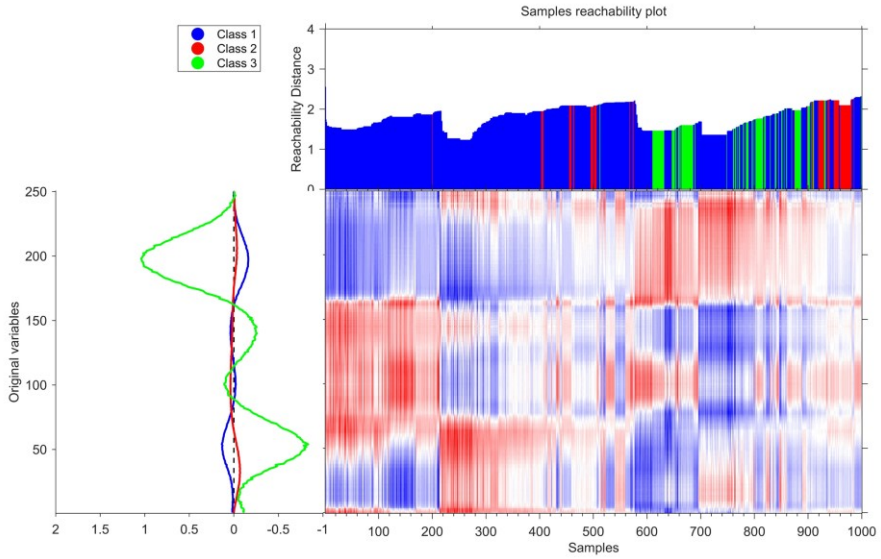


Figure 3.38. Reachability plot and reordered heatmap obtained from the optimized Fused adjacency matrix of the circles dataset. On the left, the class-average signals.

By inspecting the reachability plot of Figure 3.38, it can be noticed that the structures highlighted by OPTICS are weakly related to the designed clustering structure, even if class 3 (in green) is somehow confined within a certain range, as well as most samples of class 2, located at the end of the plot.

3.6.1.4. t48k dataset

The t48k simulated dataset is a "perfect" dataset for density-based methods like OPTICS and DBSCAN, because of its clustering structure (Figure 3.39). Indeed, as represented in Figure 3.43, the best results are obtained by running OPTICS on the raw data. This was used as a benchmark for the performance assessment of the Fused Adjacency Matrix method.

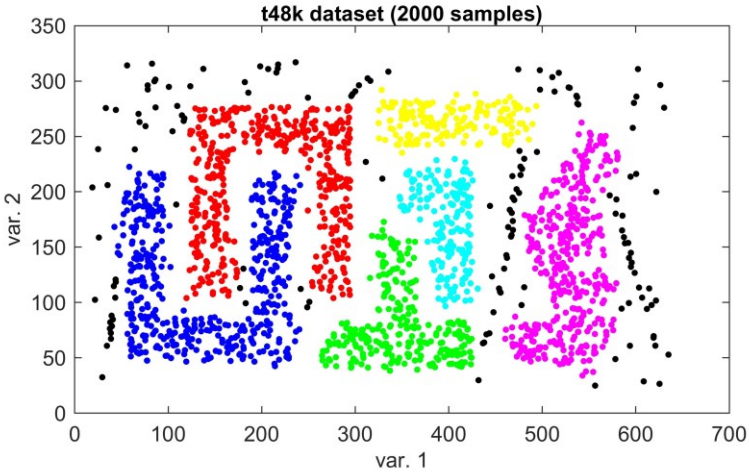


Figure 3.39. The t48k dataset, raw data and classes.

Choosing the number of thresholds for each distance measure was rather straightforward with the t48k dataset. As it can be seen in Figures 3.40d and 3.42d, both distance matrices, if reordered according to the known classes, highlight clear structures. By inspecting the AMs (Figures 3.41 and 3.43), the situation becomes even clearer: one threshold is indeed enough to capture the clustering structure both in the Euclidean and the Mahalanobis cases.

In SOM, a similar situation is found, even though there are some differences between the two types of grid. With the rectangular grid any threshold/neighbourhood lower than 5 seems to be fine. Therefore, for the sake of simplicity, only the first neighbourhood was selected. Also, for the hexagonal grid only the first threshold was selected, because the first two were identical, and something changed once the third threshold was reached.

This behaviour is linked to the nature of SOM: by sampling unevenly the original space and focusing on the zones with higher sample density, SOM was probably able to condense each cluster on few adjacent nodes, mapping the nodes further apart from each other. For this reason, in the first and closest neighbourhoods no sample “meets” samples from any other cluster than its own.

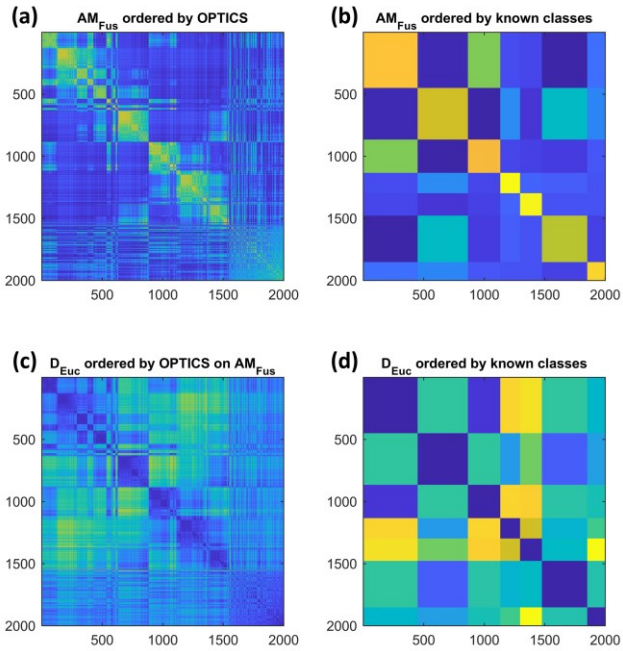


Figure 3.40. Comparison between the Fused adjacency matrix (a-b, AM_{Fus}) and the Euclidean distance matrix (c-d, D_{Euc}) computed from the t48k dataset.

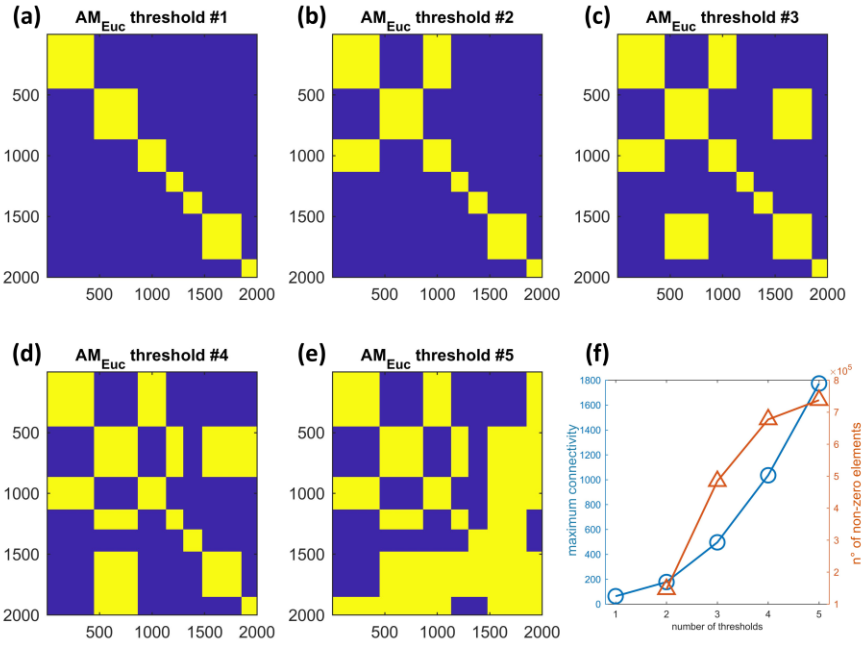


Figure 3.41. Five AMs (a-e) computed from the Euclidean distance matrix, each one corresponding to a threshold value; in (f) maximum connectivity per single sample (○) and sum of non-zero elements in differences of subsequent AM (Δ).

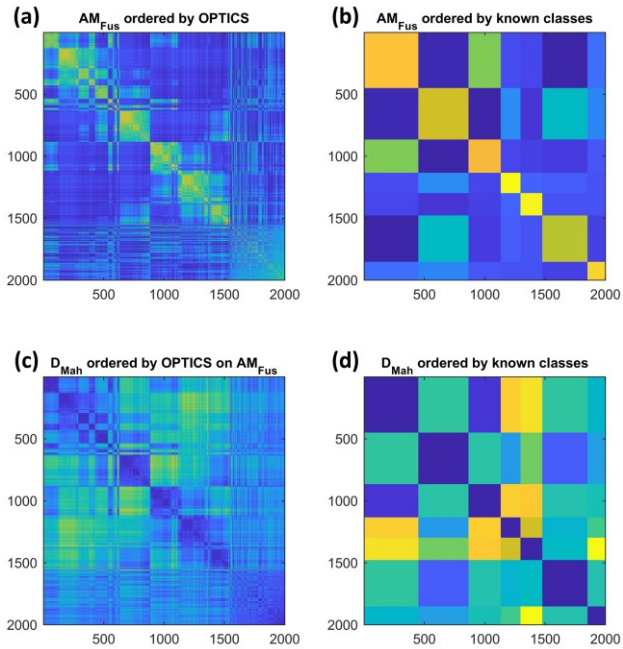


Figure 3.42. Comparison between the Fused adjacency matrix (a-b, AM_{Fus}) and the Mahalanobis distance matrix (c-d, D_{Mah}) computed from the t48k dataset.

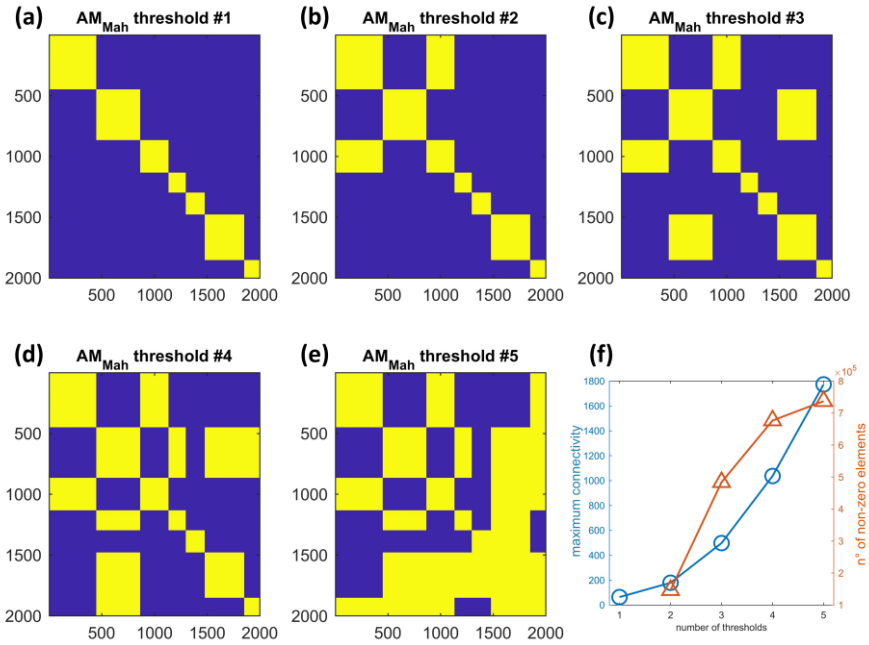


Figure 3.43. Five AMs (a-e) computed from the Mahalanobis distance matrix, each one corresponding to a threshold value; in (f) maximum connectivity per single sample (○) and sum of non-zero elements in differences of subsequent AM (Δ).

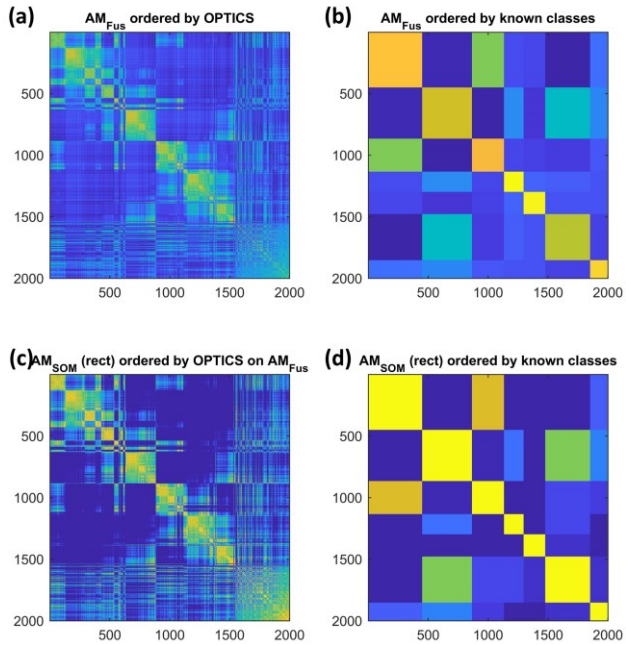


Figure 3.44. Comparison between the Fused adjacency matrix (a-b, AM_{Fus}) and the SOM adjacency matrix (c-d, AM_{SOM}) computed from the circles using a rectangular SOM grid.

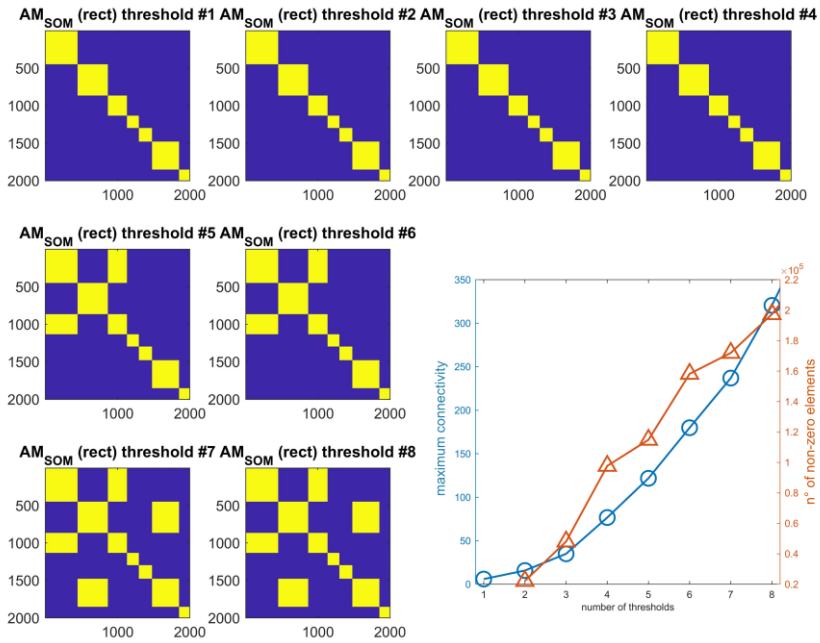


Figure 3.45. Eight AMs computed from the SOM rectangular top-map, each one corresponding to a neighbourhood level; in the bottom-right position, maximum connectivity per single sample (○) and sum of non-zero elements in differences of subsequent AM (Δ).

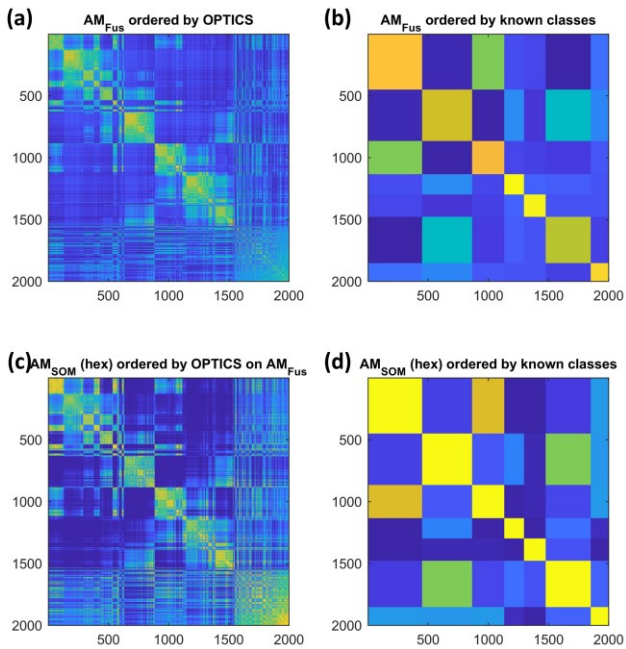


Figure 3.46. Comparison between the Fused adjacency matrix (a-b, AM_{Fus}) and the SOM adjacency matrix (c-d, AM_{SOM}) computed from the circles using a hexagonal SOM grid.

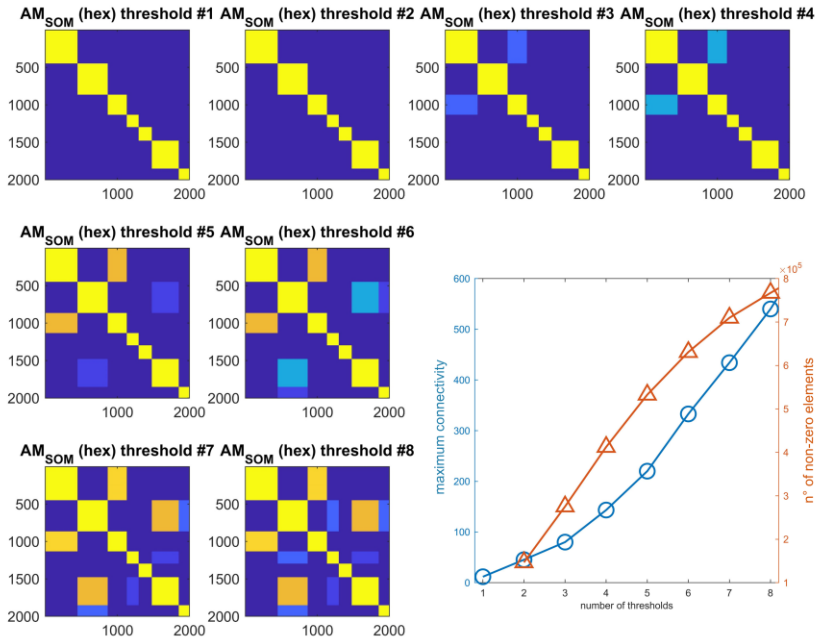


Figure 3.47. Eight AMs computed from the SOM hexagonal top-map, each one corresponding to a neighbourhood level; in the bottom-right position, maximum connectivity per single sample (○) and sum of non-zero elements in differences of subsequent AM (Δ).

The new Fused adjacency matrix of the t48k dataset, was obtained using the manually optimized threshold/neighbourhood values summarized in Table 3.5.

Table 3.5. Set of optimized parameters for the t48k dataset.

	Euclidean	Mahalanobis	SOM rectangular	SOM hexagonal
t48k	1	1	1	1

Figure 3.48 show the optimized Fused Adjacency Matrix (b) reordered according to its OPTICS sequence, which is also represented by the reachability plot (a). It is clearly visible that the highlighted structures correspond to the known classes to a very good extent.

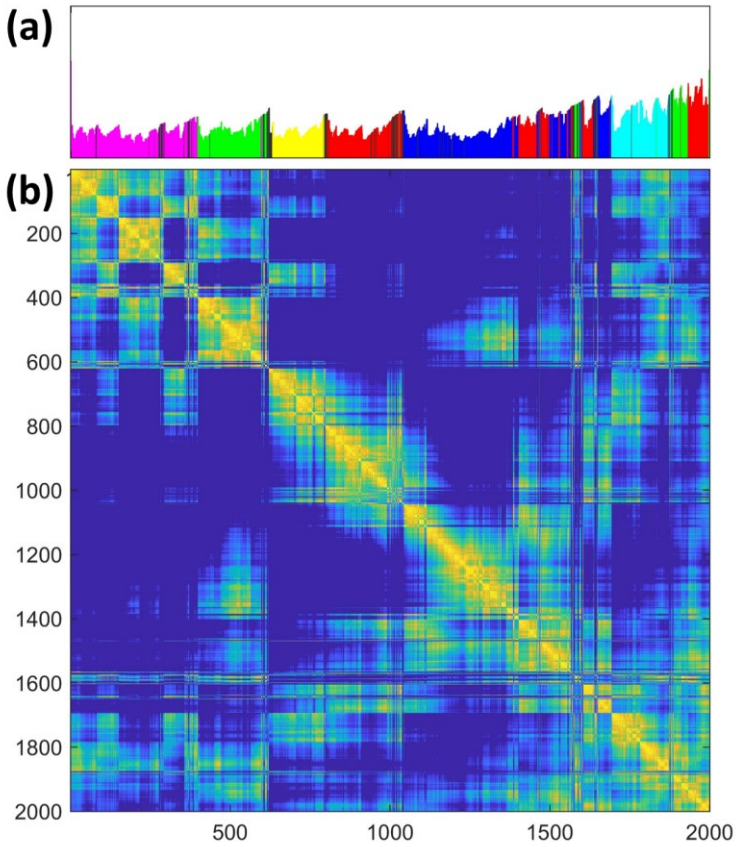


Figure 3.48. (b) Optimized Fused adjacency matrix reordered according to its (a) OPTICS reachability plot (k = 10).

OPTICS with k = 10 was chosen because of the strong clustering tendency of the dataset, where the nearest neighbours of each sample in a cluster is very likely that still belongs to that cluster: a low value of k ensures that a small local neighbourhood of each sample is explored during OPTICS.

The reachability plots of all inspected matrices are reported in Figure 3.49. It is very clear that the raw data and the distance matrices (D_{Euc} and D_{Mah}) perform equally well, being practically indistinguishable and therefore equally able to identify the clusters. Both distance AM (AM_{Euc} and AM_{Mah}) then look pretty much the same as well.

The Fused adjacency matrix actually performed well, compared to the benchmark represented by OPTICS on the raw data. The groups that were obtained are compact and only three clusters got split during the analysis (groups in green, red and blue in Figure 3.49). Even if the non-grouped set (bars in black in Figure 3.49) was not identified like in the raw case, just a few of these rogue samples fall within the recognized clusters. Many of these rogue samples happen to be however usually very close to the clusters the fall in.

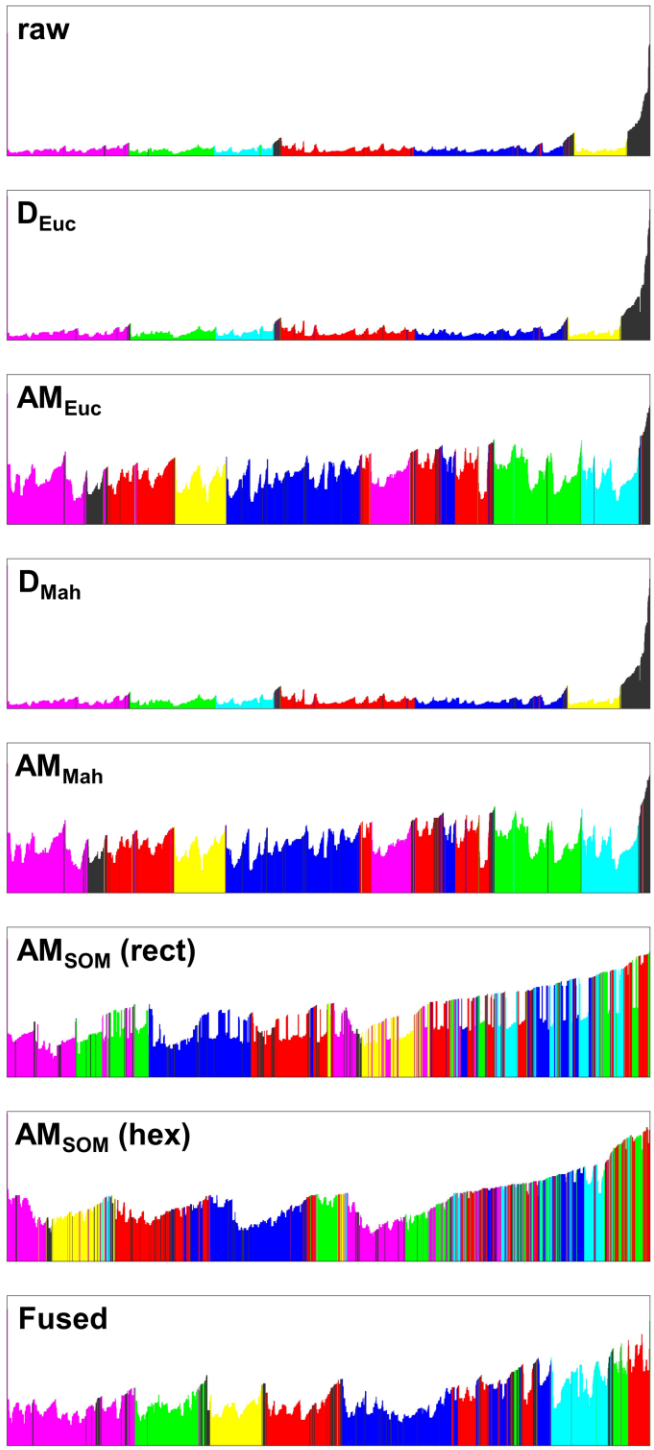


Figure 3.49. OPTICS reachability plots from the t48k dataset (k = 10).

3.6.2. Whisky dataset

This section is devoted to reporting and comparing the whisky raw data results and the results obtained applying the Fused Adjacency Matrix approach.

3.6.2.1. Coclustering results

Twenty coclusters were extracted by means of SMR (Section 2.1.1.4), and the raw data were preprocessed using *unit variance scaling* [22]. The analysis will only be focused on coclusters containing two or more samples, as those with one sample alone may be too specific for the purpose of exploratory analysis. Table 3.6 reports the coclusters of interest for the exploratory analysis of the whisky paragraphs.

