university of groningen

faculty of science and engineering

ROLAND JANNO VEEN

MACHINE LEARNING ANALYSIS OF STEROID METABOLOMICS DATA: SUBTYPES OF ADRENAL TUMOURS

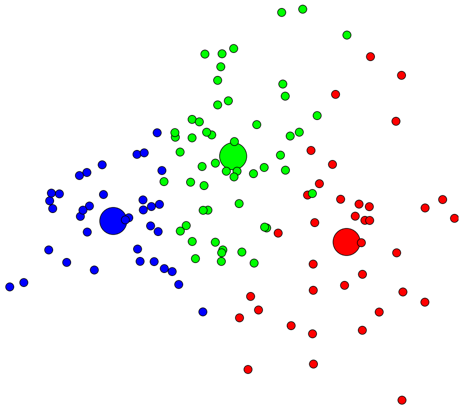
university of groningen

faculty of science and engineering

MACHINE LEARNING ANALYSIS OF STEROID METABOLOMICS DATA: SUBTYPES OF ADRENAL TUMOURS

ROLAND JANNO VEEN

s1211404, r.j.veen@student.rug.nl / roland.veen@gmail.com



An application of unsupervised clustering in the biomedical domain

Supervised by

PROF. DR. MICHAEL BIEHL

PROF. DR. KERSTIN BUNTE

Master of Science (MSc)
Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence
Faculty of Science and Engineering
University of Groningen

August 2022 – e-print

Roland Janno Veen: *Machine Learning Analysis of Steroid Metabolomics Data: Subtypes of Adrenal Tumours,* An application of unsupervised clustering in the biomedical domain, © August 2022

SUPERVISORS:
Prof. Dr. Michael Biehl
Prof. Dr. Kerstin Bunte

LOCATION:

Groningen

Ohana means family.
Family means nobody gets left behind, or forgotten.

— Lilo & Stitch

ABSTRACT

The fatality rate of Adrenocortical Carcinomas is high. Due to low prevalence we still have a lot to learn. Using unsupervised clustering of metabolomics profiles, we find compelling evidence backed by domain knowledge of a subdivision of the carcinomas in three clusters. Identifying different types of Adrenocortical Carcinomas may in the future enable targeted treatment leading to increased survival rates.

SAMENVATTING

Het sterftecijfer onder patiënten met kwaadaardige bijnierschorstumoren is hoog. Wegens de zeldzaamheid van dit type kanker is er nog veel onbekend. Met behulp van unsupervised clustering van metabolomics-profielen vinden we overtuigend bewijs dat wordt ondersteund door domeinkennis van een onderverdeling van de kwaadaardige tumoren in drie groepen. Het identificeren van verschillende soorten kwaadaardige bijnierschorstumoren kan in de toekomst gerichte behandelingen mogelijk maken die kunnen leiden tot verhoogde overlevingskansen.

We can see only a short distance ahead,
but we can see plenty there
that needs to be done.

— Alan M. Turing [1]

ACKNOWLEDGMENTS

First I would like to thank Michael Biehl for his never-ending patience, guidance and unwavering belief in my abilities. Second my thanks to Kerstin Bunte, who spontaneously agreed to become second supervisor.

Furthermore I would like to thank Wiebke Arlt, Angela Taylor, James Hawley, Warwick Dunn, Thomas Papathomas, Vasileios Chortis, Alessandro Prete, Alice Sitch, Lorna Gilligan, Brian Keevil and Fozia Shaheen for welcoming me in the research group, their contributions and support.

Special thanks to my brother Jelmer Veen.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRONYMS

SOM	Self-Organising Map
HC	Hierarchical Clustering
ACC	Adrenocortical Carcinoma
ACA	Adrenocortical Adenoma
USM	Urine Steroid Metabolomics
LC-MS/MS	Liquid Chromatography-Tandem Mass Spectrometry
GC-MS	Gas Chromatography-Mass Spectrometry
PCA	Principal Component Analysis
UCAT	Unsupervised Clustering Analysis Tool
GMLVQ	Generalised Matrix Learning Vector Quantization

LIST OF STEROID HORMONE METABOLITES

This is a list of the steroid hormone metabolite abbreviations used and their full names[1] (from: [2]).

An, Andro	Androsterone
Etio	Etiocholanolone
DHEA	Dehydroepiandrosterone
16-α-DHEA	16-α-Dehydroepiandrosterone
5-PT	5-Pregnenetriol
(sum) 5PD	(Pregnenediol and) 5-Pregnenediol
THA	Tetrahydro-11-dehydrocorticosterone
5α-THA	5α-Tetra-11-dehydrocorticosterone
THB	Tetrahydrocorticosterone
5α-THB	5α-Tetrahydrocorticosterone
3α, 5β-THALDO	3α, 5β-Tetrahydroaldosterone
THDOC	Tetrahydrodeoxycorticosterone
PD	Pregnanediol
3α5α17HP	3α, 5α-17-hydroxy-pregnanolone
17HP	17-hydroxy-pregnanolone
PT	Pregnanetriol
PTONE	Pregnanetriolone
THS	Tetrahydro-11-deoxycortisol
cortisol	Cortisol
6β-OH-cortisol	6β-hydroxy-cortisol
THF	Tetrahydrocortisol
5αTHF	5α-Tetrahydrocortisol
α-cortol	α-Cortol
β-cortol	β-Cortol
11β-OH-Andro	11β-Hydroxy-androsterone
11β-OH-etio	11β-Hydroxy-etiocholanolone
cortisone	Cortisone
THE	Tetrahydrocortisone
α-cortolone	α-cortolone
β-cortolone	β-cortolone
11-oxo-etio	11-Oxo-etiocholanolone

[1] When not specified the A-ring configuration is 3α,5α- in all abbreviations

1 INTRODUCTION

As there is still a lot to learn about Adrenocortical Carcinoma (ACC), machine learning techniques may be of use to give an unbiased and objective insight into their metabolome.

In this thesis, a medical data set, provided by our collaborators from the University of Birmingham/UK, is analysed by unsupervised machine learning. The data comprises steroid metabolomics and clinical data of patients with malignant adrenocortical tumours - ACC. The goal is to identify potential subgroups of patients in which the tumours display similar characteristics that distinguish them from other groups.

This is attempted by unsupervised learning and clustering techniques. Together with the medical researchers at the University of Birmingham, we define tumour classes based on medical insight and the results of the unsupervised analysis. Eventually, Generalised Matrix Learning Vector Quantization (GMLVQ) will be used to reproduce the hypotheses and to identify the relevances of the steroid biomarkers for the classification.

We address the following research questions:

Q1) Does unsupervised learning reveal clusters / subgroups of adrenal tumours in the data?

Q2) Does the inclusion of clinical data beyond steroid metabolomics help in Q1?

Q3) Can supervised learning / classification reproduce the clusters defined in Q1 and Q2?

Q4) What is the relevance or importance of individual steroid markers for the discrimination of tumour sub-types?

In the remainder of this chapter, we briefly discuss previous work on which our research builds, provide a summary of the biomedical embedding of the work, an overview of the followed methodology, and a brief introduction to the algorithms used for unsupervised clustering.

1.1 PREVIOUS WORK

The researchers in Birmingham, in collaboration with the European Network for the Study of Adrenal Tumours (ENSAT) undertook a proof-of-principle study [3] and a subsequent international multi-centre prospective test validation study (EURINE-ACT; [4]) of the diagnostic value of an algorithm that was obtained with supervised machine learning [5]. This algorithm differentiates incidentalomas between malignant Adrenocortical Carcinoma (ACC) and benign Adrenocortical Adenoma (ACA) on the basis of the steroid metabolome in the urine excreted by the adrenal tumour patients. For the combination of steroid metabolome profiling by mass spectrometry techniques and subsequent classification on the basis of machine learning, the authors coined the term Urine Steroid Metabolomics (USM). As a follow-on, the steroid data have been used to undertake *'Classification of Benign Adrenal Tumors Based on Steroid Metabolomics'* [6] which describes finding subclasses of ACA. In this paper, we now wish to expand upon this previous research by looking into clusters and possibly classes in ACC [7, 8].

An incidentaloma is an incidentally (hence the name) discovered mass that can either be an adenoma (benign) or carcinoma (malignant)

1.2 MEDICAL BACKGROUND

1.2.1 *Adrenocortical Carcinoma*

Adrenocortical Carcinoma is a type of cancer that originates from the cortex of the adrenal gland. The adrenal gland specialises in the synthesis of of steroid hormones, including glucocorticoids, mineralocorticoids, and precursors of sex steroid biosynthesis, namely adrenal androgens.

Adrenal Cortical Carcinoma is rare, with an incidence of 0.7 to 2.0 new patients presenting per million people per year, and has a poor prognosis, with a 5 year survival rating of below 40%. Excess secretion of steroid hormones by the tumour can lead to clinical signs and symptoms, e. g. Cushing's Syndrome or virilisation [9–13]. Treatment options are very limited, but the outcome is improved with early diagnosis. In this paper, we attempt to discover subtypes of ACC to hopefully aid in future diagnosis and treatment.

1.2.2 *Steroid Metabolomics*

A well-known diagnostic test is examining steroid hormone metabolites in urine. To account for natural fluctuations of hormone levels throughout the day and night, this is collected by the patient over 24 hours. From this volume, a number of samples are taken and converted to digital data using either Gas Chromatography-Mass Spectrometry (GC-MS) and/or Liquid Chromatography-Tandem Mass Spectrometry (LC-MS/MS). After processing, we get concentrations of the following steroid hormone metabolites, divided in four groups based on their biological function and relationships:

ANDROGENS AND A-PRECURSORS Androsterone (An, Andro), Etiocholanolone (Etio), Dehydroepiandrosterone (DHEA), 16-α-Dehydroepiandrosterone (16-α-DHEA), 5-Pregnenetriol (5-PT), (Pregnenediol and) 5-Pregnenediol ((sum) 5PD)[1]

MINERALOCORTICOIDS AND MC-PRECURSORS Tetrahydro-11-dehydrocorticosterone (THA), 5α-Tetra-11-dehydrocorticosterone (5α-THA), Tetrahydrocorticosterone (THB), 5α-Tetrahydrocorticosterone (5α-THB), 3α, 5β-Tetrahydroaldosterone (3α, 5β-THALDO), Tetrahydrodeoxycorticosterone (THDOC)

GLUCOCORTICOID-PRECURSORS Pregnanediol (PD), 3α, 5α-17-hydroxy-pregnanolone (3α5α17HP), 17-hydroxy-pregnanolone (17HP), Pregnanetriol (PT), Pregnanetriolone (PTONE)

GLUCOCORTICOIDS Tetrahydro-11-deoxycortisol (THS), Cortisol (cortisol), 6β-hydroxy-cortisol (6β-OH-cortisol), Tetrahydrocortisol (THF), 5α-Tetrahydrocortisol (5αTHF), α-Cortol (α-cortol), β-Cortol (β-cortol), 11β-Hydroxy-androsterone (11β-OH-Andro), 11β-Hydroxy-etiocholanolone (11β-OH-etio), Cortisone (cortisone), Tetrahydrocortisone (THE), α-cortolone (α-cortolone), β-cortolone (β-cortolone), 11-Oxo-etiocholanolone (11-oxo-etio)

A schematic of steroid synthesis and metabolism can be found in figure 1.1.

1.2.3 *GC-MS versus LC-MS/MS*

Gas and liquid chromatography with mass spectrometry allow us to quantify steroids in urine. Both methods have their advantages and limitations. Where LC-MS/MS is generally less expensive, offers a faster analysis time and has a higher throughput, GC-MS has higher resolving power, thus allowing for a broader palette of steroid metabolites to be examined which aid in the discovery of novel metabolomes [2, 14]. In this paper, we use both methods in tandem for the most part. The extraction methods used for GC-MS and LC-MS/MS can be found in [4, 15].

1.3 METHODOLOGY

At first, unsupervised learning techniques are used for the analysis of the steroid metabolomics data. These can include, but are not limited to, principle component analysis, hierarchical clustering, vector quantisation, and self-organising maps. The correlation of potential clusters with clinical data is studied in terms of post-labelling and, if possible, visualisation thereof. The assumed cluster structure is then

[1] Pregnenediol and 5-Pregnedienol make sum 5-PD for GC-MS due to these two products being produced in the derivatisation process. In LC-MS/MS we measure just 5-pregnenediol.

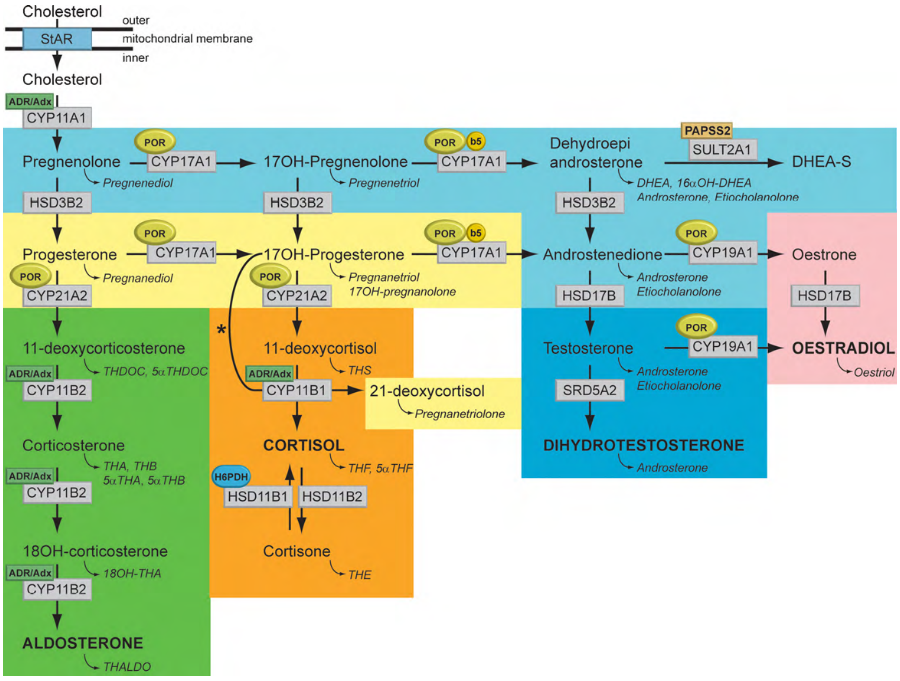


Figure 1.1: Synthesis and metabolism of hormonal steroids. This figure illustrates the formation of the major hormone classes from cholesterol. Steroid names in conventional script are steroid hormones and precursors; those in italics are urinary metabolites of the aforementioned. The major transformative enzymes are in rectangular boxes, the cofactor ("facilitator") enzymes in ovals. The pale blue area contains common intermediate steps; the yellow area, preliminary steps in glucocorticoid synthesis; the green, mineralocorticoids; the orange, glucocorticoids; dark blue, androgens and pink, oestrogens. Taken from: [2]

used to define hypothetical sub-classes of malignant tumours. Alternatively (or in addition) hypotheses formulated on the basis of expert knowledge are considered. By means of supervised techniques, i.e. classifiers like Generalised Matrix Relevance Learning Vector Quantisation, or Random Forest, we aim to reproduce the suggested clusters or classes. Emphasis is on identifying the steroid markers or clinical data that are most discriminative with respect to the clustering.

1.4 ALGORITHMS FOR UNSUPERVISED CLUSTERING

As it is not feasible to apply every clustering method in existence to this problem within the scope of a master's thesis, we selected the following candidates to be thoroughly investigated: Self-organising Map, k-means, k-medoids, and hierarchical clustering. In this section, we give a brief introduction to these algorithms.

1.4.1 *Self-Organising Map*

The Self-Organising Map (SOM) is proposed by Kohonen [16] to be a model of neighbourhood cooperativeness between the prototypes, which is motivated by the formation of ordered representations of sensory information in cortical brain areas [17]. A typical schematic of a SOM is seen in figure 1.2. Samples are taken from a high dimensional input space. Its variables are used as an input vector and - together with a weight vector - a winning node is determined.

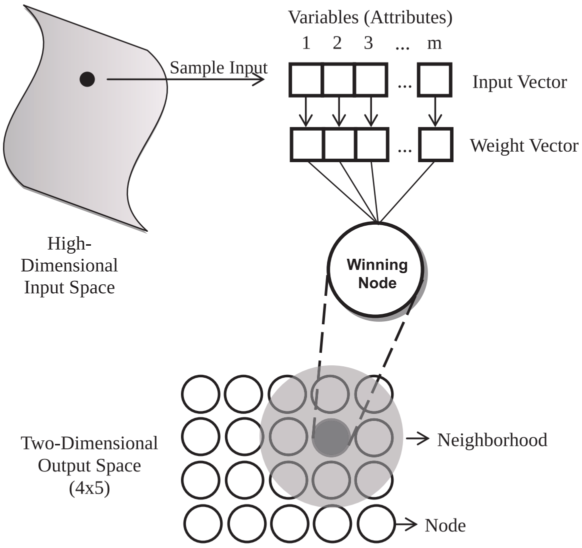


Figure 1.2: Graphical illustration of a self-organising map. Taken from: [18]

1.4.2 *k-means*

k-means is sometimes referred to Lloyd's Algorithm in honour of its inventor

The k-means algorithm was first proposed by S.P. Lloyd in *'Least squares quantization in PCM'* [19, 20]. *'It is an iterative, data-partitioning algorithm that assigns* n *observations to exactly one of* k *clusters defined by centroids, where* k *is chosen before the algorithm starts.'*, from: [21].

1.4.3 *k-medoids*

First described by Leonard Kaufman and Peter J. Rousseeuw in *'Partitioning Around Medoids (Program PAM)'* [22].

*'*k*-medoids clustering is a partitioning method commonly used in domains that require robustness to outlier data, arbitrary distance metrics, or ones for which the mean or median does not have a clear definition.*

It is similar to k*-means, and the goal of both methods is to divide a set of measurements or observations into* k *subsets or clusters so that the subsets minimize the sum of distances between a measurement and a center of the measurement's cluster. In the* k*-means algorithm, the center of the subset is the mean of measurements in the subset, often called a centroid. In the* k*-medoids algorithm, the center of the subset is a member of the subset, called a medoid.*

The k*-medoids algorithm returns medoids which are the actual data points in the data set. This allows you to use the algorithm in situations where the mean of the data does not exist within the data set. This is the main difference between* k*-medoids and* k*-means where the centroids returned by* k*-means may not be within the data set. Hence* k*-medoids is useful for clustering categorical data where a mean is impossible to define or interpret.'*, from: [23]

1.4.4 *Hierarchical Clustering*

Hierarchical clustering is described and used in many papers, e. g. in *'Algorithms for hierarchical clustering: an overview'* [24] and *'The Shallow and the Deep, An Introduction to Neural Networks and Machine Learning'* [20].

'Hierarchical clustering groups data over a variety of scales by creating a cluster tree or dendrogram (figure 1.3). The tree is not a single set of clusters, but rather a multilevel hierarchy, where clusters at one level are joined as clusters at the next level. This allows you to decide the level or scale of clustering that is most appropriate for your application.', from: [25]

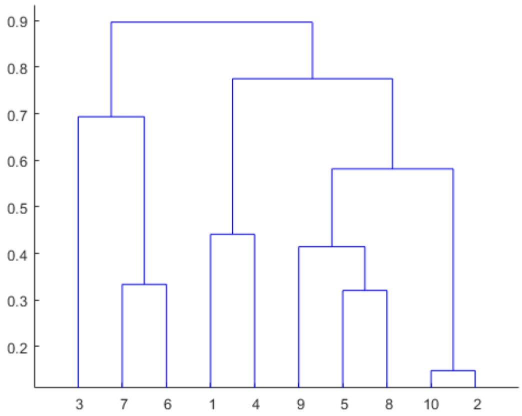


Figure 1.3: Example of a dendrogram. From: [26]

2 RESULTS

In this chapter we present the results of our experiments. We will first describe the pre-processing steps we have taken. Next, we describe the analysis framework we created and/or used. This is followed by a look into the optimal and/or desired number of clusters. We conclude this chapter with the results of the clustering for each chosen algorithm.

2.1 PRE-PROCESSING

When untreated, the data does not reveal its secrets readily (figures 2.1a and 2.1b).

First attempts of clustering after doing a customary log-transform showed that the algorithms mostly focused on differences in total metabolite concentrations which is not that interesting (see appendix for figure A.1).

Usually, a z-score-transformation is helpful. However, this time it did not produce good results, i. e. bad separation, imbalanced cluster sizes, and too many outliers (see appendix, figure A.2).

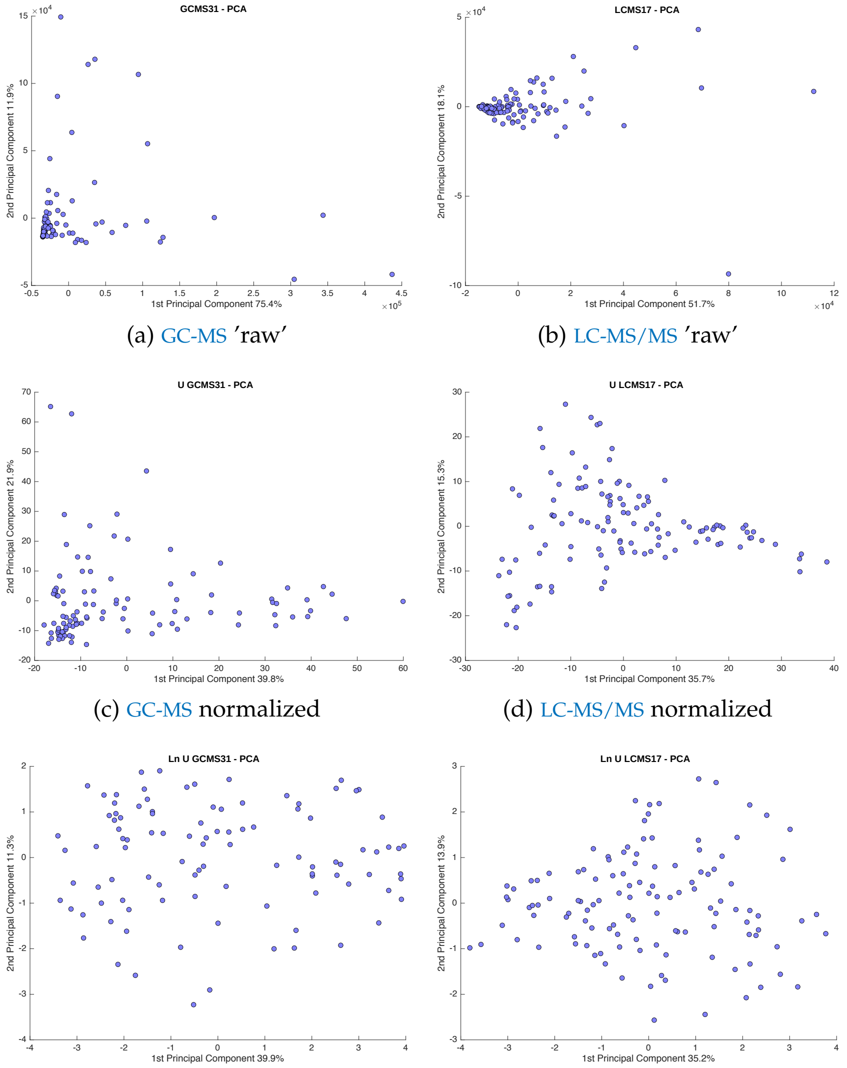
As a solution, we decided to normalize the steroid panel for each patient (sample), dividing each variable by the sum and multiplying by 100, basically unit-normalization with unit 100. Now we see a better spread of the samples across the first two components of the Principal Component Analysis (PCA) (2.1c, 2.1d). Empirically this gave an improved result over simply doing L1-normalization (unit 1).

By applying the natural logarithm to the normalized steroid concentration values, we can see that the samples are becoming even more spread out in the projection (2.1e, 2.1f). This is further illustrated in figure 2.2 where we see box charts at the different stages of pre-processing.

For the interested reader, please find histograms of the pre-processing steps in the appendix, figures A.3 - A.8, and a bar chart of the variance explained by each principal component for the transformed data in figure A.9.

2.2 ANALYSIS FRAMEWORK

For the research in this paper we use the software MATLAB (R2021b) by Mathworks. To systematically research the proposed questions, we assembled and created several libraries to process and analyse the data in different ways.



