Development and investigation of chemometric baseline correction approaches and metabonomic classification algorithms

Bhaskaran David Prakash

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF PHARMACY

NATIONAL UNIVERSITY OF SINGAPORE

2015

DECLARATION

I hereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously

----

Bhaskaran David Prakash

29 **July 2015**

Acknowledgements

I would like to thank Dr Yap Chun Wei for his distinguished advice throughout the course of my PhD. I am grateful for Dr Yap in accepting me as a part time student and believing that I would be able to complete my PhD part time despite working full time on projects that were not related to my PhD topic. Due to the hectic schedule of my full time job, Dr Yap's quick email responses to my sometimes ad hoc queries were very helpful and convenient for me. I would also like to thank Dr Yap for maximizing my freedom to explore my topics of choice within the domain of metabonomic algorithms which allowed me to contribute two new independent and uncorrelated algorithms to the field of metabonomics.

I appreciate Dr Eric Chan Chun Yong and Dr Kishore Pasikanti for acquiring the GC/MS data that was used in the evaluation of the fully automated classification algorithm, Automated Pearson's correlation change classification (APC3) [4].

Table of Contents

Summary

Metabonomic analysis has been used for classification in a diverse range of areas from toxicology and dietary effects through to parasitology and molecular epidemiology [1], including disease diagnosis and therapy monitoring [2]. Metabonomic data requires correction via pre-processing approaches followed by post-processing involving a robust modelling approach to provide accurate and fast classification. In this work, we developed novel algorithms for both phases.

For pre-processing, we developed a baseline correction algorithm, Automated Iterative Moving Averaging (AIMA) [3], which has similar accuracy as existing semi-automated algorithms but is fully automated and computationally more efficient (28.6 to 197.7 times faster). AIMA baseline correction was developed based on the idea of a moving average smoother and evaluated on both simulated and experimental data from HPLC chromatograms, Raman spectra, surfaced enhanced laser desorption ionization time-of-flight (SELDI-TOF) chromatograms, LC-MS chromatograms and NMR signals.

For post-processing, we developed a fully automated classification algorithm, Automated Pearson's correlation change classification (APC3) [4], which has similar or better prediction accuracy as the current state of art algorithms for metabonomic data but is 3.9 to 7 times faster. APC3 involves correlation

based feature selection that is integrated to form a classification algorithm catered for the classification of two-class data. APC3 was tested on the total ion chromatograms (TICs) of two sets of two-class GC/MS datasets. APC3 was evaluated against various dimensionality reduction and classification combinations. 6 transformation methods and 12 variable selection techniques were each separately combined with 3 classification approaches to form a total of 54 current dimensional reduction and classification combinations which APC3 was tested against.

Finally, we did a comparative study on four sparsity embedded classification techniques, namely shrunken centroids regularized discriminant analysis (RDA) [3], nearest shrunken centroids (NSC) [4], sparse partial least squares - discriminant analysis (sPLS-DA) [5] and penalized linear discriminant analysis (PLDA) [6]. All these four methods have been previously performed primarily on microarray. Sparsity embedded classification approaches allows embedded sparse structures and dimensional reduction to complement each other resulting in the elimination of noisy variable which does not contribute to classification to make the classifier more accurate and predictive [7] and allow automatic variable selection [7]. These classification techniques were evaluated on three-class LC-MS chromatograms to study their suitability. RDA was found to be the most accurate whereas NSC was found to be the fastest.

List of Tables

List of Figures

Fig Ia. Boxplot of computational time (both training and testing phases) across the 100 sampling for urine sample splits using APC3, DDA-NB and LC-NB. The red triangle represents the mean value of each boxplot.

Fig IIb. Boxplot of computational time (only training phase) across the 100 sampling for urine sample splits using APC3, DDA-NB and LC-NB. The red triangle represents the mean value of each boxplot.

Fig IIb. Boxplot of computational time (only training phase) across the 100 sampling for urine sample splits using APC3, DDA-NB and LC-NB. The red triangle represents the mean value of each boxplot.

Fig IIIc. Boxplot of computational time (only testing phase) across the 100 sampling for urine sample splits using APC3, DDA-NB and LC-NB. The red triangle represents the mean value of each boxplot.

List of Abbreviations

AIMA	Automated Iterative Moving Averaging
airPLS	Adaptive iteratively reweighted penalized least squares
ALS	Asymmetric Least Squares baseline correction
APC3	Automated Pearson's Correlation Change Classification
AUC	Area under the response operating characteristic curve
BC	Bladder cancer
CA	Correspondence analysis
colAUC	Column wise area under receiver operator curve
CV	Cross-validation
DCA	Detrended correspondence analysis
DDA	Diagonal discriminant analysis
DKM	Dietterich, Kearns, and Mansour
DNA	Deoxyribonucleic acid
GC/MS	Gas chromatography mass spectrometry
Gini	Gini-index
HMDB	Human Metabolome Database
HPLC	High performance liquid chromatography
IA	Iterative Averaging
IAS	Iterative Averaging Smoothing
LC-MS	Liquid chromatography mass spectrometry
LDA	Linear discriminant analysis
locLDA	Localised linear discriminant analysis
LOO	Leave one out
MDL	Minimum Description Length
mRNA	Messenger RNA
NB	Naïve Bayes
NCA	Non Component Analysis
NMR	Nuclear magnetic resonance
NSC	Nearest shrunken centroids
PCA	Principal component analysis
PLDA	Penalized linear discriminant analysis
PLS	Partial least square
PLSR	Partial least squares regression
RDA	Shrunken centroids regularized discriminant analysis
RMSEP	Root mean error of prediction
RNA	Ribonucleic acid
RNA-Seq	Transcriptome sequencing
SELDITOF	Surfaced enhanced laser desorption ionization time-of-flight
sPLS-DA	Sparse partial least squares - discriminant analysis
TICs	Total ion chromatograms

Chapter 1: Introduction

1.1 Major "Omics"

The four major "Omics" disciplines are genomics, transcriptomics, proteomics and metabonomics [2, 8, 9]. Their brief descriptions, interactions and limitations would be addressed in the sub-sections below. A first cut conceptual diagram which is an enhanced adaptation of [10] is displayed in Fig. 1.1. In total, there are almost 200 different named "omics", most of which are highly specialized and may come under the umbrella of one of the four main "omics" [2] and hence would not be reviewed. An example is lipodomics, an "omic" involving the specialized study of lipid pathways and networks which is a subset of metabonomics since lipids are in fact metabolites [11].

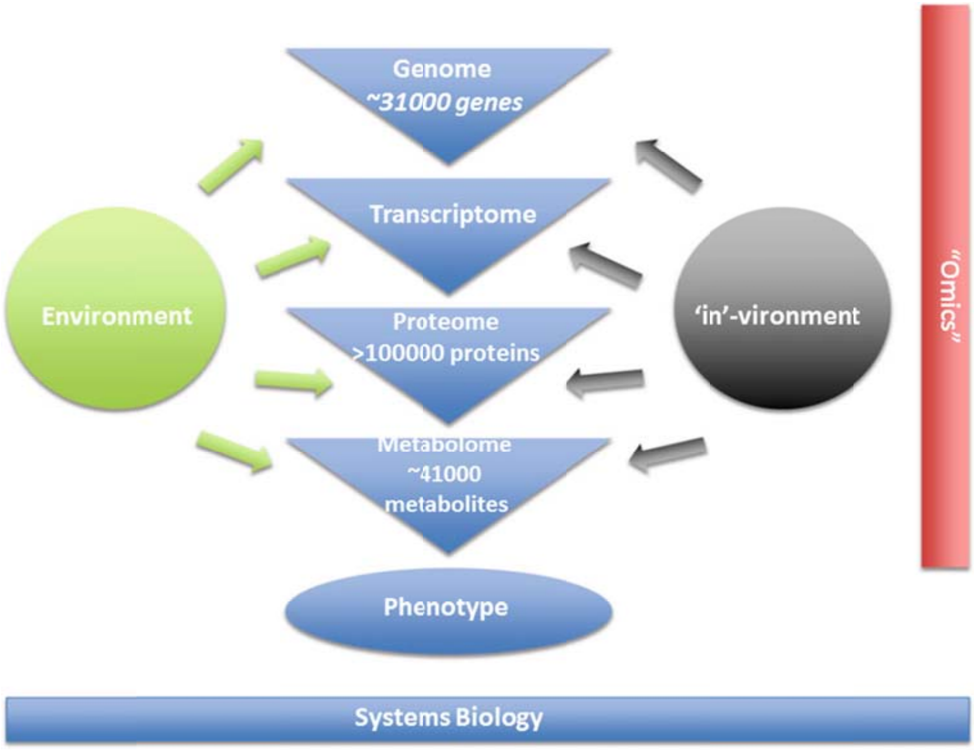


Fig. 1.1 The concept of systems biology as an integrated understanding of the immediate biological processes at play at different molecular levels or "omics" levels (genomics, transcriptomics, proteomics and metabonomics). The systems biology approach for a complete molecular characterization of the phenotype also includes taking the environmental (e.g., diet, food, and alcohol) and "in"-vironmental (e.g., gut microbiota) factors into account. Adapted and enhanced from [10].

1.1.1 Genomics

Genomics is the systematic study of the genome which is the total deoxyribonucleic acid (DNA) of a cell or organism [8]. A gene is a functional unit of DNA that controls a discrete hereditary characteristic and usually corresponds to a single protein or ribonucleic acid (RNA) [12]. In the human genome, there are about 31 000 protein encoding genes [13]. The use of DNA microarray technology permits the differential investigation of DNA sequences between individuals [8]. Genomics has enabled the discovery of markers that correlate genetic variation with biological outcomes such as new markers of the susceptibility of individuals to different diseases and response to specific therapies [9]. Understanding the response of a specific genetic profile to a specific therapy may provide insight to develop effective individualized therapy such as in multifactorial disease treatment for diseases like breast cancer [14]. The vulnerability to environmental influences is a confounder in genomics and epigenetics is an additional domain that needs to be further investigated to cope with this inherent susceptibility of genomics. Hence the exposure to epigenetic regulation complicates genomics [15]. Environmental changes may greatly affect metabolism making it difficult to dissect these influences from gene-related outcomes [16]. Fig. 1.2 graphically depicts the relationship between disease, genetic and environmental factors [17].

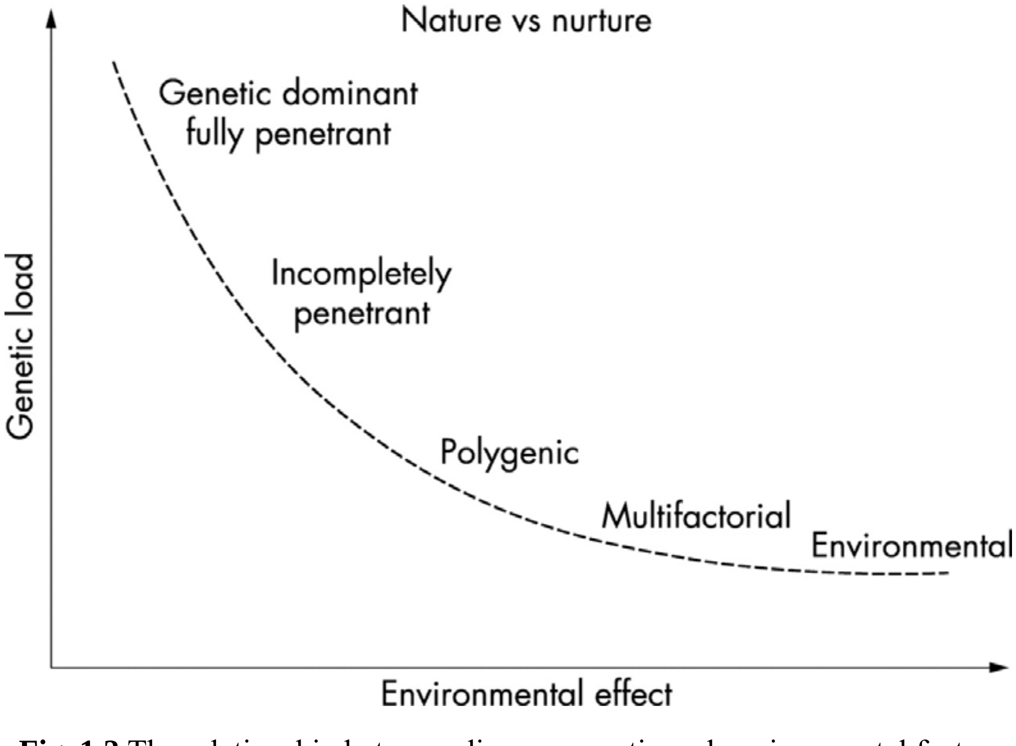


Fig. 1.2 The relationship between disease, genetic and environmental factors (adapted from [17]), which suggests that every disease has both genetic and environmental influences though the extent of each of these factors may vary.

1.1.2 Transcriptomics

Transcriptomics is the study of mRNA in a cell or organism [8] and a transcriptome is the total mRNA in a cell or organism [8]. Gene expression microarrays are used to measure gene expression or mRNA levels at a given time [8, 9]. It is important to note that gene expression microarrays used in transcriptomics differ from DNA microarrays used in genomics via the entity they measure. The former measure mRNA while the latter measure differences in DNA sequences [8].

Gene expression levels can be used to separate normal cells or tissues into their subtype classification, identify prognostic disease markers, identify disease state markers, sub-classify disorders that may appear similar on the surface and identify predictors of therapeutic response to facilitate effective individualized treatment [9]. Levels of mRNA are not always directly proportional to protein expression levels as demonstrated in mammalian culture cells [18-21] and 12 different normal human tissues [22]. This miscorrelation between mRNA levels and protein levels is possibly due to alternative splicing, post-translational modification of proteins [9] and the different time scales which gene expression and protein expression operate. These create difficulty in establishing causal linkages [23]. With levels of mRNA not always correlating to protein expression levels [18-22] and as gene expression microarrays measure changes in mRNA levels instead of protein, there is a lack of consensus relating to the interpretation of the data [8]. The transcriptome, just like the genome is exposed to environmental changes or epigenetic influences and these factors add constraints that need to be deciphered to prevent misinterpretation of gene profiling data [2, 8]. The study of epigenetic regulation of gene expression [24] is another enormous research domain that needs to be further understood for better interpretation of gene expression.

1.1.3 Proteomics

Proteomics is the large scale application of evolving technologies to study the biological functions of proteins and characterize proteins according to their appropriate functional cellular or protein pathways [9, 25, 26]. The proteome is the set of all expressed proteins in a cell, tissue or organism [27].

The presence of an immense diversity of proteins possibly due to alternative splicing and post-translational modification of proteins, is an advantage for the study of proteomics over gene expression in differentiating normal and abnormal cellular processes since more information is found in protein analysis than gene expression [9]. However, it is useful to note that the existence of a huge number of proteins (>100 000) also poses as a computational bottleneck for proteomic analysis which is further complicated by a lack of accurate detection of low-abundance proteins [8] due to the obscuring of the low-abundance protein detection by proteins with higher abundance [28-30]. The accurate detection of smaller, less-abundant proteins is itself a separate field of exploration where novel approaches of excluding large proteins from the analysis have been proposed [28-30]. Although proteomics is potentially less expensive than genomics, it can be slow and labour intensive [31]. Due to the *in-vitro* nature of genomic, transcriptomic and proteomic studies, it is difficult to correlate their time-response to drug exposure where an *in-vivo* multi-organ functional integrity in real time is preferred resulting in difficulty to relate these three domains to classical indices of toxicity or toxicological endpoints [31].

1.1.4 Metabonomics

Metabonomics involves the quantitative measurement of the global, dynamic metabolic response of living systems to biological stimuli or genetic manipulation with a focus on elucidating systemic change through time [16]. Metabolomics pursues an analytical descriptive analysis of complex biological samples, and aims to distinguish and quantify all the small molecules that are present [16]. The terms metabonomics and metabolomics were initially coined for slightly varied purposes where the former was first used in the chemometric interpretation of biological fluids and tissues analysed via nuclear magnetic resonance (NMR) spectroscopy [32] whereas the latter was first used in plant science and the study of *in vitro* systems [33]. The terms metabonomics and metabolomics have since converged and are being used interchangeably as they essentially share the same analytical and modelling procedures [2, 16] and both try to characterize the metabolome [10]. From the literature review via PubMed keyword searches as shown in Fig. 1.3, there is an increase in the number of publications for both “metabonomics” and “metabolomics” in the last decade though there appears to be a consistent preferential use of the term “metabolomics” which is also the term that is alphabetically closer to the term “metabolome”.

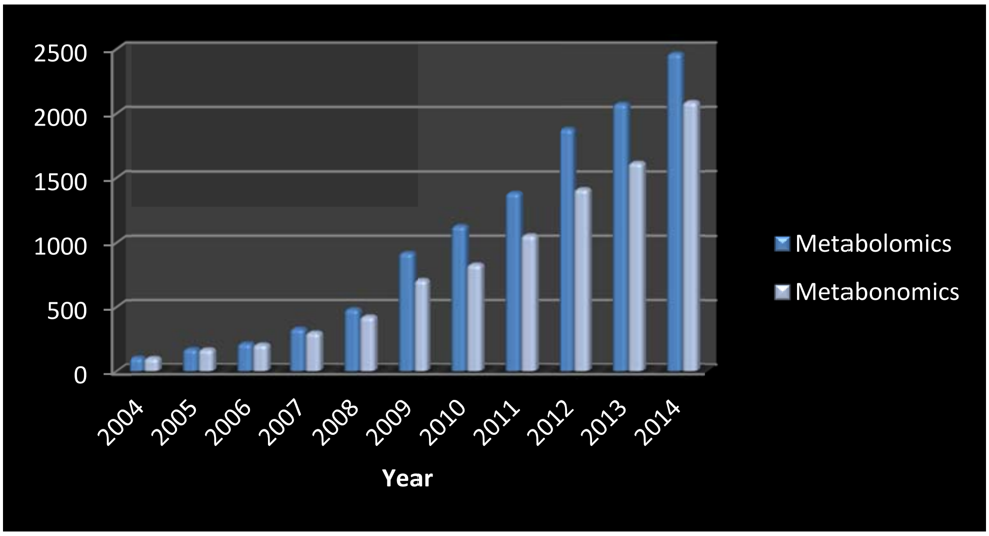


Fig. 1.3 Number of metabolomics and metabonomics publications; literature review was conducted using PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) with the keywords (a) (metabolomics) and (b) (metabolomics)

Metabolome is derived from the ancient Greek word *metabol* referring to change. The metabolome is the total quantitative collection of low molecular weight compounds (metabolites) present in a cell or organism that participates in metabolic reactions which also includes those metabolites taken in from external environments or symbiotic relationships [8]. Metabolites are small molecules that include peptides, lipids, amino acids, nucleic acids, carbohydrates, organic acids, vitamins, minerals, food additives, drugs, toxins, pollutants and just about any other chemical (with a molecular weight <2000 Da) that humans ingest, metabolize, catabolize or come into contact with [34]. The most recent Human Metabolome Database (HMDB), version 3, gives the total human metabolites count at 40153, of which 20900 has been detected and the rest belong to those for which biochemical pathways are known or human intake/exposure is frequent but the compound has yet to be detected in the body [34].

