Data-Driven Methods for Exploratory Analysis in Chemometrics and Scientific Experimentation

by

Guy Emerton

Thesis presented in partial fulfilment of the requirements for the degree of Master of Sciences in the Institute for Wine Biotechnology, Faculty of AgriSciences at Stellenbosch University

Institute of Wine Biotechnology,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Supervisor: Dr. Daniel Jacobson

March 2014

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Signature: . . . .
G. Emerton

Date: 2013/12/20 . . . .

We often forget how science and engineering function. Ideas come from previous exploration more often than from lightning strokes. Important questions can demand the most careful planning for confirmatory analysis. Broad general inquiries are also important. Finding the questions is often more important than finding the answer. Exploratory data analysis is an attitude, a flexibility, and a reliance on display, NOT a bundle of techniques, and should so be taught. Confirmatory data anlaysis, by contrast, is easier to teach and easer to computerize. We need to teach both; to think about science and engineering more broadly; to be prepared to randomize and avoid multiplicity.

– John W. Tukey, 1980

Abstract

Data-Driven Methods for Exploratory Analysis in Chemometrics and Scientific Experimentation

G. Emerton

Institute of Wine Biotechnology,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Thesis: Master of Science in Wine Biotechnology

March 2014

Background

New methods to facilitate exploratory analysis in scientific data are in high demand. There is an abundance of available data used only for confirmatory analysis from which new hypotheses can be drawn. To this end, two new exploratory techniques are developed: one for chemometrics and another for visualisation of fundamental scientific experiments. The former transforms large-scale multiple raw HPLC/UV-vis data into a conserved set of putative features - something not often attempted outside of Mass-Spectrometry. The latter method ('StatNet'), applies network techniques to the results of designed experiments to gain new perspective on variable relations.

Results

The resultant data format from un-targeted chemometric processing was amenable to both chemical and statistical analysis. It proved to have integrity when machine-learning techniques were applied to infer attributes of the experimental set-up. The visualisation techniques were equally successful in generating hypotheses, and were easily extendible to three different types of experimental results.

Conclusion

The overall aim was to create useful tools for hypothesis generation in a variety of data. This has been largely reached through a combination of novel and existing techniques. It is hoped that the methods here presented are further applied and developed.

Uittreksel

Data-Driven Methods for Exploratory Analysis in Chemometrics and Scientific Experimentation

G. Emerton

Institute of Wine Biotechnology,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Thesis: Magister in die Natuurwetenskappe in Wyn Biotegnologie

March 2014

Agtergrond

Nuwe metodes om ondersoekende ontleding in wetenskaplike data te fasiliteer is in groot aanvraag. Daar is 'n oorvloed van beskikbaar data wat slegs gebruik word vir bevestigende ontleding waaruit nuwe hipoteses opgestel kan word. Vir hierdie doel, word twee nuwe ondersoekende tegnieke ontwikkel: een vir chemometrie en 'n ander vir die visualisering van fundamentele wetenskaplike eksperimente. Die eersgenoemde transformeer grootskaalse veelvoudige rou HPLC / UV-vis data in 'n bewaarde stel putatiewe funksies - iets wat nie gereeld buite Massaspektrometrie aangepak word nie. Die laasgenoemde metode ('StatNet') pas netwerktegnieke tot die resultate van ontwerpte eksperimente toe om sodoende ân nuwe perspektief op veranderlike verhoudings te verkry.

Resultate

Die gevolglike data formaat van die ongeteikende chemometriese verwerking was in 'n formaat wat vatbaar is vir beide chemiese en statistiese analise. Daar is bewys dat dit integriteit gehad het wanneer masjienleertegnieke toegepas is om eienskappe van die eksperimentele opstelling af te lei. Die visualiseringtegnieke was ewe suksesvol in die generering van hipoteses, en ook maklik uitbreibaar na drie verskillende tipes eksperimentele resultate.

Samevatting

Die hoofdoel was om nuttige middele vir hipotese generasie in 'n verskeidenheid van data te skep. Dit is grootliks bereik deur 'n kombinasie van oorspronklike en bestaande tegnieke. Hopelik sal die metodes wat hier aangebied is verder toegepas en ontwikkel word.

Acknowledgements

I would like to express my sincere gratitude to the following people and organisations: my supervisor Daniel Jacobson for direction, guidance and tutelage; as well as the Institute for Wine Biotechnology (IWBT) for facilitating this study and providing such excellent opportunities for the application of these developed techniques. Within the institute, thanks must go to the Computational Biology research group as a whole for advice and assistance, especially Piet Jones who gave tireless advice on statistics and parallelisation. Further acknowledgments are due to Phillip Rosochacki for his enthusiastic assistance with machine learning techniques; Adriaan Haasbroek for excellent morale boosting and my family for their unconditional support. Also to Danie Coetsee for his help in translation.

Special thanks to Dr. Astrid Buica, whose excellent data collection from a very large-scale experiment was the basis for much of this thesis; also to her advice on the development of StatNet which was invaluable. My gratitude is also extended to Carien Coetzee for collaboration on the visualisation of her sensory experiment, as well as Kari du Plessis who made her data available for testing.

Contents

Declaration **i**

Acknowledgements **v**

Contents **vi**

List of Figures **ix**

1 Introduction **1**

1.1 Background 1

1.1.1 Chemometrics 1

1.1.2 Network Visualisation - *StatNet* 2

1.2 Problem Statement 2

1.3 Aims 3

1.4 Chapter Overview 3

2 Literature Review **4**

2.1 Chemometric Literature 4

2.1.1 Experimental Data 4

2.1.2 Chemometric Methods 5

2.1.2.1 Wavelets 5

2.1.2.2 Savitzky-Golay Smoothing Filter 7

2.1.2.3 Baseline Correction 8

2.1.2.4 2-Dimensional Alignment 9

2.1.2.4.1 Methods Review 9

2.1.2.4.2 AlignDE 10

2.1.2.4.3 COW 11

2.1.2.5 Feature Map Alignment 13

2.1.3 Machine Learning Techniques 15

2.1.3.1 Decision Trees 16

2.1.3.2 Principal Component Analysis 17

2.1.3.3 Network Analysis 17

2.2 StatNet Literature 18

2.2.1 Background 18

2.2.2 Network Visualisation 19
2.2.3 Temporal Data 20
2.2.4 Principal Component Analysis 21
2.2.5 Data Interaction 21
2.2.6 Statistical Methods 22
2.2.6.1 Metrics 22
2.2.6.2 Statistical Tests 22
2.2.6.3 Correction for Multiple Hypothesis Testing . . . 23
2.3 Conclusion 24
2.4 List of References 26

3 Chemometric Analysis 30
3.1 Introduction 30
3.1.1 HPLC/UV-vis 30
3.1.2 Technical Issues 31
3.1.3 Data 31
3.1.4 Purpose 32
3.2 Methodology 32
3.2.1 Preprocessing 33
3.2.1.1 Data Parsing 33
3.2.1.2 Baseline Correction and Smoothing 34
3.2.1.3 Alignment 34
3.2.1.4 Peak Detection 39
3.2.2 Feature Matrix Generation 44
3.2.2.1 Problem Statement 44
3.2.2.2 Overview 45
3.2.2.3 Aggregating Peak Features 46
3.2.2.4 Experiment-Database Alignment 49
3.3 Results and Discussion 52
3.3.1 PCA 52
3.3.2 Decision Trees 57
3.3.3 Chemical Network 65
3.3.3.1 Network Layout 65
3.3.3.2 Maximum Spanning Tree 66
3.3.3.3 Correlation Network 66
3.3.3.4 Focused Correlation Network 66
3.4 Conclusion 71
3.5 Acknowlegments 72
3.6 List of References 73

4 Network Visualisation 76
4.1 Introduction 76
4.2 Methodology 77
4.2.1 Experimental Data 77

4.2.2 Network creation 78
4.3 Results and Discussion 79
4.3.1 Sensory Data 79
4.3.1.1 Correlation Network 79
4.3.1.2 Central Composite Network 80
4.3.2 *Botrytis* Infection 82
4.3.2.1 All-against-all Network 82
4.3.2.2 ET50 Network 84
4.3.2.3 Time-centric Network 84
4.3.3 Browning Experiment 84
4.3.3.1 Time-centric Network 87
4.3.3.2 Tiered Variable Network 87
4.3.3.3 Star Network 89
4.4 Conclusion 91
4.5 Acknowledgments 91
4.6 List of References 91

5 Conclusion **93**
5.1 Summary 93
5.2 Conclusion 93
5.3 Future Perspectives 94

List of Figures

2.1 Mexican Hat Wavelet . . . 6
2.2 Savitzky-Golay Noise Filter . . . 7
2.3 Warping Optimisation . . . 12
2.4 Time Wheel Visualisation . . . 21

3.1 Baseline correction of Chromatograms . . . 35
3.2 Wavelength-wise Alignment . . . 37
3.3 TAC Alignment . . . 39
3.4 Continuous Wavelet Transform for Peak Detection . . . 40
3.5 SNR ratios . . . 41
3.6 Final Peak Plot . . . 41
3.7 Peak Detection at High Wavelength . . . 42
3.8 Minimum Noise Level Thresholding . . . 43
3.9 Peak Landscape Plots . . . 43
3.10 Peak Comparison Workflow . . . 46
3.11 Peak Detection Drift . . . 47
3.12 Peak Aggregation . . . 49
3.13 Feature Comparison . . . 50
3.14 Variance explained by PCA on entire feature matrix . . . 53
3.15 Score plot of treatment - 2D . . . 54
3.16 Score plot of treatment - 3D, no normalisation . . . 55
3.17 Score plot with labelled pH levels . . . 55
3.18 Score plot with labelled oxidation type . . . 56
3.19 Score plot of time with a reduced data set . . . 57
3.20 Loading plot of PCA on entire data set . . . 58
3.21 Loading plot of PCA on reduced data set . . . 59
3.22 Decision tree for treatment class . . . 60
3.23 Gini importance for variables in the decision tree for treatment . . . 60
3.24 Decision tree for pH class . . . 61
3.25 Gini importance for variables in the decision tree for ph . . . 61
3.26 Decision tree for oxidation class . . . 63
3.27 Gini importance for variables in the decision tree for oxidation . . . 64
3.28 Random permutation tests on decision tree models for each experimental condition . . . 64

3.29 Maximum spanning tree derived from full feature matrix 67
3.30 Correlation network with threshold of 0.5, derived from full feature matrix 68
3.31 Correlation network with threshold of 0.5, derived from a single experiment 69
3.32 Maximum spanning tree of single experiment 70
3.33 Comparison of negatively correlated features in a single experiment 70

4.1 Correlation network for orthogonal sensory data 80
4.2 Overall sensory network for central composite experiment 81
4.3 Focused sensory interaction network 83
4.4 PCA loadings of sensory experiment 83
4.5 All-againt-all comparison of infection rate 85
4.6 ET50 of infection rate 86
4.7 Subgraph of strain infection at one time point 86
4.8 Overview of time network for browning data 88
4.9 Focused view of time network for browning data 88
4.10 Tiered network centered on treatment 89
4.11 Complete 'star' network for browning data 90
4.12 Query into specific conditions in 'star' network 90

Chapter 1

Introduction

Modern developments in computation and technology have allowed for biology to become a data rich science. Advances in the ability to derive information from, for example, genetic and chemical samples have caused a deluge in the available data for analysis; such that it is necessary to continually develop new methods for handling and interpreting this influx of information.

This provides platforms for both the generation of new types of data, as well as novel insights into pre-existing data. In this study, focus is lent to the latter, whereby methods are developed to mine data for information which would otherwise be overlooked. The focus is not in the experiments themselves - how the data is generated and collected - but on methods of interpretation after the fact.

To this end, methods for two different purposes are covered: firstly, the processing and interpretation of chemometric data; secondly, the interpretation of the results of various experiments, using network analyses.

1.1 Background

1.1.1 Chemometrics

Chemometrics is the application of data-driven methods to chemical data in order to deconvolute the high-dimensional outputs of common analytical chemistry tools. It augments the trained chemist's ability to manually search and quantify target chemicals from an analysis by firstly, correcting for technical error from the machine; and secondly, offering an analysis of the resultant data in such a way as to deliver results that would otherwise not be realised by inspection.

The untargeted approach in chemometrics is a "bottom-up" approach whereby putative molecules that influence the data in interesting ways can later be identified. This is in contrast to the traditional analysis method of identifying compounds of interest before the chemical analysis, then tracking their quantitative change exclusively across experimental perturbations.

Methods are here developed to allow for this type of untargeted analysis in large-scale experiments. Within these experiments, variables and conditions are perturbed to different degrees so that compounds are not necessarily conserved across all measurements - causing chemical heterogeneity between samples. Additionally, the samples were taken over a time series, adding a further layer of complexity. This massive and multi-modal data set necessitated a relatively novel development and combination of analysis tools in order to compare chemical phenomena across samples.

1.1.2 Network Visualisation - *StatNet*

Networks are excellent tools for visualisation of complex relationships within data. One such kind of complex data is that which is typically generated from scientific experimentation - the targeted perturbation of input variables in order to gauge the level of some output. The visualisation of these kinds of scientific data is still largely facilitated by classical methods of line, scatter and box plots.

It is contended that networks can be used as an alternative for visualising the results from scientific experiments, not only to draw conclusions from the original hypotheses behind the experiment, but to generate new hypotheses as well. Not only are they amenable to interpretation by the human mind, they also lend themselves to advanced user interaction. In this way networks can represent trends on a large scope, as well as execute advanced queries through filtering, nearest-neighbor searches and subgraph generation. To this end a set of related methods are devised, dubbed *'StatNet'*.

Several data sets generated by scientific experimentation are subjected to this network analyses in order to assess their viability. Although they are all related to the field of wine biotechnology and chemistry, as data types and structures they differ widely. Different variations of similar network workflows are applied to each, with the central theme being the statistical testing for significant results followed by structured representation.

1.2 Problem Statement

The traditional scientific method prescribes a cycle of hypothesis and confirmation. While this is useful for targeted investigation of phenomena, the generation of new hypotheses is often overlooked. This is notable in chemometric analysis, where the vast majority of analyses on data is targeted on specific molecules with conjectured concentrations in the substance in question. This approach is also present to a lesser extent in general scientific investigation with experimental setups targeted towards the answer of a preconceived problem. In this case the traditional visualisations of results can be overwhelming and often confound the search for new hypotheses.

1.3 Aims

The primary and overarching aim is to develop novel methods for exploratory analysis and hypothesis generation. This common aim is pursued along two different avenues: firstly, a large scale generation of putative features from raw machine-generated data; secondly, innovative visualisation of small scale data collected through targeted scientific experimentation.

Aims specifically related to chemometrics are to develop a workflow both simple and efficient enough to process the chromatograms of a large and exhaustive experiment, and ultimately to detect putative compounds and derive experimental conclusions about them. Thereafter, to coerce the data into a format amenable to statistical and machine-learning exploration; specifically, some representation of putative features. As the chemometric data is derived from a targeted experiment, validation of the original hypotheses through such exploration forms a further aim.

Regarding the second channel of data exploration, the aim is to build on research into generalised and extensible methods of network visualisation (tentatively named 'StatNet') that can be broadly applied. The final product should be something that is intuitive to explore; able to present answers to the original hypothesis, and have the latent facility to generate new ones.

1.4 Chapter Overview

The thesis is split into five chapters. Following the present chapter, there is a single literature review chapter covering all of the pertinent literature for both the research chapters. The research chapters are split in two: the first (Chapter 3) will cover the body of chemometric work and includes a discrete introduction, results and conclusion. The second (Chapter 4), contains the majority of the research for network visualisation with a similar structure. The final chapter contains a final conclusion to the overall thesis.

Chapter 2

Literature Review

2.1 Chemometric Literature

2.1.1 Experimental Data

An interesting model case for a large-scale experiment with HPLC/UV-vis data is that of Buica (2012) into the effects and causes of browning in white wine during aging. Several conditions were directly altered in order to observe their combined effect on browning and oxygen levels in model wine. One of the main sources of variance was the addition two phenolic compounds in three separate treatments.

Phenols constitute some of the most important compounds in wine, contributing to the aroma, colour and palette. In their study into the browning of white wine, Kallithraka *et al.* (2009) observed the changes in phenolic compounds over time as well as their correlation to various browning measures. Two of the most significant phenolic compounds were Caffeic Acid and Catechin - cited as two phenolic compounds influencing browning, leading to the formation of by-products due to polymerisation of *ortho*-quinones (Guyot *et al.*, 1996). The fluctuations of these phenols also affects the flavour profile of the wine. Kallithraka *et al.* (2009) found that the concentration of Catechin decreased over time in the experiment; whereas Caffeic Acid was one of the few phenols that increased during aging.

A further effect that was studied was the addition of sulphur dioxide. This has the ability to reduce the same *ortho*-quinones created by the presence of Caffeic Acid and Catechin (Singleton, 1987). Simpson (1982), however, found that the inhibitory effects of SO_2 were fleeting; ineffective in the advanced stages of browning once depleted.

In a large study regarding the overall kinetic effects of aging in white wine, Ferreira (2002) found that the majority of chemical fluctuations occurring during the aging process were a result of the effects of oxidation, and pH-induced reactions - in that order. These two are linked by the fact that phenols can suffer autoxidation, which leads to rapid consumption of oxygen within the media. Autoxidation of phenols is extremely sensitive to pH level, as confirmed by Ferreira (2002) in the same work; a difference of 3- or 4 pH was noted to have the capacity to alter the rate of autoxidation up to 9 times for certain compounds.

2.1.2 Chemometric Methods

In a field such as chemometrics, in which there has been a long and vested data analytic interest, there are a wealth of techniques that allow for the analysis of extremely complex data types. The particular type of data commonly subjected to such analyses is High-Performance Liquid Chromatography (HPLC) with UV/vis spectra. This is a chromatographic technique coupled with absorbance spectrometry, which produces a continuous absorbance feature for each time point. Currently it is common (especially with metabolomics) to couple chromatography with mass spectrometry. This produces a discrete set of mass/charge ratios for each feature; in contrast to the continuous nature of absorbance spectroscopy, with the result that many of the algorithms and software developed are not compatible across these two different types of detectors.

The individual methods used for parts of the overall analysis are reviewed in sections 2.1.2.1 to 2.1.2.4 below. A review of some of the pertinent software and algorithms for the feature map alignment problem for mass spectroscopic analysis is given in section 2.1.2.5 for comparison to the custom feature alignment method presented in the next chapter.

2.1.2.1 Wavelets

Many of the contemporary methods used in chemometric analysis are make use of wavelet transforms in some manner. In particular, the baseline correction and peak detection implementations often used are based on these transforms. This type of analysis is gaining increased popularity due to the arrival of computational capacity allowing for it's somewhat intensive execution.

Throughout much of the history of chemometrics, Fourier analysis was the dominant peak deconvolution approach. Fourier theorems propound the hypothesis that any signal can be reduced to a series of sines and cosines in what is known as Fourier expansion. A problem with Fourier expansion is that it describes a signal in frequency space, but loses the measure of time due to Heisenberg's uncertainty principle. In signal processing terms this is expressed by the fact that it is not possible to know both the frequency and the time at which that frequency occurs in a signal simultaneously (Valens, 1999). Due to this phenomenon, it is necessary in Fourier analysis to slice the time vector into discrete frames for expansion.

Wavelets have the ability to overcome this limitation by applying what is known as multiresolution analysis. This is achieved by shifting a moving

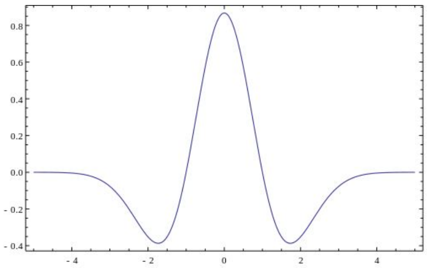


Figure 2.1: The Mexican hat wavelet (Daubechies and Others (1992))

window across the data, and calculating a wavelet-space spectrum for each shift. The window is dynamically scaled by a scaling function, and the same moving window analysis is applied at each new window size. The spectrum can then be represented by amplitude or a weighted coefficient. At the end of the analysis a time-scale representation of the signal is generated, which can be used for a number of different purposes (generally for data compression, but in this case - peak detection).

Two different types of transforms are commonly used: Discrete- and Continuous Wavelet transforms. Discrete transforms eliminate redundant coefficients, and are thus more efficient; in peak detection, however, a high resolution is desired thus continuous transforms are preferred and the redundant coefficients retained (Du *et al.*, 2006). At a high level of resolution, the wavelet coefficient matrix reflects the actual peak shapes along the signal allowing for improved interpretation of peak position.

The central equation describing continuous wavelet transforms is:

$$C(a,b) = \int_R s(t)\psi_{a,b}(t)dt, \psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right), a \in R^+ - 0, b \in R \quad (2.1.1)$$

In the above, $C(a,b)$ is the final 2D matrix of wavelet coefficients; $s(t)$ is the signal; a is the scale; b is the translation and $\psi_{a,b}(t)$ is the wavelet. This wavelet is scaled and translated from the 'mother wavelet' $\psi(t)$. The mother wavelet can be one of a number of mathematical functions, to which the signal is matched with a requisite wavelet coefficient.

The type of wavelet typically used for peak detection is the Mexican hat wavelet, developed by Daubechies and Others (1992) and expressed by the following (equivalent to the second derivative of the Gaussian probability density function):

$$\psi(t) = \frac{2}{\sqrt{3\sigma}\pi^{\frac{1}{4}}}\left(1 - \frac{t^2}{\sigma^2}\right)e^{\frac{-t^2}{2\sigma^2}} \quad (2.1.2)$$

The wavelet is appropriate for matching peak signals as it has the same basic shape, is symmetrical and also positive (Figure 2.1). Generally, the wavelet coefficients will reach a local maximum at the signal peak center. This local maximum increases as the scale is increased from $a = 1$, and itself reaches a maximum when the scale best matches the peak width, before decreasing

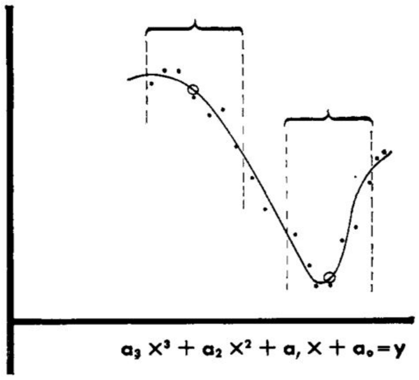


Figure 2.2: Polynomial regression and fitting of the frame center, as depicted in the original paper (Savitzky and Golay (1964)). The frames, denoted by the brackets, are fitted with separate polynomial functions and are used to predict their respective center values (denoted by the circles)

again. These local amplitude maxima resemble ridges if superimposed on the 2-D coefficient matrix, presenting a robust method for detection of peaks. The peak width is represented as the scale corresponding to the maximum value on the ridge, and its area, if desired, can be approximated from the maximum coefficient on the ridge. Refer to section 3.2.1.4 and figure 3.4 for the application of this technique.

2.1.2.2 Savitzky-Golay Smoothing Filter

Raw chromatographic data, much like any other time-series, is subject to noise. To this end a Savitzky-Golay filter can be applied before any further corrective measures. This particular smoothing method is one of the oldest and most commonly applied, and is a simple way to eliminate noise conservatively and unobtrusively. Its base algorithm is essentially unchanged since the method's original publication (Savitzky and Golay, 1964).

Parameters defining the smoothing filter include segment length, polynomial order and an optional derivative function. The algorithm then operates by considering segments of the chosen length from one side of the chromatogram to the other: for each segment, a polynomial of the chosen order is fitted by least squares. The point at the direct center of the segment is then defined by the fitted polynomial at that point. The frame shifts by one data point on either side, and a new polynomial function is regressed for the segment. The new center value adjacent to the previous is then inferred, and the frame moves one point further along the chromatogram; this is repeated until all points have been approximated. The figure from the original paper depicting this process is shown in Figure 2.2.

2.1.2.3 Baseline Correction

A set of useful tools for baseline correction is that developed by Zhang *et al.* (2011) and offered in the authors' and collaborators' open-source software alignDE. The method requires initial peak detection, and thus may have the potential to introduce bias before the rest of the analysis is performed.

The first step for the method of Zhang *et al.* (2011) is thus to apply the CWT peak detection method as described in section 2.1.2.1. For baseline adjustment, the Haar wavelet is used, as opposed to the Mexican Hat wavelet. This is due to the fact that the Mexican Hat wavelet has the tendency to underestimate the scale of a peak (Zhang *et al.*, 2010). The Haar wavelet, in contrast, has the ability to accurately detect the start and end points of a peak due to its discrete nature.

With these peak positions, a putative start and end point is assigned for each peak using a local minima algorithm. A penalised least squares algorithm is then applied, as developed by Zhang *et al.* (2010). The concept behind this algorithm is one of reaching an equilibrium between two measures: firstly, the 'roughness' of the fitting, and secondly the 'fidelity' of the fitting to the original data. These measures can be discretely defined as follows:

The fidelity of the data is measured by the difference of fitting vector z to the original chromatogram c over m points:

$$F = \sum_{i=1}^{m} (c_i - z_i)^2 \tag{2.1.3}$$

Conversely, the relative roughness is measured by the difference between neighbouring points in the fitted data:

$$R = \sum_{i=2}^{m} (z_i - z_{i-1})^2 = \sum_{i=2}^{m} (\triangle z_i)^2 \tag{2.1.4}$$

Penalised least squares attempts to maximise fidelity between the corrected and raw data, while at the same time minimising the roughness of the final fit. The trade-off between these measures can therefore be described by Q, and is regulated by an adjustable parameter λ:

$$Q = F + \lambda R = |\mathbf{c} - \mathbf{z}|^2 + \lambda|\mathbf{Dz}|^2 \tag{2.1.5}$$

$\mathbf{D}$ is the derivative of the identity matrix of size m^2, and represents delta in $\triangle z_i$. Finding for the vector of partial derivatives and solving for $(\delta Q/\delta z) = 0$ gives a linear system of equations:

$$\mathbf{z} = (\mathbf{I} + \lambda \mathbf{D'D})^{-1}\mathbf{c} \tag{2.1.6}$$

These can be simultaneously solved to arrive at a final fit. The adjustable parameter λ strengthens or attenuates the aggressiveness of the correction.

Further development of the method to account for missing values in the data is described in Zhang *et al.* (2011).

The algorithm is run in three steps: 1.) fit an initial rough estimate off the raw chromatogram using λ; 2.) apply the same method on the initial estimate to obtain a refined fit and 3.) adjust the refined fit for possible errors in peak position and width. The final corrected signal is then prepared for further analysis.

2.1.2.4 2-Dimensional Alignment

2.1.2.4.1 Methods Review

Undoubtedly one of the most challenging and contentious steps in the pre-processing of chromatographic data is that of alignment. Alignment is made necessary due to the ubiquitous phenomenon of drift in chromatographic techniques. Disparate positions of the same peak along the time axis is symptomatic of drift and can even be present between technical repeats of the same sample due to differences in basic environmental conditions, such as column temperature between runs (Tomasi *et al.* (2004)). This drift leads to mismatches in peak position, and has a significant confounding effect on multivariate analysis if vectors of the whole chromatogram are used (Nielsen *et al.* (1998)).

There are a myriad of approaches one can take for correcting drift and these are embodied in hundreds of different methods and variations. At present, these can be divided into broad categories of methods that either use the entire chromatographic signal for alignment, or first detect peaks and align according to detected peak position (Arancibia *et al.* (2012)). Another distinguishing factor is whether it is necessary to assign a target chromatogram on which to base the alignment method. This can have a significant effect on accuracy, especially with different measurements from an experiment of factorial design. In their current review of chromatographic calibration, Arancibia *et al.* (2012) state that the two most important methods currently employed are correlation optimised warping (COW) (Nielsen *et al.* (1998)) and rank alignment based on PCA of an augmented data matrix (Prazen *et al.* (1998)). COW is often cited as the most extensively used alignment algorithm and has many algorithmic implementations. It is also relatively simple and fast to execute: an important factor for mass pre-processing. COW evolved from Dynamic Time Warping, which is less constrained and allows warping of the signal over large spans of time. A comparison by Tomasi *et al.* (2004) in the original formulation of COW found that COW is a more precise method for large-scale pre-processing of chromatographic data. Another prevalent method to review is the offering from the AlignDE software, which is used for peak detection in the current work.

A further, recently developed method for alignment falls under the PyMS project (Isaac *et al.* (2012)). The alignment algorithm circumvents the bias of choosing a single target chromatogram by aligning signals between experiments in a clustered similarity tree structure. Experiments that are most similar are aligned first; their combined alignment is then set against the next closest experimental cluster and so on, until the final uppermost branch is reached. The disadvantage of this method is that it relies heavily on peak detection before alignment, which adds a potential layer of bias that COW avoids by using the full-length signal.

2.1.2.4.2 AlignDE

AlignDE is one of the methods that use detected peaks to generate an alignment. As mentioned, this has its disadvantages; however the authors conclude that the resultant alignment is more true to the original peaks of each respective chromatogram (Zhang *et al.*, 2011). It involves the alignment of each chromatogram to a single target, optimising the correlation coefficient between them in a manner similar to COW (see section 2.1.2.4.3).