Table 3.6. Inspected coclusters rom the whisky data.

cocluster	samples	most important variables
1	ARD_1, BRI_1, BPE_1	2-benzothiophene, naphthene, 1-benzofuran, 1-indene, 1,2-dimethylcyclopent-2-ene-1-carboxylic acid, 3-ethylcyclopentan-1-one, 2-methylcyclopentan-1-one
2	ARD_1, BRI_1, BRI_2, CHI_1, GFA_3, GFA_4, GLI_1, LOC_1, NIK_3, TOM_2, TOM_3	1-ethyl-3,5-dimethylbenzene, 1-ethyl-2-methylbenzene, 1-ethyl-3-methylbenzene, 1,2-xylene, 1,3-xylene, propylbenzene, 1,2,3,5-tetramethylbenzene, 1,2,4,5-tetramethylbenzene, 1,2,3,4-tetrahydronaphthalene, 1,2,3,5-tetrahydronaphthalene
11	JDA_1, JDA_2, JDA_3, WIL_1	2-methylpropyl acetate, 1-sulfanylpropan-2-one, 3-methylbut-3-enenitrile, 3-methylbutyl propanoate, 1,1-diethoxy-2-methylpropane, heptyl acetate, octan-2-one, methyl 14-(2-octylcyclopropyl)tetradecanoate

3.6.2.2. Country of origin

The first clear feature that is retained in both the raw data and the Fused Adjacency Matrix is the group of three Jack Daniel's samples. These three samples are grouped together in all the score plots reported in Figure 3.50 (sub-figures a, c and d, highlighted in orange).

Their clear grouping tendency and separation from the bulk can be explained considering the mash bill used to brew these whisky products: the production of Jack Daniel's envisages the use of corn, rye and barley, as opposed to the use of barley alone for the large majority of the remaining (mostly European) samples. Moreover, one of these products from Jack Daniel's is the traditional and very famous "Old No. 7", while the other two are derived from it by finishing them in slightly different way right before filling the casks. From the point of view of their "overall chemistry" they look therefore almost the same.

Following the grouping tendency of the Jack Daniel's samples, the Fused approach also provides a more general separation tendency, related to the American whiskeys (Figure 3.50d). The other two American samples are found close to the bulk of samples, but in the direction of the Jack Daniel's group. This tendency seems to be strongly determined by the mash bill, i.e. the type of malt. This piece of information is also contained in cocluster #11 (Table 3.6), where the only missing American sample is the farthest from the Jack Daniel's group.

Moving back to Europe, almost all samples from Ireland are better grouped by the Fused Adjacency Matrix approach (Figure 3.50c) than the raw data (Figure 3.50a).

The three Japanese whiskies are found among the Scotch samples both in the raw data and by the Fused Adjacency Matrix. This suggests a close similarity between the narrow selection from Japan and the large collection of samples from Scotland. Another fact supporting this chemical similarity comes from the history of the Nikka distillery, as its founder studied malt whisky production in Scotland[2]. It is very interesting that products from very far and different locations appear so similar, from the point of view of their chemistry.

Finally, the two Indian samples are found closer by the Fused approach (Figure 3.50c) than in the raw data (Figure 3.50a). This difference may mean that a piece of information was lost, or that the two samples are much more similar than what could be deduced by looking at the raw data alone.

[2] https://www.nikka.com/eng/story/history/

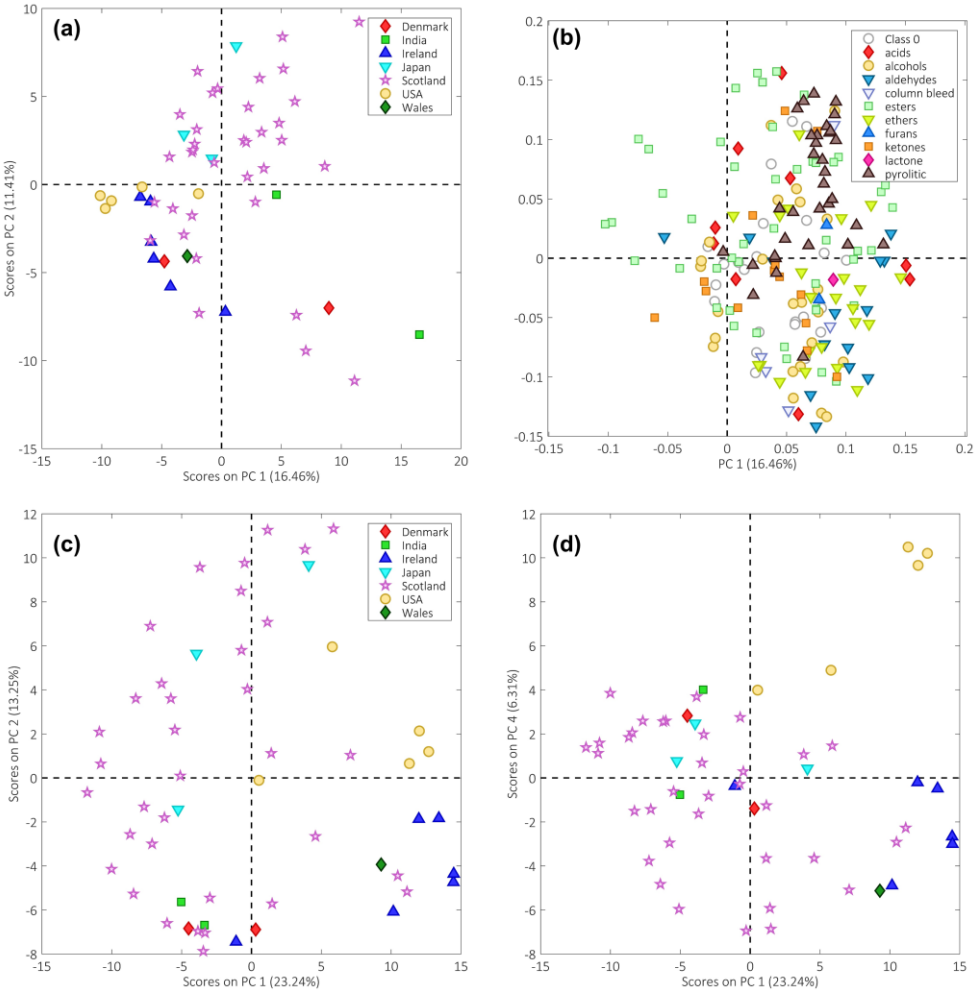


Figure 3.50. PC1-PC2 score plot (a) and loading plot (b) of the raw data; score plots of the Fused Adjacency Matrix (c, PC1-PC2; d, PC1-PC4).

3.6.2.3. Blended vs single malt

From the point of view of the blended/single malt distinction, the raw data provided more satisfactory results than the Fused approach. A PCA on the Fused Adjacency Matrix could only highlight a weak distinction tendency in the first two components, as shown in Figure 3.51c. On the contrary, even if at higher components, the raw data provided a very clear grouping tendency regarding the blended samples (Figure 3.51a). By looking at the loadings plot of the raw data (Figure 3.51b) it is not very easy to understand why

the blended samples end up being different from the single malt, but an overall feature that seems to be shared by these samples is that they tend to have low content of pyrolytic compounds (i.e. related to the smoky flavour) and esters.

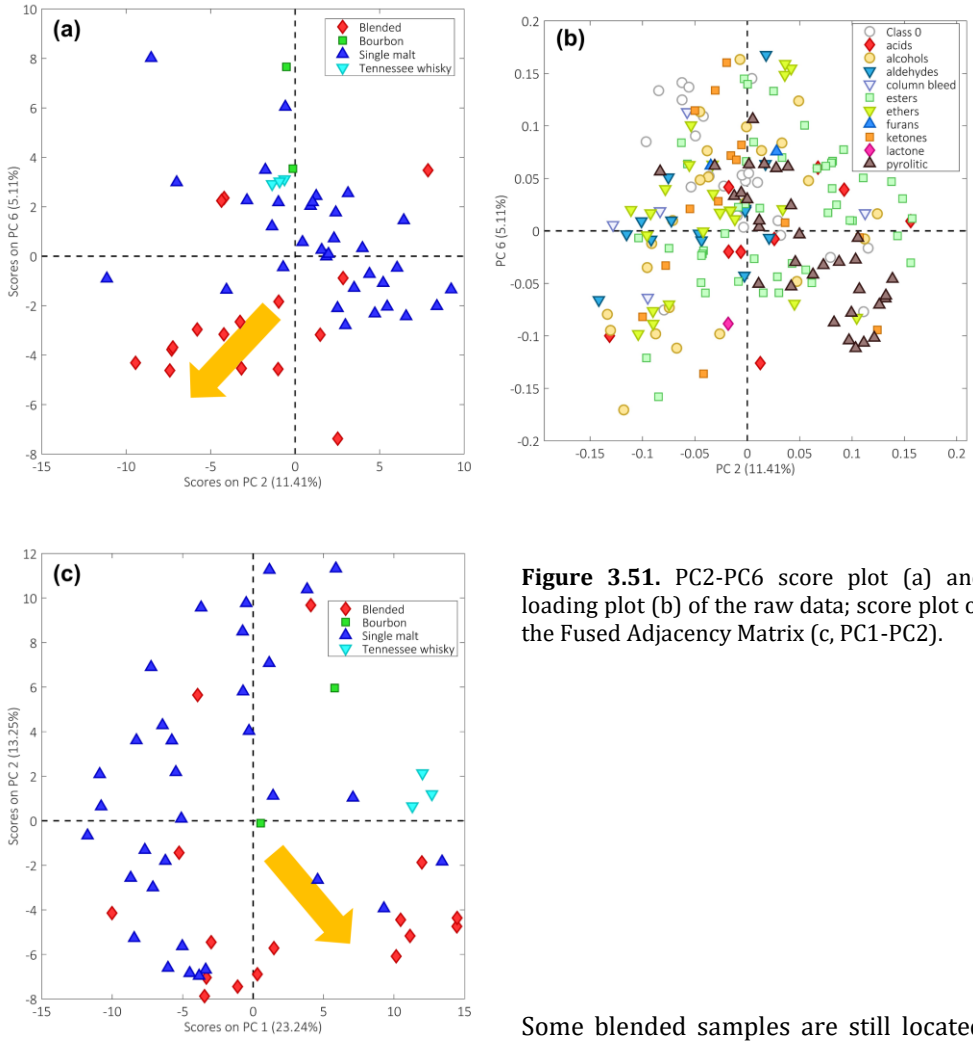


Figure 3.51. PC2-PC6 score plot (a) and loading plot (b) of the raw data; score plot of the Fused Adjacency Matrix (c, PC1-PC2).

Some blended samples are still located among the single malt whiskies. A possible explanation of why these samples seem to be more similar to the single malt ones may be found in the blending recipe: if very few whiskies are blended, then the mixture may result more similar to a "pure" whisky than a more complex blended one. Unfortunately, such a piece of information is very hard to obtain: each distillery has its own master blender who would never share the recipe, as it usually is a longstanding secret passed down from teacher to student.

3.6.2.4. Peated vs non-peated

No clear separation between the peated and the non-peated samples, neither in OPTICS nor in PCA was obtained. This may also be due to the fact that most whisky products are not explicitly labelled as "peated" or "non-peated", therefore this piece of information may suffer from some ambiguity. Moreover, depending on the type and provenience [23] of peat used for drying the malt and providing the smoky aromas, the *bouquet* of such aromas may be very different among the products.

The actual content of peat-related compounds may be better investigated by coclustering. For instance, two coclusters that can be directly related to pyrolysis products, are reported in Table 3.6. The samples belonging to these two coclusters were highlighted in the PCA score plots of Figure 3.52, for both the raw data (a) and the Fused Adjacency Matrix (b). The Fused approach was able to obtain a clearer separation of the peated samples.

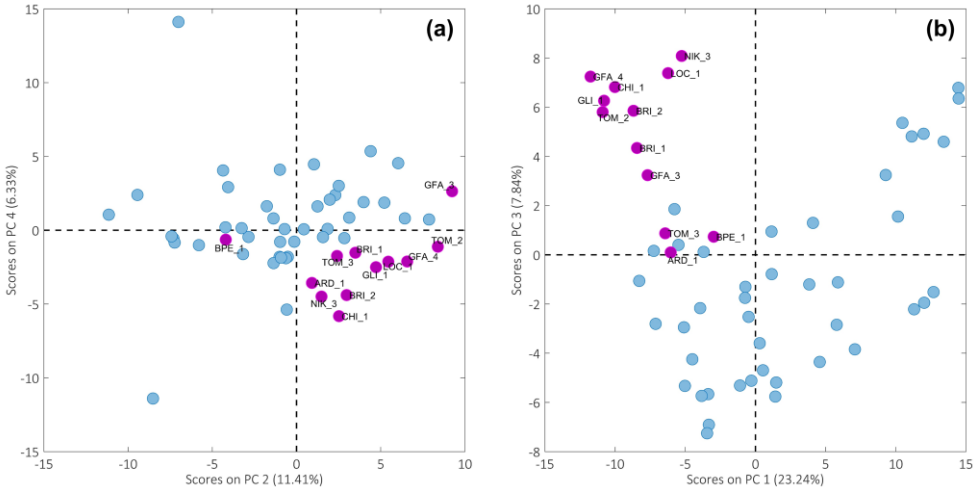


Figure 3.52. PCA score plots of the raw data (a, PC2-PC4) and the Fused Adjacency Matrix (b, PC1-PC3). The highlighted samples belong to coclusters #1 and #2, reported on Table 3.6.

On the other hand, if one focuses on the sharp peated/non-peated distinction provided by the aforementioned "label" information (which is represented in the PCA score plot of Figure 3.53), a trend in the raw data (Figure 3.53a) can be noticed: most of the peated whiskies are found at positive scores on PC1 and PC3. This direction is strongly

associated with chemical compounds originating from pyrolysis, which are naturally linked to the process of malt drying (Figure 3.53b). In the Fused Adjacency Matrix case however, a grouping direction can be seen in the PC1-PC2 score plot, highlighted with an orange arrow (Figure 3.53c). Those peated samples that are found in the opposite direction with respect to the arrow correspond to the samples lying among the non-peated samples in the PCA on the raw data (Figure 3.53a).

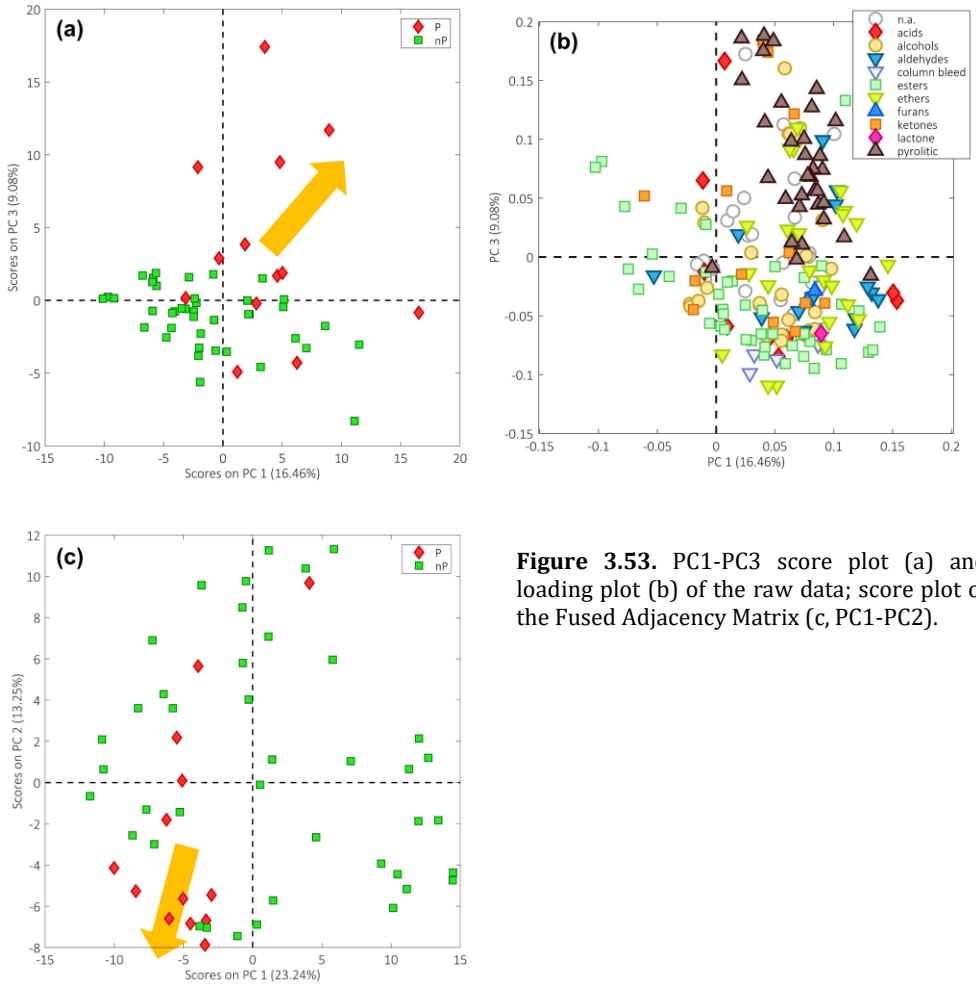


Figure 3.53. PC1-PC3 score plot (a) and loading plot (b) of the raw data; score plot of the Fused Adjacency Matrix (c, PC1-PC2).

3.6.2.5. Conclusions

The Fused Adjacency Matrix approach led satisfactory results regarding features like the country of origin and the peaty flavour of whisky, while concerning the distinction of blended and single malt whiskies, the analysis of the raw data was much more efficient in providing a clear grouping tendency.

The variables assigned to the chemical class of pyrolytic compounds surely deserve to be more deeply investigated, also to gain more knowledge about the smoky flavour, which represents one of the key features in producing and marketing a product such as whisky.

No trends related to the cask type used for aging the whisky were found, however, by selecting only a couple of cask types or by means of variable selection better results may be obtained, also with the application of the Fused Adjacency Matrix approach. For instance, an *ad hoc* variable selection approach may be applied to the data, with the aim of including only those compounds that can be directly related to the whisky-wood interaction: for instance, a study by Kew *et al.* [24] found some possible discriminant molecules for distinguishing between ex-Sherry and ex-Bourbon casks.

3.7. Mid-level data fusion application: the beer benchmark

As explained in Chapter 2, Section 2.3.2, the beer dataset consists of three data blocks, obtained by three different analytical techniques, namely Vis, NIR and NMR spectroscopies. This dataset represents the origin of the approach's development. Due to its potential richness in analytical information acquired, associated with its weak grouping structure and limited a priori knowledge (rather general such as beer style, alcohol content and colour), it is a challenging benchmark to test the approach's potential.

The approach was tested by processing each data block individually, and then merging the three output matrices $\mathbf{AM_X}$ to form the $\mathbf{AM_{Fus}}$ matrix depicted in black in Figure 3.3. The results of the Fused adjacency matrix approach were compared with those obtained from a traditional mid-level fused dataset, consisting of seventy-seven features extracted from the three data blocks, and with the exploratory analyses performed on each data block individually.

3.7.1. Mid-level fused dataset using a traditional approach

The traditional mid-level data fusion dataset was obtained by merging 7 PCA scores from the Vis dataset, 6 PCA scores from the NIR dataset and the 64 NMR features. To represent the three different blocks evenly, autoscaling followed by block-scaling was performed.

3.7.2. Results

The results reported in this section are almost identical to the paper *Fused Adjacency Matrices to enhance information extraction: the beer benchmark* (Cavallini *et al.*, [25]), except for minor parts, mostly extensions obtained by adapting parts from the supplementary materials of the paper. More specifically, Sections 3.7.2.5 and 3.7.2.6.3 were not included in the paper. Please note that all sections and figures were newly numbered and, in some cases, expanded.

3.7.2.1. Visible dataset

The Visible spectra, after preprocessing, were analysed by PCA and OPTICS. Figure 3.54 reports the results, namely the OPTICS reachability plot (RP) in Figure 3.54a, and the PC1-PC2 score plot in Figures 3.54b and 3.54c, colored according to beer style (b) and colour intensity (c).

Two main groups were identified by OPTICS. The first one, the Ales group, is mainly composed by ale-style samples and it is less homogeneous compared to the second, the Lagers group, which is largely composed by lager-style samples. The two groups also have different density: the Lagers group results denser than the Ales group, and this can be seen in both the RP (Figure 3.54a) and the score plot (Figure 3.54b). The colour scale employed in Figure 3.54c describes the beer colour intensity, that is defined as the absorption of the sample at 430 nm, taken as reference wavelength [26]. A colour intensity gradient is recognizable along PC1 (Figure 3.54c). The sample distribution along PC2 is, on the contrary, much less clear. Some of the mid-coloured samples are spread along PC2, and the four samples with the strongest absorption have negative scores on this component. These four samples belong to very different beer styles but look rather grouped in the PC1-PC2 score plot. This is not reflected by the RP, where the samples show increasingly higher distances. Actually, by inspecting the score plots of higher PCs (not shown) these non-grouped samples are always found at extreme positions with respect to the rest of the samples. Since OPTICS operates on the full spectra, the increasing RD trend is due to the piece of information that is not included in the PC1-PC2 score plot.

3.7.2.2. NIR dataset

The information that could be extracted from the NIR dataset is rather limited, and this can be seen by inspecting the RP (Figure 3.55a) and the PC1 score plot (Figure 3.55b), both obtained from the NIR preprocessed spectra.

A clear alcohol content (% alcohol by volume, ABV%) gradient is recognizable along PC1, as shown in Figure 3.55b. Ethanol content is therefore efficiently represented by

PC1, whose corresponding loadings (not shown) are characterized by two intense ethanol bands within the region 2200-2400 nm [27].

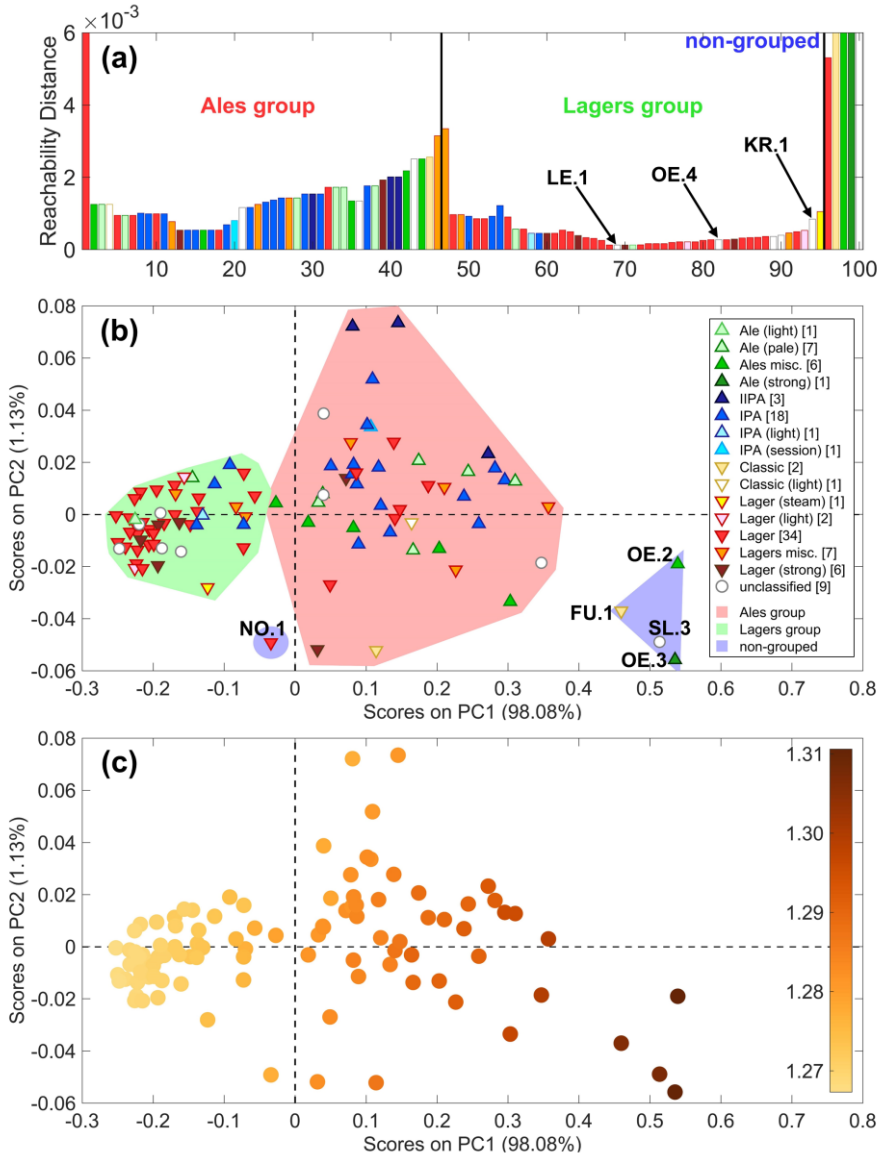


Figure 3.54. Visible spectra dataset: (a) Reachability Plot; (b) PC1 vs PC2 score plot, different symbols refer to top (▲) and bottom (▼) fermentation, while colours are by beer style, as detailed in the legend; (c) PC1 vs PC2 score plot coloured according to beer colour intensity: one intensity value for each spectrum is calculated by taking the average of intensity values in the interval 430±5 nm. The background patches in (b) highlight the OPTICS groups defined in (a).

Two main clusters of samples were identified by inspecting the RP (Figure 3.55a), a small one which contains a mix of beer types ("mixed group") and the Lagers group. The Light beer samples appear rather grouped, as it is indicated by the shaded light blue rectangular area in Figures 3.55a and 3.55b. The samples located at the right end of the plot can be considered as non-grouped. This was also found in PCA, where the two identified clusters have reduced variability along PC1 with respect to the non-grouped samples (Figure 3.55b). The non-grouped set is much more scattered, as it has both higher bars in the RP (Figure 3.55a) and a large variability range along PC1 (Figure 3.55b).