1.1.5 Advantages of Metabonomics

Metabonomics has the advantage of not mandating analyte preselection [16] whereas the converse is true for the other three major “omics” which encompasses pre-determination of the analytes in order to select the appropriate sample preparation and detection platforms [31, 35]. Moreover, via the selective use of biofluids such as urine, metabonomics convey minimal invasion since these biofluids are essentially collected in a non-invasive manner [31, 35]. The option of preselecting the analytes which are metabolites in metabonomics dissects metababononomic investigations into targeted and untargeted and this division would be elucidated in the Section 1.4 [15]. In the presence of external influencing factors such as environment, the interpretation of other “omic” data such genomics and proteomics becomes non trivial but metabonomics overcomes this challenge by monitoring the global outcome of all the influencing factors, without making assumptions about any single contributing effects to that outcome [16, 35]. Since metabolites are not directly encoded in the genome, unlike RNA and proteins, metabonomics also provides information to aid deciphering metabolic pathways, which can be used to understand biological mechanisms better [36]. As the final downstream product of gene transcription, the metabolome inherits relatively amplified changes in comparison to the transcriptome and proteome [37] and is the closest to the phenotype of the biological system investigated [8]. Furthermore, as metabolic biomarkers are closely correlated to real biological end-points, metabonomics makes hypothesis generation studies easier [31, 35].

Metabonomics has a smaller domain than proteomics giving it a computational edge since the number of features in terms of metabolites is lesser than the number of proteins that exist in nature. Despite having the smaller domain than proteomics, the metabolome contains a diverse range of biological molecules making it physically and chemically more complex than the other "omics" [8], implying possibly greater informative content. Metabolic biomarkers have higher cross species flexibility than transcriptomic or proteomic biomarkers since metabolites do not differ as frequently across species which is important for pharmaceutical studies [35]. Compared to the other "omics", metabonomics is less expensive with lower cost per sample and per analyte and also less labour intensive [8, 35]. These cost and labour efficient attributes of metabonomics are driven by its better technological advances that includes its analytical procedures being stable and robust and with high degree reproducibility [8, 35].

1.2 Applications of Metabonomics

Metabonomic analysis has been used for classification in a diverse range of areas from toxicology and dietary effects through to parasitology and molecular epidemiology [1], and including disease diagnosis and therapy monitoring [2]. We attempt to organize them into three broad spectrums, namely identifying biological targets, individual profiling and population profiling [16]. These would be discussed in the proceeding sub-sections.

1.2.1 Identifying biological targets

With regards to preclinical toxicity, metabonomics enables

- detection of toxic biomarkers while investigating the adverse effects of candidate drugs preclinically [38-45]. It should be noted that this may differ between the preclinical species [46], and there is a possible inherent and unavoidable uncertainty of these preclinical toxicity markers with respect to humans.
- identifying relevant time points for these studies [46]
- better study design to help reduce animal count, expenses and output more reliable results [47]

An interesting controversial notion that needs awareness and may appear as a possible limitation is whether the metabonomic approach is appropriate to study toxic effect, especially when non-metabonomic markers are not able to conclude metabolic disruption which is only sensitive via metabonomic markers [48].

With respect to disease diagnosis, metabonomics facilitates

- clinical disease diagnosis such as ovarian cancer [49], meningitis [50], prostate cancer [51], inborn errors of metabolism [52], coronary heart disease [53], renal cell carcinoma [54] and various brain tumours [55]

- drug efficacy study on cardiac mouse models leading to phenotyping of four mouse models of cardiac disease [56]
- developing preclinical assessments of metabolic response to drug therapy which may aid in differentiating efficacious from toxic effects [47, 57]
- clinical disease progression monitoring such as cerebrospinal fluid in aneurysmal subarachnoid haemorrhage [58] and prostate cancer [51]
- differentiating closely related disease types such as distinguishing between brain tumour types [55]

In plant science, metabonomics enables

- detection of metabolite markers for predicting crop yield, which can be used to develop transgenic strategies for yield enhancement [59]
- discovery of metabolite markers for desirable characteristics in plant breeding [60, 61]

1.2.2 Individual profiling

Pharmaco-metabonomics is known as the approach that uses metabonomics to create personalized drug treatment, which can improve efficacy and limit the instances and severity of adverse drug reactions [62]. It has been studied on both preclinical [62, 63] and clinical subjects [64].

Pharmaco-metabonomics is more superior compared to a conceptually similar approach known as pharmacogenomics, which differs by using genomics instead of metabonomics but aspires to achieve a similar end point that is individualized drug treatment [62]. The pre-eminence of pharmacometabonomics over pharmacogenomics lies simply in the choice of using metabonomics instead of genomics, which implies its added ability to factor in environmental influences [62].

1.2.3 Population profiling

A sound understanding of a normal biochemical profile would aid to connect therapeutic or toxic effects to normality or to elucidate disease associated biochemical alterations [35]. The following are examples of various effects that have been investigated via preclinical studies involving urine samples

- diurnal variation, gender, age, diet, species, strain, hormonal status and stress [65]
- age, strain, gender and diurnal variation [66]
- xenobiotics within the environment or environment toxins [67-69]

Phenotypic effects can be studied via metabonomic approaches after gene transfection in animal models [70], which may provide insights into the usefulness of these trans genetic animals as disease models or in drug efficacy studies [35].

In plant science, metabonomics contributes to

- classifying herbicidal mode of action by relating phenotypical end points with physiological processes of herbicide application so that the outcome of using new potential herbicides can be anticipated [71-74]

- study the response of plants to abiotic stresses where degrees of tolerance in species and genotype have been shown [75]
- study the response of plants to biotic stresses via the metabolic profiling of volatiles [76]

1.3 Relative vs Absolute Quantification

In genomics and transcriptomics, data generated from microarrays are relative expression ratios of the same gene under different conditions [77]. Hybridisation efficiency [78], cross-hybridisation issues, limited dynamic detection range, presence of background noise and the detection of transcripts being limited to sequences printed on the array [79] constitute to the limitations present in microarray generated data. These restrictions disallow the comparison of absolute expression levels across genes using microarray [77].

However, transcriptome sequencing (RNA-Seq) gives expression levels in terms of counts of expressed transcripts that can be related to transcripts per cell which is an absolute level. Therefore, the expression levels generated are comparable across the transcriptome and have been shown to be more indicative of protein concentrations than gene expression levels generated from microarrays [78]. In cancer genomics, absolute quantification of the copy number changes and point mutations is preferred over relative quantification for the identification of oncogenes and tumour suppressors and paints a better picture regarding the tumour subclonal architecture and evolution [80, 81].

In proteomics, relative quantitation is the comparison of the levels of a specific protein in different samples with results being expressed as a relative fold change of protein abundance [82], whereas absolute quantitation is the exact determination or mass concentration of a protein, for example, in units of ng/mL of a plasma biomarker [83]. Absolute quantification of proteins can be obtained by either labelling via spiking known amounts of stable isotope–labelled standards into the samples or by a label free approach of computationally comparing peptide signals of different samples [83, 84].

In metabonomics, when comparing across studies, having an absolute concentration to compare against would be ideal but it requires having to define a reference metabolite or standard to quantify against [85]. With a reference material of known fix concentration, the absolute concentration of the metabolites under analysis can be determined in relation to the known concentration of the reference material. However, no standard reference material (SRM) has been made available yet. Hence in place of an endogenous reference material, metabolite spike-ins are recommended for use at the moment [85].Therefore, it is still possible to perform cross-study analysis using metabonomics with metabolite spike-ins to account for technical normalization. But the search for a suitable endogenous reference metabolite which can additionally account for biological variation across studies remains an open area of research.

1.4 Targeted vs non-targeted Metabonomics

Non-targeted metabonomics can be defined as

"Non-biased identification and quantification of all metabolites in a biological system. The analytical technique(s) must be highly selective and sensitive. No one analytical technique, or combination of techniques, can currently determine all metabolites present in microbial, plant or mammalian metabolomes [86]."

Targeted metabonomics can be defined as

"Quantitative determination of one or a few metabolites related to a specific metabolic pathway after extensive sample preparation and separation from the sample matrix and employing chromatographic separation and sensitive detection [86]."

Untargeted metabonomics requires coupling to advanced chemometric techniques, such as multivariate analysis, to reduce the extensive datasets generated into a smaller set of manageable signals. This require annotation using either in silico libraries or experimental investigation and subsequent identification using analytical chemistry [87]. Untargeted analysis is suitable for novel target discovery with the metabolome coverage only restricted by the sample preparation methodologies and the inherent sensitivity and specificity of the analytical technique used [87]. The bottlenecks of this approach lie in the protocols and time required to process the huge amount of generated raw data, the complications in identifying and characterizing unknown small molecules, the dependence on the intrinsic analytical coverage of the platform employed, and the bias towards detection of high-abundance molecules [87].

On the contrary, targeted metabonomics can be constructed in a quantitative or semi-quantitative manner through the use of internal standards and takes advantage of the comprehensive understanding of a vast array of metabolic enzymes, their kinetics, end products, and the known biochemical pathways to which they contribute [87]. Furthermore, sample preparation can be optimized which reduces the dominance of high-abundance molecules in the analyses [87]. Due to a well spelt out selection of analysed metabolites, analytical artefacts do not flow through to downstream analysis and metabolic associations to specific physiological states can be studied [87].

1.5 Metabonomic Workflow

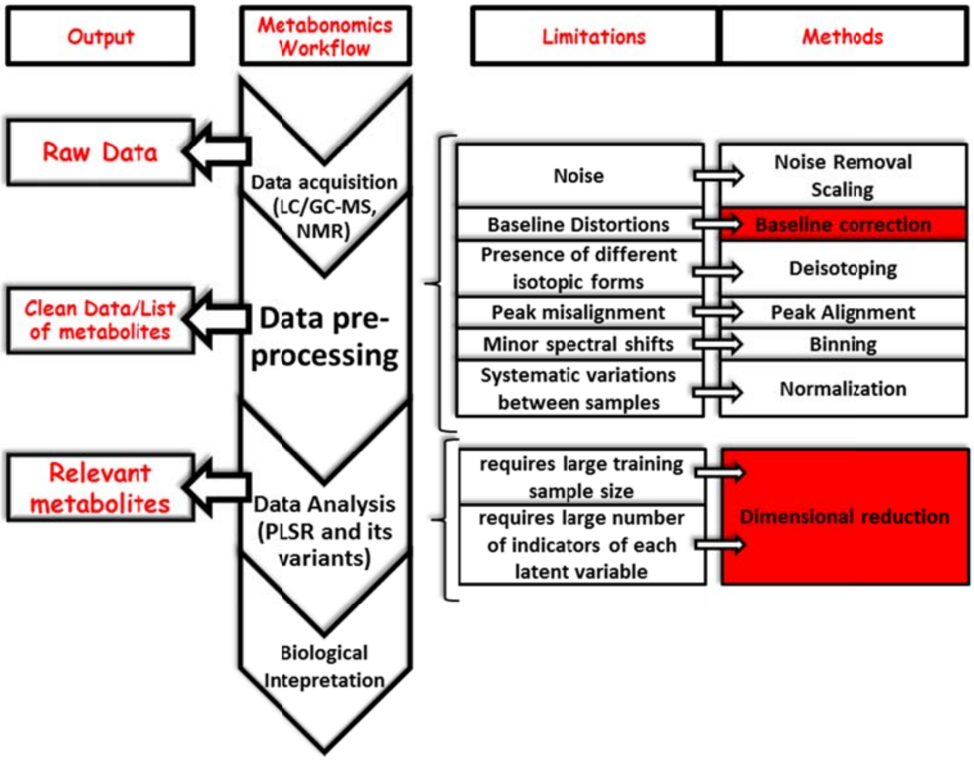


Fig. 1.4 A simple metabonomic workflow that depicts the outputs, limitations and some methods proposed to minimize these limitations. Our investigations would focus on baseline correction and dimensional reduction which are highlighted in **red background**.

The metabonomic workflow as illustrated in Fig. 1.4 begins with data acquisition followed by data pre-processing, then by data analysis and finally the biological interpretation [88].

1.5.1 Need for Pre Processing

The ideal chemical spectrum for LC-MS, GC/MS or SELDITOF should have well-resolved peaks, adequate signal-to-noise ratios, no background contribution, and a large linear response range between analyte concentration and detector signal for individual samples or runs [89]. If more than one single sample is used, having stable retention times and well-defined peak shapes is ideal [89]. However, due to sample complexity and increasing speed of the chromatographic runs, artefacts such as baseline drifts, changes in the peak shapes and elution times shifts are inherent [90].

1.5.2 Issues with current baseline correction approaches

Baseline correction is one of the components of the chemometric data preprocessing phase to counter the baseline drifts. Manual baseline correction though commonly used in vibrational spectroscopy, tends to have bias towards user experience, noise levels and baseline characteristics [91]. For automatic baseline correction, we can divide it into fully-automated [8-12] and semi-automated [13-20]. The predominant method for fully automated baseline correction is polynomial fitting. Polynomial fitting have been shown to perform badly for low signal to noise and signal to background spectrums [92, 93] for Raman spectroscopy. In NMR data, even with commercially available polynomial baseline correction, manual correction might sometimes be necessary [94]. Some variants of polynomial fitting have been shown to be suitable for only broad and smooth baseline deviation [95].

Semi-automated baseline correction, such the most recently introduced penalized least squares variant, the adaptive iteratively reweighted penalized least squares (airPLS) [96], generally have better accuracy than current fully automated baseline correction. One general disadvantage of semi-automated methods is the need to optimize parameters. Other semi-automated penalized square approaches resulted in negative valued regions [8, 14]. Although the default values for these parameters are available for different signals such as NMR, Raman and HPLC chromatograms [96], the accuracy of baseline correction depends on the careful optimization of these parameters.

Semi-automated wavelet baseline correction techniques transform the signals into different frequency components, followed by the removal of the varying low-frequency background to finally reconstruct the signal from the wavelet coefficient. This reconstruction results in some loss of spectra information and can cause distortion at some part of the spectra [92]. Wavelet based algorithms assumes that the background is well separated in the transformed domain from the signal, which may not be correct for real-world spectra [97].

1.5.3 Limitations of current data analysis methods

Currently, partial least squares regression (PLSR) and its variants are the preferred approach in metabonomic data modelling and classification due to their flexibility and accuracy in catering to the complexity of these data [98] including their suitability in handling the issue of multicollinearity [99]. However, PLSR typically requires large training sample size and large number of indicators of each latent variable [100, 101] which may be disadvantageous for rare metabonomic datasets such as those of rare diseases. In addition, it would be of interest to reduce PLSR's training complexity and hence the processing time when dealing with metabonomics data such as gas chromatography mass spectrometry (GC/MS) total ion chromatograms (TICs), which tend to be very large [102]. A common method to reduce the computational complexity of classification is to use dimensionality reduction approaches prior to classification. Dimensionality reduction techniques can be broadly divided into variable selection and transformation. Variable selection approaches can identify the significant variables but may not perform well when the data is highly correlated. Transformation based approaches tend to combine variables without selecting a subset of significant variables. There are many different dimensional reduction approaches and this increases the complexity of finding an optimum dimensionality reduction approach for PLSR and its variants for each metabonomics data. Hence it would be useful to develop a simpler modelling approach to address these problems.

1.5.4 Approaching dimensional reduction via sparsity embedded approaches

Metabonomic data such as TICs being high dimensional in nature exposes themselves to an array of inherent issues. A TIC is high dimensional because there are many time points where the total intensity is measured and each time point is a feature or dimension. Firstly, high dimensional data have been shown to have an inverse exponential relationship between the optimal rate of convergence and the dimension of the data under regularity assumptions, thereby impeding learning [103]. Next is the concentration phenomenon, which describes the difficulty in performing inference due to the presence of an inverse proportionality linking the Euclidean distances between feature vectors and the data's dimension [104].

Amongst the plethora of approaches being proposed for dimensional reduction, to export important features necessary for modelling is a recent concept that introduces the use of sparse representations [105]. A sparse representation uses a basis to transform the dimensions into a linear combination of a few dimensions. The basis can either be predefined using functions such as wavelets or adapted directly from the data presented. The application of bases learnt directly from the data being modelled has been shown to be more compact with improved performances in comparison to predefined bases [63-65].

The high dimensionality of the data and in some instances the small number of available training samples for modelling proposes the use of dimensional reduction to ease the adaptive learning of the bases. However many existing dimensional reduction methods are not known to complement the embedded sparse structures. Even dimensional reduction approaches that use sparse linear models are not recommended for sparse learning as the basis has been predefined [66,67]. Another drawback of many existing sparse representations is their inability to handle non-linear relationships. The addressing of non-linearity has been shown to improve classification [68-71].

1.6 Significance of Project

Metabonomic data requires correction via pre-processing approaches followed by post processing involving a robust modelling approach to provide accurate and fast prediction.

Vast developments were observed in metabonomics over the last decade. These developments were particularly focused in modelling methods, rather than simply via advances in the supporting analytical platforms and biosampling modalities. However, further contributions are needed in these areas, including enhanced mathematical analysis [1].

Our project intends to contribute ideas to both phases beginning with pre-processing. The study of these two consecutive phases would provide a more cohesive evaluation which would open up the possibility of re-designing of concepts back and forth from pre and post processing thereby providing complementary enhancements to each phase. The main aim is to develop at least one new fully automated algorithm to pre-processing phase, another new fully automated algorithm to the post-processing phase and a survey on new novel approaches that are relatively new to metabonomics for classification. Metabonomic data would include predominantly gas chromatography mass spectrometry (GC/MS), liquid chromatography – mass spectrometry (LC-MS) and nuclear magnetic resonance (NMR).

Current metabonomic pre-processing and post processing techniques are focused on semi-automated approaches, which tend to have better accuracy due to the flexibility of having user defined parameters but these have to be optimized. There are two ways to perform parameter optimization. One is via an exhaustive search and the other is through the use of modelling [106]. The exhaustive search involves a grid search of the entire parameter space and is hence time consuming. Modelling approaches such as the shrinking hypercube method [107, 108] are less time consuming as they utilize scores from previous optimization tuning rounds as a guide to obtain a local maximum. The grid search approach ensures the global maximum is reached since the full parameter space is evaluated at the cost of a larger computational load. In contrast, modelling approaches estimate a local maximum at the cost of a reduced computational overhead. Either way, parameter optimization is an additional computational cost that is avoided in automated algorithms, which is beneficial when processing bulk metabonomic datasets. With the advent of Big Data initiatives such as more funding for larger studies and clinical trials [109], the advantage of having computationally efficient automated pre and post processing algorithms is slowly becoming a necessity.

Moreover, automation reduces user interactivity, thus allowing efficient processing of extremely large datasets in a high throughput approach [102], and can reduce any bias that may be present manual processing [110]. With reference to baseline correction, it has been acknowledged that there does not exist any perfect baseline correction method that performs well for all regions of a spectra and hence the best approach would be to test several baseline correction techniques to achieve the best prediction [111, 112]. Thus, fully automated and fast baseline correction approaches would greatly facilitate such multiple testing by reducing the amount of user interactivity that is required.

For post processing, we also intend to study the possibility of introducing new novel classification techniques that has not been extensively used in the metabonomics realm but have reduced complexity so that they could be applied to deal the issues of high dimensionality of extremely large datasets [113] for the first time on three-class and two-class TIC datasets.

The introduction of new algorithms in literature usually does not include readily downloadable software for communities to easily utilize them. In our developments, we assure the availability of plugins to ease their evaluation and usage by the metabonomic communities. Presentation would conform to open source technologies and wherever possible, plugins to open source metabonomics processing tools such as MZmine [114] would be made publicly available to benefit the metabonomics community.