The way it achieves this optimisation is through differential evolution (DE). This is a variation of a genetic algorithm; a population-based optimiser for which fitness is determined for a number of vectors in degenerate generations. Each vector is populated with peak positions, the variance of which is assigned an upper and lower bound.

The algorithm is initialised with random values for each target position (either a positive or negative slack for the relative shift of a peak position). Each of these vectors are then subjected to mutation (the random alteration of parameters), crossover (the 'mating' of vectors up to a set fraction - a section of one replacing that of another), and selection, whereby the vectors with the best correlation value with the target chromatogram are kept for the next generation.

Once this process is completed, the peaks are aligned according to their respective slacks and the space in-between the peaks are subjected to linear interpolation.

In the author's comparison with COW, they found that while COW had a higher correlation coefficient, it tended to transform peak features more aggressively such that the determination of peak width became difficult.

However, AlignDE was not used for several reasons. The most important of these is that it does not scale as well as COW to large data sets (seeing that its algorithm is semi-dynamic). Due to its use of peak position for alignment, it is not easy to extend into 2-Dimensions, as well as introducing some bias into the data.

Finally, due to the fact that the methods developed here are primarily for the purposes of hypothesis generation, and not for the exact and accurate quantification and identification of features in each chromatogram, the peak height was used as a proxy quantity for feature intensity. Thus the peak distortion seen with COW due to its aggressive optimisation of correlation is a reconcilable shortcoming.

2.1.2.4.3 COW

For the above reasons, as well as methodological restrictions documented in COW is often chosen as the primary alignment technique in the chemometric analysis. A brief description of its operation follows.

COW works primarily on 1-Dimensional data. Attempts have been made to extend the algorithm to 2-D data (an example of this is found in Zhang *et al.* (2011)), however the complexity of warping data in 2 dimensions increases greatly. The method used to apply COW to HPLC/UV-vis 2-D data is discussed in section 3.2.1.3.

The basic operation of COW is to warp a sample chromatogram along the time axis so that the intensity pattern most closely matches that of some other target chromatogram. This warping of the intensity vector is performed by linear interpolation, and the measure for the match parity is linear correlation (Nielsen *et al.* (1998)).

More specifically, given a target T and a sample P of length L, to be warped to P', the sample is split into a set number of segments N each of equivalent length m given by $N = P/m$. For each section with starting value x_s and final value x_e, a warping is applied to each intensity value p of P after warping of x_s to x'_s and x_e to x'_e:

$$p_j = \frac{j}{x'_e - x'_s}(x_e - x_s) + x_s, j = 0, 1..., x'_e - x'_s \tag{2.1.7}$$

The value of $P'(x'_s + j)$ is then calculated by interpolating between the points in P adjacent to p_j. Each warping can be done to within a certain set magnitude. Giving a finite limit to the number of possible warpings for each segment is an important aspect of COW. This number is referred to as the 'slack', t. Given this limit - that each segment has a set number of possible warpings 0...t - the global alignment problem can be reduced to optimisation of the warpings for each segment i in N. If the original segment positions are

$$x_0 = 0 < x_1 < ... < x_{N-1} < x_N = L \tag{2.1.8}$$

and the warpings u are

$$u_i \in [\triangle - t; \triangle + t]; i = 0, ..., N - 1 \tag{2.1.9}$$

so

$$x_{i+1} = x_i + m + u_i; i = 0, ..., N - 1 \tag{2.1.10}$$

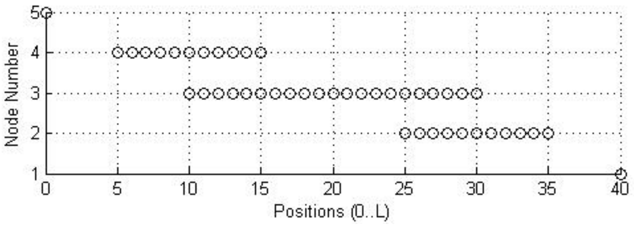


Figure 2.3: The possible positions of nodes x_0 to x_4 with a slack of 5, in the example covered in Nielsen *et al.* (1998).

then the correlation ρ for the segment defines the optimal node position $\mathbf{x*}$ by

$$\mathbf{x*} = \arg\max_x \left(\sum_{i=0}^{N-1} \rho(P'[x_i;x_{i+1}], T[x_i;x_{i+1}]) \right) \tag{2.1.11}$$

$$= \arg\max_x \left(\frac{\mathrm{Cov}(P'[x_i;x_{i+1}], T[x_i;x_{i+1}])}{\sqrt{V(P'[x_i;x_{i+1}], T[x_i;x_{i+1}])}} \right) \tag{2.1.12}$$

Using this correlation as a penalty function, the optimal solution can be arrived at through dynamic programming. This is done by iterating through all segments starting at x_0, keeping the optimal warpings and discarding all other suboptimal warpings. While sequentially considering the position of every x_i (referred to as a node) two matrices are constructed - U, the optimal warping of each node (numerically between $-t..0..t$), and F, the cumulative benefit function. Both these matrices have the same dimensions: the number of nodes i along the rows, and all possible node positions along the column - $(N+1) \times (L+1)$.

A crucial aspect of the optimisation process is that the benefit function is determined for the current node as well as the previous node in the iteration. This optimisation variant is known as backward dynamic programming. An example in the original paper by Nielsen *et al.* (1998) is based on the following simple warping: $L = 40$, $m = 10$ and $t = 5$, giving three warping segments with five nodes in total. The first (x_0) and last (x_4) nodes are constrained at the beginning and ends of signal, and thus x_1 and x_4 can only be warped 5 positions either way of their origin. The further a node is away from these constraints, however, the more possible warping positions are available due to the cumulative nature of the warping; thus the middle nodes x_2 and x_3 have the highest span of possibilities (refer to figure 2.3).

Consider the warping of x_3. If x_2 is placed at position 29 then the position of x_3 is constrained to two values: 34 or 35. Position 34 represents the maximum warping for the 3rd segment, with warping $u_2 = 5$ and $u_3 = 4$; whereas position 35 is the maximum allowable warping for segment 4 ($u_2 = 4$ and $u_3 = 5$). Each of these possibilities has a corresponding cumulate benefit function value $f([x_2, x_3]) + f([x_3, x_4])$. The highest of these is the optimal, and is stored in $F_{2,29}$ with the requisite optimal warping u_2 stored at $U_{2,29}$.

Once all the nodes have been treated in this way, backtracking through U will give the final optimal warping.

It can be deduced from the above demonstration that the only parameters that are necessary to set are the segment length m, and the 'slack' t. Extensive review of the choice of these two parameters, and how they effect the final correlation between sample and target, is covered in Nielsen *et al.* (1998). It was determined that the segment length should be set to around the width of the smallest peak in the signal; smaller segments allow for a higher-resolution warping, though do not significantly add to the final correlation values. The choice of slack is less defined: the larger the slack, the more possibilities exist for warping and therefore the more computationally intensive the alignment. One therefore needs to find a balance between alignment flexibility and time. Another cost to take into account is over-fitting of the data.

It was found by the authors that a slack of just 10% of the segment length was sufficient for a reasonable level of accuracy. Anything above this value did not significantly add to the accuracy of their model alignment, only increasing the computational time gratuitously.

2.1.2.5 Feature Map Alignment

When aligning between multiple measurements along 3 Dimensions, there are two types of approaches (Lange and Tautenhahn, 2008). The first is know as 'Raw Map Alignment', which is the global correction of retention times across multiple measurements; followed by their superposition and subsequent simultaneous analysis. This type of approach is, however, extremely computationally intensive on a large scale. A second type of approach is known as 'Feature Map Alignment', which usually involves a dewarping step followed by feature detection. Finally, alignment of these features is performed before final analyses.

In the latter method, 'features' are manifestations of chemical compounds in chemometric data - typically a peak region in retention with a signature along a second dimension (depending on what is attached to the chromatographic column). A 'feature map' is the collection of features for a single dataset from a particular run. Generally 'consensus features' are obtained through the feature map alignment process. They represent unique features that are common to feature maps within the larger data set. The 'consensus map' is, in turn the map of these global features.

The existing algorithms and software to achieve the above are almost exclusively for cases in which the second dimension is described by the mass-charge ration (m/z) of mass spectroscopy - GC or LC-MS data. There are, however, few to no existing algorithms suitable for the solution of the feature map alignment problem with HPLC/UV-vis data as is used in this study. Nevertheless, it is still informative to review the existing LC/GC-MS methods as the underlying problem remains similar. A brief review of the most prominent of these methods is presented in a comparative study by Lange and Tautenhahn (2008); the methods and software suites compared remain some of the most commonly used.

Typically, as Lange and Tautenhahn (2008) describes, there are 6 stages of achieving a feature map alignment:

1. Signal pre-processing and centroidisation
2. Detection of the 2-Dimensional features or putative compounds
3. Normalization
4. Warping to correct for drift in retention times
5. Computation of a 'consensus map' by multiple comparisons of features across maps
6. Statistical analysis and interpretation

Items 1 through 4 are explained in the subsection above; 6 in the section below. While there are many different methods that can be applied for these steps, they are relatively standard in comparison to the 5th step; for which there are as many algorithms as there are software packages.

Two common distinctions between algorithms are firstly, whether a global correction or 'warping' in retention time is applied (either linear or non-linear); secondly whether clustering or sequential star-wise iteration is used for step 5 above. Notable dangers in feature map alignment are that corresponding features across maps are not grouped into the same consensus feature; secondly that consensus features include multiple features instead of a single unique feature. The most prominent of these software packages, as well as a brief explanation of their approaches, are listed below:

- X-Align (Zhang *et al.*, 2005). The algorithm is reliant on pre-defined 'windows', into which detected features are binned for each feature map The most intense feature for each of these windows is then compared across maps; features found in all maps are deemed significance and an 'average' mapping is created. The map having the features closest to this average mapping is then used as the reference map, to which all other maps are aligned. A final 'consensus' map is the micro-alignment of all the resultant features.

- XCMS (Smith *et al.*, 2006). Part of the R *bioconductor* package (Gentleman *et al.*, 2004). XCMS also employs a window 'binning' technique. Features in the same bin are matched by their mass-spectra signatures. Matching features in the same bin are then resolved using a kernel density estimator, using a probabilistic approach to assign final final feature retention times.

- msInspect (Bellew *et al.*, 2006). Combines the features from multiple experiments into what it calls a single 'peptide array'. It is assumed that warping in the measurements occurs due to a global linear effect, which is first estimated using the most intense features with similar m/z values. After warping according to this linear transformation, it uses a method of 'divisive clustering' to compare and assign the features into the final peptide array. User-defined parameters to achieve this include a window threshold for both retention time and m/z ratio.

- MZmine (Katajamaa *et al.*, 2006). The method used in this software is subtly different from the ones above due to the fact that it does not assume a global trend from which individual experiments must be dewarped. Rather, a 'master list' of features is created. Each map is compared in turn to this master list of features within a retention time window; if the compared feature is deemed similar to the master feature according a set tolerance (both in retention time or m/z value) using some similarity score, the feature is assigned to the master feature. If not, it is appended to the master list as a new feature. The final consensus map becomes this master list once the analysis is complete.

Lange and Tautenhahn (2008) performed quality checks on all of the above methods by obtaining a 'ground truth' of consensus feature maps using MS/MS data that was excluded from the respective software's analysis. The final results from each software suite was then compared to the ground truth in two ways: firstly precision, the probability that an assigned feature is correct; secondly recall, the probability that an assigned feature is found.

This comparison was performed on both protein and metabolic-centric data sets. For the latter, MZmine generally performed best according to both measures.

A more recently developed implementation is an iteration of peak map alignment in PyMS. The algorithm employs an unsupervised clustering technique; building a tree-like structure from feature maps and performing a bottom-up alignment (Isaac *et al.*, 2012). It relies on a 'common ion' to indicate similarity between features in different data sets; once again precluding its use with the UV-vis data at hand. No comparisons in literature of this method with the aforementioned were found; however there does appear to be conceptual promise in this unbiased approach.

2.1.3 Machine Learning Techniques

The analysis of data processed by the above means should be statistically analysed for both the purpose of validation and hypothesis generation for the un-targeted analysis. To this end, several machine learning techniques can be applied; namely, decision trees, principal component analysis and network analysis methods.

2.1.3.1 Decision Trees

Decision trees are one of the most popular machine learning methods for classifying data (Rokach and Maimon, 2005). They are consistently used for their easy interpretability; simplicity and the fact that little to no preprocessing is necessary on the data. It sequentially divides the data into discrete classes using optimal binary partitioning, based on some metric. This process constructs a network in the form of a directed tree. The nodes - or 'leaves' - represent classes, while the edges ('branches') are partitions of the data. Each branch is created from a test on one of the variables in the input data. The test will result in a binary split.

At the top of the decision tree is the first variable criterion by which the data is split - at the bottom are the final class assignments once the model has reached its conclusion. Each level of the tree from the apex downward constitutes a refinement of the model - a lower mis-classification rate - until the data is classified with perfect fidelity. This top-down approach is known as 'recursive partitioning'; and this type of algorithm is greedy - aggressively finding local optima, aiming for a globally optimum solution (Rokach and Maimon, 2005).

The metric by which the variable test is chosen is most commonly the gini index. The gini index measures the divergent probabilities of the binary split in the data based on a given variable test. Concretely, it is the likelihood that a random sample will be mis-classified within its sample subset at that point in the tree, given all the previous binary conditions.

If samples can take on class labels $(1..m)$, and f_i is the number of samples with label i in the data subset at that point in the tree, then the probability that it is misclassified at that point is as follows:

$$P = \sum_{i=1}^{m} f_i(1 - f_i) = 1 - \sum_{i=1}^{m} f_i^2 \tag{2.1.13}$$

This heuristic is greedily estimated for all possible variable splits at each level; in this way the algorithm is NP-complete, which can become restrictive with large scale data.

A useful aspect of decision trees lies in the ability to determine a quick variable importance metric from the model. This is known as gini importance, and is calculated for each variable by simply adding up the reduction in gini impurity at each branch in which the variable is the tested.

This measure has already been used to good effect in chemometrics by Menze and Kelm (2009) for feature selection on several spectral data sets, using random forests - essentially an ensemble method combining multiple decision tree models.

2.1.3.2 Principal Component Analysis

Principal component analysis is a ubiquitous method for data mining in chemometrics (Wold *et al.*, 1987). It is most often used for the purposes of clustering as well as the discovery of new- or validation of known latent variables in data.

At its base, it is a technique that maps a dataset from its existing set of variables onto a new set of axes or 'principal components'. These new axes coincide with the direction of most variance within the data set, in decreasing order; in this way they form an orthogonal set of vectors.

Typically only a few of these principal components are needed to explain most of the data's variance, so that the dimensionality is greatly decreased. It is often found that the first few components are reflective of latent variables. The process of mapping the data onto principal components results in two useful matrices: the loading and score matrices. The scores are essentially the distance of each sample from the principal components; the loadings are vectors of the relative 'direction', or transformation from each of the original variables to the principal components.

The scores are useful in deriving how principal components relate to latent variables, while the loadings inform how original variables influence the principal components; combining these two sources of information, one can draw qualitative and quantitative conclusions as to how original variables influence latent variables.

2.1.3.3 Network Analysis

Networks are a relatively novel tool in the analysis of metabolomic data on a large scale. Recent work by Jacobson *et al.* (2013) used a method of network reconstruction to represent underlying chemical reactions in the aging of port wine. Seeing that the data used in this study is also wine-related and time-series based, much the same techniques were applied for statistical analysis on the final preprocessed data. Network reconstruction is particularly useful for the un-targeted approach used, as it maps out the underlying chemical relationships between detected features in a manner that is visually stimulating to the analyst, aiding in hypothesis generation.

Networks have the ability to model the correlation between chemical features. A simple metric such as Pearson correlation can be used, although the method is open to other statistical metrics should these be more applicable. An all-against-all calculation of correlation can then be performed between features. The nodes of the networks thus consisted of the features; while the edges were weighted by the correlations between them.

Two ways of depicting the resultant network is either to make an arbitrary threshold of correlation, so that only the significant interactions are shown; or constructing a Maximum Spanning Tree (Jacobson *et al.*, 2013). While the former is capable of showing a more complete view of the interactions in a data set, its interpretability can suffer from an abundance of information. In this way, a Maximum Spanning Tree can reduce the data set to only its most salient components, and in so doing represent the skeleton of the network's strongest lines of communication.

A Maximum Spanning Tree is simply the inverse of a Minimum Spanning Tree, a network construct often seen in literature - originally for the solution of the classic 'Travelling Salesman' problem (Kruskal, 1956). A spanning tree is a subgraph of any connected graph in which there are no cycles, and all nodes within the graph are connected. In weighted graphs, the Minimum Spanning Tree is the spanning tree for which the overall sum of edge weights is the minimum possible.

The same algorithm devised by Kruskal is used; the inverse of the edge weights are simply taken. Briefly, the algorithm works as follows (Kruskal, 1956): firstly, a 'forest' is initialised from the graph at hand by adding each individual node as a separate tree. A set of all the edges from the original graph is then created. At each iteration, the edge with minimum weight is removed from this set. If this edge connects two of the trees in the forest, it is included into the growing minimum spanning tree; else it is discarded. The iterations cease once all nodes are connected (there is only a single tree in the former forest).

Jacobson *et al.* (2013) stated that a further advantage of the Maximum Spanning Tree when applied to models of chemical reactions is that it is robust to missing data - intermediate steps within chemical reactions are not reflected in the tree as the strongest correlations over time will be between initial substrates and final products. Additionally, it was proposed that with time-series data the maximum spanning tree has a kinetic element; the flow through the tree representing consecutive reactions in a directed chemical evolution of the media.

2.2 StatNet Literature

2.2.1 Background

Wong and Bergeron (1994) compiled an historical review of the advancement of scientific visualisation, especially regarding that of multi-dimensional, multivariate data (MDMV). The first stage of analysis, dubbed the 'Searching Stage', was characterised by small datasets usually visualised in 2-dimensional plots, occasionally augmented by other graphics denoting categories (in one case the display of cartoon faces with differing expressions on each data point).

The second stage of data analysis ('The Awakening', assigned to the years 1977-1985), was fomented by Tukey's exploratory data analysis (EDA). This was more a foundational paradigm, as enshrined by a brief paper entitled 'We Need Exploratory and Confirmatory' (Tukey, 1980). The principal idea behind this movement was to generate hypotheses instead of only confirming pre-existing ones. Naturally, the visualisation of experimental results was at the center of this ideal. This stage was also aligned with the advancement of computing power, and the advent of the personal computer, allowing for widespread adoption and development of techniques related to EDA. The data sets were generally two- or three dimensional at most; however many of the techniques developed remain the most prevalent today.

Included in Tukey's book on EDA (Tukey, 1977) are typical graphical exploratory techniques; ones that are still applied today with great success. Examples include boxplots, histograms, pareto charts, scatter plots, and stem-and-leaf plots. All of these methods draw their power from classical statistical measures, by which they are defined.

Wong et al. describe the third stage of scientific visualisation (1986-1991) as that of discovery. Studies into interpretation of mdmv data moved away from statistical metrics in two dimensions, and attempted to describe all dimensions of the data in a single plot. This type of analysis relies heavily on the drive of graphical computing, which was gradually facilitating this shift. The final stage was described, at that time, as being one of elaboration - combining the techniques developed up to that time into new methods.

Although there have been many advances in the field of visualisation, in the interpretation of experimental results often the simplest methods are still employed. The plotting of several overlaid line graphs over a time axis; surface response plots and bar graphs are still prevalent in literature.

2.2.2 Network Visualisation

Networks have been used extensively in the field of data visualisation. A review of the various manifestations of networks in this realm is done by Herman *et al.* (2000). The author claims that most information systems in which there are inherent relationships between data elements are susceptible to being rendered into a network.

The application areas listed, however, are generally defined by relationships of extant - not putative - knowledge. Included are systems of predetermined interactions such as computer filing systems; object-oriented programming representations such as UML diagrams and various other hierarchical formats. In addition to this, biology is a field at the fore of large-scale network analysis for phylogenetics; biochemical pathways; metabolomics and genomics.

Probably the most prevalent example of network visualisation for data mining is that of decision trees (Rokach and Maimon, 2005). This machine-learning technique is typically applied to high-dimensional data sets in order build a supervised model of the data and ascertain which variables are most significant towards the prediction of a (single) outcome. In simplistic data sets, the application of this method would not lead to sensible results; it is more appropriate for large-scale data sets.

Some of the concepts within decision trees may, however, be applicable. The topography of a decision tree - with inputs leading from a root node to leaves representing the outputs - is a fundamentally concise and representative way of presenting relationships for dependent data.

2.2.3 Temporal Data

Two of the three experimental data sets analysed in this study were based on time-series (temporal) data. As one of several fundamental types of data (Shneiderman, 1996) it often requires distinct methods for its presentation. Aigner *et al.* (2008) performed a review on contemporary methods for visualising time data specifically. The authors distinguish between analyses that include time as an incidental variable, or simply integrate it into the depiction of others. The latter is more common for scientific analysis; the former for the purposes of planning.

Furthermore, three important distinguishing factors are listed for visualisation of temporal data: firstly, whether the time measurement is linear or cyclical. Both instances of the temporal data analysed were linear; the most common and easiest to visualise. The second distinction is whether the data involves discrete time points or time intervals; only the former is used in the current study. The last is whether time is organised or branches; branching cases were not encountered in the current development.

According to these criteria, the most applicable method covered by the authors is 'TimeWheel' analysis (Tominski *et al.*, 2004). This analysis relies on a 2-Dimensional multi-axis view. Time occupies a central axis, around which axes related to output variables are evenly spaced, reminiscent of the spokes of a wheel. At each discrete time point (the method does not lend itself to time intervals), a line is drawn between its position on the temporal axis and the corresponding level on the variable axis. For each attribute, therefore, the fluctuation of an attribute over time can be characterised the relation to the central axis; for example - if the attribute decreases over time, the formation of the parallel lines will be upper-triangular; if the opposite is true it will be lower triangular (depending on the orientation of the time axis).

There is value to combining several attribute graphs in this way, however the inclusion of so many parallel lines can be overwhelming to the end-user. While the overall trend may be characterised, it is difficult to identify potentially interesting edge cases. All of the data is included in the visualisation, making the method exhaustive, but there is no elimination of insignificant comparisons.

Although it may not be the express purpose of the technique, visualising experimental data where there are perturbed variables and several measured outputs remains difficult. Essentially these are two conceptual layers in the data - an 'input' and 'output' paradigm which is challenging to view simultaneously.

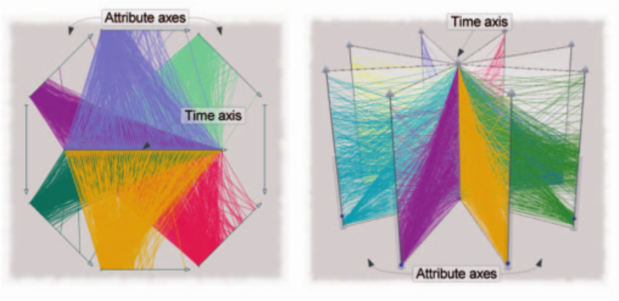


Figure 2.4: Time Wheel Method for visualising temporal data as depicted in Aigner *et al.* (2008)

2.2.4 Principal Component Analysis

Probably the most prevalent technique, however, in visualising and interpreting high-dimensional data is Principal Component Analysis. This has the ability to reduce the pertinent information in a multivariate data set into a few components that can be analysed in 2-dimensional plots. The technique itself is discussed more fully in section 2.1.3.2; however, many of the interactions and subtleties between variables in a data set are often not sufficiently described in loading plots. In data sets with relatively small numbers of variables, this technique also loses it's power of description; often classical visualisations from the above-mentioned second stage of development (line graphs, box plots and surface plots) are reverted to in order to describe outcomes.

2.2.5 Data Interaction

While the nature of how data is presented is the primary concern in interpretation, an important augmentation of any visualisation is the ability to interact with the results. Keim (2002) summarised the various ways in which interactive data representation can assist in visualisation.

The first is 'dynamic projection'. This includes methods of projecting high-dimensional data onto low-dimensional spaces in order to render the information amenable to human interpretation, which is at most capable of three dimensions. The most prevalent of these techniques it the 'Grand Tour' method, which projects subsequent representations of high-dimensional scatter plots onto 2-dimensional planes.

Another aspect of interactivity is the ability to filter the data. Splitting the data into subsets is an extremely useful tool for any end-user; the ability to extract meaningful segments from the overall analysis to arrive at logical syllogisms. These can also be defined as advanced data 'queries' that reflect specific questions related to the data. Consummate with filtering should be an automated re-organisation of the visualisation so that interpretability is retained.

The third factor listed by Keim (2002) is interactive zooming. The visualisation should have an overarching structure from which conclusions can be drawn; however, users should have the freedom to focus on particular areas of interest, so that detailed conclusions can be drawn.

Two further types of data interactivity are interactive distortion (simultaneous presentation of differing levels of detail) and interactive linking (combinations of disparate visualisation techniques); neither of which are applicable to the present method.

2.2.6 Statistical Methods

The statistical methods used in StatNet are relatively straightforward. Most of the power of visualisation lies in the topographical structure of a network than in the descriptive ability of the statistics themselves. Nevertheless, an overview of the statistical measures, tests and corrections is discussed below.

2.2.6.1 Metrics

The metrics used to quantify the relationships in the data were generally either fold change and Pearson correlation. Fold change F is simply the symmetrical ratio of two measures a and b; such that their relative change is centered at 1 and -1:

$$r = \frac{a}{b}; F = \left\{ \begin{array}{ll} r & : r \geq 1 \\ \frac{-1}{r} & : r < 1 \end{array} \right\} \tag{2.2.1}$$

Pearson correlation is defined by the familiar equation; the ratio between covariance and combined standard deviations:

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \tag{2.2.2}$$

Where covariance is defined as the combined expected deviations from respective means:

$$E\left[(X - \mu_X)(Y - \mu_Y)\right] \tag{2.2.3}$$

2.2.6.2 Statistical Tests

Different tests are appropriate to determine significance for variable comparison, depending on whether an underlying probability distribution is assumed; and if so, which distribution. Two examples of a parametric and non-parametric test are illustrated below.

In general for testing of significance differences where a normal distribution is assumed, a Student's t-test can be used. The T-Test is used for normal distributions, and has the advantages of speed and ease of application. The t-statistic is given by the following equation for the comparison of two independent samples of identical length n:

$$t = \frac{\overline{X_1} - \overline{X_2}}{s_{X_1X_2} \cdot \sqrt{\frac{2}{n}}} \tag{2.2.4}$$

where

$$s_{X_1X_2} = \sqrt{\frac{1}{2}\left(s^2_{X_1} + s^2_{X_2}\right)} \tag{2.2.5}$$

$s_{X_1X_2}$ is the pooled standard deviation; $s^2_{X_1}$ and $s^2_{X_2}$ estimators of the variances of the two samples respectively. This is essentially a normalisation for the combined samples.

This t-statistic is assumed to follow a normal distribution. A test is therefore performed at the requisite defined thresholds in the normal distribution to establish whether the null hypothesis - that the samples' means are not significantly different - is true.

A non-parametric alternative to the t-test can also be used, especially when the vectors being compared were of a small length N. This was in the form of the Wilcoxon Rank Sum test (Wilcoxon, 1945). Its appeal is that the only assumptions needed in order to perform the test was that the samples are randomly drawn from the same population and are amenable to sorting - i.e. they vectors have an ordinal scale.

The test compares the requisite pairs of values ($x_{1,i}$ and $x_{2,i}$) for each sample in the ordinal ranking. For each one of these pairs in $i = 1..N$, $|x_{2,i} - x_{2,i}|$ and $\text{sgn}\,(x_{2,i} - x_{1,i})$ are calculated. The N_r pairs are then ranked by the absolute difference measure, after which the test statistic W is calculated:

$$W = |\sum_{i=1}^{N_r} [\text{sgn}\,(x_{2,i} - x_{1,i}) \cdot R_i]\,| \tag{2.2.6}$$

The null hypothesis is then rejected if W is below a set threshold, as is the case with the p-value generated from the t-test.

2.2.6.3 Correction for Multiple Hypothesis Testing

The danger in making many simultaneous hypothesis tests is that it is possible to propagate false positives, or 'Type 1' errors. This is also known as familywise error rate (FWER). A correction for this is in the form of the Holm-Bonferroni Correction (Holm, 1979). It is based on the Bonferroni Correction method, but is cited by Holm et al. as being statistically more powerful.

The multiple hypotheses are corrected by first rank-ordering the p-values $P_1..P_m$ from all hypotheses $H_1..H_m$. All probabilities are then iterated through sequentially in order of lowest to highest; at each iteration the probability P_k is compared to a new probability threshold, adjusted from the original as follows:

$$P_k > \frac{\alpha}{m+1-k} \tag{2.2.7}$$

Where α is the selected significance level. If a p-value fails this test along the bottom-up search, all subsequent hypotheses with higher p-values are voided and the algorithm terminates.

2.3 Conclusion

A body of literature has been reviewed to establish a platform for untargeted chemometric analysis and network visualisation.

In terms of model data for untargeted chemometric analysis, HPLC/UV-vis is a good candidate for novel methods. The data generated by the extensive experiments into browning phenomena by (Buica, 2012) is conserved, has experimental duplicates and presents an interesting challenge due to its scale and diversity of experimental conditions. There are also clear targets for building models of the data, in the form of distinct classes of experimental variables.