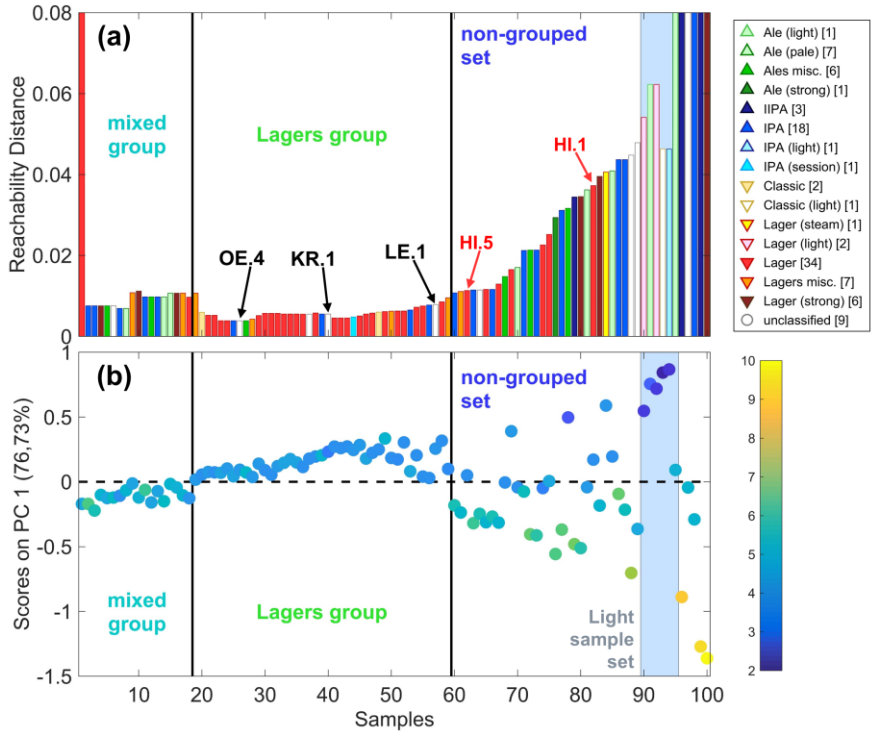


Figure 3.55. NIR spectra dataset: (a) Reachability Plot, bars are colored by beer style, as detailed in the legend; and (b) PCA score plot colored by ABV content. Samples in both in (a) and (b) were reordered according to OPTICS order.

3.7.2.3. NMR dataset

A data representation from the field of sensomics [28,29], was used for inspecting the NMR features and the results are shown in Figure 3.56. The heatmap [29] in the central

part of the figure represents the data values. The columns of the heatmap represent the samples while the rows represent the variables (concentrations of MCR-resolved features in the different samples). Rows and columns were reordered according to the sequences obtained by running OPTICS first in the samples' direction (RP on top) and then also in the variables' direction (RP on the left side). This allows highlighting both groups of samples and variables, making it easier to relate the most influent groups of variables to each group of samples [29].

To obtain clearer groupings in the variables' direction, correlation among the NMR features was used, instead of distance, to calculate the reachability distance for the RP plot. Three main groups of variables can be identified (Figure 3.56 variables' RP, on the left side): the first group mainly contains amino acids, together with uridine and gallate; the second group is composed of yet unassigned variables, and the third group is partially related to maltose and to two unassigned variables.

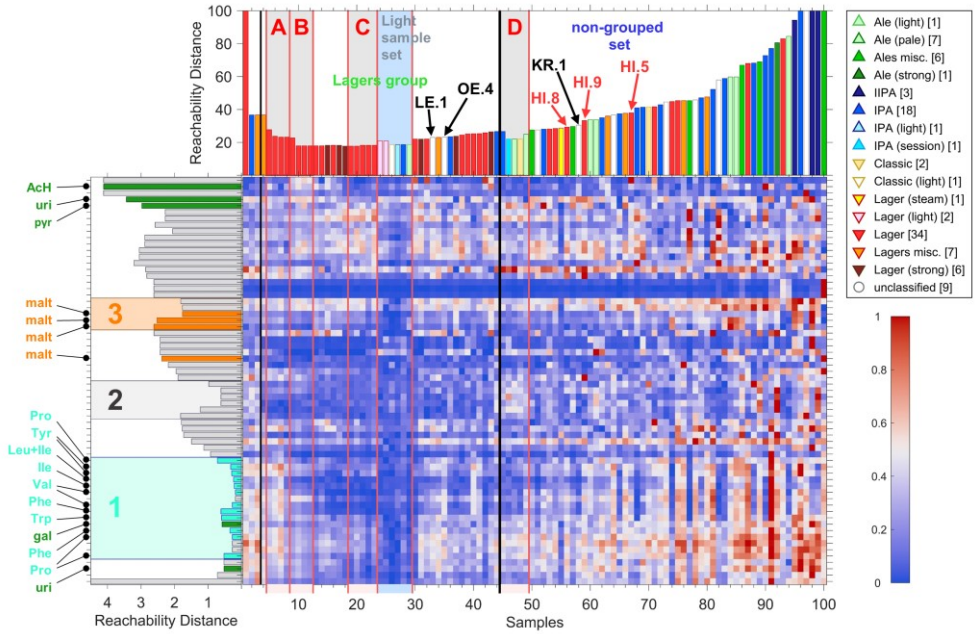


Figure 3.56. Heatmap of NMR features with Reachability Plots: variable's RP on the left side (k = 3), samples' RP on top (k = 5). OPTICS in the variables' direction was performed on the correlation matrix, instead of the variables themselves. In the central part of the figure it is shown the heatmap obtained by reordering both the samples and the variables according to the respective OPTICS sequences. The dataset was normalized between zero and one to enhance its visual representation and interpretability.

The samples' RP shows a cluster that can be identified as the Lagers group. The rest of the plot is rather uninformative from a group-spotting point of view, since its largest part consists of a sequence of increasing RDs (non-grouped set). Interestingly, the Light beer samples constitute a recognizable sub-group which, as expected, has generally low values for all the variables. Also, a small group can be spotted at the centre of the RP plot (group D in Figure 3.56), and it is characterized by medium-low values in amino acids and medium values for the second group of variables. The non-grouped set contains very different beer styles. The samples belonging to this group generally have higher amino acids content, but also maltose (third group of variables).

3.7.2.4. Traditional mid-level fused dataset

The PCA and OPTICS results obtained from the preprocessed mid-level fused dataset are shown in Figure 3.57. The OPTICS results resemble those of the NMR features dataset: a slightly defined Lagers group at the beginning of the RP, followed by a tail of slowly increasing RDs forming a non-grouped set (Figure 3.57a). However, the sample distribution obtained by PCA (score plot in Figure 3.57b) is mainly determined by few variables, according to the loadings plot (Figure 3.57c). Features related to ABV ("Scores PC1–NIR") and colour ("Scores PC1–Vis", "Scores PC2–Vis") are the most influential.

All the Light beer samples are located at negative PC1 and positive PC2 scores, while two of the strongest samples lie far away in the opposite direction. This defines an ABV direction (light blue arrow in Figure 5b). Even though the Light beer samples seem to be rather grouped in PCA, they are not found grouped in the RP. Again, an explanation for this discrepancy can be found in the different amount of information described by the RP (the whole preprocessed data) and the first two PCs shown in Figure 5b, which only account for 29.63% of the total variance of the mid-level fused dataset. Almost perpendicularly to the ABV direction, the variable "Scores PC1–Vis" (Figure 3.57c) tends to separate the most coloured samples (Figure 3.57b, highlighted in orange), and helps to separate along PC1 the Lagers from the Ales, which usually have more intense colours.

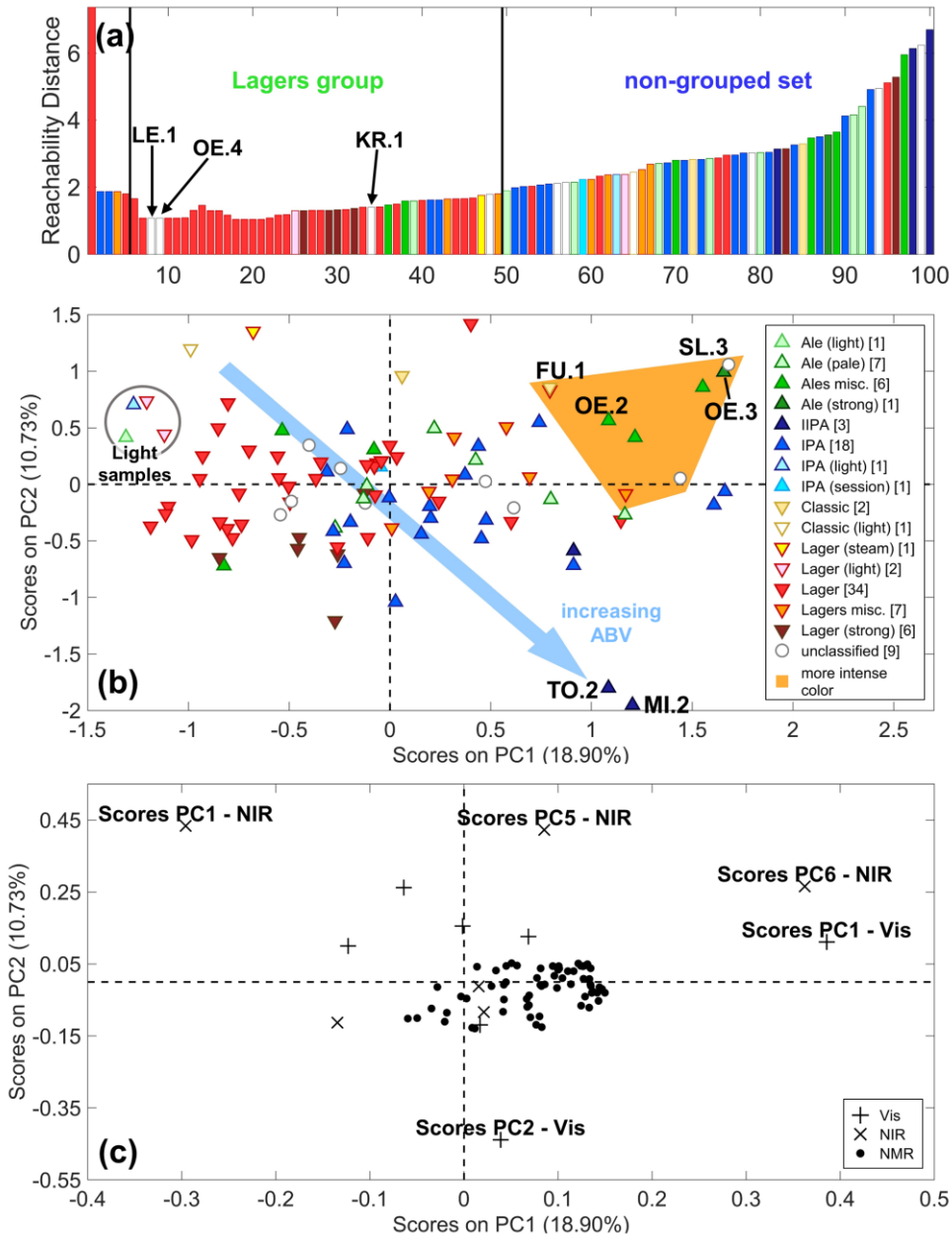


Figure 3.57. Mid-level fused dataset: (a) Reachability Plot, (b) PC2 vs PC1 score plot, (c) PC2 vs PC1 loadings plot; colours and symbols explained in the legend on the plot. The area highlighted in orange corresponds to the most coloured beer samples.

3.7.2.5. Fused Adjacency Matrix results

The results obtained by OPTICS and PCA on the Fused Adjacency Matrix preprocessed as explained in Section 3.5.4 are discussed here and shown in Figure 3.58.

Two clusters of samples and a non-grouped set can be identified in the RP (Figure 3.58a). These three groups have a correspondence in the PC3-PC1 score plot of the same matrix (Figure 3.58b) The non-grouped set is more scattered in PCA (blue patch in Figure 3.58b), and it contains the strongest one and three of the five Light beer samples. The Ales and Lagers groups are much more defined compared to the results found with the single techniques and the mid-level data fusion approach. It is also interesting to notice the sample distribution within the Lagers group, where the "simple" lager samples (in red in Figure 3.58b) are very grouped on the right side, which is in an opposite position compared to the Ales group.

PC1 is related to the colour, and when combined with PC4 the samples adopt an arch-like distribution (Figure 3.58c). The PC1-PC4 score plot not only shows the colour trend, but also suggests new groups of samples, which are highlighted in grey in Figure 3.58c. To gather which characteristic features are shared within these sub-groups the sub-group average NIR spectra (Fig 3.59) and NMR resolved features (Figure 3.60) were compared. Most of the groups have some distinctive regions, e.g. sub-groups 6 and 7 have higher content of amino acids content, while the three close IPAs (sub-group 4) have high values in NMR for maltose and a set of features not yet completely identified, among which ethanal, isopentanol and higher alcohols were tentatively assigned.

Based on our current knowledge, it is not possible to fully explain these groupings, however work is in progress analysing a database of consumer preferences obtained from the website ratebeer.com to assess if some of the grouping may be related to such information. Preliminary results show that PC1 of the Fused Adjacency Matrix seems to have a strong inverse relationship ($R^2 = -0.973$) with the overall score computed by the website from the users' evaluations (Figure 3.61).

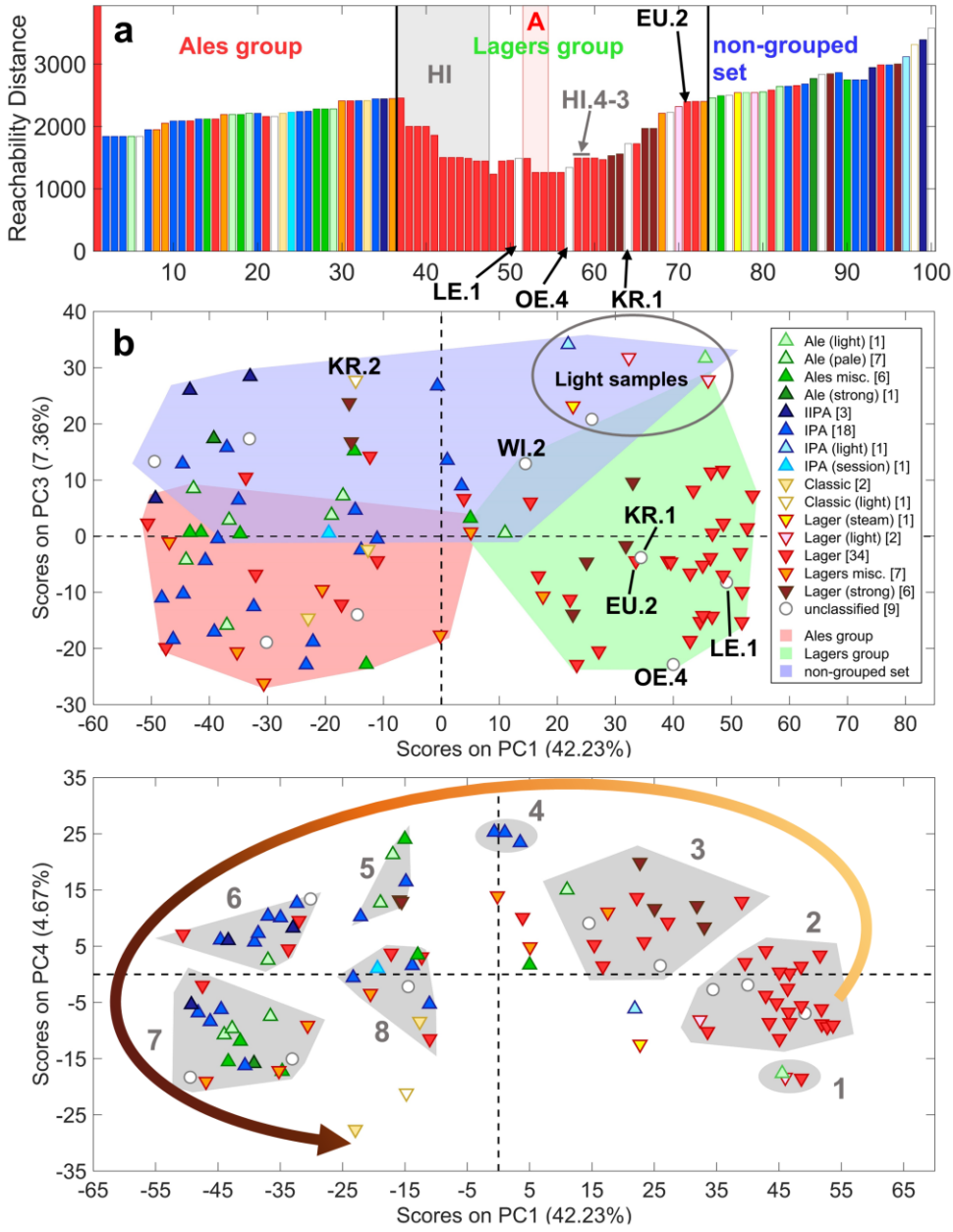


Figure 3.58. Fused Adjacency Matrix: (a) Reachability Plot; (b) PC3 vs PC1 score plot, colours and symbols explained in the legend on the plot; the background patches in (b) highlight the OPTICS groups defined in (a). (c) PC4 vs PC1 score plot, colours and symbols explained in the legend on the plot; the curved arrow in (c) describes the beer colour intensity trend; the red background patches in (c) highlight possible new groups.

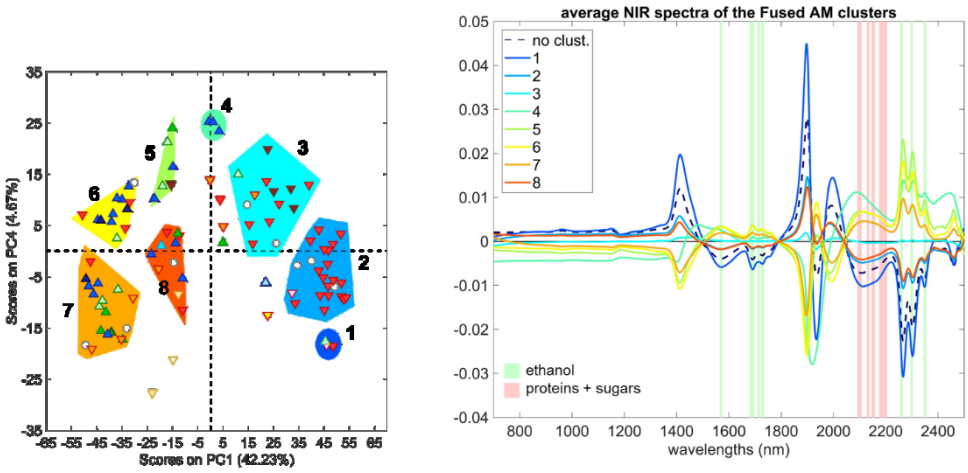


Figure 3.59. Sub-groups identified in Figure 3.58c, and their corresponding average NIR spectra.

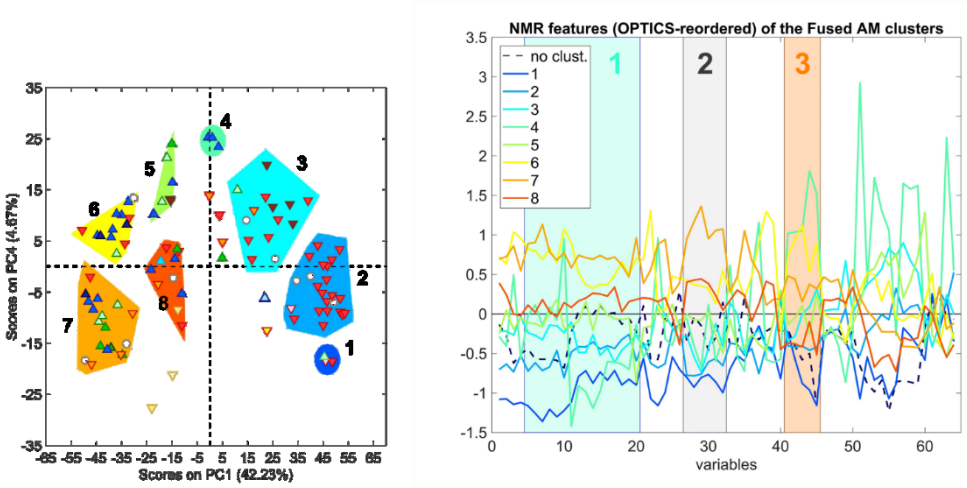


Figure 3.60. Sub-groups identified in Figure 3.58c, and their corresponding average NMR features.

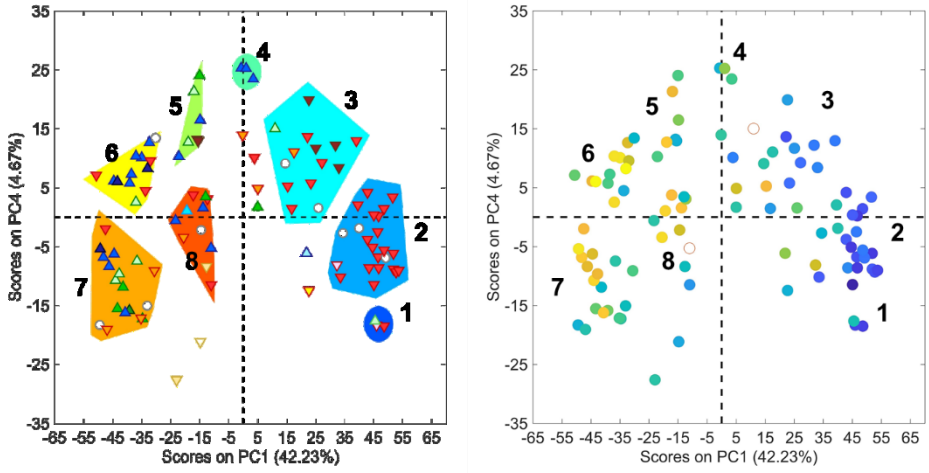


Figure 3.61. Sub-groups identified in Figure 3.58c, and their corresponding average RateBeer ratings.

3.7.2.6. Beer features comparison summary

In this section, more detailed comparisons among the results obtained by the different data blocks and data fusion approaches are reported. Table 3.7 is organized as a summary of these comparisons. Some overall samples' sets and beer features were tracked along the single data blocks.

3.7.2.6.1. Lagers group

The Lagers group was identifiable in all representations of the data, and it appears to be rather stable. The Vis and **AM$_{Fus}$** datasets showed the best results in terms of samples grouping, which is probably reflected by their similarity, as highlighted by Procrustes Analysis (Section 3.7.2.8).

An interesting group of lager-style samples is the HI samples set, which includes beer products from the same brand, Hite. This set of samples is organized in couples of replicates: "Pale Lager" (HI.1-2, HI.3-4), "Dry Finish" (HI.6-7), "Golden" (HI.8-9) and "Fresh" (HI.10-11-12-13), where the second replicate underwent thermal treatment to simulate ageing. Only sample HI.5 does not have a replicate and it is also a different beer product ("MAX"). The HI samples were generally found in the Lagers group, with some

exceptions: HI.1 and HI.5 in NIR (Fig.3a); HI.8-9 and HI.5 in NMR (Fig.4). No fixed order related to thermal treatment was found, neither with OPTICS nor with PCA, in any dataset. Moreover, no consistent order of the replicates was found neither in the spectral datasets, nor in the mid-level fused dataset, even though in the NMR case some of the HI samples were found gathered in two sub-groups: group B (HI.10-11 and HI.12-13) and group C (HI.4-3, HI.6-7) in Figure 3.56. Group B has higher content of some amino acids, acetate, uridine and an unassigned variable between the two last ones. On the contrary, this piece of information clearly emerged by analysis of $\mathbf{AM_{Fus}}$ dataset. In fact, the HI samples were found very well grouped together in the RP (HI in Figure 3.58a), forming a rather ordered sequence of couples of HI replicates; couple HI.3-4 was not found among the other HI samples, but some positions further in the sequence of the RP (Figure 3.58a).

Another interesting set of samples is represented by the EU beers. They belong to the same brand and three of them are the same product (EU.1-2-3, "Brüger Premium Pils"), while EU.4 ("Servus") is different. However, sample EU.2, differently from the other three EU samples, did not undergo thermal treatment. These samples were not found grouped in the Vis and NIR cases, while in NMR, mid-level data fusion and $\mathbf{AM_{Fus}}$ the EU group was recovered in the RPs, albeit to different extents. In the NMR case, the samples are ordered (group A in Figure 3.56) as EU.1, EU.3 ("Brüger" treated), then EU.2 ("Brüger" non-treated) and finally EU.4 ("Servus" treated). In the case of mid-level data fusion, a similar situation was found, but EU.4 was found further in the RP. Interestingly, in the $\mathbf{AM_{Fus}}$ case, the three thermal treated samples (EU.1, EU.3 and EU.4) were found grouped together (group A in Figure 3.58a), while EU.2 one was found further in the OPTICS sequence, suggesting that, only by this approach, a clearer difference based on the treatment was recovered.

Three "unclassified" samples (LE.1, OE.4, KR.1) were consistently found in the Lagers group. These products are described as "summer beers", therefore their presence in the Lagers groups is not unforeseen: this product type is intended to be refreshing and easy-to-drink, and it usually is lighter in aromas and alcohol content. For these reasons it can be expected to find these summer beers more similar to the lagers than the ales.

3.7.2.6.2. Light samples set

The Light samples set includes five beers of different styles (KR.2, Classic light / LE.2, IPA light / FB.2, Lager light / TO.4, Lager light / NO.2, Light Ale). These beers are labelled as "light" and they are produced with the aim of obtaining a lower content of ethanol and flavours.

The NIR and the NMR datasets gave the best results in terms of grouping the Light samples set. In the NIR case the Light samples were found grouped both in the RP and the PCA scores (light blue patches in Figure 3.55). They lie at extreme positive values along PC1, which is a component that describes ethanol content. A confirmation of the generally lower content in flavours was found from the NMR results: all the Light samples share a similar pattern of very low values along all the variables of the dataset (Light sub-group in Figure 3.56).

The Light samples set was found rather grouped in the data fusion cases (Figures 3.57b and 3.58b), but only in PCA. In the Vis case, the Light samples are neither grouped in RP or PCA but belong to the Lagers group: lighter beers are usually less processed/fermented, so they tend to develop less intense colour.

3.7.2.6.3. ABV trend

No ABV trend was evident in the Vis case. This is naturally present in the NIR case (Figure 3.55b), since PC1 describes the ethanol content. The trend is also present in the mid-level data fusion case, since variable PC1 from NIR is highly influential (Figure 3.57c). No clear ABV trend was found in the RP for the NMR case, even if it was found in PCA, which is reported in Figure 3.62.

The ABV trend in Figure 3.62 is clearly recognizable, even though sample FB.3 is in a quite strange position, among the 4-6% alcohol beers. This sample has the highest ABV content (10%), and its position could be explained by the fact that no direct information about ethanol is present in the NMR dataset (i.e. no peaks directly related to ethanol are included). The ABV trend could be related to the sugar content: during fermentation, the yeasts consume sugars to produce ethanol, therefore the higher the ABV content, the lower the sugar content. To reach very high ABV values more sugar is needed, and this

can be achieved by adding fermentable sugars. This addition may break the inverse "balance" between sugar and ethanol content: if sugar was added to sample FB.3, the residual amount of sugars could be the cause of its position, far from the low-sugars/high-ABV samples.