Sparsity embedded classification algorithms introduce shrinkage as an integral part of their classification procedure. By embedding sparsity within the classification algorithm, we may have the best of both worlds via the embedded sparse structures and dimensional reduction complementing each other. In other words, the dimension is reduced while promoting the embedded sparse structures. The two-fold benefits of embedding shrinkage within a classification technique are:

1. the elimination of noisy variable which does not contribute to classification to make the classifier more accurate and predictive [7]
2. to achieve automatic variable selection [7]

We have three hypotheses. The first hypothesis is it is possible to develop a fully automated baseline correction algorithm that has similar accuracy as semi-automated algorithms. The second hypothesis is it is possible to develop a fully automated classification technique that has similar prediction accuracy as the current state of art algorithms for metabonomic data. The third is current sparsity embedded classification techniques can classify multi-class metabonomic TIC datasets.

1.7 Thesis Structure

Chapter 1 would deal with the aspects of providing a basic simplified but yet informative introduction to the field of metabonomics, its place in comparison to the other "omics", its advantages over the other "omics", its various attempted applications, its sub categories such as the use of relative versus absolute quantification and targeted versus non-targeted metabonomics and its processing workflow with the limitations that were addressed in this thesis.

Chapter 2 would describe the fully automated baseline correction technique which we developed [115] and its results where it had been shown to have similar accuracy as current semi-automated baseline correction methods when tested on HPLC chromatograms, Raman spectra, surfaced enhanced laser desorption ionization time-of-flight (SELDI-TOF) chromatograms, LC-MS chromatograms and NMR signals.

Chapter 3 would introduce the fully automated classification algorithm which we proposed [116] and the results of its evaluation on two sets of two-class GC/MS TICs against other classification algorithms, classification algorithms in combination with transformation techniques and classification algorithms in combination with variable selection approaches.

Chapter 4 would display the evaluation of four current sparsity embedded classification approaches in terms of prediction accuracy, variable optimization and time complexity on a three-class metabonomic TIC dataset and another two-class metabonomic TIC dataset that was used in Chapter 3.

Chapter 5 would summarize the main contributions, state the limitations of my work and propose future direction based on the limitation.

Chapter 2: A fully Automated Iterative Moving Averaging (AIMA) technique for baseline correction

2.1 Introduction

The presence of baseline drifts in chemometric data establishes the need to employ techniques such as baseline correction. One way to segregate the various baseline methods in literature is via their feasibility to be automated. Those that come under the umbrella of being non-automated, also referred to as manual baseline correction, tend to be user biased [91]. As for automated baseline correction, further subdivision into fully-automated [8-12] and semi-automated [13-20] is possible. However, the former has issues for certain types of spectrum such as those with low signal to noise and signal to background [92, 93], though showing suitability for broad and smooth baseline deviation [95]. The latter, hovering midway between being fully automated and manual, requires the optimization of parameters at the cost of generally producing better accuracy than current fully automated baseline correction methods [96].

Hence, in an attempt to attain both the higher accuracy of semi-automated and the redundancy of parameter optimization of fully automated baseline correction methods, we developed a novel fully automated baseline correction method which has similar accuracy as semi-automated baseline correction methods.

2.2 Methodology

The moving average is an univariate spectral filtering method used in chemometrics [117]. The most recent use of the moving averaging as a baseline correction technique was in a computational tool, LIMPIC [118] where the baseline was estimated using a simple linear interpolation of the average values of signals with selected segments. AIMA was developed based on this idea of moving average smoother.

The algorithm is divided into two steps. The first involves getting a baseline of a spectrum where the peaks are not maximized. Starting with an array of intensities with equal interval $y = [y_1, y_2, \dots y_N]$, the first iteration updates the even intensities as follows:

$$y_{i+1} = \min(y_{i+1}, (y_i + y_{i+2})/2) \qquad (2.1)$$

where i=1,3,5,…,N-3,N-1

The next iteration updates the odd intensities as follows:

$$y_{i+1} = \min(y_{i+1}, (y_i + y_{i+2})/2) \qquad (2.2)$$

where i=2,4,6,…,N-4,N-2

The following iteration updates the even intensities but leaves out the first and last update as follows:

$$y_{i+1} = \min(y_{i+1}, (y_i + y_{i+2})/2) \tag{2.3}$$

where i=3,5,…,N-5,N-3

In a similar note, the next iteration involves the updating the odd intensities with the first and last updates being left out as follows:

$$y_{i+1} = \min(y_{i+1}, (y_i + y_{i+2})/2) \tag{2.4}$$

where i=4,6,…,N-6,N-4

The stopping criteria is when the first update, i, reaches the floor (N/2) where N is the number of intensities. This is known as the Iterative Averaging (IA) procedure, which will be reused with slight modification in Step 2.
Next, consecutive segments are formed from the initial intensity array where the first and last intensity in each segment is equal to the corresponding intensity value of the derived intensity array y. The first and last intensity positions of these segments are noted and used to update the initial intensity array with linear interpolated intensity values.

Step 2 involves an iterative procedure to maximize the peak. Fig. 2.1 shows the maximization of the peak after the first and second calls to the Iterative Averaging Smoothing (IAS) function. Step 2 starts with a call to the IAS function using the output array from Step 1. The first part IAS involves a loop creating a copy of y known as y'. In the loop, the first iteration updates the even intensities as follows:

$$y'_{i+1} = (y'_i + y'_{i+2})/2 \tag{2.5}$$

where i=1,3,5,…,N-3,N-1

The next iteration updates the odd intensities as follows:

$$y'_{i+1} = (y'_i + y'_{i+2})/2 \tag{2.6}$$

where i=2,4,6,…,N-4,N-2

The following iteration updates the even intensities but leaves out the first and last update as follows:

$$y'_{i+1} = (y'_i + y'_{i+2})/2 \tag{2.7}$$

where i=3,5,…,N-5,N-3

In a similar note, the next iteration involves the updating the odd intensities with the first and last updates being left out as follows:

$$y'_{i+1} = (y'_i + y'_{i+2})/2 \tag{2.8}$$

where i=4,6,…,N-6,N-4

The stopping criteria is when the first update, i, reaches the floor (N/2) where N is the number of intensities.

Next a new array y_{max} is created using

$$y_{max,i} = max(y'_i\, , y_i) \tag{2.9}$$

where i=1,2…N

Another array y_{min} is created as

$$y_{min,i} = min(y_{max,i}\, , y_i) \tag{2.10}$$

where i=1,2…N

Using y_{min} and y_{max} another array y_{diff} is created using

$$y_{diff,i} = |(y_{max,i} - y'_{min,i})| \tag{2.11}$$

where i=1,2…N

Then we update every consecutive segments of y' where $y_{diff,i} = 0$ with a linear interpolation using the first intensity, y_f and the last intensity, y_l of that particular segment as follows:

$$y'_i = y_f + k(y_f - y_l)/(M) \tag{2.12}$$

where k=0,1,2….M

This is known as IAS.

Suppose the return intensity array, of IAS is y', another array y'' is created using

$$y''_i = min(y_i, y'_i) \tag{2.13}$$

where i=1,2,…,N

Next we get a value A_{abs} using

$$A_{abs} = \sum |(y''_i - y_i)| \tag{2.14}$$

Then we repeat the IAS except instead of creating a copy in the start of IAS, y'' is used as y' and y is reused.

Suppose the return intensity array, of IAS is a modified y', another array y'' is created using

$$y''_i = min(y_i, y'_i) \tag{2.15}$$

where i=1,2,…,N

Next we get a value B_{abs}using

$$B_{abs} = \sum |(y''_i - y_i)| \tag{2.16}$$

Using a1 and a2, we get a value

$$C_{abs} = A_{abs}/B_{abs} \tag{2.17}$$

Step 2 is repeated until the current C_{abs} is less than the previous C_{abs}.

Fig. 2.1 shows a demonstration of AIMA on an NMR spectrum from [110].

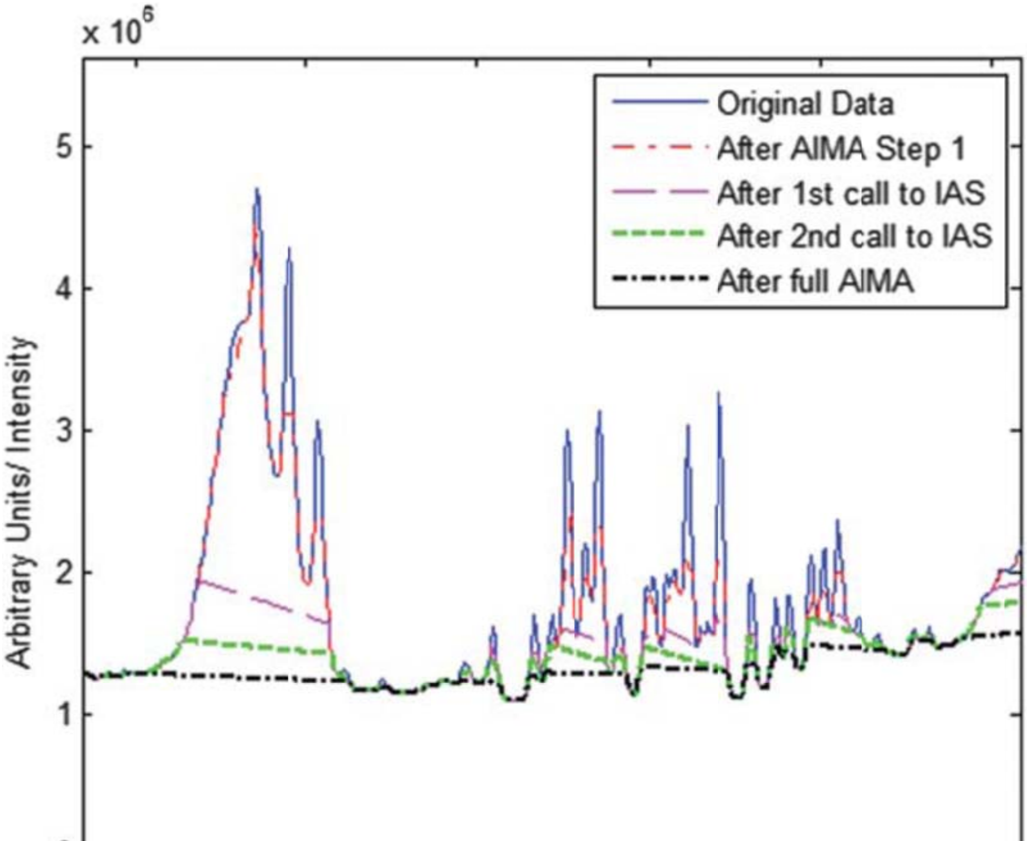


Fig. 2.1 Zoom in on a sample NMR spectrum from [110]. Baseline correction after step1, 1st call to IAS functions in step 2, 2nd call to IAS function in step 2 and after full AIMA (steps 1 and 2).

Both simulated and experimental data were used to evaluate and compare the performance and speed of the AIMA algorithm. All data were compared with three other semi-automated baseline correction techniques, airPLS, Asymmetric Least Squares baseline correction (ALS) [119, 120] and a parametric baseline correction [121]. airPLS is the most recently introduced baseline correction technique and it is shown to give better accuracy than ALS [122]. Although ALS was shown to have poorer accuracy than airPLS, we decided to include it in our comparison because our set of experimental data was larger than in the earlier study. Thus, it will be interesting to further compare airPLS and ALS on this larger experimental data. Parametric baseline correction [121] is a recent NMR baseline correction method, which has been shown to give better results than a commercial automatic baseline correction function in XWINNMR 3.5.

2.2.1 Simulated data

Simulated data were used because the actual peak heights were known and thus it is possible to compute the baseline correction relative error of the AIMA algorithm. Data were simulated using three different baselines, which are convex curved, concave curved and linear. Pure signals of three Gaussian peaks were used. Each peak varied in intensity. Random noise was also added to the spectrum. Mathematically, the spectrum can be expressed as follows:

$$s(x) = a(x) + b(x) + r(x) \tag{2.18}$$

where $s(x)$is the simulated signal, $a(x)$is pure signal peaks, $b(x)$ is the baseline (either convex, concave or linear), and $r(x)$is random noise.

A range of noise factor was multiplied to the random noise created to evaluate the ability of the algorithm to perform baseline correction in both high and low noise environment. Values from the sets {0.01, 0.02.., 1} and {1.1, 0.1.., 11} were used for low and high noise ranges respectively. For each noise factor the baseline correction was performed with 10 newly generated random noises to minimize bias.

As the simulated data and parameters ($\lambda = 10, p = 0.001\ and\ d = 2$) were similar to an earlier study using airPLS and ALS, the optimum parameters for both of these methods were obtained from that study [96]. The parametric method required estimation of the standard deviation of noise, σ_p. A systematic search showed that a value of $1\ * 10^3$ for this parameter was optimal for the simulated data of low noise and a larger value of $1\ * 10^1$ was suitable for high noise data. Appendix - Fig. S1 and S2 shows a plot of a curved convex baseline spectrum simulated with noise factors of 0.01 and 11 respectively with various baselines estimated using different estimated standard deviations of noise.

2.2.2 Experimental data

Experimental data from HPLC chromatograms, Raman spectra, surfaced enhanced laser desorption ionization time-of-flight (SELDI-TOF) chromatograms, LC-MS chromatograms and NMR signals were used to show the applicability of the AIMA algorithm in actual data sets. Information about these data sets is shown in Table 2.1.

For HPLC and Raman spectra, comparison of the three methods for such univariate data were done by measuring the reduction of convex hull of the PCA plots. This is because the compactness and separation in principle components pattern space would improve clustering and classification results [96, 123]. λ values of 30 and 50 which were previously used in airPLS for the HPLC and Raman datasets respectively [96] were assumed to be optimal. For ALS, we performed a grid search using p values from the set {0.001, 0.011... 0.081, 0.091} and λ values from the set {10, 20… 490, 500} and determined the optimal parameters that have the minimum reduction in convex hull for both HPLC and Raman datasets. The parametric method required the optimization of the estimated σ_p which was obtained by doing a grid search on the set of values of {10000, 1000, 100, 10, 1} and {10, 20, 30… ,C} and assuming the optimal to have the minimum reduction in convex hull for both HPLC and Raman datasets. C is the ceiling of the maximum of the standard deviation of every spectrum and was calculated to be 150 for HPLC datasets and 7250 for the Raman datasets.

For both SELDI-TOF and LC-MS spectra analysis, we compared the ability of the algorithms to improve the prediction performance of partial least square (PLS) models. For the SELDTI-TOF data, we selected 64 spectra per class to form a training set for developing PLS models. The remaining spectra were used as a validation set. Root mean error of prediction (RMSEP) [124] determined using 10-fold cross-validation (CV) was used to determine the optimum number of latent variables for the PLS models and the optimal parameter combination for airPLS, ALS and parametric method. Once the optimum PLS model was determined for each algorithm, the prediction performance of these models were assessed by computing the area under the response operating characteristic curve (AUC) [125] using the validation set. The entire process of selecting a training set and validation set, developing, optimizing and validating PLS models was repeated 30 times. For the LC-MS data, a similar procedure was used except that the full data set was used for training and the prediction performance of the optimum PLS model was determined using the cross-validated RMSEP. This is because each class contains only 6 spectra and thus it is not practical to divide the dataset into a training set and validation set. During the optimization of the parameters, we used different λ values from the set {10, 20… 490, 500} for airPLS and a grid search using p values from the set {0.001, 0.011... 0.081, 0.091} and lambda values from the set {10, 20… 490, 500} for ALS. For the parametric method, we used σ_p values from the union set of {10000, 1000, 100, 10, 1} and {10, 20, 30… 2700}.

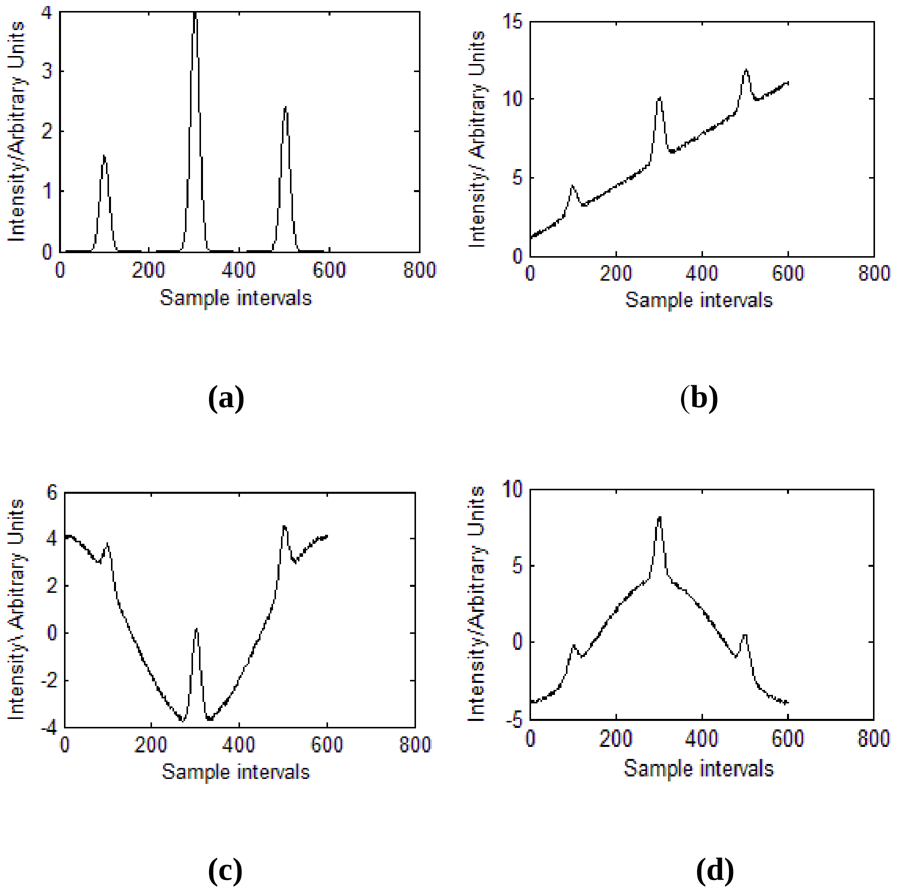
Table 2.1 Experimental data used to evaluate AIMA algorithm.

Spectra type	Description
HPLC	Eight chromatograms of Red Peony Root [96], which has varying baseline drifts from sample to sample. The Red Peony Root was collected from different producing areas in China, and a standard sample was also bought from the National Institute for control of Pharmaceutical and Biological Products. Two UV spectra per second from 200 nm to 600 nm with a bandwidth of 4 nm resulted in 100 data points in each UV spectrum. The "most peaks rich" wavelength 230 nm was then selected.
Raman	Spectra of Prednisone Acetate Tablets (PATs) from 10 different pharmaceutical factories [96]. The spectra were measured using a laser of 785 nm wavelength for excitation by BWTEK i-Raman-785 spectrometer with a 2048 elements thermoelectric cooled linear charge-coupled device (TEC-CCD) arrays and recorded with 5000ms integration times.
SELDI-TOF	One set of chromatograms containing mouse pancreas protein analysis, with 101 spectra from control cells and 80 spectra from cancerous cells [34].
LC-MS	Subset of the data from 200-600 m/z and 2500-4500 seconds of the spinal cords of 6 wild-type and 6 FAAH knockout mice [126]

2.2 Results

2.2.1 Comparing simulated data using our AIMA and other algorithms

Simulated data using three different baselines are shown in Fig. 2.2.



(a) **(b)**

(c) **(d)**

Fig. 2.2 Simulated data. (a) 3 pure Gaussian signals; (b) pure signal with linear baseline and low random noise; (c) pure signal with convex curved baseline and low random noise; (d) pure signal with concave curved baseline and low random noise.