After reviewing a number of pre-processing techniques, it is proposed that the following should be incorporated into a large-scale analysis: smoothing using the standard Savitzgy-Golay filter (Savitzky and Golay, 1964); baseline correction using wavelet methods as implemented in the alignDE package (Zhang *et al.*, 2011) and peak alignment using the simple yet effective COW (Tomasi *et al.*, 2004).

HPLC/UV-vis data has a rich feature map, with continuous features along wavelengths. Untargeted analysis using the full map is something not often attempted, and may yield useful results. The incorporation of information across both dimensions would be preferable as features can differ across wavelengths. As feature map alignment algorithms are generally only implemented for discrete MS data, a new implementation would need to be developed that is memory efficient and can be applied sequentially for a group of UV-vis chromatograms. A brief review of some of the more prominent MS implementations was expounded; the closest method to the stated requirements is most likely MZmine as developed by (Katajamaa *et al.*, 2006).

Several candidate machine learning and statistical methods were reviewed, the applications of which to the results of an untargeted analysis may give insight and validation.

Opportunities for the application of network methods were also explored. While there has been much progress in the field of exploratory data analysis and visualisation, there is still room for new methods to explore the multivariate space of scientific experimentation - especially with temporal data, which has an added layer of complexity.

Pure network representations are not often applied to this field, but rather used as a motif for conceptual design. It may be possible to employ networks as a multivariate comparison tool for hypothesis generation using experimental results. Several statistical metrics for feature comparison were also reviewed as possible bases for network visualisations.

2.4 List of References

Aigner, W., Schumann, H., Tominski, C., Miksch, S. and Mu, W. (2008). Visual Methods for Analyzing Time-Oriented Data. *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 1, pp. 47–60.

Arancibia, J.a., Damiani, P.C., Escandar, G.M., Ibañez, G.a. and Olivieri, A.C. (2012 December). A review on second- and third-order multivariate calibration applied to chromatographic data. *Journal of chromatography. B, Analytical technologies in the biomedical and life sciences*, vol. 910, pp. 22–30. ISSN 1873-376X.
Available at: http://www.ncbi.nlm.nih.gov/pubmed/22365532

Bellew, M., Coram, M., Fitzgibbon, M., Igra, M., Randolph, T., Wang, P., May, D., Eng, J., Fang, R., Lin, C., Chen, J., Goodlett, D., Whiteaker, J., Paulovich, A. and McIntosh, M. (2006). A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics*, vol. 22, no. 15, pp. 1902–1909.
Available at: http://bioinformatics.oxfordjournals.org

Buica, A. (2012). Oxidation and Browning in Model Wine Solutions: Effect of pH, SO2 and Phenolic Content. In: *South African Society for Enology and Viticulture Conference*.

Daubechies, I. and Others (1992). *Ten lectures on wavelets*, vol. 61. SIAM.

Du, P., Kibbe, W.a. and Lin, S.M. (2006 September). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics (Oxford, England)*, vol. 22, no. 17, pp. 2059–65. ISSN 1367-4811.
Available at: http://www.ncbi.nlm.nih.gov/pubmed/16820428

Ferreira, A.S. (2002). Kinetics of oxidative degradation of white wines and how they are affected by selected technological parameters. *Journal of Agricultural ...*, pp. 5919–5924.
Available at: http://pubs.acs.org/doi/abs/10.1021/jf0115847

Gentleman, R., Carey, V., Bates, D., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, vol. 5, no. 10, p. R80. ISSN 1465-6906.
Available at: http://genomebiology.com/2004/5/10/R80

Guyot, S., Vercauterent, J., Cheyner, V.I.R., Viala, P., Pharmacognosie, D., Ii, U.B. and Saignat, L. (1996). Structural determination of colourless and yellow dimers resulting from (+)-catechin coupling catalysed by grape polyphenoloxidase. *Elsevier Phytochemistry*, vol. 42, no. 5, pp. 1279–1288.

Herman, I., Melancon, G. and Marshall, M.S. (2000). Graph visualization and navigation in information visualization: A survey. *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS*, vol. 6, no. 1, pp. 24–43.
Available at: `http://ieeexplore.ieee.org`

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70.

Isaac, A., O'Callaghan, S. and Likic, V. (2012). PyMS.
Available at: `http://code.google.com/p/pyms`

Jacobson, D., Monforte, A.R. and Ferreira, A.C.S. (2013 March). Untangling the Chemistry of Port Wine Aging with the Use of GC-FID, Multivariate Statistics, and Network Reconstruction. *Journal of agricultural and food chemistry*. ISSN 1520-5118.

Kallithraka, S., Salacha, M. and Tzourou, I. (2009 March). Changes in phenolic composition and antioxidant activity of white wine during bottle storage: Accelerated browning test versus bottle storage. *Food Chemistry*, vol. 113, no. 2, pp. 500–505. ISSN 03088146.
Available at: `http://linkinghub.elsevier.com`

Katajamaa, M., Miettinen, J. and Orešič, M. (2006). MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics*, vol. 22, no. 5, pp. 634–636.
Available at: `http://bioinformatics.oxfordjournals.org`

Keim, D. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, no. 1, pp. 1–8. ISSN 10772626.
Available at: `http://ieeexplore.ieee.org`

Kruskal, J. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society*, vol. 7, no. 1, pp. 48–50.
Available at: `http://www.jstor.org/stable/10.2307/2033241`

Lange, E. and Tautenhahn, R. (2008 January). Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics*, vol. 9, p. 375. ISSN 1471-2105.
Available at: `http://www.pubmedcentral.nih.gov`

Menze, B. and Kelm, B. (2009 January). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, vol. 10, p. 213. ISSN 1471-2105.
Available at: `http://www.pubmedcentral.nih.gov`

Nielsen, N.-P.V., Carstensen, J.M. and Smedsgaard, J.r. (1998 May). Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A*, vol. 805, no. 1-2, pp. 17–35. ISSN 00219673.
Available at: `http://linkinghub.elsevier.com/`

Prazen, B.J., Synovec, R.E. and Kowalski, B.R. (1998 January). Standardization of Second-Order Chromatographic/Spectroscopic Data for Optimum Chemical Analysis. *Analytical Chemistry*, vol. 70, no. 2, pp. 218–225. ISSN 0003-2700.
Available at: `http://dx.doi.org/10.1021/ac9706335`

Rokach, L. and Maimon, O. (2005). Top-Down Induction of Decision Trees Classifiers - A Survey. vol. 35, no. 4, pp. 476–487.

Savitzky, A. and Golay, M. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In: *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pp. 336–343. IEEE.

Simpson, R. (1982). Factors affecting oxidative browning of white wine. *Vitis*.
Available at: `http://www.vitis-vea.de/admin/volltext/e020660.pdf`

Singleton, V.L. (1987). Oxygen with phenols and related reactions in musts, wines, and model systems: observations and practical implications. *American Journal of Enology and Viticulture*, vol. 38, no. 1, pp. 69–77.

Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R. and Siuzdak, G. (2006). XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Analytical Chemistry*, vol. 78, no. 3, pp. 779–787.
Available at: `http://pubs.acs.org/doi/abs/10.1021/ac051437y`

Tomasi, G., van den Berg, F. and Andersson, C. (2004 May). Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics*, vol. 18, no. 5, pp. 231–241. ISSN 0886-9383.
Available at: `http://doi.wiley.com/10.1002/cem.859`

Tominski, C., Abello, J. and Schumann, H. (2004). Axes-based visualizations with radial layouts. In: *Proceedings of the 2004 ACM symposium on Applied Computing*, pp. 1242–1247. ACM.

Tukey, J. (1980). We need both exploratory and confirmatory. *The American Statistician*.
Available at: `http://www.tandfonline.com`

Tukey, J.W. (1977). Exploratory data analysis. *Reading, Ma*, vol. 231.

Valens, C. (1999). Really Friendly Guide to Wavelets.
Available at: `http://www.robots.ox.ac.uk/ parg/mlrg/papers/arfgtw.pdf`

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83.
Available at: `http://www.jstor.org/stable/10.2307/3001968`

Wold, S., Esbensen, K. and Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 3, pp. 37–52. ISSN 0169-7439.
Available at: `http://www.sciencedirect.com`

Wong, P. and Bergeron, R. (1994). 30 Years of Multidimensional Multivariate Visualization. *Scientific Visualization*.
Available at: `http://wwwx.cs.unc.edu`

Zhang, X., Asara, J.M., Adamec, J., Ouzzani, M. and Elmagarmid, A.K. (2005). Data pre-processing in liquid chromatographyâmass spectrometry-based proteomics. *Bioinformatics*, vol. 21, no. 21, pp. 4054–4059.
Available at: `http://bioinformatics.oxfordjournals.org`

Zhang, Z.-M., Chen, S. and Liang, Y.-Z. (2010 May). Baseline correction using adaptive iteratively reweighted penalized least squares. *The Analyst*, vol. 135, no. 5, pp. 1138–46. ISSN 1364-5528.
Available at: `http://www.ncbi.nlm.nih.gov/pubmed/20419267`

Zhang, Z.-M., Chen, S. and Liang, Y.-Z. (2011 January). Peak alignment using wavelet pattern matching and differential evolution. *Talanta*, vol. 83, no. 4, pp. 1108–17. ISSN 1873-3573.
Available at: `http://www.ncbi.nlm.nih.gov/pubmed/21215845`

Chapter 3

Chemometric Analysis

A method for untargeted analysis of multiple HPLC/UV-vis chromatograms is developed. Approximately one thousand chromatograms taken from a single experiment are preprocessed en masse for smoothing, baseline correction and alignment. Using a newly developed algorithm, feature map alignment is performed to create a conserved list of chemical features.

A reduced data set using these features is then validated and explored using Machine-Learning techniques. It was found through permutation tests that the generated data had integrity with regards to the experimental set-up. Finally, hypotheses were drawn regarding the role of different features in the experiment.

3.1 Introduction

3.1.1 HPLC/UV-vis

The data used for the development of this method were approximately 1000 HPLC/UV-vis chromatograms from a single large-scale experiment. To summarise this separation technology: HPLC exploits the difference of interaction strenghts between molecules to effect a separation of compounds within a sample. This separation is achieved by pumping a liquid solvent containing the sample through a column loaded with solid compounds. The column solids (stationary phase) interact with the compounds within the solvent (mobile phase) to different magnitudes of attraction, inducing a separation. Compounds will therefore elute from the column at different times (retention times) throughout the run interval. HPLC differs from regular liquid chromatography due to the higher pressure induced by a pump within the column; whereas ordinary liquid chromatography relies mostly on gravitational forces.

The end of the column is fitted with a UV-vis spectrophotometer. This detector measures the absorbance of the the liquid exiting the HPLC column at specified time intervals. Absorbance is measured at wavelengths within the

Ultra-Violet range (in this case specifically - 190 to 560 nanometres), created by diffracting a single light source though a prism. The detector operates by measuring the intensity of light after it passes through the sample, and comparing this to either the light before sample absorbance, or an appropriate separate reference material; the ratio of these two measures is known as the transmittance.

The data from a single HPLC-UV-vis run is therefore separated across two dimensions (time and wavelength), and measured along a third (spectral absorbance at a wavelength, based on transmittance). This high-dimensional data presents a challenge for traditional data analysis, for a number of reasons.

3.1.2 Technical Issues

Perhaps some of the most significant challenges to overcome are the issues arising from technical errors from the machinery itself. Several phenomena related to the process of chromatography can translate into artefacts in the data that need to be corrected before any meaningful information can be extracted. These types of artefacts include baseline drift, which is common in any type of chromatography; as well as shifts in retention time between technical or sample repeats.

3.1.3 Data

The test data used for the chemometric analysis was from a study into the effects of various conditions related to browning in white wine. There were a number of experimental conditions thought to relate to the phenomenon of browning that were independently perturbed in a factorial-like manner. HPLC/UV-vis readings were then taken for each experimental duplicate over the duration of the experiment.

The media used was a standard synthetic media for white wine, with 12% alcohol and 5g/L tartaric acid.

The experiments were divided into two main groups: slow- and rapid oxidation. Slow oxidation was preformed in a sealed bottle with restricted O_2 and a constant headspace - small in comparison to the volume of liquid (0.5 mL/135 mL). For rapid oxidation, the bottle was left open and there was no restriction in available O_2. Beyond oxidation conditions, three other experimental conditions were tested: phenolic treatment, pH and SO_2 levels.

The phenolic treatment involved the addition to the media of two of the main phenolic compounds typically present in white wine: Caffeic Acid (CA) and Catechin (C). Three different treatment types were tested: the addition of Catechin on its own (150 mg/L); Caffeic Acid on its own (200 mg/L), and a combination of both phenolic compounds together (150 mg/L and 200 mg/L for C and CA, respectively).

pH is typically at 3.6 in white wine. It was tested at 1.9, 3.1, 3.6, 4.1 and 7.2 in order to observe its effects on oxygen and browning. The addition of the phenolic compounds was timed to well after the pH had stabilised at the desired value as their effect on pH level was strong and fast-acting in terms of observed browning.

Similarily, SO_2, which is ordinarily added at 25 ppm, was also tested at 0- and 50 ppm. This was postulated to have a direct effect on the oxidation process of the wine - typically SO_2 is added to suppress the effects of oxidation.

Each of the conditions were tested at their respective levels in a factorial design experiment. A batch was created for each combination of conditions, from which three separate repeats were decanted. The oxidation reactions were started in both cases with the addition of catalysts: Fe at 5 mg/L and Cu at 0.3 mg/L. Two of these repeats were measured using HPLC/UV-vis on days 0, 1, 3, 7, 15, 45 and 60 for each separate condition.

3.1.4 Purpose

The purpose of the method development presented in this section is to create a platform for an untargeted analysis of HPLC/UV-vis data. This approach is entirely different from that of a targeted analysis, where the compounds of interest are known before the time and systematically identified after the chemical analysis. Rather, the analysis is first performed and interesting features - hopefully related to compounds - are identified for further investigation.

This kind of analysis can lend itself to novel research because it removes the factor of bias towards a specific outcome for a project. The fact is that many substances have a high degree of chemical complexity, the entirety of which is not currently known. Wine is a good example of this potential complexity; having a huge variety of different compounds. The approach taken within this project is to first find significant putative compounds within single experiments, with the intention of finding common compounds within all experiments, as expressed by the feature map alignment problem mentioned by Lange and Tautenhahn (2008). By mapping the data to putative features and their relative intensities, one essentially reduces the 2-Dimensional HPLC/UV-vis data to a 2-Dimensional matrix - the exact methodology for which is described in 3.2.2. Once this is achieved, it should be possible to apply statistical methods to the resultant data in order to identify the function and significance of the putative compounds with regards to experimental variations.

3.2 Methodology

The methodology was used to analyse all experiments on a large scale. This can be roughly split into two parts: firstly, the preprocessing of the data to create feature maps. This part is mostly the combination of several existing methods, applied en-masse and in parallel to each experiment. The second part is the solution of the feature map alignment problem, most of which was devised and coded independently of any existing solution.

3.2.1 Preprocessing

Preprocessing of the HPLC/UV-vis runs comprised much of the computational effort involved in the chemometric analysis. As explained in the previous section, a large portion of the analysis of chemometric data involves the correction of technical artefacts. The premise behind this kind of preprocessing is, firstly, that it allows for improved qualitative analysis further down the pipeline; for example improved detection of peaks. Secondly, it allows for a dataset that can subsequently be more accurately and meaningfully compared to other datasets in feature map alignment.

With this in mind, several methods were applied in serial to correct and standardise the data. These included data parsing; baseline correction; smoothing and alignment. While these are fairly standard tasks in any analysis of chemometric data, and are easily performed on single datasets, the mass processing of approximately one thousand of these high-dimensional datasets required more involved solutions with regards to computational and algorithmic complexity. All of the steps after the initial data parsing were performed in parallel on a computing cluster through the application of open-source libraries. While there is an additional layer of complexity in this regard, it is more than compensated for by the great increase in processing throughput.

3.2.1.1 Data Parsing

The initial data was in the raw, proprietary format of the chromatographic system, in this case Agilent's Chemstation. It was desired that the data be in a ASCII text format, as this is more universal and can be used for whichever multitude of languages may be required along the preprocessing pipeline.

Due to the fact that there were 918 datasets in total, and that manually navigating the Chemstation menu system for each one was laborious, some Chemstation macros were written to automate the process. These macros were based on the file structure of the chemstation .D files. The macro would search through a specified folder and load each of the chromatogram files in series. Once loaded, the file registry would be searched for the methods label - present in each run, describing which levels of pH and SO_2; the applied treatment, as well as the time point at which the sample was taken.

The full spectra was then written to a textfile using Chemstation macro commands. For each folder, an index linking the methods label to the run name was generated. This ensured that unique experimental names were maintained.

A perl script was then made to collect, rename and move the text files to a central folder. A wrapper was also created, using a bash script to utilise the GNU iconv library to re-encode the text files to UTF-8. Once all this was achieved, there was a single folder containing methods-labeled HPLC/UV-vis data for each experimental run, including technical repeats.

3.2.1.2 Baseline Correction and Smoothing

The data was first smoothed using the Savitzky-Golay noise filter (Savitzky and Golay, 1964). The implementation used was that developed in the MassSpecWavelet library developed by Du *et al.* (2006), which forms part of the bioconductor package in R. A quadratic fitting was used on the data, with a frame/segment size of 15 intervals. This produced an adequate and not overly-aggressive initial smoothing of the data.

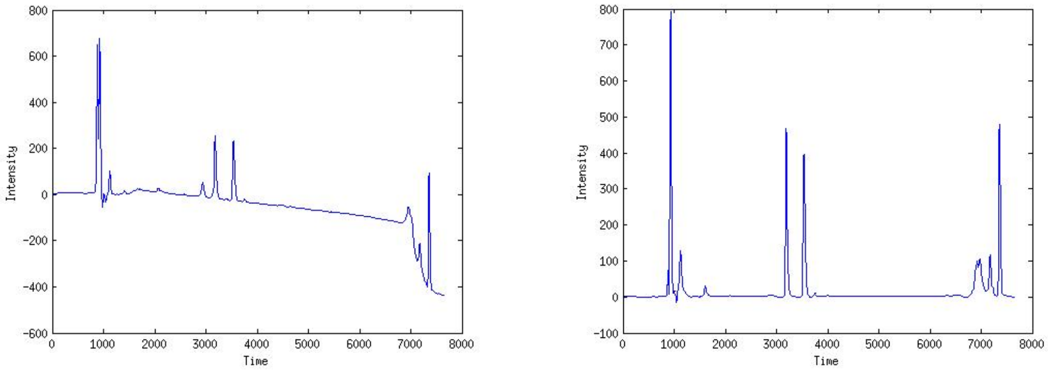
Baseline correction is an essential step in preprocessing. Background sources of light and interference can lead to an artificially raised baseline in UV-vis spectra. An additional source of baseline drift is the possible change in solvent composition over time. This can happen either deliberately or as a consequence of differential, possibly unintentional interaction of the mobile and stationary phases. Whatever the cause, correction of baseline drift is essential, especially with regards to linear correlation - the cost function of COW.

The baseline correction method was that developed by Zhang *et al.* (2011). The implementation in the R package alignDE was used. It is based on the initial detection of peaks and peak widths; preserving these features while warping the chromatographic space between them. Visually, the default parameters of alignDE worked well with the data (based on a small representative spread of experiments), allowing for a reasonable balance between roughness and fidelity.

The steps of smoothing and baseline correction were both implemented in R. Due to the huge scale of processing all text data for each experiment, it was found to be easiest to parse and organise the data in perl; an interface was therefore used whereby R was called within a perl script using the CPAN module Statistics::R (Graciliao, 2011). The output of the R code is then parsed again through perl and a final CSV file is made with the corrected data of each experiment. This script was executed in parallel for each dataset on a High-Performance Computing (HPC) cluster.

3.2.1.3 Alignment

Two different alignment techniques were tested on single chromatograms before one was chosen to be applied to the data large-scale. Considerations included constraints in their implementation - as it was desired that the final scripts be compatible with the Linux-based HPC cluster; their computational efficiency, or their conceptual compatibility. Another important factor is that the method



(a) Before correction

(b) After correction

Figure 3.1: Correction of baseline using alignDE

should be extendible to 2-Dimensional separation data. Additionally, only open-source implementations were used in the interest of reproducibility.

Among these were methods that relied on peak detection before alignment. As mentioned by Tomasi *et al.* (2004), this has the potential to introduce unwanted bias. Seeing, however, that the downstream analyses depend heavily on peak detection, this would not be such a great disadvantage with this particular approach.

AlignDE was one such method. The alignment works by first detecting all peaks in a signal; grouping clusters of peaks together; then using Differential Evolution (analogous to a modified genetic algorithm) to match the peaks according to a global optimal correlation. Similar to COW, it warps the spaces between the peaks in order to reach this optimum.

While it served excellently in baseline correction of the signals, there were several problems with alignment on a larger scale; most notably the fact that it is difficult to extend to 2-Dimensional detection techniques - whereas for COW this is a relatively simple extrapolation (as will be demonstrated below). Additionally, it is more appropriate for data that comes from a similar source as the algorithm allows for the peaks to be matched over a large spans of the signal non-sequentially (Zhang *et al.* (2011)). In measurements from heterogeneous experimental sources these shifts may be done spuriously if some peaks are not present in both chromatograms. COW is designed for more heterogeneous data alignment as it is insensitive to peak features, and will not gratuitously shift a peak that is not present in the target (doing so will not improve the correlation - see equation 2.1.12).

Thus, COW was the algorithm of choice for alignment of the multiple chromatograms under question. Conceptually it was the most appropriate for the data, which is heterogeneous and continuous. Implementing the algorithm on a large scale was not too costly in computational time, and parallelising it in an open-source environment, while requiring some adjustment, was not infeasible. Additionally a relatively simple extension to 2-D data was possible. Parallelisation of the alignment process involved simply running a separate job for each dataset CSV file.

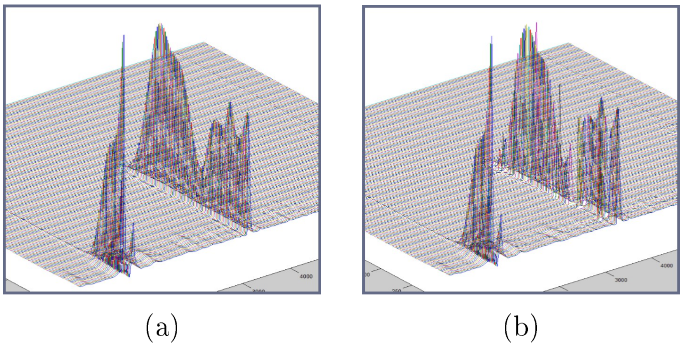
An implementation of COW was written in MATLAB by the authors of Tomasi *et al.* (2004). The code was translated into Octave, which required little alteration, and run on an HPC cluster through another octave script. To run multiple alignment jobs simultaneously and without employing a differential alignment such as in PyMS, a simplification had to be made as to which datasets were compared to each other in order for all experiments to be globally comparable. To this end a single HPLC/UV-vis data set was chosen against which all other data sets were aligned (centroidisation).

This is a simplification with many issues, and one should take special care in relatively disparate data sets. The downstream validation of peaks (refer to section 3.2.2) should overcome some of the potential errors in alignment that may result from this; however an attempt was made to at least select for a run that is rich in peak features. As stated above, COW should theoretically avoid falsely aligning peaks in a sample when none are present in nearby segments of the target (as this has no increase in the benefit function in the optimisation step); the target was thus chosen for the number of significant peaks present as it is believed that misalignment of different features could be corrected in the next step.

An initial run of peak detection therefore needed to be performed to characterise the peak richness within each data set. It was performed in parallel with the method underscored in section 3.2.2 below. Each job was split up on the HPC cluster per HPLC/UV-vis data set and peak detection was run on every wavelength. A summative score S over each wavelength w was calculated for all N peaks along the wavelengths, using the peak height p and average peak height $\bar{p}$:

$$S = \sum_{w=190}^{560} \frac{N_w}{\bar{p}_w} \sum_{i=1}^{N_w} p_{w,i} \tag{3.2.1}$$

The data set with the highest score was an experiment with both treatments (caffeic acid and catechin); highest pH of 7.2; highest SO_2 of 50 ppm, and taken on day 15 of the experiment - which is half way. This makes intuitive sense for a number of reasons: having both treatments present will naturally result in a higher frequency of peaks - not only due to the treatment compounds themselves but due to their probable combined impact on the media's chemical nature. High levels of pH and SO_2 are also the cause of significant variation - as depicted in the following chapter. Sampling half way through an experiment will most likely capture intermediate compounds in the kinetic evolution of the sample, so one could hypothesise that the peak density at this stage should be highest. This sample was therefore kept as the 'centroid' for alignment.



(a) (b)

Figure 3.2: Alignment by consecutive wavelengths. Observe the 'drift' in peaks from wavelength to wavelength in (b)

The segment length parameter m was estimated by observation of the smallest peak width. All wavelengths were considered, and it was set to a conservative value a little lower than the smallest width: 20 time intervals. In Nielsen *et al.* (1998), the slack parameter for COW was set to approximately one-tenth of the segment size. In order to avoid under-fitting the alignment, a slack slightly greater than this was chosen: 3.

The question then remained of how to align 2-Dimensional separation data using 1-Dimensional COW. The first approach was to align each data to the centroid simply by consecutively aligning the corresponding chromatograms at each wavelength. An octave script was written as a wrapper for COW and a separate job was run for CSV data file on the HPC cluster. The results from this analysis were not promising (see Figure 3.2). It seems that the optimal correlation for each wavelength resulted in different alignments. This is indicated by a shift in the peak features across wavelengths. It is postulated that the cause of this could be the over-fitting danger suggested by Nielsen *et al.* (1998); lowering the slack and lengthening the segments, however, did not solve the problem. The issue of 2-D COW alignment is addressed by Zhang *et al.* (2008); however the results of the 'grid-warping' method used are complex; unvalidated and have not been implemented in code. A method was independently developed in the form of a 'TAC' (Total Absorbance Count), commonly used in mass spectrometry in the form of a 'TIC' (Total Ion Count). A TIC is simply a summation of intensities over all mass/charge ratios for each retention time point. Precisely the same method can be used for UV/vis detection, summing all absorbance values over all wavelengths for each time point. The resultant pseudo-chromatogram is then representative of peak positions for the entire spectrum.

The summation chromatograms were then used as a proxy for every HPLC/UV-vis data set alignment. This was thought to be more robust than the former approach, as optimisation of the correlation between such signals is unlikely to encounter the localised variations over wavelengths. This can potentially lead to better comparisons across aligned samples later on in the analysis. Naturally, this is a simplification and may be subject to some error. One of the assumptions behind this is that peak features that span many wavelengths do not vary significantly in time over said span. While this is not strictly true - corrections for this had to be made in section 3.2.2 - in general this variation for peak apexes were small (in the region of 3-7 time intervals).

Once the alignment had been done, it was extrapolated back to the original data using a custom-built transformation script. This is made possible by COW's inherent simplicity: the input of COW is a vector of the beginning and end-points of the segments in order (for $L = 100$ and $m = 4$: $[x_0, x_1, ..., x_{24}, x_{25}] = [0, 4, ..., 96, 100]$), and the output is simply the segment nodes after warping (e.g. $[0, 3, ...92, 100]$). These parameters can be retro-fitted back to the original data by warping the chromatogram at each wavelength in the same way using linear interpolation. The chromatograms of each wavelength are divided into segments of m; the new points p'_i at newly warped times t'_i between t'_0 and t'_n of the new segment are calculated as follows.

The p'_0 and p'_n simply inherit the original p_0 and p_n at the respective original segment bounds of t_0 and t_n. An approximate original time t_i is calculated for new times t'_i by:

$$t_i = t_0 + (t_i - t_0) \times \left(\frac{t'_i - t'_0}{t'_n - t'_0} \right) \tag{3.2.2}$$

The original height at this approximate time is then used as the new height p'_i, interpolated between heights p_i^- and p_i^+ at the integer time values above and below it:

$$p'_i = p_i^- + (p_i^+ - p_i^-) \times \left(\frac{p_i - p_i^-}{p_i^+ - p_i^-} \right) \tag{3.2.3}$$

This interpolation was added to the alignment script, so that the generation of the 'TAC', the subsequent COW alignment and the interpolative adjustment were performed as part of one job. This was done in Octave, calling the alignment implementation by Tomasi *et al.* (2004).

The interpolation of the alignment back across all peaks proved to be an effective tactic for alignment, preserving the integrity of the features and eliminating the drift seen in Figure 3.2. A demonstration of the 'TAC' alignment is found in Figure 3.3. This is fairly representative of most of the alignment processes; for the many cases visually inspected, it appeared that the time shifts were not extreme and only minor adjustments were necessary.

Because the alignment was only done on a single pseudo-wavelength, the running time of the alignment was fairly low (about 70s per job on the HPLC cluster). Once alignment had been performed on each HPLC/UV-vis data, peak detection could be performed towards the end of multiple feature maps for comparison.