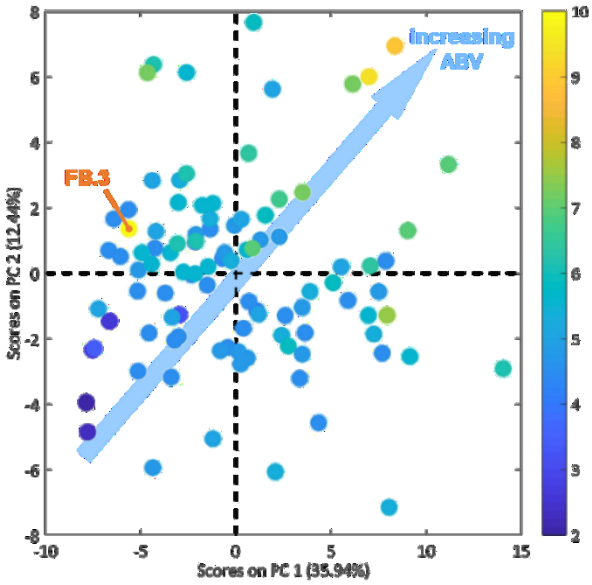


Figure 3.62. NMR PC1-PC2 score plot, colored according to ABV content.

The **AM$_{Fus}$** case is rather different. The ABV trend is present in PC1-PC3 (score plot reported in Figure 3.63), but in a transformed way. The strongest and the lightest beers all lie in the top part of the plot and they all belong to the non-grouped set (as in Figure 3.58b). These samples represent the extremes in ABV, so their position is probably due to the fact that the approach is just able to detect their dissimilarity from the bulk of "ABV-average" samples.

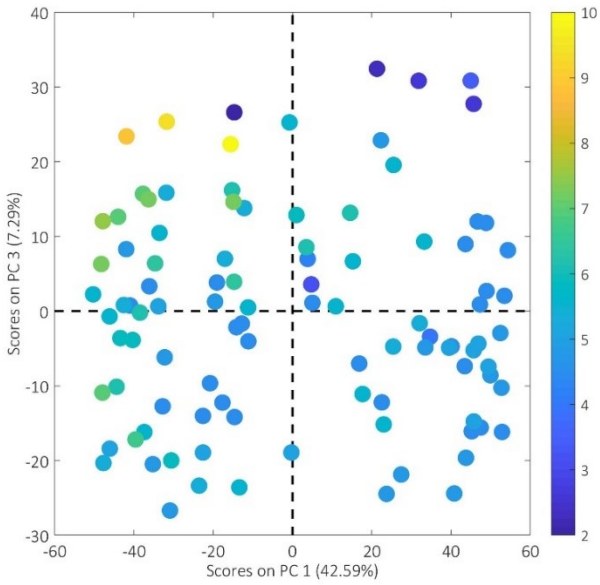


Figure 3.63. Fused Adjacency Matrix PC1-PC3 score plot colored according to ABV content.

3.7.2.6.4. Lagers Strong set

The Lagers Strong set includes six beers (ordered by increasing ABV, MA.3, SI.9, MA.5, MA.6, MA.2, FB.3) and it is interesting to track their position because of their style: lagers strong are beers brewed with lager yeasts, but more alcohol is obtained during the brewing process.

The Lagers Strong set was generally found split into two groups: four "low-ABV" and two "high-ABV" samples. The low-ABV samples (MA.3, SI.9, MA.5, MA.6) were found in the Lagers group in the cases of Vis, mid-level data fusion and $\mathbf{AM_{Fus}}$, while the NIR and NMR cases provided two different situations. In the NIR case, the three lowest ABV samples were found in the mixed group, closer to the Lagers than the three highest ABV samples (Figure 3.55a). On the contrary, in the NMR case, the Lager Strong samples are all in the Lagers group and do not follow any ABV order (Fig.4). Both the data fusion approaches, in RP by OPTICS (Figure 3.57a and Figure 3.58a) is clearly highlighted that the four low-ABV samples are more similar to the lagers (they belong to the Lagers group) but are also located closer to each other within the RP sequence. However, the

separation between high- and low-ABV samples is much better appreciable in the PCA of the $\mathbf{AM_{Fus}}$ (Figure 3.58b) than in the mid-level data fusion score plot (Figure 3.57b). In $\mathbf{AM_{Fus}}$, moving along PC1 from the Lagers group towards the Ales group, the four low-ABV samples are found, while the two high-ABV samples are much more distant, and closer to the strongest samples in the dataset. On the contrary, the same samples in the mid-level data fusion score plot (Figure 3.57b) are located in the same area.

3.7.2.6.5. Colour trend

The colour trend naturally originates from the Vis dataset (Figure 3.54c). No trace of it was found neither in the NIR nor the NMR cases. Both the data fusion methods were able to recover this piece of information, even though the $\mathbf{AM_{Fus}}$ (Figure 3.58c) provides a clearer trend than the mid-level data fusion (Figure 3.57b).

3.7.2.6.6. Summary Remarks

The trends and groupings described above generally correspond to the main known traits of the beer styles under examination. While the single spectral data blocks can primarily provide one aspect each, both the data fusion approaches were able to collect and keep most pieces of information. The Fused Adjacency Matrix, however, could capture finer structures in the main groups, for instance the very well-ordered HITE group, with the replicates of each product found in a sequence by OPTICS, or the EU set, where the treated samples were found grouped together and the non-treated one was found much further away. Trends like colour intensity and lager/ales distinction were recovered more clearly by the Fused Adjacency Matrix, while others like ABV content and the Light samples set were slightly better retrieved by the mid-level data fusion approach.

It is also very promising that the Fused Adjacency Matrix approach can highlight small sub-groups (Figure 3.58c) which may be worth further investigation of their chemical/sensory characteristics. A deeper characterization of these sub-groups may, for instance, provide new inspiration in beer production, helping to define intersections between established and more general styles.

Table 3.7. Comparison summary (*ordered by increasing ABV)

	Visible	**NIR**	**NMR** (Fig.3.56)	**Mid-level data fusion**	**Fused Adjacency Matrix AM_{Fus}**
Lagers group	Dense cluster in RP. (Fig.3.54a) Grouped in PCA. (negative scores, Fig.3.54b)	Slightly defined in RP. (Fig.3.55a) At positive PC1 scores, close to zero. (Fig.3.55b)	Slightly defined in RP. Medium to low variable values in general. Some sub-groups; contains the Light samples set as a sub-group.	Slightly defined in RP. (Fig.3.57a) At negative PC1 scores. (Fig.3.57b)	Defined cluster in RP. (Fig.3.58a) HI samples grouped and well-ordered together in RP. (Fig.3.58a) Grouped in PCA. (Fig. 3.58b)
Unclassified ○ fresh/summer beers in the Lagers group (most frequent ones: LE.1, OE.4, KR.1)	LE.1, OE.4, WI.2, SK.4, KR.1 (Fig.3.54a)	OE.4, UG.3, KR.1, LE.1 (Fig.3.55a)	LE.1, OE.4 KR.1 is in the non-grouped set.	LE.1, OE.4, KR.1, TY.3 (Fig.3.57a)	LE.1, OE.4, KR.1, WI.2 (Fig. 3.58b)
Light samples set (KR.2, LE.2, FB.2, TO.4, NO.2)*	All in the Lagers group. (Fig.3.54b) Generally lighter colours. (Fig.3.54c)	Quite grouped in RP. (Fig.3.55a) All extreme on PC1. (Fig.3.55b)	Grouped in RP. Included in the Lagers group. Low values in general.	Not grouped in RP. (Fig.3.57a) Grouped in PCA. (Fig.3.57b)	Not grouped in RP. (Fig. 3.58a) Grouped in PCA. (Fig. 3.58b)
Lager Strong four low-ABV: MA.3, SI.9, MA.5, MA.6 two high-ABV: MA.2, FB.3	Four low-ABV in the Lagers group, low-colour. (Fig.3.54a-b) Two high-ABV in the non-grouped set, mid-colour. (Fig.3.54-b)	Three in the mixed group. (Fig.3.55a / SI.9, MA.5, MA.3) Three in the non-grouped set. (Fig.3.55a / *MA.6*, MA.2, FB.3)	All in the Lagers group.	Four low-ABV in the Lagers group. (Fig.3.57a) Two high-ABV quite far in the non-grouped set. (Fig.3.57a)	Four low-ABV close to the Lagers group in PCA. (Fig. 3.58b) Two high-ABV close to the Ales. (Fig. 3.58b)
ABV trend	Not found.	Very well described by PC1. (Fig.3.55b)	Found in PCA (Fig.3.62); probably reflecting the sugar content.	Found in PC1-PC2 score plot. (Fig.3.57b)	Found in a transformed way. (Fig.3.63)
Colour trend	Clearly found along PC1. (Fig.3.54c)	Not found.	Not found.	In PCA the stronger colored samples lie at positive PC1 and PC2 scores. (Fig.3.57b)	Nicely represented by PC1 and PC4. (Fig. 3.58c)

3.7.2.7. Comparisons by means of Procrustes Analysis

In Sections from 3.7.2.6.1. to 3.7.2.6.6. we have graphically inspected and compared the information gathered by the different data blocks as depicted in the principal components space, with the aim of highlighting similarities and differences among them. This way of visually exploring the data easily allows spotting trends and peculiarities, but subjectivity and limited availability of metadata (i.e. additional information such as the beer style or the ABV content) can sometimes be a drawback.

A more objective evaluation of how similar/different are the results obtained from the different data blocks by comparing their PCA spaces can be obtained by means of Procrustes Analysis (PA, [30,31]). Like in our beer benchmark case, the same set of objects can be described by two distinct sets of PC scores, obtained for instance from two different analytical sources. The aim of PA is to obtain the closest match between these two PC spaces by applying operations such as scaling, rotation, reflection and translation. The similarity of the two spaces is expressed using a dissimilarity parameter d, ranging from zero to one [31].

In this work, the PCA spaces obtained from the different blocks (i.e. each single analytical platform, the mid-level fused data set and the $\mathbf{AM_{Fus}}$ data set, referred to as inter-block comparison) are compared by PA analysis. Also, the data obtained from the different steps of the procedure, going from the raw data to the AMs for each single data set (which will be named $\mathbf{AM_X}$, with the suffix X being Vis, NIR and NMR, in turn) have been compared by PA. The latter case is referred to as intra-block comparisons. An overview of the results is given hereinafter, while the visual representation is reported in Figures 3.64 and 3.65.

Inter-block comparisons were made, in pairs, using the PC scores of the Visible spectra (7 PCs), the NIR spectra (6 PCs), the NMR features (6 PCs), the mid-level fused data (5 PCs) and the Fused Adjacency Matrix ($\mathbf{AM_{Fus}}$, 7 PCs). The same number of principal components as that considered to build the mid-level fused dataset were used in PA, to keep it constant, and the results are shown in Figure 3.64, where the dissimilarity value between each pair of data sets is reported. $\mathbf{AM_{Fus}}$ is substantially different (dissimilarity higher than 0.5) from the mid-level fused data, which suggests that these two datasets carry different information. $\mathbf{AM_{Fus}}$ was also found rather different from the other

datasets: this is a desirable situation, since we are dealing with a data fusion approach. A too strong resemblance with any single source dataset would have meant that the fusion process was giving too much importance to that source, while a too loose similarity would have meant that the information was either too reduced or not captured by the approach.

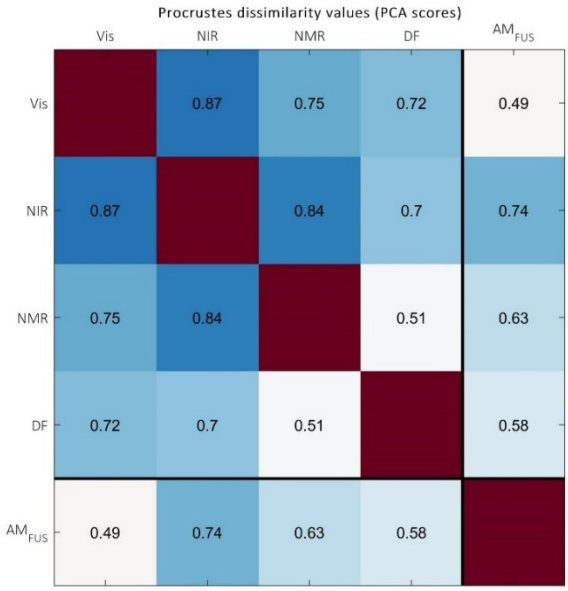


Figure 3.64. Inter-block comparisons by means of Procrustes Analysis. The numbers reported at each comparison's cross are the Procrustes' dissimilarity values. Vis, NIR and NMR represent the spectral (features in the NMR case) data from which the PCA scores were obtained.

The effect of the different fusion steps was also assessed. These intra-block comparisons were made for each data block individually (using the same number of PCs as specified above), and the results are shown in Figure 3.65. One interesting point is the transition from the distance information to its correspondent $\mathbf{AM_X}$. The Euclidean distance $\mathbf{D_{Euc}}$ resulted consistently similar to the Euclidean $\mathbf{AM_{Euc}}$ meaning that the "coded" AM version of the data is keeping a large part of the original distance information. The same was observed with the Mahalanobis distance, albeit for the NMR case the similarity between $\mathbf{D_{Mah}}$ and $\mathbf{AM_{Mah}}$ was found lower (Figure 3.65). By inspecting the corresponding score plot it appears that this difference is due to a limited number of samples which have extreme values on the second component in PCA of $\mathbf{D_{Mah}}$ and are not in $\mathbf{AM_{Mah}}$ (adjacency being assigned on interval values is less sensitive to extreme

values). Another interesting relation is between the Euclidean and SOM AMs: the matrices $\mathbf{AM_{Euc}}$ and $\mathbf{AM_{SOM}}$ are very similar, either because the samples pattern in the beer data can be well described by a linear model or because the Euclidean distance (which is a non-linear transform) is sufficient to model the non-linearity present in the data pattern. These two AMs also represent the two major contributions to the single-data block $\mathbf{AM_X}$. The Mahalanobis distance was consistently found rather different from $\mathbf{AM_X}$ and the other distance measures. This is probably due to the fact that higher PCs bring in rather different information with respect to the first ones, as in order to avoid singularities we have calculated the Mahalanobis distance on PCA-compressed data and thus it corresponds to Euclidean distances on the autoscaled PCs. However, a systematic different behaviour of the Mahalanobis distance with respect to other metrics (including Euclidean) has been previously observed in a study considering several data sets [6].

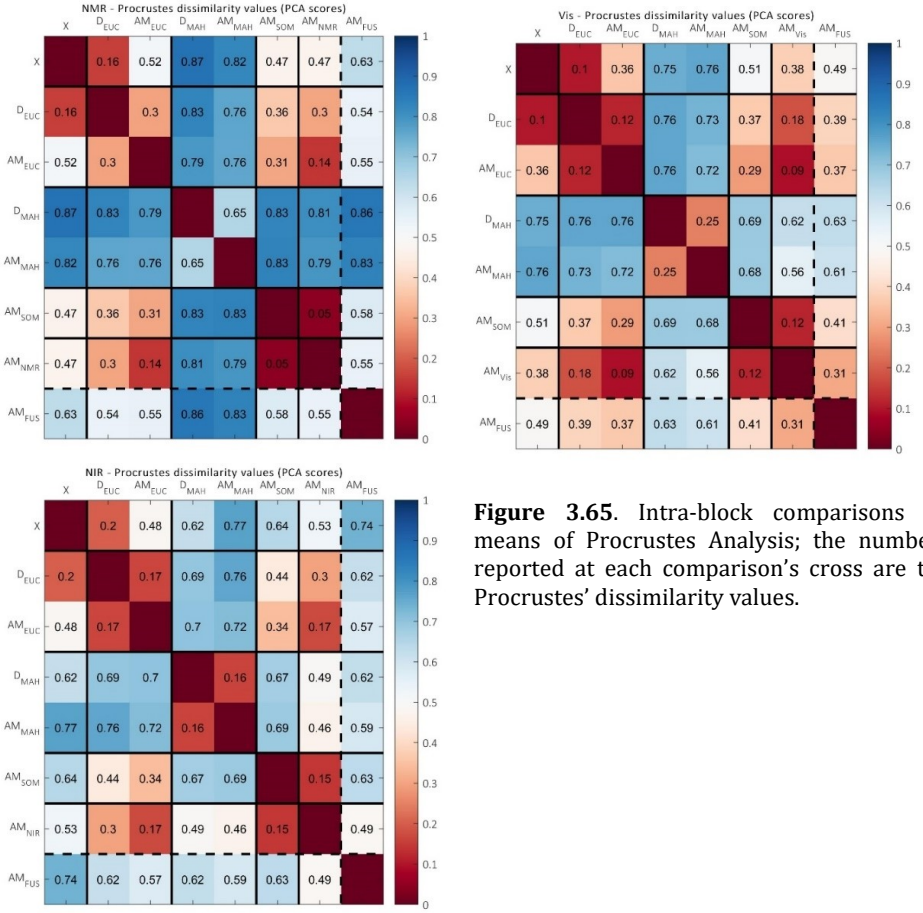


Figure 3.65. Intra-block comparisons by means of Procrustes Analysis; the numbers reported at each comparison's cross are the Procrustes' dissimilarity values.

3.7.3. Link to the original variables

One of the major issues when dealing with adjacency matrices is that the link with the original variables is lost. When an adjacency matrix is built, the "adjacency condition" for each pair of samples is evaluated, therefore the focus is on how distant the two samples are: the original variables are only used to compute the distances.

A way for linking back the Fused Adjacency Matrix results to the original variables is presented in Figure 3.66 using the NMR features dataset as an example. By using the same representation used in Figure 3.56, the samples were reordered using the RP sequence obtained from the Fused Adjacency Matrix. Therefore, the heatmaps of the two figures only differ in the order of their columns. Such a new column sorting allows a direct comparison between the observed sample clusters and the chemical features linked to specific class of compounds, as detailed in the following section.

The Ales group in Figure 3.66 shows medium-high values in correspondence of the amino acids. The non-grouped set also has some samples with comparable values for the amino acids, but the Ales group has a more uniform composition. The amino acids region also represents the main difference between the Ales and the Lagers groups. This is in accordance with the results obtained by Duarte *et al.* [32], who suggested that the aromatic region could be used to distinguish between ales and lagers.

Two sub-groups can be noticed within the Ales group (A and B in Figure 3.66). The first sub-group (A) is mixed, and consists of seven ales, four lagers and one unclassified beer. These samples have medium values for variables from 3 to 11, which include compounds such as tryptophan, gallate, phenylalanine, uridine and two signals from proline. Their amino acid content is on the other hand much lower if compared to the other samples belonging to the Ales group. The second sub-group (B in Figure 3.66) consists of five ales and two lagers. This sub-group is characterized by high values related to the first 20 variables, which include all the identified amino acids together with gallate and uridine.

The Lagers group generally has medium-low values, especially in the case of the second group of variables and the amino acids group. Several sub-groups can be identified within the Lagers group (C, D, E, F and G in Figure 3.66). A couple of samples at the

beginning of the group (C) have almost identical patterns, especially for the amino acids content. These two samples are the same beer product, but the second one underwent thermal treatment. Some differences can be spotted along the two patterns, and the second sample always has higher values at these points. A second sub-group (D) consists of four lager samples of the same brand, which are among the poorest in amino acids content. Their patterns look very similar to sub-group E, which contains two beers of the previous brand, two more lagers and one lager strong. Sub-groups F and G also have similar patterns, but the samples in F tend to have higher values in amino acids, but lower values for the variables in the upper part of the map. At the boundary between the Lagers group and the non-grouped set, a sub-group of four samples (H) can be found. This small group is characterized by high values in amino acids and medium values for the maltose group.

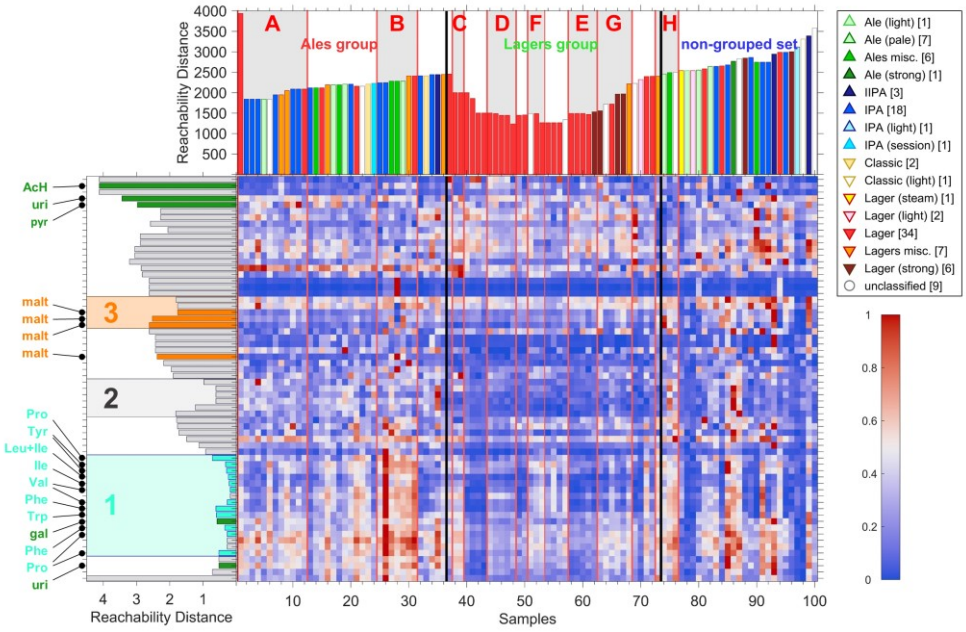


Figure 3.66. Heatmap of NMR features with Reachability Plots: variables' RP on the left side (OPTICS performed as described in the caption of Figure 4), samples' RP on top (k = 5). The samples are reordered according to the OPTICS sequence obtained from the Fused Adjacency Matrix (as in Figure 6). The dataset was normalized between zero and one to enhance its visual representation and interpretability.

This visualization approach is very efficient when dealing with data such as extracted features, while in the case of continuous data (e.g. spectra, chromatograms) reordering the original variables would make the visual interpretation very difficult.

An example with the Vis and NIR cases is given in Figure 3.67 and 3.68 respectively, without having performed variables reordering. In the case of Vis (Figure 3.67) different intensity of the absorption bands between the two main Ales and Lagers group can be observed, while for the NIR case (Figure 3.68) the pattern is not so clear to interpret and differences in absorption intensity, for most of the spectral regions, are highlighted only for the non-grouped set.

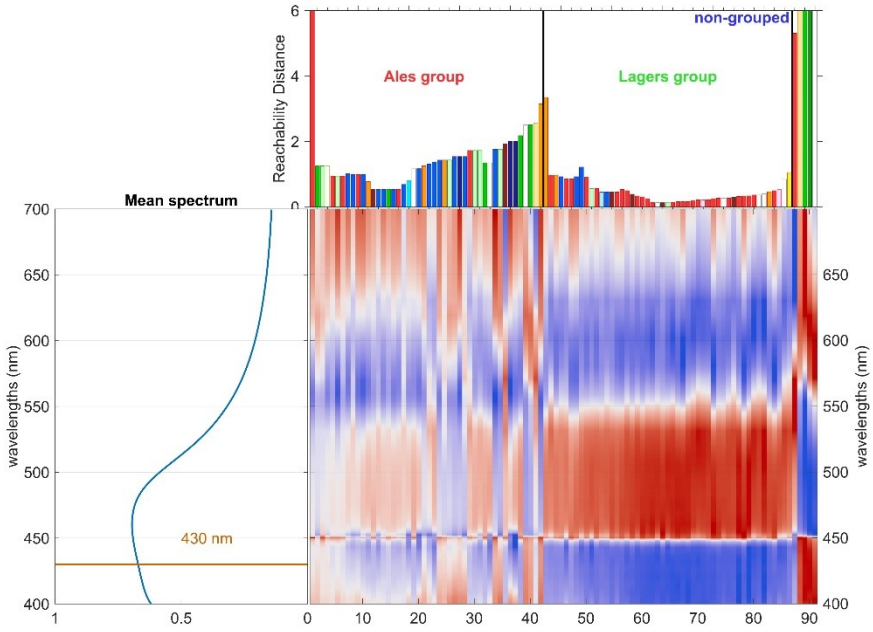


Figure 3.67. Heatmap of the Visible spectra with the samples' RP on top (k = 5). In the central part of the figure it is shown the heatmap obtained by reordering the samples according to the OPTICS sequence. The dataset was normalized between zero and one to enhance its visual representation and interpretability.

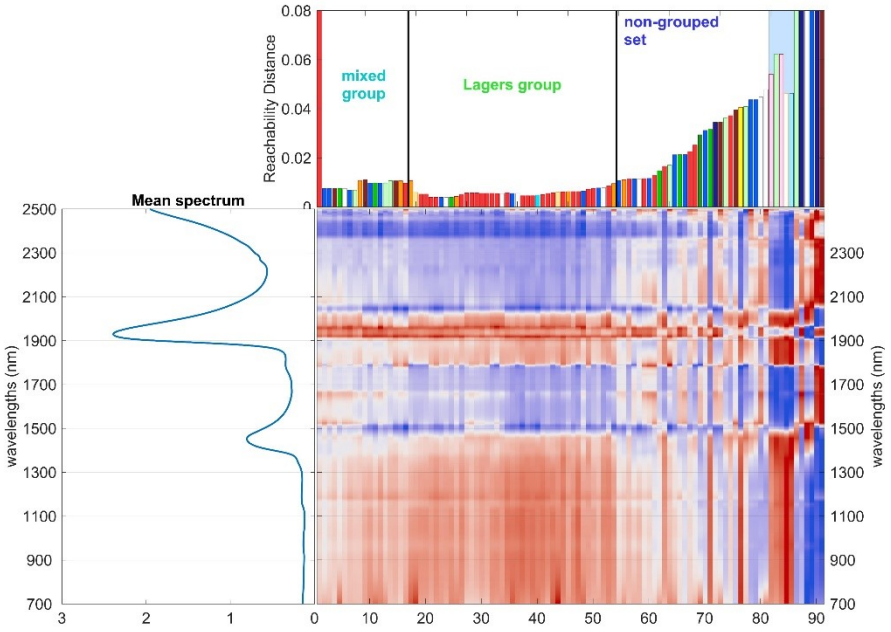


Figure 3.68. Heatmap of the NIR spectra with the samples' RP on top (k = 5). In the central part of the figure it is shown the heatmap obtained by reordering the samples according to the OPTICS sequence. The dataset was normalized between zero and one to enhance its visual representation and interpretability.