The difference between the expected and corrected peak height were calculated for the four algorithms and expressed as a percentage difference to the actual peaks as well as for the overall spectrum. Table 2.2 shows the percentage error the different algorithms for individual peaks for the various baselines used. For each row, the top performer is highlighted in bold. Overall, AIMA outperforms all the other three algorithms in high noise environment for all peaks except peak 2 where ALS outperformed AIMA by 1.24 %. In the low noise environment, airPLS and AIMA were better than ALS and parametric method. airPLS tends to work better for convex baseline and AIMA tends to work better for concave baseline. The poorer performance of AIMA on convex baseline may be due to the inherent inflexibility of the fully-automated approach of AIMA. It is interesting to note that in the spectra where airPLS outperformed AIMA, airPLS outperformed by a maximum error reduction of 1.49%. We derived the value of 1.49% as the difference at peak 2 of the convex baseline (low noise) between airPLS and AIMA had values of 6.21% and 7.70% as error reduction respectively. This is the largest difference between AIMA and airPLS where airPLS had the smaller error of reduction. However, where AIMA outperformed airPLS, the maximum error reduction was 59.39% which occurred at the peak 1 of the linear baseline with high noise. Therefore it should be noted the AIMA tends to outperform airPLS with a wider margin compared to when airPLS outperforms AIMA.

Table 2.2 Comparison of percentage error of individual peak heights for all the algorithms and various baselines.

For each baseline, we have calculated the percentage error of individual peak heights for both low and high noise for all the algorithms
* P stands for Parametric

		Peak 1				Peak 2				Peak 3			
		airPLS	ALS	P	AIMA	airPLS	ALS	P	AIMA	airPLS	ALS	P	AIMA
Concave baseline	Low noise	22.72	52.27	53	**9.18**	21.64	56.76	6.78	**3.59**	22.28	55.03	48.82	**6.18**
	High noise	148.4	114.55	134.71	**107.95**	58.57	48.42	55.26	**45.24**	99.61	73.41	94.33	**71.59**
Convex baseline	Low noise	**12.87**	52.28	62.12	14.14	**6.21**	56.75	28.56	7.7	**10.18**	55.03	47.48	11.35
	High noise	155.38	114.56	137	**106.61**	72.17	**48.42**	62.81	49.66	105.18	73.41	90.94	**69.91**
Linear baseline	Low noise	**12.44**	52.27	21.5	13.2	10.3	56.77	44.21	**8.69**	**8.45**	55.02	13.94	9.28
	High noise	168.82	114.55	143	**109.44**	59.47	48.42	52.96	**42.62**	114.03	73.4	97.96	**72.56**

2.2.2 HPLC Chromatogram

AIMA was ranked second in terms of the reduction of the convex hull area and was 34.83% behind the top performer, ALS as shown in Table 2.3. The much better performance of ALS compared to the other 3 methods suggests that ALS is more suitable for baseline correction of HPLC chromatograms. However, more studies are necessary as the number of HPLC chromatograms used in this study is small.

Table 2.3 Comparison of percentage reduction in area of convex hull of airPLS, ALS, parametric methods and AIMA

Method	Optimum parameters	Percentage reduction of area of convex hull
ALS	$\lambda = 10, p = 1 \, * 10^3$	83.83%
AIMA	N.A.	49.00%
Parametric method	$\sigma_p = 1$	37.77%
airPLS	$\lambda = 30$	35.58%

2.2.3 Raman

Table 2.4 showed that the performance of the 4 methods is comparable for baseline correction of Raman spectra. AIMA was ranked second in terms of the reduction of the convex hull area and was only 1.72% behind the top performing method, ALS.

Table 2.4 Comparison of percentage reduction in area of convex hull of airPLS, ALS, parametric methods and AIMA

Method	Optimum parameters	Percentage reduction of area of convex hull
ALS	$\lambda = 10, p = 0.031$	99.75%
AIMA	N.A.	98.02%
airPLS	$\lambda = 50$	96.00%
Parametric method	$\sigma_p = 1$	95.59%

2.2.4 SELDI-TOF

The box plot for the AUCs determined using the validation sets for the 30 optimum PLS models of each algorithm is shown in Fig. 2.3. The parametric method and AIMA had the highest and lowest median AUC respectively. The 3 semi-automated algorithms were able to outperform AIMA through a very careful parameter optimization with a median AUC improvement ranging from 1.13 to 1.17 folds. However, it is to be noted that the time taken to optimize these 3 semi-automated algorithms ranged from 28.6 to 197.7 times that required for AIMA.

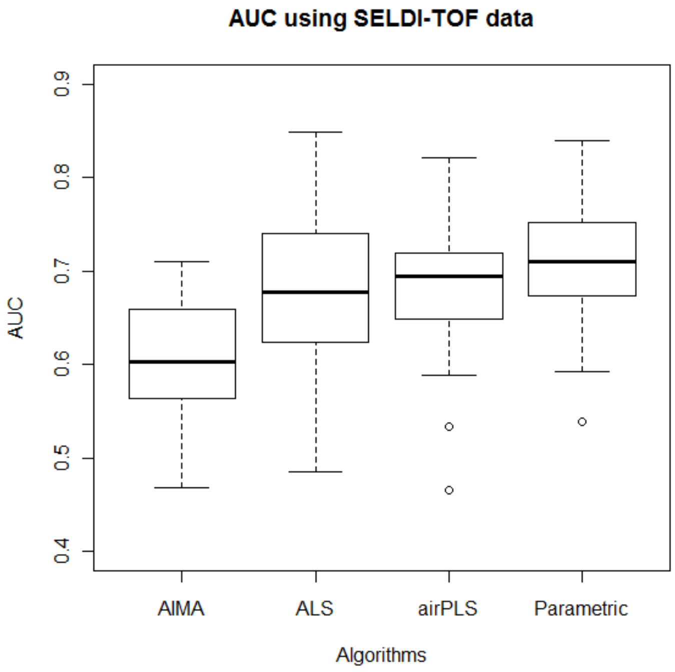


Fig. 2.3 Box plot of AUC for airPLS, ALS, parametric method and AIMA using the SELDI-TOF data.

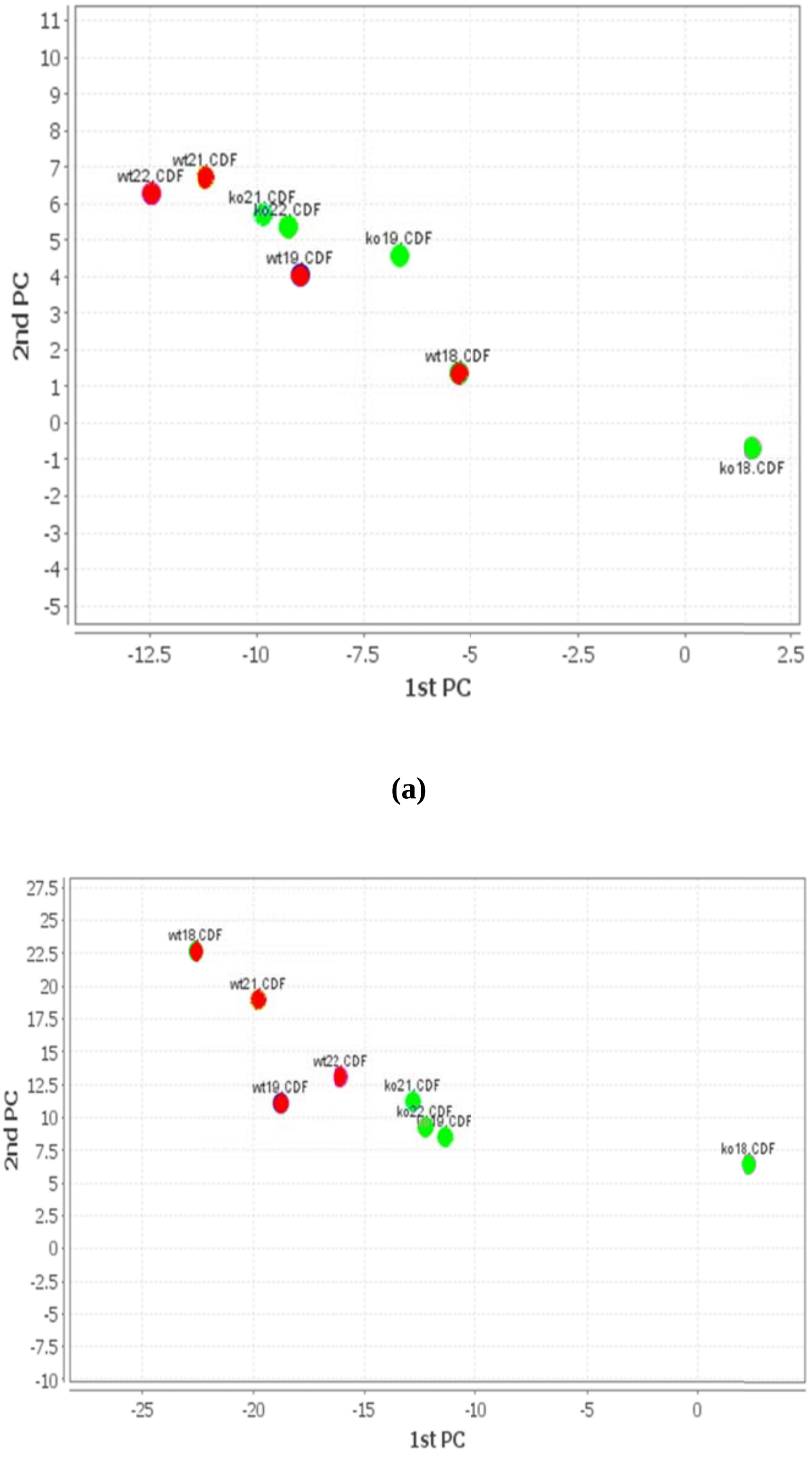
2.2.5 LC-MS

The RMSEP of the optimum PLS models for airPLS, ALS, parametric method and AIMA using LC-MS data are given in Table 2.5. A similar trend is seen as in the SELDI TOF data where the best performer was the parametric method followed by airPLS, ALS and finally AIMA. All the 3 semi-automated algorithms outperform AIMA by 1.03 folds for ALS to 1.23 folds for the parametric method. It is important to note again that the time needed to carefully optimized the parameters of the 3 semi-automated algorithms can be prohibitive compared to AIMA which do not require optimization of parameters.

Table 2.5 Comparison of RMSEP of optimum PLS models for airPLS, ALS, parametric methods and AIMA

Method	Optimum parameters	Optimum PLS latent variables	RMSEP
Parametric method	$\sigma_p = 490$	3	0.445
AirPLS	$\lambda = 120$	5	0.461
ALS	$\lambda = 370, p = 0.001$	3	0.531
AIMA	N.A.	3	0.549

PCA plots of uncorrected spectra and spectra corrected using AIMA are shown in Fig. 2.4. The results were processed in MZmine using our AIMA plugin and post processing was done in an image editor to give additional colouring for clearer differentiation of the two classes of scores. The results show that the corrected spectra has a clearer separation of wild type and knock out compared to uncorrected spectra.



(a)

(b)

Fig. 2.4 Zoom in of PCA Plots with 1st two principal components. Wild type are in red and Knock out in green (a) uncorrected data. (b) using AIMA

2.2.6 Speed

Table 2.6 shows the time taken in minutes for processing the full SELDI-TOF data to create baseline corrected spectra for all parameter combinations and the optimization of PLS models. Baseline correction was performed on Duo Core 2.53GHz Windows Vista Business laptop with 4GB RAM using Matlab. PLS analysis was performed on a Xeon E5530 2.40GHz Windows Server 2008 R2 with 40GB RAM using R. Both the baseline correction followed by PLS was averaged from two runs with very similar timings. AIMA's clear advantage is seen when parameter optimization of the other algorithms is needed as it does not require any optimization. The total time needed for the 3 semi-automated algorithms range from 28.6 to 197.7 times that required by AIMA.

Table 2.6 Time taken to process SELDI-TOF data with baseline correction for all parameter combination and PLS optimization of parameter and number of latent variable for airPLS, ALS, parametric methods and AIMA

Algorithm	Time for baseline correction (mins)	Time for PLS optimization (mins)	Total time (mins)
AIMA	5.86	8.33	14.19
airPLS	6.82	398.78	405.6
Parametric	184.49	2621.07	2805.56
ALS	98.92	2957.81	3056.73

2.3 Conclusion

The results show that AIMA is generally comparable to semi-automated algorithms like airPLS, ALS and the parametric algorithm. Using simulated data where the actual peaks are known, it revealed that AIMA was the overall best performer. However, for experimental data, we do not know the real data and hence the different performance metrics that were used to hypothesis the improvements like area of the convex hull, RMSEP, AUC may not truly reflect the real performance of the baseline correction. Based solely on ranking for the individual experimental datasets, ALS would be the best-performer. Hence in our study, our comparability was based on AIMA being the best-performer in the simulated data, ranking 2nd for 2 out of 4 experimental datasets and finally even though it was ranked 4th in the LCMS dataset, the PCA plots showed a clear separation between the two sample types after AIMA correction indicating it effectiveness.

The AIMA algorithm is a fully-automated baseline correction technique whereas other algorithms required optimization of its parameters which would considerably increase the time taken. We acknowledge that further tuning of the parameters for the individual spectrum in each type of spectra was possible. However, individual spectrum parameter optimization would further exponentially increase the computational time for the semi-automated techniques and thus was not done in this study. When processing large data sets, a fully-automated algorithm such as AIMA would be desirable as it is not necessary to optimize any parameters. Thus, the AIMA algorithm is a potentially useful baseline correction method for a variety of spectra types.

Chapter 3: An Automated Pearson’s Correlation Change Classification (APC3) approach for GC/MS metabonomic data using Total Ion Chromatograms (TIC)

3.1 Introduction

A study has shown variable ranking via the correlation based feature selection [127] which uses the magnitude of the Pearson's correlation coefficient between the class values and variable values for each feature to be promising. In this study, we extended from correlation based feature selection [127] and created a new automated Pearson's correlation change classification (APC3) technique which have high computational efficiency. The aim of this study is to evaluate the performance of APC3 by comparing it with other classification algorithms, classification algorithms in combination with transformation techniques and classification algorithms in combination with variable selection approaches using TICs of binominal GC/MS data.

3.2 Methodology

3.2.1 Automated Pearson's correlation change classification (APC3) algorithm

Suppose there are *N* training samples, let matrix ***v*** and vector ***w*** contain the intensity values for the retention times (variables) and class values of the training samples respectively. Each class is assigned either a value of 1 or -1 since we are dealing with binary classification. The magnitude of the Pearson's correlation coefficient for each variable with the class value will be calculated to derive a vector, ***a***.

To classify a test sample, the test sample will be added to the training samples and the Pearson's correlation coefficient of each variable will be recalculated. Since the class value of the test sample is unknown, there will be two possible new Pearson's correlation coefficients vectors ***b*** and ***c***; each vector corresponding to the case when the test sample is assumed to have a class value of 1 or a class value of -1 respectively. The values in ***a***, ***b*** and ***c*** are then sorted based on descending order of the values in ***a***. The first position *i* where a_i lies between the b_i and c_i is determined and the test sample will be classified as belonging to class 1 if $c_i < a_i < b_i$ or belong to class -1 if $b_i < a_i < c_i$.

We will use the following simple example to illustrate how the algorithm works. Suppose we have 4 training samples per class with 4 retention times for each sample. N would have the value of 8. Assume that vector $\boldsymbol{a}$ has the following values {0.812, 0.988, 0.608, 0.709}, which means that the second retention time is the most highly correlated with the class values. The value of 0.988 is the Pearson's correlation coefficient for the intensity values for the second retention time ({123, 200, 182, 132, 458, 456, 460, 480}) with the class values of the training samples ({1, 1, 1, 1, -1, -1, -1, -1}). Next assuming we have a test sample with intensity values of {110, 130, 393, 293}. For the second retention time, we will add this test sample's intensity value of 130 to that of the training samples to get {123, 200, 182, 132, 458, 456, 460, 480, 130}. Then b_i for the second retention time will be calculated using the new vector and the new class vector {1, 1, 1, 1, -1, -1, -1, -1, 1} by assuming the test sample belongs to class 1. Similarly c_i is calculated by assuming the test sample belongs to class -1, with the new class vector as {1, 1, 1, 1, -1, -1, -1, -1, -1}. In this example, the values of b_i and c_i would be 0.989 and 0.714 respectively. Since $c_i < a_i < b_i$, the test sample will be classified as belonging to class 1.

3.2.2 Evaluation of APC3 algorithm

We compared our APC3 algorithm with different dimensionality reduction and classification combinations on the TIC of two GC/MS datasets. One urinary metabonomics data of 24 bladder cancer (BC) patients and 35 healthy (H) subjects which had been introduced in our previous paper [128]. The second set was red wine samples harvested from four different geographical locations [129]. The wine data TICs were already available online and hence no pre-processing was required to extract the TIC. For our urinary GC/MS dataset, we used a custom script integrated into MZmine [130] using MZmine's library functions. The script would first get the number of scans per raw data file and derive a value, *m*, which is the maximum of these values. For every raw data file, an array with value 0 at every index and size *m* would be initialized. Next, the script would iterate again through every raw data file and within a raw data file, it would further iterate through the consecutive scan numbers to get a TIC value for each scan number. The scan number for each raw data file would correspond to the array index for that raw data file. Hence the TIC value for each scan number in a raw data file would be updated to the value at that corresponding array index. The wine data was divided into three sets of two-class data using combinations of those from Chile, Australia and South Africa with sample size of 15, 12 and 11 respectively. The Argentinian samples had a sample size of 6 which is too small for dividing our set into training and testing sets and hence were not used. Sample information utilized for training and test sets and the distribution of classes for wine sets are shown in Table 3.1. Each split was to ensure that a minimum of half of each class was in the training set and the number of samples per class was equal in the training set. Models were developed using the training set and their performance were evaluated using the testing set. To reduce bias due to the splitting of the dataset, we repeated the experiments 100 times with different training and test sets for the three wine data sets. The same 100 training and test sets were used for the APC3 algorithm and the dimensionality reduction and classification combinations.

Table 3.1 No. of samples per class for the training and testing sets and the location of the individual classes in the wine sets.

	No of samples				**Location of Wine Sets**	
	Training set		**Testing set**			
	Class 1	**Class 2**	**Class 1**	**Class 2**	**Class 1**	**Class 2**
Wine Set A	8	8	4	7	Australia	Chile
Wine Set B	6	6	6	5	Australia	South Africa
Wine Set C	8	8	7	3	Chile	South Africa

Using the performance on the wine data sets as a preliminary evaluation, we selected the top 2 dimensionality reduction and classification combinations to compare with APC3 on the urine data using various training and testing sizes as shown in Table 3.2. For each training and testing size, just as in the wine data evaluation, we repeated the experiments 100 times with different training and testing sets. For an extended evaluation on computational efficiency, we used the urine sample split A and bootstrapped both the training and testing sets so the BC and H for both training and testing sets each have the same number of samples which was varied from the set {50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 4000}.

Additionally, we also tested for performance in the presence of outliers in the urine data. We randomly assigned one of the BC TIC in the training set to have the class value of H to make that TIC appear as a H outlier and repeated the experiments of the urine data. Next, we randomly assigned one of the H TIC in the training set to have the class value of BC to make that TIC appear as a BC outlier and similarly repeated the urine data experiments.

Table 3.2 No. of samples per class for the training and testing sets in the urine data set.