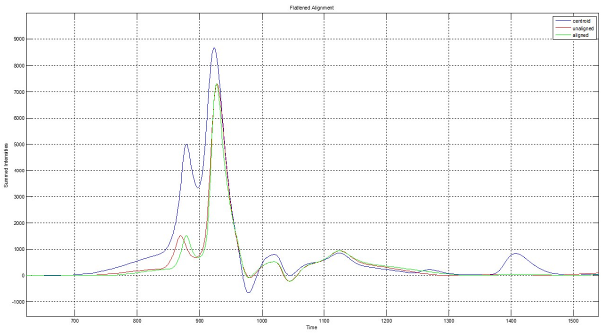


Figure 3.3: Alignment of the summative 'TAC' over all wavelengths.

3.2.1.4 Peak Detection

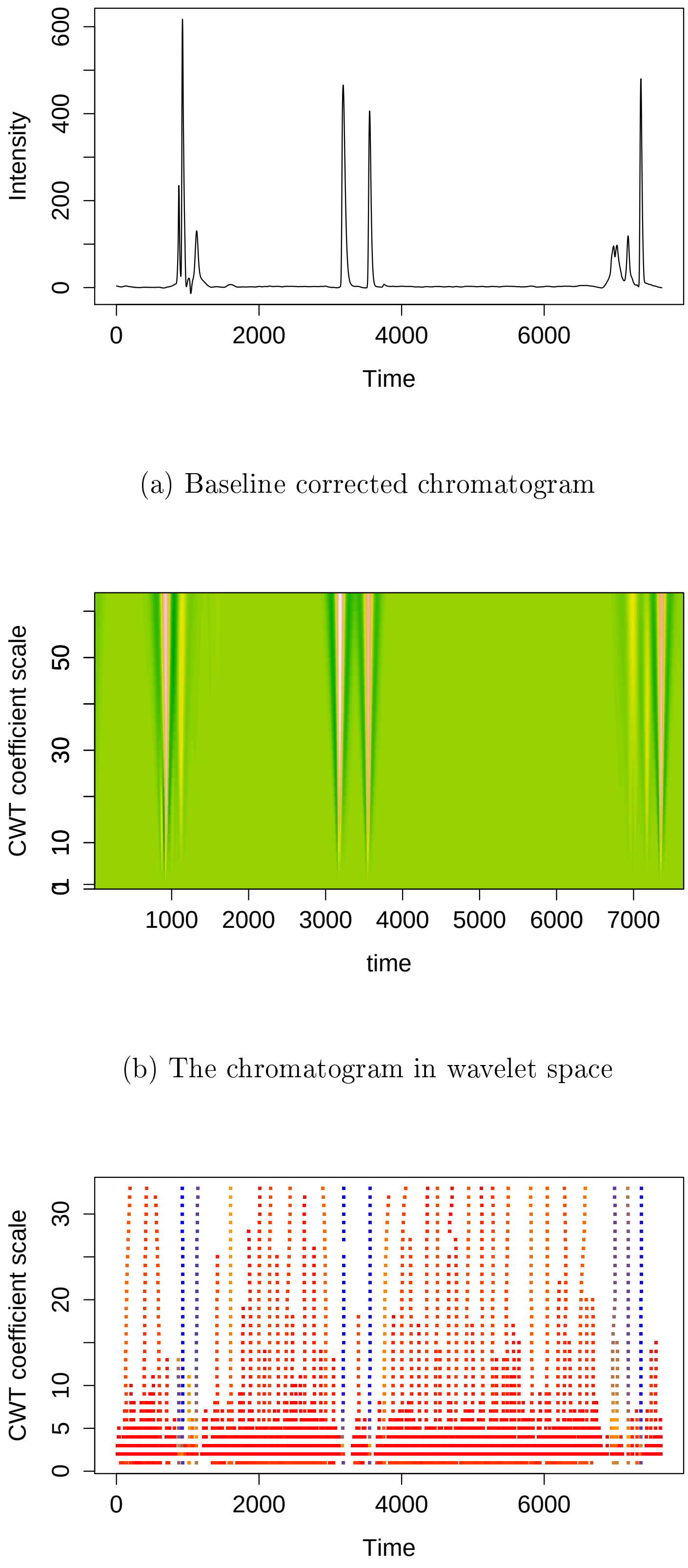
Peak detection was done using an implementation of Continuous Wavelet Transforms . The library used for this was the MassSpecWavelet package, which is part of the Bioconductor project in R and was developed by Du *et al.* (2006). Due to the scale of the data, the peak detection process was again done in parallel on the HPC cluster through an R interface with perl. The peak detection was performed in two steps: firstly, the continuous wavelet transform was created; then peak identification was performed using the transformed output. The first step is depicted in Figure 3.4.

The wavelet scales used were those recommended by Du *et al.* (2006): spanning from 1 to 64 with an interval of 2. It is clear from the high resolution coefficient matrix that there are around six or seven peaks that dominate the wavelet space, having large local maxima at high scales.

In order to define these local maxima, a ridge object is created from the coefficient matrix. This is formed algorithmically from the original wavelet coefficient matrix as follows: firstly, the local coefficient maxima are detected using a sliding window approach, with the window the size of the wavelet support region at the scale. These local maxima must then be approximated as ridge lines for subsequent peak identification. The algorithm is applied by simply creating ridge features within the sliding window, starting at the largest scale in the coefficient matrix and moving downwards to scale one. A maximum gap threshold is set for the ridge feature; any ridge with a gap larger than the threshold is discarded.

This results in another 2-D matrix, as depicted in Figure 3.4. The ridges are coloured according to the coefficient strength at each scale point - blue being the strongest, down to yellow - close to zero. The major peaks are easily identifiable as ridge features with the requisite coefficient strength values.

The SNR ratio threshold is perhaps the most important parameter to set, and has a significant effect on which level of peaks are detected. As further stated by Du *et al.* (2006), noise can be assumed to be either negative or positive peaks with very small width, and can thus be approximated by the



(a) Baseline corrected chromatogram

(b) The chromatogram in wavelet space

(c) Resultant ridge lines

Figure 3.4: Continuous Wavelet Transform for Peak Detection

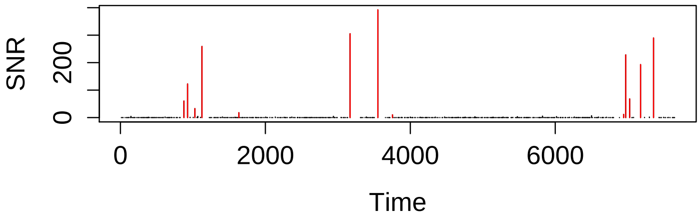


Figure 3.5: Depiction of the SNR ratio for each detected peak in the chromatogram (same signal as for Figure 3.4). The SNR ratios in red are those above the set ratio of 7.

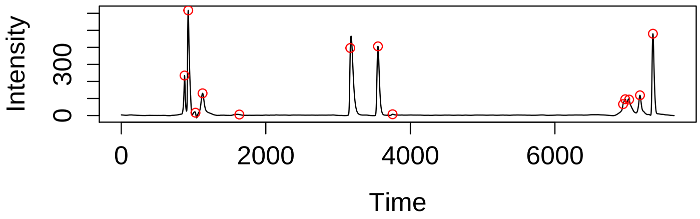


Figure 3.6: The final result of the peak finding algorithm on the CWT coefficient matrix.

wavelet coefficients at small scale values. They further define the local noise surrounding a peak to be the 95th percentile of absolute wavelet coefficients at scale $a = 1$, measured within a set window surrounding that peak. This can be displayed in the form of a histogram for each peak (refer to Figure 3.5).

Besides for the threshold of SNR ratio, two additional restrictions are applied in order for a ridge feature to be identified as a peak. Firstly, the maximum amplitude on a ridge (which reflects the width of the peak) must lie above a set wavelet scale; secondly, the overall ridge line length must be above a certain value. The latter eliminates the small peaks that are often found near major peaks, which are simply artefacts of the latter. Default values for both of these thresholds proved effective. The final peak position is estimated by the position of the ridge among its lowest wavelet scales. The result of the peak detection process is shown in 3.6.

The above figures depict the peak finding algorithm for a single sample at one wavelength. The data at hand, on the contrary, spans just under a thousand samples and is measured at 186 wavelengths. The method therefore had to be generalised to account for this. The aforementioned perl-R interfaced program was set to run through all wavelengths within a sample, generating a

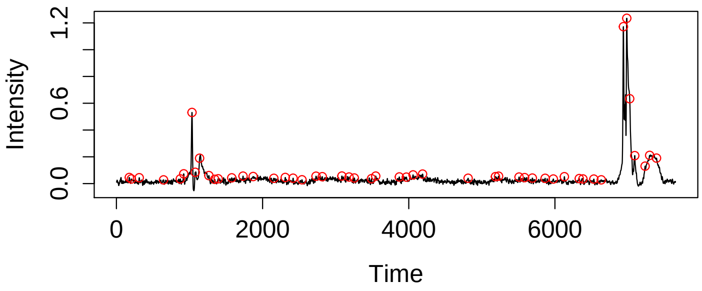


Figure 3.7: Peak Finding with SNR ratio 8 at 550 nm (near highest) as compared to that at 190 nm in Figure 3.6

list of peaks, saved in text format for later processing. The program was run separately for each sample in parallel.

Seeing as the method was applied across all wavelengths, it was important to optimise the signal to noise ratio carefully. The danger of a high signal to noise ratio is that smaller peaks, which may still be of chemical significance, may be unwittingly eliminated from the resultant peak list. If it is set too low, then noise is incorporated into the peak list, which results in spurious comparisons and false results in the analysis downstream. In the case of the HPLC/UV-vis data, clearly defined peaks were visible at the lower wavelengths in the UV spectrum; however for the higher wavelength values the peaks were closer to noise and the features were not clearly delimited over spectra. A fairly low SNR ratio was appropriate at the lowest wavelength of 190 nm, however at the higher wavelength of 550 nm the small peak height demanded a much higher threshold (refer to 3.7).

One solution for this issue would be to dynamically change the SNR threshold for higher length wavelengths, defining spectra beyond which the threshold is reduced. While easy to implement, there is no guarantee of where the low-lying peaks begin on the spectrum for each HPLC injection, and checking this property for every dataset defeats the purpose of mass processing. The SNR threshold for the lower wavelengths, where the peaks had consistently higher amplitudes across all datasets, is more definitive.

The problem was reconciled by applying a simple height filter across all wavelengths. The filter was set low enough to eliminate the noise seen at the higher wavelengths and depicted in Figure 3.7. The level of this noise was consistent across many of the datasets, and its elimination led to a much more structured and unconvoluted peak landscape as shown in 3.8. While this is a very simple abstraction to apply, it was deemed appropriate given the nature of the data and the scale of the analysis.

Once this threshold was set to a conservative level, the SNR ratio was adjusted to its appropriate level respectively. This involved viewing sample

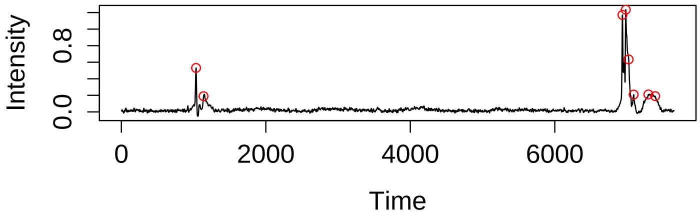
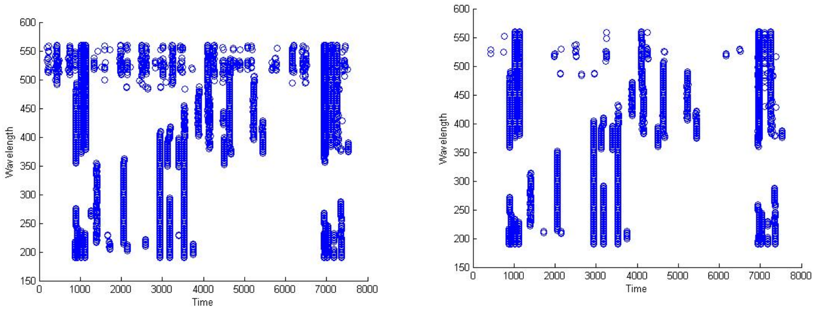


Figure 3.8: High wavelength signal with a minimum noise level threshold of 0.05



caption

(a)

Figure 3.9: Identified Peak landscapes with SNR threshold set to (a.) 5 and (b.) 8 respectively

datasets either in a peak 'landscape' plot - where the detected peaks were plotted against wavelength and time in a 2-Dimensional scatter plot (Figure 3.9); and alternating between conventional peak views at lower and higher wavelengths as in Figures 3.6 and 3.8 respectively. Scripts were created in R and MATLAB to facilitate this.

An important aspect of the final peak landscape is that it should not include noisy features that are inconsistent in retention time through the spectrum. In theory, a feature at a particular retention from HPLC should be absorbed through a number of spectra with little deviation on the time axis. Essentially, this refers to the straightness of the feature line; a concept that is algorithmically formalised in the following section 3.2.2. A good compromise between this ideal, and the ability to detect low-lying peaks at all, was found in an SNR threshold of 8 for the HPLC/UV-vis data. This was arrived at through systematic trial-and-error, using the centroid data set as a primary test and validating with other edge-case experiments.

3.2.2 Feature Matrix Generation

3.2.2.1 Problem Statement

As mentioned above, the ultimate aim of this analysis is to have a comparable list of features, where each feature represents a compound and is either present and quantified with a single value; or completely absent from a sample. This can be represented as a matrix, where the columns are putative compounds and the rows the samples for comparison. Once a conserved matrix is obtained, it can then be subjected to any number of statistical modelling techniques.

Once again, this method of dimensionality reduction applies a simplified abstraction to complex data, but this can be justified both from its means and end. One may ask why entire HPLC/UV-vis data sets cannot be compared to each other without being forced into the context of peak detection. Indeed, this would be the least bias comparison, and as the chromatograms are aligned, this can be a valid approach. An implementation of this was attempted on the data, but it failed on a practical front. Due to the fact that the data for each sample is 2-Dimensional (2-Dimensional separation coupled with intensity values), it needed to be vectorised in order for each sample to occupy a single row in the resultant matrix. This was achieved by 'unfolding' the separation values; making a column entry for each wavelength at each respective retention time. The number of columns were therefore the retention time range (7650 points - a resolution below which one loses information) multiplied by the absorption spectra (186 wavelengths total) - totalling 1,422,900. Unfortunately, this column number is simply too large to be parsed into most mathematical languages through text files, much less used for the proposed techniques. A reduction of the resolution to achieve a manageable scale would also detract too severely from the quality of the data.

There are, of course, ways of splitting up the data into more digestible subsets. A script was made that iterates through each chromatogram's CSV file, extracting the signal for each wavelength and writing to separate individual wavelength CSV files. The problem with this is that wavelengths are highly correlated, and features can span many wavelengths (consider the lines in Figure 3.9). Therefore, comparing samples across single wavelengths, while overcoming the practical limitation of vector length, was inconsistent with the nature of the analysis, which is to compare complete features.

Describing chemometric data within the context of peak features is important for later interpretation by chemists. Thus no matter the method of comparison, the recording of feature position is important. From this point of departure, an efficient method was devised to firstly consolidate the peak features detected by the wavelet algorithms described in section 3.2.1.4; secondly to compare these consolidated peak lists to each other such that putative peak features can be compared across samples in a novel implementation of a feature map alignment algorithm.

Interpretability is also essential. If a compound is thought to exist and have some kind of impact on the kinetics of a sample, for example, then that feature must be traceable to an existing compound. Another requirement is that the feature should ideally have its magnitude described by a single value, so that all the data can be reduced to a single matrix.

Furthermore, the information contained in the feature matrix, while represented by a single-dimension number, should contain the information of the 2-D extraction. Using a 'TAC' to asses peak similarity would constitute a glaring oversight in that (especially in a summative feature) two slightly different chemical features may co-exist at the same time point, and exhibit different absorption patterns across wavelengths.

3.2.2.2 Overview

The question of how to sequentially build this feature matrix arises. Some of the requirements for this matrix have been outlined above: the features must be globally comparable and there must be a conserved list to compare across samples.

A number of approaches to fulfil these requirements were considered. The order and manner by which chromatograms were cross-compared was a crucial detail. Underlying this consideration was the fact that the alignment and peak position detection procedures must be imperfect; thus peaks for equivalent putative molecules (sharing identical absorption patterns) may reside at slightly different time points in the chromatogram.

The most common and simple method of multiple comparisons is an all-against-all approach. Its simplicity is often outweighed by its 'brute force' computational intensity; in this context, however, there are larger conceptual concerns. The peaks need to be globally comparable and an all-against-all analysis is by nature dual-comparable. While this will be accurate and interesting on a local level, the comparison is not extensible to a global scale.

The developed solution was in the form of a dynamic peak database. The concept behind this was to retain a 'master list' of peak features against which all chromatograms are compared, similar to the method of MZmine (Katajamaa *et al.*, 2006). This list or database of peaks is initially populated with a 'seed' chromatogram, and if a feature in a subsequent chromatogram is not found in the database it is dynamically added to the database for future comparisons. Thus if the alignment has a predictable level of accuracy, and new peak features are added as discovered, the sequential manner of comparison should not miss meaningful feature matches between chromatograms.

To satisfy the stated requirement of meaningful chemical interpretation, a second database was maintained that stored each unique peak added to the comparison database, as well as the identity of the original chromatogram in which it was found. The list is then written to a text file for later reference.

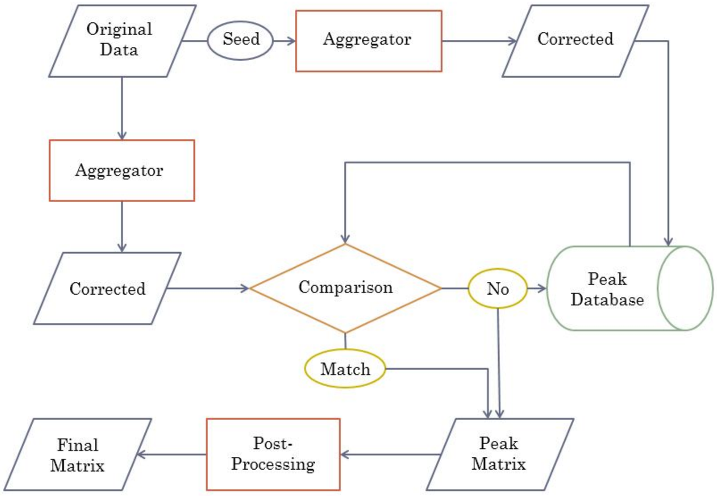


Figure 3.10: Depiction of workflow for comparison of all peaks in the generation of a feature matrix

While this is a relatively concise concept, the implementation was quite involved given the number of experiments compared, and to this end a perl program was written from scratch to import the data, manipulate it and write out the results. A depiction of the workflow is included in Figure 3.10 below.

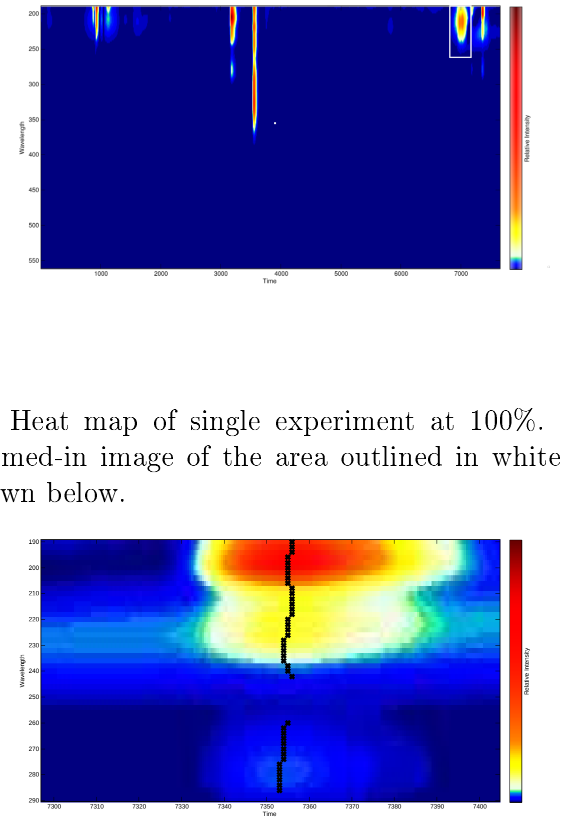
Chronologically, there is an initialisation step (traced from left to right along the top part of figure 3.10) whereby the database is created, followed by the main comparison step from the top left down and to the right. A final pruning and correction step is then taken from right to left at the bottom of the flowchart.

The feature 'aggregator' is described in detail in section 3.2.2.3 below. The dynamic database comparison is then elaborated in section 3.2.2.4.

3.2.2.3 Aggregating Peak Features

For the sake of clarity with regards to this analysis, a feature is defined as the apex of a peak along all its wavelengths in the spectrum.

Before feature map alignment can take place, the feature map for each experiment must be generated from the detected peaks. This was achieved using some custom algorithms, using window searching techniques as is common in feature map alignment software (Lange and Tautenhahn, 2008). Figure 3.9 depicts the detected peaks across the separation dimensions. The slight 'distortions' of the peak features over wavelengths are clearly visible over time, in the form of deviations from the straight lines one expects in compound features - effectively small drifts in time over wavelengths. These deviations can be attributed to any number of causes; exploring the features visually in 2-Dimensional space gives an idea of how this happens. This is demonstrated in heatmaps in Figure 3.11. From this and several other observed examples,



(a) Heat map of single experiment at 100%. A zoomed-in image of the area outlined in white is shown below.

(b) Enlarged peak to the far right of (a) with detected peaks

Figure 3.11: 'Drift' of detected peaks along a single feature.

it was clear that this it not an artefact of the peak detection algorithms, but rather a change in the shape of the peak itself over wavelengths.

Naturally, when one wants to compare peak features between different samples, it is advantageous to compare features along a single time point; comparing along a time frame increases the complexity enormously. Thus the peaks features exhibiting drift were aggregated and re-assigned to a central time point. The peaks, with the exception of low-lying peaks at higher wavelengths, adhered to this aggregation well as the drift was generally not severe. A few parameters are needed to direct the aggregation, and are shown in algorithm 1.

The algorithm iterates through a moving window of set size that shifts along the retention time one interval at a time. At each time iteration, the window is defined by an interval either side of the central time (in this case, an interval of two on either side was considered). The algorithm then passes

Algorithm 1 Define peak features

```
Import Peak list and data matrix for HPLC run
Select a time window size of $n$ and apply over all $N$ times
Keep reference structure of done peaks ($D$)
Keep reference structure of found peaks ($F$)
for $n \in 1 : N$ do
    keep count of number of consecutive peaks ($c$) along each wavelength ($w$) over entire spectrum ($S$)
    Store last peak wavelength: $w_l \leftarrow w_1$
    for $w \in S$ do
        for $t \in n$ do
            Select a consecutive tolerance gap $g$ between peaks
            Select a minimum number $m$ for $c$ to add to $F$
            if $w - w_l < g$ then
                $c \leftarrow c + 1$
            else if $w - w_l > g$ and $c > m$ then
                Add wavelength range to peak feature in $F$
                Start new putative feature at $w$
            else if $a < b$ then
                if $c > m$ then
                    Add wavelength range to peak feature in $F$
                end if
            end if
        end for
    end for
end for
```

through each wavelength in order, checking for peak points within the window. If one is found, it is added to a growing list of consecutive peaks. The minimum number of peak points considered to constitute a feature is set. Additionally, the maximum gap between points is defined, so that separate peaks or spurious detections that appear farther down the spectrum are considered separately.

If the number of consecutive peaks along the spectrum for a certain time window is above this minimum feature length, and the algorithm does not find any further points within the set minimum spectral gap, then that peak feature is added to a list of features for the experiment (feature map). If these requirements are not fulfilled then that consecutive peak list is destroyed and the search along the time point begins anew at the current wavelength.

Only the retention time of the feature and its wavelength interval are stored during the search. The final output includes entire features with corresponding intensity values along all wavelengths. Depiction of the result of the algorithm is displayed for a sample in figure 3.12.

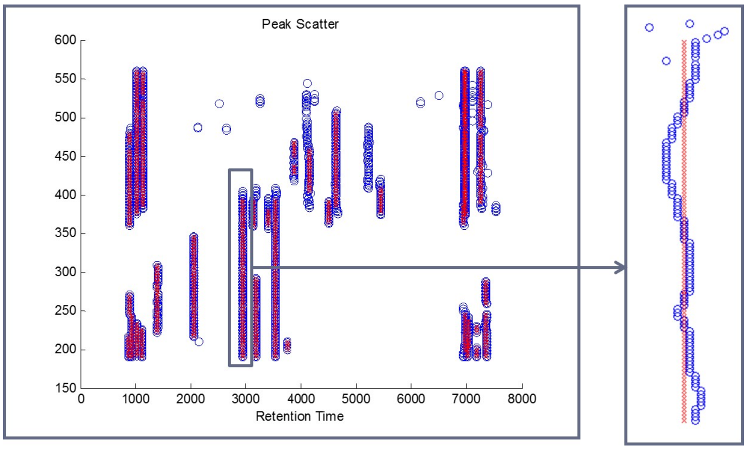


Figure 3.12: Aggregation algorithm as applied to a sample. Focus on one particular feature within the sample is shown on the right.

3.2.2.4 Experiment-Database Alignment

The first decision to be made for initialisation of the database comparison algorithm is which chromatogram is to be used for the initialisation step; the 'seed'. In the case of the demonstrated set of data, the most logical choice was the chromatogram chosen as the centroid for alignment (refer to section 3.2.1.3). This is both because as it was used as the alignment centroid, its peak features should be approximately at the average time of all other peaks; it also contains the most peak features as calculated in equation 3.2.1.

The database itself is stored in the form of a perl reference structure. The reference structure holds tiered information for a feature in order: firstly, the retention time of the feature; secondly, a unique number for the feature at that time (bearing in mind that more than one feature can appear at the same retention time). Attached to this is a hash table of matching wavelengths and intensities, so that the entire feature is stored. As the reference structure is stored in the form of tiered hash tables, it is efficient to search, sort and compare.

After initialisation of the database, each experiment is analysed in succession. Feature extraction is the first step of analysis, after which the feature set is compared to the features in the database. If it is found that, within some tolerance, the feature matches one of the features found in the database, the maximum point in the feature is recorded. If a significant feature is, however, extracted from the experiment but is not found in the database, it is added to the database for future comparisons. This process is outlined in algorithm 2 below.

When a new feature is considered by the algorithm, a linear coefficient (Pearson) is calculated between the two features. A linear comparison between the two peaks was chosen as it is a simple and efficient means of determining qualitative similarity between features, and is insensitive to the peaks'

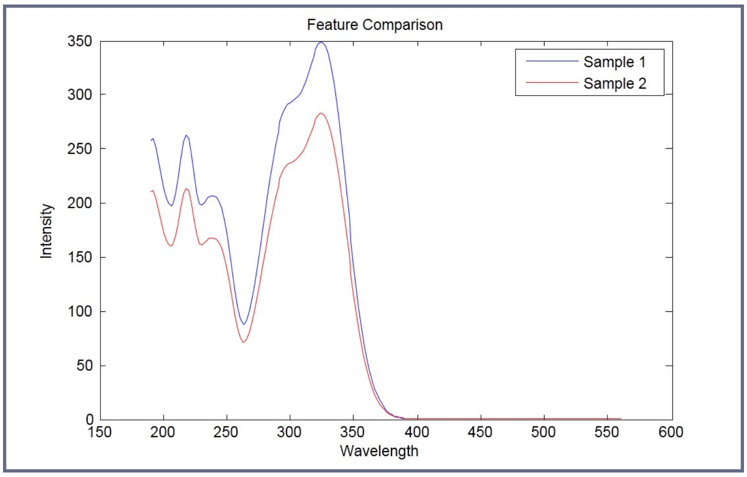


Figure 3.13: The comparison of two features to obtain a correlation coefficient

magnitude. The coefficients between features that are visually congruent was generally observed to be very close to one. A figure depicting a comparison between peak features over length is shown in 3.13.

One of the complications surrounding this comparison included the fact that some peaks at the same approximate retention time will only have a very small overlap in the spectrum, and may thus introduce a spurious comparison. A minimum number of overlapping wavelengths was therefore imposed. Another was that parts of features with very low intensity may exhibit disjointed sections, where a continuous feature will exist for a compound at a higher concentration.

An important parameter to set is the retention time tolerance Δt_{min}, particularly with regards to the phenomenon of elution order change. Lange and Tautenhahn (2008) encountered elution order changes in at least one instance for most of the aligned features in their data sets. They attribute this to pressure fluctuations or changes in column temperature. The closeness of features to one another also has a significant influence. Additionally, one of the main phenolic treatment compounds, Catechin, exhibits chirality. Enantiomers are known to swap elution order in HPLC, a fact which is addressed by Okamoto (2002). The possibility of elution order changes were therefore accounted for in fairly relaxed Δt_{min}.

Naturally, the widening of this tolerance foments the probability that none-related features are compared; it should therefore be coupled with a stricter threshold for feature comparison across wavelengths, R_{min}. Several corrections for possible misalignment in this way are also accounted for in the post-processing on the feature matrix at the conclusion of the feature map alignment.