(a) GC-MS 'raw'

(b) LC-MS/MS 'raw'

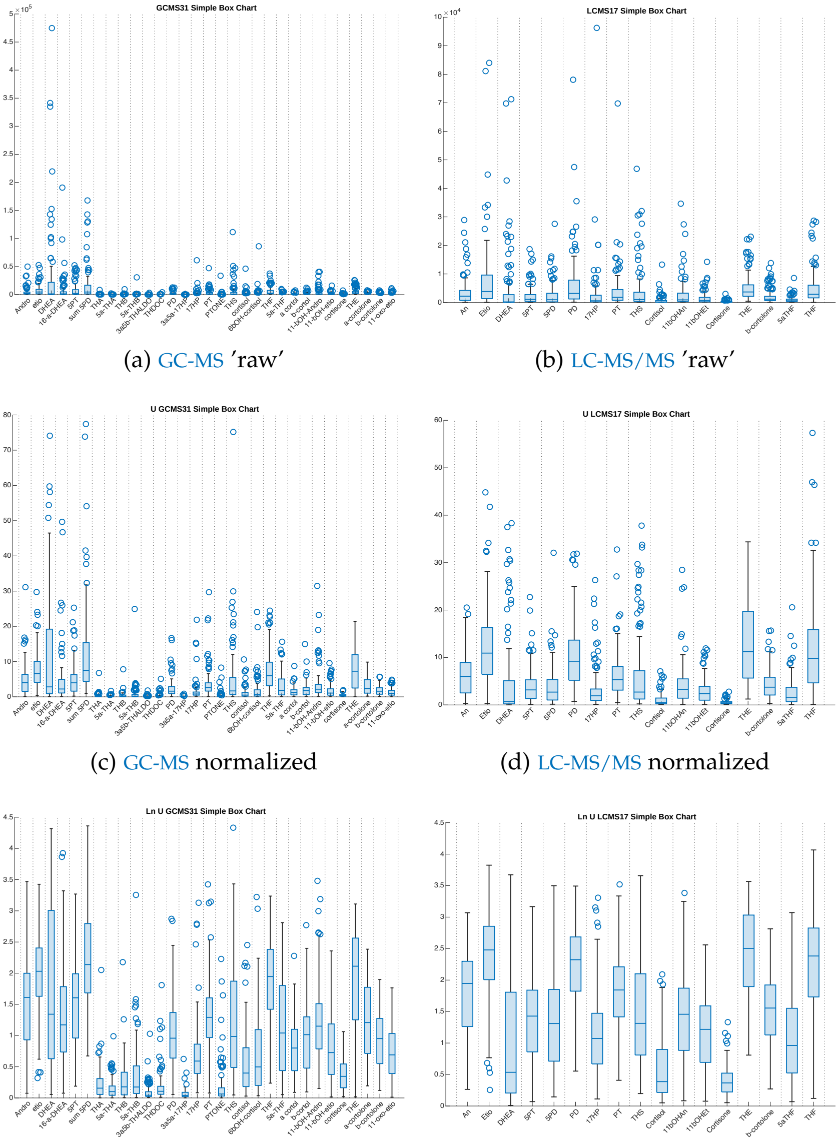
(c) GC-MS normalized

(d) LC-MS/MS normalized

(e) GC-MS Log-transformed normalized

(f) LC-MS/MS Log-transformed normalized

Figure 2.1: Data pre-processing steps: Principal Component Analysis. From top to bottom we see the 'raw' data, normalized data and finally log-transformed normalized data, projected on the first two principal components.



(a) GC-MS ’raw’

(b) LC-MS/MS ’raw’

(c) GC-MS normalized

(d) LC-MS/MS normalized

(e) GC-MS Log-transformed normalized

(f) LC-MS/MS Log-transformed normalized

Figure 2.2: Data pre-processing steps: Box charts. Here we see the effects of the pre-processing on the metrics of the individual steroids.

2.2.1 *Clustering evaluation*

We used native MATLAB implementations of the following criteria to evaluate the number of clusters: Davies-Bouldin [27], Calinski-Harabasz [28], Silhouette [29], A modified version of the Gap statistic [30] as proposed by prof. dr. Nicolai Petkov (unpublished) and the simple elbow method [31]. Each of these (except the elbow method) try to either minimize (Davies-Bouldin) or maximize (all the other methods) a cluster separation measure.

2.2.2 *SOM Toolbox*

For our experiments with SOM we used the SOM-Toolbox by by Esa Alhoniemi, Johan Himberg, Jukka Parviainen and Juha Vesanto[1] [32]. We chose a hexagonal lattice on a toroid. This is a popular choice, as it is isotropic and generally works well. We let the toolbox determine the size of the map. After training, we apply a k-means on the neural lattice to cluster the results.

2.2.3 *Unsupervised Clustering and Analysis Tool*

To make the experiments easy to reproduce and rapidly change parameters, we have written a library we call Unsupervised Clustering Analysis Tool (UCAT). it consists of the class `UCAT` with supporting classes and functions. This neatly encapsulates data, pre-processing steps, clustering results and visualisation functions.

2.2.4 *GMLVQ Toolbox*

The no-nonsense GMLVQ Toolbox [33] was first published by Michael Biehl in 2015. It was extended and upgraded by Floris Westerman in 2019. The latest feature changes and bug fixes were done by the author of this thesis in 2021 and is currently being maintained by the same.

GMLVQ was first described in [34]. It adds a matrix scheme to Relevance Learning Vector Quantisation (RLVQ) [35, 36] which in turn is based on Learning Vector Quantisation (LVQ) by Kohonen [16].

The GMLVQ toolbox, like UCAT, is object oriented in design and integrates neatly with UCAT. They use the same formats for data and labels for easy exchange of data. Unsupervised clustering results can thus be turned into labelled samples for supervised classification with minimal effort.

In this paper we use GMLVQ to further investigate properties of the unsupervised clustering by looking at the relevances, separability, accuracy and projection.

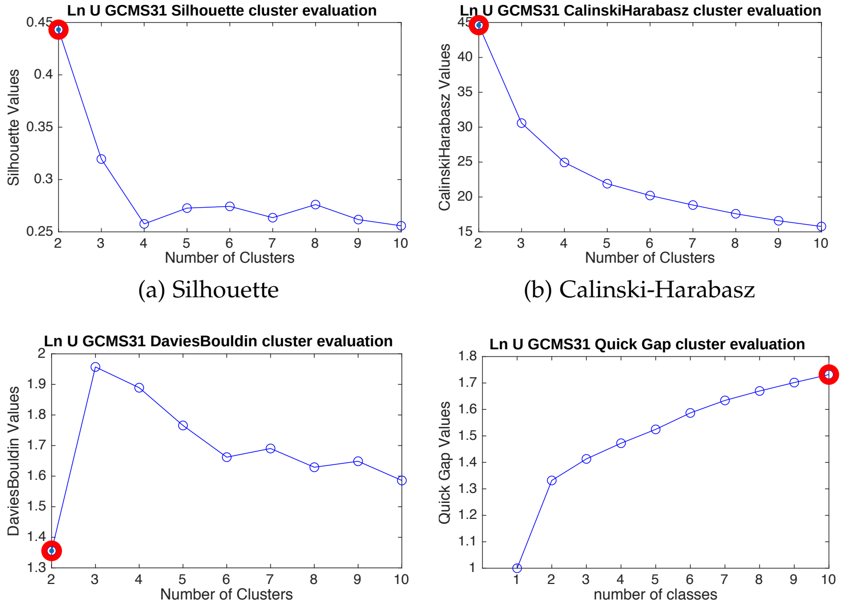
[1] https://github.com/ilarinieminen/SOM-Toolbox

2.3 MACHINE VERSUS BIOLOGY - NUMBER OF CLUSTERS

In this chapter we give an overview of the results and how they fit with expert domain knowledge. First we look at objective measures for clustering evaluation, followed by a short intermediate discussion about the number of clusters we should focus on.

2.3.1 *Clustering evaluation results*

We performed an objective cluster analysis using four criteria on the processed data (Silhouette, Calinski-Harabasz, Davies-Bouldin and Quick Gap) for both GC-MS (see figure 2.3 and LC-MS/MS (see figure 2.4). The criteria show similar results across both the data sets individually (except Davies-Bouldin), but are not in agreement when comparing the criteria to each other. They either recommend two clusters or an unrealistic number. A fifth criterion is the elbow method, as seen in figure 2.5. Unfortunately this result is not very clear either, with a rather shallow "bend".



(a) Silhouette

(b) Calinski-Harabasz

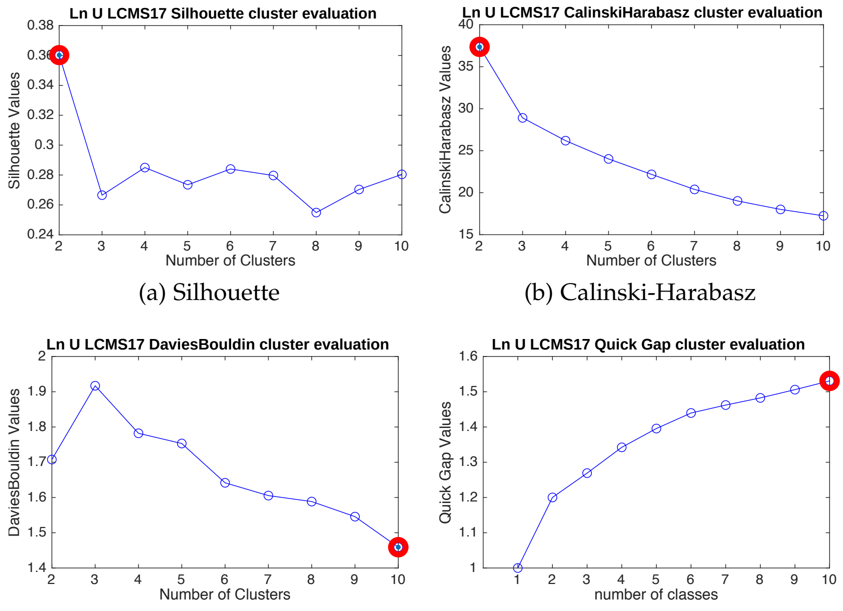
(c) Davies-Bouldin

(d) Quick Gap

Figure 2.3: GC-MS cluster evaluation results. Red circle marks the "best" value for the respective algorithm.

2.3.2 *Two clusters or three?*

Based on the objective cluster analysis, a point can be made for two clusters. In fact, a reasonable two cluster separation can be achieved



(a) Silhouette

(b) Calinski-Harabasz

(c) Davies-Bouldin

(d) Quick Gap

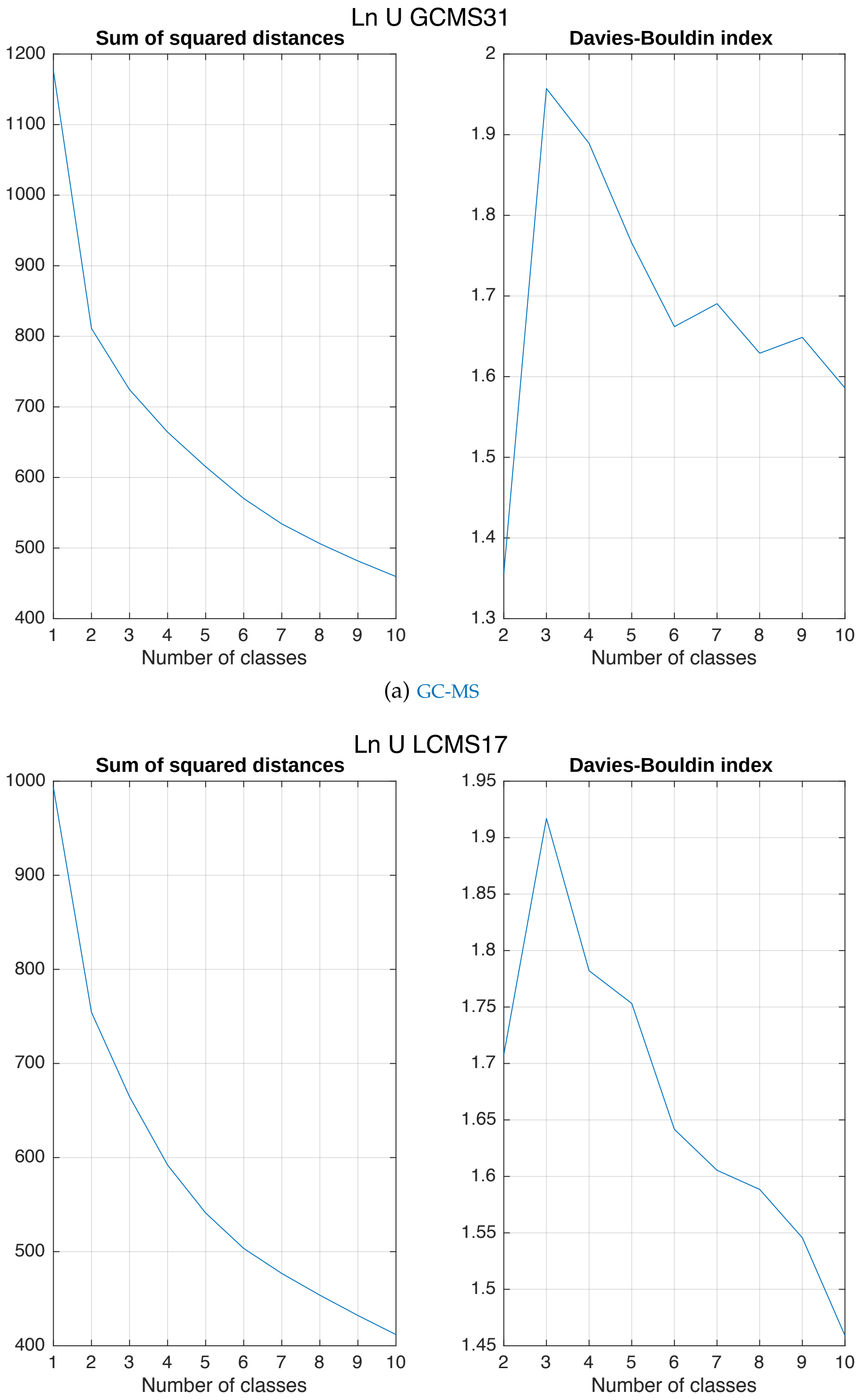
Figure 2.4: LC-MS/MS cluster evaluation results. Red circle marks the "best" value for the respective algorithm.

by using only four variables: DHEA versus the sum of THF, 5αTHF, and THE (see figure 2.6).

When looking at a PCA projection (figure 2.7), post-hoc coloured with k-means cluster labels, we already see some clustering. However, domain knowledge suggests there is merit to consider three clusters. The alignment of the three clusters with the established grouping of steroids and their precursors strongly suggests that this is more useful from a biomedical perspective.

When looking at the cluster characteristics in figure 2.8, we see the red cluster mostly unchanged, staying dominant in the Androgens and A-precursors, where the blue cluster splits up into blue and green, blue being mostly dominant in Glucocorticoids, and green Mineralocorticoids / MC-precursors and GC-precursors.

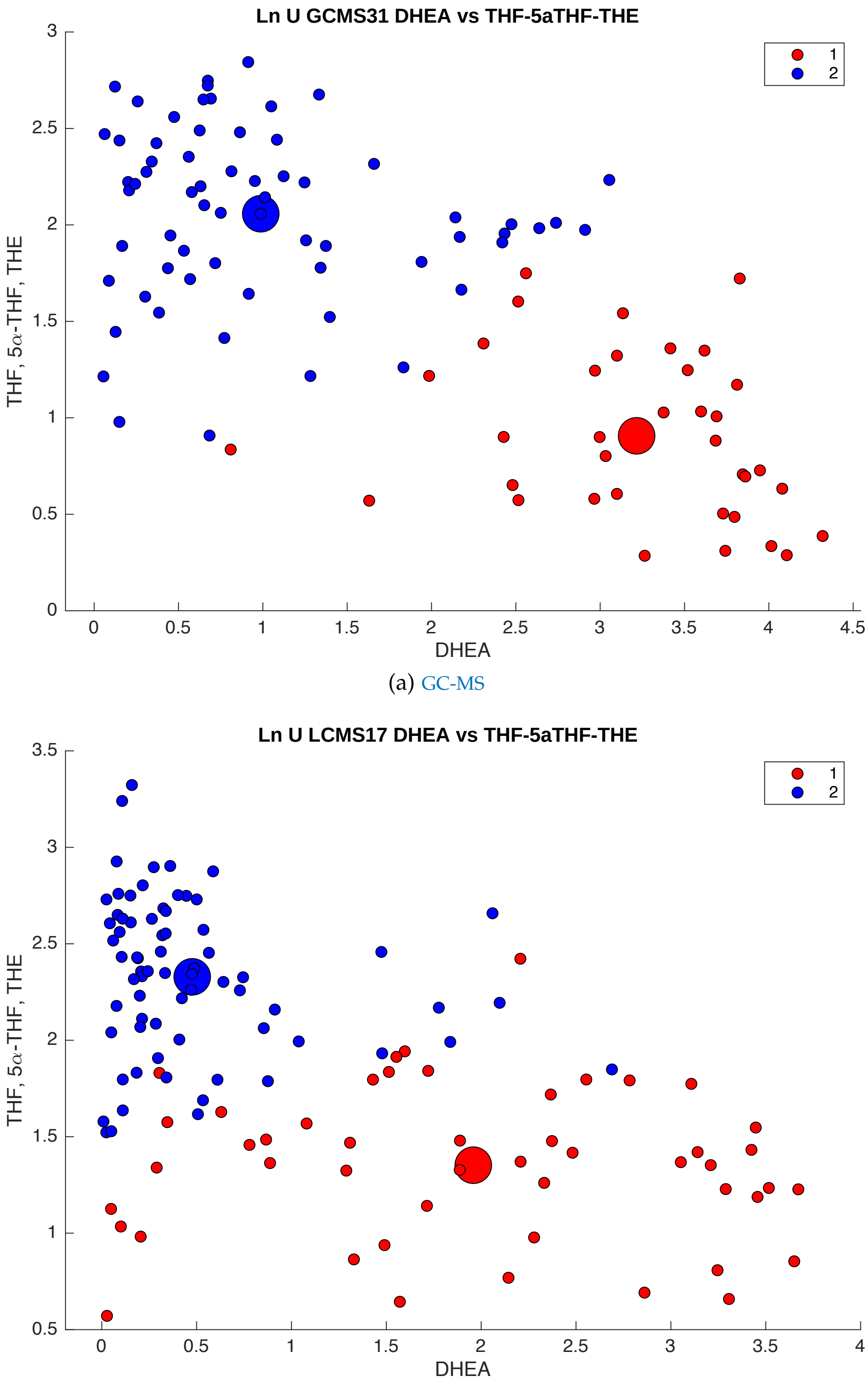
Therefore, after careful consideration and consulting endocrinology experts, we decided to primarily focus on three clusters for the remainder of this thesis.



(a) GC-MS

(b) LC-MS/MS

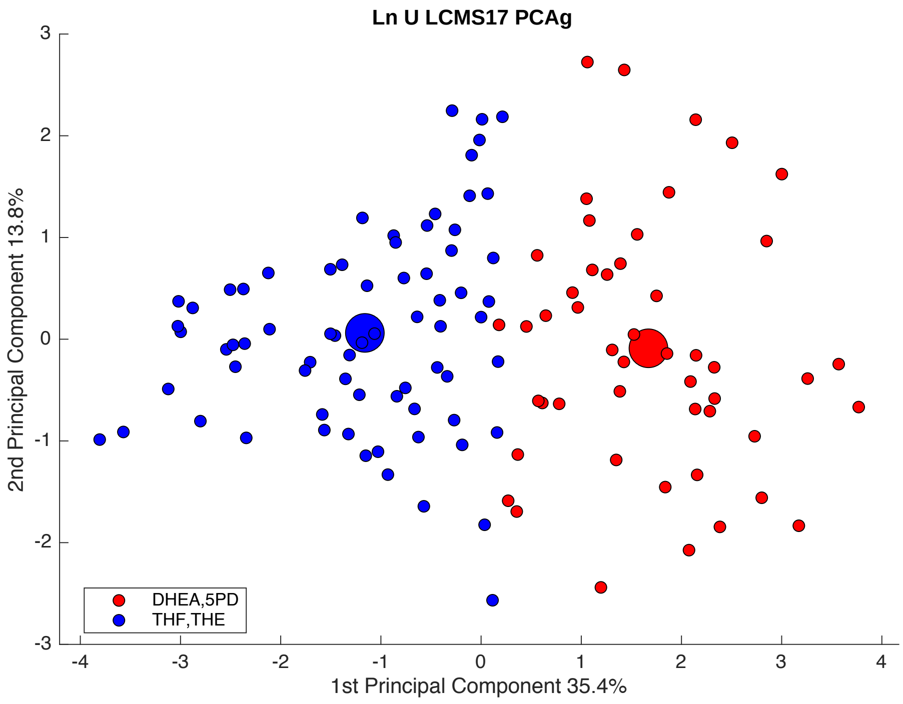
Figure 2.5: Elbow method vs Davies-Bouldin



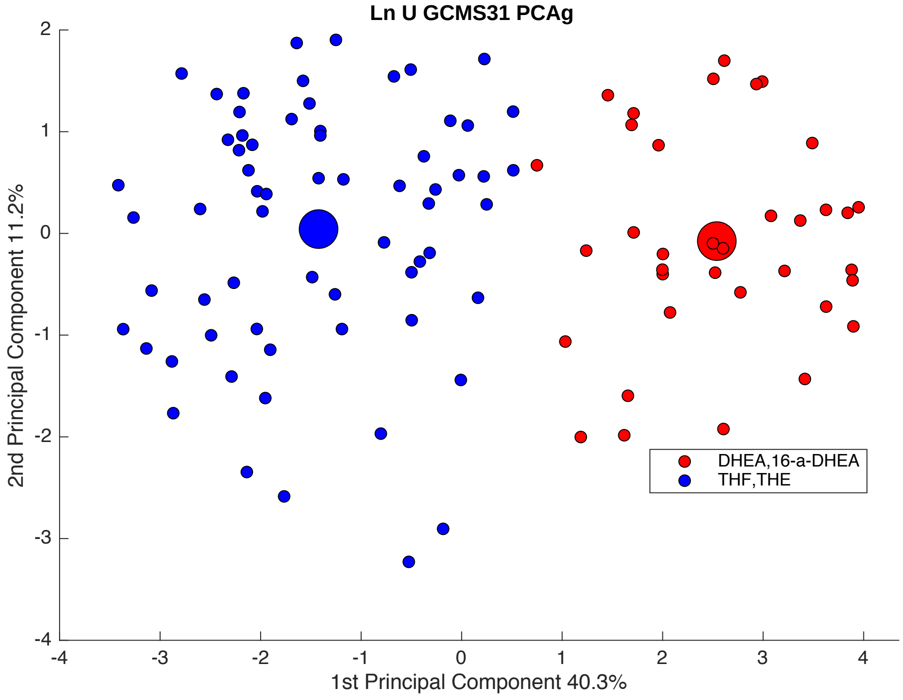
(a) GC-MS

(b) LC-MS/MS

Figure 2.6: Custom clustering view of DHEA vs THF-5α-THF-THE. K-means for 2 clusters.

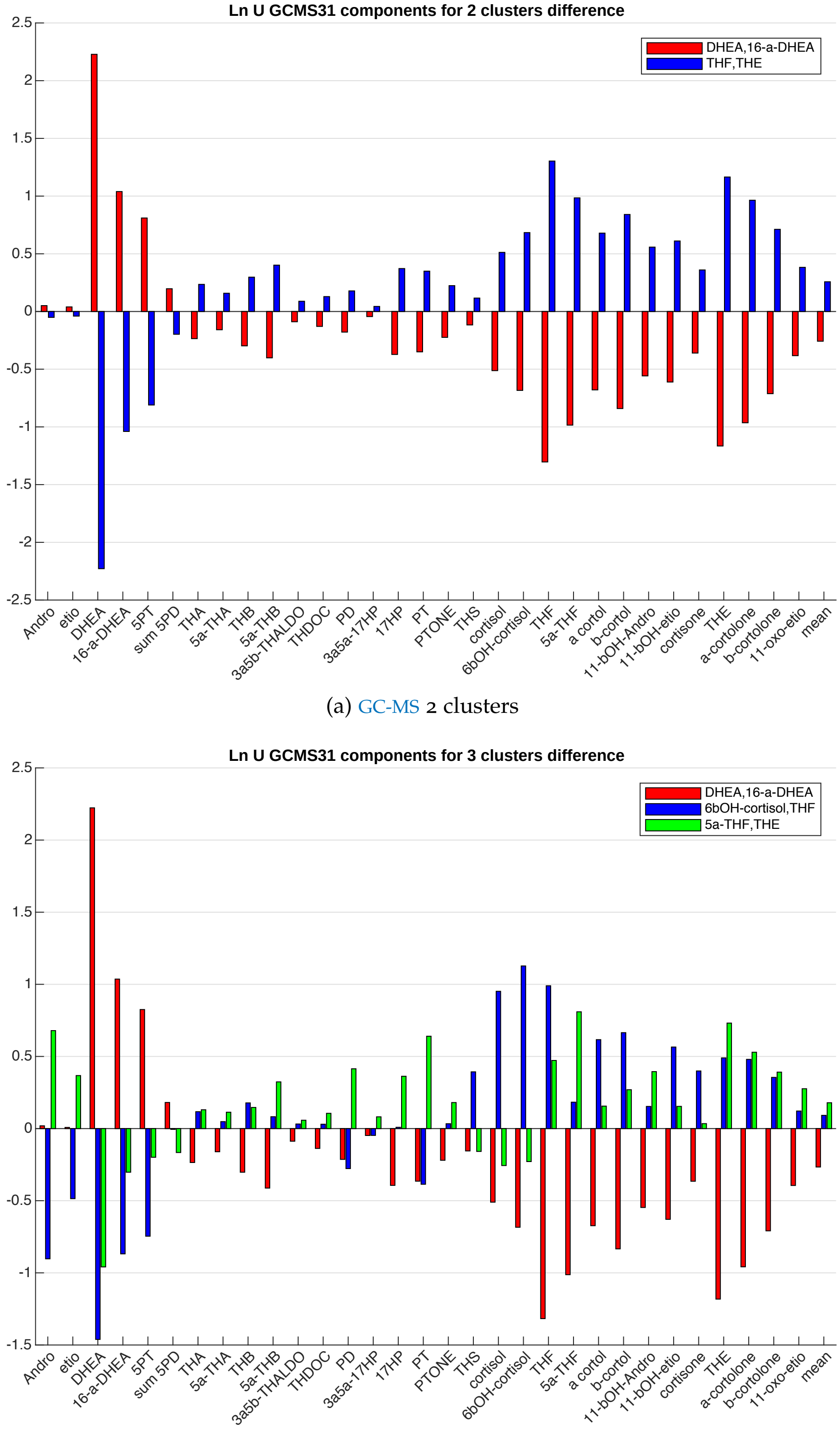


(a) GC-MS



(b) LC-MS/MS

Figure 2.7: PCA projection for 2 clusters k-means. The multidimensional data is projected on the first two principal components.