	No. of samples			
	Training set		**Testing set**	
	BC	**H**	**BC**	**H**
Urine Sample Split A	18	18	6	17
Urine Sample Split B	16	16	8	19
Urine Sample Split C	14	14	10	21
Urine Sample Split D	12	12	12	23
Urine Sample Split E	10	10	14	25
Urine Sample Split F	8	8	16	27

For transformation techniques, we chose to study the original untransformed matrix of TICs which would be referred to as Non Component Analysis

(NCA), principal component analysis (PCA) [131] using the covariance matrix (PCA_1), principal component analysis (PCA) [131] using the correlation matrix (PCA_2), correspondence analysis [132] (CA_1), correspondence analysis with scaling (CA_2) and detrended correspondence analysis (DCA) [133], which are all available from the vegan [134] package. CA is a reciprocal averaging ordination technique whereas DCA is an extension of CA with the ability to correct the distortion caused by unimodal distributions along gradients when using CA. PCA usually performs better than DCA and CA when there is a monotonic distributions along gradients. However, DCA and CA are more suited for unimodal distributions compared to PCA. In total, we experimented with 6 transformation approaches including NCA.

For variable selection algorithms, we used Pearson's correlation from the FSelector [135] package, column wise area under receiver operator curve (colAUC) [136], linear discriminant analysis (LDA) [137] and diagonal discriminant analysis (DDA) [137] using the sda [138] package and variants of Relief [139], Minimum Description Length (MDL) [140], Gini-index (Gini) and a measure named Dietterich, Kearns, and Mansour (DKM) [141] from the CORElearn [142] package. For the variants of Relief [139], we used the original Relief [139], ReliefFexpRank, ReliefFequalK, ReliefFbestK, MyopicReliefF. ReliefFexpRank is a version of Relief where K nearest instances have weight exponentially decreasing with increasing rank and the rank of nearest instance is determined by the increasing Manhattan distance from the selected instance. ReliefFequalK is a version of Relief using equally weighted K nearest instances. ReliefFbestK is a version of Relief where all possible K, representing K nearest instances, are tested and for each feature the highest score is returned and nearest instances have equal weights. MyopicReliefF is a myopic version of Relief resulting from assumption of no local dependencies and attribute dependencies upon class. Hence in total, we studied 12 different variable selection algorithms.

The three classification methods to combine with the dimensionality reduction methods were naïve Bayes (NB) and localised linear discriminant analysis (locLDA) from the klaR [143] package and PLSR from the pls [144] package. NB serves to compute the conditional a-posterior probabilities of a categorical class variable given independent predictor variables using the Bayes rule. locLDA is a localized version of LDA using the localization concept for classification [145] thereby introducing the flexibility of classifying test samples individually which would cater for non-linearity otherwise restricted by conventional LDA. Although variants of PLSR exist, we only choose the improved kernel PLSR [146] since it was shown to be both stable and fast [147]. For the dimensionality reduction and classification combinations, the number of latent variables of PLSR can be optimized via cross validation methods such as a 10-fold cross-validation (CV) or a double cross validation [148]. We verified which cross validation approach is significantly better for 10 different training and testing sets across the PLSR combinations for the optimization of latent variables. Then, this cross validation approach was employed for latent variable optimization of the PLSR combinations for the full 100 different training and testing sets. In the double cross validation, we used k=7 for the inner loop and k=8 for the outer loop which is the most commonly used partition for double cross validation when applied to metabonomic data sets [148]. For the variable selection and classification combinations, the number of important variables was optimized for each algorithm combination using the accuracy of the training set as a criterion for NaiveBayes, locLDA and PLSR classification approaches. A maximum of 50 variables were imposed on the variable selection algorithms.

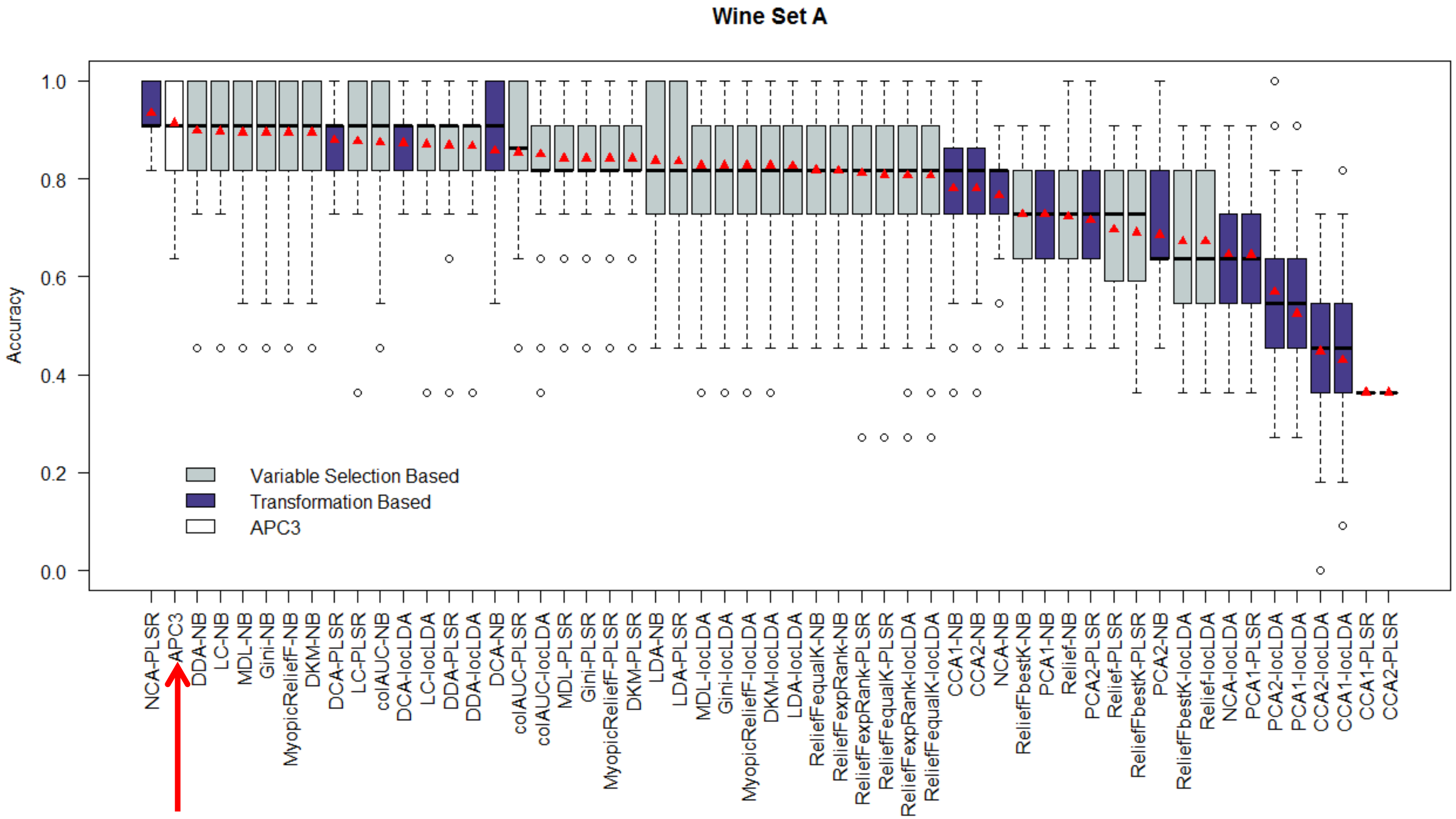
APC3 was initially implemented in R [149] but ported to JAVA for faster processing in the evaluation. The top 2 dimensionality reduction and classification combinations were also implemented in JAVA by integrating with the WEKA [150] library to have a consistent platform when assessing their computational time against that of APC3. All other evaluations were performed in R [149]. In all, we compared APC3 with 54 dimensional reduction and classification combinations which consist of 18 transformation and classification combinations and 36 variable selection and classification combinations.

3.3 Results

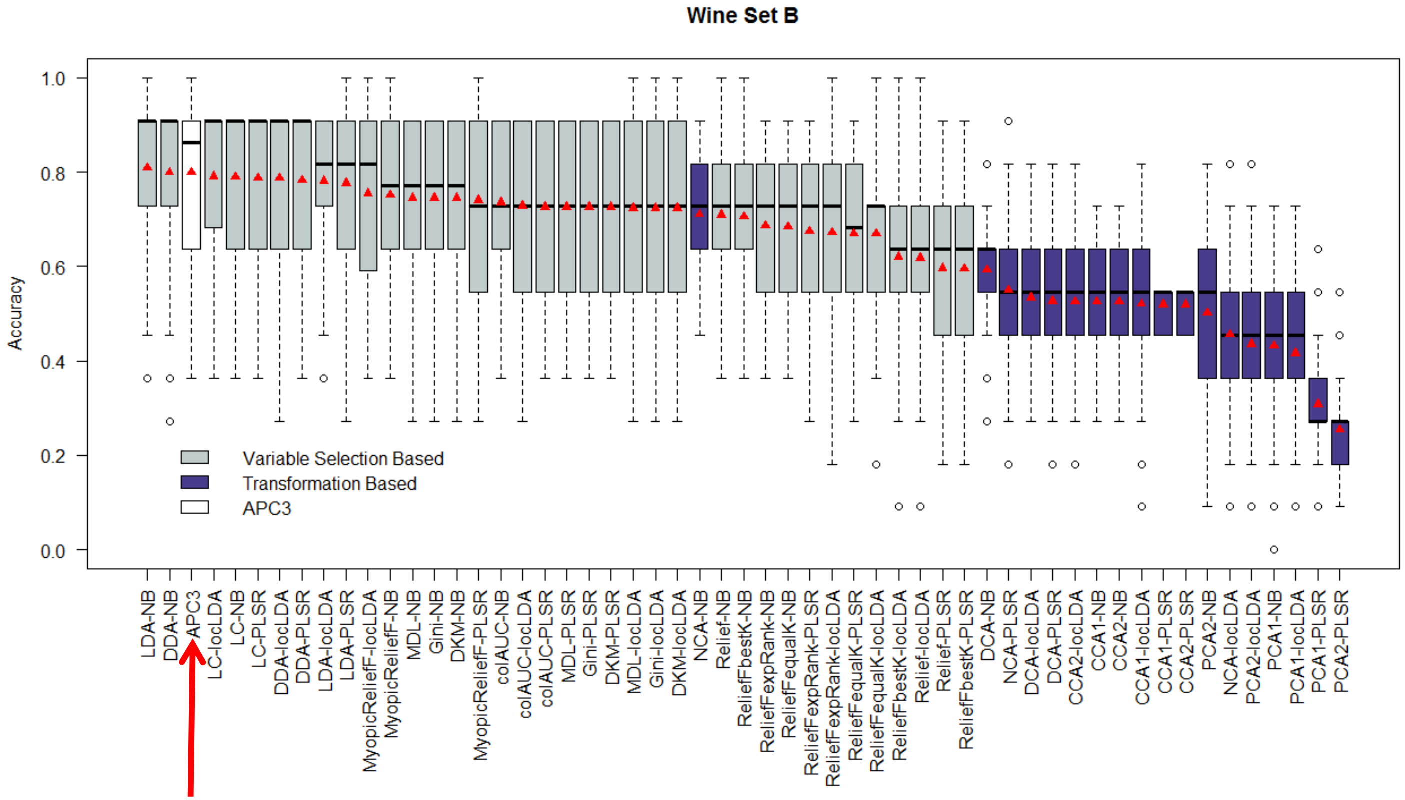
For the PLSR latent variable optimization, 10-fold cross validation had significantly better performance than double cross validation for most of the PLSR combinations as shown in the Appendix - Tables I and II [116]. Hence, we choose to use the 10-fold cross validation for PLSR latent variable optimization for the rest of the experiments.

In general, the variable selection and classification combinations outperformed the transformation and classification combinations. Fig. 3.1 shows the boxplot of the accuracies and AUC of the wine data sets using the dimensionality reduction and classification combinations and APC3 in decreasing order of the average of the accuracies and average of the AUC for the 100 experiments respectively. More details on the overall average accuracy and overall average AUC for each dimensionality reduction and classification combinations using the wine testing sets are provided in the Appendix - Table III [116].

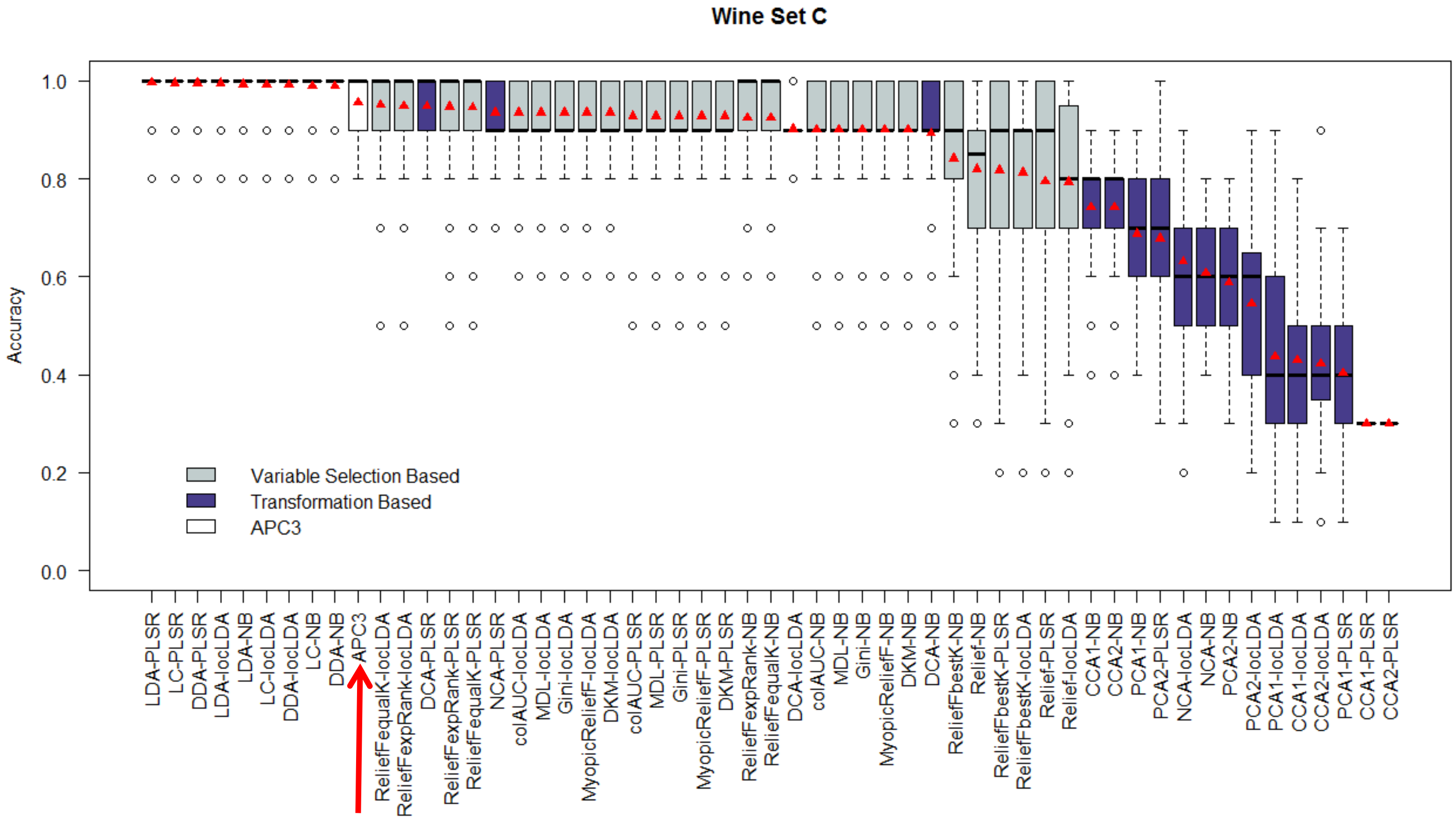
For each wine dataset the variable length for the original dataset was 2700. The variable length after transformation is shown in the Appendix - Table IV. For transformation, the variable length does not change further for both NB and locLDA classifications while building the training model. However, for PLSR, the training model is optimized for the number of latent variables and the average number of latent variables for the 100 training and testing sets each wine data set is given in the Appendix - Table V. In the variable selection combinations, the average variable length of the training models for the 100 training and testing sets for the wine datasets using PLSR, NB and locLDA are given respectively in the Appendix - Tables VI, VII and VII [116].


Wine Set A
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
Variable Selection Based
Transformation Based
APC3
NCA-PLSR
APC3
DDA-NB
LC-NB
MDL-NB
Gini-NB
MyopicReliefF-NB
DKM-NB
DCA-PLSR
LC-PLSR
colAUC-NB
DCA-locLDA
LC-locLDA
DDA-PLSR
DDA-locLDA
DCA-NB
colAUC-PLSR
colAUC-locLDA
MDL-PLSR
Gini-PLSR
MyopicReliefF-PLSR
DKM-PLSR
LDA-NB
LDA-PLSR
MDL-locLDA
Gini-locLDA
MyopicReliefF-locLDA
DKM-locLDA
LDA-locLDA
ReliefFequalK-NB
ReliefFexpRank-NB
ReliefFexpRank-PLSR
ReliefFequalK-PLSR
ReliefFexpRank-locLDA
ReliefFequalK-locLDA
CCA1-NB
CCA2-NB
NCA-NB
ReliefFbestK-NB
PCA1-NB
Relief-NB
PCA2-PLSR
Relief-PLSR
ReliefFbestK-PLSR
PCA2-NB
ReliefFbestK-locLDA
Relief-locLDA
NCA-locLDA
PCA1-PLSR
PCA2-locLDA
PCA1-locLDA
CCA2-locLDA
CCA1-locLDA
CCA1-PLSR
CCA2-PLSR