Once each experiment has been analysed in this way, a final 'sweep' through all the experiments is done with the full feature database. For the dataset used, no additional feature matches were found. This is an encouraging result, as it

Algorithm 2 Comparing chromatograms to the peak database, while adding to the database itself

```
Set retention time tolerance for peak similarity $\Delta t_{min}$
Set minimum correlation between features $R_{min}$

for Experiment retention times $t_e \in P$ do
    for Database retention times $t_d \in D$ do
        if $|t_e - t_d| < \Delta t_{min}$ then
            Extract features $f_e, f_d$
            Record maximum intensity $i_{max} \in f_e$
            if Correlation $(f_e, f_d) > R_{min}$ then
                Add entry $[f_d, i_{max}]$ to feature matrix $M$
            else
                Add $f_e$ to $D$
                Add entry $[f_e, i_{max}]$ to $M$
                Record experiment name, $t_e$ in $O$
            end if
        end if
    end for
end for
```

indicates that the database accrual method may be exhaustive.

Further refinement of the final matrix was performed by comparing the intensities of features in an all-against-all manner, using a similar method to algorithm 2 in an attempt to ensure that the same putative peak is not repeated.

Thereafter pruning exercises were done on the final feature matrix to ensure that spurious entries were eliminated: firstly, features with only a single experimental entry (singletons) were eliminated. Additionally, in several instances it was found that two putative features shared very similar values across experiments. For each pair of features in the matrix, if only 5% of the entries were dissimilar - and to only a small degree, the features were merged by averaging between them. Generally, the name of the original feature with the most entries was kept, and a record of the duplicate features recorded in a final text file for future reference. The possible repercussions of the choice of these thresholds are revealed in section 3.3.

Finally, text files were created to store the final feature matrix, as well as the record of the original feature intervals with their requisite experiment.

This method of feature map alignment is relatively simple, relying on two properties of the data: firstly, a reproducible alignment across all spectra using reliable preprocessing techniques, and secondly, the fidelity of the signature of a feature through the wavelengths of the UV spectra. The validity of this approach is addressed in the results section that follows.

3.3 Results and Discussion

As mentioned in 3.1, the aim of the analysis was to identify putative compounds and thereafter to identify their significance across the experiments. What remains after the method described above for data reduction is a matrix with experiments represented as rows and putative compounds as columns. The intensity of these compounds for each experiment constitute the entries in the matrix.

This feature matrix can be useful in its own right. Firstly, it can be compared to the original HPLC/UV-vis data as a reference. The small shifts due to the alignment and peak database matching procedure should not be large enough to obscure a comparison. In this way the feature matrix can serve as a kind of common mapping for all experiments, so that visible features in the original data from one experiment can be compared to the features of another. Secondly, the feature matrix can be used to identify known compounds. As mentioned in section 3.2.2, a list of the detected peaks is kept along with the original experiment in which the feature was found, as well as the range of wavelengths that the feature spans. An index of comparison for all significant features across experiments has therefore been constructed.

Naturally, validation of the representative nature of the feature matrix is necessary. Two approaches can be taken to this end: firstly, the features can be identified empirically by a qualified chemist. This is outside the scope of this project; so the second approach was adopted: verify if the feature matrix is related to the original data through correlating it with what is known of the experimental conditions.

The latter is more in keeping with the untargeted philosophy of finding significance first, identity later. Additionally it has the advantage of finding the more influential features towards the experimental fluctuations, so that feature identification - a labour-intensive task - can be prioritised on a per-feature basis. A further advantage in reducing the dataset to a set of common peaks is that ordinary machine learning techniques can be applied on the entire dataset, whereas before this was prohibitively computationally intensive.

Three separate techniques were used as data exploration and validation: Principal Component Analysis (PCA), Decision Trees and network analysis.All analyses were performed using python numpy and scipy (Oliphant, 2007), as well as scikit-learn (Pedregosa and Varoquaux, 2011) libraries. The experimental data was kept in a pandas multi-indexed array (McKinney, 2012), and results were visualised in matplotlib (Hunter, 2007).

3.3.1 PCA

A PCA model was fitted to the entire feature matrix for some initial data exploration. The nature of the data is such that there are many confounding

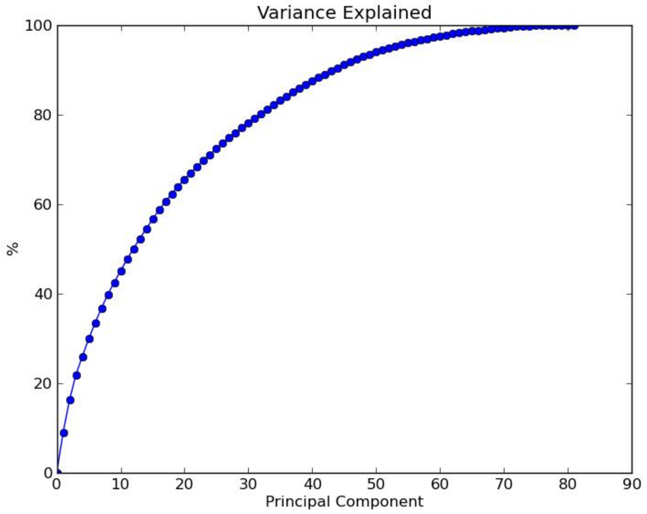


Figure 3.14: Variance explained by PCA on entire feature matrix

factors - from the use of rapid oxidation to the time of the experiment. This is the type of data that PCA is especially good at de-convoluting, by describing orthogonal directions of highest variance. If that variance coincides with some known attributes of the data, then the validity of the data itself can be implicitly verified. Additionally, the variables (putative compounds/features in the current case) most responsible for the observed variance can then be ascribed to the observed attribute.

The question of whether to normalise the data should also be addressed. It was found that distinctions between attribute groups were more apparent if the data was normalised using the standard scale (by standard deviation) and mean-center technique. However, when PCA was applied to smaller subsets of the data, it often did not converge if normalisation was applied. This may be due to the fact that, upon reducing the data to, for example, a single set of specific conditions over the experimental time period; many of the peaks are not present that appear in other conditions. This results in a relatively sparse matrix, which may high collinearity between samples.

The results of the overall PCA analysis with normalisation are shown below. The variance explained by each principal component is first depicted in figure 3.14. The trend of the variance explained is not of the kind generally desired from a PCA model; ideally the cumulative explained variance of the first few principal components should reach near 100%. In the current case, the first two components combined constitute less than 20% of the variance. While this is far from ideal if one wants to exhaustively explain a data set, the purpose here is to validate the dataset itself - if meaning can still be extracted from the first few principal components then some degree of validation can still be drawn.

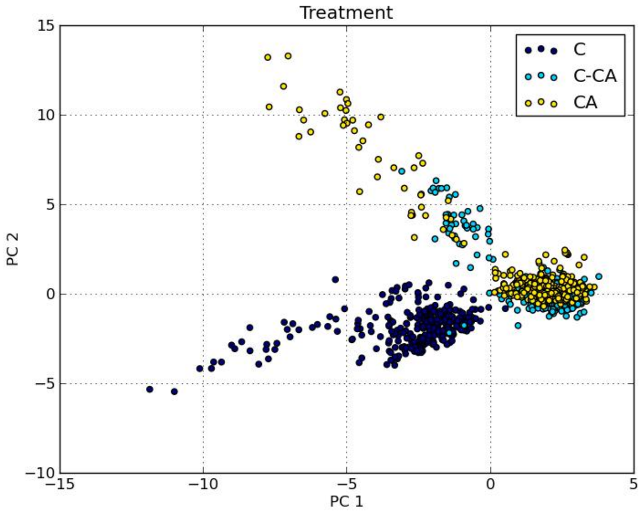


Figure 3.15: Score plot of treatment - 2D

Labelled plots of the principal component scores were analysed thereafter to identify if there is correlation between the principal components and the original experimental conditions. Within the first two principal components, a clustering or separation of conditions was apparent for both treatment type and pH.

The first attribute analysed with PCA was the treatment type - see figure 3.15 with the requisite labels. It appears that the type of treatment is roughly separated across the first principal component. It is also interesting to note that the combined treatment (caffeic acid with catechin) clusters closely with the caffeic acid, and is completely separated from the catechin treatment on its own. This could indicated that the chemical effects of caffeic acid addition far outweigh that of catechin; as their combination seems to vary in more or less the same direction as caffeic acid and not at all with catechin.

In figure 3.16 the data was not normalised, and a 3-Dimensional view is taken with the first three components. In contrast to the previous figure, the combined treatment does not entirely coincide with caffeic acid. Rather, it only partly intersects with caffeic acid while remaining quite separate from both clusters (a thorough rotation of the figure indicates this more clearly). This stands to reason, as removing the normalisation procedure will increase the effects of the respective treatment molecules themselves, while minimising their latent effects on other compounds. If the combined treatment has similar carry-over effects to one of the treatments on its own, this phenomenon will be masked by the concentration of the treatment compounds themselves.

Another attribute that was clearly identifiable through the PCA model was the controlled pH of the medium. The second principal component seem to approximately coincide with the pH labels in figure 3.17. Wine typically

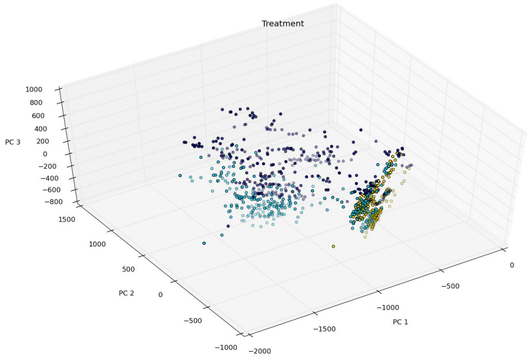


Figure 3.16: Score plot of treatment - 3D, no normalisation

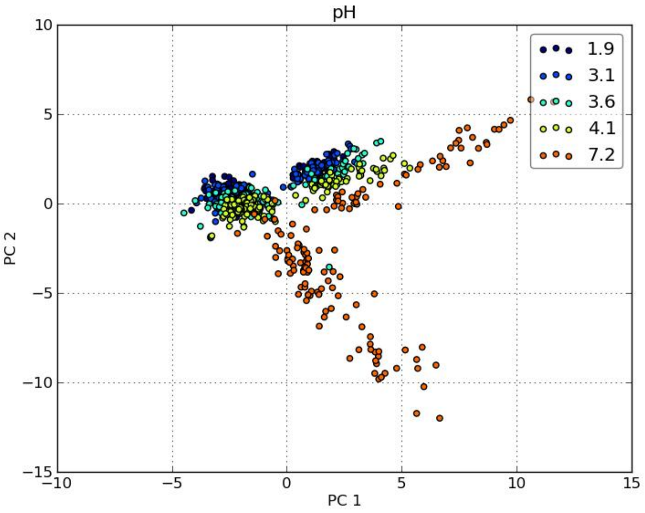


Figure 3.17: Score plot with labelled pH levels

has a pH nearer to 3.6, so it seems logical that the pH far beyond this (7.2) represents most of the outlying scores. In addition, raising the pH far below normal levels seems to have more of an effect on the chemical nature of the medium than extreme lowering of the pH (1.9). Stratification of all levels of the pH is clear, however there is an outlier effect of contracting the lower pH clusters by the pH 7.2 experiment scores.

The oxidation technique also exhibited some separation across principal component two, as seen in figure 3.18. It is possible, however, that this is due to the fact that the rapid oxidation experiments' pH levels were not fluctuated to the full range of the experiments with ordinary oxidation. However, the region spanned by the rapid oxidation experiments does seem to coincide with that of the lower pH levels - which poses the question of the mutual effects of

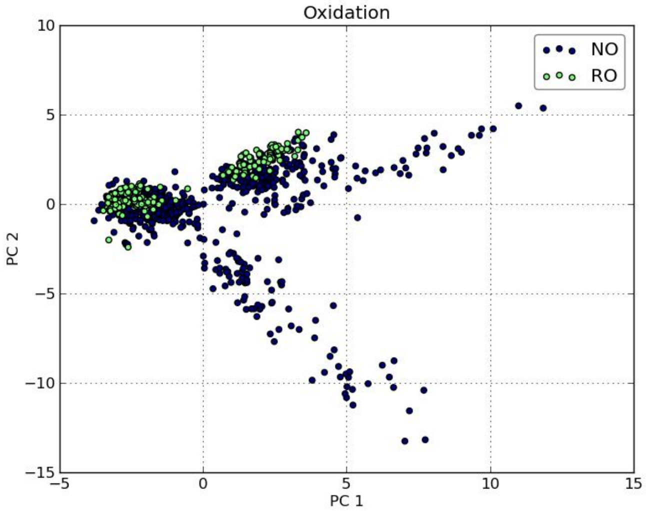


Figure 3.18: Score plot with labelled oxidation type - 'Normal oxidation' (NO) or 'Rapid oxidation' (RO)

pH and rapid oxidation.

This was originally addressed by Ferreira (2002), who found that these are the two most significant mechanisms in the kinetics of the wine media in the aging process. The link between pH level and autoxidation of phenols could be a further cause of this covariance.

Other attributes of the data to be considered were the time of the experiment and the SO_2 content. Analysing all of the data within a single PCA model does not reveal any clustering by time (even for the higher-dimension principal components). A reduction in the scope of the data was therefore attempted in order to see whether time can be ascribed to variance within the data at a smaller scale. A single experiment was chosen (at the conditions that were chosen for the centroid in the alignment - combined treatment at 25 ppm SO_2 and pH 3.6) to apply the small scale PCA. The results are shown below in figure 3.19. One observes that there is a distinct separation of time values across both the first and second principal component. Two separate clusters of time points seem to be split at approximately one week in the experiment. To some extent this does validate that the feature matrix varies according to the time of the experiment; however it appears that across all experiments it is not as influential a source of variance compared to the other experimental conditions.

The relative insignificance of SO_2 regarding the oxidation levels at advanced stages of ageing is mentioned by Simpson (1982) in their own study on the causes of browning. This appears to be confirmed with the absence of clear SO_2 clustering of the data with any combination of principal components.

Finally, the loading plot from the PCA on the entire data set (figure 3.20)

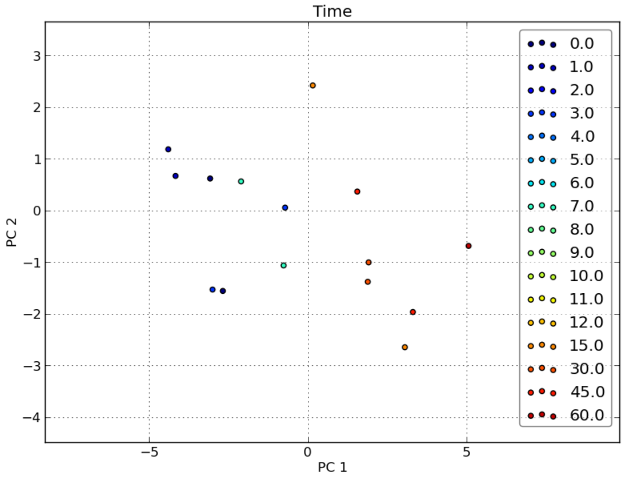


Figure 3.19: Score plot of time with a reduced data set. The legend refers to days after the experiment's commencement

can be analysed to infer the importance of certain putative compounds towards experimental conditions. As mentioned above, it may be possible to prioritise the investigation of putative compounds related to experimental phenomena using PCA. The loading plot can serve as a guide to this end. The compounds most contributing towards the variance of the first principal component, for example, are probably related to the different treatment types. They can either represent the catechin or caffeic acid molecules themselves; or some of the compounds most influenced by their addition. Along the second principal component, compounds possibly related to the alteration of pH levels should be found at either end of the scale along the second axis.

The loading plot corresponding to the reduced PCA model (the scores of which are in Figure 3.19) is shown in figure 3.21. The variance across the first and second principal components - both of which seem to separate the samples over time - is again explained in relative strength by the putative compounds at the far ends of the axes. In this way one can focus on particular experiments to explore chemical changes over time in a single experiment.

3.3.2 Decision Trees

It is now considered whether different experimental conditions for each sample can be used as targets against which to fit a predictive model; a 'supervised' learning approach in contrast to PCA, which is traditionally an 'unsupervised' learning or clustering approach. This has the advantage of using the target classes in the decomposition of the original matrix, which is a more direct approach of trying to infer significance of putative compounds. There are a plethora of machine learning methods available to achieve this; several of

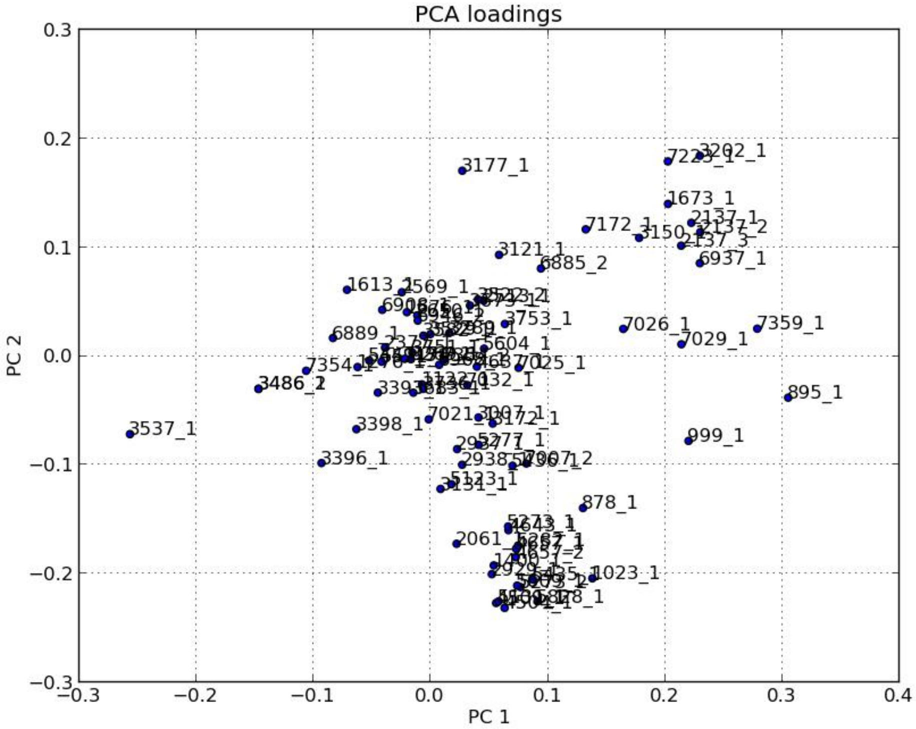


Figure 3.20: Loading plot of PCA on entire data set. Point labels refer to putative compounds

the methods in the scikit-learn library were applied to the data with differing levels of success. The ability to interpret the outcome of a machine learning method is the over-ruling factor with this type of investigation. With this in consideration, decision trees were selected as the primary classifier for their simplicity and effectiveness.

Similar to the PCA analysis, each experimental condition was analysed separately; a decision tree model was fitted using the feature matrix as the predictor and the class labels as the predicted values. An important parameter when fitting and displaying decision trees is the maximum depth. Here the maximum depth chosen is 7; this simply constituted a reasonable compromise between accuracy and interpretation: any higher, and the tree was over-simplified; lowering the depth resulted in an abundance of spurious branches. A program was written to generate decision trees using scikit-learn from the feature matrix pandas DataFrame (McKinney, 2012); thereafter to generate a graph file using graphvis (Ellson *et al.*, 2003).

The decision tree for the treatment type is shown in figure 3.24. The levels of the tree from the top downwards exhibit decreasing effectiveness at splitting the data into the target classes, as measured by gini impurity. The top entry in the decision tree is therefore the variable (putative compound in this case)

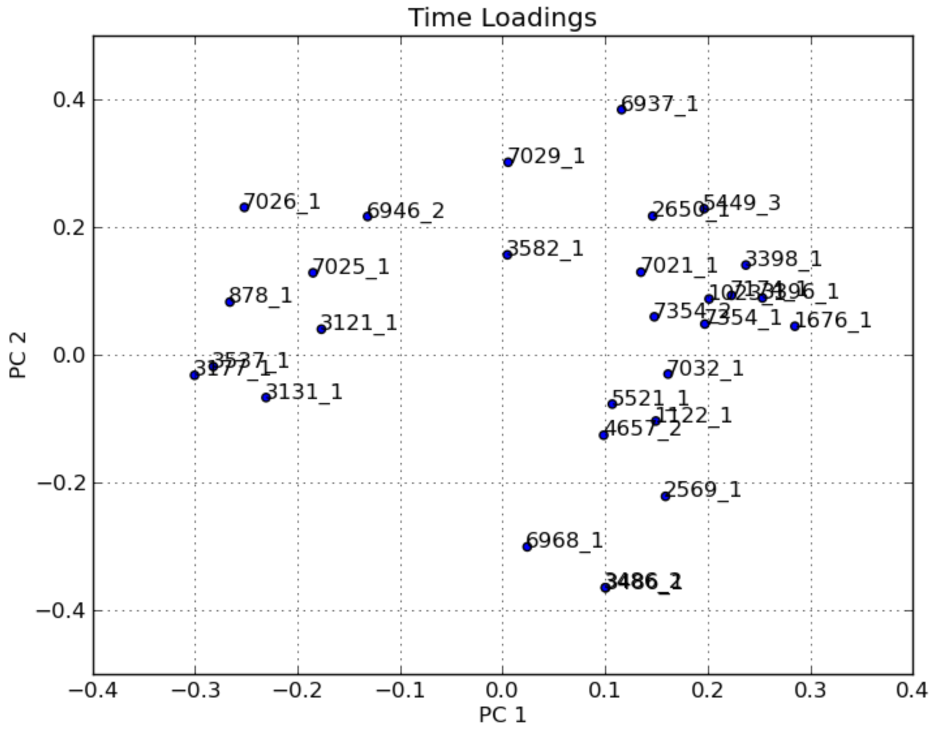


Figure 3.21: Loading plot of PCA on reduced data set

by which the data can be most effectively split into the respective classes. The second level of tree indicate the two next most efficacious compounds with regards to pH classification.

In this way, compounds can again be identified for their relative importance with regards to experimental conditions. In the case of decision trees, the putative compounds are conveniently ranked for importance by their classification efficacy.

The first few levels of figure 3.24 exhibit an effective split in the data. Almost all of the experiments with separate treatments (Catechin and Caffeic Acid) are split down the right hand side of the tree, with zero gini impurity at a very low level. This is in congruence with our expectations of the data; the experiments with different treatments should be readily split by feature intensity while the experiments with the combined treatments should be much more difficult to differentiate. Note that the left side of figure 3.24 extends far beyond the frame; gradually splitting off small numbers of experiments with single treatments until a relatively deep level.

A notable result of this tree is that the compound with the highest splitting efficacy is 3177_1, which is one of the most influential compounds in principal component 2 as seen in figure 3.20. In a similar way, on the second level of the tree, compounds 3537_1 and 7223_1 are at the far ends of the left and right of

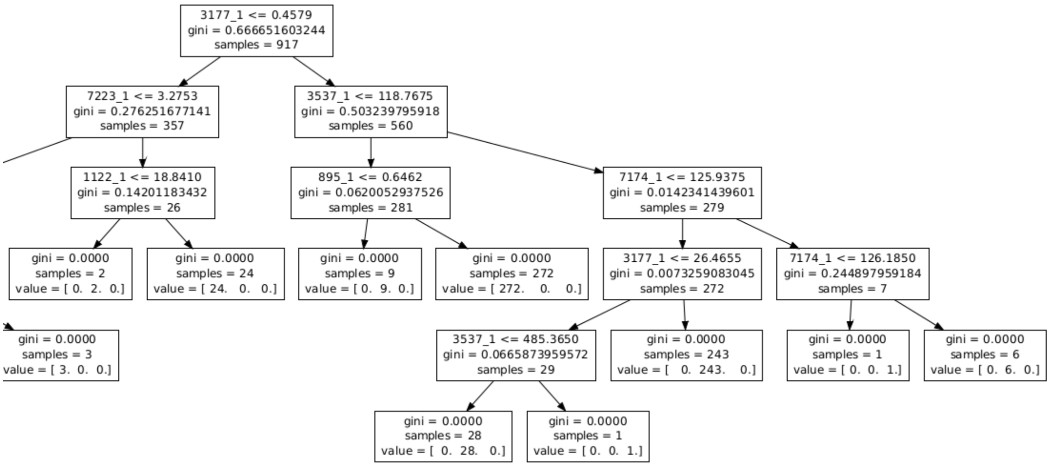


Figure 3.22: Decision tree for treatment class - groups represent (C, CA, C-CA)

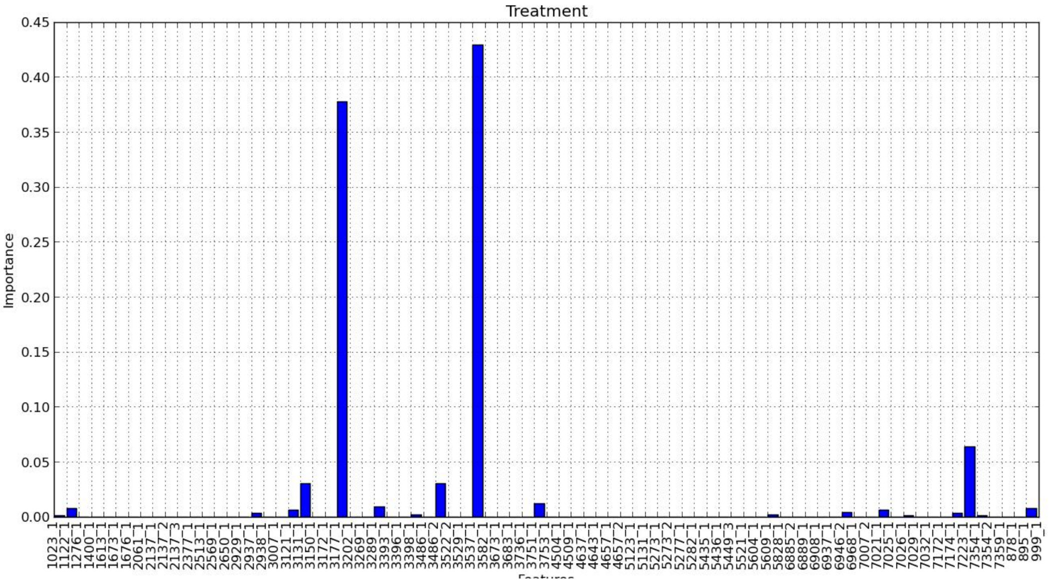


Figure 3.23: Gini importance for variables in the decision tree for treatment

principal component 1. The ranking of the importance of putative compounds can be re-enforced by cross-referencing in this way. This also served as an additional layer of validation for both machine learning techniques, as well as the integrity of the data itself. The methods for the application of PCA and Decision Tree learning are completely different, as can be evidenced in the descriptions in the literature review; their agreement is therefore a non-trivial occurrence.

The gini importance (Rokach and Maimon, 2005) for the features, as related to the model for treatment type is shown in figure 3.23. This validates

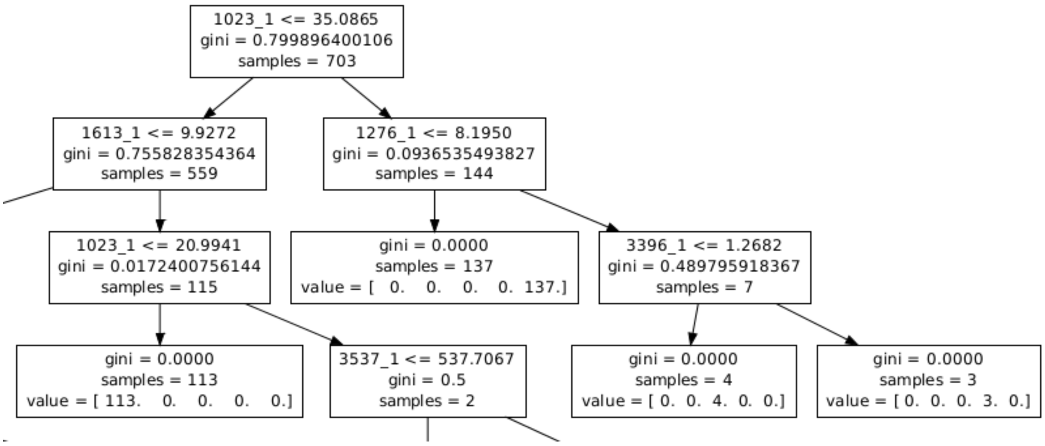


Figure 3.24: Decision tree for pH class - groups represent (1.9, 3.1, 3.6, 4.1, 7.2)

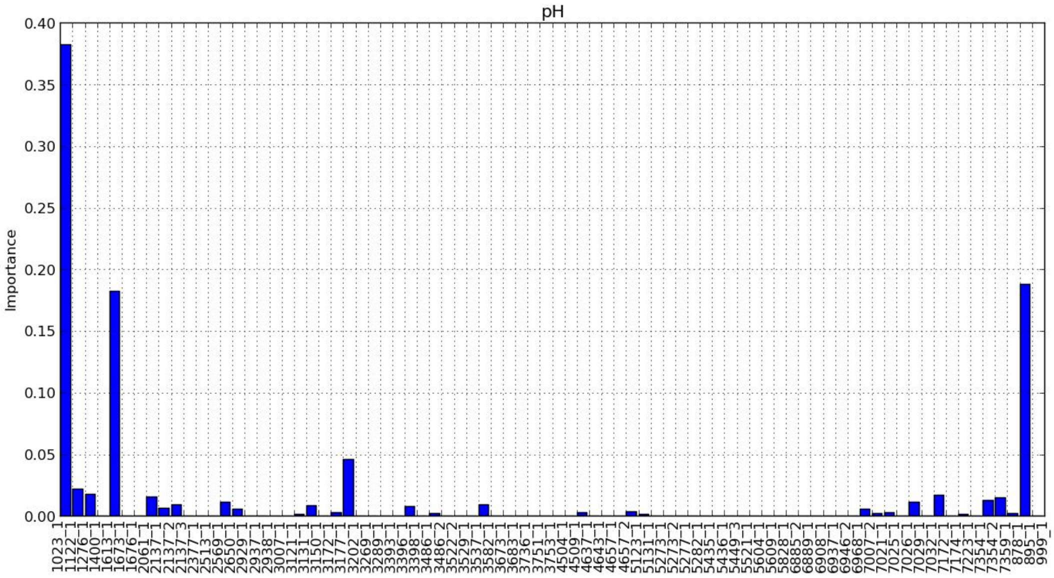


Figure 3.25: Gini importance for variables in the decision tree for ph

the observations already drawn from figure 3.24. As expected, the first three variables used in the classification have the highest gini importance; in fact according to scale they appear to dominate the feature space completely.