(a) GC-MS 2 clusters

(b) GC-MS 3 clusters

Figure 2.8: Cluster Characteristics comparison 2 versus 3 clusters.

2.4 UNSUPERVISED CLUSTERING - RESULTS PER METHOD

Here we present an overview of results obtained by using different methods. For each method, we show relevant figures accompanied by a brief explanation.

2.4.1 *Self-organising Map (SOM)*

In figure 2.9 we see the U-matrix (unified distance matrix [37]) of the SOM. The U-matrix is a greyscale image visualising the distance between neurons. In the image, the brightness values of the neurons correspond to the average distance to its neighbours. Lighter shades signify close distances whereas darker shades represent larger distances.

On top of the U-matrix, a hit-map of the samples is projected post-hoc, the size of the patches is relative to the number of hits. The neurons of the lattice are clustered using a k-means algorithm, resulting in three clusters. These are visualised in figure 2.10 (PCA projection), figure 2.11 (heat map), figure 2.12 (cluster characteristics), and figure 2.13 (box chart).

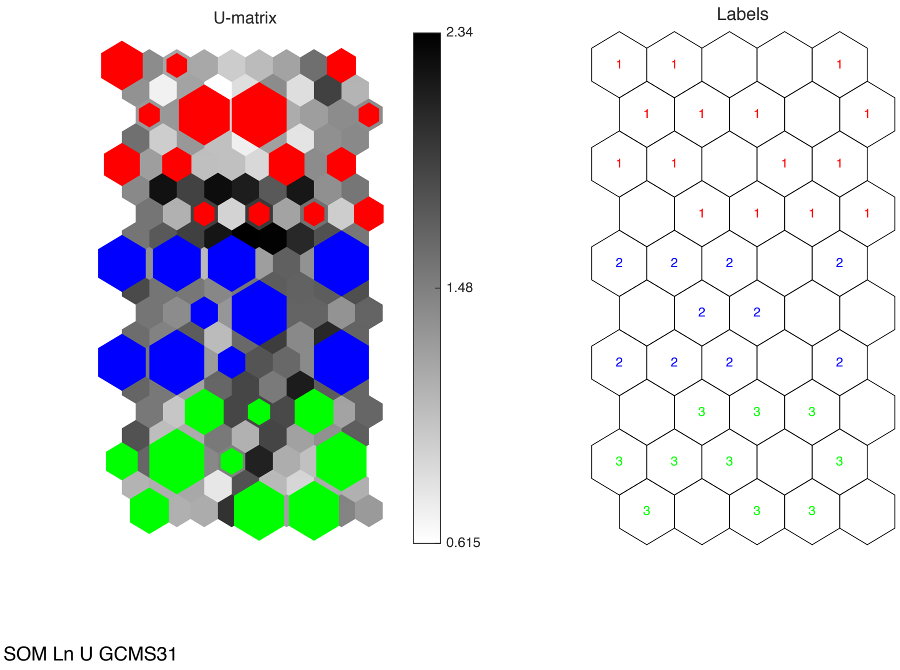
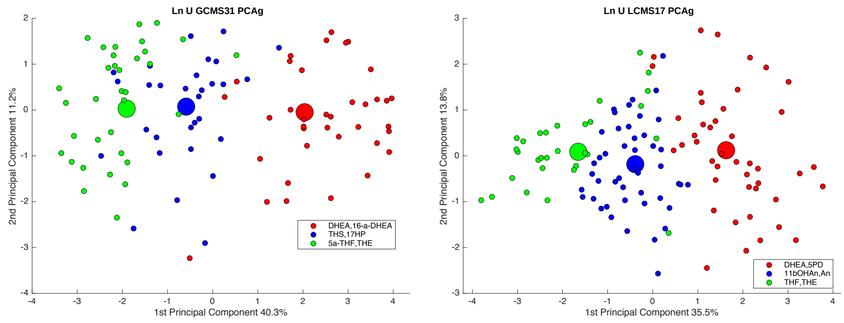


Figure 2.9: SOM k-means lattice U-matrix with hit-map overlay and labels per node

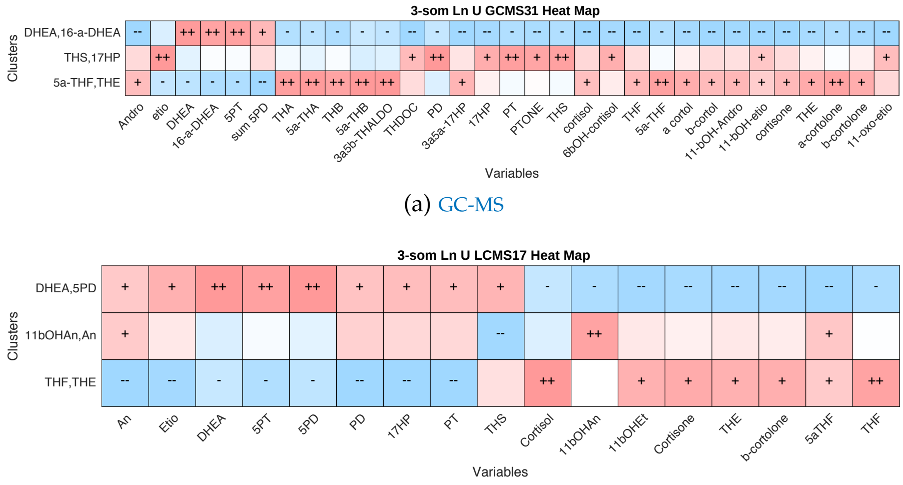
2.4.2 *Hierarchical Clustering*

The dendrograms which are typical for Hierarchical Clustering (HC) can be seen in figure 2.14. Further characteristics of the clusterings



(a) GC-MS (b) LC-MS/MS

Figure 2.10: Self-organising Map PCA projection. Colouring post-hoc according to k-means.



3-som Ln U GCMS31 Heat Map

Clusters	Andro	etio	DHEA	16-a-DHEA	5PT	sum 5PD	THA	5a-THA	THB	5a-THB	3a5b-THALDO	THDOC	PD	3a5a-17HP	17HP	PT	PTONE	THS	cortisol	6bOH-cortisol	THF	5a-THF	a cortol	b-cortol	11-bOH-Andro	11-bOH-etio	cortisone	THE	a-cortolone	b-cortolone	11-oxo-etio
DHEA,16-a-DHEA	--		++	++	++	+	-	-	-	-	-	--	-	--	--	-	--	-	--	--	--	-	--	--	--	--	--	--	-	--	--
THS,17HP		++										+	++		+	++	+	++		+						+					+
5a-THF,THE	+	-	-	-	-	--	++	++	++	++	++			+					+		+	++	+	+	+	+	+	+	++	+	

Variables

(a) GC-MS

3-som Ln U LCMS17 Heat Map

Clusters	An	Etio	DHEA	5PT	5PD	PD	17HP	PT	THS	Cortisol	11bOHAn	11bOHEt	Cortisone	THE	b-cortolone	5aTHF	THF
DHEA,5PD	+	+	++	++	++	+	+	+	+	-	-	--	--	--	--	--	-
11bOHAn,An	+								--		++					+	
THF,THE	--	--	-	-	-	--	--	--		++		+	+	+	+	+	++

Variables

(b) LC-MS/MS

Figure 2.11: Self-organising Map - Heat Maps

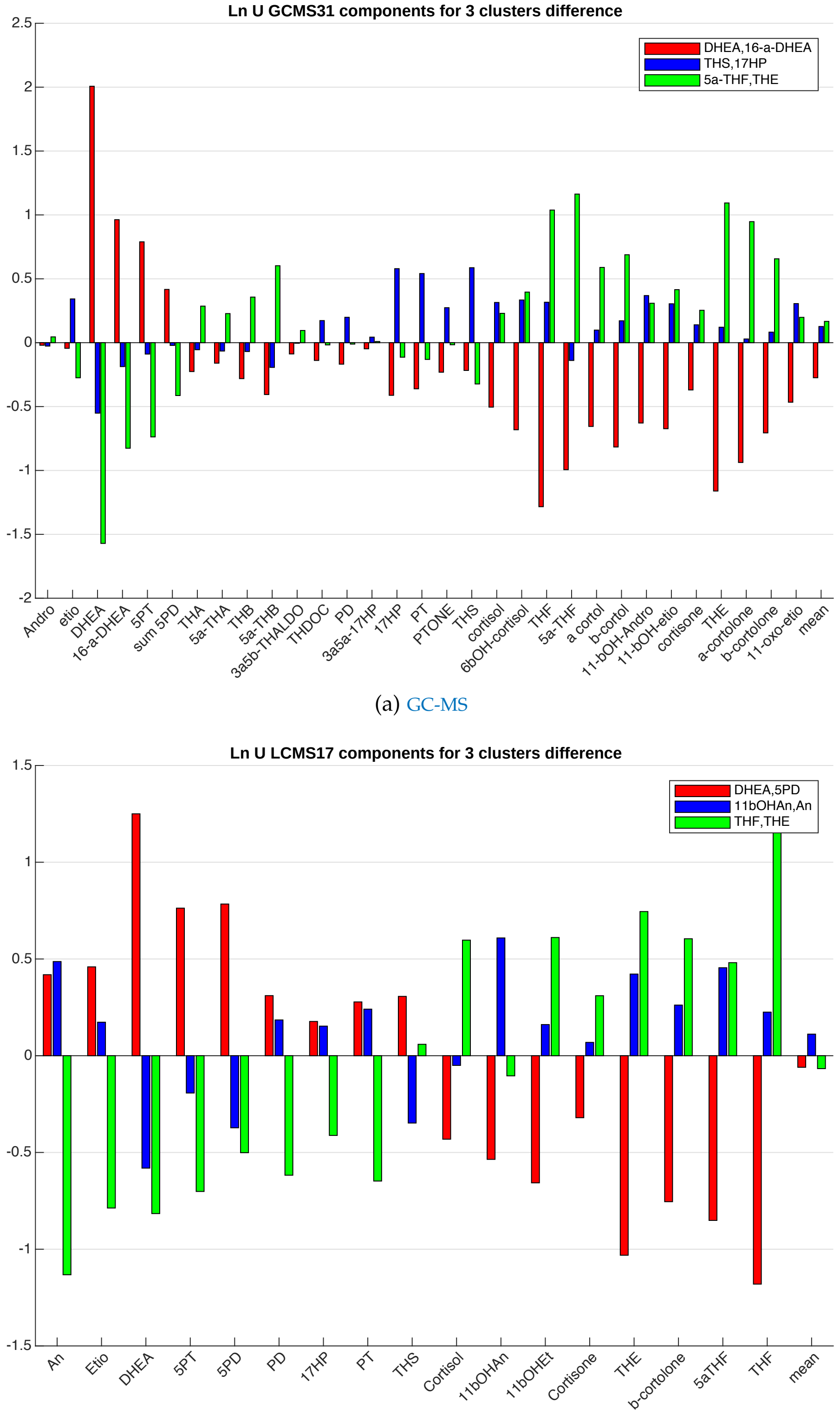
are visualised in figure 2.15 (PCA projection), figure 2.16 (heat map), figure 2.17 (cluster characteristics), and figure 2.18 (box chart). While at first glance this clustering seems fine, the cluster sizes are imbalanced and the projection separations lacking.

2.4.3 *K-medoids*

The clusterings for k-medoids are visualised in figure 2.19 (PCA projection), figure 2.20 (heat map), figure 2.21 (cluster characteristics), and figure 2.22 (box chart). Overall, it looks good; however, we see poor separation of the blue and green clusters in the PCA projection.

2.4.4 *K-means*

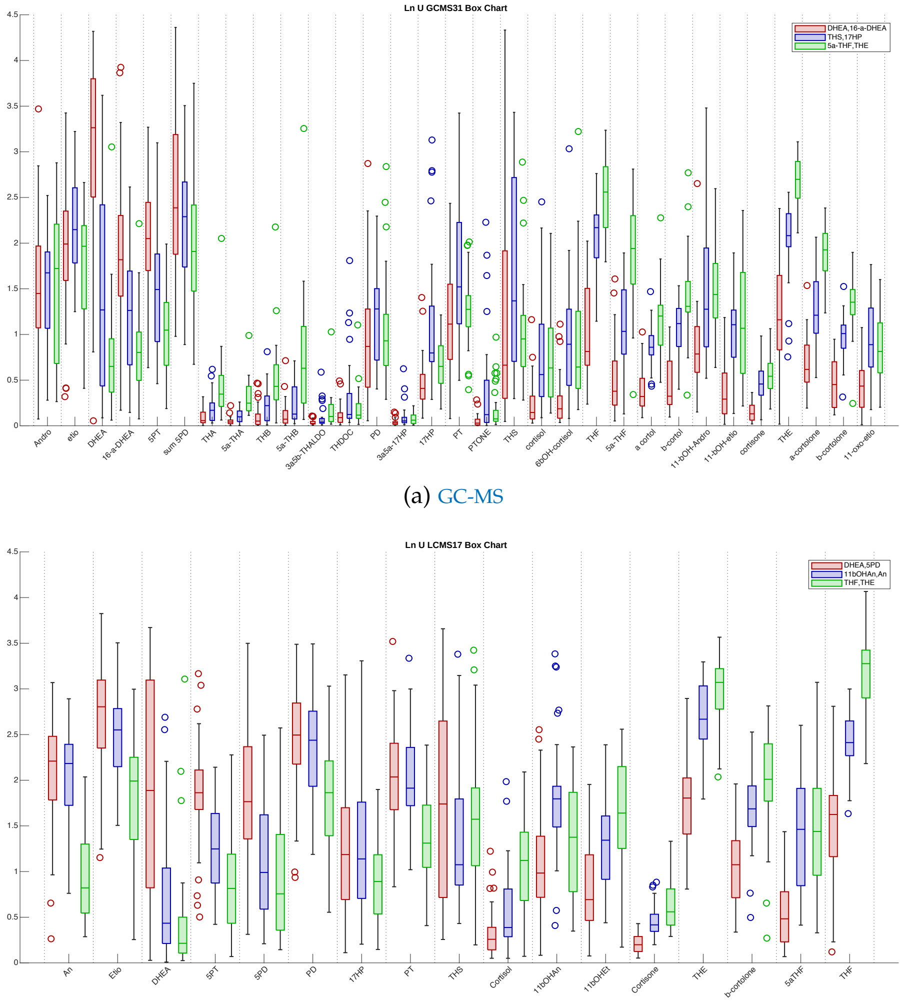
The clusterings for k-means are visualised in figure 2.23 (PCA projection), figure 2.24 (heat map), figure 2.25 (cluster characteristics), and figure 2.26 (box chart). We see good separation in the PCA projection. The LC-MS/MS does have a different 'opinion' how to divide the green/blue clusters. The heat map as well as the cluster difference chart give a good view of the centroids.



(a) GC-MS

(b) LC-MS/MS

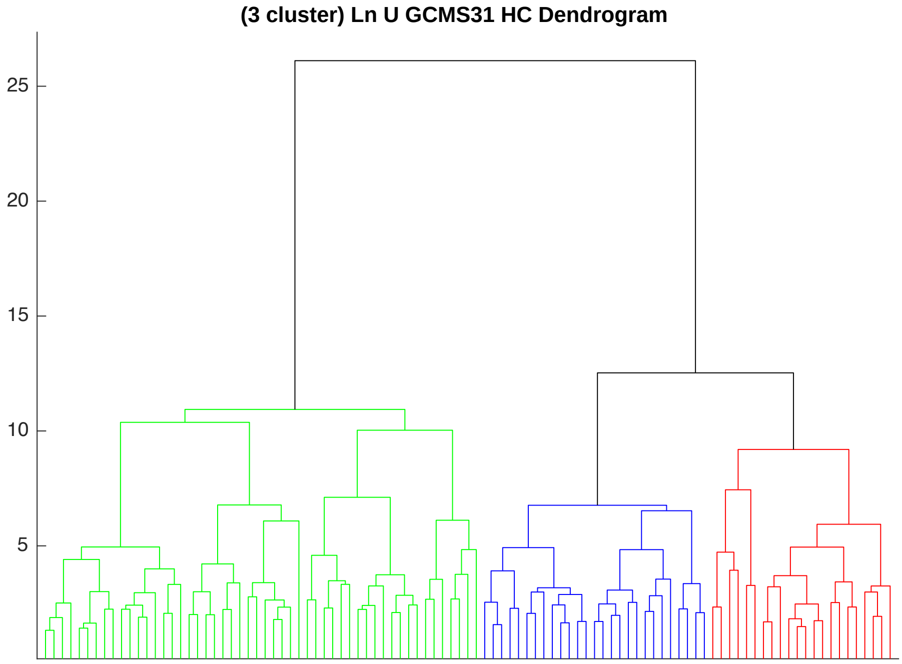
Figure 2.12: Self-organising Map Cluster Characteristics



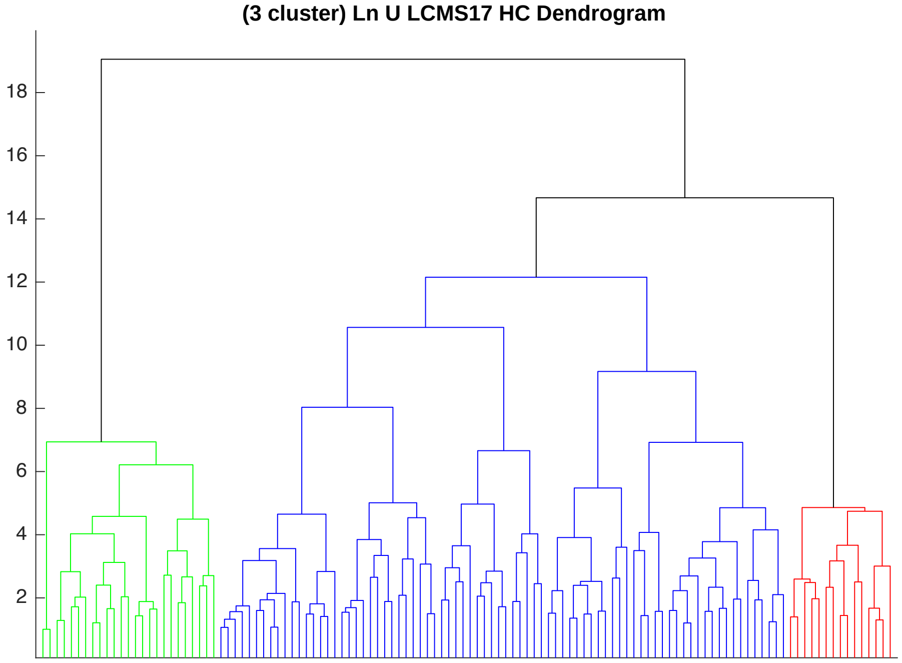
(a) GC-MS

(b) LC-MS/MS

Figure 2.13: Self-organising Map Box Plots

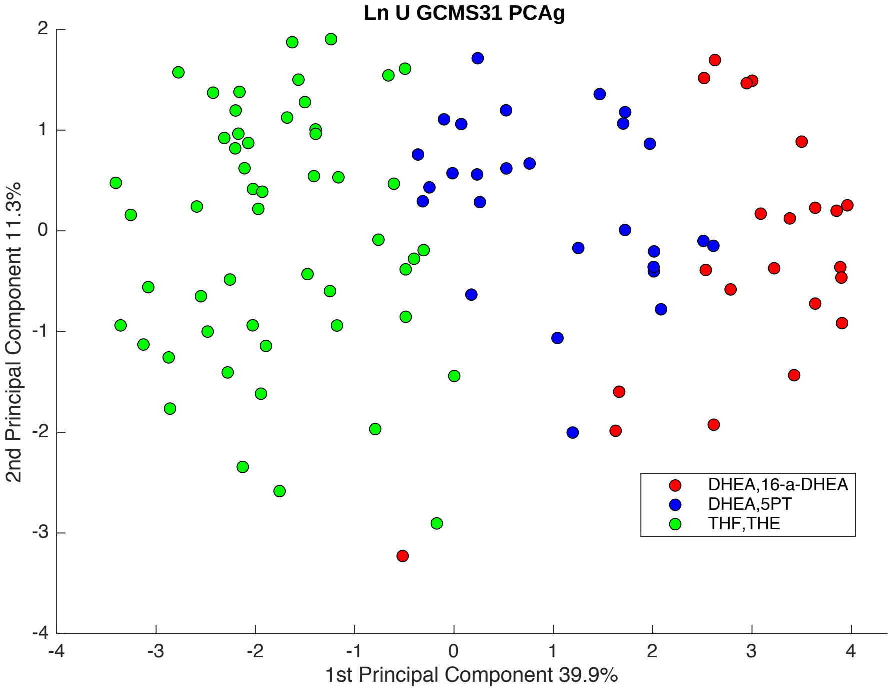


(a) GC-MS dendrogram. Left to right: THF,THE (52); DHEA,5PT (27); DHEA,16-a-DHEA (22).

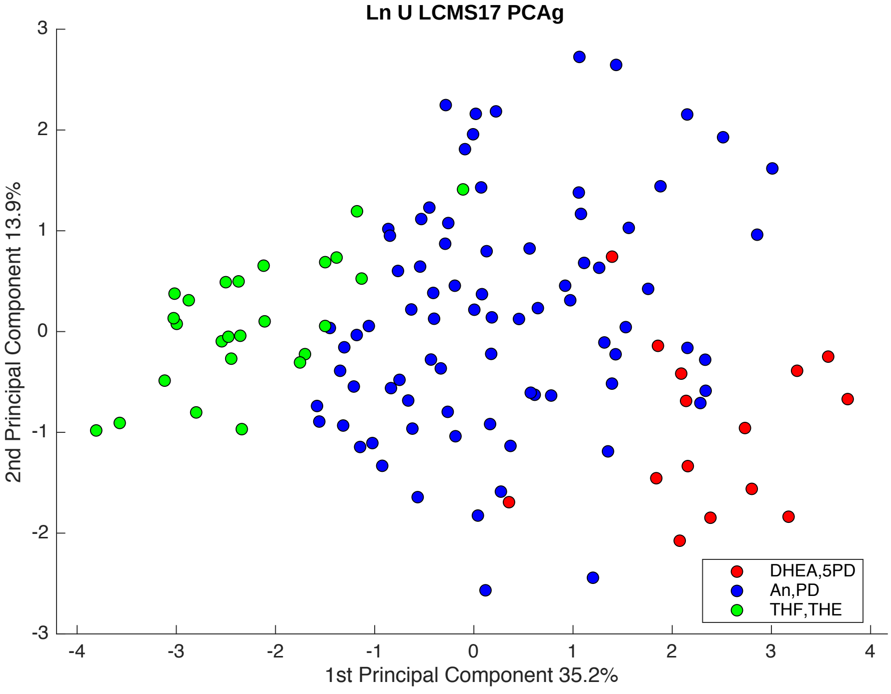


(b) LC-MS/MS dendrogram. Left to right: THF,THE (25); An,PD (80); DHEA,5PD (15)

Figure 2.14: Hierarchical Clustering Dendrograms.