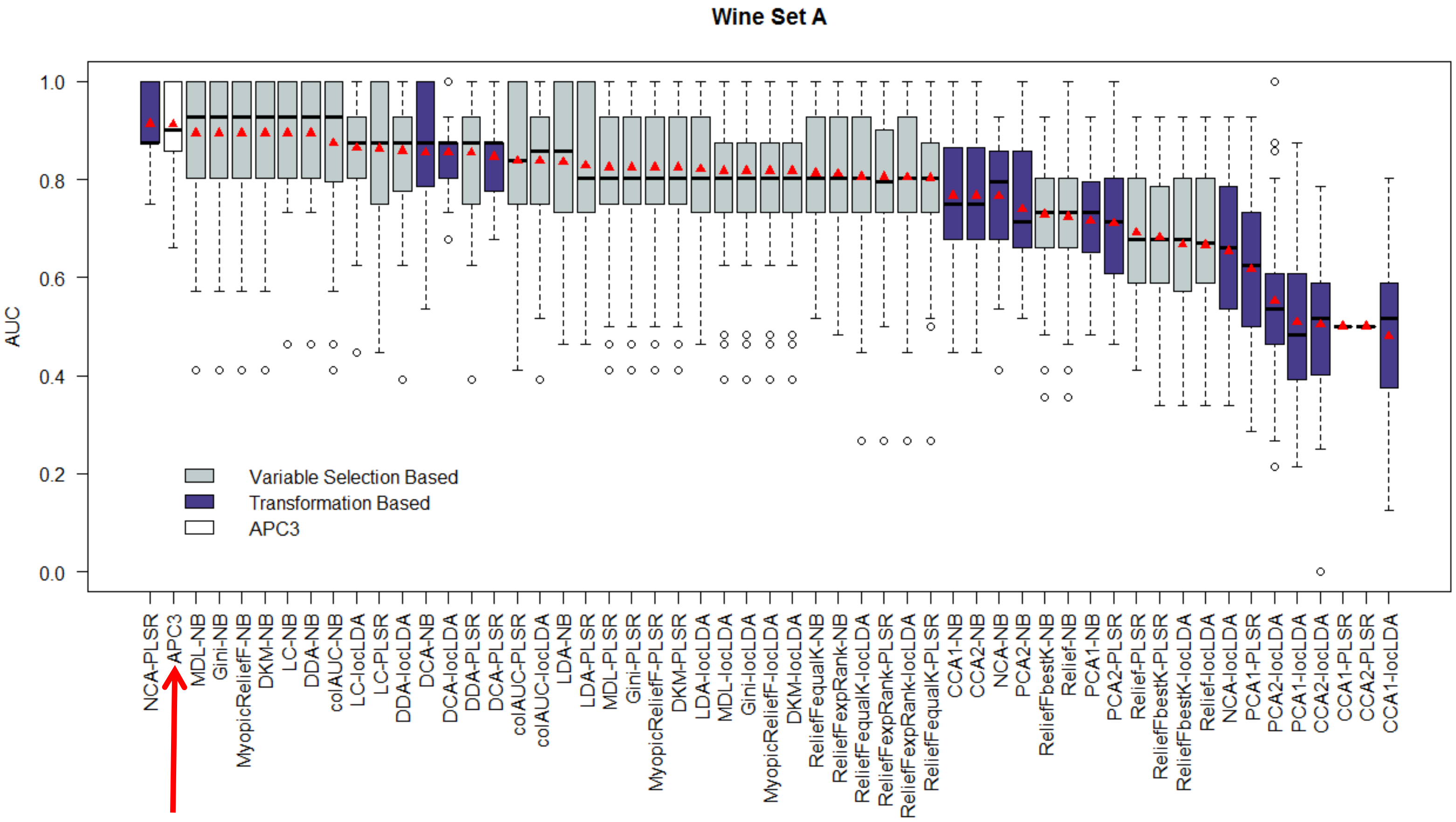
(a)


Wine Set B
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
Variable Selection Based
Transformation Based
APC3
LDA-NB
DDA-NB
APC3
LC-locLDA
LC-NB
LC-PLSR
DDA-locLDA
DDA-PLSR
LDA-locLDA
LDA-PLSR
MyopicReliefF-locLDA
MyopicReliefF-NB
MDL-NB
Gini-NB
DKM-NB
MyopicReliefF-PLSR
colAUC-NB
colAUC-locLDA
colAUC-PLSR
MDL-PLSR
Gini-PLSR
DKM-PLSR
MDL-locLDA
Gini-locLDA
DKM-locLDA
NCA-NB
Relief-NB
ReliefFbestK-NB
ReliefFexpRank-NB
ReliefFequalK-NB
ReliefFexpRank-PLSR
ReliefFexpRank-locLDA
ReliefFequalK-PLSR
ReliefFequalK-locLDA
ReliefFbestK-locLDA
Relief-locLDA
Relief-PLSR
ReliefFbestK-PLSR
DCA-NB
NCA-PLSR
DCA-locLDA
DCA-PLSR
CCA2-locLDA
CCA1-NB
CCA2-NB
CCA1-locLDA
CCA1-PLSR
CCA2-PLSR
PCA2-NB
NCA-locLDA
PCA2-locLDA
PCA1-NB
PCA1-locLDA
PCA1-PLSR
PCA2-PLSR

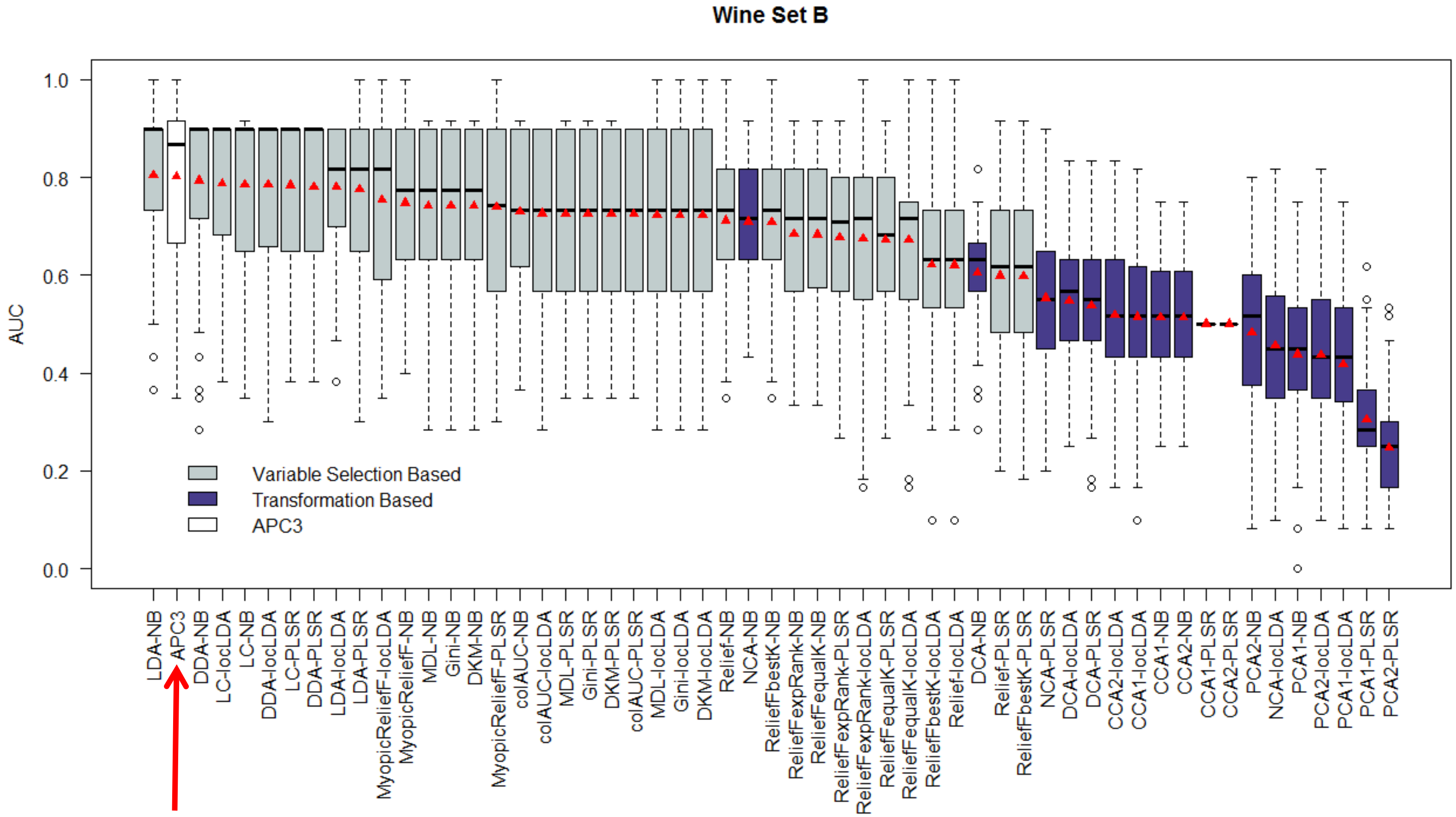
(b)


Wine Set C
Accuracy
0.0
0.2
0.4
0.6
0.8
1.0
Variable Selection Based
Transformation Based
APC3
LDA-PLSR
LC-PLSR
DDA-PLSR
LDA-locLDA
LDA-NB
LC-locLDA
DDA-locLDA
LC-NB
DDA-NB
APC3
ReliefFequalK-locLDA
ReliefFexpRank-locLDA
DCA-PLSR
ReliefFexpRank-PLSR
ReliefFequalK-PLSR
NCA-PLSR
colAUC-locLDA
MDL-locLDA
Gini-locLDA
MyopicReliefF-locLDA
DKM-locLDA
colAUC-PLSR
MDL-PLSR
Gini-PLSR
MyopicReliefF-PLSR
DKM-PLSR
ReliefFexpRank-NB
ReliefFequalK-NB
DCA-locLDA
colAUC-NB
MDL-NB
Gini-NB
MyopicReliefF-NB
DKM-NB
DCA-NB
ReliefFbestK-NB
Relief-NB
ReliefFbestK-PLSR
ReliefFbestK-locLDA
Relief-PLSR
Relief-locLDA
CCA1-NB
CCA2-NB
PCA1-NB
PCA2-PLSR
NCA-locLDA
NCA-NB
PCA2-NB
PCA2-locLDA
PCA1-locLDA
CCA1-locLDA
CCA2-locLDA
PCA1-PLSR
CCA1-PLSR
CCA2-PLSR

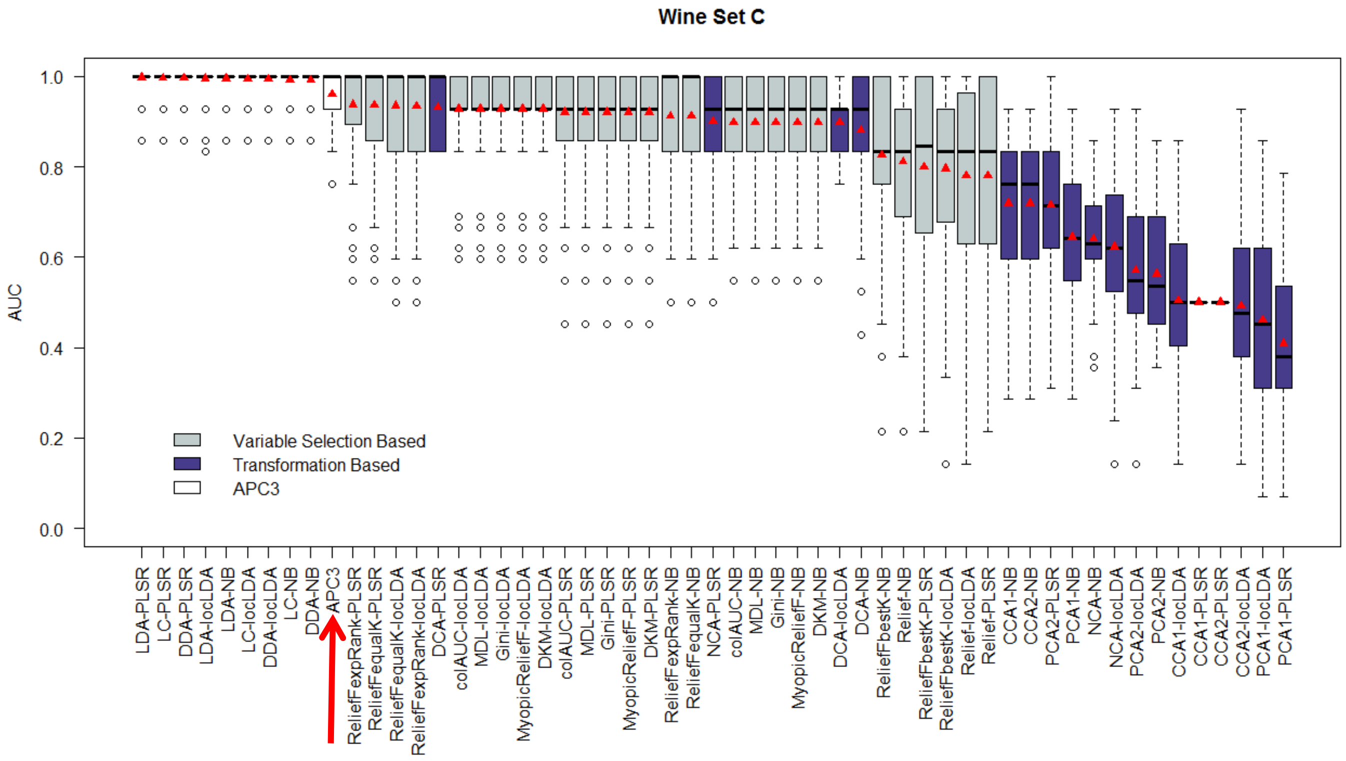
(c)


Wine Set A
AUC
1.0
0.8
0.6
0.4
0.2
0.0
Variable Selection Based
Transformation Based
APC3
NCA-PLSR
APC3
MDL-NB
Gini-NB
MyopicReliefF-NB
DKM-NB
LC-NB
DDA-NB
colAUC-NB
LC-locLDA
LC-PLSR
DDA-locLDA
DCA-NB
DCA-locLDA
DDA-PLSR
DCA-PLSR
colAUC-PLSR
colAUC-locLDA
LDA-NB
LDA-PLSR
MDL-PLSR
Gini-PLSR
MyopicReliefF-PLSR
DKM-PLSR
LDA-locLDA
MDL-locLDA
Gini-locLDA
MyopicReliefF-locLDA
DKM-locLDA
ReliefFequalK-NB
ReliefFexpRank-NB
ReliefFequalK-locLDA
ReliefFexpRank-PLSR
ReliefFexpRank-locLDA
ReliefFequalK-PLSR
CCA1-NB
CCA2-NB
NCA-NB
PCA2-NB
ReliefFbestK-NB
Relief-NB
PCA1-NB
PCA2-PLSR
Relief-PLSR
ReliefFbestK-PLSR
ReliefFbestK-locLDA
Relief-locLDA
NCA-locLDA
PCA1-PLSR
PCA2-locLDA
PCA1-locLDA
CCA2-locLDA
CCA1-PLSR
CCA2-PLSR
CCA1-locLDA

(d)


Wine Set B
AUC
1.0
0.8
0.6
0.4
0.2
0.0
Variable Selection Based
Transformation Based
APC3
LDA-NB
APC3
DDA-NB
LC-locLDA
LC-NB
DDA-locLDA
LC-PLSR
DDA-PLSR
LDA-locLDA
LDA-PLSR
MyopicReliefF-locLDA
MyopicReliefF-NB
MDL-NB
Gini-NB
DKM-NB
MyopicReliefF-PLSR
colAUC-NB
colAUC-locLDA
MDL-PLSR
Gini-PLSR
DKM-PLSR
colAUC-PLSR
MDL-locLDA
Gini-locLDA
DKM-locLDA
Relief-NB
NCA-NB
ReliefFbestK-NB
ReliefFexpRank-NB
ReliefFequalK-NB
ReliefFexpRank-PLSR
ReliefFexpRank-locLDA
ReliefFequalK-PLSR
ReliefFequalK-locLDA
ReliefFbestK-locLDA
Relief-locLDA
DCA-NB
Relief-PLSR
ReliefFbestK-PLSR
NCA-PLSR
DCA-locLDA
DCA-PLSR
CCA2-locLDA
CCA1-locLDA
CCA1-NB
CCA2-NB
CCA1-PLSR
CCA2-PLSR
PCA2-NB
NCA-locLDA
PCA1-NB
PCA2-locLDA
PCA1-locLDA
PCA1-PLSR
PCA2-PLSR

(e)



(f)

Fig. 3.1 Boxplot of accuracies using (a) wine set A (b) wine set B (c) wine set C and AUC using (d) wine set A (e) wine set B (f) wine set C. The red triangle represents the mean value of each boxplot.

Comparing APC3 with the dimensionality reduction and classification combinations, the results in Fig. 3.1 showed that the APC3 algorithm, which had both an overall average accuracy and an overall average AUC of 0.89 ± 0.08, has similar performance as the top few dimensionality reduction and classification combinations. For the overall results of the wine data sets, which is given in Fig. 3.2, the top two dimensionality reduction and classification combinations were DDA-NB and LC-NB. Hence the evaluation of the urine samples were only performed using DDA-NB, LC-NB and APC3.

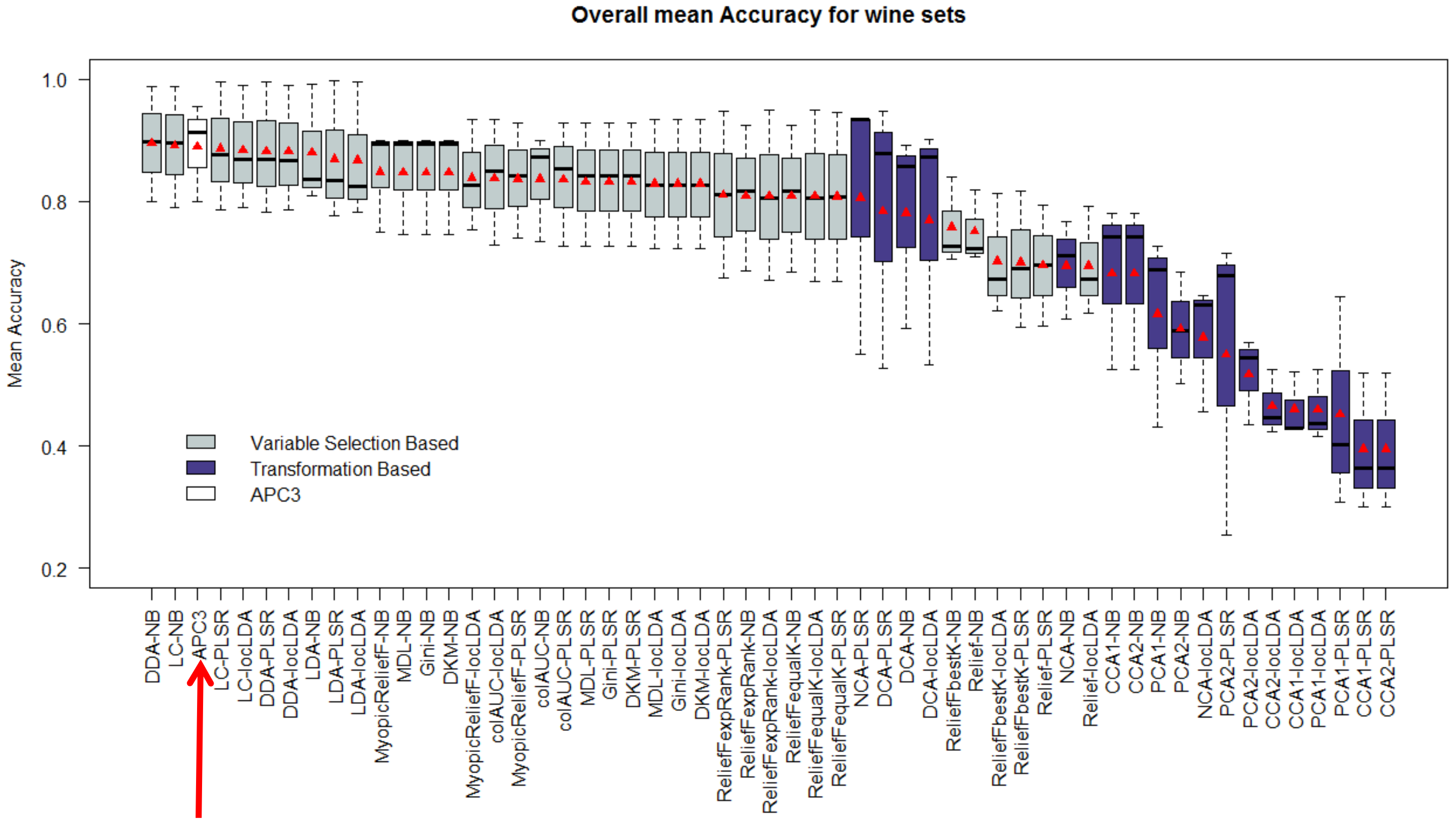


Fig. 3.2.a Accuracy boxplot of the average accuracy for the wine data sets. The red triangle represents the mean value of each boxplot.