3.8. Conclusions

The Fused Adjacency Matrix approach and some of its applications were described in this chapter. It has been shown that the approach can recover coherent information from datasets of different nature with highly complex structures, highlighting groups and trends.

Two different (but somehow linked, see Section 2.3.1) food datasets were used to test the approach, one for exploratory analysis, and the other both for exploratory and mid-level data fusion purposes. As it should be expected from a data fusion approach, the Fused Adjacency Matrix is able to retain the information of the original datasets, but also to reveal other features arising from the combination of the fused sources.

Four simulated datasets were also used for testing the approach, with the aim of assessing the influence of its different steps on the final output, i.e. the Fused Adjacency Matrix. The approach performed quite well in almost all cases, but even if these results are promising, it is no secret that the further tests and improvements are needed.

For instance, the issue of linking back to the original variables should be investigated, as the representations given in Figures like 3.66, 3.67 and 3.68 can be of help, but do not address the core of the problem: when distances are computed, the link with the original variables is lost, therefore a way of mathematically linking the obtained clusters to the starting variables would be the best direction to investigate. Automatic selection of the optimal parameters for building the AMs should be also considered, as tools like maximum connectivity and the number of non-zero elements in differences of subsequent AMs have proven to be useful, but in this way the selection still needs to be done manually.

Finally, a series of tests with new and even more various types of data can be recommended, also for studying how the different distance measures behave and influence the final output. Different data type may be better suitable for certain distance measures or for SOM, and the different possibilities should be investigated as well. An example of this variability due to the type of data can be found in Figure 3.65, where the different steps of the approach are compared using Procrustes analysis: it was found that the NMR data block used in that part of the study yielded a Mahalanobis adjacency matrix very different from the parent Mahalanobis distance matrix and the rest of the distance and adjacency matrices too. A possible explanation was found in the nature of the data block, since the NMR information was present as resolved features, instead of original preprocessed spectra, like the other two Visible and NIR data blocks.

References | Chapter 3

[1] M. Li Vigni, C. Durante, M. Cocchi, Exploratory Data Analysis, in: Data Handl. Sci. Technol., Elsevier, 2013: pp. 55–126. doi:10.1016/B978-0-444-59528-7.00003-X.

[2] P. Latinne, O. Debeir, C. Decaestecker, Combining Different Methods and Numbers of Weak Decision Trees, Pattern Anal. Appl. 5 (2002) 201–209. doi:10.1007/s100440200018.

[3] Chuanyi Ji, Sheng Ma, Combinations of weak classifiers, IEEE Trans. Neural Networks. 8 (1997) 32–42. doi:10.1109/72.554189.

[4] E. Borràs, J. Ferré, R. Boqué, M. Mestres, L. Aceña, O. Busto, Data fusion methodologies for food and beverage authentication and quality assessment – A review, Anal. Chim. Acta. 891 (2015) 1–14. doi:10.1016/j.aca.2015.04.042.

[5] R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart, The Mahalanobis distance, Chemom. Intell. Lab. Syst. 50 (2000) 1–18. doi:10.1016/S0169-7439(99)00047-7.

[6] R. Todeschini, D. Ballabio, V. Consonni, Distances and Other Dissimilarity Measures in Chemometrics, Encycl. Anal. Chem. Appl. Theory Instrum. (2015) 1–34. doi:10.1002/9780470027318.a9438.

[7] F. Marini, R. Bucci, A.L. Magrì, A.D. Magrì, Artificial neural networks in chemometrics: History, examples and perspectives, Microchem. J. 88 (2008) 178–185. doi:10.1016/j.microc.2007.11.008.

[8] T. Kohonen, Essentials of the self-organizing map, Neural Networks. 37 (2013) 52–65. doi:10.1016/J.NEUNET.2012.09.018.

[9] F. Marini, Artificial neural networks in foodstuff analyses: Trends and perspectives A review, Anal. Chim. Acta. 635 (2009) 121–131. doi:10.1016/J.ACA.2009.01.009.

[10] M. Silvestri, A. Elia, D. Bertelli, E. Salvatore, C. Durante, M. Li Vigni, A. Marchetti, M. Cocchi, A mid level data fusion strategy for the Varietal Classification of Lambrusco PDO wines, Chemom. Intell. Lab. Syst. 137 (2014) 181–189. doi:10.1016/j.chemolab.2014.06.012.

[11] M. Bevilacqua, R. Bucci, A.D. Magrì, A.L. Magrì, F. Marini, Data Fusion for Food Authentication. Combining near and Mid Infrared to Trace the Origin of Extra Virgin Olive Oils, NIR News. 24 (2013) 12–15. doi:10.1255/nirn.1355.

[12] A. Biancolillo, R. Bucci, A.L. Magrì, A.D. Magrì, F. Marini, Data-fusion for multiplatform characterization of an italian craft beer aimed at its authentication, Anal. Chim. Acta. 820 (2014) 23–31. doi:10.1016/j.aca.2014.02.024.

[13] L. Breiman, Random Forests, Mach. Learn. 45 (2001) 5–32. doi:10.1023/A:1010933404324.

[14] Yan-Yong Xu, Xian-Zhong Zhou, Zhong-Wei Guo, Weak learning algorithm for multi-label multiclass text categorization, in: Proceedings. Int. Conf. Mach. Learn. Cybern., IEEE, 2002: pp. 890–894. doi:10.1109/ICMLC.2002.1174511.

[15] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998) 226–239. doi:10.1109/34.667881.

[16] L.I. Kuncheva, Combining pattern classifiers - methods and algorithms, 2014.

[17] A.-O. Boudraa, A. Bentabet, F. Salzenstein, Dempster-Shafer's Basic Probability Assignment Based on Fuzzy Membership Functions, ELCVIA Electron. Lett. Comput. Vis. Image Anal. 4 (2004) 1. doi:10.5565/rev/elcvia.68.

[18] B. Brownfield, T. Lemos, J.H. Kalivas, Consensus Classification Using Non-Optimized Classifiers, Anal. Chem. 90 (2018) 4429–4437. doi:10.1021/acs.analchem.7b04399.

[19] F. Marini, A.L. Magrì, F. Balestrieri, F. Fabretti, D. Marini, Supervised pattern recognition applied to the discrimination of the floral origin of six types of Italian honey samples, in: Anal. Chim. Acta, 2004. doi:10.1016/j.aca.2004.01.013.

[20] V. Simon, J. Gasteiger, J. Zupan, A combined application of two different neural network types for the prediction of chemical reactivity, J. Am. Chem. Soc. 115 (1993) 9148–9159. doi:10.1021/ja00073a034.

[21] R. Vitale, O.E. de Noord, A. Ferrer, A kernel-based approach for fault diagnosis in batch processes, J. Chemom. 28 (2014) S697–S707. doi:10.1002/cem.2629.

[22] L. Eriksson, E. Johansson, N. Kettaneh-Wold, S. Wold, Multi- and Megavariate Data Analysis: Part I – Basic Principles and Applications, Umetrics Acedemy, 2013.

[23] B.M. Harrison, F.G. Priest, Composition of Peats Used in the Preparation of Malt for Scotch Whisky Production-Influence of Geographical Source and Extraction Depth, J. Agric. Food Chem. 57 (2009) 2385–2391. doi:10.1021/jf803556y.

[24] W. Kew, I. Goodall, D. Clarke, D. Uhrín, Chemical Diversity and Complexity of Scotch Whisky as Revealed by High-Resolution Mass Spectrometry, J. Am. Soc. Mass Spectrom. 28 (2017) 200–213. doi:10.1007/s13361-016-1513-y.

[25] N. Cavallini, F. Savorani, R. Bro, M. Cocchi, Fused Adjacency Matrices to enhance information extraction: the beer benchmark, Anal. Chim. Acta. (2019). doi:10.1016/J.ACA.2019.02.023.

[26] T.H. Shellhammer, Beer color, in: Beer, Academic Press, 2009: pp. 213–227. doi:10.1016/B978-0-12-669201-3.00007-5.

[27] S. Engelhard, H.-G. Löhmannsröben, F. Schael, Quantifying Ethanol Content of Beer Using Interpretive Near-Infrared Spectroscopy, Appl. Spectrosc. 58 (2004) 1205–1209. doi:10.1366/0003702042336000.

[28] D. Intelmann, G. Haseleu, A. Dunkel, A. Lagemann, A. Stephan, T. Hofmann, Comprehensive Sensomics Analysis of Hop-Derived Bitter Compounds during Storage of Beer, J. Agric. Food Chem. 59 (2011) 1939–1953. doi:10.1021/jf104392y.

[29] I. Stanimirova, C. Boucon, B. Walczak, Relating gas chromatographic profiles to sensory measurements describing the end products of the Maillard reaction, Talanta. 83 (2011) 1239–1246. doi:10.1016/J.TALANTA.2010.09.018.

[30] J.M. Andrade, M.P. Gómez-Carracedo, W. Krzanowski, M. Kubista, Procrustes rotation in analytical chemistry, a tutorial, Chemom. Intell. Lab. Syst. 72 (2004) 123–132. doi:10.1016/J.CHEMOLAB.2004.01.007.

[31] P.D. Wentzell, S. Hou, C.S. Silva, C.C. Wicks, M.F. Pimentel, Procrustes rotation as a diagnostic tool

for projection pursuit analysis, Anal. Chim. Acta. 877 (2015) 51–63. doi:10.1016/j.aca.2015.03.006.

[32] I.F. Duarte, A. Barros, C. Almeida, M. Spraul, A.M. Gil, Multivariate Analysis of NMR and FTIR Data as a Potential Tool for the Quality Control of Beer, J. Agric. Food Chem. 52 (2004) 1031–1038. doi:10.1021/jf030659z.

Chapter 4 |
Beer's linguistics and chemistry: an investigation

4.1. Introduction

During the last decade the consumers' interest in how food is produced and prepared has strongly increased. Consumers tend nowadays to be more aware about the different aspects regarding food consumption and, in line with this trend, new-concept restaurants, new food production techniques and experiments on recipes and food pairings are constantly developed. This phenomenon is driven by high quality standards and often speaks a language based on what can be called the "craft rhetoric", for which craft/handmade is opposed to industrial [1], and mass production is opposed to artisanal [2]. Certainly, this kind of rhetoric is not confined to food consumption, but the food world provides clear examples of this trend [3]. During the last two decades, the beer industry has undergone massive changes, led by the explosion of craft and micro-breweries [4,5] and the spread of home brewing.

Analytical chemistry in synergy with advanced data analysis can be profitably used to build new tools to aid consumers when choosing and pairing foodstuff, and producers to meet the consumers' expectations. In this perspective, the aim of the present study is to investigate the links between the "objective" world of analytical chemical profiling – e.g. using spectroscopy – and the "subjective" world of consumers tasting and describing food.

Beer has been investigated both from the point of view of its chemistry and composition [6–8] and also from the point of view of the consumers' preferences [1,9]. Consumer's preferences are traditionally assessed by directly interviewing small groups of people, but with the growth of the Internet and its web communities, mining online-posted reviews has become an interesting approach for assessing product appreciation and reception [10,11]. Huge amounts of user-generated data are available today in very different formats, such as numeric scores, logical scores (in the form of like/dislike), geotags and written descriptions.

The case under examination is about the Beer datasets described in Section 2.3.2 and their combinations with user-generated reviews mined from a website[1]. Text analysis methods [12,13] were applied to process the user-generated reviews and convert them into numeric format, by the *bag-of-words* approach [12]. Principal component analysis–generalized canonical analysis (PCA–GCA, [14]) was used to investigate the links between spectral and text data. To select subsets of terms from the text data two approaches were used: topics extraction [15] using penalized matrix decomposition [16] and manually-defined sets of terms related to specific aspects of beer making and tasting.

4.2. Materials and methods

This materials and methods section is structured as follows: the first part is devoted to introducing the text analysis methods used for processing (Section 4.2.1.3) and visualizing (Section 4.2.1.1) the user-generated reviews; then, the text data collection will be described in Section 4.2.1.2 and some notes about the spectral datasets used in this study will be given in Section 4.2.2; finally, the method used for linking the spectral data to the text data will be introduced in Section 4.2.3.

4.2.1. Text analysis

The aim of text analysis [12,17] methods is to extract information from text documents by converting the text data to a format suitable for analysis. With these methods, it is possible to operate on different levels of detail [12,17], from the simplest approach of counting the occurrences of words or groups of words (i.e. their count or frequency of appearance in the processed documents) to the analysis of whole sentences with methods like latent semantic analysis ([18] i.e. extraction of the semantic structure of text by considering relationships and relative positions among words) and sentiment analysis [19,20].

In the present study, the text data were converted to wordcounts. This approach envisages the creation of a so-called *bag-of-words* model [12], which consists of a list of

[1] https://www.ratebeer.com/ (accessed: 16/02/2019)

terms or *vocabulary*, and their counts in each document used for building the model. The wordcounts are arranged in an array whose rows usually represent the documents (or samples) and each column is associated with one term. This matrix can be analysed with common multivariate methods, and the results can be interpreted according to the most influent terms.

This section is dedicated to describing the text data collection (4.2.1.2) and the processing methods used in the present work (Section 4.2.1.3), as well as the visualization techniques useful for inspecting and representing the extracted information (Section 4.2.1.1). Then, the preprocessing of the wordcounts matrix is described (Section 4.2.1.4) and the method for extracting meaningful topics from the data is reported (Section 4.2.1.5).

4.2.1.1. Text data visualization using word clouds

Word clouds are a visualization tool [21] that allows to clearly highlight the most important terms of a collection of words. The importance of a term is usually determined by its frequency of appearance within the text corpus.

Many online tools[2] and software-specific toolboxes[3] allow to generate word clouds, but the resulting representations may be very different from the points of view of interpretation and clarity. Online and basic tools usually provide a representation in which the words assume a quite cluttered distribution and may also end up being rotated, depending on the packing scheme of the tool. Such a representation is usually difficult to read and interpret [22], especially if a colour- or size-code is not implemented. The word clouds of this chapter were generated using the MATLAB Text Analytics toolbox, which provides a clear, uncluttered visualization: the terms' relative importance is depicted by using both a colour code and a size code, with the most important/frequent words in orange (by default) and bigger sizes, and the less frequent/important ones smaller and in black.

---

[2] Online word clouds generator tools: https://www.wordclouds.com/, https://www.jasondavies.com/wordcloud/

[3] MATLAB function: https://se.mathworks.com/help/matlab/ref/wordcloud.html (accessed: 11/02/2019)
R function: https://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf (accessed: 11/02/2019)
Python function: https://github.com/amueller/word_cloud (accessed: 11/02/2019)

4.2.1.2. Text data collection

All the user's comments and ratings were obtained from Ratebeer.com, a free online community consisting of a website on which the users can review any beer they have tasted by writing a comment and giving scores according to five parameters (aroma, appearance, taste, palate, overall, described in Table 4.1). Many users, especially the most prolific in tasting and reviewing, also use descriptors taken from the sensory "standard" terminology in their written comments, probably following the approach of considering the different aspects of the rating parameters. A more detailed description of the standard terminology of beer tasting assessment is given in Section 4.3.

Table 4.1. Definitions of the rating parameters of RateBeer.com

parameter	definition
aroma /10	what can be appreciated with the smell (both ortho- and retro-nasal)
appearance /5	anything concerning the beer look (colour, liquid's visual texture, head)
taste /10	what can be appreciated with the tongue (sweet, bitter, sour, salt and umami)
palate /5	what can be physically sensed with the lips, tongue, gums and roof of the mouth (also called "mouthfeel")
overall /20	evaluation of the overall sensory experience, as a measurement of the person's own appreciation

The text dataset contains 88 samples instead of 100 (as the Beer spectral datasets described in Section 2.3.2) because some of the beer samples were replicates that underwent thermal treatment. The thermally treated samples were removed, since the non-treated samples can be considered more like the common products that consumers can buy, taste and review.

4.2.1.3. Text processing

Text processing consists of a series of tools aimed at converting the input text corpus into a format suitable for further analysis. To this aim, tools from Natural Language Processing (NLP, [23,24]) are generally used. A "natural language" is any naturally

evolved language, as developed by humans through its use and repetition, without conscious premeditation or planning.

The result of converting text into numbers is a text dataset characterized by its own *vocabulary*, i.e. all the single words/entities contained in the dataset. The approach of counting how many times each word has occurred in the input text corpora is called *bag-of-words* modelling [10,12,25,26]. If only the occurrence of single words is considered (e.g. “warpigs”, “galaxy”, “awesome”), then the distance relationships among words are neglected (“warpigs galaxy is awesome”, “warpigs”, “galaxy” and “awesome” are be considered separately) and the model would be a bag-of-words of *unigrams*. If instead the occurrences of groups of *n* words are considered (e.g. “warpigs galaxy awesome”), the model would be a bag-of-words of on *n*-grams. Each bag-of-words has its own vocabulary and wordcount matrix, an array whose rows correspond to the input single documents and the columns correspond to one term of the vocabulary each. In the present work the unigrams bag-of-words model was used.

The information stored in a bag-of-words model can be represented using word clouds, which are visual devices for representing the most frequent terms (Section 4.2.1.1). Figure 4.1 shows the data processing pipeline, with one word cloud corresponding to each step. The first step (a) shows very clearly how important the cleaning processing is, since if nothing was removed, punctuation and very common and uninformative particles (e.g. “and”, “the”, “is”, “with”) would be strongly predominant.

Figure 4.1. Text processing pipeline, (a) from the raw data to the clean data and (b) from the clean data to the clean data after selecting the English language. Because of the strong predominance of English over the other languages, no apparent difference is detectable before and after step (b).

4.2.1.3.1. Text cleaning

To convert the data from text to a numeric format suitable for analysis, a series of cleaning steps is required. Operations such as conversion to lower case and removal of numbers, punctuation, stop words, very short or long words are generally implemented.

In the present study, the following list of cleaning steps was followed (references suggesting each operation are also provided):

1. remove HTML tags [25,26];
2. remove HTML entities [25,26];
3. remove URLs [25];
4. correct special characters;
5. remove numbers [10];
6. convert to lower case [10,26];
7. erase punctuation [10,25];
8. remove very common words, stop words and syntactic particles [10,13,27];
9. remove short (<3 characters) and long words (>10 characters);
10. words normalization[4] or *lemmization* (Porter stemmer, [28]) [13];
11. remove uninformative terms;
12. select English (see Section 4.2.1.3.2);
13. remove more uninformative terms;
14. remove infrequent terms (document-wise at least 5% of the documents) [13,26,27];
15. remove infrequent terms (<50 overall) [13].

This cleaning sequence resulted in a bag-of-words consisting of 662 terms. The procedure for selecting the English terms is described in the next Section.

4.2.1.3.2. English selection

In the present work it was decided to focus on the English language only. This allows to work with a more "homogeneous" dataset, avoiding spreading parts of the same information across different terms/variables. For instance, the occurrences of the words "hop", "hops" and "hoppy" would be gathered under the term "hop" by the

[4] In this context, "normalization" has the meaning of "reducing to the same root". Also, the used MATLAB function for doing this type of lemmization reduced all words ending in *-y* to their root, ending in *-i*. This is why some of the terms represented in the wordclouds of this chapter strangel*i* end in *-i*.

lemmization preprocessing step (Porter stemmer, bullet point 10 in Section 4.2.1.3.1). However, the words for "hop" in other languages can be very different: "luppolo" in Italian, "humle" in Danish, "hopfen" in German, "houblon" in French, "chmiel" in Polish. Without translating and merging all these variants of the same word information may end up quite fragmented over many variables.

The reason why the English language is strongly predominant is found in the source website, which is US-based: it is therefore very likely the most people using it are English native speakers. The success of the platform has also attracted foreign users, who in many cases decided to write in English as well. A cross-section of the languages used on the website is given in Figure 4.2.

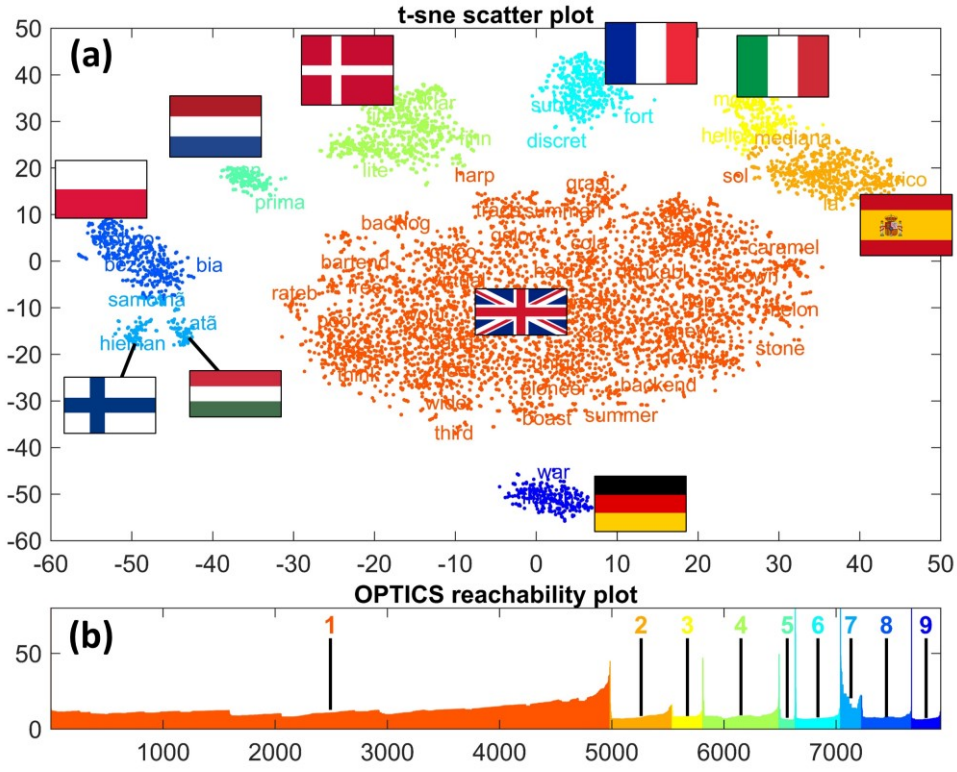


Figure 4.2. Language clusters form the text data: (a) scatter plot obtained by t-SNE, colored according to the clusters identified by OPTICS (b), operated on the two-dimensional output of t-SNE.

Each point on the scatter plot represents one term from the text dataset, without any regard to its count (as opposed to the word cloud representation). The recognizable languages rank in this decreasing order: English (62,83%), Danish (8,57%), Spanish (6,98%), Polish (5,61%), French (5,08%), Italian (3,44%), German (3,29%), Dutch (1,83%), Hungarian and Finnish (2,36%).

The procedure for selecting the English terms consisted of three steps:

i. Train a word embedding

   First, a *word embedding* was trained using the *word2vec* function from MATLAB's Text Analytics toolbox. Word embeddings[5] are a family of techniques for language modelling in the NLP context: in a word embedding, each word or phrase from the input text corpus is mapped to a vector of real numbers. The *word2vec*[6] [29] function is based on the original algorithm developed by Mikolov *et al.* [30] while working at Google in 2013. The word embedding model dimensionality was set to 100, meaning that each word was represented by a vector consisting of 100 elements.

ii. Plot the word embedding using t-SNE

   Then, the t-distributed stochastic neighbour embedding (t-SNE, [31]) algorithm was used for dimensionality reduction on the word embedding, resulting in the two-dimensional representation of Figure 4.2a. The very clearly defined clusters of terms corresponded to the individual languages used of the RateBeer website.

iii. Cluster identification and English selection using OPTICS

   To efficiently select the English cluster of terms, OPTICS (described in Section 2.1.1.5.1), which is a density-based method, was used. Each detected cluster (reachability plot in Figure 4.2b) was investigated and matched with the t-SNE scatter plot, leading to the identification of cluster #1 as the set of English terms.

All the terms belonging to the other clusters/languages were therefore removed from the dataset.

[5] Marcus Sahlgren - *A brief history of word embeddings (and some clarifications)*: https://www.linkedin.com/pulse/brief-history-word-embeddings-some-clarifications-magnus-sahlgren/

[6] *word2vec* can be freely downloaded from: https://code.google.com/archive/p/word2vec/ (accessed: 30/01/2019)

4.2.1.4. Wordcount data preprocessing

The traditional preprocessing of word counts matrices goes under the acronym of *tf-idf* (term frequency–inverse document frequency, [32]). This approach is usually employed for machine learning purposes and in the classification context [33,34]. The approach adopted in the present work is however based on a "chemometric perspective", in which one of the fundamental aims is to interpret the data and understand their structure, instead of "only using" them.

Therefore, a preprocessing consisting of row-normalization and autoscale was chosen. The wordcounts matrix was normalized using the sum along the rows, for accounting for the different numbers of reviews associated with each sample.

The actual number of reviews was also used for scaling when studying the most suitable preprocessing method, and it was found less efficient than the row sum. This is probably due to the fact that not only the number of reviews is important, but also their length and complexity, whose information may be better captured by the word counts. However, an effect related to the number of reviews was still found on PC1, but with reduced magnitude (Figure 4.3).

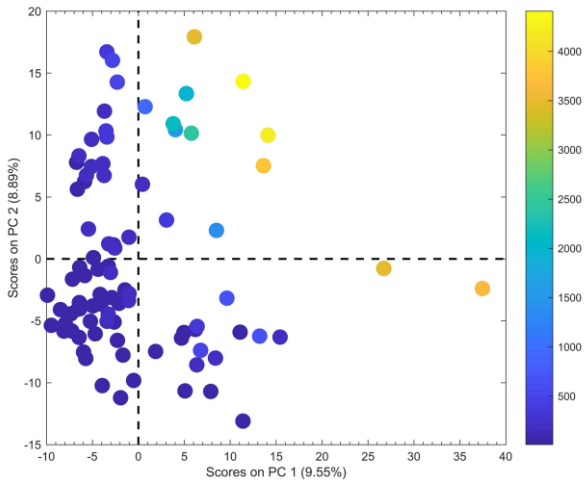


Figure 4.3. PCA on preprocessed word counts, colored according to the number of reviews.

4.2.1.5. Extraction of meaningful topics via penalized matrix decomposition (PMD)

In order to work with subsets of the wordcount matrix, many different ways for automatically defining groups of related terms can be used. A set of related words can be defined as a *topic* [15].