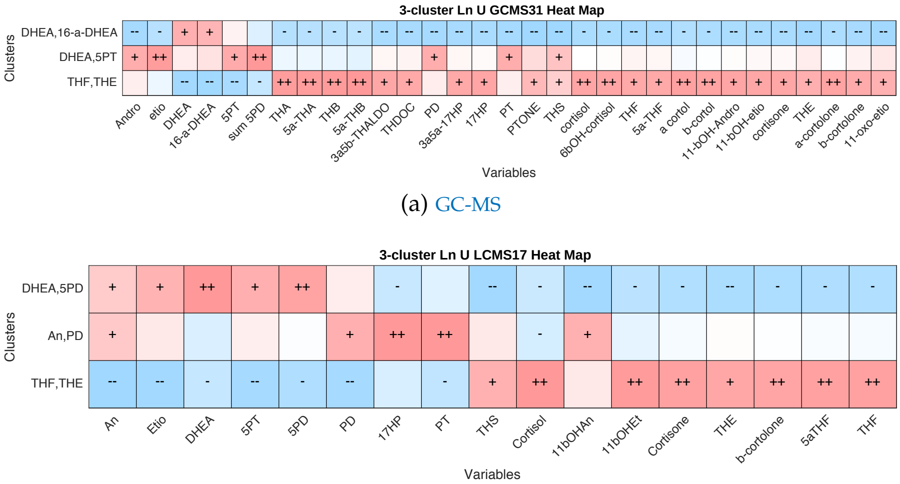


(a) GC-MS



(b) LC-MS/MS

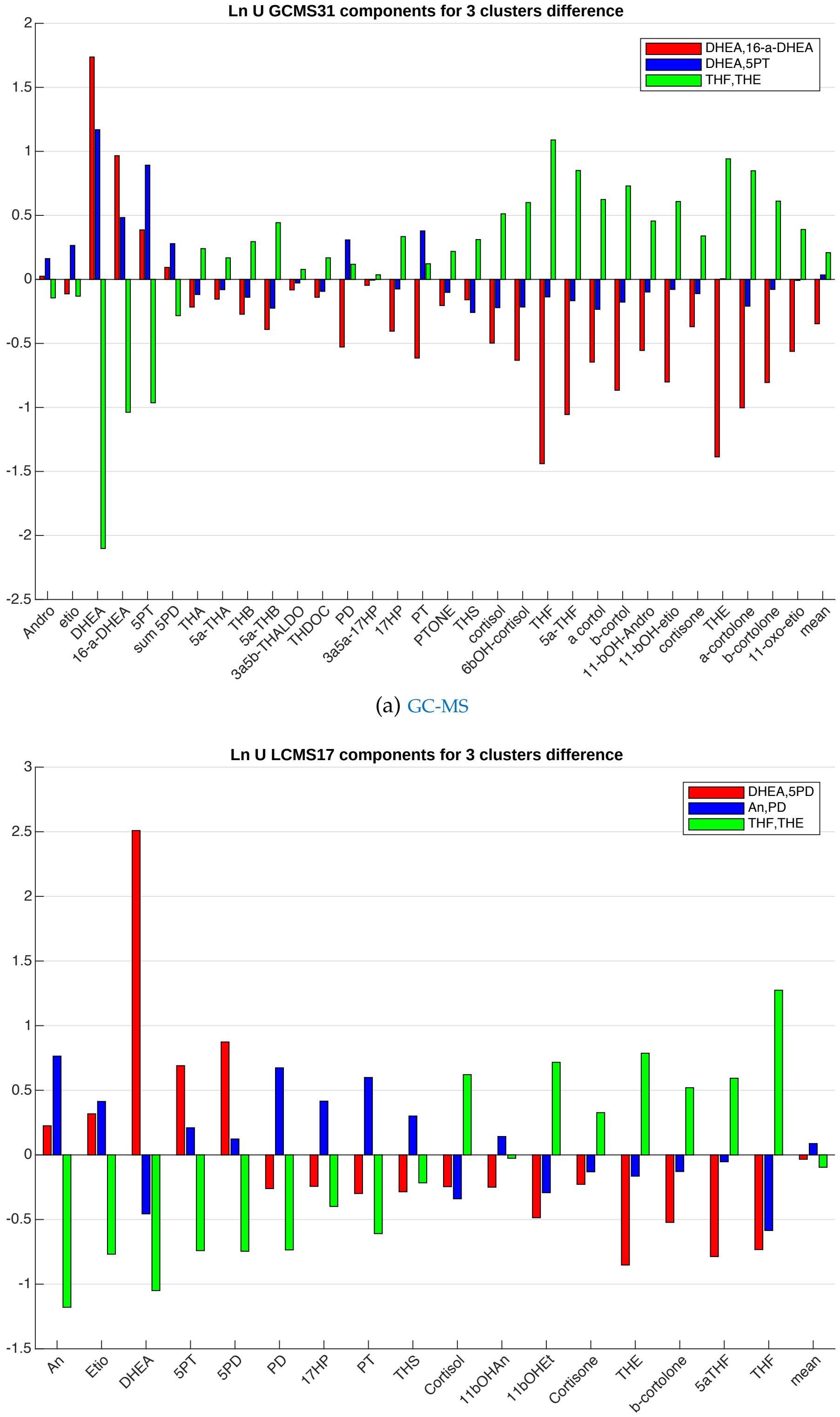
Figure 2.15: Hierarchical Clustering PCA projection. Colouring post-hoc according to hierarchical clustering.



(a) GC-MS

(b) LC-MS/MS

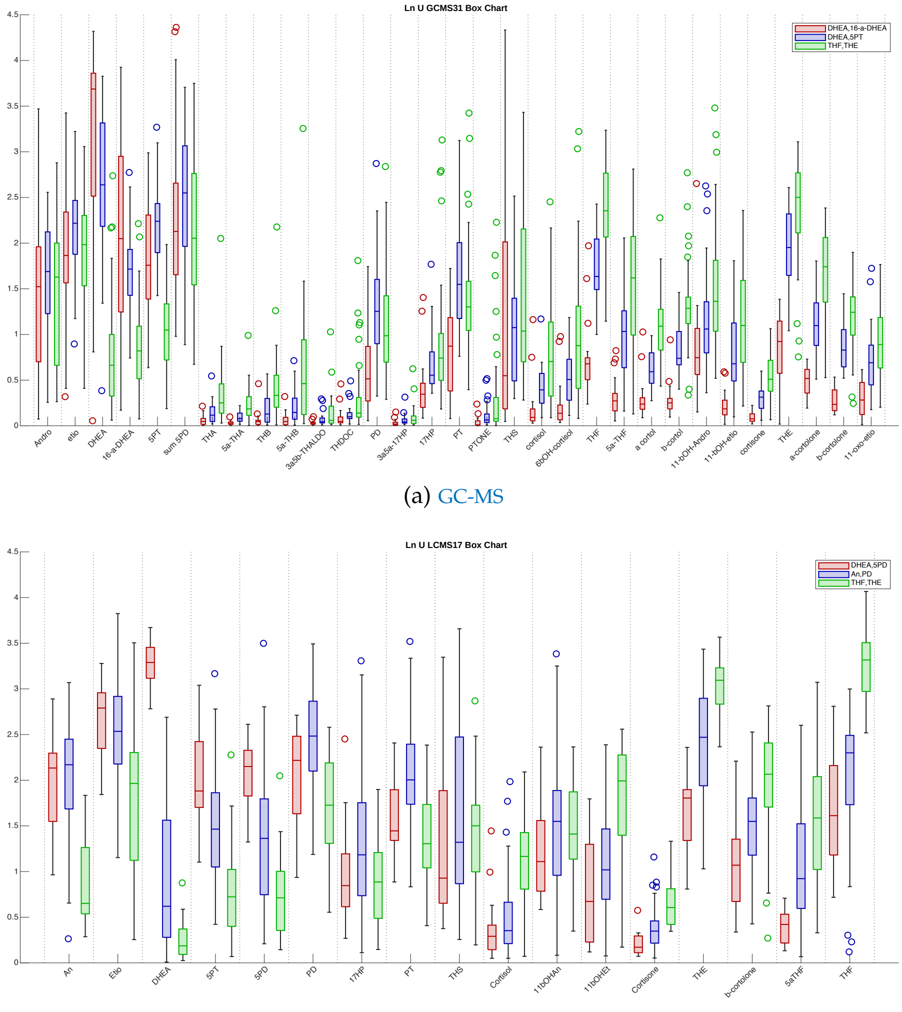
Figure 2.16: Hierarchical Clustering - Heat Maps.



(a) GC-MS

(b) LC-MS/MS

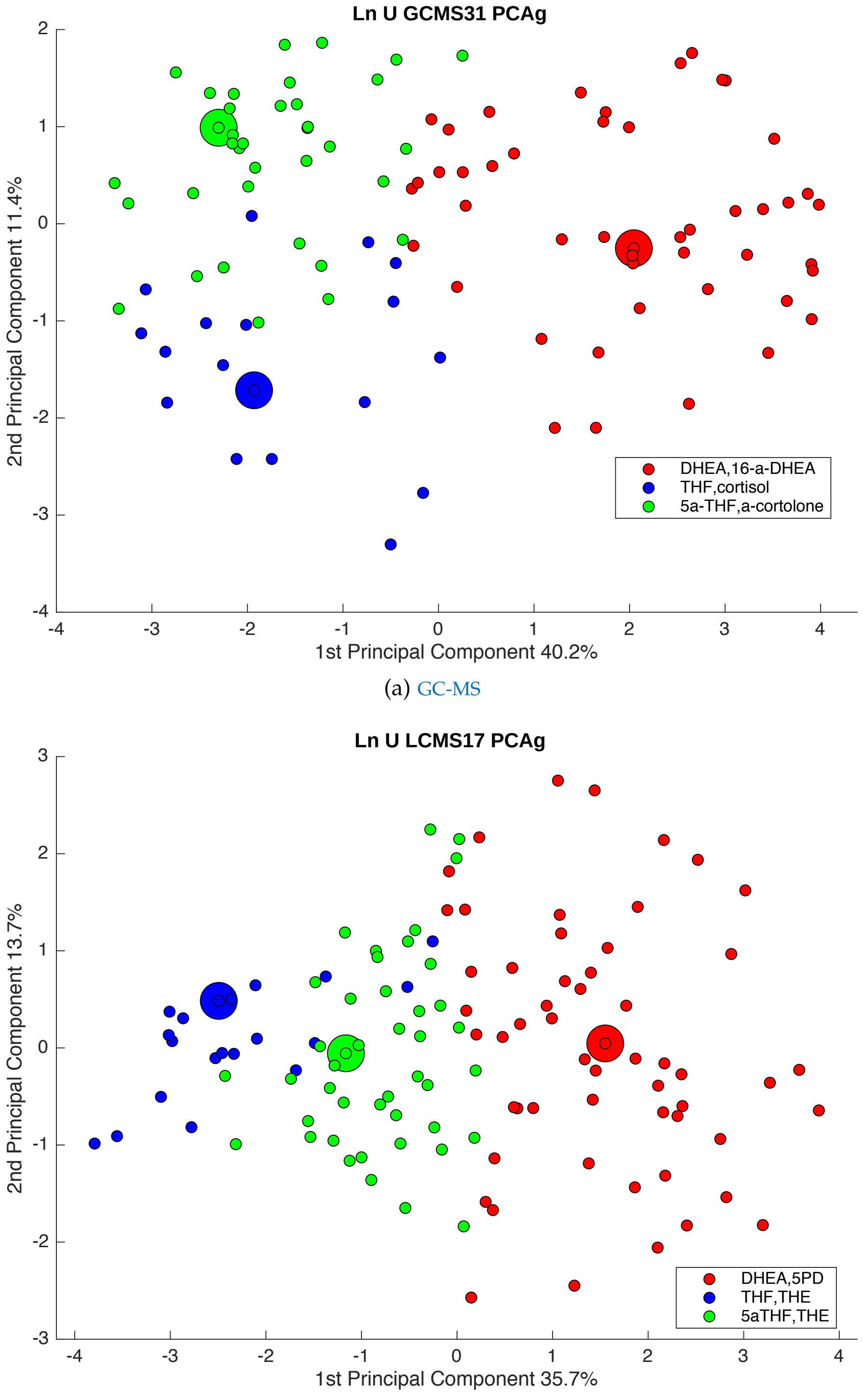
Figure 2.17: Hierarchical Clustering Cluster Characteristics



(a) GC-MS

(b) LC-MS/MS

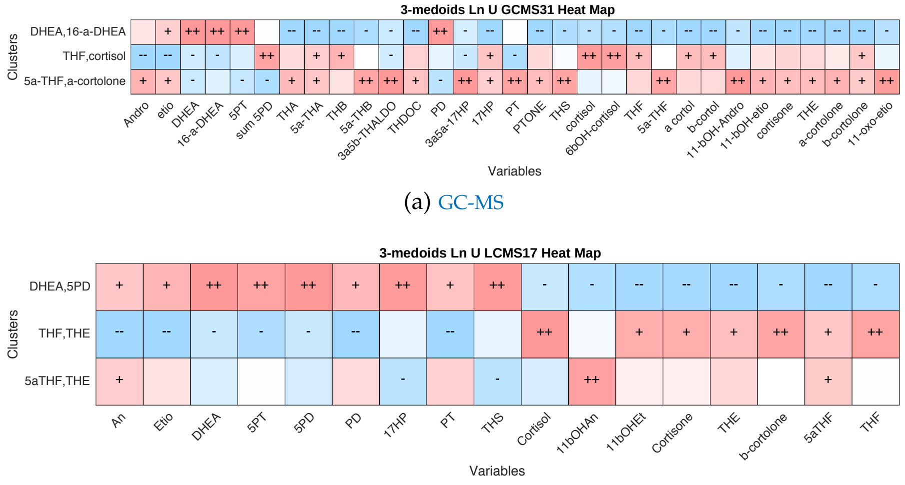
Figure 2.18: Hierarchical Clustering Box Plots



(a) GC-MS

(b) LC-MS/MS

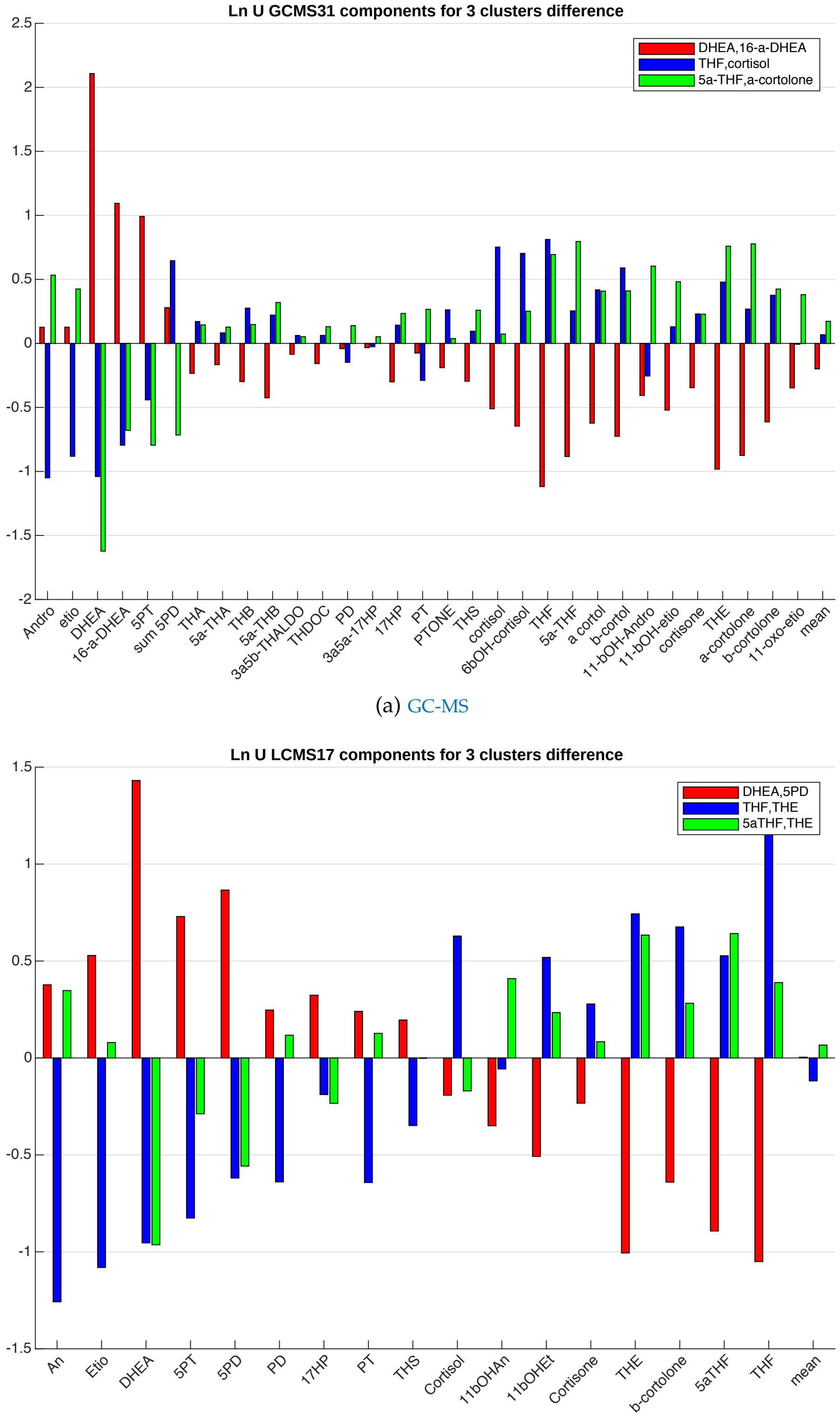
Figure 2.19: K-Medoids PCA projection. Colouring post-hoc according to hierarchical clustering.



(a) GC-MS

(b) LC-MS/MS

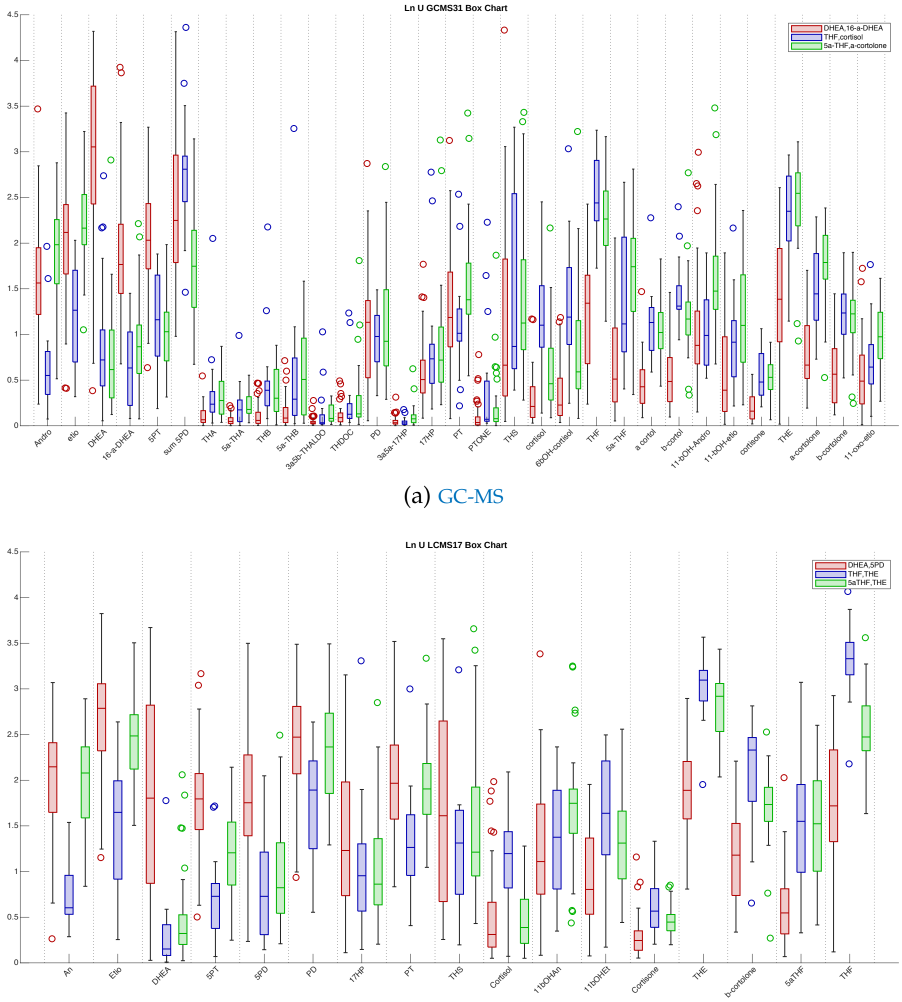
Figure 2.20: K-Medoids - Heat Maps.



(a) GC-MS

(b) LC-MS/MS

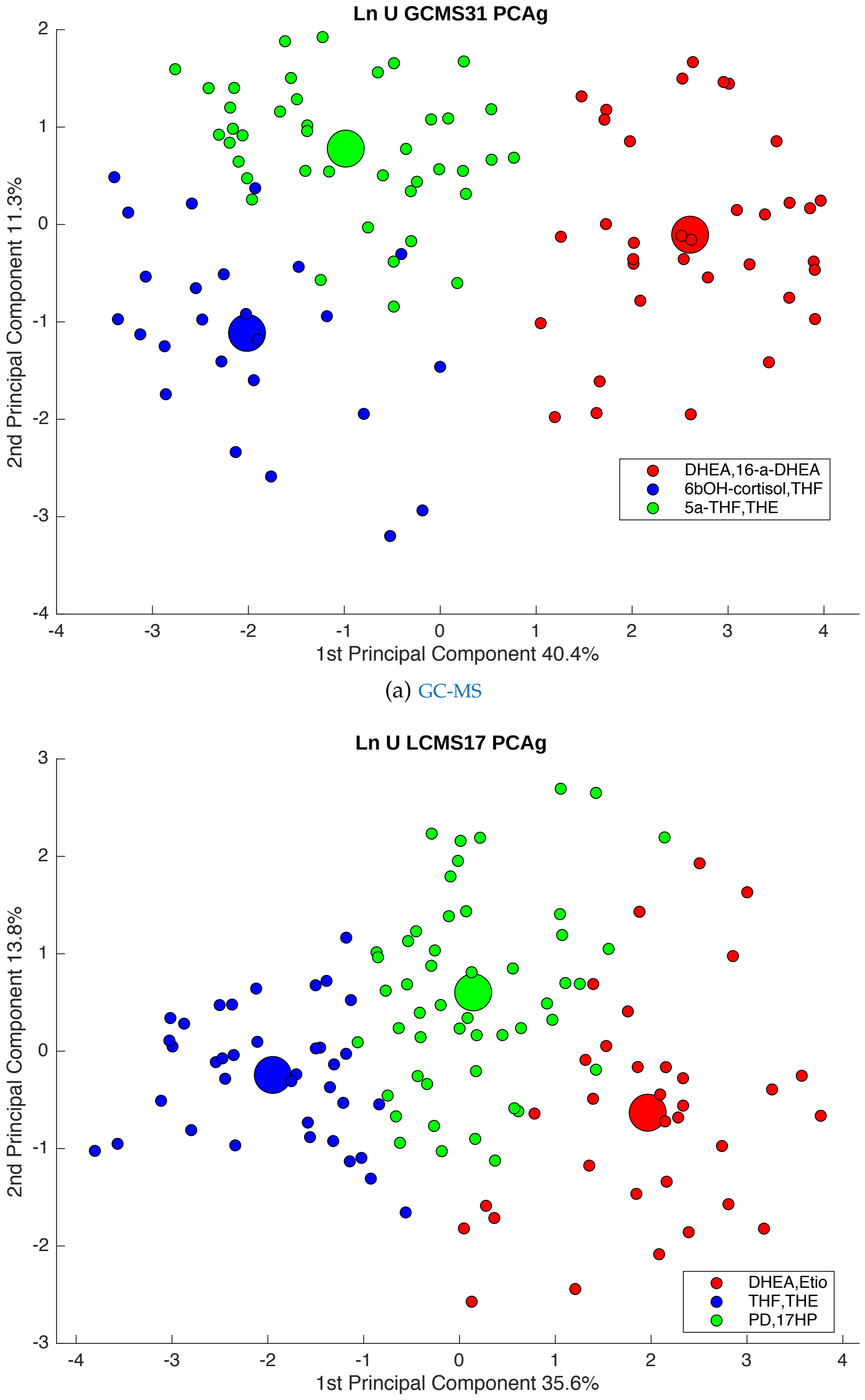
Figure 2.21: K-Medoids Cluster Characteristics



(a) GC-MS

(b) LC-MS/MS

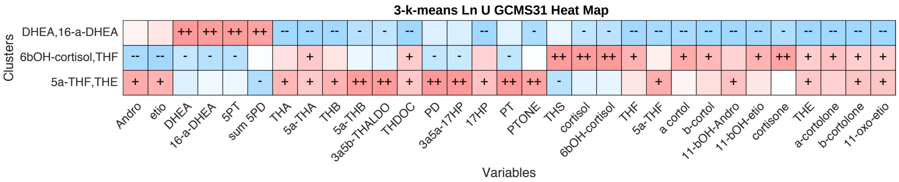
Figure 2.22: K-Medoids Box Plots.



(a) GC-MS

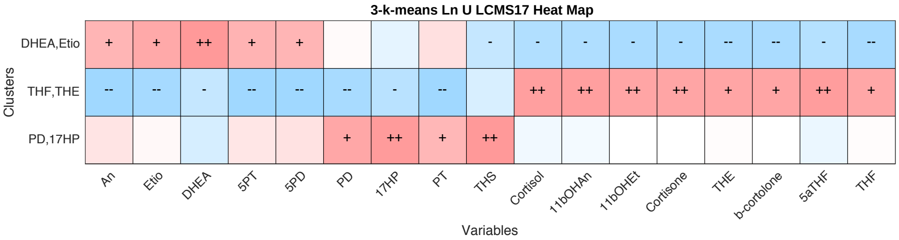
(b) LC-MS/MS

Figure 2.23: K-means PCA projection. Colouring post-hoc according to hierarchical clustering.