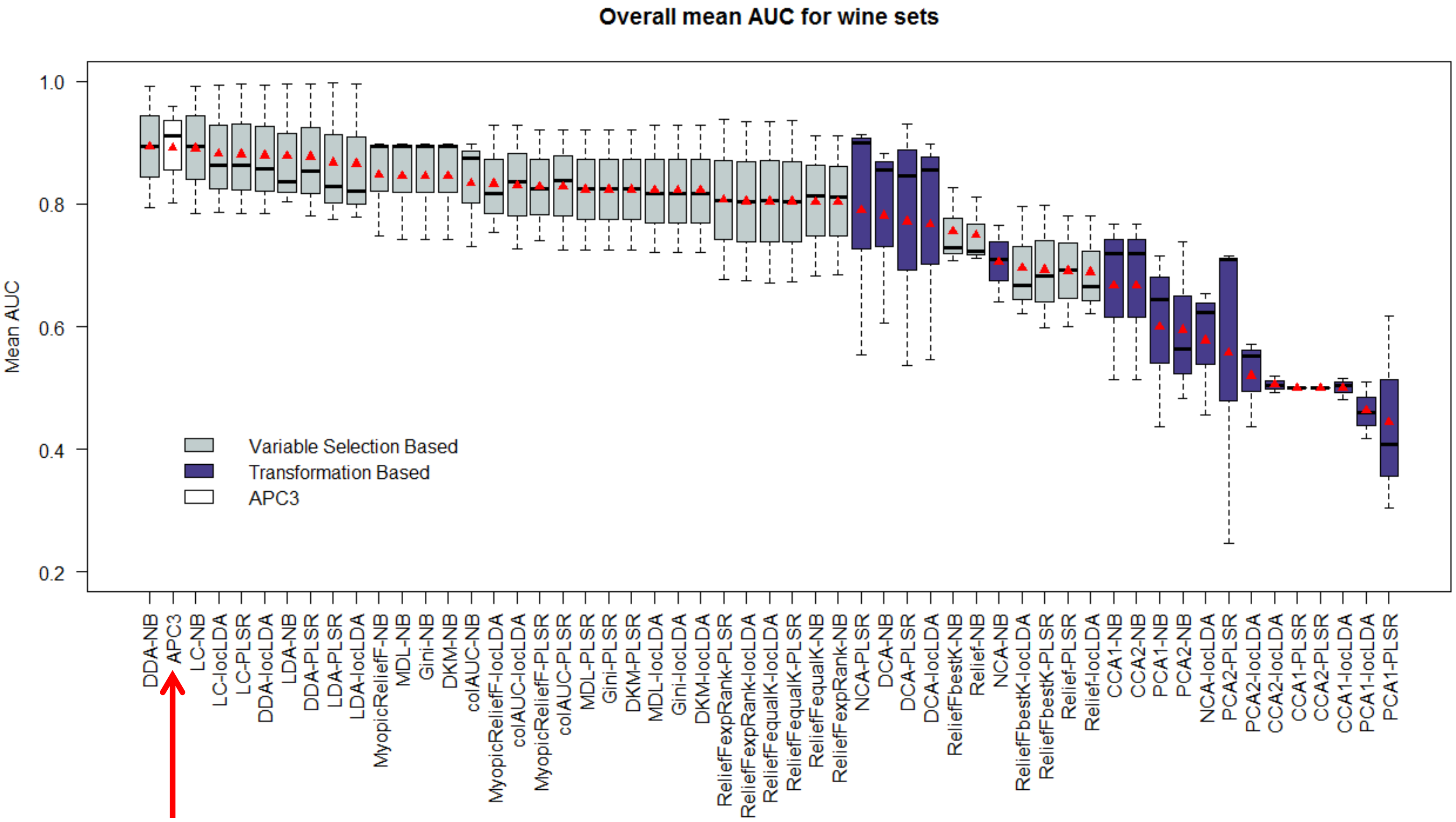


Fig. 3.2.b AUC boxplot of the average AUC for the wine data sets. The red triangle represents the mean value of each boxplot.

The mean accuracies and mean AUC across the 100 experiments for the urine data for APC3, DDA-NB and LC-NB is shown in Fig. 3.3. The results show that APC3 generally had slightly higher or similar performance as DDA-NB and LC-NB. Furthermore, only APC3's performance was almost independent of the presence of outliers in the training set.

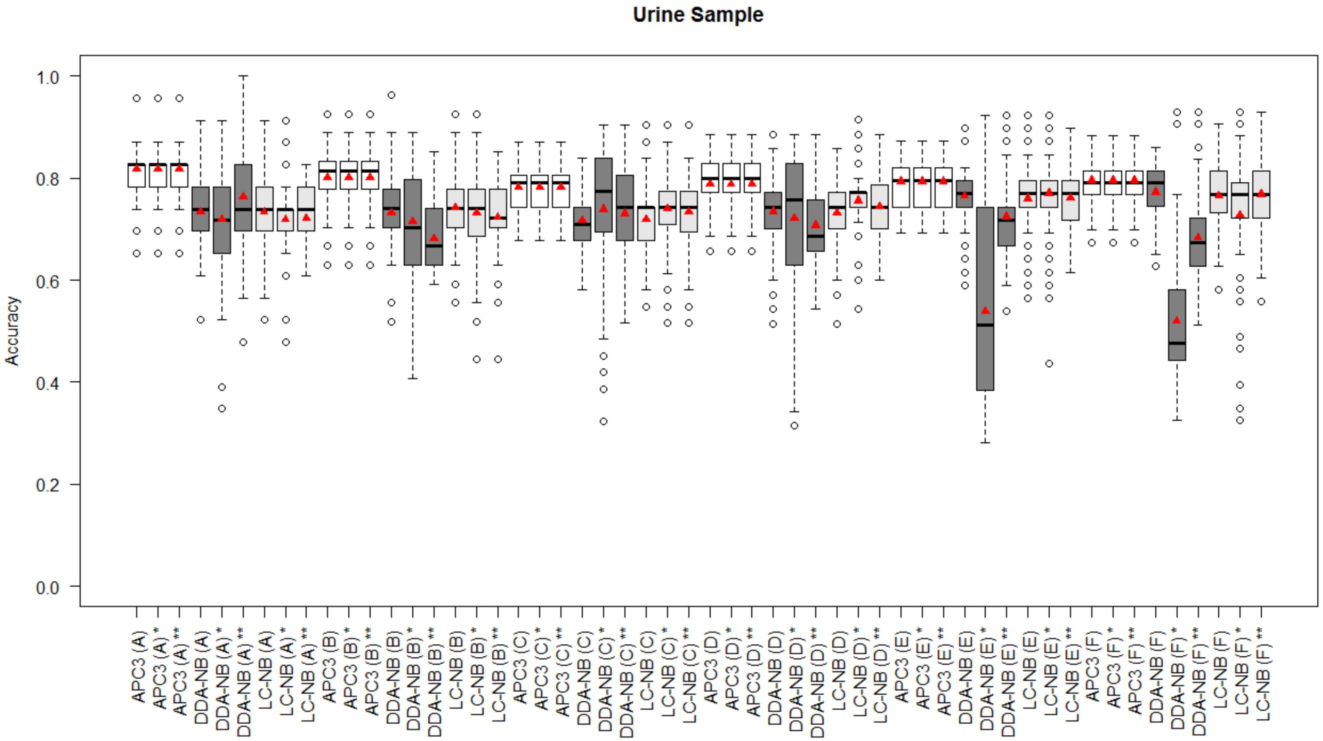


Fig. 3.3.a Accuracy boxplot of the average accuracy for the urine data set splits. The red triangle represents the mean value of each boxplot.

* one of the BC TIC in the training set assigned the class value of H.

** one of the H TIC in the training set assigned the class value of BC.

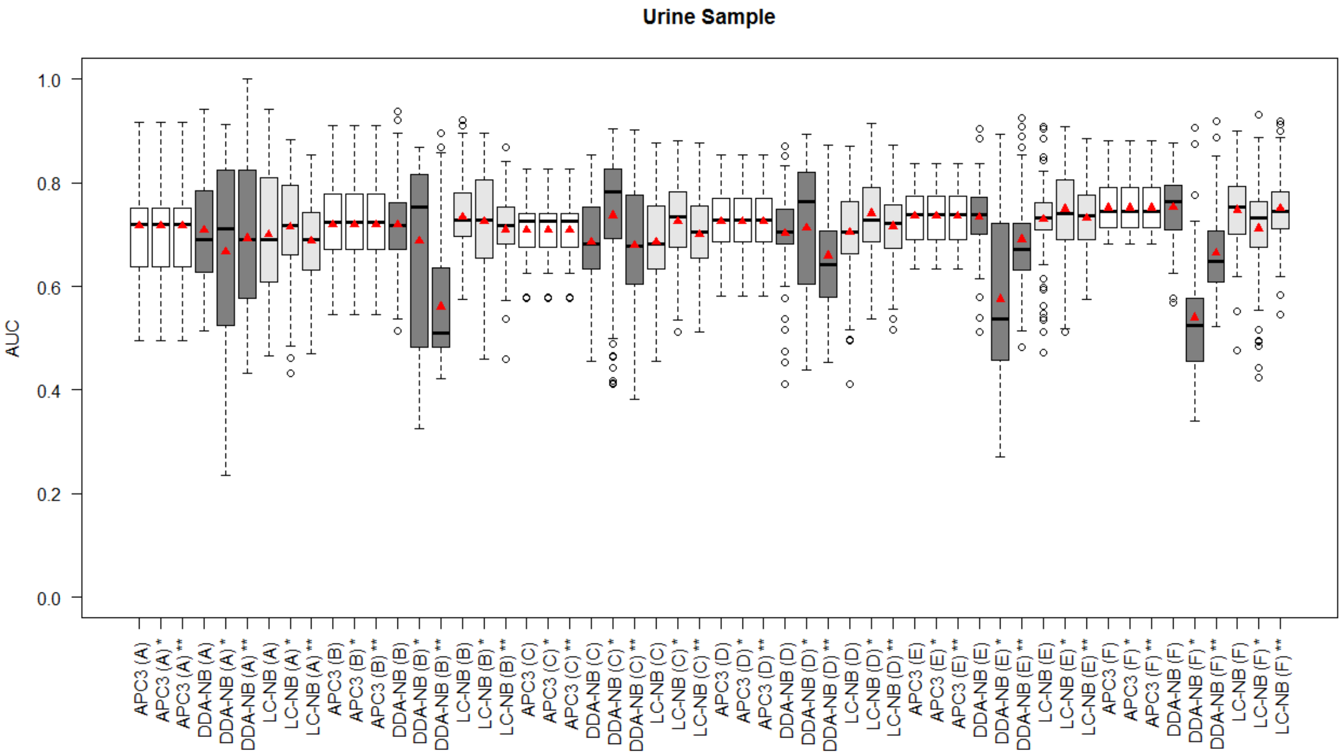


Fig. 3.3.b AUC boxplot of the average accuracy for the urine data set splits. The red triangle represents the mean value of each boxplot.

* one of the BC TIC in the training set assigned the class value of H.

** one of the H TIC in the training set assigned the class value of BC.

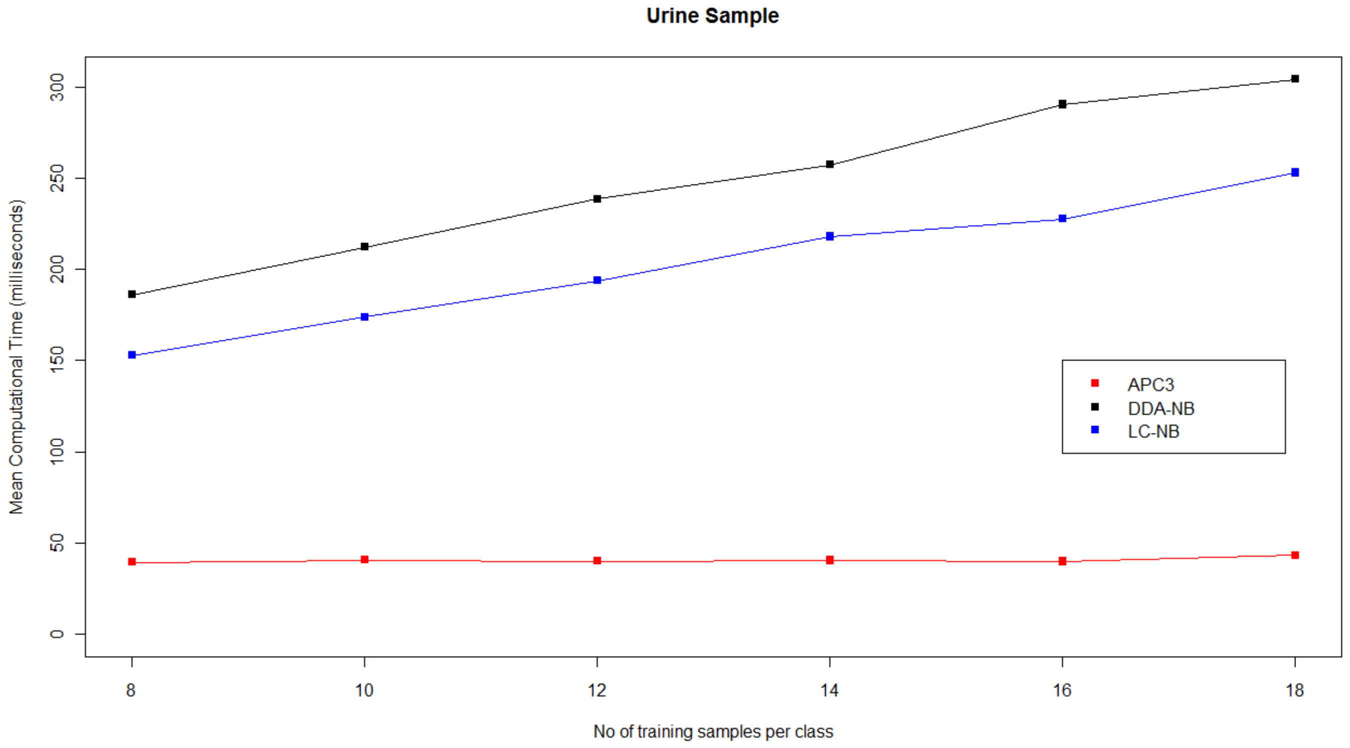


Fig. 3.4.a Average computational time (both the training and test phases) APC3, DDA-NB and LC-NB across the different training sample sizes for the urine data.

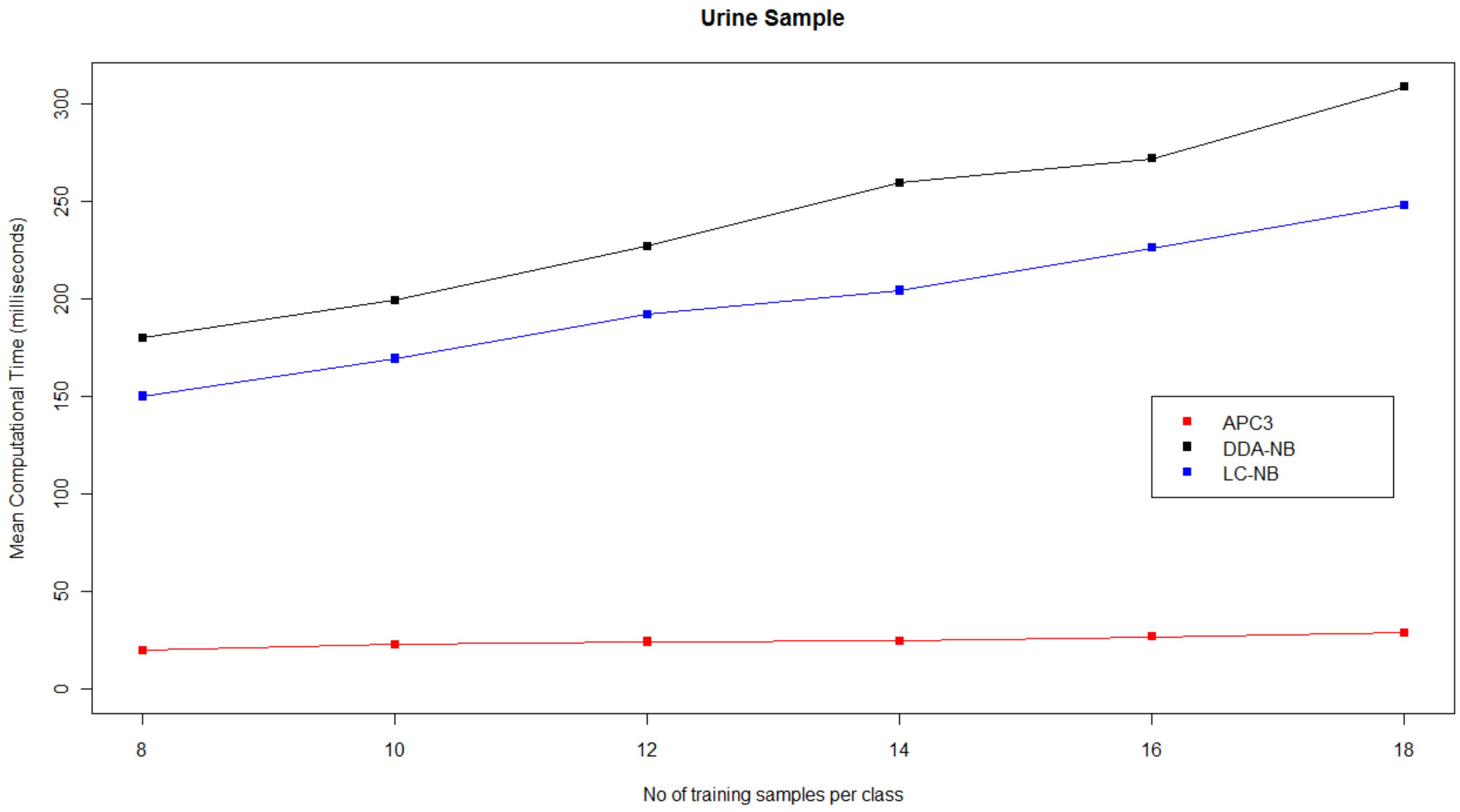


Fig. 3.4.b Average computational time only for the training phase APC3, DDA-NB and LC-NB across the different training sample sizes for the urine data only.