Many methods for automatically extracting meaningful topics were considered and tested, e.g. PCA ([35], Section 2.1.1.1), MCR ([36], Section 2.1.1.2), archetypical analysis [37,38], coclustering ([39], Section 2.1.1.4), latent Dirichlet allocation [40] and penalized matrix decomposition (PMD, [16]). Since the two main aims were obtaining meaningful groups of terms and that these groups should have reduced dimensionality, PMD was chosen for the quality of the obtained topics and the sparseness of the groups (groups with maximum 10 terms were obtained).

PMD is a method for obtaining meaningful clusters based on non-negative matrix factorization [41] which applies two-sided sparsity. This means that both scores and loadings are sparse, and this corresponds to the desired situation: the groups of terms that we are looking for should correspond to groups of beers, capturing the peculiar aspect that groups them. In other words, a topic can emerge if a specific feature or set of features is shared among groups of beers (i.e. the scores are sparse).

4.2.2. Spectral data

As described in Section 2.3.2, the beer spectral datasets were obtained using Visible, NIR and NMR spectroscopies. For the present study, a more refined version of the NMR features dataset than the one used in the data-fusion application of the Fused Adjacency Matrix approach (Chapter 3, Section 3.7) was used. Sixty-one instead of sixty-four features were newly extracted, and 56 out of 61 of them were given chemical names. This refinement was done at a later stage, for improving the interpretability of the present study, where the link between the sensory-like consumers' descriptions and the chemical features of beer is investigated.

Procrustes Analysis (Section 2.1.3) was done on the two versions of the NMR features dataset, resulting in an average Procrustes dissimilarity value of 0.15 (over twenty

principal components), which means that the two datasets carry similar information. The dissimilarity values according to the number of components is depicted in Figure 4.4.

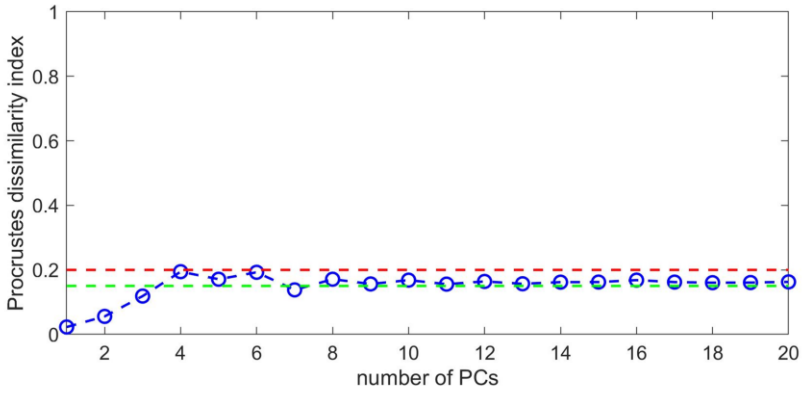


Figure 4.4. Procrustes similarity between the NMR datasets, over 20 principal components.

As stated in Section 4.2.1, eighty-eight samples were included in the text dataset and the same samples were selected from the original spectral datasets for the analysis of this study. The identified metabolites are reported on Table 4.2, together with their assignment, chemical shift, chemical class and the references that were used for the assignment.

Table 4.2. List of assigned compounds, and literature references. (*m = multiplicity, Chenomx is cited in Section 2.3.2.1.5.)

compound	m*, chemical shift (δ, ppm), assignment	chemical class	references
2-butanone	s, 2.19	misc.	Chenomx
2'-deoxyuridine	d, 7.84	nucleoside	Chenomx
2'-deoxyuridine	t, 6.28	nucleoside	Chenomx
2-octenoic acid	m, 5.79	acids	Chenomx
acetaldehyde	q, 9.67	aldehydes	[6]
acetaldehyde	d, 2.23, CH3	aldehydes	[6]
acetic acid	s, 2.22	acids	[42]
adenine	s, 8.21	nucleoside	[43], Chenomx
adenine	s, 8.18	nucleoside	[43]
alanine	d, 1.47	amino acid	[42]
alcohols	t, 0.88, CH3	alcohols	[6,44]
	d, 0.88, CH3		

compound	m*, chemical shift (δ, ppm), assignment	chemical class	references
	d, 0.88, CH3		
arginine	m, 1.95, γ-CH2	amino acid	[45]
choline	s, 3.18	cholines	[43]
dextrins		carbohydrates	[6,42]
dextrins	d, 5.09, α(1→6) glycosidic linkages	carbohydrates	[46,47]
dextrins	d, 5.08, α(1→6) glycosidic linkages	carbohydrates	[46,47]
dextrins	d, 5.01, α(1→6) glycosidic linkages	carbohydrates	[46,47]
dextrins	m, 4.57	carbohydrates	Chenomx
dextrins		carbohydrates	Chenomx
glucose	dd, 3.23	carbohydrates	[47], Chenomx
guanosine	s, 8.0	nucleoside	Chenomx
histidine	s, 7.99, C2H	amino acid	[6,42,48]
histidine	s, 7.03, C4H	amino acid	[6,42,48,49]
inosine	s, 8.28, CH4	nucleoside	[6], Chenomx
isoleucine	d, 0.995	amino acid	Chenomx
isopentanol	1.42, CH	alcohols	[6], Chenomx
lactic acid	d, 1.35, CH3	acids	[6,42,43,49]
leucine+isoleucine	bt, 0.96, δ-CH3	amino acid	[45]
	t, 0.95, δ-CH3		
maltose	d, 5.22, α-C1H	carbohydrates	[6,42]
maltose	d, 4.64, β-C1H	carbohydrates	[6,42]
maltose	t, 3.42	carbohydrates	[6], Chenomx
maltose	dd, 3.26, β-C2H	carbohydrates	[6]
methionine	m, 2.26, beta-CH2	amino acid	[45]
N-acetyltyrosine	m, 6.83	amino acid	Chenomx
phenylalanine	m, 7.35, C2H, C6H	amino acid	[45,48]
phosphocholine	s, 3.21, N-CH3	cholines	[43]
polyphenols	polyphenols	polyphenols	[48,49]
polyphenols	polyphenols	polyphenols	[48,49]
polyphenols	polyphenols	polyphenols	[48,49]
proline	m, 2.39, β-CH2	amino acid	[43,45]
proline	m, 2.34	amino acid	[42], Chenomx
proline	m, 2.34, β-CH2	amino acid	[6,42], Chenomx
propanol	m, 1.53, CH2	alcohols	[6]
pyruvate hydrate	s, 2.36, CH3	acids	[6,42,45,49]
pyruvate hydrate	s, 1.58, CH3	acids	[45]
trehalose	d, 5.18	carbohydrates	Chenomx
trigonelline	s, 9.11	misc.	[50], Chenomx
trigonelline	m, 8.82	misc.	[50], Chenomx
tryptophan	bd, 7.7, Ar-H	amino acid	[45]
tyrosine	d, 7.17, Ar-H	amino acid	[6,42,43,45,48,49]
tyrosine	d, 6.88, C3H, C5H	amino acid	[6,42,43,45,49]

compound	m*, chemical shift (δ, ppm), assignment	chemical class	references
unknown 1	s, 10.2		
unknown 2	s, 9.445		
unknown 3	s, 6.35		
unknown 4	s, 2.21		
unknown 5	s, 2.12		
uracil	d, 5.79	nucleoside	Chenomx
uridine	d, 7.86	nucleoside	[6,42,48,49]
uridine	m, 5.89, C1'H	nucleoside	[6,42,43,48]
valine	d, 1.06, γ-CH3	amino acid	[6,45,48,50]
valine	d, 1.02, γ-CH3	amino acid	[6,45,48,50]

4.2.3. Principal component analysis–generalized canonical correlation (PCA–GCA)

The issue of linking two datasets is at the core of the present study. Different approaches can be taken, and one of the most established chemometric methods for inspecting the relations between two data blocks is partial least squares (PLS, [51,52]) regression. However, even if PLS regression models using topics extracted from the text data were attempted, very poor results were obtained, suggesting that a strict prediction of single terms or even topics is probably not attainable.

For this reason, it was decided to investigate the connections between spectral and text data using a method able to find common subspaces among data blocks, which is a combination of principal component analysis and generalized canonical correlation (PCA–GCA, [14]).

The aim of PCA–GCA is to find linear combinations of the blocks under examination, fitting as well as possible a set of orthogonal common components. This set of common components is not necessarily contained in the column space of any data block, but it is instead in the combined space represented by the concatenated matrix obtained by joining the data blocks in the variables' direction [14].

GCA aim at finding the trends in the data blocks that correlate the strongest. Since this method only focuses on correlation, within-block variability may be not captured at best. PCA operated on the individual blocks prior to GCA helps enhancing the stability of the extracted common components [14]. Combining PCA with GCA leads to a stepwise procedure in which the first step is devoted at determining the subspaces'

dimensionalities (each block having its own dimensionality). Then, the correlation coefficients and the block-wise explained variances are obtained by GCA and are used for deciding the number of common components. Each common component is in fact obtained by separately estimating one component from each data block, in a way that all the block-related components are as similar as possible (i.e. they are as strongly correlated as possible).

The number of common components is chosen according to how strong the correlation is and by evaluating the amount of variance explained from each data block.

4.2.4. Software

All analyses were carried out under MATLAB environment (2017b, Mathworks, MA, USA).

Text data processing was done using the functions provided by the Text Analytics Toolbox (version 1, Mathworks, MA, USA) and in-house written scripts.

Word embedding visualization was performed using the **t-SNE** function contained in the MATLAB's Statistics and Machine Learning Toolbox (version 11.2).

The **PMD algorithm** used for topics extraction was written by Jose Camacho Paez and it is based on the work by Witten *et al.* [16].

The **PCA-GCA toolbox for MATLAB** [14] was written by Ingrid Måge and can be found at:
https://nofimamodeling.org/software-downloads-list/pca-gca-toolbox-for-matlab/
(last access 13/02/2019)

4.3. Terminology of beer flavour and aroma

Sensory description of foodstuff is generally done by means of a *lexicon* [53,54], which is a set of terms used for documenting and describing the sensory perceptions of a selected food [53]. By defining a lexicon, it is possible to perform sensory studies with a panel of evaluators who are "talking the same language", so that the results of their assessments can be compared and quantitatively elaborated. During the years, the beer

flavour terminology [55] has been constantly enriched, updated and used for characterizing [56] different beer products, but it also made easier the communication between stakeholders, scientists and also consumers: a clear example of how this terminology has spread and reached the consumers is given in Section 4.3.2 and in Figure 4.6, where the experimental *beer vocabulary* is described. In text analysis, a set of terms is usually referred to as a *vocabulary*.

4.3.1. The beer flavour wheel

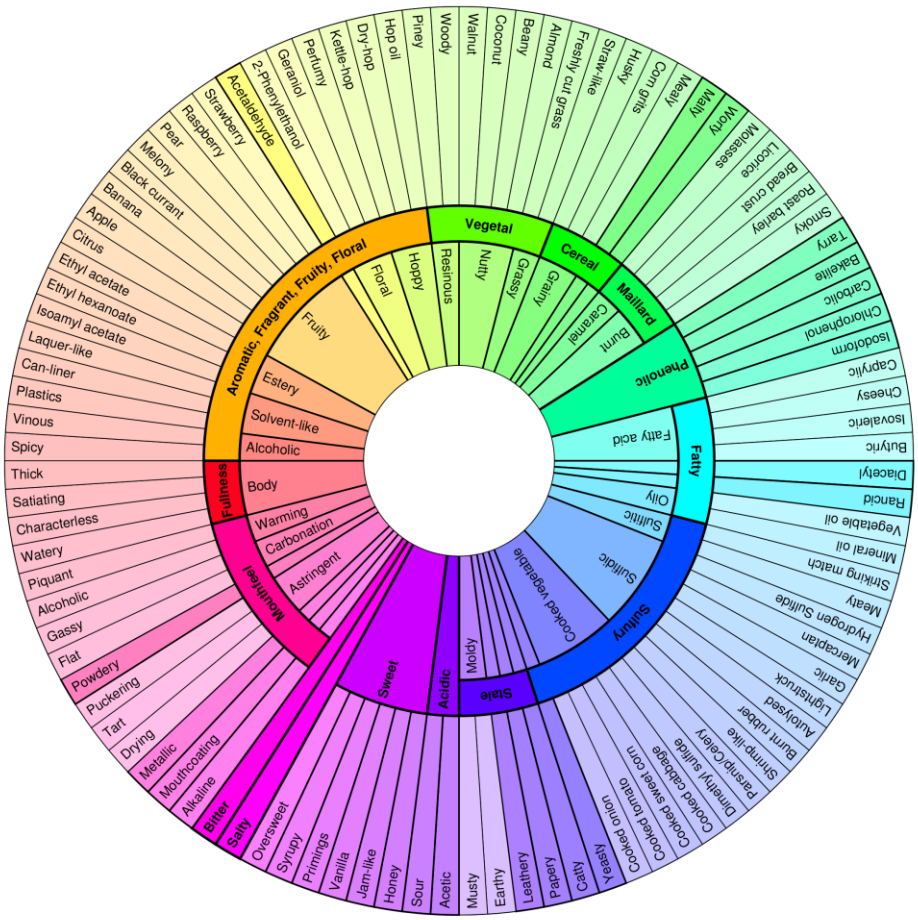


Figure 4.5. The beer flavour wheel[7].

[7] Downloaded from: http://www.beerflavorwheel.com/ (accessed: 16/02/2019)

Figure 4.5 shows an example of the beer flavour wheel, which was first introduced by Meilgaard *et al.* [55]. The purpose of this visual device is to allow locating quickly and easily any sensory term.

Meilgaard grouped the sensory terms into classes and organized them in tiers within each class. The middle ring with stronger background colour in Figure 4.5 represents the classes. On the external ring of the wheel are located the second-tier terms, which represent the highest level of detail in this classification system: as defined by Meilgaard, each "separately identifiable flavour characteristic has its own name" [55].

If the second-tier terms represent the most detailed description of a flavour characteristic, the first-tier terms represent the family to which these terms belong to: for instance, within the *Aromatic, fragrant, fruity, floral* class, the first-tier term *fruity* represents the flavour characteristics directly related to typical fruit flavours, like *citrus*, *apple*, *banana*, *blackcurrant*, *melon*, *pear*, *raspberry* and *strawberry*.

4.3.2. The experimental beer vocabulary

Using the cleaning procedure described in Section 4.2.1.3, the text data were reduced to a vocabulary of 662 terms, corresponding to a reduction of 99.09% from the starting text corpus. Figure 4.6 shows the top one-hundred terms of the experimental beer vocabulary, by frequency of use: many of these terms are clearly derived from the standard beer flavour terminology, as introduced with the flavour wheel of Section 4.3.1. In this study however, the vocabulary also includes many other "sensory-like" terms (which often are just synonyms of the reference sensory terms), but also many other descriptors covering the different aspects of beer tasting, e.g. its appearance, the situation/occasion of consuming it and the consumer's opinion about the product.

A more detailed representation of these different aspects is given in Figure 4.7. Seven categories were manually defined, according to macro aspects of beer tasting, such as appearance/colour (4.7d), recognition of malt- or hops-related flavours (4.7b and 4.7c), general sensory-like terms (4.7a), personal judgement (4.7f) and experience-memory (4.7e) linked to the consumption of beer.

Figure 4.6. Top one-hundred terms of the experimental beer vocabulary, by frequency.

Figure 4.7. The experimental beer vocabulary divided into sub-groups: (a) sensory-like, (b) malt-related, (c) hops-related, (d) appearance/colour-related, (e) situation/experience-related, (f) judgement-related and (g) mixed/confused terms.

As it will be clarified in the next section, of these seven manually-defined groups of terms, only the hops- and appearance/colour-related groups (c and d in Figure 4.7) were used in the modelling steps, mainly as a sort of benchmark for assessing some of the topics automatically extracted by PMD (Section 4.2.1.5). A summary of the use of groups c and d is given in Table 4.3.

4.3.3. Extracted topics

Figure 4.8. Twenty topics extracted using PMD. Highlighted in yellow the topics that are discussed in the Results section.

Twenty topics were extracted from the wordcounts data using PMD and are represented in Figure 4.8. Twelve of them (#1, #5, #6, #7, #8, #10, #11, #12, #15, #16, #18 and #20) have clear links to the sensory world, being characterized by terms such as *pine*,

hophead, *wood*, *roast*, *syrup*, *earth*, *copper*, *alcohol*, *mint*, *vanilla*, *lime*, *lemon*, *caramel*, *pear*, *papaya*, *clove*, *spice*, *yeast*, *banana*, *wheat*, *caramel*, *coffee* and *pepper*. Many of these terms come directly from the standard beer flavour terminology.

Other topics, like #2 and #9, seem to be related to negative experiences, with characteristic terms, sometimes vulgar, such as *waste*, *ass*, *piss*, *urine*, *suck*, *headache*, *blah* or *cost*, probably originated from products that did not quite match quality/price expectations. On the contrary, topics like #4, #17 and #19 seem to be related to positive experiences (*goto*, *gateway*, *reliable*, *staple*, *favorit*, *flagship*), bringing back memories (*memory*, *reliable*) and specific situations in which thirst needs to be quenched (*summer*, *beach*). Finally, topics like #3 and #13 include very mixed terms that hinder the identification of a common theme.

4.3.4. Use of the text data and summary on data analysis workflow

At this point, it can be useful to explain how the text data will be used as well as summarise the data analysis flow, which is represented in Figure 4.9. After the vocabulary was refined by applying the text cleaning steps of Section 4.2.1.3.1, subsets of words were identified with two methods:

a. manual definition of *groups of terms* (Figure 4.7);
b. automatic *topic* extraction by PMD (Section 4.2.1.5).

Please notice that even if *groups of terms* and *topics* are basically synonyms, it was decided to refer to the manually-defined topics as "groups of terms" and to use "topics" for referring to the ones automatically obtained by PMD.

All topics (Figure 4.8) were associated to the spectral data by PCA–GCA, and only those showing correlations above 0.7–0.75 were further inspected. The results of four of them (in yellow in Figure 4.8) were then reported in this Chapter. Since two of the topics, namely #1 and #5, had an "overall theme" logically connected with two of the manually defined groups, it was decided to compare them, in relation to the same spectral data. A summary of these comparisons is given in Table 4.3.

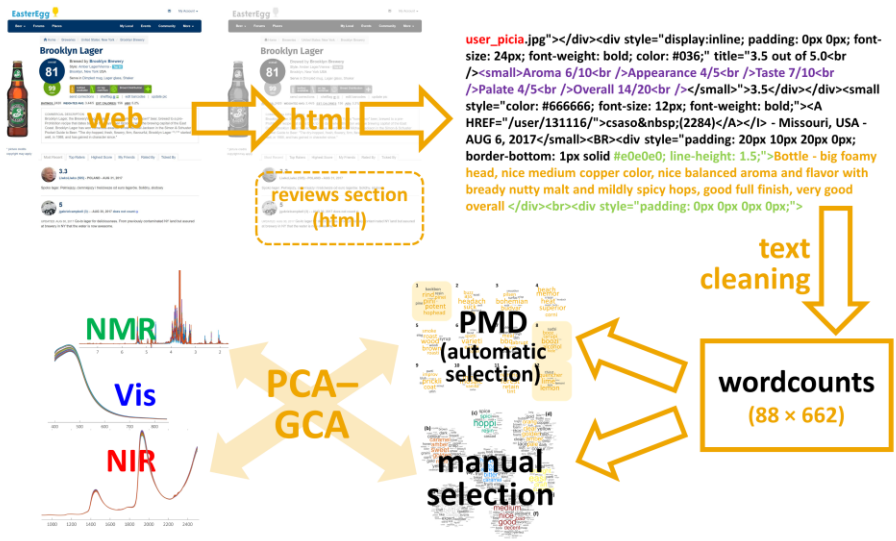


Figure 4.9. Data analysis workflow.

4.4. Results

The results obtained by PCA–GCA linking the spectral datasets with subsets (i.e. the topics) of the text data are reported in this section. First, twenty topics were extracted by PMD (Section 4.4.1, Figure 4.8) from the preprocessed wordcounts (Section 4.2.1.4).

4.4.1. Linking the spectral data to the topics

In general, only results with common component correlations > 0.75 were considered, with a minor exception of topic #8, whose PCA–GCA models resulted in correlations of 0.74 (Figures 4.16 and 4.15). In this section, four topics from Figure 4.8 are discussed, in relation to their correlation with the spectral datasets.

4.4.1.1. *Hops*

Topic #1 provided the best common component correlation result with the NMR data (0.91) and seems related to the hops [58]. Terms like *resin* and *pine* (in the Counts loadings) naturally refer to the resins extracted from the hops' cones during the boiling

step of beer wort [59]. *Piney* is also a term from the beer flavour terminology, and it can be found under the *vegetal* class and the first-tier term *resinous*, in the beer flavour wheel (Figure 4.5). This also confirms the topic's association with the hops.

Trigonelline, from the NMR loadings, is a plant metabolite generally found and studied in relation to coffee [60,61] for its pharmacological [62,63] and health benefits [61]. It has recently been found in beer and described as a plant-associate metabolite whose concentration increases with boiling [50]. Hops are generally added right before boiling the beer wort, so that heat allows converting the hop acids to become soluble and extracting them. For these reasons, trigonelline can also be associated with the hops.

Finally, since the loadings associated with *resin*, *pine* and *trigonelline* share the same direction, a connection between them can be deduced, confirming that topic #1 can be related to the hops.

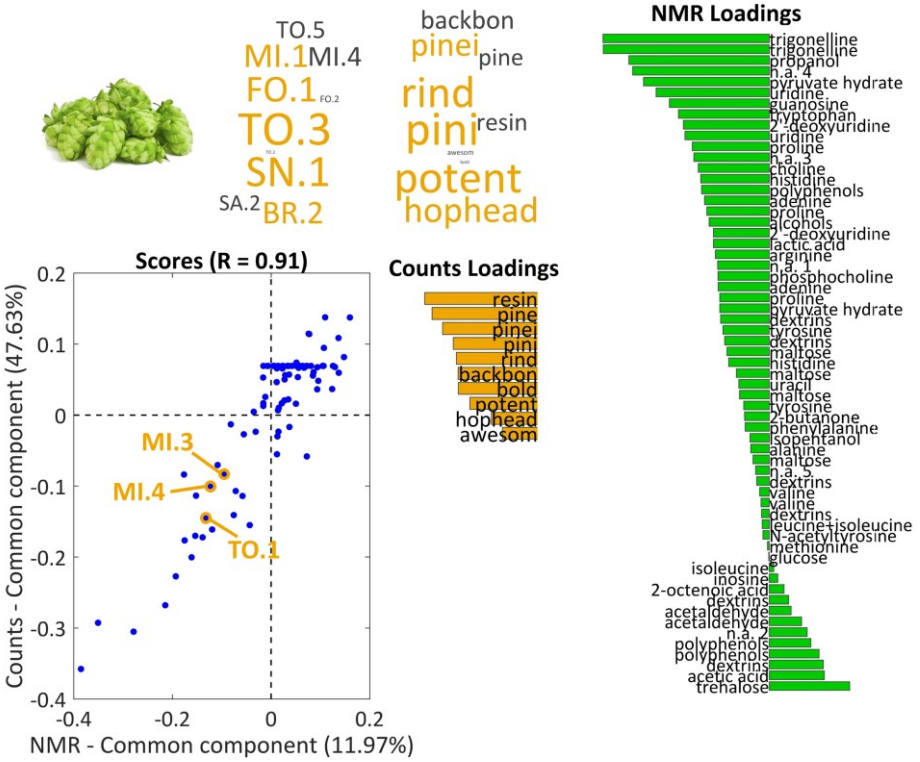


Figure 4.10. *Hops*. PCA–GCA scores obtained by comparing the NMR dataset (88×61) and topic#1 (88×10).

By inspecting the samples' distribution in the score plot of Figure 4.10, at negative scores values mostly IIPAs, IPAs and ales are found. The presence of three lagers can be explained with their recipes, which are rich in hops: sample MI.3 (*Helping Hand*, Mikkeller) is described as a "hoppy pilsner" [8]; sample TO.1 (*Hop Love Pils*, To Øl) has "hop" in the name and is described as "brewed with lots of hops" [9]; sample MI.4 (*American Dream*, Mikkeller) is described as "packed with American hops" [10]. Terms like *bold* and *potent* also contribute to the samples' separation along the scores: the most extreme samples at negative scores in Figure 4.10 belong to IIPA style, which is an acronym for Imperial India Pale Ale. The attribute "imperial" is generally used for very strong beers, both from the point of view of alcoholic strength and flavour richness. A clear link with the terms *bold* and *potent* can therefore be recognized.

On the opposite end of the plot, at positive scores, only lagers from producers such as Hite, Heineken, Budwiser, Pilsner Urquell and San Miguel are present. These are very widespread products, and their style does not involve much addition of hops or spices. They seem to be mainly characterized by variables mostly related to sugars (*dextrins* and *trehalose*) and malt (*polyphenols*), as if in absence of a rich/peculiar bouquet of flavours the most basic taste of beer emerges. This is also confirmed by the opposite direction of the topic's terms *bold* and *potent*, which are logically distant from beers with more common flavours.

An interesting contrast can be identified between two metabolites from the NMR loadings: trehalose and pyruvate (hydrate). *Trehalose* [64] is a disaccharide that is involved in the anaerobic carbohydrate metabolism in yeast cells, as an intermediate on the path for the formation of glycogen [65], an "energy storage" compound for yeast cells. *Pyruvate*, on the contrary is an intermediate on the path that leads to the production of ethanol, alcohols, aldehydes and esters. The opposite directions that *trehalose* and *pyruvate* have also correspond to the two main beer style families, ales and lagers. Ales beers tend to be richer in flavour and have higher alcoholic strength, a

[8] "Helping Hand" on Untappd: https://untappd.com/b/mikkeller-helping-hand/818743 (accessed: 15/02/2019)
[9] "Hop Love Pils" on RateBeer: https://www.ratebeer.com/beer/to-ol-hop-love-pils/250560/ (accessed: 15/02/2019)
[10] "American Dream" on RateBeer: https://www.ratebeer.com/beer/mikkeller-american-dream/110815/ (accessed: 15/02/2019)

product that can be related to a fermentation process in which the yeasts produce a larger variety of metabolites. A link with the *pyruvate* path can therefore be traced, as opposed to the production of lagers, where the yeasts may also express to a large extent the metabolic path related to *trehalose* and glycogen.

It is interesting to notice that in the Counts scores there is a set of samples that all share the same score value: the words belonging to this topic may have been used in an extremely similar way for these samples, which may have ended up practically identical from the terms' point of view.