	Andro	etio	DHEA	16-a-DHEA	5PT	sum 5PD	THA	5a-THA	THB	5a-THB	3a5b-THALDO	THDOC	PD	3a5a-17HP	17HP	PT	PTONE	THS	cortisol	6bOH-cortisol	THF	5a-THF	a cortol	b-cortol	11-bOH-Andro	11-bOH-etio	cortisone	THE	a-cortolone	b-cortolone	11-oxo-etio
DHEA,16-a-DHEA			++	++	++	++	--	--	--	-	-	--			--		-		-	-	--	--	--	--	--	--	-	--	--	--	--
6bOH-cortisol,THF	--	--	-	-	-			+				+	-	-		-		++	++	++	+		+	+		+	++	+	+	+	+
5a-THF,THE	+	+				-	+	+	+	++	++	+	++	++	+	++	++	-				+			+			+		+	+

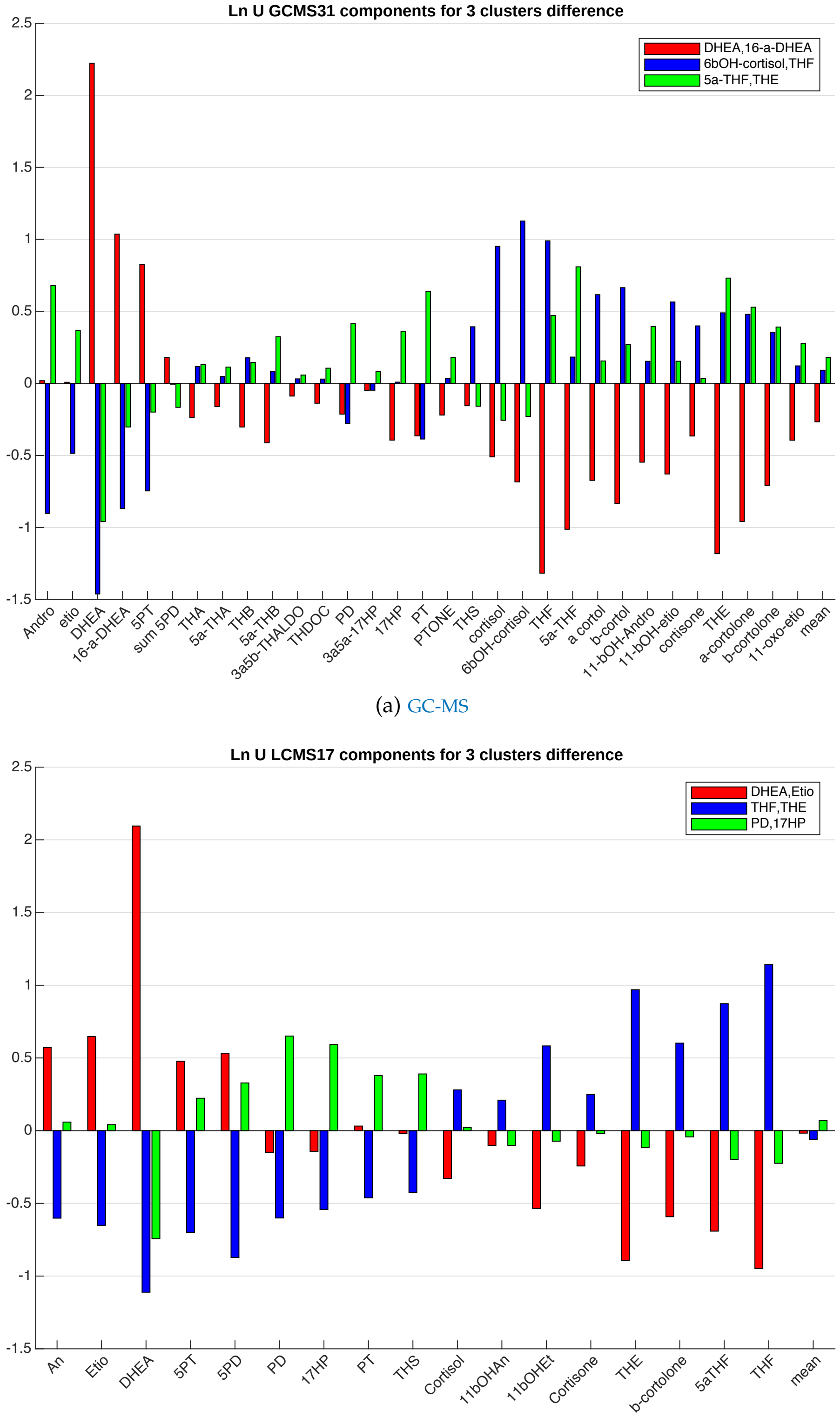
(a) GC-MS



	An	Etio	DHEA	5PT	5PD	PD	17HP	PT	THS	Cortisol	11bOHAn	11bOHEt	Cortisone	THE	b-cortolone	5aTHF	THF
DHEA,Etio	+	+	++	+	+				-	-	-	-	-	--	--	-	--
THF,THE	--	--	-	--	--	--	-	--		++	++	++	++	+	+	++	+
PD,17HP						+	++	+	++								

(b) LC-MS/MS

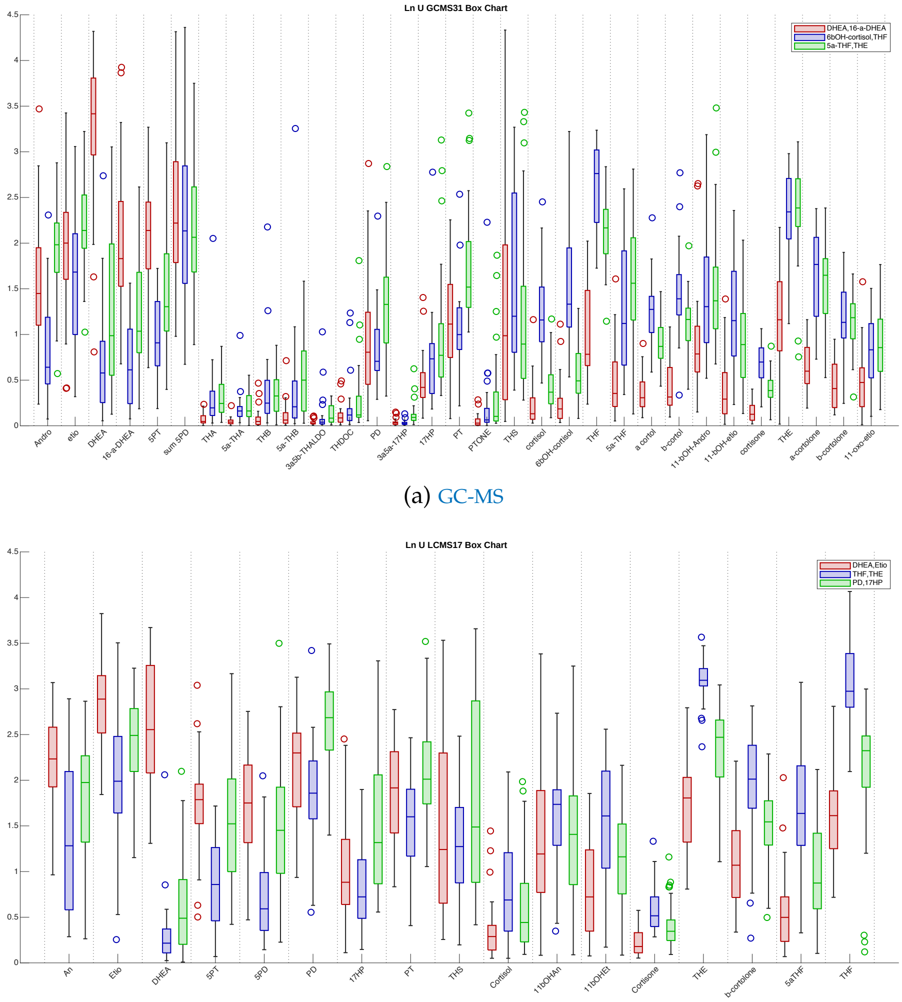
Figure 2.24: K-means - Heat Maps.



(a) GC-MS

(b) LC-MS/MS

Figure 2.25: K-means Cluster Characteristics



(a) GC-MS

(b) LC-MS/MS

Figure 2.26: K-means Box Plots.

3 DISCUSSION

In this chapter, we first discuss which algorithm provides the best clustering. Using this best method, we add domain knowledge for a phenotypical analysis of the clusters. Then we perform a GMLVQ on the data, revealing relevances we compare with our previous cluster analysis. Finally we decide whether LC-MS/MS or GC-MS gives the best clustering at this time.

3.1 UNSUPERVISED CLUSTERING - BEST CANDIDATE

When comparing the different algorithms, we see a lot of similarities between the clusters they find (see table 3.1). Cluster A (DHEA) is quite stable, with the exception of HC. Clusters B (Cortisol) and C (5a-THF) show a bit more variation. When looking at clustering similarity of the algorithms per dataset (GC vs. LC), in figures 3.1, 3.2, 3.3 and 3.4, we see that for k-means the clusters produced between GC-MS and LC-MS/MS are the most similar.

All things considered, we choose to go with k-means. It proved to be a stable clustering, has balanced cluster sizes, aligns best along the steroid groups, and, as we see in the next section, has interesting phenotypical characteristics.

ALGORITHM	CL. A		CL. B		CL. C		TOTAL	
SOM	34	97%	16	64%	21	51%	71	70%
HC	21	60%	24	96%	13	32%	58	57%
k-medoids	35	100%	14	56%	25	61%	74	73%
k-means	35	100%	25	100%	41	100%	101	100%

Table 3.1: Cluster comparison between methods. All comparisons are done against k-means for GC-MS. Clusters are ordered according to k-means, compared against the cluster that is most similar for the other algorithm.

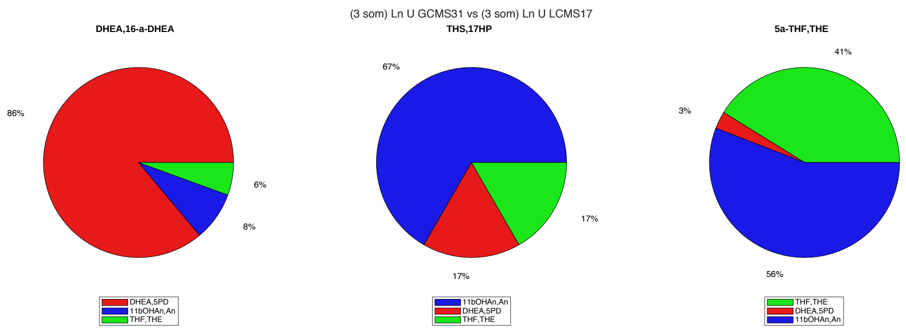


Figure 3.1: GC versus LC: SOM.

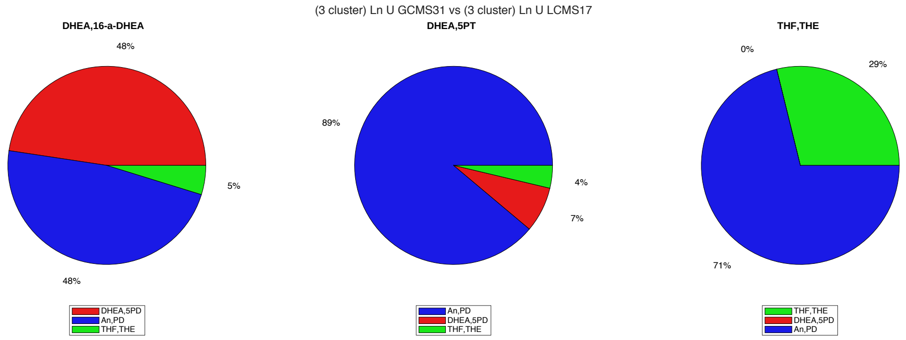


Figure 3.2: GC versus LC: Hierarchical Clustering.

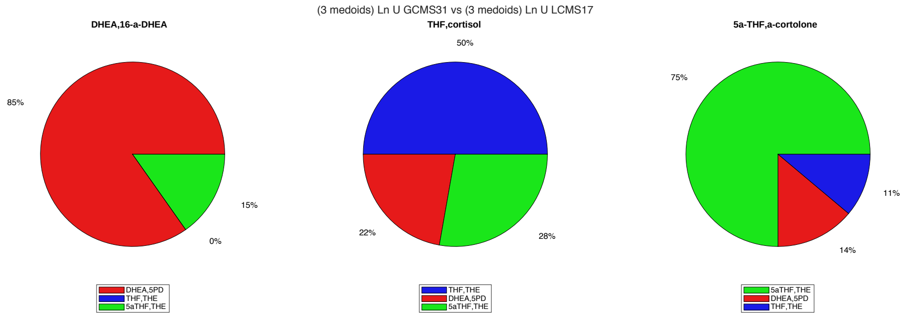


Figure 3.3: GC versus LC: k-medoids.

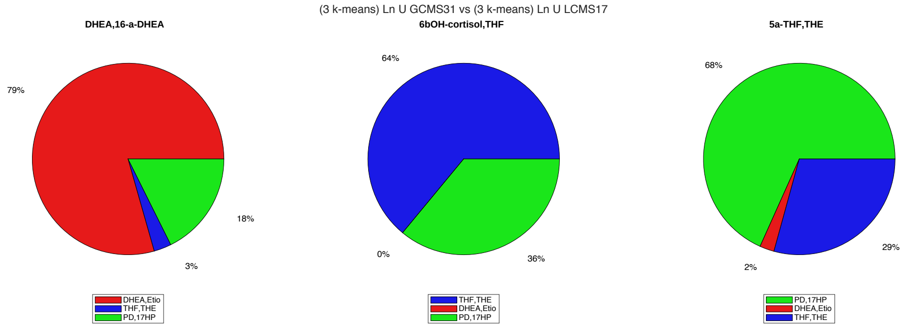


Figure 3.4: GC versus LC: k-means.

3.2 PHENOTYPICAL ANALYSIS

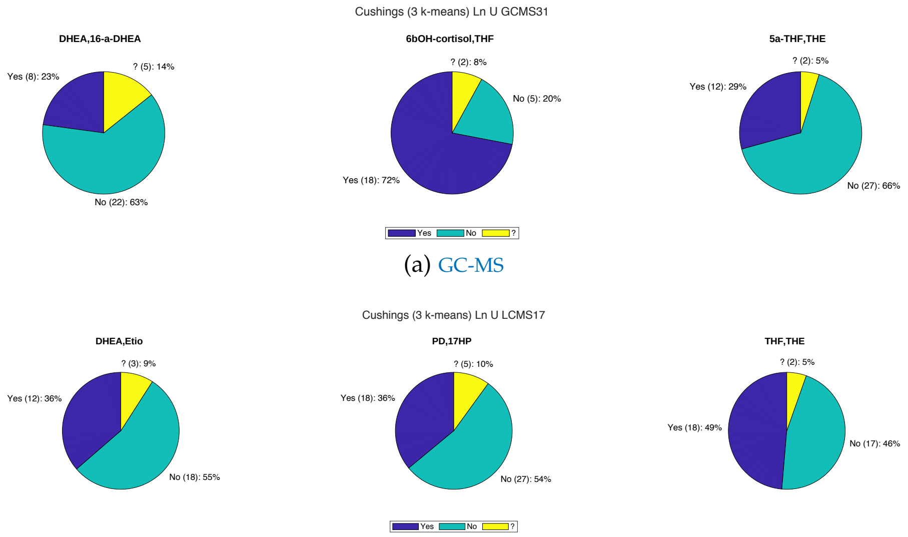
To further our insight into the clustering, we performed a brief phenotypical analysis of the clusters we found using k-means. We looked at the prevalence of Cushing's Syndrome, whether there was recurrence after R_0 resection, the KI-67 index, and distribution of patient sex. We visualise the phenotype distributions with pie charts. Each pie corresponds to one cluster. Since we do not have data available for all patients, we add a category labelled with '?' for patients we don't have data for.

Cushing's Syndrome is a rare disorder that results from prolonged and pathologic exposure to excess glucocorticoids. [38] We see in figure 3.5 that for GC-MS the hydroxycortisol-thf cluster (2) has the most patients with Cushing's syndrome, which makes sense. For LC-MS/MS this distinction is less significant.

The KI-67 antibody binds to the KI-67 protein most dominantly (though not exclusively) found in proliferating (cancerous) cells. As such, it is a method that is often used to detect and track cancer [39, 40]. In figure 3.7 we see for GC-MS an increase of the KI-67 index in the hydrocortisol-thf (2) cluster and a slight decrease in the 5a-THF-THE (3) cluster. The clustering based on LC-MS/MS does not show significant differences.

The R_0 resection recurrence describes the case where the tumour is resected in its entirety with negative margins at the microscopic level, but still the cancer returns. [41–43]. We see in figure 3.6 that, for GC-MS, we have a higher percentage of recurrence for the hydroxycortisol (2) cluster, which is striking. Note, however, that we only have a small number of known samples. Also, for the R_0-resection recurrence, the group '?' includes patients who had a worse resection, i. e., R_1 where the margin is only macroscopically free of cancer cells or R_2 leaving macroscopic (pieces of) tumours, giving them a higher chance of recurrence [42]. Therefore, the potential to obtain more patient labels for R_0-resection recurrence is low. For LC-MS/MS, again, the distinction is insignificant.

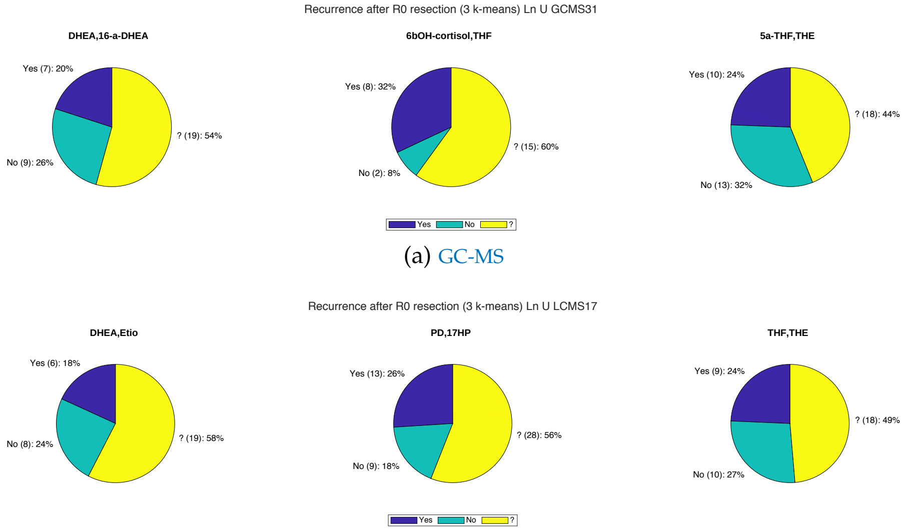
Patient sex is registered as either male or female. As we can see in figure 3.8, we have data for most of the patients. We see no significant difference between the clusters or between LC-MS/MS and GC-MS when it comes to patients sex.



(a) GC-MS

(b) LC-MS/MS

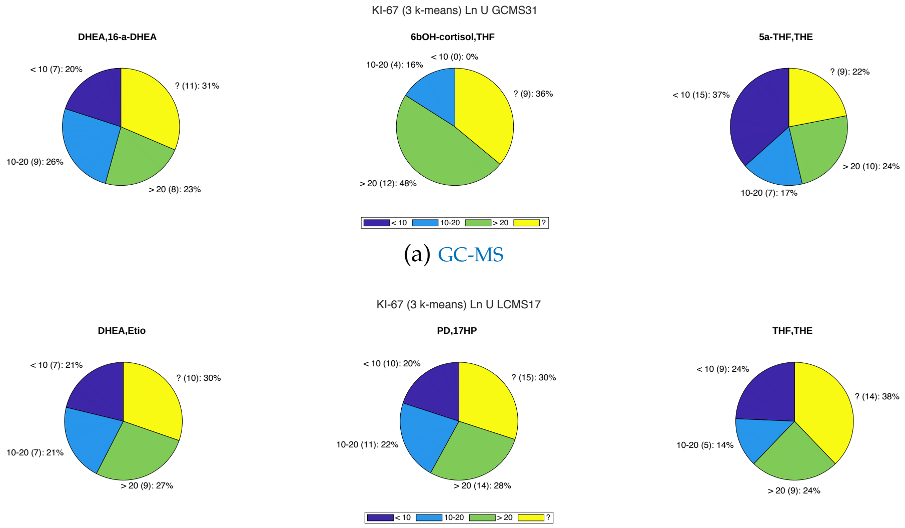
Figure 3.5: Phenotype information: Cushing's Syndrome.



(a) GC-MS

(b) LC-MS/MS

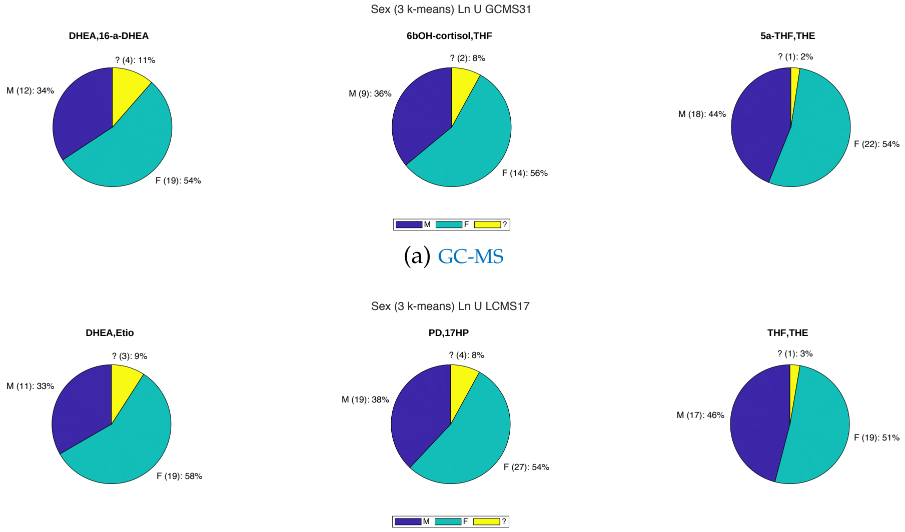
Figure 3.6: Phenotype information: Recurrence after R_0 resection.



(a) GC-MS

(b) LC-MS/MS

Figure 3.7: Phenotype information: KI-67 index.



(a) GC-MS

(b) LC-MS/MS

Figure 3.8: Phenotype information: Patient sex.

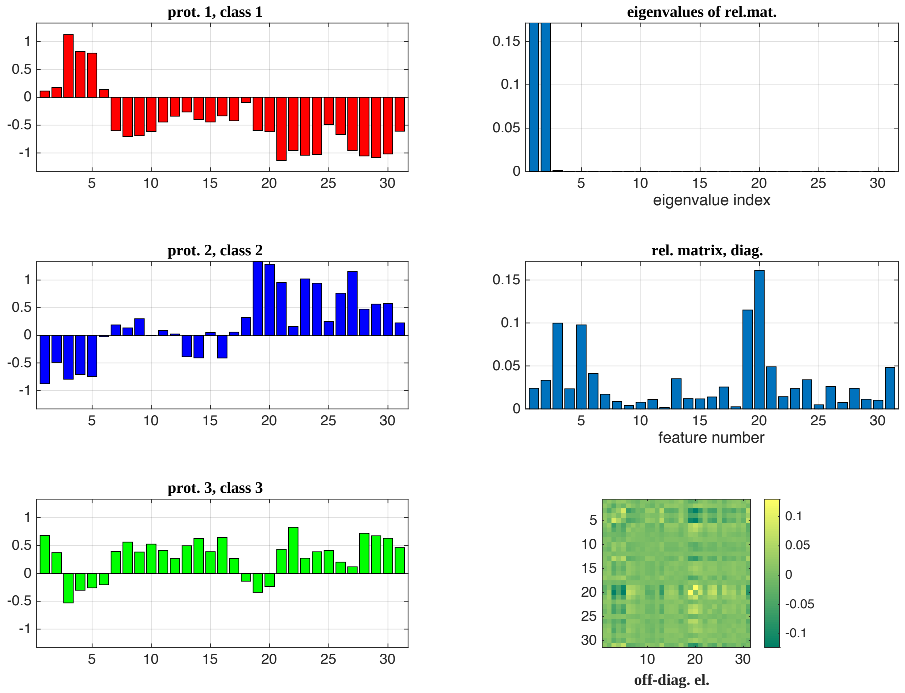


Figure 3.9: GMLVQ GC-MS prototypes and relevance matrix.

3.3 SUPERVISED CLASSIFICATION - GMLVQ

To further investigate the clusterings, we use the cluster assignment labels we found using k-means and use them to train a GMLVQ classifier [33]. Our main interest here is to look at the relevances of the steroid metabolites and not reliable classification. As such, we do not discuss performance in depth here. We did perform a quick cross-validation of 100 runs with a random 10% of the samples reserved for testing each run, the results of which can be seen in the appendix, figure A.12 and onwards.

We see the results for GC-MS in figure 3.9 (prototypes, eigenvalues, relevance matrix), figure 3.10 (2-D projection on the first two eigenvectors of Λ), and figure 3.11 (the confusion matrix). For LC-MS/MS we have the same series of figures, prototypes 3.12, data projection 3.13, and the confusion matrix 3.14. We see that the relevances are very similar to the previously calculated cluster characteristics.

3.4 LC-MS/MS OR GC-MS?

When looking at the phenotype information, the GC-MS based clustering shows significant phenotypical differences between the clusters, whereas the LC-MS/MS based clustering fails to do so. Taking into account the results of the GMLVQ, we see a better separation between the clusters and more steroid metabolites driving the separation. We are therefore going forward with GC-MS to further analyse the clustering.

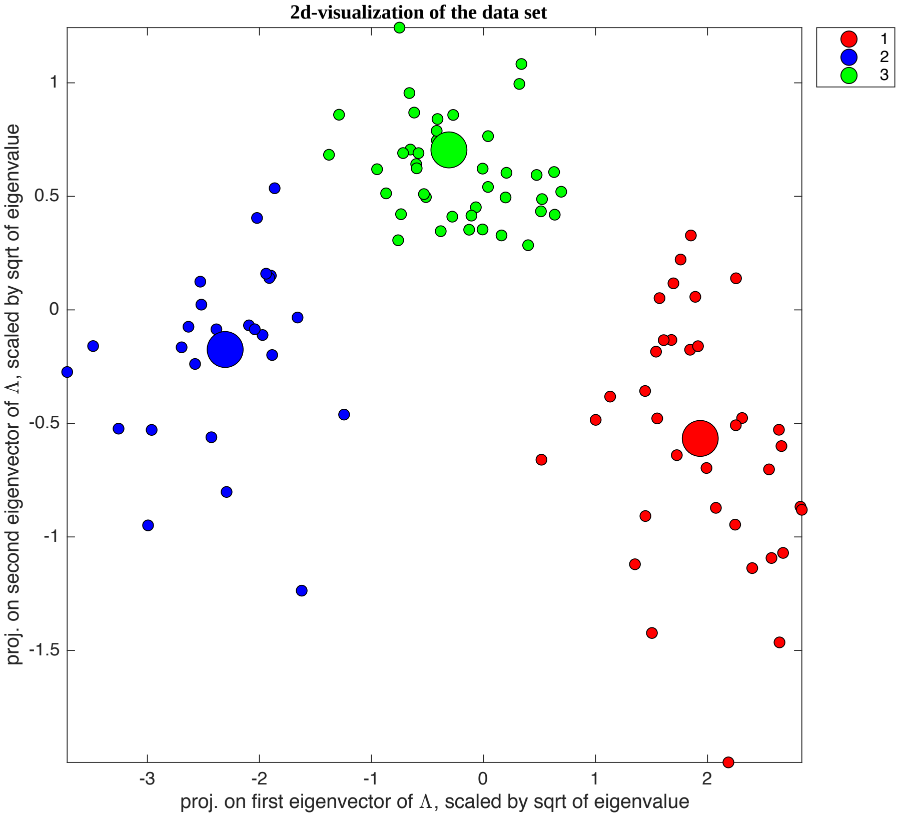


Figure 3.10: GMLVQ GC-MS Data visualisation.