The second decision tree, classifying the data into groups of pH, was also effective at splitting the data into classes in relatively few recursions. A question that arises with this kind of data is whether to fit either a regression or classification tree, seeing that the levels of pH can either be distinct labels or continuous values. In this case a classification tree was selected; this was due to the fact that the pH levels were an a-priori controlled condition - not a definitively measured variable during the experiment.

The first group of experiments to be cleanly classified with a gini impurity of zero are those with a pH of 7.2. Again, this is in line with what we would expect given the PCA score plots; as mentioned above, these experiments are sources of high levels of variance within the data. As such it is unsurprising that they are the first substantial subset to be pared of from the rest of the data - almost all of the experiments with pH of 7.2 are accounted for in the first large-scale pure classification. This is after the model specifying the levels of only two separate feature.

The second large group to be isolated with zero impurity are all the experiments with the lowest pH - 1.9. As the other experiments are at pH levels close to ordinary in wine, it seems natural that the two extremes are most easily identified with the fewest number of significant compounds.

The feature importances in 3.25 show more of a range than that for the features in the treatment tree - if features are selected for further investigation using gini importance, then a larger range such as this could facilitate a more interesting study.

The decision tree constructed for oxidation is shown in figure 3.26. It is clear that the model is not as effective as for treatment and pH; wherein most of the significant classes were almost distinguished in the first few levels. However it should still be possible to derive meaning from the tree - one can follow the branches with the highest impurity decrease toward the graph's outer leaves.

An observation that is of potential interest is that two of the more important features at the beginning of the spectrum in Figure 3.27 are shared with pH in Figure 3.25. The overlap in the score plot with oxidation and low levels of pH mentioned in section 3.3.1 may have a link with this phenomenon.

While the top node is not identifiable in the loadings of the first two principal components as shown in figure 3.20 above, it is worth noting that features at retention times near 7032 are at at the far ends of the third principal component.

The decision tree model for SO_2 is not shown due to the fact that its accuracy and effectiveness was not high, the same difficulty in distinguishing SO_2 classes that was found for the PCA model.

The development of models to describe variable outputs can result in falsely high accuracy if there is an underlying bias in the data. In the present case, where the data is constructed using previously untested methods, it is crucial that the model is validated to verify that it is not simply fitting collinear, dependent variables to the output classes.

Validation tests were performed for each of the experimental conditions in the form of random class permutations. This involves the random shuffling of class labels in the target vector before fitting a test model, for a set number of iterations.

The results are shown in figure 3.28. The accuracy is defined as the fraction correctly labelled classes using a decision tree at a given depth of 7. Displayed

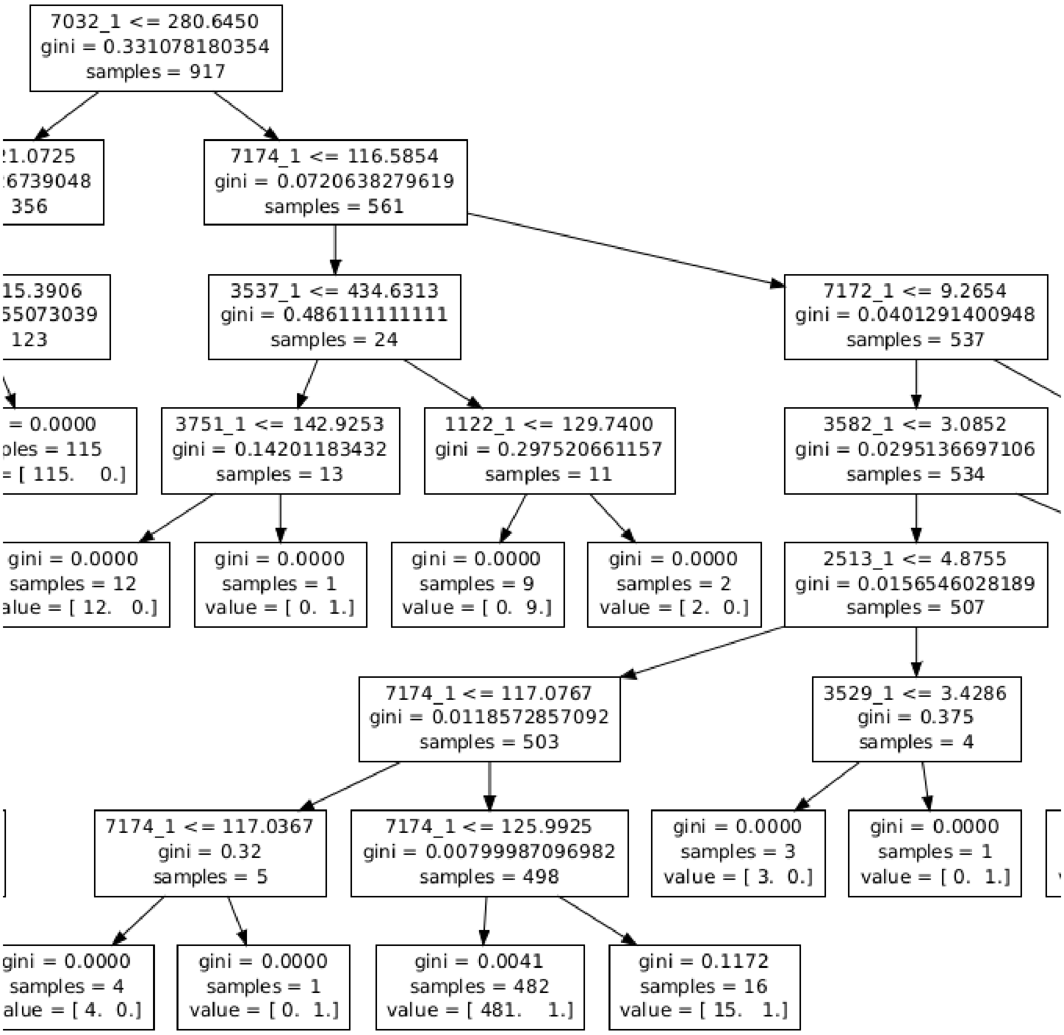


Figure 3.26: Decision tree for oxidation class - groups represent (Normal Oxidation, Rapid Oxidation)

in each figure are the results for the random permutations in blue bars; a vertical green line denoting the accuracy of the model with a correctly labelled target vector; and a black line for what the accuracy should be in a completely random assignment of class labels. The probability given in each figure legend is for how likely it is that a random class label permutation will give an improved answer over the original label ordering.

In each case, this probability is so small as to reach it's finite limit. The correct alignment of class labels always results in an improved model accuracy. This is a very positive result - since essentially it means that there is some underlying integrity to the data with respect to the experimental inputs. Should the feature matrix represent the original data poorly, the shuffling of class labels in the target vector would have no effect on the accuracy of the

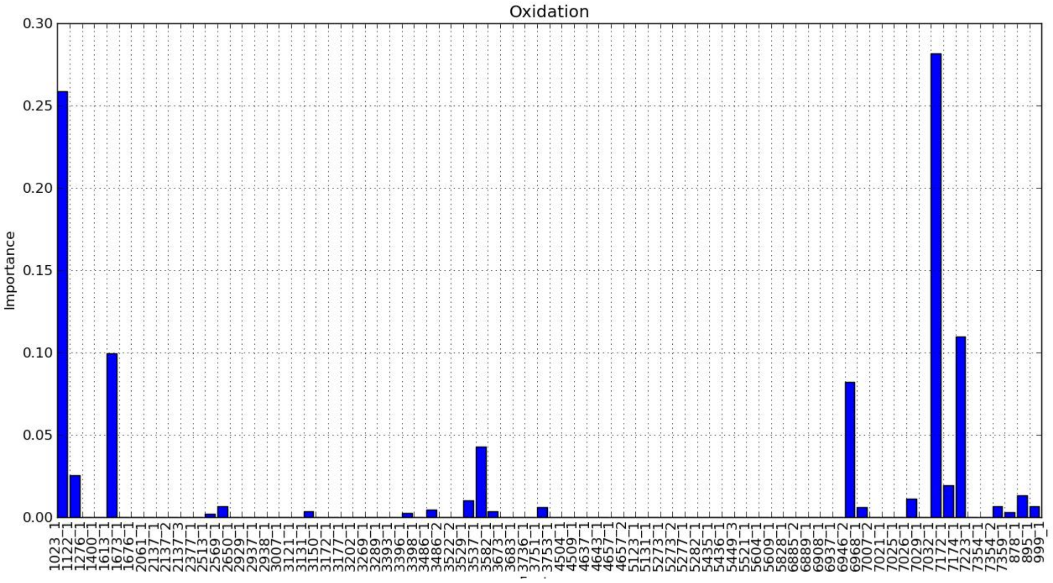
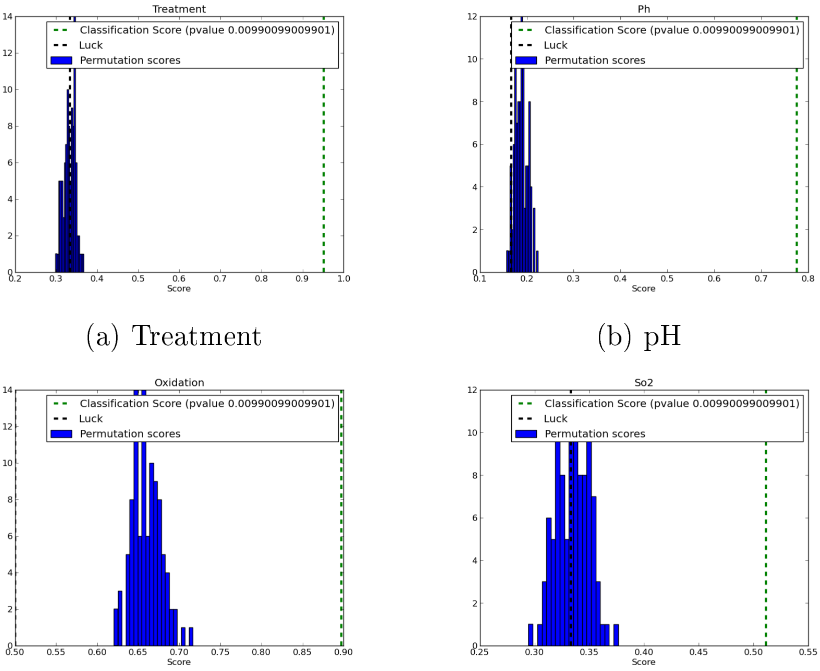


Figure 3.27: Gini importance for variables in the decision tree for oxidation



(a) Treatment

(b) pH

(c) Oxidation

(d) SO_2

Figure 3.28: Random permutation tests on decision tree models for each experimental condition

outcome.

The figures are also instructive as to which experimental condition is most easily predicted using a decision tree; in accordance with all previous findings it appears that treatment, pH and oxidation (to a lesser extent) are most effectively modelled, while SO_2 is less easily predicted from the feature matrix. However there still appears to be some power of prediction, seeing that the correct label ordering categorically out-performs the random permutations.

3.3.3 Chemical Network

While the previous analysis methods have been instructive in indicating which are the significant features either with regards to chemical conditions or in general (in the case of PCA); the interactions between the features themselves have not been explored.

One way to analyse the feature matrix to this end is in the form of a network. One could view the chemical features as a set of interacting nodes with relative strength. The edges could be weighted according to a number of metrics. In this way chemical flux over the span of the experiments can be represented.

A network was constructed to this end using the python module networkx (Hagberg *et al.*, 2008), and visualized in Cytoscape (Shannon *et al.*, 2003). The network was built by first calculating the Pearson correlation between all of features using the full data set over all experimental conditions. These constituted the edge weights between features. In this case, it was postulated that negative Pearson correlations still indicate an interesting result. For example, the reactants and products in a reaction taking place over time in an experiment will exhibit an inverse correlation - as one is depleted, the other accrues.

Node attributes were also added for the mean value of the maximum intensity for each feature. These are used simply to display the relative levels of the compounds. Due to the fact that some compounds were present in the media at levels in multiple orders of magnitude higher than others, the natural logarithm of this value was used for ease of interpretation. A guide for the interpretation of these figures is given in the section that follows.

3.3.3.1 Network Layout

The networks presented throughout this work generally have the same layout for interpretation. There is a central colour scheme: red denotes a positive value and blue a negative. These colours have a range of shading intensity proportional to their value. Values around zero are white, however these are not often displayed if a significance threshold is imposed.

Nodes have a colour between red, white and blue depending on the chosen metric assigned to entities being displayed in the network. Lines are coloured according to the metric assigned to the interactions between these entities. The thickness of a line reflects this same interaction metric (Centered around zero; interactions with high negative- or positive values will have the same thickness).

3.3.3.2 Maximum Spanning Tree

There are several methods available to present a network constructed from weighted edges in this way; one of the simplest and most elegant is a maximum spanning tree. Keeping the provisos about the simplifications of an MST in mind (Jacobson *et al.*, 2013), the maximum spanning tree for the full feature matrix is shown in figure 3.29.

One feature of this network that is immediately apparent is that the nodes with the highest log-average seem to cluster together. The correlation between these high-concentration compounds is seldom high, suggesting that they may be independently present in the media. The fact that the full matrix is used in figure 3.29 does mean that temporal trends cannot be traced.

3.3.3.3 Correlation Network

The correlation network can be pruned by imposing a simple threshold on the edge weights for clarity. This was done with the full feature matrix in an attempt to identify global relationships between putative compounds. The result of imposing a very loose 0.5 threshold on the data is shown in figure 3.30. It is evident from the presence of several sub-graphs that global relationships between putative compounds are not common. Additionally, many of the strongest correlations are between features at very close retention times. This could mean one of two things: either the putative compounds are the same, or different features at different wavelengths for the same retention time are correlated.

One interesting trend to note from this network is that putative compound 3537_1 is the only compound to exhibit a negative correlation in the network. This same feature is the highest outlier in principal component 2 in figure 3.20; and is one of the most important features for treatment in the decision tree analysis. As mentioned above it is possible that the negative correlation describes the relationship between reactant and substrate, perhaps indicating that 3537_1 is an important pre-cursor to significant substrates in the wine media.

3.3.3.4 Focused Correlation Network

Due to the large variability between experiments - especially considering that experiments at all times are simultaneously compared, it is not surprising that there are few interesting global correlations between features. Focusing on

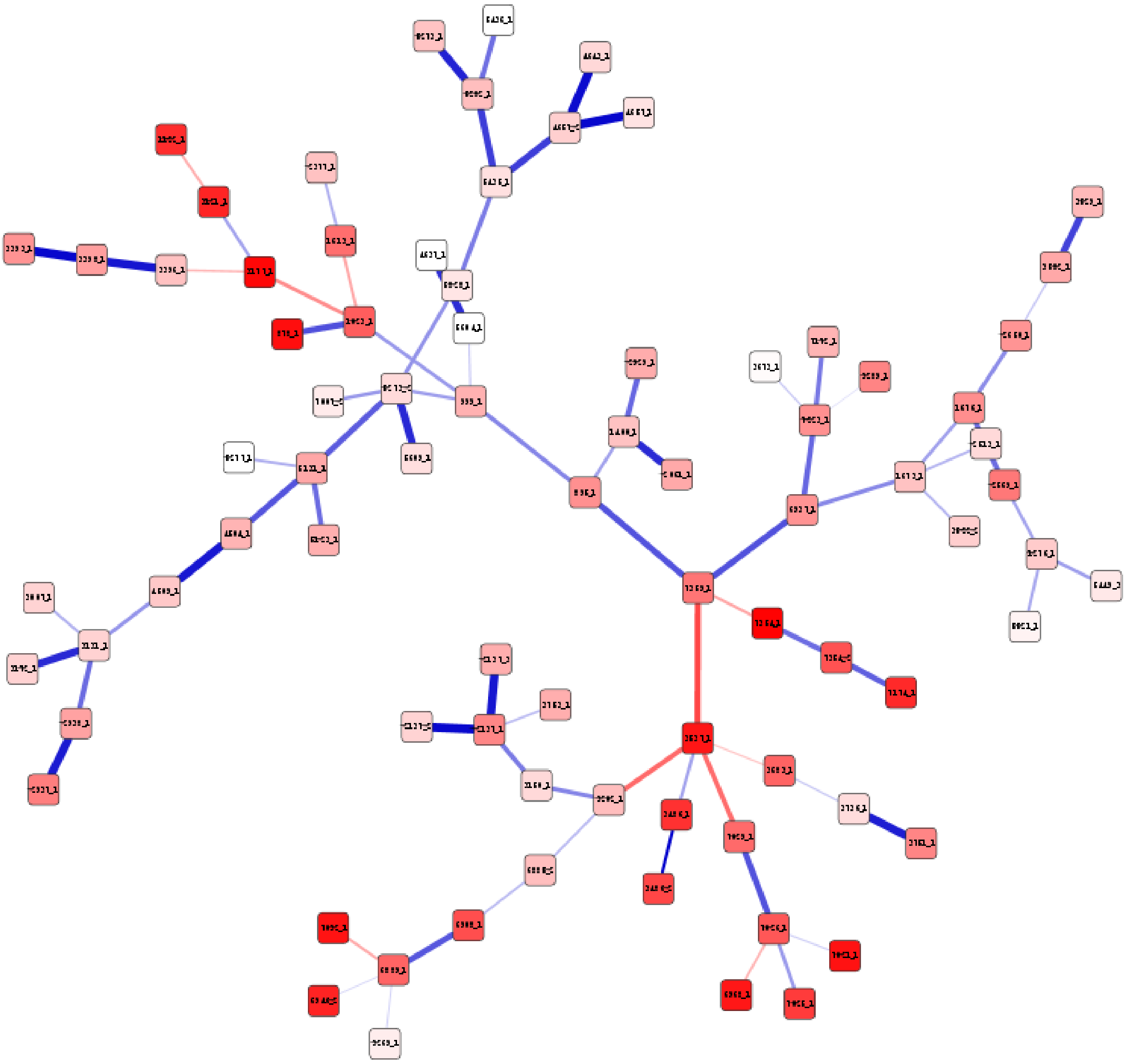


Figure 3.29: Maximum spanning tree derived from full feature matrix. The colour of the nodes represents fold change; the line colour the Pearson correlation. The layout is described in section 3.3.3.1

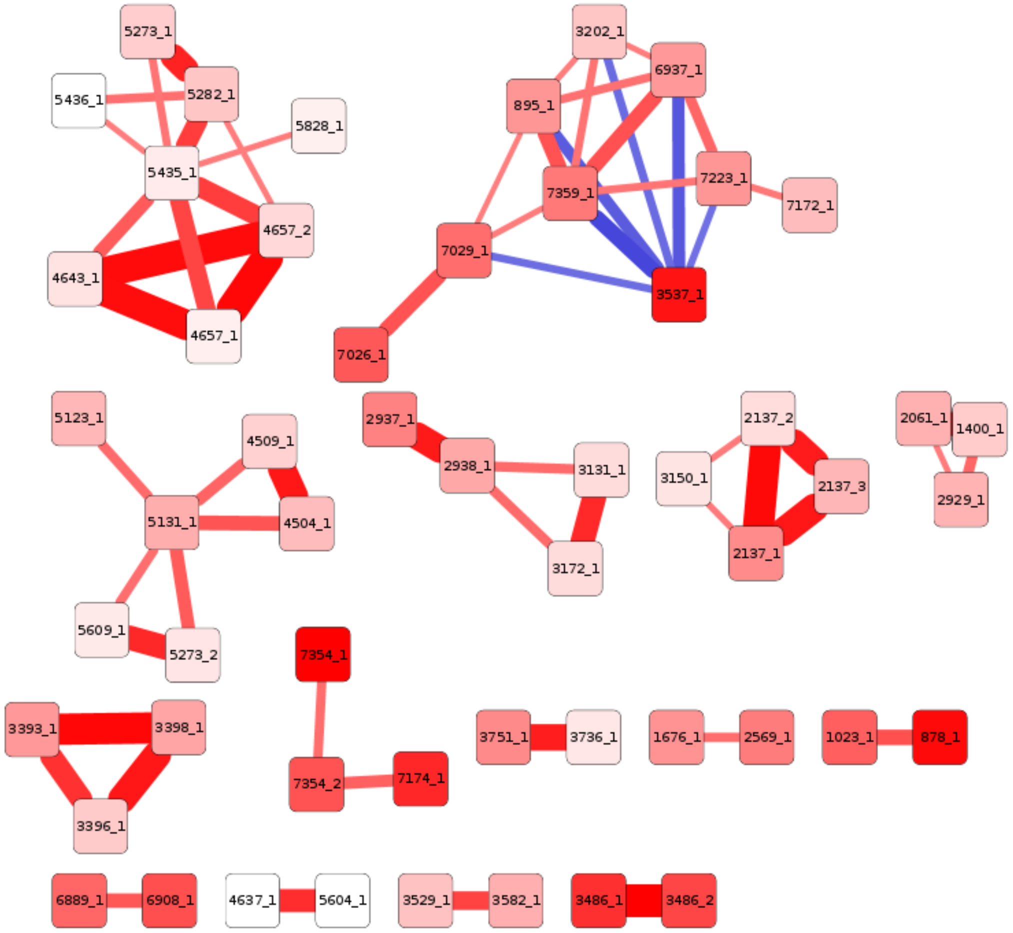


Figure 3.30: Correlation network with threshold of 0.5, derived from full feature matrix

a single experiment should therefore shed a better light on the kinetic relationships between compounds. To this end a network was constructed for the 'centroid' experiment. Similar graphs constructed with threshold-correlation and maximum spanning trees are shown in figures 3.31 and 3.32 below.

The amount of 'significant' correlations between putative compounds is much higher for this type of localised network, evidenced by its high level of connectedness in figure 3.31. Additionally, high correlations are found between compounds further apart in retention time. This could simply indicate that the possibly spurious correlations between similar features found in figure 3.30 are not as present in a single experiment. Due to the fact that the only perturbed variable in this network is time, the correlations between the putative compounds should be strictly kinetic; whereas for the entire feature matrix, feature correlation across other experimental perturbations have a compound contribution.

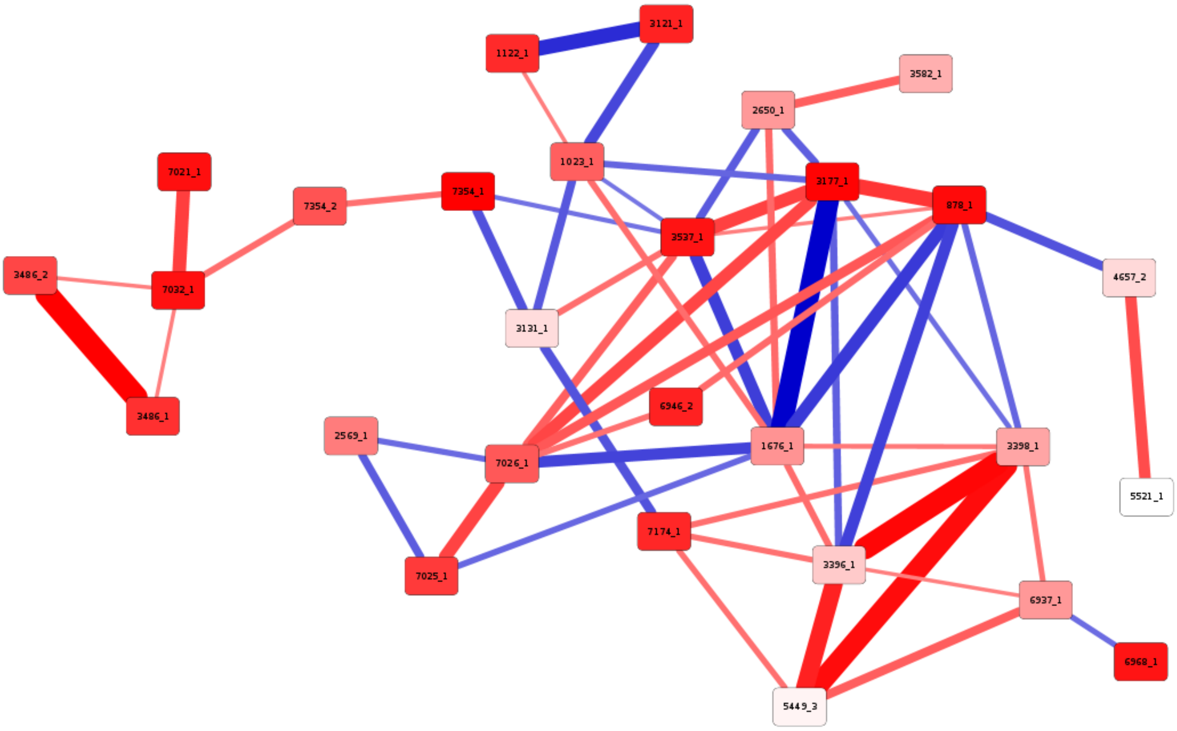


Figure 3.31: Correlation network with threshold of 0.5, derived from a single experiment

If one considers the loading plot for the same experiment in figure 3.21, where time appears to be reflected in the first principal component, the significant features have interesting properties in figures 3.31 and 3.32. At the ends of the range in PCA 1 are features 3177_1, 3537_1 and 878_1 to the left of the axis; 1676_1 to the right. Presumably these are some of the features responsible for the most variance over the passage of time. In figure 3.31, these are 'hub' nodes at the centre of the network, with respectively high degrees. Interestingly, node 1676_1 has a high apparent importance despite its relatively low concentration. It is highly negatively correlated with 3177_1; this pair also has the highest distance across PCA 1 as seen in figure 3.21.

This correlation between 3177_1 and 1676_1 is shown in a simple plot in figure 3.33. The nature of the negative correlation is immediately apparent. Feature 1676_1 is only present after day 15, and appears to be completely dependent on the presence of feature 3177_1 thereafter.