The results shown in Figure 4.10 closely resemble those obtained from the model built using all the terms related to the hops: as shown in Figure 4.11, similar common component correlation values were obtained, and the inclusion of more terms (not limited to 10, like in the topic's case) made possible to break the group of samples with very similar values along the Counts scores (Figure 4.10).

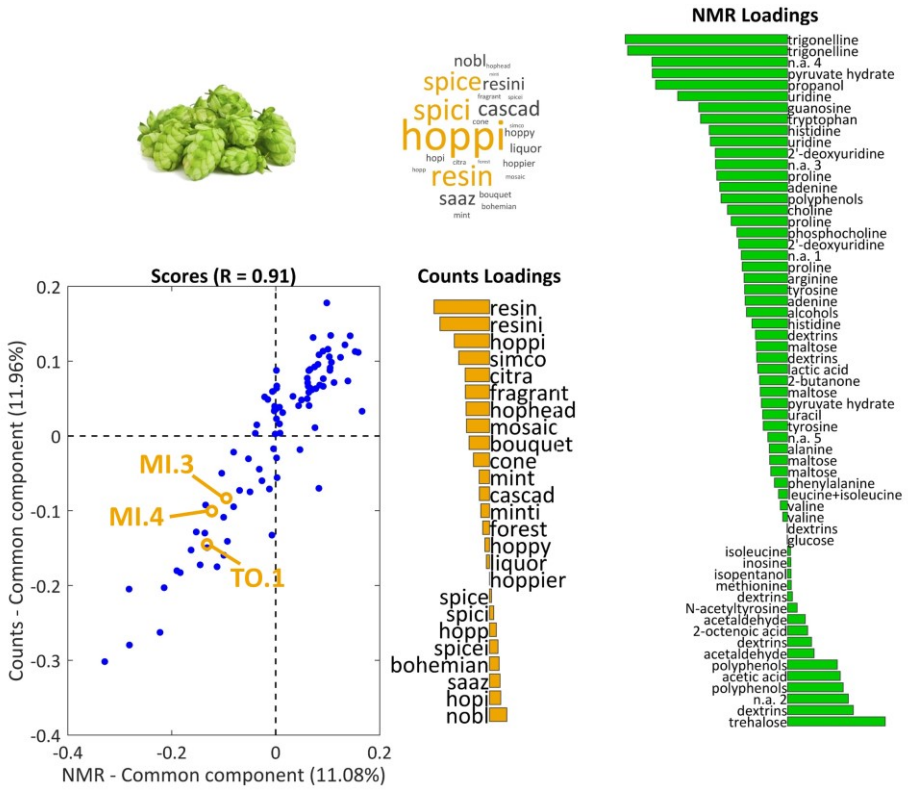


Figure 4.11. *Hops*. PCA–GCA scores obtained by comparing the NMR dataset (88×61) and the hops-related terms (88×25, group c in Figure 4.7).

4.4.1.2. *Brown colour*

Topic #5 is characterized by interesting terms such as *wood*, *brown*, *roast*, *syrup* and *dark*. This combination suggests that features related to beers with darker colours and brownish hues were captured by the topic. At first glance, the common component correlation value looks good as well, but just inspecting the distribution in the score plot of Figure 4.12 it is clear that one sample may be driving the correlation.

As a matter of fact, if sample SL.1 (at very negative scores in both components) is removed, the correlation value drops to 0.51, meaning that even if a connection between *glucose* and *maltose* and the brown/roasted colour of topic #5 seems plausible, the situation described in the figure may not be real.

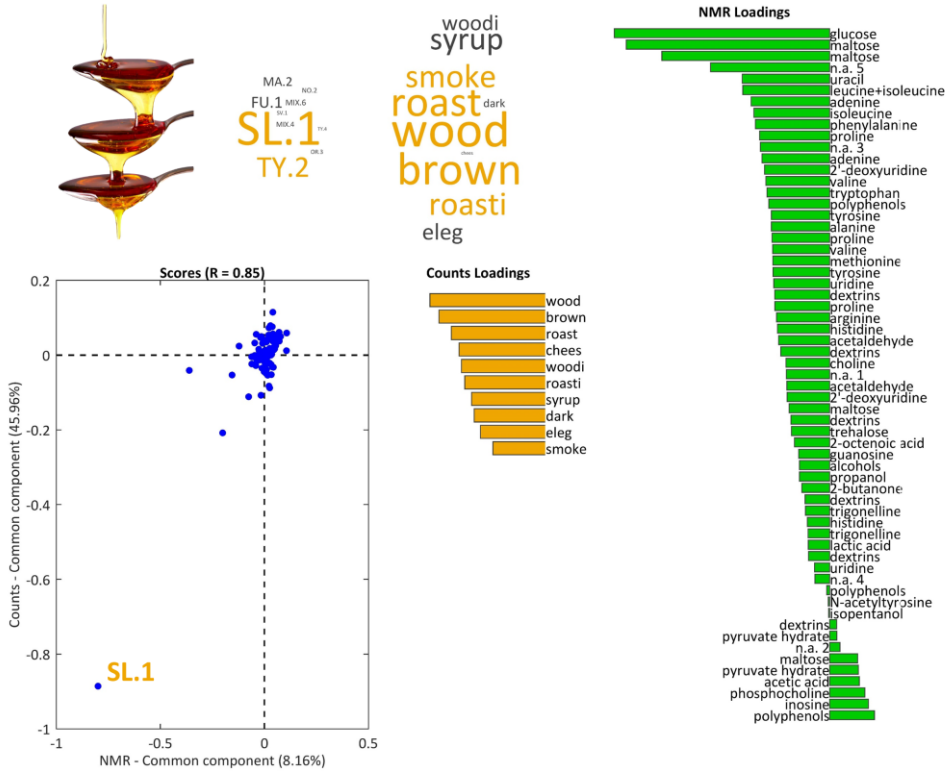


Figure 4.12. *Brown colour*. PCA–GCA scores obtained by comparing the NMR dataset (88×61) and topic#5 (88×10).

However, by inspecting the same topic in relation to the Vis data, a completely different situation is found: the result is a common component correlation of 0.75, as shown in Figure 4.13. *Dark*, *roast*, *brown* and *roasty* are the most correlated terms to the Vis data, whose interpretation results quite difficult.

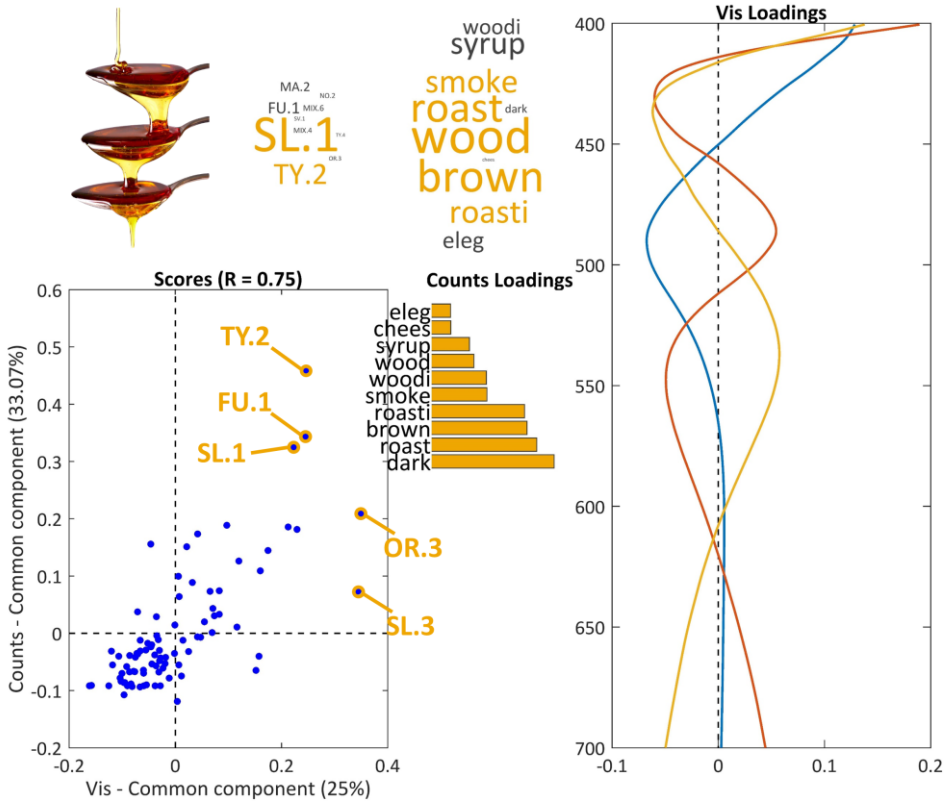


Figure 4.13. *Brown colour*. PCA–GCA scores obtained by comparing the Vis dataset and topic #5 (88×10). The Vis data were compressed by PCA, and 3 PCs were used for computing the PCA–GCA model; for this reason, there are three loading vectors in this figure.

In this case, the common component correlation is mainly driven by the samples at positive scores in Figure 4.13: OR.3 strong ale, FU.1 amber lager, TY.2 amber lager, SL.1 brown ale, MA.2 lager strong, SL.3 Oktoberfest. These samples are mainly strong and darker beers, which is in line with the terms they are associated with.

Topic #5 seems therefore to be related to some extent to the appearance of the beer, mainly to its colour. If all the appearance related-terms of the experimental beer

vocabulary are considered, both the NMR and Vis datasets perform well: NMR provides a correlation of 0.86 (Figure 4.14), while Vis provides a slightly better correlation of 0.88 (Figure 4.15).

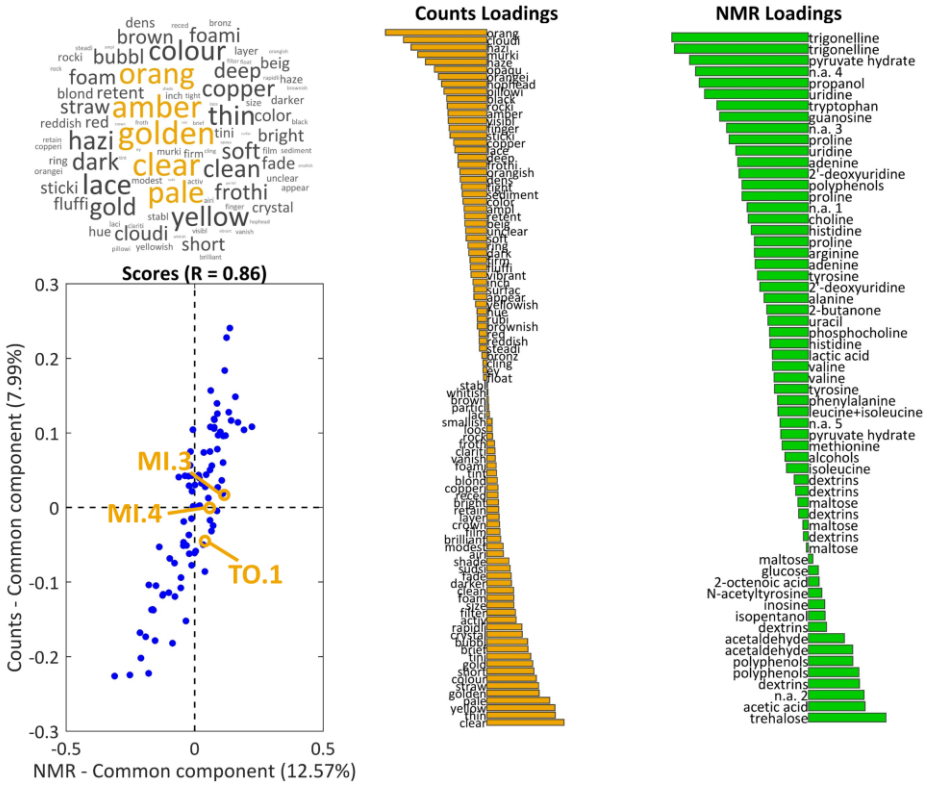


Figure 4.14. *Brown colour*. PCA–GCA scores obtained by comparing the NMR dataset (88×61) and the appearance/colour-related terms, manually selected (88×95, group d in Figure 4.7).

In the case of all the appearance/colour-related terms (group d in Figure 4.7), the most frequent ones are directly linked to beer colour: *orange*, *amber*, *golden*, *clear*, *pale* and *colour* followed by more specific terms such as *hazy*, *foamy*, *dark*, *lace* (the web-like pattern produced by the beer's foam when it dries on the glass' walls), *copper*, *clean*, *yellow* and so on. For the NMR dataset, terms referring to darker (*orange*, *black*, *amber*, *copper*) and hazy (*cloudy*, *hazy*, *murky*, *opaque*) beers results related to metabolites typical of ales and hopped beers (*trigonelline*, *pyruvate* and *propanol*). On the contrary, NMR signals such as *trehalose*, *dextrins* and *polyphenols* result more linked to terms like *clear*, *thin*, *yellow*, *pale*, *golden*, *straw* and *gold*, which are characteristic of lighter and

clearer beers. The direction of the loadings of Figure 4.14 corresponds to the trend in which at negative scores mainly ales are found, and on the other end of the distribution almost only lagers are found.

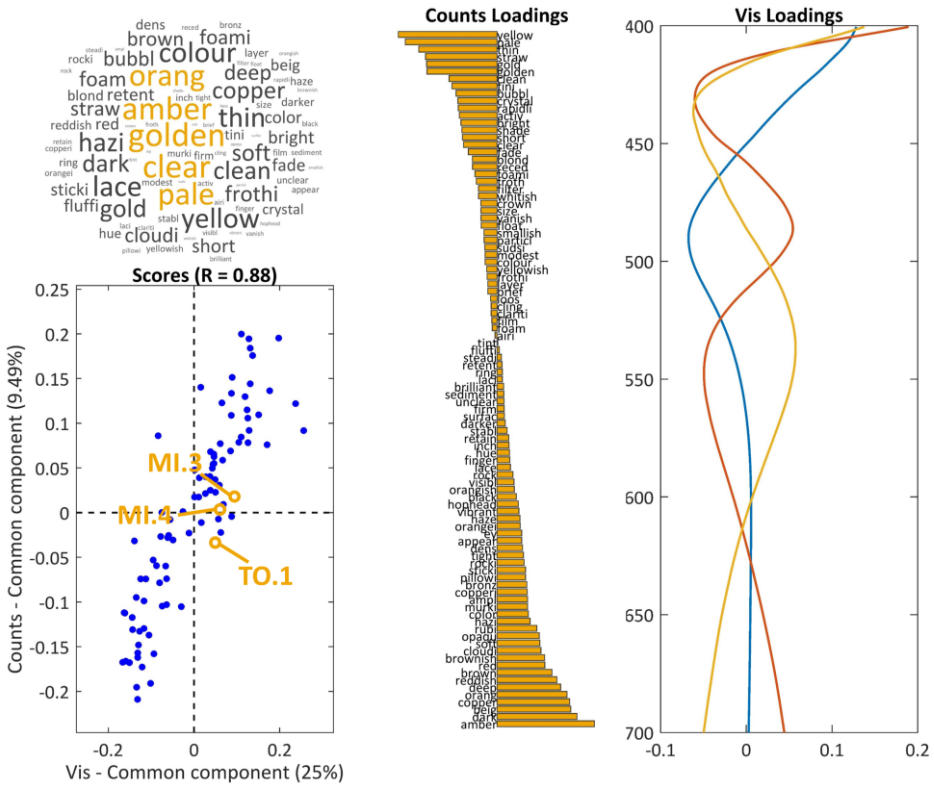


Figure 4.15. *Brown colour*. PCA–GCA scores obtained by comparing the Vis dataset and the appearance/colour-related terms, manually selected (88×95, group d in Figure 4.7).

A situation similar to Figure 4.14 is reported in Figure 4.15: the Count loadings are ordered in almost the same way, with the light/clear beer characteristic terms on one end, and the terms related to darker colours on the other end. The Vis loadings seem to indicate that stronger absorption occurs in the 400–500 nm region, which corresponds to the blue/violet absorption interval, whose observed colour is yellow/orange, the colour of beer [66]. Positive association with the Vis loadings may therefore be linked to stronger absorption of light, which means darker colour; on the contrary, in the case of terms like *yellow*, *pale*, *straw* and *golden*, which are negatively correlated with the Visible loadings, this means that less light is absorbed, and therefore the observed

colour should results less intense, as it is generally observed with the lagers and light beers linked to these terms.

4.4.1.3. *Booze*

Topic #8 is very interesting because of its most important terms: *boozy*, *alcohol* and *syrupy*. According to Urban Dictionary[11], the top definition for the term *booze* in everyday English slang is: "*An alcoholic beverage, specifically any type of beer. It doesn't matter which […]*". This suggests that the information captured by topic #8 may be related both to the sensory-like detection of alcohol and to the beer drinking aimed to drunkenness. It was not found any meaningful interpretation for terms like *hidden* or *hide*.

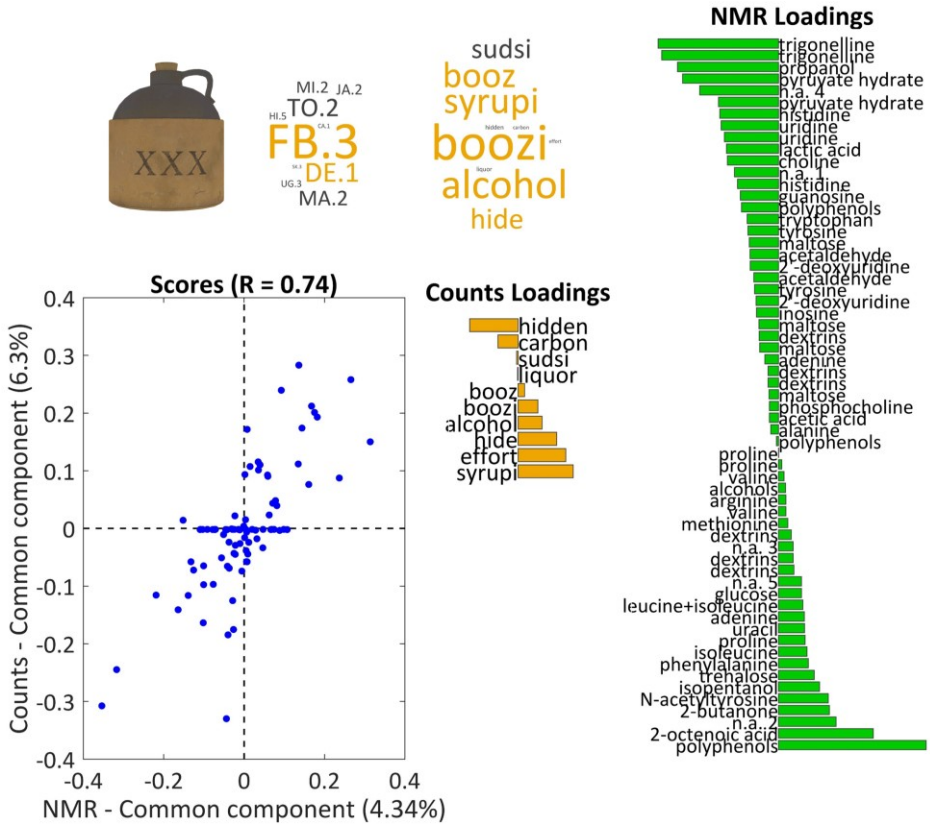


Figure 4.16. *Booze*. PCA–GCA scores obtained by comparing the NMR dataset (88×61) and topic#8 (88×10).

[11] https://www.urbandictionary.com/define.php?term=Booze (accessed: 16/02/2019)

The NMR dataset performs quite well, with a common component correlation of 0.74. However, the strongest beers in the dataset do not end up at positive scores, as the loading sign of the terms *boozy* and *alcohol* may suggest. A quite strong association with *polyphenols* is however found, and since the source of most polyphenols in beer is barley [67], it is possible that the *syrupy* term is related to the sweet/malty taste of beer. However, no terms such as *sweet*, *malty* or *barley* are associated with this topic, therefore this link is just hypothetical.

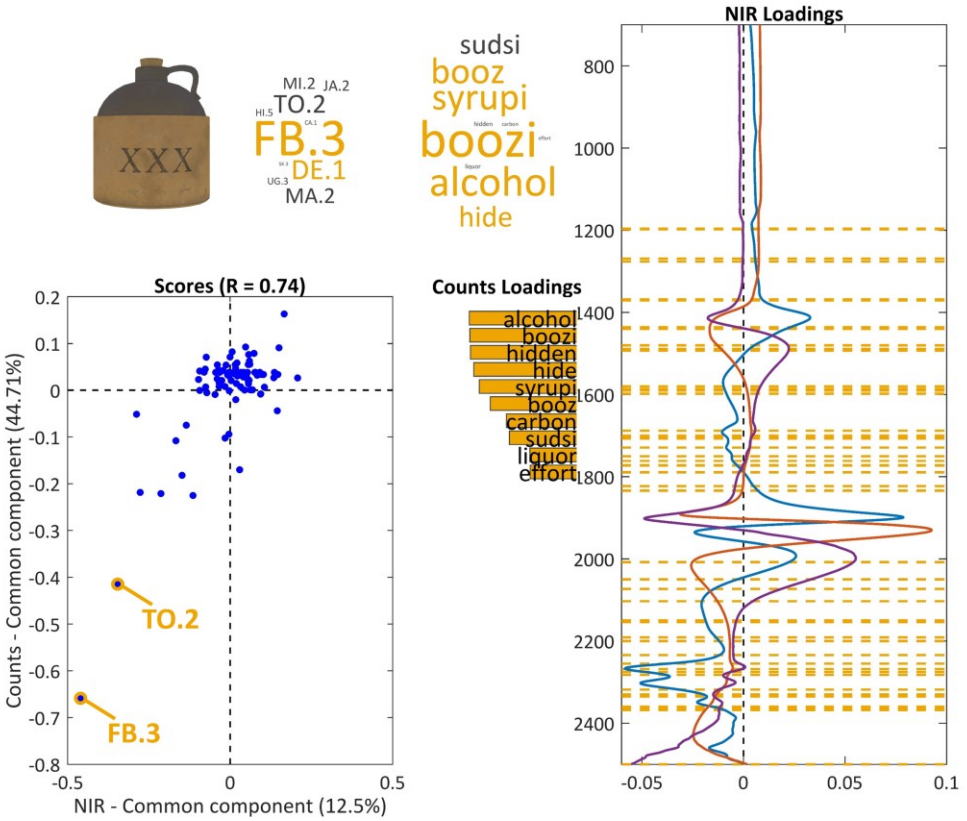


Figure 4.17. *Booze*. scores obtained by comparing the NIR dataset and topic#8 (88×10). The NIR data were compressed by PCA, and 7 PCs were used for computing the PCA–GCA model. For the sake of clarity, only the first three loading vectors are depicted together with the assignments of various carbohydrates.

The same situation is found with the NIR dataset, with a common component correlation of 0.74. However, the samples' distribution suggests that this correlation is highly influenced by few samples, which are located at negative scores values. These

samples are FB.3 and TO.2 and are the strongest beers in the dataset (ABV respectively 10% and 9.3%). If these two samples are excluded, the common component correlation drops to 0.56. The rather grouped set of samples close to the origin of the score plot represent the bulk beers that have ethanol content lover than 7%.

This unbalanced situation is probably the direct result of having very few extreme samples and a substantial group of "average" samples. However, the fact that terms like *alcohol* and *boozy* are the most related to these samples and the NIR signals of ethanol (they both have negative loadings) suggests that this correlation may actually exist.

4.4.1.4. *Refreshment*

Topic #12 is mainly characterized by terms like *lemony*, *chill*, *thirst*, *quencher* and *lime*. Fresh and light beers can be found at positive scores, corresponding to the direction of these terms. Metabolites such as *acetaldehyde*, *dextrins* and *trehalose* also share this direction, as opposed to the strongly hops-related metabolites *trigonelline*, *propanol* and *pyruvate*. *Acetaldehyde* is a key component of lemon [68], and may be associated to freshness. Samples like LE.1 – *Sommerøl* (= "summer beer"), MA.4 – San Miguel Fresca (= "fresh"), TO.4 – Sun Dancer and NO.2 – Lemon Ale are found at positive scores, in line with this "freshness" trend.

In slight opposition to this groups of freshness-related terms is *zesty*, a term generally related to the citrus flavour, but also very used in the beer flavour description in association with the flavour of hops.

At negative scores are most ales and IPAs, in the same direction as trigonelline, but also *propanol*, which is linked to *alcohol*, *ripe fruit* aromas [69]. If topic #12 is about getting refreshment by looking for fresh, lemony flavours, IPAs and ales do not fit for this purpose, being more spiced and stronger in general (higher ABV, but also richer in flavour).

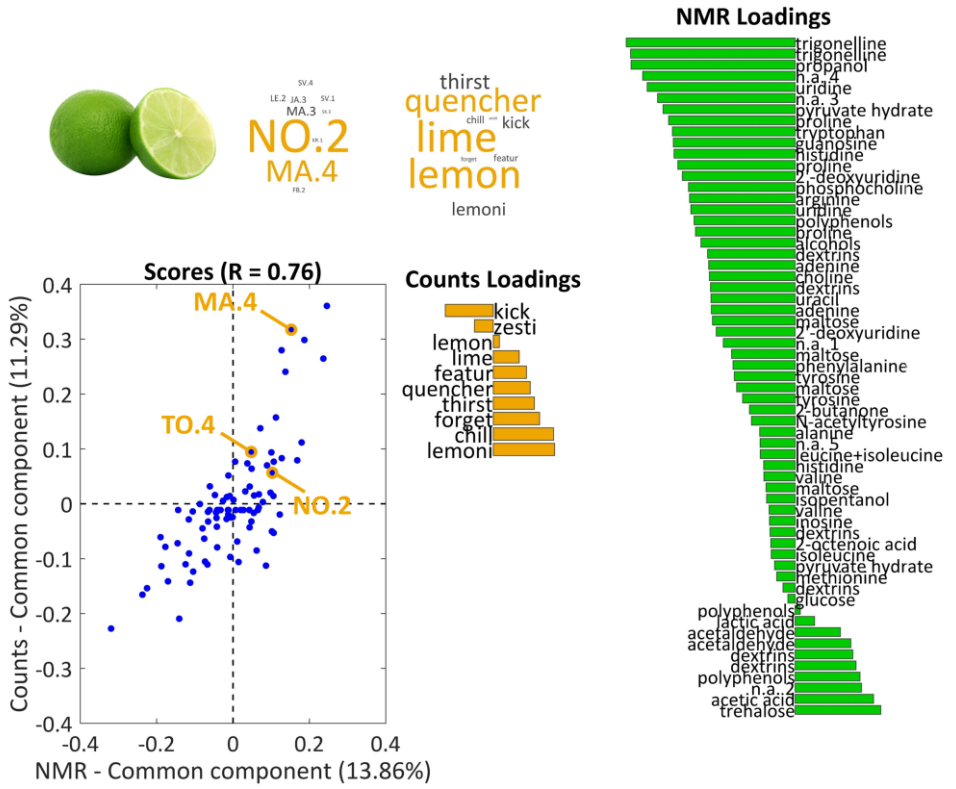


Figure 4.18. *Refreshment*. PCA–GCA scores obtained by comparing the NMR dataset (88×61) and topic#12 (88×10).