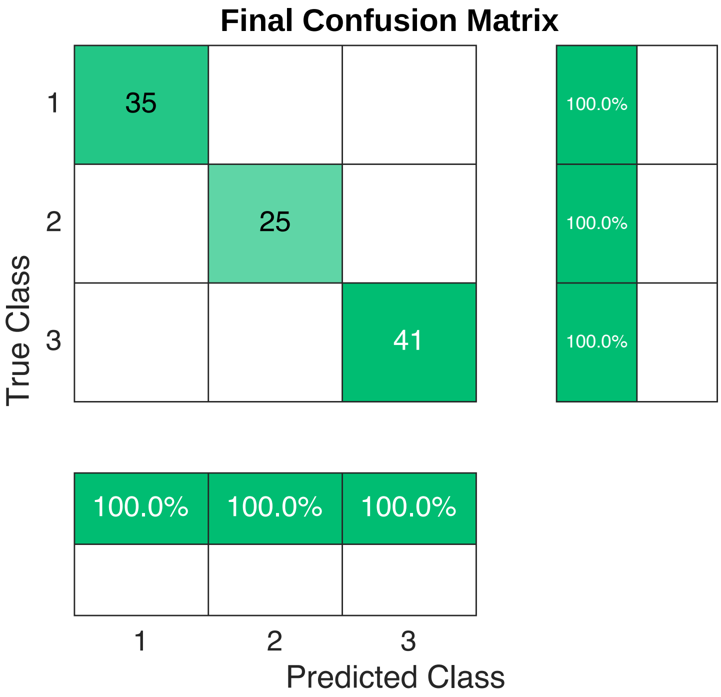


Figure 3.11: GMLVQ GC-MS Confusion matrix.

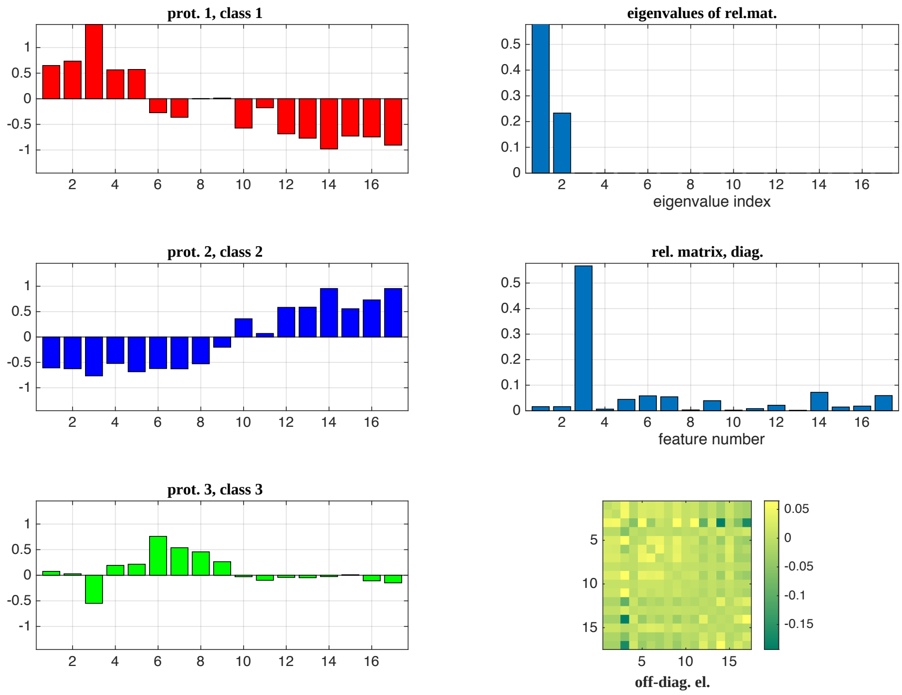


Figure 3.12: GMLVQ LC-MS/MS prototypes and relevance matrix.

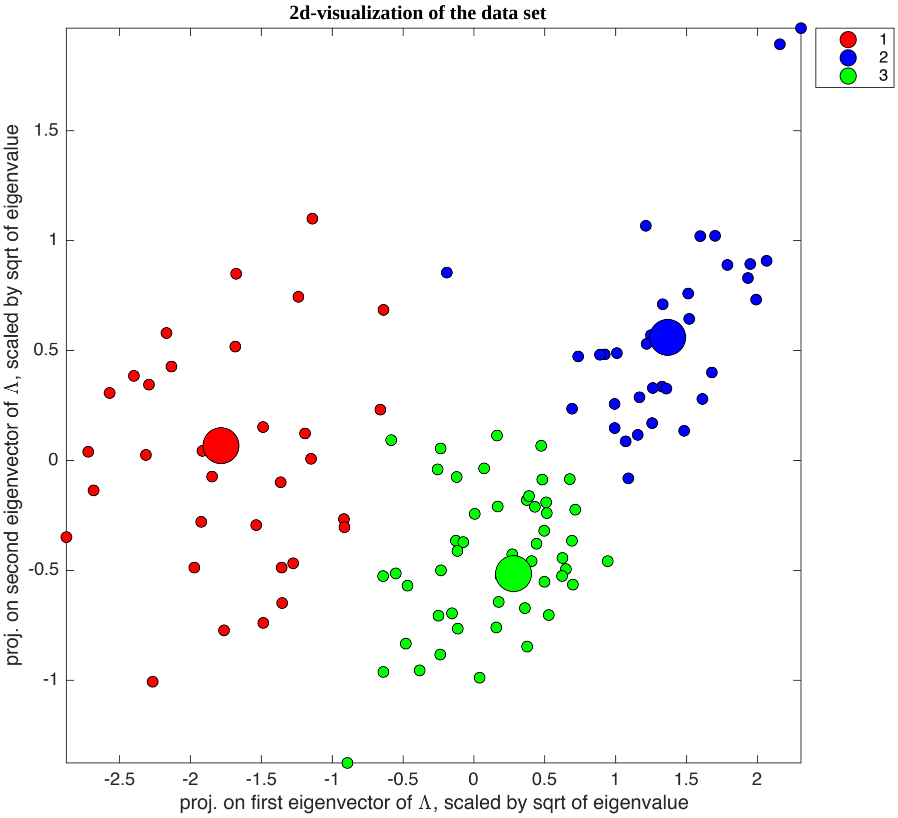


Figure 3.13: GMLVQ LC-MS/MS Data visualisation.

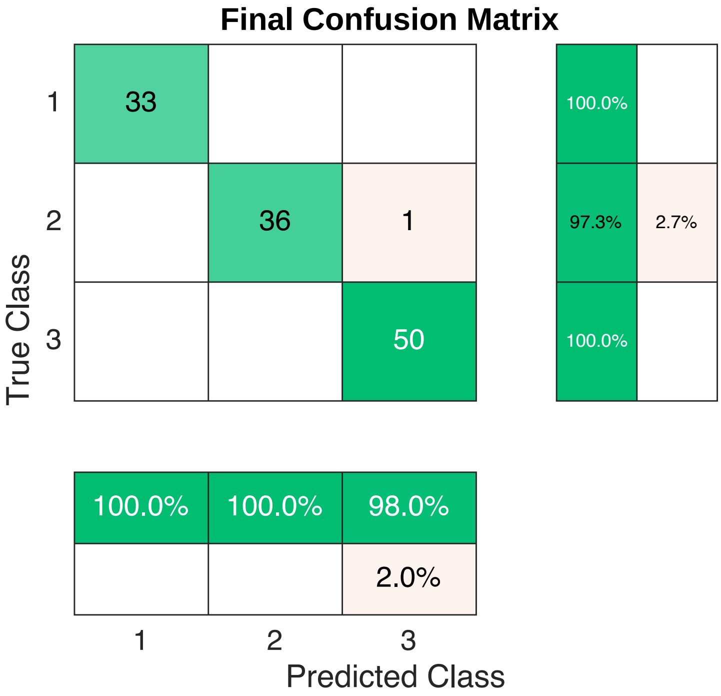


Figure 3.14: GMLVQ LC-MS/MS Confusion matrix.

4 CONCLUSION

The analysis of the raw urinary steroid metabolome data of patients using different pre-processing methods shows that a unit transformation followed by a natural logarithmic transformation prepares the data sufficiently so that a Euclidean k-means is able to separate the patients into meaningful clusters. Where mathematical analysis suggests two clusters, it does not rule out that there could be more. Domain knowledge suggests three clusters is viable and of interest. The other clustering methods show a similar clustering to k-means; however, the latter produces reliable and robust results. This is supported by the PCA projection and Phenotypical analysis. These show that the clustering based on GC-MS is superior to LC-MS/MS. GMLVQ is able to reproduce the clustering in a convincing way, further strengthening our case. Now we go back to our research questions and answer them to the best of our ability.

Q1) Does unsupervised learning reveal clusters / subgroups of adrenal tumours in the data?

A1) *Yes, it does. We observe three clusters that are interesting from the domain perspective.*

Q2) Does the inclusion of clinical data beyond steroid metabolomics help in Q1?

A2) *It helps to underline the validity of the clustering by highlighting a non-uniform distribution of patient phenotypes.*

Phenotype: the set of characteristics of a living thing, resulting from its combination of genes and the effect of its environment [44]

Q3) Can supervised learning / classification reproduce the clusters defined in Q1 and Q2?

A3) *Yes, using GMLVQ to classify according to the clustering, we see excellent (GC-MS) and good (LC-MS/MS) results in separation, AOC and accuracy (see section 3.2).*

Q4) What is the relevance or importance of individual steroid markers for the discrimination of tumour sub-types?

A4) *As we see in figure 4.1, 6β-OH-cortisol, cortisol, DHEA, and 5-PT are the most relevant steroid markers.*

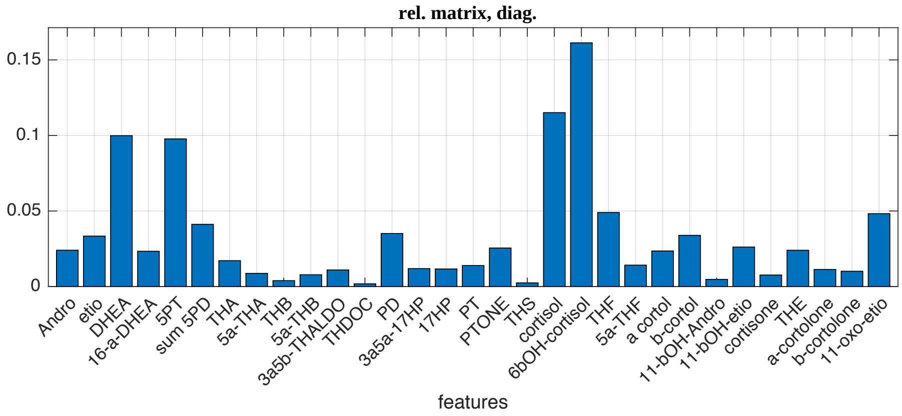


Figure 4.1: GMLVQ GC-MS Relevances. Enlarged view of the subfigure from figure 3.9. Labelled with the steroid names.

5 OUTLOOK, FURTHER WORK

During the writing phase and after this thesis is concluded, the author continues to work with the research group of Professor Dr. Wiebke Arlt, MD in Birmingham, both on this subject as well as other medical applications that benefit from unsupervised clustering analysis and additional machine learning techniques.

The outlook of this research is that it could lead to a differential diagnosis of ACC, increased understanding of the underlying mechanisms through analysis of the steroid metabolome, ultimately resulting in targeted treatment to reduce mortality rate.

Further work includes, but is not limited to

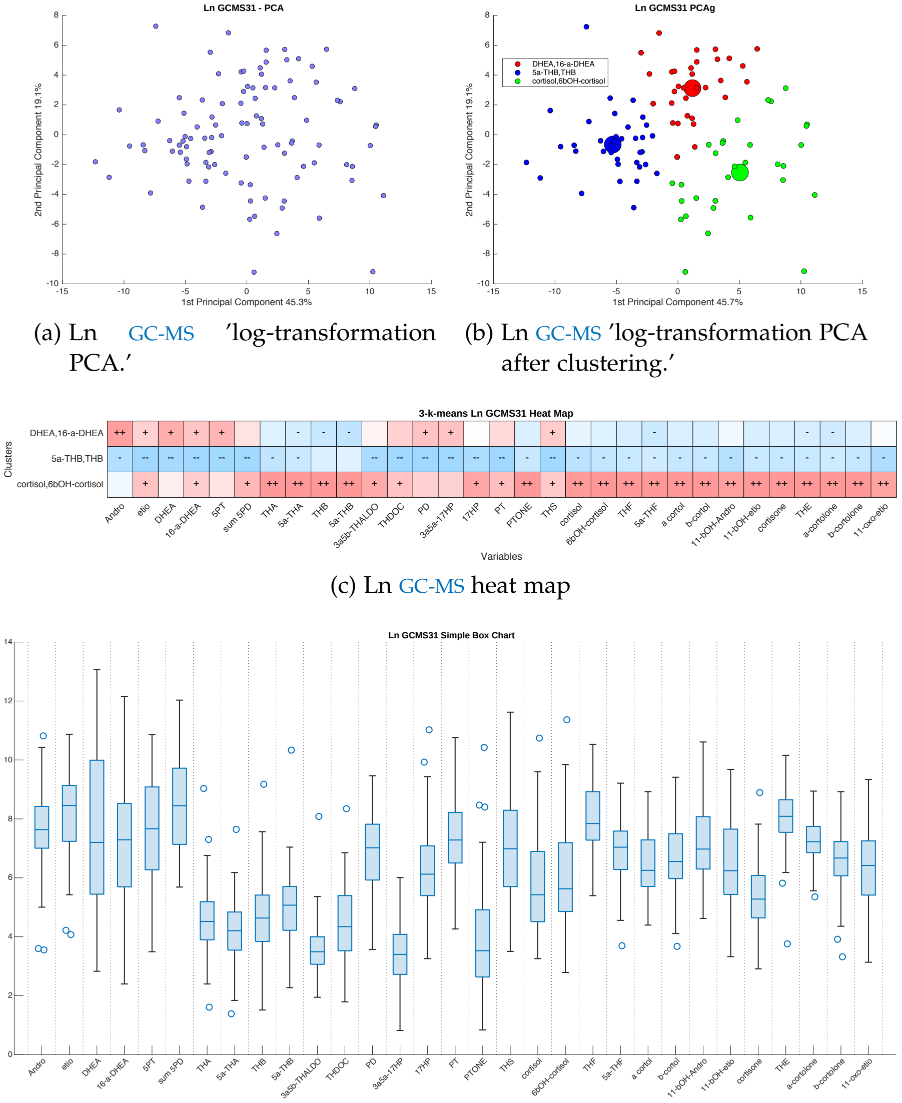
- Looking into additional clustering algorithms to be used. *E. g. Spectral clustering, Generative topographical Mapping (GTM), DBScan, Gaussian Mixture Models.*
- Analysing different data sets. *E. g. spot urine or blood serum.*
- Supervised learning follow up. *When we have properly labelled classes, a supervised classification follow-up would be very interesting.*
- Extending and enhancing UCAT. *As with any software project, there is always room for improvement in terms of functionality, code quality, adaptation to different use cases.*
- Applying unsupervised clustering to different medical problems. *A start has been made to apply unsupervised clustering to urine metabolome data of patients with Polycystic Ovary Syndrome (PCOS) with encouraging first results.*

APPENDIX

A

ADDITIONAL FIGURES

These figures are not directly relevant to the main text, but for some readers they may be of interest.



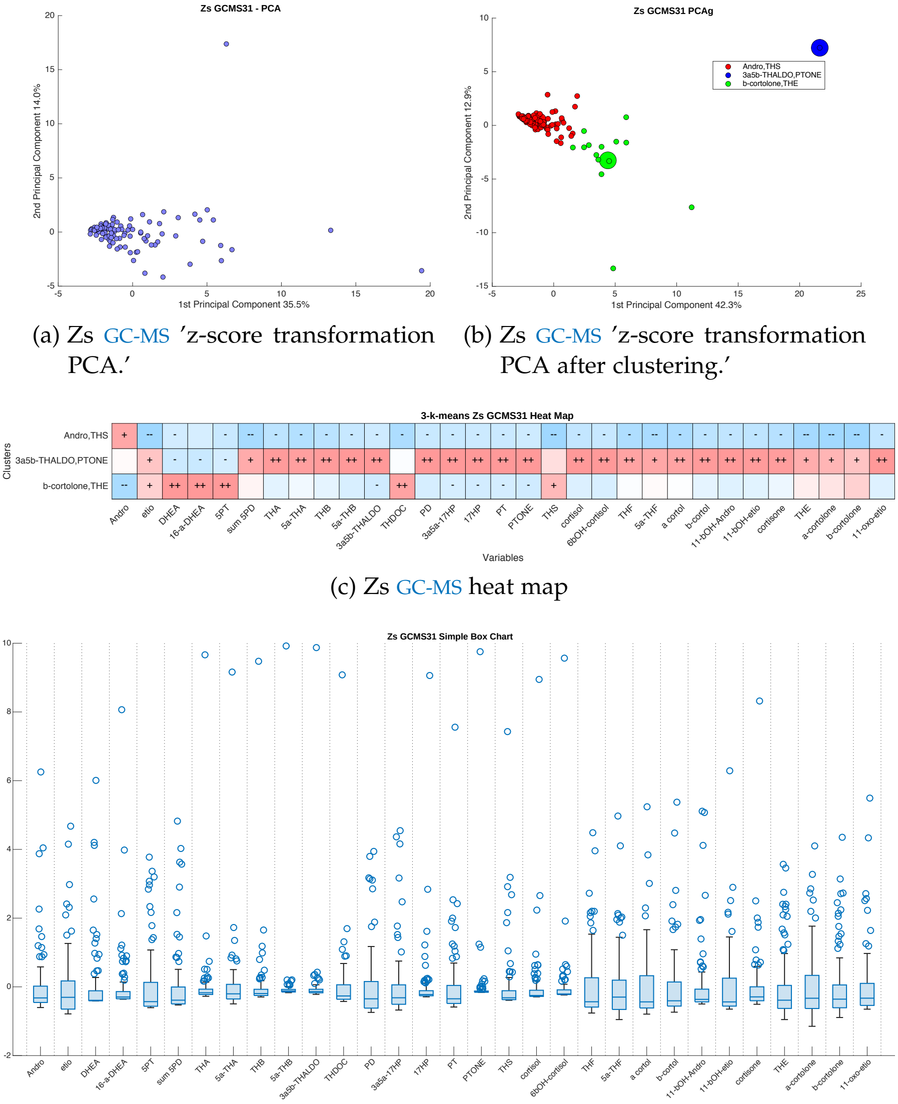
(a) Ln GC-MS 'log-transformation PCA.'

(b) Ln GC-MS 'log-transformation PCA after clustering.'

(c) Ln GC-MS heat map

(d) Ln GC-MS box chart

Figure A.1: Data pre-processing: log-transformation. See section 2.1.



(a) Zs GC-MS 'z-score transformation PCA.'

(b) Zs GC-MS 'z-score transformation PCA after clustering.'

(c) Zs GC-MS heat map

(d) Zs GC-MS box chart

Figure A.2: Data pre-processing: z-score transformation. See section 2.1.

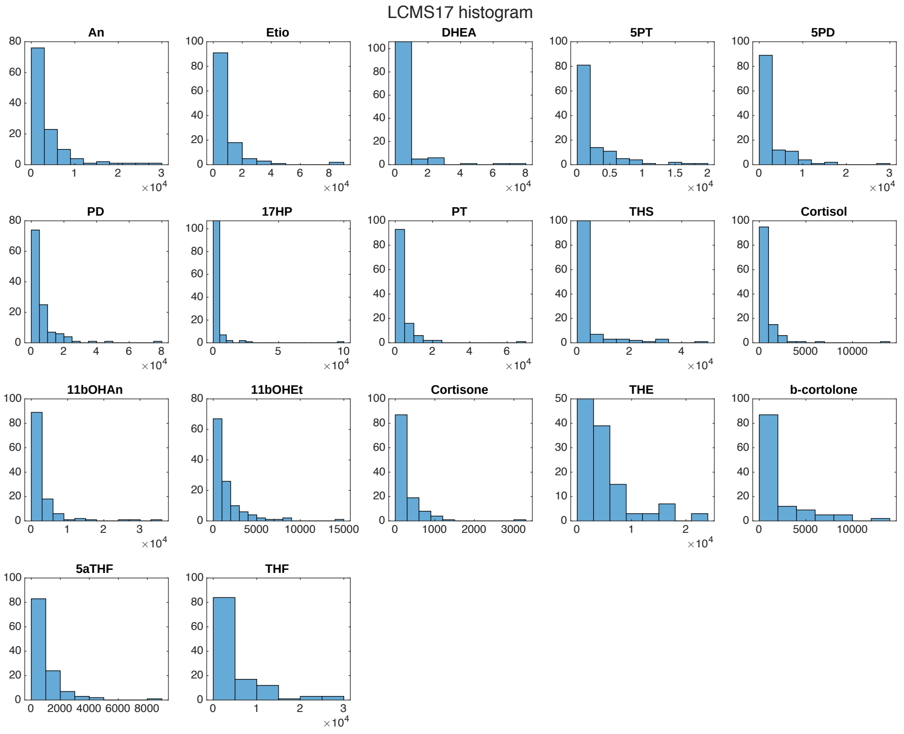


Figure A.3: LC-MS/MS 'Raw' histogram. They look rather skewed. See section 2.1.

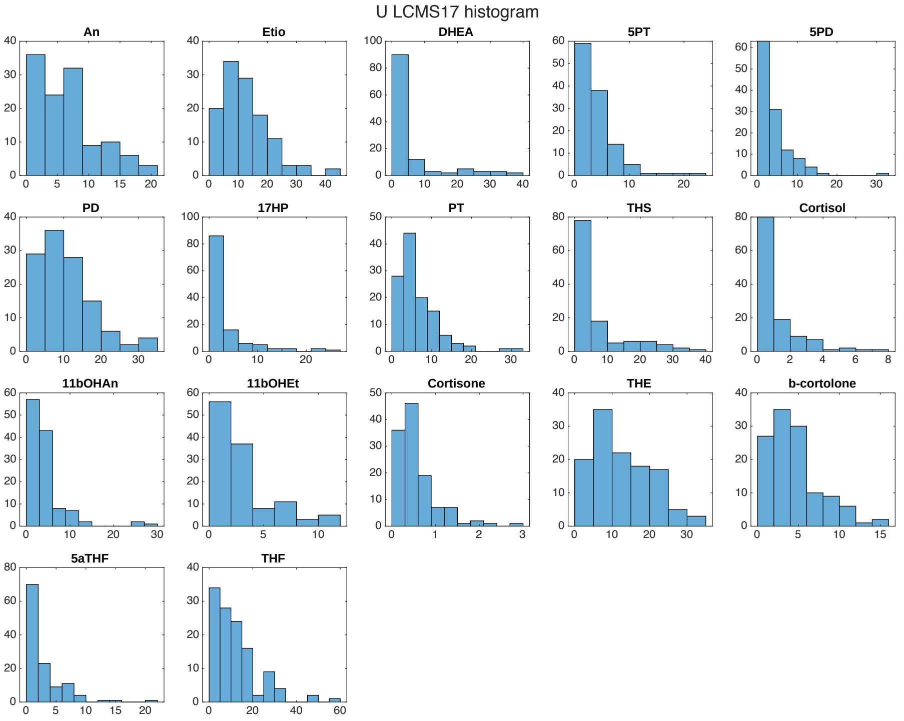


Figure A.4: LC-MS/MS Unit-Normalized histogram. This looks better, but it doesn't look very Gaussian yet. See section 2.1.