There are several cases, evident in both networks and PCA loading plots, where two features are very tightly coupled. It is not impossible that the methods used to create the feature maps would report the same feature as separate entities. While this may expose flaws in the method, it does not greatly detract from the original purpose of the investigation, which is to create a platform for hypothesis generation from which further investigation

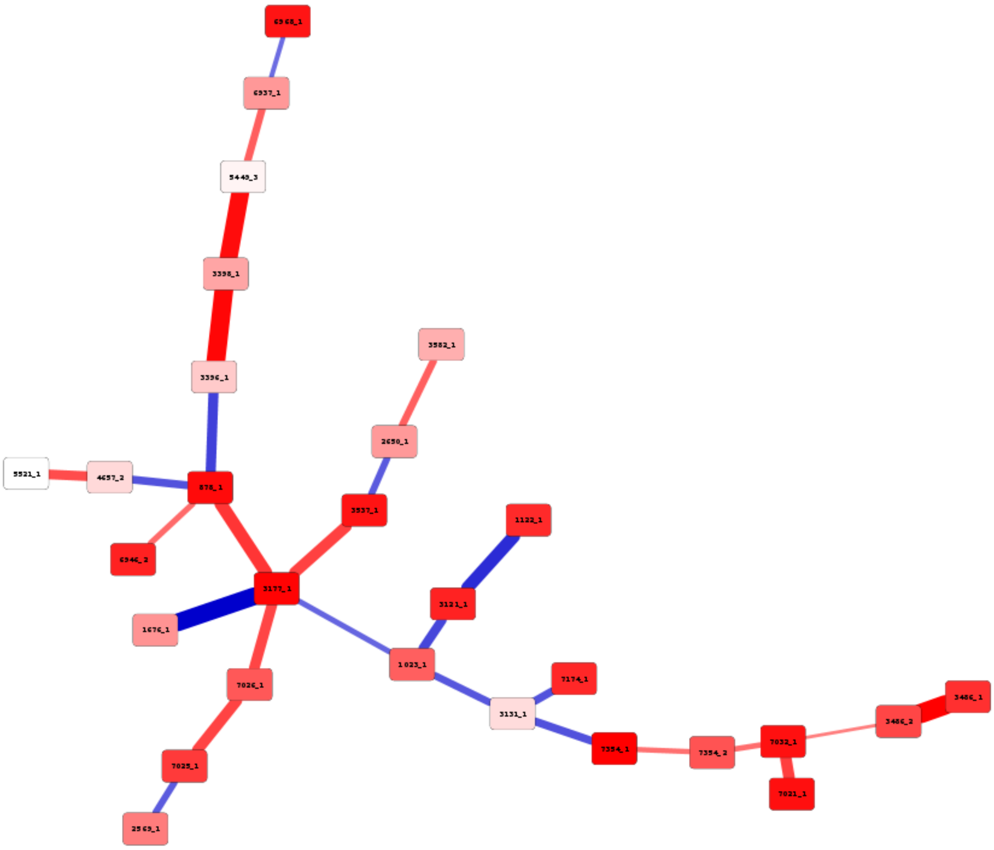


Figure 3.32: Maximum spanning tree of single experiment

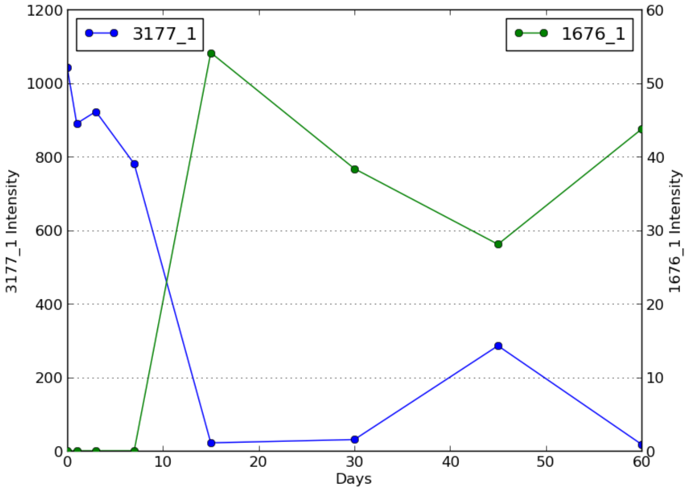


Figure 3.33: Comparison of negatively correlated features in a single experiment

into putative compounds can be conducted.

A further interesting phenomenon when linking PCA with networks, is that the two compounds at the far ends of the second principal component - 3486_1/2 and 6937_1 - are respectively leaf and leaf-adjacent nodes at the opposite ends of the maximum spanning tree. As the only known latent variable across this experiment is time, it is difficult to ascribe a second; however the dataset does include experimental replicates at each time point. Variance between replicates could be explained by this principal component. If the maximum spanning tree is reflective of the kinetic evolution of the media over time, then the leaf nodes could represent final products the experiment after the experiments conclusion. These types of compounds are naturally the most likely to be different between experiments, as they are the final result of a high number of pre-requisite reactions; decreasing their combinatorial probability. Indeed, feature 6937_1 is only present in one of the experiments at day 60, appearing as a relatively low-intensity peak.

3.4 Conclusion

A possible solution to the original problem - to reduce an enormous and complex data-set to a interpretable and meaningful form from which to conduct further investigation, has been presented. The nature of the analysis was untargeted and blind to pre-determined compounds of interest.

Much of the energy - both cognitive and computational, is invested in the mass preprocessing of this data in order that it might be mined for new hypotheses. It can be argued that the preprocessing was largely successful, preparing the data for feature map alignment. An unconventional approach of 'TAC' alignment using a modified reversed COW interpolation was used; this allowed for a more conserved feature alignment between experiments.

A method for feature map alignment of HPLC-UV/vis data was developed in the absence of any existing software. The algorithm is based on a dynamic feature database or 'master list' approach; comparison between features was done on a linear basis. The signature of a feature across the UV-spectra was exploited in the comparison of features from different experiments. It was found that the use of simple correlation was sufficient to infer congruency between these features.

While there is certainly room for improvement of the algorithm, it served its purpose in the generation of a comparable list of features for each experiment upon which data exploration techniques could be applied. It was found by various means of validation that the resulting peak matrix had statistical significance with regards to various experimental conditions to which the wine media was subjected. Decision tree models used to predict experimental conditions based on the feature matrix were highly effective; furthermore, permutation tests performed on the experiment labels revealed that these correlations were far from random.

Both decision trees and PCA were instrumental in identifying important putative compounds with regards to experimental conditions; however relationships between the putative compounds were explored more effectively using network models based on linear correlation. Networks constructed on data where time was the only perturbed variable were potentially instrumental in the mapping of the kinetics of reactions occurring in the media between putative compounds.

It was, however, found that networks focusing on a smaller scale reveal some tightly coupled features that could be repeated reports of the same putative compound. While these occurances are not ideal, the aim of this type of analysis is not to achieve a perfect and conserved accuracy for the reporting of results; rather to generate new hypothesis and focus further investigation into compounds of interest. Additionally, the occurance of these possible repeats is small relative to the number of putative compounds.

This is symptomatic of the fact that there are still many potential streams for improvement in the analysis pipeline presented here. Unfortunately, as mentioned in the literature review, there are no standard feature map alignment procedures for this particular type of data with which to compare the present outcome. An exhaustive investigation into the translation of the MS feature map alignment algorithms to UV-vis would be a large undertaking; further work into this type of analysis would be hugely beneficial as it remains a prevalent method in analytical chemistry. Many of the algorithms reviewed in for feature map alignment ((Zhang *et al.*, 2005), (Smith *et al.*, 2006), (Katajamaa *et al.*, 2006), (Bellew *et al.*, 2006)) have the potential to be applied, and could possibly be integrated into the work-flow developed in this study; however an undertaking of this nature would be outside the present scope.

While imperfect in its application, it is proposed that this kind of untargeted mass-analysis, alignment and feature comparison will be useful in generating hypotheses regarding heretofore disregarded putative compounds; in discovering new kinetic relationships which would otherwise have gone unnoticed, and to explore global trends in significant compounds related to experimental perturbations.

3.5 Acknowlegments

I would like to acknowledge most especially the researchers responsible for providing the excellent model data on which this analysis is based: Astrid Buica, as well as her supervisor Wessel du Toit. Dr. Buica also performed all HPLC/UV-vis measurements. This project would also not have been possible without the help and support of the students and staff of the IWBT, and the advice and guidance of the Computational Biology group within the institute.

3.6 List of References

Bellew, M., Coram, M., Fitzgibbon, M., Igra, M., Randolph, T., Wang, P., May, D., Eng, J., Fang, R., Lin, C., Chen, J., Goodlett, D., Whiteaker, J., Paulovich, A. and McIntosh, M. (2006). A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics*, vol. 22, no. 15, pp. 1902–1909.
Available at: `http://bioinformatics.oxfordjournals.org`

Du, P., Kibbe, W.a. and Lin, S.M. (2006 September). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics (Oxford, England)*, vol. 22, no. 17, pp. 2059–65. ISSN 1367-4811.
Available at: `http://www.ncbi.nlm.nih.gov/pubmed/16820428`

Ellson, J., Gansner, E.R., Koutsofios, E., North, S.C. and Woodhull, G. (2003). Graphviz and dynagraph â static and dynamic graph drawing tools. In: *GRAPH DRAWING SOFTWARE*, pp. 127–148. Springer-Verlag.

Ferreira, A.S. (2002). Kinetics of oxidative degradation of white wines and how they are affected by selected technological parameters. *Journal of Agricultural ...*, pp. 5919–5924.
Available at: `http://pubs.acs.org/doi/abs/10.1021/jf0115847`

Graciliao, M. (2011). Statistics::R.
Available at: `http://search.cpan.org/ gmpassos/`

Hagberg, A., Swart, P. and Chult, D. (2008). Exploring network structure, dynamics, and function using NetworkX. , no. SciPy, pp. 11–15.
Available at: `http://www.osti.gov`

Hunter, J.D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, pp. 90–95.

Jacobson, D., Monforte, A.R. and Ferreira, A.C.S. (2013 March). Untangling the Chemistry of Port Wine Aging with the Use of GC-FID, Multivariate Statistics, and Network Reconstruction. *Journal of agricultural and food chemistry*. ISSN 1520-5118.

Katajamaa, M., Miettinen, J. and Orešič, M. (2006). MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics*, vol. 22, no. 5, pp. 634–636.
Available at: `http://bioinformatics.oxfordjournals.org`

Lange, E. and Tautenhahn, R. (2008 January). Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics*, vol. 9, p. 375. ISSN 1471-2105.
Available at: `http://www.pubmedcentral.nih.gov`

McKinney, W. (2012). pandas: A Python Data Analysis Library.
Available at: `pandas.pydata.org`

Nielsen, N.-P.V., Carstensen, J.M. and Smedsgaard, J.r. (1998 May). Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A*, vol. 805, no. 1-2, pp. 17–35. ISSN 00219673.
Available at: `http://linkinghub.elsevier.com/`

Okamoto, M. (2002 January). Reversal of elution order during the chiral separation in high performance liquid chromatography. *Journal of Pharmaceutical and Biomedical Analysis*, vol. 27, no. 3-4, pp. 401–7. ISSN 0731-7085.
Available at: `http://www.ncbi.nlm.nih.gov/pubmed/11755741`

Oliphant, T.E. (2007). Python for Scientific Computing. *Computing in Science & Engineering*, vol. 9, no. 3.

Pedregosa, F. and Varoquaux, G. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830.
Available at: `http://dl.acm.org/citation.cfm?id=2078195`

Rokach, L. and Maimon, O. (2005). Top-Down Induction of Decision Trees Classifiers - A Survey. vol. 35, no. 4, pp. 476–487.

Savitzky, A. and Golay, M. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003 November). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, vol. 13, no. 11, pp. 2498–504. ISSN 1088-9051.
Available at: `http://www.pubmedcentral.nih.gov`

Simpson, R. (1982). Factors affecting oxidative browning of white wine. *Vitis*.
Available at: `http://www.vitis-vea.de/admin/volltext/e020660.pdf`

Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R. and Siuzdak, G. (2006). XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Analytical Chemistry*, vol. 78, no. 3, pp. 779–787.
Available at: `http://pubs.acs.org/doi/abs/10.1021/ac051437y`

Tomasi, G., van den Berg, F. and Andersson, C. (2004 May). Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics*, vol. 18, no. 5, pp. 231–241. ISSN 0886-9383.
Available at: `http://doi.wiley.com/10.1002/cem.859`

Zhang, D., Huang, X., Regnier, F.E. and Zhang, M. (2008 April). Two-dimensional correlation optimized warping algorithm for aligning GC x GC-MS data. *Analytical chemistry*, vol. 80, no. 8, pp. 2664–71. ISSN 1520-6882.
Available at: `http://www.ncbi.nlm.nih.gov/pubmed/18351753`

Zhang, X., Asara, J.M., Adamec, J., Ouzzani, M. and Elmagarmid, A.K. (2005). Data pre-processing in liquid chromatographyâmass spectrometry-based proteomics. *Bioinformatics*, vol. 21, no. 21, pp. 4054–4059.
Available at: `http://bioinformatics.oxfordjournals.org`

Zhang, Z.-M., Chen, S. and Liang, Y.-Z. (2011 January). Peak alignment using wavelet pattern matching and differential evolution. *Talanta*, vol. 83, no. 4, pp. 1108–17. ISSN 1873-3573.
Available at: `http://www.ncbi.nlm.nih.gov/pubmed/21215845`

Chapter 4

Network Visualisation

Extensible methods of visually exploring experimental data with networks were sought. Three different scientific experiments were individually modelled to develop general and specialised instances of these methods.

In order to pare down the information presented in a network, statistical tests were used to gauge significance. Significant relationships were then mapped to networks, along with quantification of results. Networks assumed different topological formats depending on the nature of the data. Loading the networks into Cytoscape allowed for interactive viewing.

It was found that these networks allowed for advanced queries and hypothesis generation. The diversity of the data on which the methods were used suggests that they are broadly applicable.

4.1 Introduction

The visual presentation of data has evolved over time to accommodate an increasing demand for exploratory data analysis. Generation of new hypotheses from innovative representations is instrumental if one desires to fully utilise data at hand.

In particular, data generated from scientific experiments benefit from comprehensive and interpretable visualisations. This is both because the generation of new hypotheses from targeted analyses constitutes a 'free score'; also the cost of performing a scientific study is usually high in materials, facilities and labour.

While the classical techniques of scientific visualisation - line and scatter plots, boxplots and surface responses - are exhaustive, they can also stymie the researcher through an overload of information. Tracking global relationships in this way is also difficult for the human mind. If the data is time-dependent, the complexity of its interpretation is severely increased.

Network visualisation is traditionally applied to extant data in order to map relationships between objects (Herman *et al.*, 2000). For scientific visualisation of generated data, the most common instance is in the form of decision trees. These are typical for classification and regression solutions in applications with many predictor variables; for data generated from small-scale experiments it is generally not appropriate.

Given this status in the visualisation of scientific data, this study aims to develop extensible methods of scientific visualisation using networks; dubbed 'StatNet' in lieu of any existing description. Networks have the potential to provide an intuitive framework for inference and hypothesis generation, as the human mind has a natural aptitude for topography. This property should be exploited to maximise the usefulness of scientific data.

The developed methods should embrace and expound the ideals of exploratory data analysis, as well as fulfil the requirements for clear data analysis as described in Kelleher and Wagener (2011). The beneficial properties of interactive visualisation as listed by Keim (2002) also serve as targets of utility.

4.2 Methodology

The methodology of the developed visualisation technique generally follows similar patterns of data collection; storing; statistical testing and network generation. Three different types of data were processed in this manner from three different experiments; these are briefly outlined in the section below.

4.2.1 Experimental Data

The first data set on which network methods were applied was drawn from an experiment on aroma in white wine. Several known compounds related to aroma in white wine were added to a model wine media. These were namely 3-Mercaptohexanol (3MH); 2-Methoxy-3-Isobutylpyrazine (IBMP); Methional (Meth) and Phenylactaldehyde (Phen). They were added in five respective concentrations, corresponding to empirical sensorial threshold levels found in literature (for example, level 1 is at the perception threshold, 3 at normal levels in wine and 5 at the most extreme).

The aim of the experiment was to observe what the effects of these concentrations are on both the intensity of certain aromas, as well as the aroma profile as a whole. A further aim was to observe the aromatic relationships between these compounds - for example, whether some compounds may have additive or suppressive effects on the sensorial effects of others.

To this end the aroma compounds were added in two different ways: firstly, by 'spiking' the wine with the compounds separately - adding each to all of its five levels in the absence of any other aroma compounds. Secondly, the compounds were added in different combinations of concentration with each other. Naturally, the potential combinatorial subspace for this type of analysis is huge; therefore, a central composite design was chosen. This choice is also aligned with the aims of the experiment as it is a minimal design that still retains the ability to test for interactions between experimental variables.

The measurement of the aroma profiles and intensities was done by a panel of judges. Every precaution was taken to ensure that the judgment was unbiased. A standardised list of aroma descriptors was used between all judges for both the spiking and composite design experiments - for example 'earthy' and 'grassy'. The judges scored the intensity of each of these descriptors using a 100 mm unstructured line scale - 'quantitative descriptive analysis' as prescribed by sensory literature. Each test was performed in triplicate.

The second data set was the most straightforward. It was derived from an experiment assessing the resilience of various cultivars of *Vitis vinefera* (grape vines) against fungal infection in the form of *Botrytis cinerea*. Each strain, represented by a single plant, was infected with the fungus on four of its leaves with four targeted infection points on each leaf.

The level of infection was then tracked and recorded for each leaf at set intervals. In some cases, the infection was so rapid that a leaf was completely destroyed before the experiment's conclusion.

The last data set on which the method was developed was from an experiment on the browning in white wine. This is the same experiment as that analysed in Chapter 3. In summary: various conditions were perturbed in samples of white wine media in order to ascertain what the effects are on browning and oxidation.

The experimental design was factorial, with the result that there were approximately one thousand samples in total (including replicates). The perturbed experimental conditions included slow- and rapid oxidation (open or sealed wine bottles) and several levels of pH and SO_2. In addition, samples were subjected to one of three phenolic treatments - the addition of catechin, caffeic acid or both. The measured outputs for each of the samples were concentrations of dissolved O_2 and SO_2 as well as the absorbance of the samples at 420 nm (a proxy for the overall effect of browning).

4.2.2 Network creation

In each case, a custom script was created in either perl or python in order to render a particular type of network. Each data set originated from csv files of differing formats, necessitating flexible parsers to be written. The data was coerced into a format amenable for statistical testing - mostly in the form of vectors attached to tiered reference structures in order of compared attribute. In general, replicates were collected into separate vectors if time-independent; for temporal data a number of vector combinations were tested. These vectors were then compared either to all other vectors, or to a single centroid or reference vector using the chosen statistical test or metric. The data sets analysed, being derived from fundamental experiments, were quite small and therefore computational intensity was not a concern.

Relationships between vectors that were deemed statistically significant according to the pre-determined threshold were then stored in a hash or reference structure. Statistical tests were performed with either using some of the statistic modules in perl, or scipy in python (Oliphant, 2007). Attributes of the vectors themselves, or the tested relationship, are stored separately in a similar manner. These hash structures are then the basis for the final network.

The network is then written to a compatible format in order to be visualised in the chosen software. In this case, Cytoscape (Shannon *et al.*, 2003) was used for visualisation as it provides many of the interactive features mentioned in 2.2.5.

4.3 Results and Discussion

4.3.1 Sensory Data

The sensory data used is time-independent and split into two different experiments: orthogonal (aroma 'spiking') and central-composite (aroma interaction studies), as mentioned in Section 4.2.1.

4.3.1.1 Correlation Network

As the first type of experiment is relatively straightforward, a simple network was chosen to represent it: the familiar correlation network. The question surrounding the orthogonal study was essentially which - and to what degree - aromatic properties are effected by particular compounds. The interaction between compounds cannot be mapped as they were individually spiked; likewise, the relationships between aromas are not of particular interest.

A simple Pearson correlation was therefore calculated between the levels (1 through 5) of the aromatic compounds and the aromatic properties recorded at each respective level. The intensity of each aromatic descriptor was averaged across all panelists. The generated network is shown in Figure 4.1. Aromatic compounds and properties are depicted together and form the nodes of the network; correlations the edges between them. The width of the edges is mapped to the strength of the correlation and the colour its sign (red for positive, blue for negative).

Due to the fact that in the absence of some of the aromatic compounds the requisite score for a sensorial property was zero, and given the resultant simplicity of the network, there was no need to impose a correlation threshold. The aromatic profiles are first mapped to the network, followed by their correlation (should it exist) to the aromatic compounds. The linkage of compounds through common sensory profiles is therefore incidental.

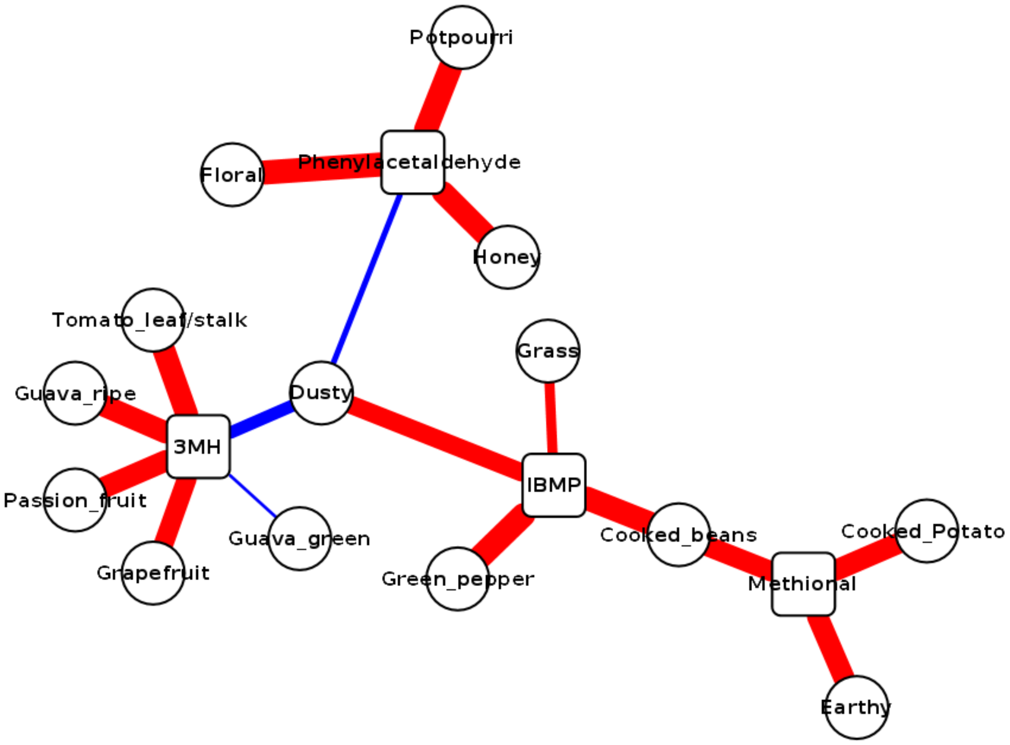


Figure 4.1: Correlation network for orthogonal sensory data. Elliptical nodes represent aroma descriptors; square nodes aromatic compounds. The thickness of the line indicates the magnitude of the correlation and the colour the direction: red for a positive correlation and blue for a negative.

The representation of the data in this way is beneficial both for the identification of the direct aromatic effects of the added compounds, and the linkages of shared aroma profiles between compounds. The 'dusty' descriptor, for example, has a shared correlation (albeit sometimes negative) with three different compounds, a finding possibly significant in the composite design study.

It is proposed that this type of representation is more concise and informative than typical presentations of this data in tables, line plots and bar graphs. The relationships between compounds and descriptors are easy and clear to infer; the incidental linkages between compounds immediately apparent.

4.3.1.2 Central Composite Network

Due to the relative complexity of the central composite experiment, a more exhaustive network representation was developed, shown in Figure 4.2. For the central composite design, there were three reference vectors for which all four of the compounds were at the same level (2, 3 and 4). Three subgraphs were

Table 4.1: Aroma compound index for reference levels. For instance the reference level '1-2-3-4' refers to levels in Meth, Phen, 3MH and IBMP respectively

Aroma Compound	3MH (ng/L)	IBMP (ng/L)	Meth (ug/L)	Phen (ug/L)
Index	1	2	3	4
Level 1	40.0	1.0	0.3	0.5
Level 2	60.0	2.0	0.5	1.0
Level 3	500.0	10.0	3.0	30.0
Level 4	2000.0	20.0	6.0	15.0
Level 5	6000.0	40.0	15.0	130.0

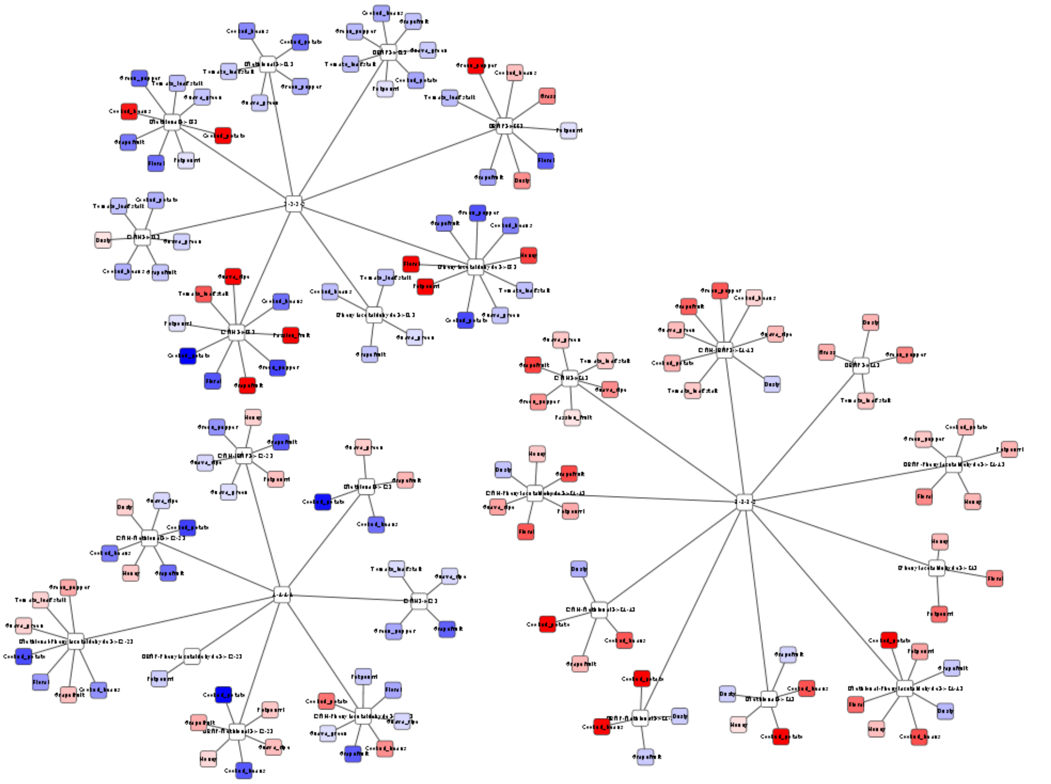


Figure 4.2: The complete network for the central composite experiment. Each subgraph represents a statistical reference point. The colour intensity of the node represents fold-change from the reference levels: red for positive; blue for negative fold-change. A more focused figure is presented in Figure 4.3

therefore made with each of these references. Note that the labelling of the aroma compounds follows the table given in table 4.1.

Any deviations from these reference points, whereby one- or two compounds were raised or lowered from the reference level, are subjected to a statistical test to ascertain how the aroma profile differs. For example (refer to Figure 4.3), from the reference point of '2-2-2-2' - all compounds at a level just below the average commonly found in wine - if 3MH and IBMP are both raised to level four ('2-2-4-4') then several changes in the aromatic properties of the wine are observed. Each of the aroma descriptor vectors for the conditions at '2-2-4-4' are statistically tested against the requisite descriptor vector at reference '2-2-2-2'. If the difference between these two vectors are deemed to be significant, the descriptor is included in the network.

For this network, a T-Test was used to gauge significance. If the null hypothesis is rejected at a significance level of 0.05, the node is included in the final network (connected to the reference level with which it was tested). The fold-change of the descriptor between the reference point and the tested sample is then used to colour the node in relative intensity (red for a positive- and blue for a negative change as is the convention).

In this way, only salient changes from the reference points are included in the final network, which allows for a scalable level of complexity. The individual and combined responses to the level of each compound are included, consummate with the original aim of a central composite design.

Figure 4.3 is a subgraph created from selecting a subset of nodes and edges from the original graph (a task easily accomplished using Cytoscape). From a brief overview, several conclusions regarding aromatic responses can be drawn. Firstly, their combination seems to have a compound effect on the perception of green pepper (a difference in sensory scores between 11 and 13 respectively, and 18 in combination). Secondly - and more interestingly - IBMP when raised on its own increases the dusty aroma; however, when combined with 3MH this effect is reversed, decreasing its intensity. These types of conclusions are easily drawn if the overall network is sensibly divided in this way.

The interactive features of this visualisation are aligned with many of those mentioned by Keim (2002): in particular, the freedom to partition the data and zoom in on logical subsets. Hypothesis generation can be aided by these abilities.

As an exercise in validation, the networks are compared to the results of principal component analysis on the same data. The PCA was performed on the central composite component of the experiment only, and the loadings are depicted in Figure 4.4. The clustering of the aromatic compounds with the descriptors mirrors the relationships mapped out in Figure 4.1 closely. This is both a confirmation of the integrity of the data - the orthogonal and central composite experiments matching - and the exhaustive ability of the network method to describe the aromatic effects.