4.5. Conclusions and further developments

The present study was aimed at assessing the links between the analytical information and the user-generated description of a set of beer samples. From the point of view of text data, both automatic and manual approaches for selecting subsets of terms, i.e. the *topics*, were employed, with different outcomes.

In the case of topic #1 (*hops*-related topic) it was shown that the automatic topic extraction provided very good results, which were comparable with the results obtained by manually selecting all the hops-related terms in the experimental beer vocabulary. In other cases, like with topic #5 (*brown colour*) the Vis dataset was more correlated with the topic's information with respect to the NMR dataset. However, even

if the correlation improved, interpreting the Vis loadings is more complex, since Vis bands cannot be directly related to specific chemical compounds.

Even though many of the twenty extracted topics made sense, no significant correlation with any of the spectral datasets was found. However, this part of distinct information surely deserved to be more deeply investigated.

Manually selecting the terms generally provided good correlation results: NMR correlated well with the hops-related terms and the Vis data correlated with the appearance/colour-related terms. A summary of the correlation findings is given in Table 4.3.

From this basis, different directions may be taken. For instance, different automatic topic extraction methods may be evaluated with the same approach employed in the present study. Moreover, many topics may be combined based on some sort of correlation index, as well as combinations of the spectral dataset through low- or mid-level data fusion approaches may be worth investigating.

Table 4.3. List of discussed comparisons.

title	**data blocks (spectra + text)**	**correlation value**	**Figure**
Hops	NMR + topic #1	0.91	4.10
	NMR + hop-terms (group c*)	0.91	4.11
Brown colour	NMR + topic #5	0.85 (0.51**)	4.12
	Vis + topic #5	0.75	4.13
	NMR + appearance/colour terms (group d*)	0.86	4.14
	Vis + appearance/colour terms (group d*)	0.88	4.15
Booze	NMR + topic #8	0.74	4.16
	NIR + topic #8	0.74 (0.56**)	4.17
Refreshment	NMR + topic #12	0.76	4.18

*groups of terms in Figure 4.7; **highly influential sample(s) removed

Following on from Section 4.4.3.3, the issue of working with a more homogeneous beer dataset should also be tackled. The aim could be expanding the view, either by collecting a much larger pool of beer samples (with particular attention to balancing the beer styles, production sites, producers, etc.) and replicate the whole study on a larger scale, and/or by gathering a larger text dataset to study how the current dataset is related to the "rest of the world" of beer, from the point of view of the consumer.

Finally, improvements to the text analysis procedure should be investigated as well, focusing on further refinement and deeper study of the text data of the present study. For instance, *n*-grams instead of unigrams may be used for creating the bags-of-words, so that the relations linking more words can also be included in the text data for modelling.

Another direction for improving the current data may also be the "recovering" of the languages removed when English was selected. Each language may be individually analysed to detect language-specific patterns and relevant terms, which could then be translated and merged with their corresponding English term. At the same time, by translating into the different languages those English terms that were identified as important, more relevant pieces of information may be recovered. This could also lead to the recovery of neglected English terms, which may be brought back into the dataset by translating relevant terms of other languages with which they may match.

References | Chapter 4

[1] C. Gómez-Corona, H.B. Escalona-Buendía, M. García, S. Chollet, D. Valentin, Craft vs. industrial: Habits, attitudes and motivations towards beer consumption in Mexico, Appetite. 96 (2016) 358–367. doi:10.1016/J.APPET.2015.10.002.

[2] J. Rice, Craft Rhetoric, Commun. Crit. Stud. 12 (2015) 218–222. doi:10.1080/14791420.2015.1014186.

[3] H. Everett, Craft Food and Drink: The Movement, Guard. Lib. Voice. (2014). https://guardianlv.com/2014/08/craft-food-and-drink-the-movement/.

[4] D.W. Murray, M.A. O'Neill, Craft beer: penetrating a niche market, Br. Food J. 114 (2012) 899–909. doi:10.1108/00070701211241518.

[5] K.G. Elzinga, C.H. Tremblay, V.J. Tremblay, Craft Beer in the United States: History, Numbers, and Geography, J. Wine Econ. 10 (2015) 242–274. doi:10.1017/jwe.2015.22.

[6] I. Duarte, A. Barros, P.S. Belton, R. Righelato, M. Spraul, E. Humpfer, A.M. Gil, High-Resolution Nuclear Magnetic Resonance Spectroscopy and Multivariate Analysis for the Characterization of Beer, J. Agric. Food Chem. 50 (2002) 2475–2481. doi:10.1021/jf011345j.

[7] S. Rossi, V. Sileoni, G. Perretti, O. Marconi, Characterization of the volatile profiles of beer using headspace solid-phase microextraction and gas chromatography-mass spectrometry, J. Sci. Food Agric. 94 (2014) 919–928. doi:10.1002/jsfa.6336.

[8] F.A. Iñón, S. Garrigues, M. de la Guardia, Combination of mid- and near-infrared spectroscopy for the determination of the quality properties of beers, Anal. Chim. Acta. 571 (2006) 167–174. doi:10.1016/j.aca.2006.04.070.

[9] C. Gómez-Corona, M. Lelievre-Desmas, H.B. Escalona Buendía, S. Chollet, D. Valentin, Craft beer representation amongst men in two different cultures, Food Qual. Prefer. 53 (2016) 19–28. doi:10.1016/j.foodqual.2016.05.010.

[10] K. Christensen, K.H. Liland, K. Kvaal, E. Risvik, A. Biancolillo, J. Scholderer, S. Nørskov, T. Næs, Mining online community data: The nature of ideas in online communities, Food Qual. Prefer. 62 (2017) 246–256. doi:10.1016/J.FOODQUAL.2017.06.001.

[11] G.D. Jacobsen, Consumers, experts, and online product evaluations: Evidence from the brewing industry, J. Public Econ. 126 (2015) 114–123. doi:10.1016/J.JPUBECO.2015.04.005.

[12] R.E. Banchs, Text Mining with MATLAB®, Springer New York, New York, NY, 2013. doi:10.1007/978-1-4614-4151-9.

[13] A. Hotho, A. Nürnberger, G. Paaß, F. Ais, A Brief Survey of Text Mining, 2005. http://www.crisp-dm.org/Process/index.htm.

[14] A.K. Smilde, I. Måge, T. Naes, T. Hankemeier, M.A. Lips, H.A.L. Kiers, E. Acar, R. Bro, Common and distinct components in data fusion, J. Chemom. 31 (2017) e2900. doi:10.1002/cem.2900.

[15] T.L. Griffiths, M. Steyvers, Finding scientific topics, Proc. Natl. Acad. Sci. (2004). doi:10.1073/pnas.0307752101.

[16] D.M. Witten, R. Tibshirani, T. Hastie, A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, Biostatistics. 10 (2009) 515–534. doi:10.1093/biostatistics/kxp008.

[17] M. Radovanović, M. Ivanović, Text mining: Approaches and applications, Novi Sad J. Math. 38 (2008) 227–234. https://www.emis.de/journals/NSJOM/Papers/38_3/NSJOM_38_3_227_234.pdf (accessed June 20, 2017).

[18] T.K. Landauer, P.W. Foltz, D. Laham, An introduction to latent semantic analysis, Discourse Process. 25 (1998) 259–284. doi:10.1080/01638539809545028.

[19] B. Pang, L. Lee, Opinion Mining and Sentiment Analysis, Found. Trends® Inf. Retr. 2 (2008) 1–135. doi:10.1561/1500000011.

[20] W. Medhat, A. Hassan, H. Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams Eng. J. 5 (2014) 1093–1113. doi:10.1016/J.ASEJ.2014.04.011.

[21] A.W. Rivadeneira, D.M. Gruen, M.J. Muller, D.R. Millen, Getting our head in the clouds, in: Proc. SIGCHI Conf. Hum. Factors Comput. Syst. - CHI '07, ACM Press, New York, New York, USA, 2007: p. 995. doi:10.1145/1240624.1240775.

[22] C. Price, BrewFinder – An Interactive Flavor Map Informed by Users, in: Springer, Cham, 2018: pp. 342–354. doi:10.1007/978-3-319-91521-0_25.

[23] T. Winograd, Understanding natural language, Cogn. Psychol. 3 (1972) 1–191. doi:10.1016/0010-0285(72)90002-3.

[24] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural Language Processing (Almost) from Scratch, J. Mach. Learn. Res. 12 (2011) 2493–2537. http://www.jmlr.org/papers/v12/collobert11a.html (accessed February 15, 2019).

[25] B. Braun, R. Timpe, Text based rating predictions from beer and wine reviews, 2015.

[26] N. Diakopoulos, D. Elgesem, A. Salway, A. Zhang, K. Hofland, Compare Clouds: Visualizing Text Corpora to Compare Media Frames, in: 2015.

[27] R. Sjögren, K. Stridh, T. Skotare, J. Trygg, Multivariate patent analysis-Using chemometrics to analyze collections of chemical and pharmaceutical patents, J. Chemom. (2018). doi:10.1002/cem.3041.

[28] M.F. Porter, An algorithm for suffix stripping, Program. 40 (2006) 211–218. doi:10.1108/00330330610681286.

[29] Y. Goldberg, O. Levy, word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method, (2014). http://arxiv.org/abs/1402.3722 (accessed January 31, 2019).

[30] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, (2013). http://arxiv.org/abs/1301.3781 (accessed January 31, 2019).

[31] L. Van Der Maaten, G. Hinton, Visualizing Data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2625.

[32] A. Rajaraman, J.D. Ullman, Data Mining, in: Min. Massive Datasets, Cambridge University Press, Cambridge, 2011: pp. 1–17. doi:10.1017/CBO9781139058452.002.

[33] W. Zhang, T. Yoshida, X. Tang, A comparative study of TF*IDF, LSI and multi-words for text classification, Expert Syst. Appl. 38 (2011) 2758–2765. doi:10.1016/J.ESWA.2010.08.066.

[34] M. Van Zaanen, P. Kanters, Automatic mood classification using tf*idf based on lyrics, in: 2010. http://www.crayonroom.com/ (accessed February 15, 2019).

[35] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods. 6 (2014) 2812–2831. doi:10.1039/C3AY41907J.

[36] S.C. Rutan, A. de Juan, R. Tauler, Introduction to Multivariate Curve Resolution, in: Compr. Chemom., Elsevier, 2009: pp. 249–259. doi:10.1016/B978-044452701-1.00046-6.

[37] A. Cutler, L. Breiman, Archetypal Analysis, Technometrics. 36 (1994) 338–347. doi:10.1080/00401706.1994.10485840.

[38] M. Mørup, L.K. Hansen, Archetypal analysis for machine learning and data mining, Neurocomputing. 80 (2012) 54–63. doi:10.1016/J.NEUCOM.2011.06.033.

[39] R. Bro, E.E. Papalexakis, E. Acar, N.D. Sidiropoulos, Coclustering-a useful tool for chemometrics, J. Chemom. (2012). doi:10.1002/cem.1424.

[40] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet Allocation, J. Mach. Learn. Res. 3 (2003) 993–1022. http://www.jmlr.org/papers/v3/blei03a.html (accessed August 22, 2018).

[41] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature. 401 (1999) 788–791. doi:10.1038/44565.

[42] J.-H. Jeong, S.-J. Cho, Y. Kim, High-Resolution NMR Spectroscopy for the Classification of Beer, Bull. Korean Chem. Soc. 38 (2017) 466–470. doi:10.1002/bkcs.11113.

[43] A. Khatib, E.G. Wilson, H.K. Kim, A.W.M. Lefeber, C. Erkelens, Y.H. Choi, R. Verpoorte, Application of two-dimensional J-resolved nuclear magnetic resonance spectroscopy to differentiation of beer, Anal. Chim. Acta. 559 (2006) 264–270. doi:10.1016/J.ACA.2005.11.064.

[44] J.A. Rodrigues, A.S. Barros, B. Carvalho, T. Brandão, A.M. Gil, Probing beer aging chemistry by nuclear magnetic resonance and multivariate analysis, Anal. Chim. Acta. 702 (2011) 178–187. doi:10.1016/j.aca.2011.06.042.

[45] L.I. Nord, P. Vaag, J.Ø. Duus, Quantification of Organic and Amino Acids in Beer by 1H NMR Spectroscopy, Anal. Chem. 76 (2004) 4790–4798. doi:10.1021/ac0496852.

[46] A. Jodelet, N.M. Rigby, I.J. Colquhoun, Separation and NMR structural characterisation of singly branched α-dextrins which differ in the location of the branch point, Carbohydr. Res. 312 (1998) 139–151. doi:10.1016/S0008-6215(98)00241-9.

[47] I.F. Duarte, M. Godejohann, U. Braumann, M. Spraul, A.M. Gil, Application of NMR Spectroscopy and LC-NMR/MS to the Identification of Carbohydrates in Beer, J. Agric. Food Chem. 51 (2003) 4847–4852. doi:10.1021/JF030097J.

[48] A.M. Gil, I.F. Duarte, M. Godejohann, U. Braumann, M. Maraschin, M. Spraul, Characterization of the aromatic composition of some liquid foods by nuclear magnetic resonance spectrometry and liquid chromatography with nuclear magnetic resonance and mass spectrometric detection, Anal. Chim. Acta. 488 (2003) 35–51. doi:10.1016/S0003-2670(03)00579-8.

[49] C. Almeida, I.F. Duarte, A. Barros, J. Rodrigues, M. Spraul, A.M. Gil, Composition of Beer by 1H NMR Spectroscopy: Effects of Brewing Site and Date of Production, J. Agric. Food Chem. 54 (2006) 700–706. doi:10.1021/JF0526947.

[50] A.R. Spevacek, K.H. Benson, C.W. Bamforth, C.M. Slupsky, Beer metabolomics: molecular details of the brewing process and the differential effects of late and dry hopping on yeast purine metabolism, J. Inst. Brew. 122 (2016) 21–28. doi:10.1002/jib.291.

[51] K. Varmuza, P. Filzmoser, Calibration, in: Introd. to Multivar. Stat. Anal. Chemom., CRC Press, 2009. doi:10.1201/9781420059496.ch4.

[52] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130. doi:10.1016/S0169-7439(01)00155-1.

[53] M.A. Drake, G.V. Civille, Flavor Lexicons, Compr. Rev. Food Sci. Food Saf. 2 (2003) 33–40. doi:10.1111/j.1541-4337.2003.tb00013.x.

[54] L.J.R. Lawless, G. V. Civille, Developing Lexicons: A Review, J. Sens. Stud. 28 (2013) 270–281. doi:10.1111/joss.12050.

[55] M.C. Meilgaard, C.E. Dalgliesh, J.F. Clapperton, Beer flavour terminology, J. Inst. Brew. 85 (1979) 38–42. doi:10.1002/j.2050-0416.1979.tb06826.x.

[56] V. Daems, F. Delvaux, Multivariate analysis of descriptive sensory data on 40 commercial beers, Food Qual. Prefer. 8 (1997) 373–380. doi:10.1016/S0950-3293(97)00012-8.

[57] A. Schmelzle, The beer aroma wheel: Updating beer flavour terminology according to sensory standards, BrewingScience. 62 (2009) 26–32.

[58] C. Schönberger, T. Kostelecky, 125th Anniversary Review: The Role of Hops in Brewing, J. Inst. Brew. 117 (2011) 259–267. doi:10.1002/j.2050-0416.2011.tb00471.x.

[59] S.R. Palamand, J.M. Aldenhoff, Bitter tasting compounds of beer. Chemistry and taste properties of some hop resin compounds, J. Agric. Food Chem. 21 (1973) 535–543. doi:10.1021/jf60188a005.

[60] J. Kidrič, I.J. Košir, Characterization of the Chemical Composition of Beverages by NMR Spectroscopy, in: Mod. Magn. Reson., Springer Netherlands, Dordrecht, 2008: pp. 1597–1603. doi:10.1007/1-4020-3910-7_177.

[61] M.T. Ayseli, Y. İpek Ayseli, Flavors of the future: Health benefits of flavor precursors and volatile compounds in plant foods, Trends Food Sci. Technol. 48 (2016) 69–77. doi:10.1016/J.TIFS.2015.11.005.

[62] J. Zhou, L. Chan, S. Zhou, Trigonelline: A Plant Alkaloid with Therapeutic Potential for Diabetes and Central Nervous System Disease, Curr. Med. Chem. 19 (2012) 3523–3531. doi:10.2174/092986712801323171.

[63] N. Mohamadi, F. Sharififar, M. Pournamdari, M. Ansari, A Review on Biosynthesis, Analytical Techniques, and Pharmacological Activities of Trigonelline as a Plant Alkaloid, J. Diet. Suppl. 15 (2018) 207–222. doi:10.1080/19390211.2017.1329244.

[64] S. Ohtake, Y.J. Wang, Trehalose: Current Use and Future Applications, J. Pharm. Sci. 100 (2011) 2020–2053. doi:10.1002/JPS.22458.

[65] A.J. Buglass, M. Mckay, C.G. Lee, Beer, in: Handb. Alcohol. Beverages, John Wiley & Sons, Ltd, Chichester, UK, 2010: pp. 132–210. doi:10.1002/9780470976524.ch9.

[66] E. Pretsch, P. Bühlmann, M. Badertscher, UV/Vis Spectroscopy, in: Struct. Determ. Org. Compd., Springer Berlin Heidelberg, Berlin, Heidelberg, 2009: pp. 1–20. doi:10.1007/978-3-540-93810-1_9.

[67] N. Whittle, H. Eldridge, J. Bartley, G. Organ, Identification of the Polyphenols in Barley and Beer by HPLC/MS and HPLC/Electrochemical Detection, J. Inst. Brew. 105 (1999) 89–99. doi:10.1002/j.2050-0416.1999.tb00011.x.

[68] M.G. Moshonas, P.E. Shaw, Analysis of flavor constituents from lemon and lime essence, J. Agric. Food Chem. 20 (1972) 1029–1030. doi:10.1021/jf60183a019.

[69] C. Sánchez-Estébanez, S. Ferrero, C.M. Alvarez, F. Villafañe, I. Caballero, C.A. Blanco, Nuclear Magnetic Resonance Methodology for the Analysis of Regular and Non-Alcoholic Lager Beers, Food Anal. Methods. 11 (2018) 11–22. doi:10.1007/s12161-017-0953-8.

Chapter 5 | Conclusions

5.1. Final remarks and perspectives

The main aim of the present PhD project was to develop new methods for unravelling similarity and dissimilarity in data with highly complex structure, by developing and applying new chemometric tools for multivariate exploratory analysis. Starting from a dataset of Visible, NIR and NMR spectra of beer samples, its weak clustering structure was the inspiration for trying a different approach: adjacency matrices were brought on the scene, and were used to combine different distance measures and a non-linear approach such as SOM.

The developed approach was named *Fused Adjacency Matrix* and is intended for exploratory analysis and as a mid-level data fusion strategy, if more data blocks are available [1]. Chapter 3 of the present dissertation was devoted to describing the method, providing the theoretical background and some test applications. The exploratory results provided insight in both the examined data and the approach itself. Based on that, a strategy for assessing the influence of the different fusion steps that are involved in the approach was devised and applied to four simulated datasets, used as benchmarks.

The Fused Adjacency Matrix approach surely needs to be further developed and tested on more datasets and the major proposed research lines are provided in the next section. An embryo of a toolbox able to compute the Fused Adjacency Matrix was also developed, and it is available at http://www.models.life.ku.dk/algorithms.

One of the most important lessons that I have learned during my thesis work in the chemometrics field, is that a good chemometrician must have a *feeling* about the data he/she handles. I luckily had the chance of starting early working in the laboratory to carry out experiments and thus generating the data to be analysed, in first person. When this PhD project started, I required that at least some laboratory work was also included in it, and both my supervisors agreed on this point.

By taking care of the whole arch of the beer and whisky experiments, I was able to gain deep insight into the data, keeping at the same time the link with reality alive and well. Chapter 4 of the present thesis is a clear example of this link: the peak integration of the NMR beer data not only taught me how to manage this type of data, but also forced me to study and understand the role of the identified molecules. The beer's linguistics study of Chapter 4 represents a first conclusion of a complete arch of data analysis: from the NMR laboratory to the connection with what consumers taste and recognize in beer. A link between two seemingly distant worlds was drawn, and the language that it speaks is a mixture of data analysis and chemistry. That is, chemometrics.

5.2. Further developments

Future lines of research in the **Fused Adjacency Matrix** framework will include:

- more tests with datasets of different nature and comparisons with other benchmark datasets [2];
- automate how parameters such as the threshold values, the SOM grids' sizes and shapes are chosen;
- thorough assessment of the different fusion steps, concerning their effect on the final output and how information is captured and embedded in the adjacency matrices;
- investigation of different distances measures [3]and/or similarity indexes [4];
- comparison with kernel methods [5,6].

Concerning the investigation of the links between analytical signals and consumer's preference, that is **beer's linguistics and chemistry**, some directions were already anticipated at the end of Chapter 4. Among these are included:

- evaluate different automatic topic extraction methods [7–9];
- combine different topics based on some sort of correlation index;
- combine the spectral datasets through different data fusion approaches [10];
- generate new wordcounts including also *n*-grams [11], in order to try to capture expressions and descriptors with stronger characterization.

Finally, based on the beer's results, an interesting research line would be to try to replicate the same study with the **whisky dataset**. Since whisky represents a kind of niche market often driven by quality and experts in the field, whisky enthusiasts tend to be interested in all aspects regarding this product, making it likely that user-generated text data harvested from the whisky community would work well in connection to our analytical data.

The collection of online reviews about the whisky data was recently started, but it was then decided to put it on hold due to time schedule priorities and to the fact that even if a website about whisky similar to RateBeer was found, the procedure for downloading the review data was much more time consuming than expected.

References | Chapter 5

[1] N. Cavallini, F. Savorani, R. Bro, M. Cocchi, Fused Adjacency Matrices to enhance information extraction: the beer benchmark, Anal. Chim. Acta. (2019). doi:10.1016/J.ACA.2019.02.023.

[2] P. Fränti, S. Sieranoja, K-means properties on six clustering benchmark datasets, Appl. Intell. 48 (2018) 4743–4759. doi:10.1007/s10489-018-1238-7.

[3] R. Todeschini, D. Ballabio, V. Consonni, Distances and Other Dissimilarity Measures in Chemometrics, Encycl. Anal. Chem. Appl. Theory Instrum. (2015) 1–34. doi:10.1002/9780470027318.a9438.

[4] P. Zerzucha, B. Walczak, Concept of (dis)similarity in data analysis, TrAC Trends Anal. Chem. 38 (2012) 116–128. doi:10.1016/J.TRAC.2012.05.005.

[5] B. Schölkopf, The kernel trick for distances, Proc. 13th Int. Conf. Neural Inf. Process. Syst. (2000) 283–289. https://dl.acm.org/citation.cfm?id=3008793 (accessed December 3, 2018).

[6] W. Wu, D.L. Massart, S. de Jong, The kernel PCA algorithms for wide data. Part I: Theory and algorithms, Chemom. Intell. Lab. Syst. 36 (1997) 165–172. doi:10.1016/S0169-7439(97)00010-5.

[7] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet Allocation, J. Mach. Learn. Res. 3 (2003) 993–1022. http://www.jmlr.org/papers/v3/blei03a.html (accessed August 22, 2018).

[8] R. Bro, E.E. Papalexakis, E. Acar, N.D. Sidiropoulos, Coclustering-a useful tool for chemometrics, J. Chemom. (2012). doi:10.1002/cem.1424.

[9] A. Cutler, L. Breiman, Archetypal Analysis, Technometrics. 36 (1994) 338–347. doi:10.1080/00401706.1994.10485840.

[10] E. Borràs, J. Ferré, R. Boqué, M. Mestres, L. Aceña, O. Busto, Data fusion methodologies for food and beverage authentication and quality assessment – A review, Anal. Chim. Acta. 891 (2015) 1–14. doi:10.1016/j.aca.2015.04.042.

[11] R.E. Banchs, Text Mining with MATLAB®, Springer New York, New York, NY, 2013. doi:10.1007/978-1-4614-4151-9.

Acknowledgements

I would like to thank quite a lot of people, but here we only have space for some specific acknowledgements. For all the others, beer will be provided.

Thank you, Marina and Rasmus, for your guidance and for having believed in me all the way through my Master and PhD experience. I am and always will be grateful to you both for this.

Thank you, Eleonora, for having changed *my* life to *our* life.

Thank you, Dillen, for being yourself and one of my (few) favourite weirdos.

Thank you, Viola, for our brunches, our hamburgers in Halifax, our wurstels in Tivoli, our Brazilian choirs and your smile, never missing a chance to explode from your office.

Thank you, CAT group, for at have været min danske familie i København. Jeg savner dig meget.

Thank you, to my Brazilian siblings, Carol, Thomaz, Vitor, Neirivaldo and Fran. And all the Mammeta People!

Thank you, to the Kaley people and all my friends from Modena, you were and still are the best I got from my whole university experience.

Thank you, to my Pociari and all my friends from Trentino. It is amazing how we ended up so scattered all around the world (well, mostly Europe / Northern Italy) but any meeting is just pure chemistry.

Thank you, to my family, always there for me, and always here with me.

Grazie allo zio Steven e alla musica difficile.

Grazie Picia, dovunque tu sia. Mi manchi da morire.

And finally, *thank you*, Nicola, for your non-stop effort to reach a better version of yourself, but please, please! sometimes just remember that you are already quite exceptional.

Torino, March 29th, 2019

DEPARTMENT OF FOOD SCIENCE
FACULTY OF SCIENCE · UNIVERSITY OF COPENHAGEN
PHD THESIS 2019

NICOLA CAVALLINI

New tools for exploratory analysis fusing information from different sources