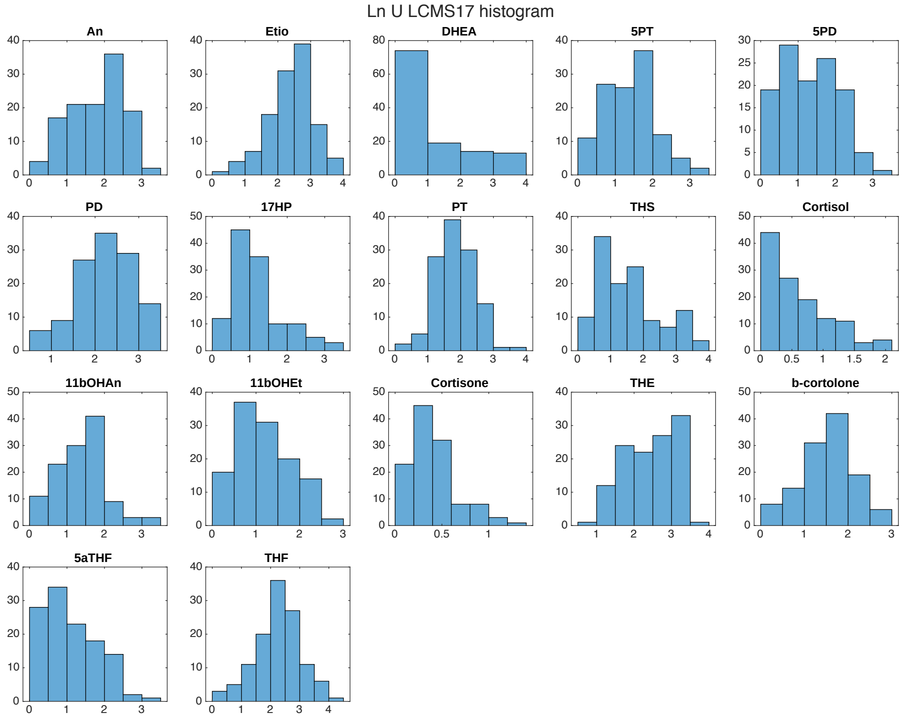


Figure A.5: LC-MS/MS Log-transformed histogram. Now we see Gaussian like distributions for some of the steroids. See section 2.1.

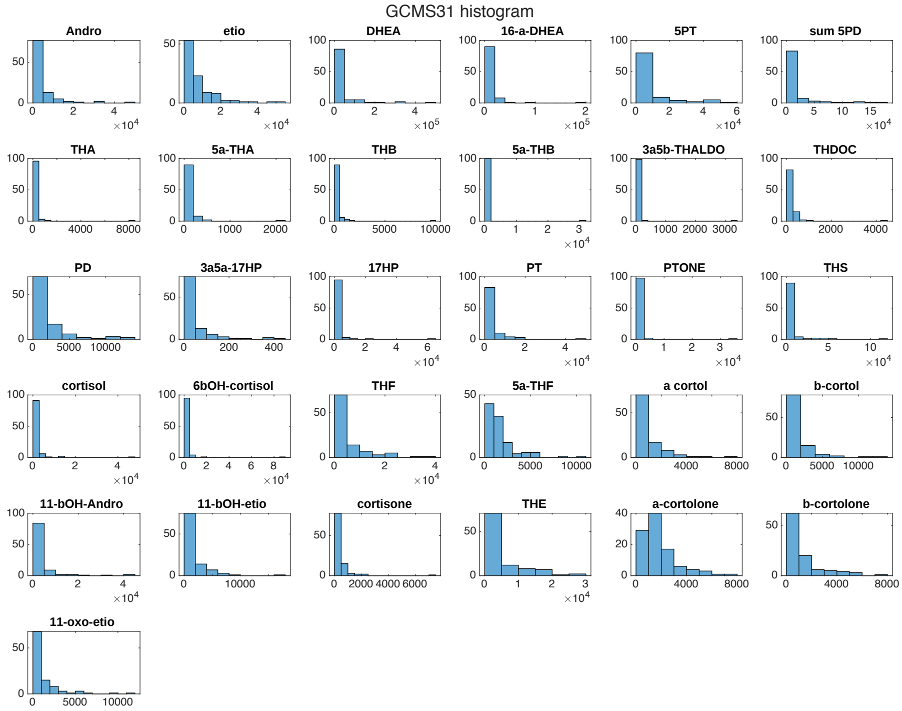


Figure A.6: GC-MS 'Raw' histogram. They look rather skewed. See section 2.1.

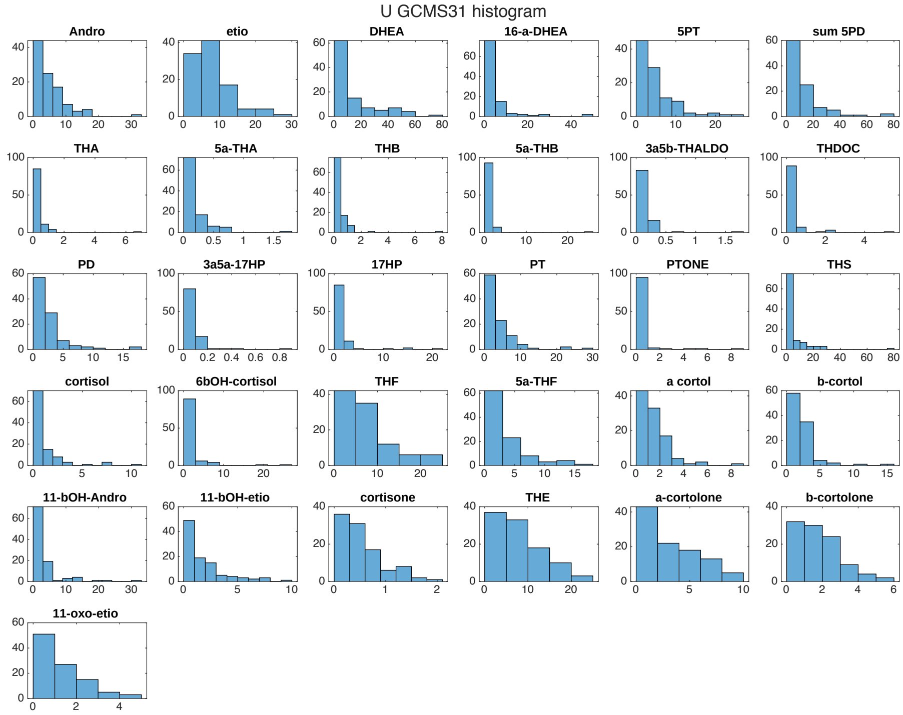


Figure A.7: GC-MS Unit-Normalized histogram. This looks better, but it doesn't look very Gaussian yet. See section 2.1.

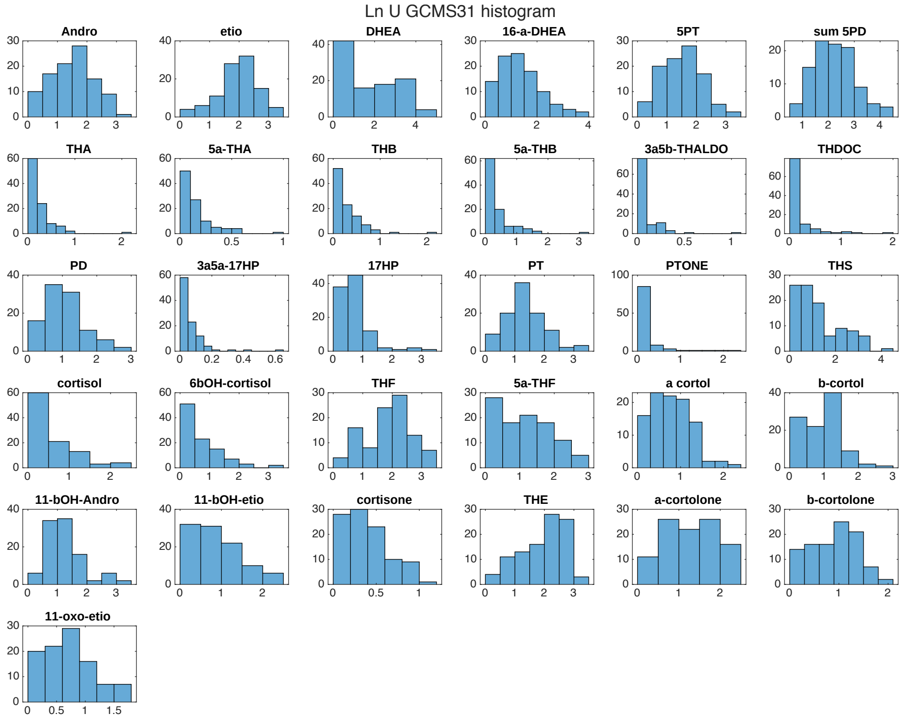
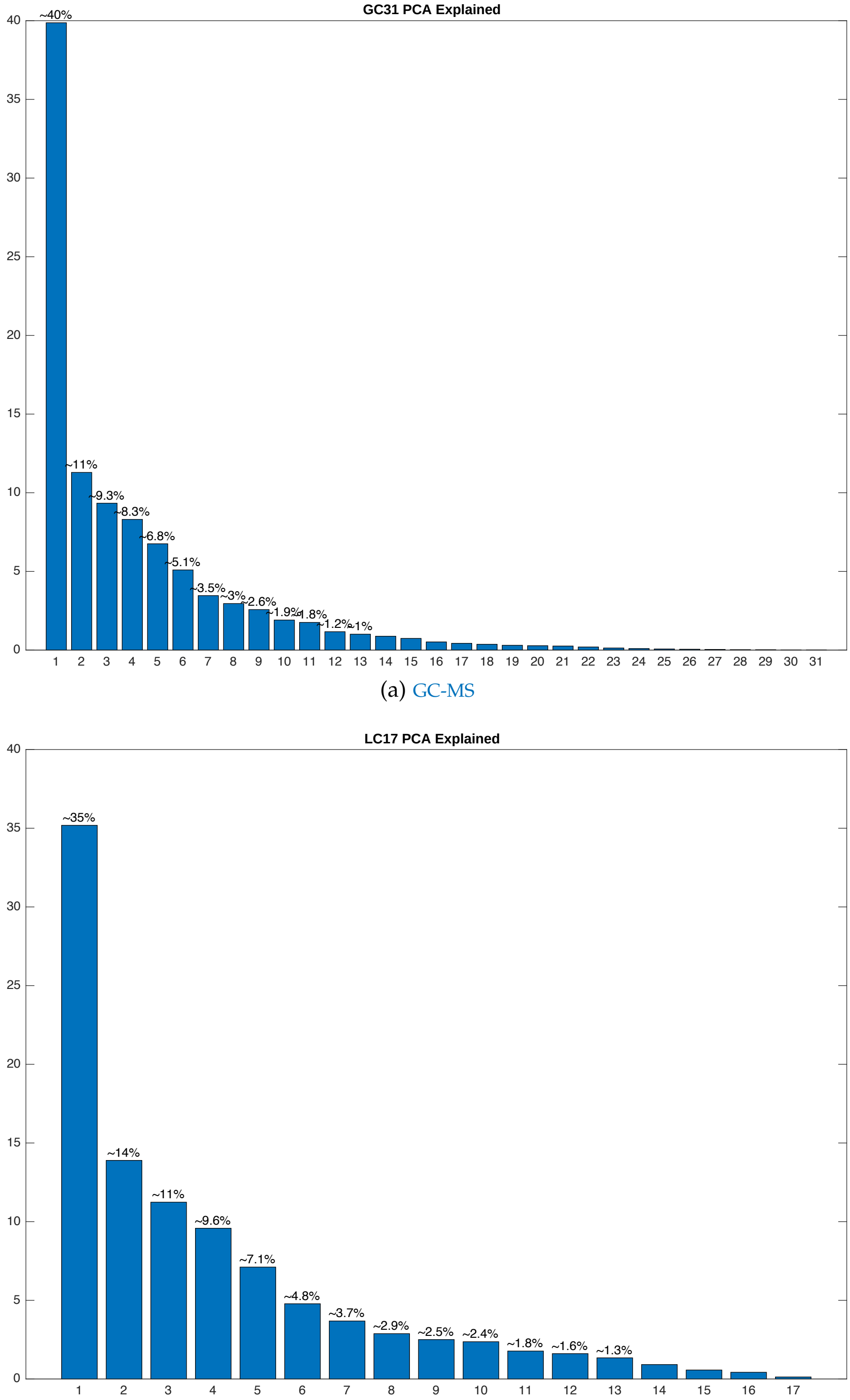


Figure A.8: GC-MS Log-transformed histogram. Now we see Gaussian like distributions for some of the steroids. See section 2.1.



(a) GC-MS

(b) LC-MS/MS

Figure A.9: PCA Variance explained by each component for the normalized log-transformed datasets. See section 2.1.

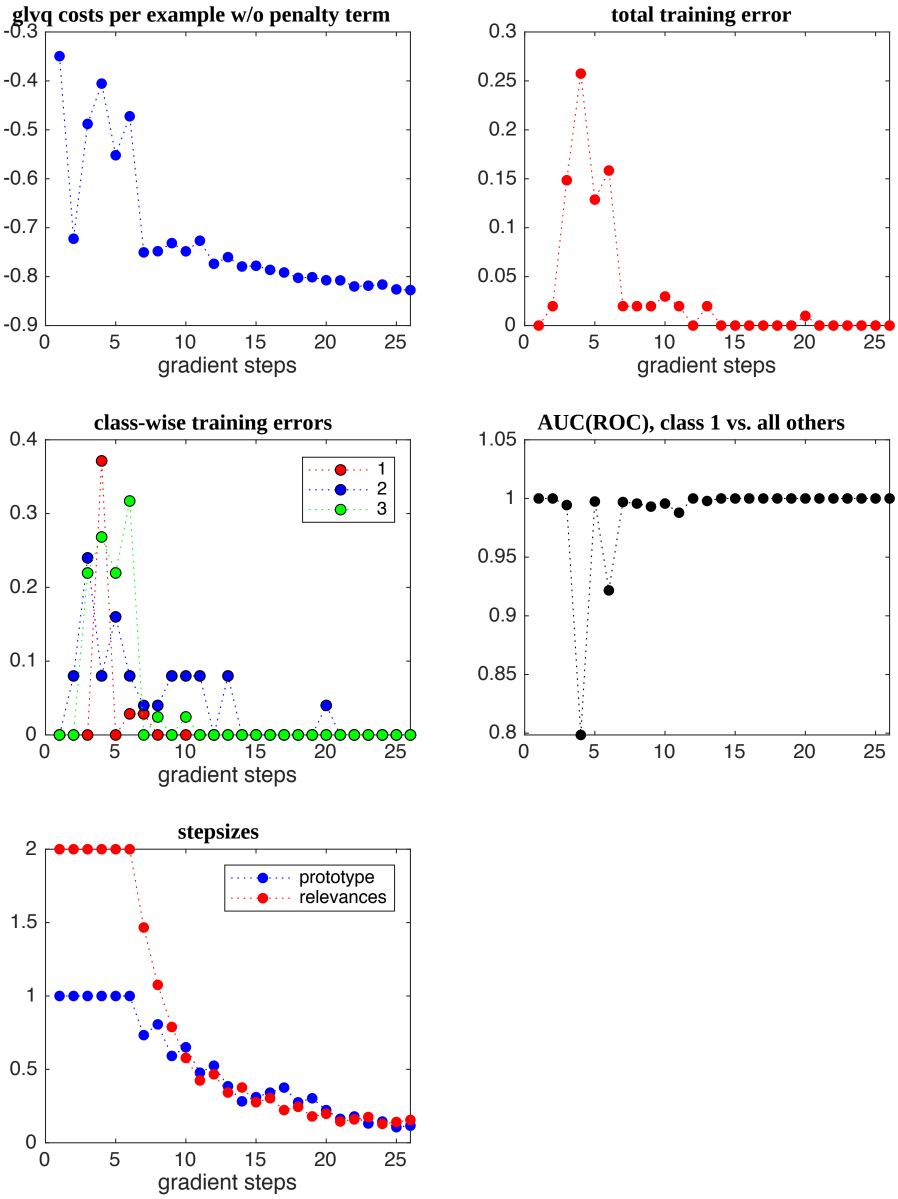


Figure A.10: GMLVQ GC-MS Training overview. See section 3.3.

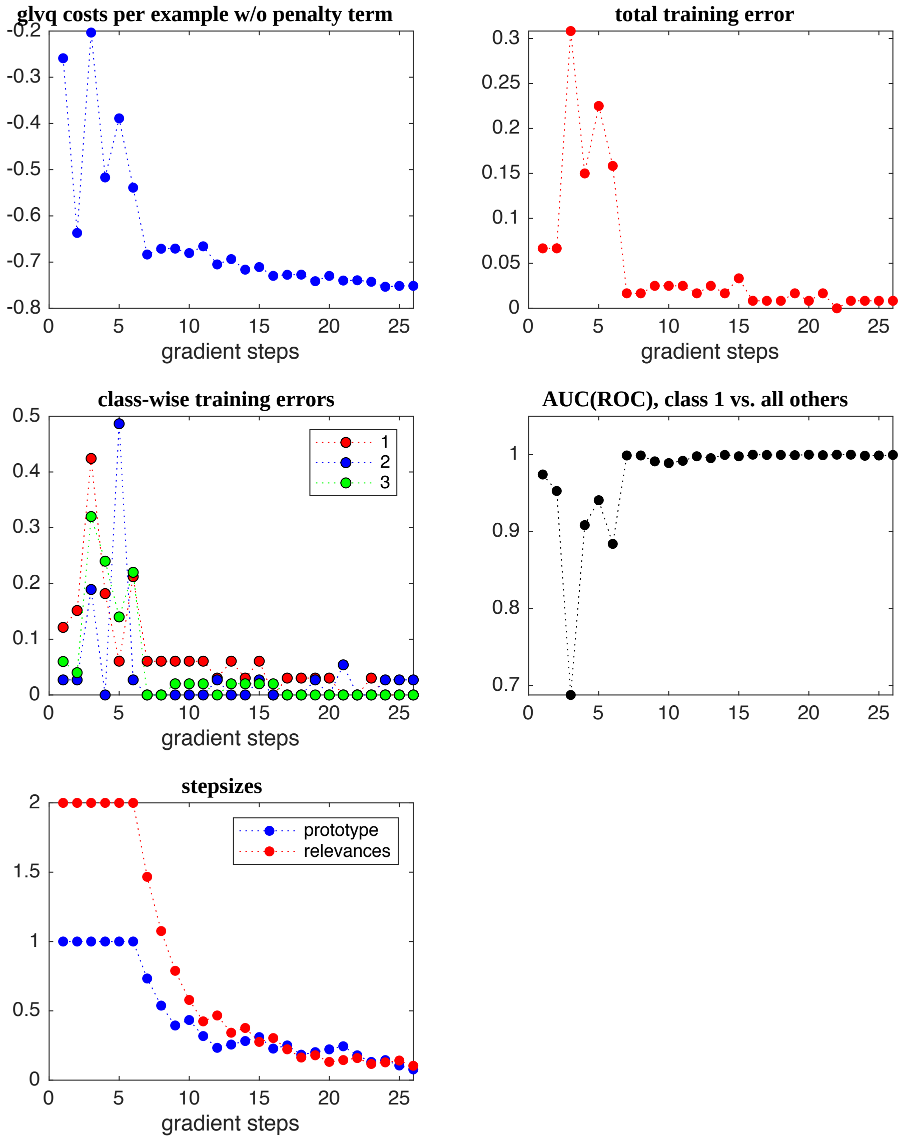


Figure A.11: GMLVQ LC-MS/MS Training overview. See section 3.3.

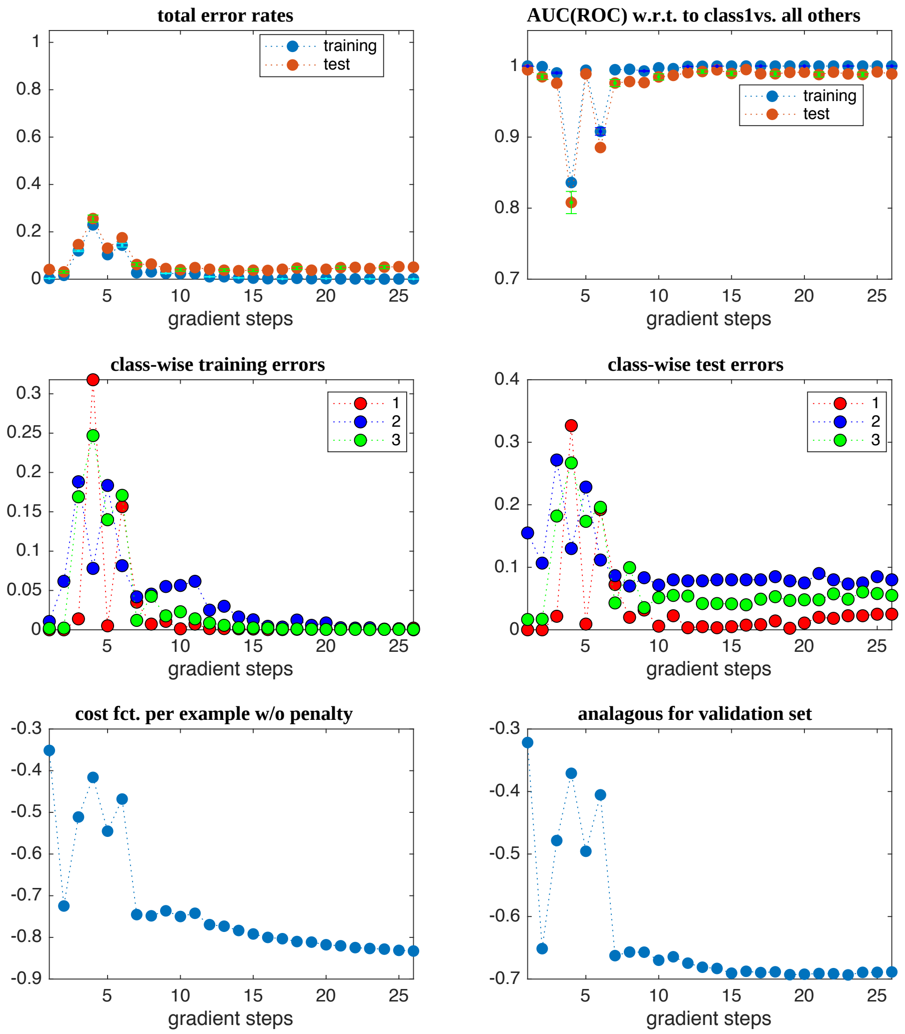


Figure A.12: GMLVQ GC-MS Cross validation overview. Average of 100 runs with random 10% of samples used for testing. See section 3.3.

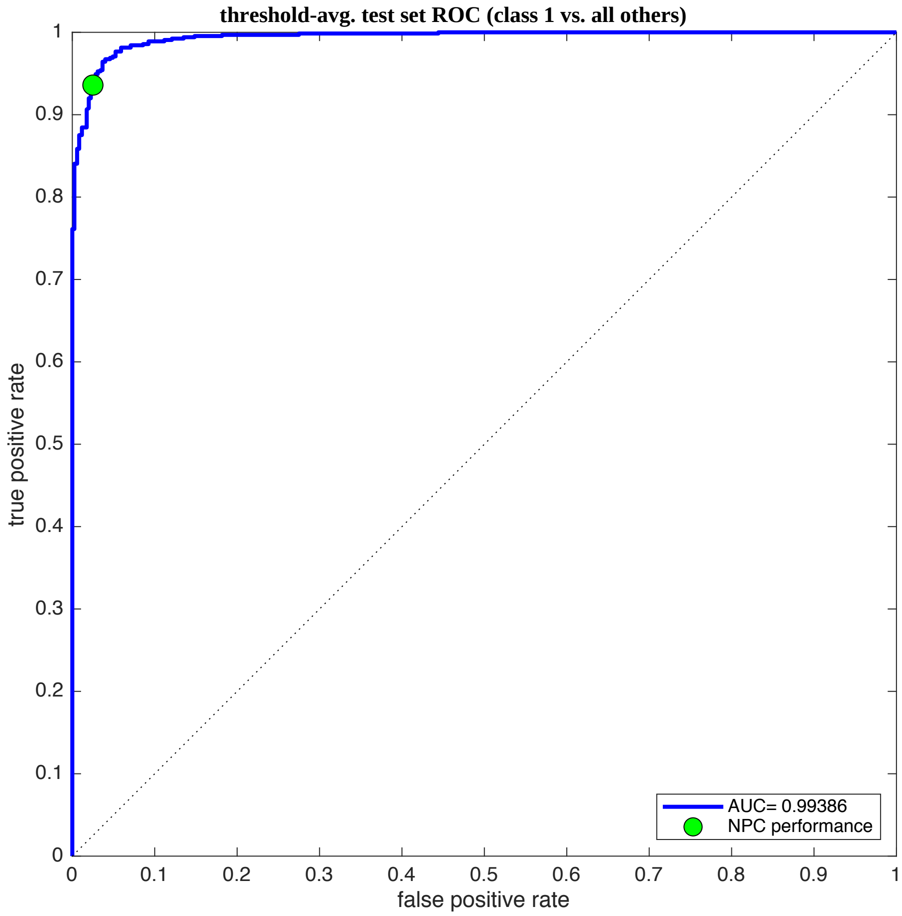


Figure A.13: GMLVQ GC-MS Cross validation ROC. Average of 100 runs with random 10% of samples used for testing. See section 3.3.