4.3.2 *Botrytis* Infection

4.3.2.1 All-against-all Network

Three simple networks were used to describe the *Botrytis* infection experiment. It was difficult to conceptualise time in a comparative network, thus for the first instance, the infection rate was compared at a single time point near the end

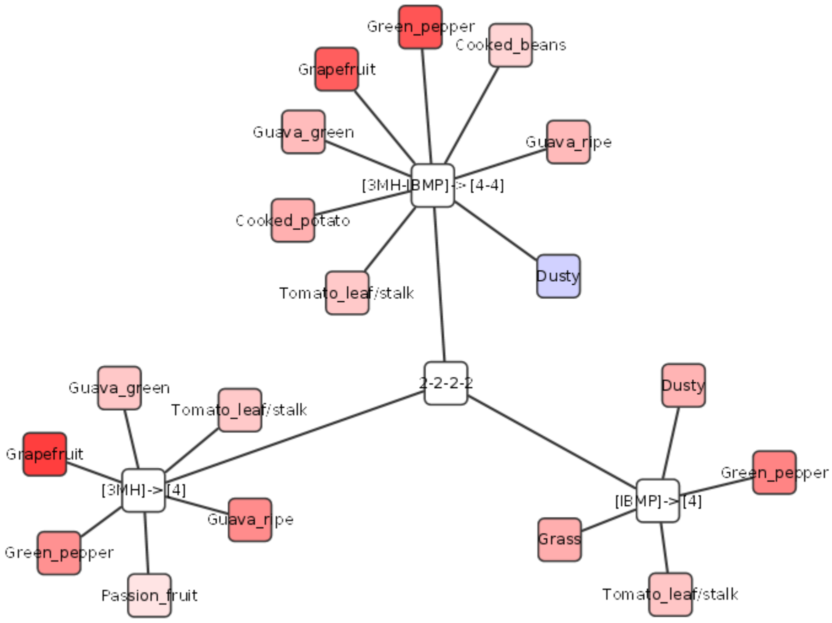


Figure 4.3: A focused view of the interaction network, from the reference level at slightly below the average found in wine. The node colourings follow the same rules as for Figure 4.2

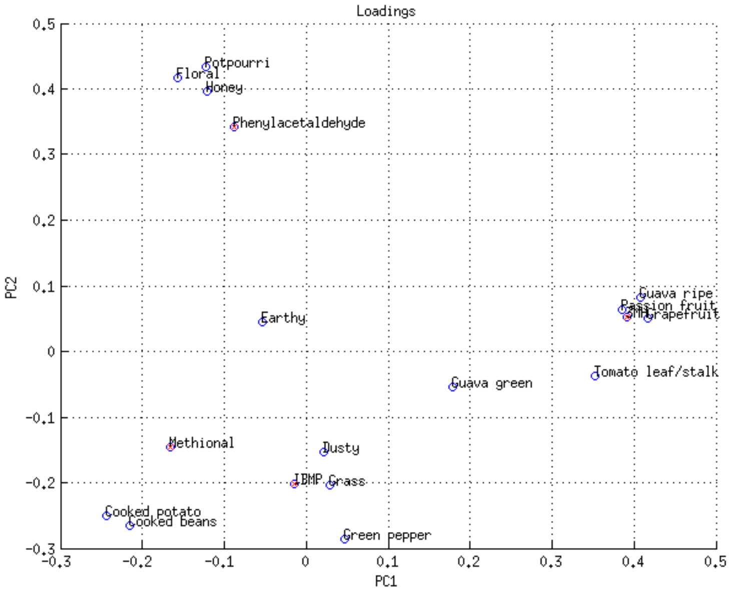


Figure 4.4: PCA loadings of the central composite experiment. Both the aromatic compounds and the sensory descriptors are included.

of the experiment in an all-against-all manner (see Figure 4.5). Note that this figure details the interactions between individual leaves in the experiment, so that the naming convention is '<cultivar>_<leaf>'. Each leaf was compared against all other leaves using a T-Test to assess the difference in means. The vectors for comparison were built from the infection levels at the four infection points on each leaf. Only significant edges were mapped to the network. Edges are coloured according to the difference in infection between the nodes; the comparison is read in the direction of the arrow.

The information from an all-vs-all type of comparison can be overly convoluted. It is possible to create sub-graphs from nearest-neighbor selection; however the most value from this network view is to identify the 'hub' leaves - those with the most significant differences globally. To this end, a degree-sorted circular layout was applied to Figure 4.5, such that the node with highest degree (ControlA_4) is placed at the bottom with successively significant nodes arranged anti-clockwise.

4.3.2.2 ET50 Network

A second network was constructed with time as a focal object. As the central question of the experiment is the rapidity with which the leaves of different cultivars of *Vitis vinifera* succumb to infection, time was quantified for each infection point. The time at which the infection reached half of its final value ('ET50') was used as an approximation for the rapidity of infection. If the ET50's for each infection point on a leaf are combined into a vector and statistically tested against each other, a new network can be generated (Figure 4.6). As the ET50 values varied much less across cultivars and leaves, the network is much simpler. This is presented as a simple way of gauging the relative infection rates.

4.3.2.3 Time-centric Network

The entire strain may also be taken as a vector, including all leaves and infection points. A network was generated to observe the differences in *Vitis vinifera* cultivars on each day of the experiment. A subgraph of this is shown in Figure 4.7 for day 4 of the experiment. The difference in the infection of strain 14A on this particular day is clearly indicated.

4.3.3 Browning Experiment

For the browning experiment, measurements were taken in triplicate for the oxygen and browning levels on set days for each combination of parameters in the factorial design. Therefore, if one experiment is compared to another at a single point in time, small vectors of length three are subjected to statistical

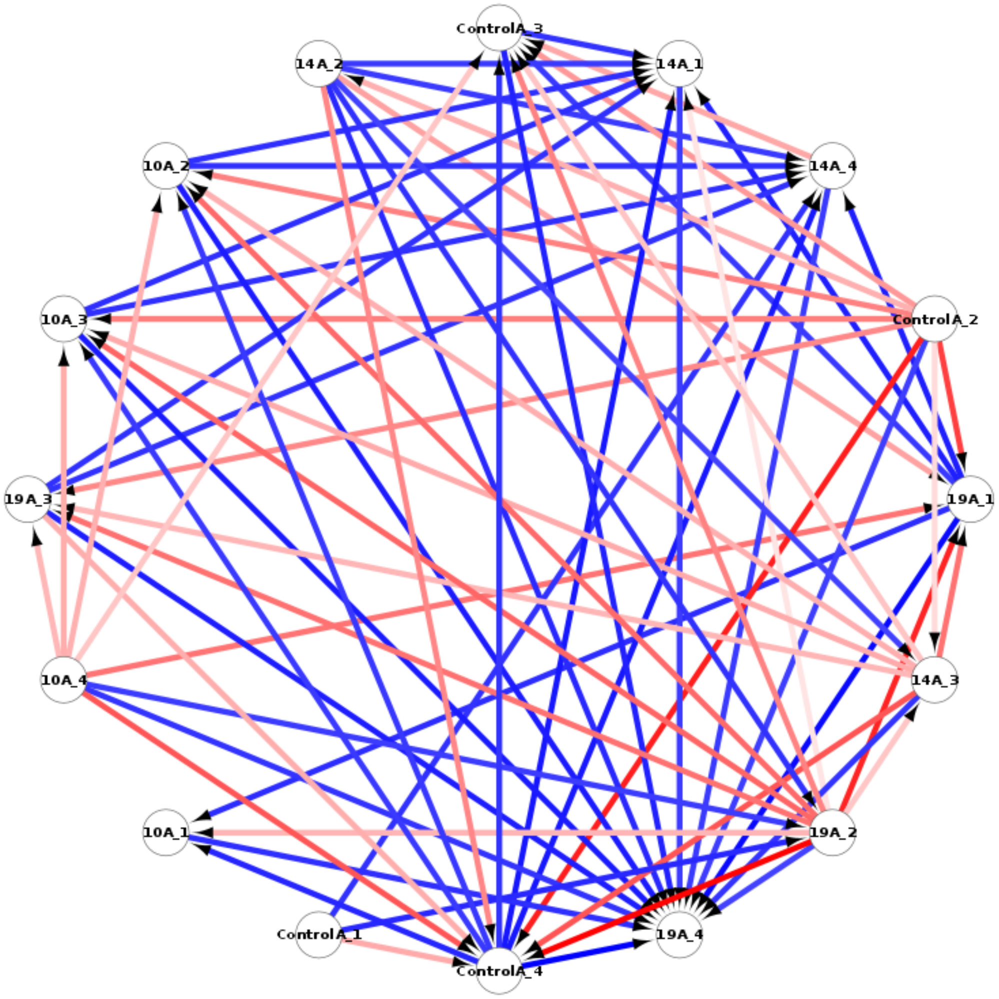


Figure 4.5: An all-against-all comparison of the infection rate near the termination of the experiment. The nodes represent individual leaves, and the interactions between the nodes significant interactions. The interactions are coloured according to the degree of correlation, scaled to intensity: red for positive, and blue for negative.

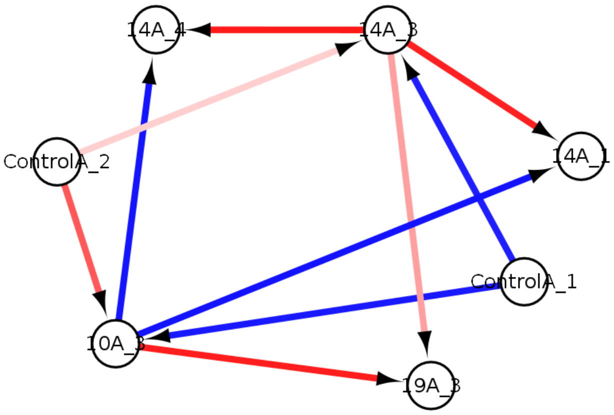


Figure 4.6: All-against-all ET50 metric for infection rate. The naming convention for the leaves as well as the colouring of the interactions follow the same pattern as in 4.5.

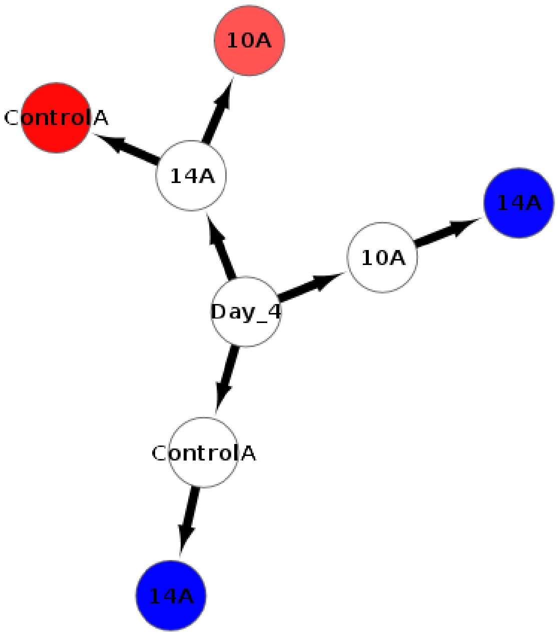


Figure 4.7: Subgraph of strain infection on day four of the experiment. The three nodes attached to the central node are individual leaves; the outer nodes are the leaves with which there is a significant difference for that particular day. The intensity of the node is coloured by the absolute difference: red for positive and blue for negative.

testing. With this in mind, the Wilcoxon test was most frequently used as it is difficult to assume a distribution with such low statistical power.

Due to the fact that the experiment was factorial there were a large number of comparisons to be made at any time point. This has the potential to increase the family-wise error rate (FWER), an issue often encountered in analyses with high numbers of hypothesis testing (Holm, 1979). A Holm-Bonferroni correction for FWER was therefore consistently applied.

In contrast to the all-against-all approach used in Section 4.3.2, a central set of experimental variables was selected as a reference against which all other combinations were tested. This was a natural selection given the nature of the experiment: the combination of variables most often associated with normal levels in wine were chosen (pH of 3.6 and SO_2 levels of 25 ppm). The combined phenolic treatment was chosen, against which the absence of one of the two compounds could be compared.

4.3.3.1 Time-centric Network

The first type of network generated focused on time as the central variable. Separate subgraphs were created for each point in time, as the data was measured at discrete time points of 0, 1, 3, 7, 15 and 30 days after initialisation. Samples at each combination of conditions were compared to the reference sample using the abovementioned test and correction, and if deemed significant was added to the network, the first portion of which is shown in Figure 4.8.

The leaves of these time subgraphs are the fold-change of the measured outputs (browning and O_2 level) as compared to the samples at the reference conditions at that time point.

As is clear from the above, the deviations of the outputs from the reference values grow more numerous as time increases; almost nothing on day '0', but plentiful by day 3. If one performs a search-and-select for a particular subset of samples (for example, Catechin-only treatment with no SO_2 but including all variations in pH); perform a nearest-neighbor selection and create a new network, a fairly powerful query is executed. Figure 4.9 refers.

One can quickly infer from this view that, for this treatment and SO_2 level, the consumption of O_2 is lower over time (except in the case of pH 7.2), and that pH 7.2 samples tended to quicken the browning process up to day 15, beyond which point it is no longer significantly different from the reference value. These type of global and local trends are intuitively displayed across the time subgraphs.

4.3.3.2 Tiered Variable Network

If one desires to eliminate the variance of one of the variables, then it is possible to create networks centered on that variable, while holding all others constant

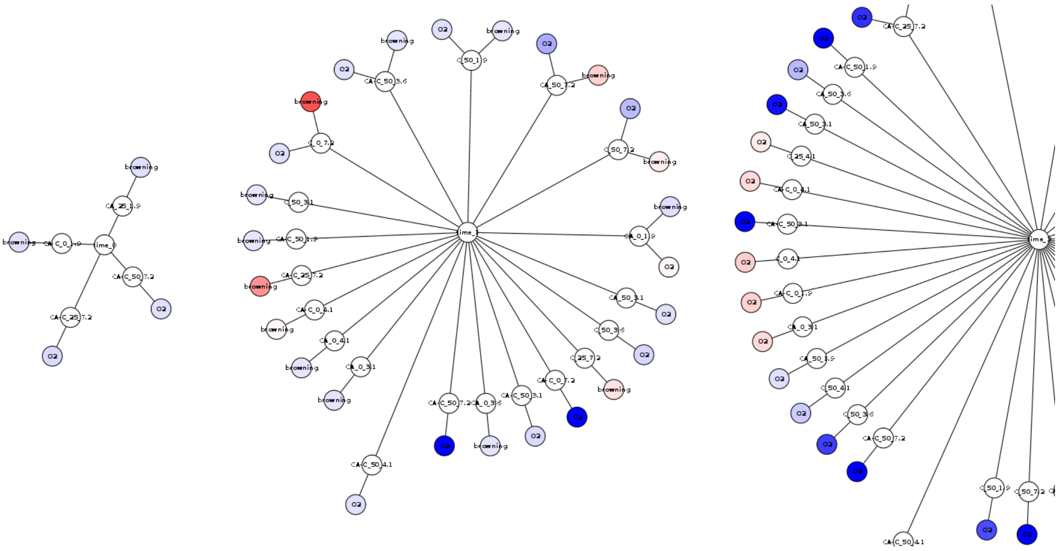


Figure 4.8: An overview of the time network generated for browning data. The fold change of the measurements from the reference levels on that particular day are depicted by the colour intensity: red for positive, blue for negative.

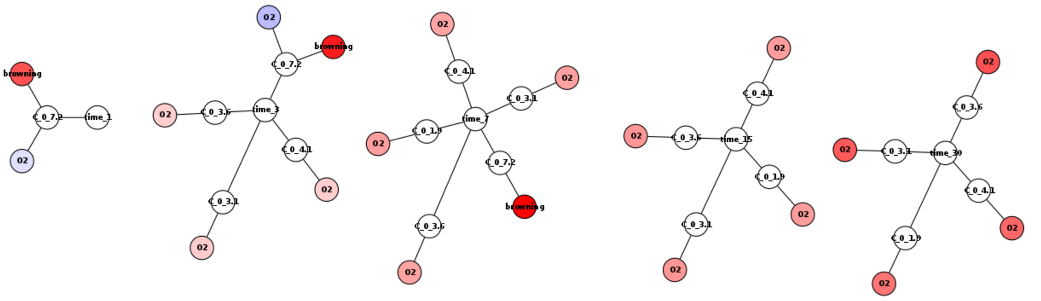


Figure 4.9: Focused subset of conditions over time subgraphs in 4.8.

as a reference. A 'tiered' network is demonstrated to this end, with each condition branching off from the center, terminating in the measured outputs of browning level and O_2 for each time point. An example of this approach is shown in Figure 4.10, where separate subgraphs are created for each of the three treatment types.

Varying levels for pH, then SO_2 form subsequent branches from the constant treatment at the center (the order is variable and can be altered for interpretability). The final variable of time is then compared to the reference sample for pH and SO_2 at the chosen reference levels for the treatment type of that subgraph.

From this network one can infer that a high pH of 7.2 introduces large deviations from the reference sample in O_2 and browning. Additionally, it is clear that within samples of this high pH value, successively higher levels of SO_2 inhibit the onset of browning, but exacerbates O_2 depletion in the early

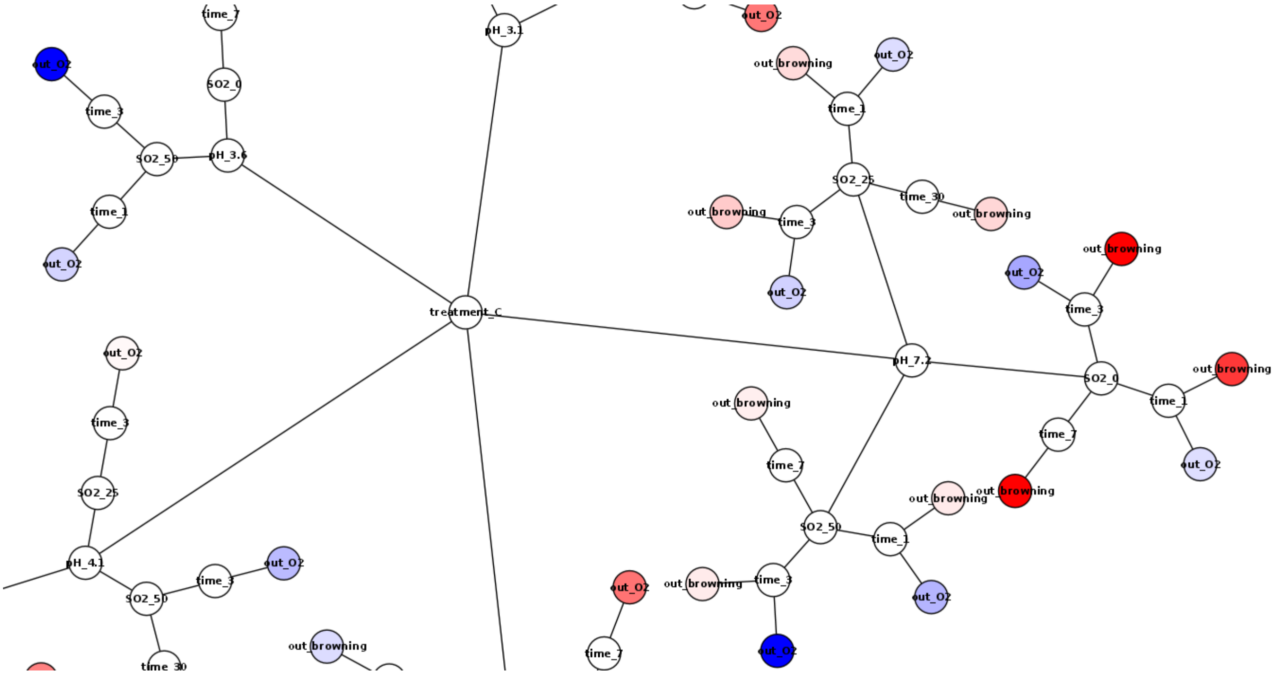


Figure 4.10: Tiered network centered on a constant treatment. The outermost leaves in the network are the experimental results for browning and oxygen levels. They have a colour intensity according to the fold change as compared to the reference level for that time and set of conditions: red for positive and blue for negative fold change.

stages of the experiment.

4.3.3.3 Star Network

A further type of network that can be constructed uses edges instead of nodes to describe the experimental outputs. The 'star' network shown in Figure 4.11 is more compact than the other figures, and also lends itself to advanced queries. To build such a network, two conditions are selected as nodes and a third the edges between them. In the case of Figure 4.11, the inner nodes are treatment types; outer nodes pH and the third variable SO_2, mapped alongside browning and O_2 levels in the network edges.

A disadvantage to simplifying the data in this way is that a single time point has to be chosen. In this case, the average ET50 for browning in the reference samples was used as the comparison time, as it was assumed that this would be the point at which variation would be highest across samples.

Any significant changes in experimental outputs are then mapped along with the third condition, so that its direct effects on the experiment are made clear. The overall network shows, predictably, a high concentration of strong differences with pH 7.2. If one selects a subgraph in a similar manner to the

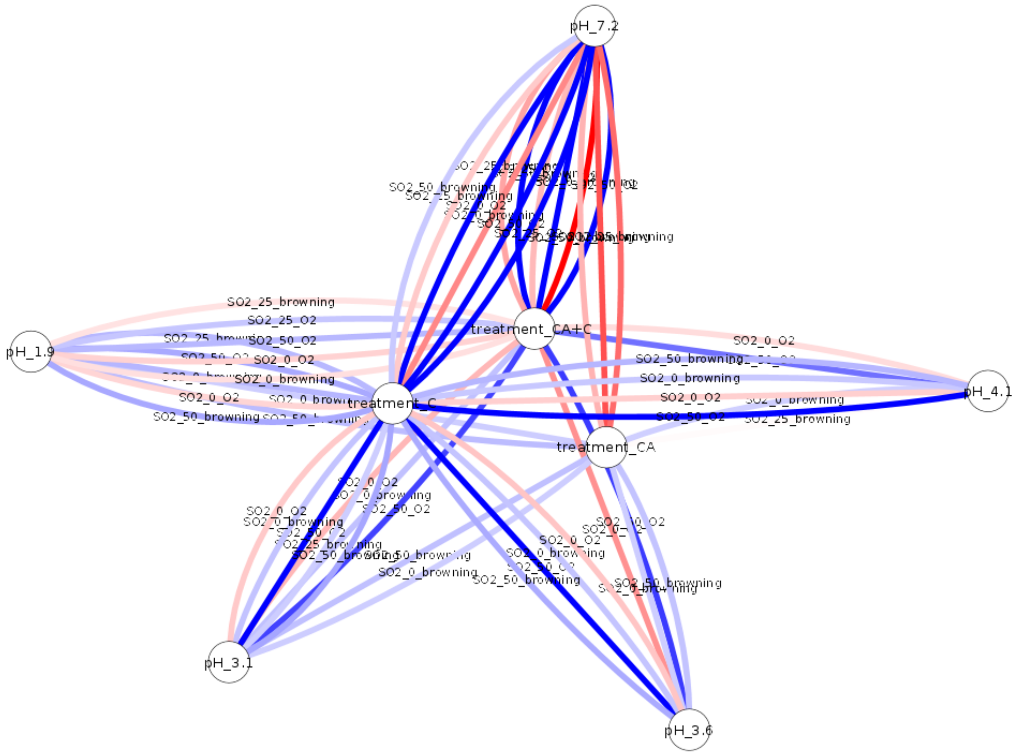


Figure 4.11: 'Star' network for more concise representation of browning data. The inner and outer nodes are for two respective experimental variables. The edges between these nodes represent the significant differences for the levels of a third variable, for the outputs of browning- and oxygen levels. The intensity is reflective of the fold change from the reference level: red for positive and blue for negative.

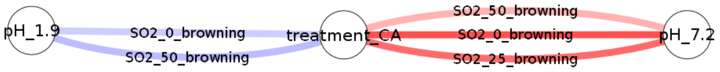


Figure 4.12: Query into specific conditions in 'star' network - a subgraph selection of Figure 4.11

tiered network, a direct comparison of conditions can be done as depicted in figure 4.12.

The subgraph is essentially a query into the relative effects of changing pH when the treatment is constant. At the assumed midway point in the experiment, it would appear that the browning rate is more rapid for the higher pH level, something that has been observed in previous networks. Many such queries can be made of the experiment in this way.

4.4 Conclusion

Several new techniques were presented for the visualisation of scientific data in networks. A basic framework of data import, statistical testing and visualisation through Cytoscape interactive software was used; this basic method is widely extensible and permutations can be built on an ad-hoc basis for different types of data.

Through its application to three different data sets, all derived from disparate sources and with differing structure, the broad nature of the method was demonstrated. It can be applied to time-dependent and independent data alike; if the data is temporal then different views can either use time points as a corporeal variable or as a background feature about which simplifying assumptions are made.

The potential of these methods to generate new hypotheses and facilitate exploratory data analysis was also summarily described. The presentation of the graphical model in an interactive topographical space assists in these ends. In particular, there is fulfilment of many of the interactive features described by Keim (2002): the ability to 'zoom', as well as to split the data into logical subsets and re-order. The advantage in having these abilities is that advanced queries into specific relationships and features of the data can be made with relatively little effort.

It is therefore proposed that statistical network models of the data - StatNet - be used in conjunction with traditional techniques in order to fully explore both global and localised trends and relationships within scientific experimental data.

4.5 Acknowledgments

I would like to thank the students and supervisors at the IWBT for their support in this project. Most especially the following researchers: Dr. Astrid Buica, who conducted the experiment on browning and whose collaboration was invaluable to the development of StatNet techniques; Kari du Plessis for the experiment on *Botrytis* infection of *Vitis vinifera* and Carien Coetzee for her sensory experiment. Finally, the Computational Biology group at IWBT for their guidance and advice.

4.6 List of References

Herman, I., Melancon, G. and Marshall, M.S. (2000). Graph visualization and navigation in information visualization: A survey. *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS*, vol. 6, no. 1, pp. 24–43. Available at: `http://ieeexplore.ieee.org`

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70.

Keim, D. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, no. 1, pp. 1–8. ISSN 10772626.
Available at: `http://ieeexplore.ieee.org`

Kelleher, C. and Wagener, T. (2011 June). Ten guidelines for effective data visualization in scientific publications. *Environmental Modelling & Software*, vol. 26, no. 6, pp. 822–827. ISSN 13648152.
Available at: `http://linkinghub.elsevier.com/`

Oliphant, T.E. (2007). Python for Scientific Computing. *Computing in Science & Engineering*, vol. 9, no. 3.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003 November). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, vol. 13, no. 11, pp. 2498–504. ISSN 1088-9051.
Available at: `http://www.pubmedcentral.nih.gov`

Chapter 5

Conclusion

5.1 Summary

Two research avenues were explored within the context of data exploration. In the first, a new method for untargeted analysis was developed for HPLC/UV-vis data. The data set used was a large-scale analysis of the effects of various experimental conditions on the browning effects of white wine.

The method was derivative of known techniques of data pre-processing, which were often parallelised to accommodate the massive scale of the data. This pre-processing included baseline adjustment, smoothing and alignment. After pre-processing, the data was subjected to a new feature map alignment technique based on feature similarity across wavelengths.

The putative features collected in this way were then mined for information related to the original experimental setup using three different methods: PCA, decision tree analysis and network modelling. It was found through validation techniques (especially related to the decision tree analysis and PCA) that there were indeed significant relationships between putative features and aspects of the experimental conditions.

For the second research avenue, there was a focus on a conserved network modelling technique. This technique was broadly applied and adapted to several disparate data sets. These included a sensory analysis of wine; fungal infection of *Vitis vinifera* leaves, and the same browning study as for the first section of research. The example data included both temporal and static data, for which the adaptability of the method was demonstrated. For each case study, several network views were generated with respective structure and topography related to experimental variables.

5.2 Conclusion

The primary aim of this thesis is to build novel platforms for new insights and hypothesis generation in scientific data. To this end two different development projects were undertaken - one more fundamentally associated with raw data generated from chemometrics; the other focusing entirely on meaningful representation and visualisation of existing experimental data.

Though these two avenues of research are entirely different in content, they demonstrate the value in the same fundamental truth: that, as suggested by Tukey, 'finding the questions is often more useful than finding the answers'. HPLC/UV-vis chromatograms are ubiquitous in research, however the purpose of generating the data is almost always confirmatory - not exploratory. Methods to detect and contextualise unexplored compounds can refocus the direction of an experiment and facilitate novel directions of research. Much of the time such pre-existing data is used for confirmatory analysis is available to be mined, a cost effective measure indeed for scientific research.

Likewise, experimental data is almost always value-enriched by simply visualising the results from a different angle. When the question underlying an experiment is broad, such as in the case of factorial- or composite design with multiple input variables, the ability to map variable relationships; query outputs and formulate conclusions and hypotheses about causality is pivotal. In this way questions and answers must be simultaneously presented.

It is contended that the primary aim of the thesis has, to a great extent, been fulfilled. Novel platforms were developed from combinations of existing and developed techniques to condense a myriad of chromatograms into a common matrix of putative compounds; thereafter to identify possible significance of the compounds in the overall experiment. The resultant data was validated through machine learning techniques.

The network views generated for application to scientific experimental data also constitutes a moderately successful attempt at a new paradigm for data interpretation. While the proof of its usefulness is more anecdotal, it is proposed that viewing data in this way can be less convoluted than traditional techniques and provide scientists with a powerful tool for delving the complex relationships of dependent variables.

5.3 Future Perspectives

Many of the studies presented here can be considered preparatory forays into the respective fields of UV-vis feature map alignment and network visualisation. There is certainly room for improvement of the feature map alignment problem, the possible shortcomings of which have been outlined. As a next step, the existing methods for MS feature map alignment could be systematically extended to UV-vis data and the results compared. A probabilistic approach is probably worth exploring first, as an alternative to much of the existing deterministic algorithms of the MS methods.

While the current work seeks to explore an untargeted space in chemometrics, the actual interpretation of the results by a chemist should still be done.

This would include identifying the significant putative compounds of interest and checking whether their proposed relationships can be kinetically validated. Furthermore, empirical confirmation of their significance with regards to the experimental conditions can be attempted.

Other machine learning methods can also be attempted for classification of experimental variables using the putative compounds (besides PCA and decision trees). For multiple classification, an ensemble method such as random forests could be applied. Non-linear methods such as Support Vector Machines could also be tested, as the assumption of linearity in the feature map could be an over-simplification.

NetStat could also be further extended. The network visualisation methods were only tested on 3 different data sets, however there is no theoretic limitation to its application with experimental data (besides perhaps a very large number of experimental variables). As demonstrated, network visualisation of experimental data can be adapted to suit fundamentally different formats - though often some creative thought is required to devise an appropriate topology.

Similar to the above, there is also a need for interpretation and validation by the researchers that performed the experiments. Much of the work in which NetStat features is already under review.