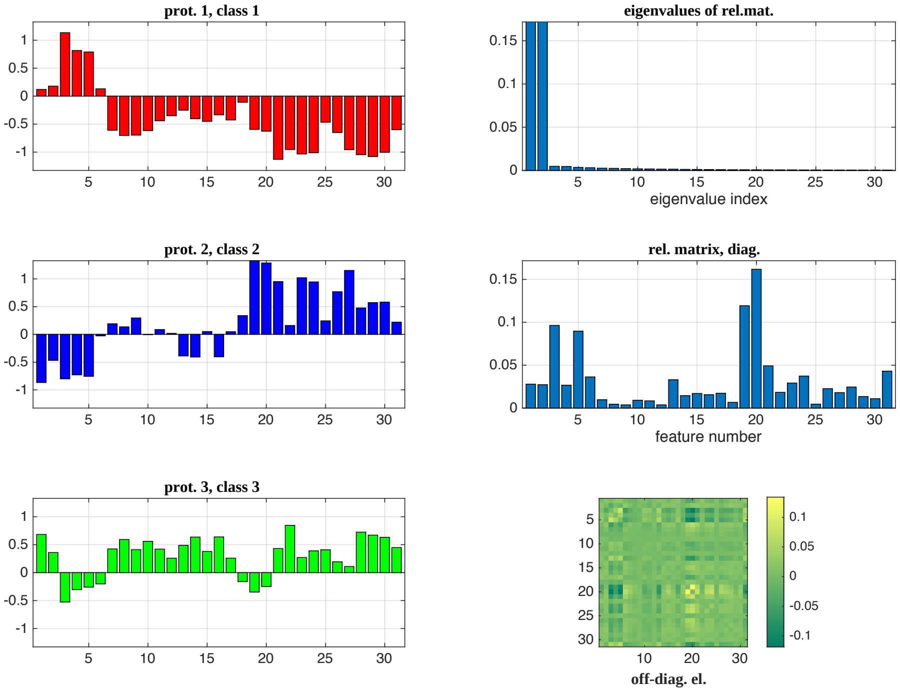


Figure A.14: GMLVQ GC-MS cross validation prototypes and relevance matrix. Average of 100 runs with random 10% of samples used for testing. See section 3.3.

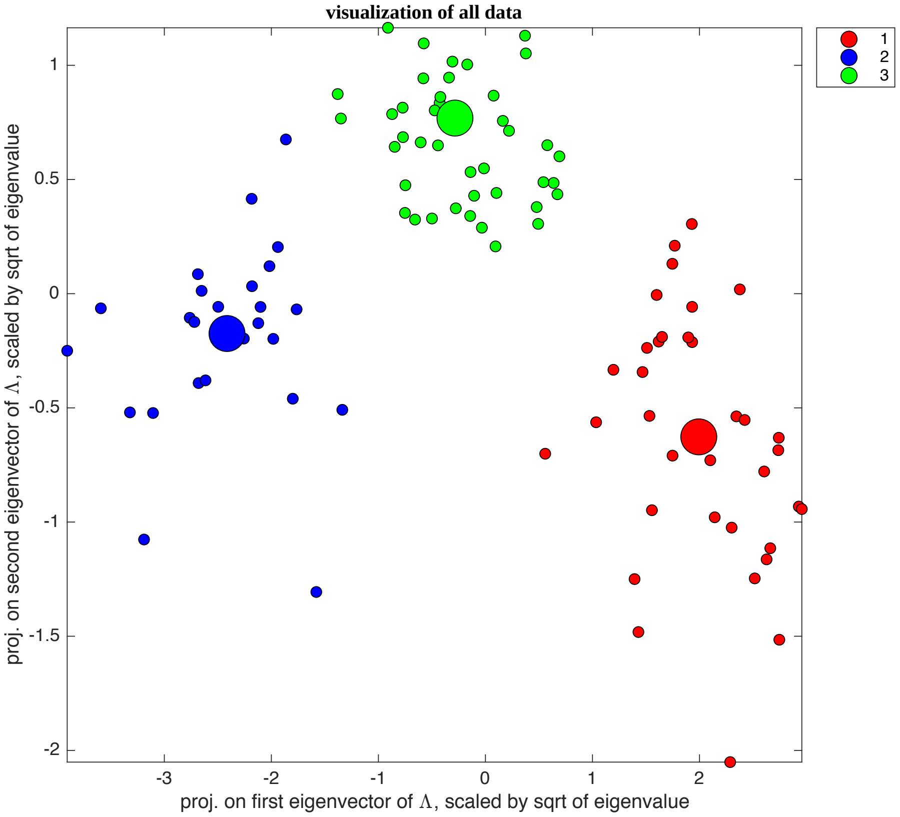
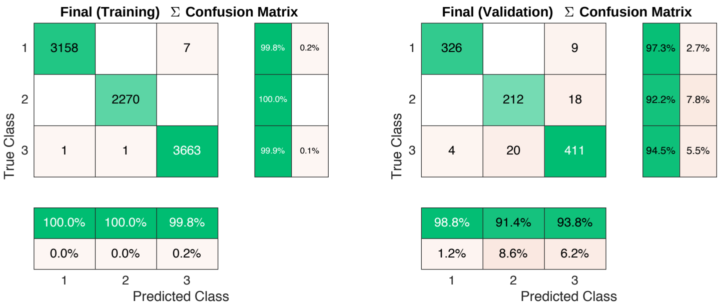


Figure A.15: GMLVQ GC-MS Data visualisation. Average of 100 runs with random 10% of samples used for testing. See section 3.3.



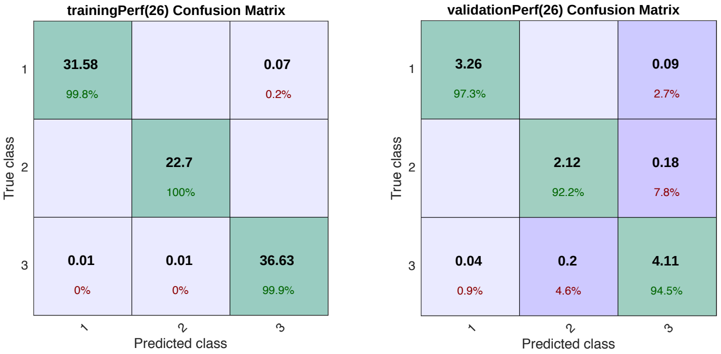
Final (Training) Σ Confusion Matrix

True Class	1	2	3		
1	3158		7	99.8%	0.2%
2		2270		100.0%	
3	1	1	3663	99.9%	0.1%
	100.0%	100.0%	99.8%		
	0.0%	0.0%	0.2%		

Final (Validation) Σ Confusion Matrix

True Class	1	2	3		
1	326		9	97.3%	2.7%
2		212	18	92.2%	7.8%
3	4	20	411	94.5%	5.5%
	98.8%	91.4%	93.8%		
	1.2%	8.6%	6.2%		

Figure A.16: GMLVQ GC-MS Validation Confusion matrix. Sum of 100 runs with random 10% of samples used for testing. See section 3.3.



trainingPerf(26) Confusion Matrix

True class	1	2	3
1	31.58 99.8%		0.07 0.2%
2		22.7 100%	
3	0.01 0%	0.01 0%	36.63 99.9%

validationPerf(26) Confusion Matrix

True class	1	2	3
1	3.26 97.3%		0.09 2.7%
2		2.12 92.2%	0.18 7.8%
3	0.04 0.9%	0.2 4.6%	4.11 94.5%

Figure A.17: GMLVQ GC-MS Validation Confusion matrix. Average of 100 runs with random 10% of samples used for testing. See section 3.3.

LISTINGS

Here you find a few snippets of code to give you an idea of the code. Some of the code shown has since been refactored (e. g. `KType` is currently a proper enumeration, the `kmeans` method is renamed to `performClustering`), but the essence remains the same.

Listing A.1: An example of invoking the `UCAT` library, slightly abbreviated

```
% Load the data
load gcms31.mat, load lcms17.mat

% Some patients diagnosis changed; exclude them:
excludelist = {'FRPA2-0135' 'FRPA2-0183' 'NLEI-85' 'GYWU-1235' '
    FRBO-0025' 'FRBO-0033' 'GYWU-0655'};

ktype = 'pure'; nClusters = 3; %som, pure, cluster, medoids

lc17 = rjv.UCAT(lcms17, 'LCMS17', lcms17names, lcms17patients);
gc31 = rjv.UCAT(gcms31, 'GCMS31', gcms31names, gcms31patients);

lc17.excludepatients(excludelist);
gc31.excludepatients(excludelist);

ucats = [lc17 gc31];
prep = @(x) x.unitnormalize(100).logtransform(true);
arrayfun(prep,ucats);

% Run and visualize all the ucats
arrayfun(@(x) makerun(x, nClusters, ktype), ucats);
arrayfun(@(x) makeviz(x), ucats);

function g = makerun(g, nClusters, ktype)
    g.KType = ktype; g.kmeans(nClusters);
end

function makeviz(g)
    figure('Name', [g.name ' t-SNE']);
    g.tsne(); g.tsnegscatter();
    figure('Name', [g.name ' PCA']); g.pcagscatter();
    figure('Name', [g.name ' Cluster Difference']); g.
        clusterDiffErrBar(false);
    figure; g.clusterDiffBoxChart(false);
    figure; g.heatmap(false);
end
```

Listing A.2: An example of invoking the `GMLVQ` library

```
% add acess to GMLVQ library
addpath(genpath("../gmlvq-v3-1/"));
% fvec: feature vectors, lbl: labels, 50:total number of steps
gmlvq = GMLVQ.GMLVQ(fvec,lbl, GMLVQ.Parameters(), 50);
result = gmlvq.runSingle();
result.plot();
```

BIBLIOGRAPHY

[1] Alan M. TURING. 'I.—COMPUTING MACHINERY AND INTELLIGENCE'. In: *Mind* LIX.236 (Oct. 1950), pages 433–460. ISSN: 0026-4423. DOI: 10.1093/mind/LIX.236.433. eprint: https://academic.oup.com/mind/article-pdf/LIX/236/433/30123314/lix-236-433.pdf. URL: https://doi.org/10.1093/mind/LIX.236.433.

[2] Nils Krone, Beverly A. Hughes, Gareth G. Lavery, Paul M. Stewart, Wiebke Arlt and Cedric H.L. Shackleton. 'Gas chromatography/mass spectrometry (GC/MS) remains a pre-eminent discovery tool in clinical steroid investigations even in the era of fast liquid chromatography tandem mass spectrometry (LC/MS/MS)'. In: *The Journal of Steroid Biochemistry and Molecular Biology* 121.3 (2010). Steroid profiling and analytics: going towards Sterome, pages 496–504. ISSN: 0960-0760. DOI: https://doi.org/10.1016/j.jsbmb.2010.04.010. URL: https://www.sciencedirect.com/science/article/pii/S0960076010002104.

[3] Wiebke Arlt, Michael Biehl, Angela E. Taylor, Stefanie Hahner, Rossella Libé, Beverly A. Hughes, Petra Schneider, David J. Smith, Han Stiekema, Nils Krone, Emilio Porfiri, Giuseppe Opocher, Jerôme Bertherat, Franco Mantero, Bruno Allolio, Massimo Terzolo, Peter Nightingale, Cedric H. L. Shackleton, Xavier Bertagna, Martin Fassnacht and Paul M. Stewart. 'Urine Steroid Metabolomics as a Biomarker Tool for Detecting Malignancy in Adrenal Tumors'. In: *The Journal of Clinical Endocrinology & Metabolism* 96.12 (Dec. 2011), pages 3775–3784. ISSN: 0021-972X. DOI: 10.1210/jc.2011-1565. eprint: https://academic.oup.com/jcem/article-pdf/96/12/3775/20288130/jcem3775.pdf. URL: https://doi.org/10.1210/jc.2011-1565.

[4] Angela E. Taylor, Irina Bancos, Vasileios Chortis, Alice J. Sitch, Carl Jenkinson, Caroline J. Davidge-Pitts, Katharina Lang, Stylianos Tsagarakis, Magdalena Macech, Anna Riester, Timo Deutschbein, Ivana D. Pupovac, Tina Kienitz, Alessandro Prete, Thomas G. Papathomas, Lorna C. Gilligan, Cristian Bancos, Giuseppe Reimondo, Magalie Haissaguerre, Ljiljana Marina, Marianne A. Grytaas, Ahmed Sajwani, Katharina Langton, Hannah E. Ivison, Cedric H. L. Shackleton, Dana Erickson, Miriam Asia, Sotiria Palimeri, Agnieszka Kondracka, Ariadni Spyroglou, Cristina L. Ronchi, Bojana Simunov, Danae A. Delivanis, Robert P. Sutcliffe, Ioanna Tsirou, Tomasz Bednarczuk, Martin Reincke, Stephanie Burger-Stritt, Richard A. Feelders, Letizia Canu, Harm R. Haak, Graeme Eisenhofer, M. Conall Dennedy, Grethe A. Ueland, Mi-

omira Ivovic, Antoine Tabarin, Massimo Terzolo, Marcus Quinkler, Darko Kastelan, Martin Fassnacht, Felix Beuschlein, Urszula Ambroziak, Dimitra A. Vassiliadi, Michael W. O'Reilly, Jr Young William F, Michael Biehl, Jonathan J. Deeks, Wiebke Arlt and E. N. S. A. T. E. U. R. I. N. E.-A. C. T. Investigators. 'Urine steroid metabolomics for the differential diagnosis of adrenal incidentalomas in the EURINE-ACT study: a prospective test validation study'. eng. In: *The lancet. Diabetes & endocrinology* 8.32711725 (Sept. 2020), pages 773–781. ISSN: 2213-8587. DOI: 10.1016/S2213-8587(20)30218-7. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7447976/.

[5] M Biehl, P Schneider, D Smith, H Stiekema, A Taylor, B Hughes, C Shackleton, P Stewart and W Arlt. 'Matrix relevance LVQ in steroid metabolomics based classification of adrenal tumors'. English. In: *Proc. 20th European Symposium on Artificial Neural Networks*. Edited by Michel Verleysen. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, ; Conference date: 25-04-2012 Through 27-04-2012. d-side publishing, 2012, pages 423–428. ISBN: 978-2-87419-049-0.

[6] Elina Laura van den Brandhof. 'Classification of Benign Adrenal Tumors Based on Steroid Metabolomics'. Master's thesis. Technical University of Munich, 2020.

[7] ENS@T. *What are adrenal tumors?* URL: http://www.ensat.org/page-1317250 (visited on 29/05/2022).

[8] ENS@T. *Adrenocortical Carcinomas*. URL: http://www.ensat.org/page-1317312 (visited on 29/05/2022).

[9] Martin Fassnacht, Olaf M Dekkers, Tobias Else, Eric Baudin, Alfredo Berruti, Ronald R de Krijger, Harm R Haak, Radu Mihai, Guillaume Assie and Massimo Terzolo. 'European Society of Endocrinology Clinical Practice Guidelines on the management of adrenocortical carcinoma in adults, in collaboration with the European Network for the Study of Adrenal Tumors'. In: *European Journal of Endocrinology* 179.4 (2018), G1–G46. DOI: 10.1530/EJE-18-0608. URL: https://eje.bioscientifica.com/view/journals/eje/179/4/EJE-18-0608.xml.

[10] Anne Jouinot and Jérôme Bertherat. 'MANAGEMENT OF ENDOCRINE DISEASE: Adrenocortical carcinoma: differentiating the good from the poor prognosis tumors'. In: *European Journal of Endocrinology* 178.5 (2018), R215–R230. DOI: 10.1530/EJE-18-0027. URL: https://eje.bioscientifica.com/view/journals/eje/178/5/EJE-18-0027.xml.

[11] James F H Pittaway and Leonardo Guasti. 'Pathobiology and genetics of adrenocortical carcinoma'. In: *Journal of Molecular Endocrinology* 62.2 (2019), R105–R119. DOI: 10.1530/JME-18-0122. URL: https://jme.bioscientifica.com/view/journals/jme/62/2/JME-18-0122.xml.

[12] Sina Jasim and Mouhammed Amir Habra. 'Management of Adrenocortical Carcinoma'. In: *Current Oncology Reports* 21.3 (Feb. 2019), page 20. ISSN: 1534-6269. DOI: 10.1007/s11912-019-0773-7. URL: https://doi.org/10.1007/s11912-019-0773-7.

[13] Y S Elhassan, B Altieri, S Berhane, D Cosentini, A Calabrese, M Haissaguerre, D Kastelan, M C B V Fragoso, J Bertherat, A Al Ghuzlan, H Haak, M Boudina, L Canu, P Loli, M Sherlock, O Kimpel, M Laganà, A J Sitch, M Kroiss, W Arlt, M Terzolo, A Berruti, J J Deeks, R Libé, M Fassnacht, C L Ronchi and the ENSAT. 'S-GRAS score for prognostic classification of adrenocortical carcinoma: an international, multicenter ENSAT study'. In: *European Journal of Endocrinology* 186.1 (2022), pages 25–36. DOI: 10.1530/EJE-21-0510. URL: https://eje.bioscientifica.com/view/journals/eje/186/1/EJE-21-0510.xml.

[14] Karl-Heinz Storbeck, Lina Schiffer, Elizabeth S Baranowski, Vasileios Chortis, Alessandro Prete, Lise Barnard, Lorna C Gilligan, Angela E Taylor, Jan Idkowiak, Wiebke Arlt and Cedric H L Shackleton. 'Steroid Metabolome Analysis in Disorders of Adrenal Steroid Biosynthesis and Metabolism'. In: *Endocrine Reviews* 40.6 (July 2019), pages 1605–1625. ISSN: 0163-769X. DOI: 10.1210/er.2018-00262. eprint: https://academic.oup.com/edrv/article-pdf/40/6/1605/30843306/er.2018-00262.pdf. URL: https://doi.org/10.1210/er.2018-00262.

[15] C.H.L. Shackleton. 'Profiling steroid hormones and urinary steroids'. In: *Journal of Chromatography B: Biomedical Sciences and Applications* 379 (1986), pages 91–156. ISSN: 0378-4347. DOI: https://doi.org/10.1016/S0378-4347(00)80683-0. URL: https://www.sciencedirect.com/science/article/pii/S0378434700806830.

[16] T. Kohonen. *Self-Organizing Maps*. Springer, Berlin, Germany, 1997.

[17] Michael Biehl, Barbara Hammer and Thomas Villmann. 'Prototype-based models in machine learning'. In: *WIREs Cognitive Science* 7.2 (2016), pages 92–111. DOI: https://doi.org/10.1002/wcs.1378. eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcs.1378. URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcs.1378.

[18] Umut Asan and Secil Ercan. 'An Introduction to Self-Organizing Maps'. In: Jan. 2012, pages 299–319. ISBN: 978-94-91216-76-3. DOI: 10.2991/978-94-91216-77-0_14.

[19] S.P. Lloyd. 'Least squares quantization in PCM'. In: *IEEE Transactions on Information Theory* 28 (1982). first published in Bell Telephone Laboratories Paper 1957, pages 129–137.

[20] Michael Biehl. *The Shallow and the Deep, An Introduction to Neural Networks and Machine Learning*. University of Groningen, Groningen, The Netherlands, Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence. Mar. 2022. ISBN: 9789403429656. eprint: https://research.rug.nl/en/publications/the-shallow-and-the-deep-a-biased-introduction-to-neural-networks.

[21] Mathworks. *k-means clustering*. 2022. URL: https://nl.mathworks.com/help/stats/kmeans.html (visited on 21/06/2022).

[22] Peter J. Rousseeuw Leonard Kaufman. 'Partitioning Around Medoids (Program PAM)'. In: *Finding Groups in Data*. John Wiley & Sons, Ltd, 1990. Chapter 2, pages 68–125. ISBN: 9780470316801. DOI: https://doi.org/10.1002/9780470316801.ch2. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470316801.ch2. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470316801.ch2.

[23] Mathworks. *k-medoids clustering*. 2022. URL: https://nl.mathworks.com/help/stats/kmedoids.html (visited on 21/06/2022).

[24] Fionn Murtagh and Pedro Contreras. 'Algorithms for hierarchical clustering: an overview'. In: *WIREs Data Mining and Knowledge Discovery* 2.1 (2012), pages 86–97. DOI: https://doi.org/10.1002/widm.53. eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.53. URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.53.

[25] Mathworks. *Hierarchical Clustering - an introduction*. 2022. URL: https://www.mathworks.com/help/releases/R2021b/stats/hierarchical-clustering.html (visited on 21/06/2022).

[26] Mathworks. *Dendrogram plot*. 2022. URL: https://nl.mathworks.com/help/stats/dendrogram.html (visited on 21/06/2022).

[27] David L Davies and Donald W Bouldin. 'A cluster separation measure'. In: *IEEE transactions on pattern analysis and machine intelligence* 2 (1979), pages 224–227.

[28] Tadeusz Caliński and Jerzy Harabasz. 'A dendrite method for cluster analysis'. In: *Communications in Statistics-theory and Methods* 3.1 (1974), pages 1–27. DOI: 10.1080/03610927408827101.

[29] Peter J Rousseeuw. 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis'. In: *Journal of computational and applied mathematics* 20 (1987), pages 53–65.

[30] Robert Tibshirani, Guenther Walther and Trevor Hastie. 'Estimating the number of clusters in a data set via the gap statistic'. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2 (2001), pages 411–423. DOI: https://doi.org/10.1111/1467-9868.00293. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9868.00293. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00293.

[31] Robert L. Thorndike. 'Who belongs in the family?' In: *Psychometrika* 18.4 (1953), pages 267–276. ISSN: 1860-0980. DOI: 10.1007/BF02289263. URL: https://doi.org/10.1007/BF02289263.

[32] Esa Alhoniemi, Johan Himberg, Jukka Parviainen and Juha Vesanto. *SOM-Toolbox*. 2012. URL: https://github.com/ilarinieminen/SOM-Toolbox (visited on 29/05/2022).

[33] R.J. Veen, F. Westerman and M. Biehl. *A no-nonsense beginner's toolbox for GMLVQ, Version v3.1*. url: http://www.cs.rug.nl/~biehl/gmlvq, accessed 12/2021. 2021.

[34] Petra Schneider, Michael Biehl and Barbara Hammer. 'Adaptive relevance matrices in learning vector quantization'. In: *Neural computation* 21.12 (2009), pages 3532–3561.

[35] Thorsten Bojer, Barbara Hammer, Daniel Schunk and Katharina Tluk Von Toschanowitz. 'Relevance determination in Learning Vector Quantization.' In: *ESANN*. Volume 1. Citeseer. 2001, pages 271–276.

[36] B. Hammer and T. Villmann. 'Generalized Relevance Learning Vector Quantization'. In: *Neural Networks* 15.8-9 (2002), pages 1059–1068.

[37] Alfred Ultsch. 'Kohonen's self organizing feature maps for exploratory data analysis'. In: *Proc. INNC90* (1990), pages 305–308. URL: https://archive.org/details/innc90parisinter0001inte/page/305/mode/2up.

[38] Susmeeta T. Sharma and Lynnette K. Nieman. 'Cushing's Syndrome: All Variants, Detection, and Treatment'. In: *Endocrinology and Metabolism Clinics of North America* 40.2 (2011). Endocrine Hypertension, pages 379–391. ISSN: 0889-8529. DOI: https://doi.org/10.1016/j.ecl.2011.01.006. URL: https://www.sciencedirect.com/science/article/pii/S0889852911000077.

[39] D. M. Schonk, H. J. H. Kuijpers, E. van Drunen, C. H. van Dalen, A. H. M. Geurts van Kessel, R. Verheijen and F. C. S. Ramaekers. 'Assignment of the gene(s) involved in the expression of the proliferation-related Ki-67 antigen to human chromosome 10'. In: *Human Genetics* 83.3 (Oct. 1989), pages 297–299. ISSN: 1432-1203. DOI: 10.1007/BF00285178. URL: https://doi.org/10.1007/BF00285178.

[40] Jörn Bullwinkel, Bettina Baron-Lühr, Anja Lüdemann, Claudia Wohlenberg, Johannes Gerdes and Thomas Scholzen. 'Ki-67 protein is associated with ribosomal RNA transcription in quiescent and proliferating cells'. In: *Journal of Cellular Physiology* 206.3 (2006), pages 624–635. DOI: https://doi.org/10.1002/jcp.20494. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcp.20494. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcp.20494.

[41] Paul Hermanek and Christian Wittekind. 'Residual tumor (R) classification and prognosis'. In: *Seminars in surgical oncology*. Volume 10. 1. Wiley Online Library. 1994, pages 12–20.

[42] Alberto Biondi, Roberto Persiani, Ferdinando Cananzi, Marco Zoccali, Vincenzo Vigorita, Andrea Tufo and Domenico D'Ugo. 'R0 resection in the treatment of gastric cancer: room for improvement'. eng. In: *World journal of gastroenterology* 16.27 (July 2010). PMC2904881[pmcid], pages 3358–3370. ISSN: 2219-2840. DOI: 10.3748/wjg.v16.i27.3358. URL: https://doi.org/10.3748/wjg.v16.i27.3358.

[43] Mahul B. Amin, Stephen B. Edge, Frederick L. Greene, David R. Byrd, Robert K. Brookland, Mary Kay Washington, Jeffrey E. Gershenwald, Carolyn C. Compton, Kenneth R. Hess, Daniel C. Sullivan, J. Milburn Jessup, James D. Brierley, Lauri E. Gaspar, Richard L. Schilsky, Charles M. Balch, David P. Winchester, Elliot A. Asare, Martin Madera, Donna M. Gress and Laura R. Meyer, editors. *AJCC Cancer Staging Manual*. 8th. Springer International Publishing, 2017. ISBN: 978-3-319-40617-6. URL: https://link.springer.com/book/9783319406176.

[44] Oxford Advanced Learner's Dictionary at OxfordLearnersDictionaries.com. *Phenotype adjective - Definition, pictures, pronunciation and usage notes*. URL: https://www.oxfordlearnersdictionaries.com/definition/english/phenotype?q=phenotype (visited on 29/05/2022).

[45] Robert Bringhurst. *The Elements of Typographic Style*. Version 4.0: 20th Anniversary Edition. Point Roberts, WA, USA: Hartley & Marks Publishers, 2013.

colophon

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede and Ivo Pletikosić. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*" [45]. `classicthesis` is available for both LaTeX and LyX:

https://bitbucket.org/amiede/classicthesis/

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

http://postcards.miede.de/

Thank you very much for your feedback and contribution.

Final Version as of 4th August 2022 (e-print).