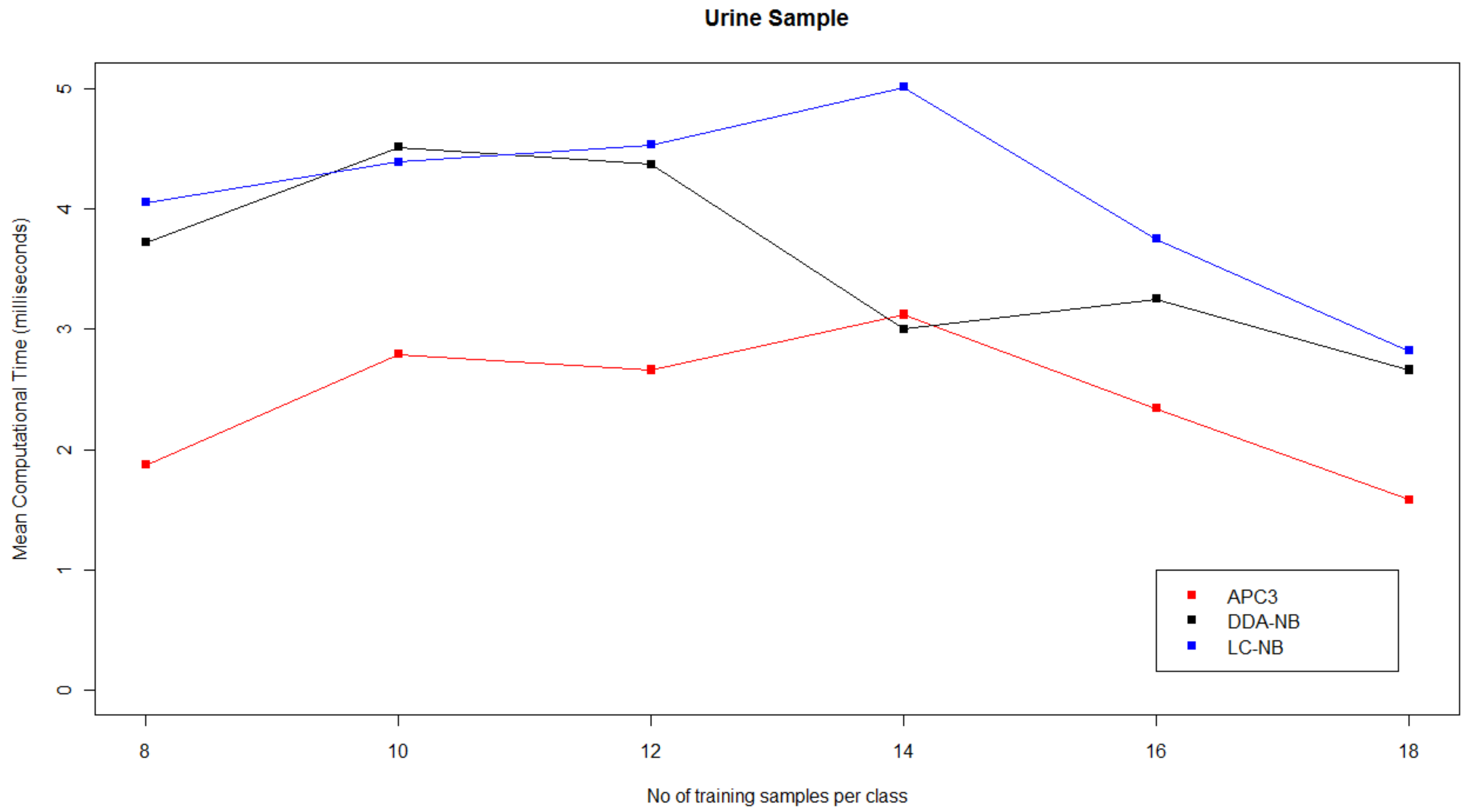


Fig. 3.4.c Average computational time only for the testing phase APC3, DDA-NB and LC-NB across the different training sample sizes for the urine data only.

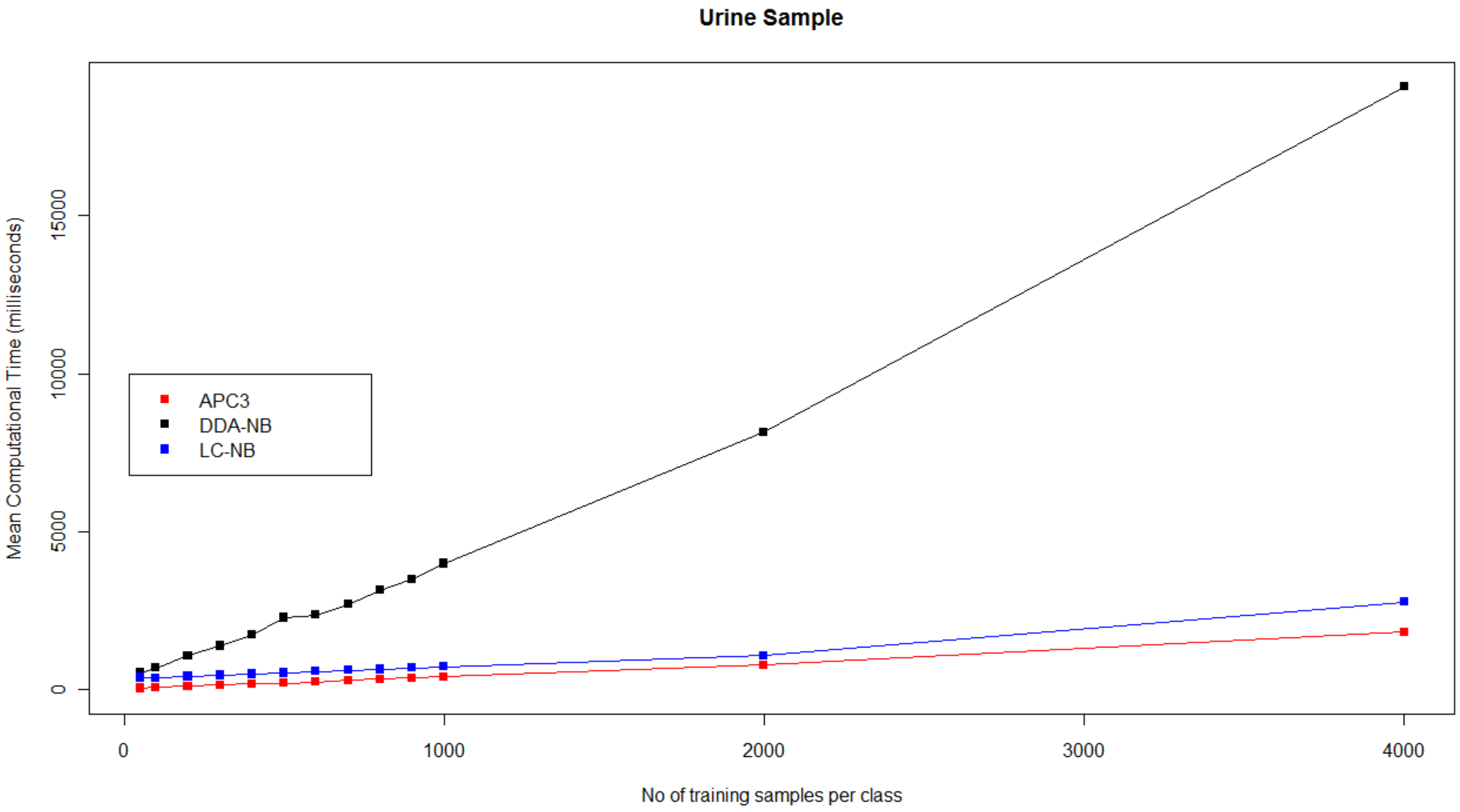


Fig **3.5 a** Average computational time (both the training and test phases) APC3, DDA-NB and LC-NB across larger training sample sizes for the urine data.

The variable length for the urine dataset before dimension reduction was 28641. The average variable length of the training models for the 100 training and testing sets for each urine sample split using DDA-NB and LC-NB are given in the Appendix - Table IX [116]. The mean variable length of the training model appears directly proportional to the number of training samples for both DDA-NB and LC-NB possibly due to increase in sample space.

The average computational time for APC3 as shown in Fig. 3.4.a is almost independent of the change in training sample size for the urine data while both DDA-NB and LC-NB increased in a linear fashion with increasing training size. This is similar to the average computational time of the modelling building phase as shown in Fig. 3.4.b. However for the testing phase, all three approaches were almost independent of the training sample size with APC3 being the fastest for all the various training sample sizes except for training sample size of 14 where DDA-NB was on the average slightly faster than APC3. The boxplot of the computational time for the various urine sample splits for APC3, DDA-NB and LC-NB is shown in the Appendix - Fig. I. DDA-NB was generally computationally more complex than LC-NB. APC3's mean computational speed ranged between 4.7 to 7 times faster than DDA-NB and between 3.9 to 5.9 times faster than LC-NB.

Further evaluation using various sizes for bootstrapping urine sample split A in Fig 3.5 a showed that with increasing sample size, DDA-NB starts to show an exponential increase in computational speed whereas LC-NB's exponential increase is more pronounced after the training set exceeds 2000 samples. APC3 remains computationally efficient with the least computational time for all the training sample sizes compared with DDA-NB and LC-NB and an almost linear computational time with respect to the larger range of training sample sizes of Fig 3.5a.

In our preliminary studies, we also explored the possibility of using non-parametric correlation coefficients such as the Spearman rank's correlation coefficient and the Kendall rank correlation coefficient [151]. However, they did not show good performances. A possible reason is because these methods were designed for use with ordinal variables but our TIC values were real numbers. So the use of these non-parametric methods results in loss of useful information, leading to poor performances.

It is to be noted that data preprocessing plays a key role in pattern recognition. The ideal chemical spectrum for a GC/MS TIC should have well-resolved peaks, adequate signal-to-noise ratios, no background contribution, and a large linear response range between analyte concentration and detector signal for individual samples or runs [89]. If more than one single sample is used, having stable retention times and well-defined peak shapes is ideal [89]. However, due to sample complexity and increasing speed of the chromatographic runs, artefacts such as baseline drifts, changes in the peak shapes and elution times shifts are inherent [90]. Hence the use of data preprocessing is essential to minimize these artefacts and thus aid pattern recognition. In this study, we did not explore the use of such data preprocessing methods because there is no single set of optimal data preprocessing algorithms for TICs and it is not the focus of this study to compare different data preprocessing methods. Evaluating a classification algorithm without the use of data preprocessing would provide a performance baseline for the classification approach which should improve with the appropriate inclusion of data preprocessing.

Another limitation of this study is that it was focused on two-class problems. Extensions to multi-class or continuous response are possible but will require modifications to the algorithm.

3.4 Conclusion

In this study, we developed APC3, which is a fully automated, computationally efficient method based on correlation based feature selection for the development of models in metabonomics. We compared APC3 with various common dimensionality reduction and classification combinations and the results show that APC3 has similar performance as the top few dimensionality reduction and classification combinations. The advantage of APC3 over these dimensionality reduction and classification combinations is that it is fully automated and is 3.9 to 7 times faster than dimensionality reduction and classification combinations. This would minimize user interactivity and allow efficient processing of extremely large datasets in a high throughput approach [102].

APC3 also has better tolerance for outliers in the training set compared to the best two dimensional reduction and classification combination pairs, DDA-NB and LC-NB. This would complement its computational efficiency in high throughput processing as the larger the dataset becomes, the more error prone it will be, according to Shannon's information theory [152]. In addition, considering outliers as a subset of the possible errors that may be present in a dataset, having an approach with a reasonable tolerance to outliers would be advantageous. Moreover, we introduced mislabeling as an attempt to create outliers. In reality, mislabeling of control subjects can commonly occur when they are undiagnosed [153-155].

The successful application of APC3 in processing GC/MS data suggests its potential application in analysing other forms of biological chromatographic data such as LC/MS TICs.

Chapter 4: Study on sparsity embedded classification approaches to classify three-class and two-class datasets

4.1 Introduction

We studied the prediction accuracies and time complexity of various optimizations of the following four state of the art sparsity embedded approaches:

- Shrunken centroids regularized discriminant analysis (RDA) [3]
- Nearest shrunken centroids (NSC) [4]
- Sparse partial least squares - discriminant analysis (sPLS-DA) [5]
- Penalized linear discriminant analysis (PLDA) [6]

RDA

RDA [3] is a modification of the original LDA. Its first use in metabonomics classification was confined to the use of metabonomics profiles [156]. Our study depicts its first evaluation on the classification of total ion chromatograms (TICs). RDA can be understood via the 3 steps below.

Step 1

Apply a general regularization of the within-class covariance matrix to solve the singularity issue. In contrast to LDA which uses a maximum likelihood estimate of the within-class covariance matrix, RDA presents a slightly biased covariance estimate, which not only solves the singularity problem but also stabilizes the sample covariance estimate as follows:

$$\overset{\sim}{\sum_{w}} = \alpha \overset{\wedge}{\sum_{w}} + (1-\alpha)I \, , \alpha \in [0,1] \qquad (4.1)$$

where $\sum_{w}^{\wedge}$ is the standard estimate for within-class covariance matrix. As parameter α shifts towards 0, $\sum_{w}^{\sim}$ becomes the identity covariance matrix, I. However as α moves towards to 1, $\sum_{w}^{\sim}$ is equal to $\sum_{w}^{\wedge}$. The varying of α provides an avenue of adaptability for RDA.

Step 2

Use a shrinkage estimator to gain sparsity. Using the following formula:

$$\mu_{i}^{\wedge *} = \sum_{w}^{\sim -1} \mu_{i}^{\wedge} \qquad (4.2)$$

the class mean, $\mu_{i}^{\wedge}$ is transformed to $\mu_{i}^{\wedge *}$. Then, we shrink $\mu_{i}^{\wedge *}$ towards 0 via:

$$\mu_{i(s)}^{\wedge *} = \mathrm{sign}(\mu_{i}^{\wedge *})(|\mu_{i}^{\wedge *}| - \Delta)_{+} \qquad (4.3)$$

where Δ is the tuning parameter to obtain a new shrunken centroid. The subscript plus indicates positive part, that is,

$(t)_{+} = t \; if \; t > 0 \; and \; zero \; otherwise$).

Then $\mu_{i(s)}^{\wedge *}$ is transformed back to $\mu_{i(s)}^{\wedge}$ using

$$\overset{\wedge}{\mu}_{i(s)} = \sum\nolimits_{w}^{\sim} \overset{\wedge *}{\mu}_{i(s)} \tag{4.4}$$

Step 3

The class of a new sample $\tilde{x}$ is predicted by computing the discriminant score for class i:

$$\begin{aligned} \delta_i(\tilde{x}) &= \tilde{x}^T \sum\nolimits_{w}^{\sim -1} \overset{\wedge T}{\mu}_{i(s)} - \frac{1}{2} \overset{T}{\mu}_{i(s)} \sum\nolimits_{w}^{\sim -1} \overset{T}{\mu}_{i(s)} + \log(\pi_i) \\ &= \tilde{x}^T \overset{*}{\mu}_{i(s)} + \frac{1}{2} \overset{T}{\mu}_{i(s)} \overset{*}{\mu}_{i(s)} + \log(\pi_i) \end{aligned} \tag{4.5}$$

NSC

Unlike RDA, NSC [4] assumes independence among the variables and uses the diagonal estimate of the within-class covariance matrix:

$$\mathrm{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_1^2) \tag{4.6}$$

where $\hat{\sigma}_1^2$ is the jth diagonal element of the within class covariance.

Let x_{ij} be the TIC intensity at position *i* in sample *j*, NSC uses a soft-thresholding rule to shrink the class centroids $\bar{x}_{ik}$ toward the overall centroid $\bar{x}_i$ similar to RDA. Then, via cross-validation (CV), tuning parameter Δ is used to obtain a new shrunken centroid $\bar{x}'_{ik}$ and the discriminant score for class k is defined as

$$\delta_k(x^*) = \sum\nolimits_{i=1}^{p} \frac{(x_i^* - \bar{x}'_{ik})^2}{s_i^2} - 2\log(\pi_k) \tag{4.7}$$

where $x^* = (x_1^*, \dots, x_p^*)$ is a new test sample, S_i is the pooled within-class standard deviation for variable i, and $\pi_k = {}^{n_k}/_{n}$ is the estimated class prior probability. The classification rule is then

$$C(x^*) = l \; if \; \delta_l(x^*) = \min_k \delta_k(x^*) \tag{4.8}$$

sPLS-DA

sPLS-DA is a natural extension of the sPLS [157, 158].

sPLS

sPLS is based on the based on Singular Value Decomposition (SVD) of the cross product:

$$M_h = X_h^T Y_h \tag{4.9}$$

sPLS was first used in identifying subsets of correlated variables coming from a data matrix, *X* and a response matrix, *Y* of sizes (n × p) and (n × q) respectively.

The left and right singular vectors derived from the SVD are denoted as u_h and v_h respectively for the iteration *h* where *h* = 1… *H* where *H* is the number of performed deflations also called the chosen dimensions of the PLS. These singular vectors can be referred to as loading vectors in the PLS context. Sparse loading vectors were then obtained by applying l_1 penalization on both u_h and v_h. The optimization problem of the sPLS minimizes the Frobenius norm between the current cross product matrix and the loading vectors as follows:

$$\min_{u_h, v_h} ||M_h - u_h v_h'||^2 F + P_{\lambda 1}(u_h) + P_{\lambda 2}(v_h) \quad (4.10)$$

where $P_{\lambda 1}(u_h) = sign(u_h)(|u_h| - \lambda_1)_+$ and $P_{\lambda 2}(v_h) = sign(v_h)(|v_h| - \lambda_2)_+$ are applied componentwise in the vectors u_h and v_h and are the soft thresholding functions that approximate Lasso penalty functions [159]. They are simultaneously applied on both loading vectors. The problem (4.10) is solved with an iterative algorithm [158] and the $X_h$and Y_h matrices are subsequently deflated for each iteration h. For practical purposes, the user can input the number of variables to select on each data set rather than the penalization parameters λ_1 and λ_2.

sPLS extended to sPLS-DA

sPLS can be extended to sPLS-DA by coding the response matrix, Y of size (n × K) with dummy variables to indicate the class membership of each sample. Variable selection is only performed on the X data set to select the discriminative features that can help predicting the classes of the samples. The Y dummy matrix remains unchanged. Therefore, we still use

$$M_h(x^*) = X_h^T Y_h \quad (4.11)$$

but the optimization of sPLS-DA has one term less the sPLS and is defined as:

$$\min_{u_h, v_h} ||M_h - u_h v_h'||^2 F + P_{\lambda}(u_h) \quad (4.12)$$

Here we only need to tune a single term, λ.

sPLS-DA [5] has been implemented to choose the number of variables to select instead of choosing λ for practical reasons. For the class prediction of test samples, we can use distance measures such as maximum, centroids or Mahalanobis distances.

PLDA

Similar to NSC, PLDA [6] assumes independence among the variables and uses the diagonal estimate of the within-class covariance matrix. Sparsity is achieved via imposing a penalty function such as lasso or fused lasso penalties on the discriminant vectors by the following

$$maximize_{\beta_k} \left(\beta_k^T \overline{\sum\nolimits_b^k} \beta_k - \lambda_k \sum\nolimits_{j=1}^p \left| \hat{\sigma}_k \beta_{jk} \right| \right) subject\ to\ \beta_k^T \underline{\sum\nolimits_w \quad} \beta_k \qquad (4.13)$$

where β_k is the k-th discriminant vector, $\underline{\sum_w \quad}$ is diagonal estimate of the within-class covariance matrix $\sum_w$, and $\overline{\sum_b \quad}$ is the standard estimate for the between-class covariance matrix $\sum_b$. Using a large value for the tuning parameter λ_k will result in some elements of $\overline{\beta_k}$ becoming equal to 0. As $\hat{\sigma}_k{}^2$ is used in the penalty function, the variables which vary more within each class would undergo greater penalization. Finally, an optimization algorithm is applied to maximize the objective function which is not concave.

4.2 Methodology

The wine classes were chosen according to their location of origin namely Australia, Chile and South Africa with sample sizes of 12, 15 and 11 respectively. The wine data was split into training and testing set, with each having all three classes. For simplicity, wine from Australia, Chile and South Africa were reference to belong to class A, B and C respectively. To ensure that at least half of each class was used in the training set, each training set consisted of 8 samples per class. To reduce bias due to the splitting of the dataset, we repeated the experiments 100 times with different training and test sets. The training set was further split into modelling and validation sets during cross validation (CV) to optimize the model parameters. The optimal model parameters were used to create a model using the full training set consisting of both entire validation and modelling sets. This model was used to predict the classes of the testing set.

From the results of the wine datasets, we selected the top 2 algorithms to compare against APC3 on the urine data using a similar methodology as outlined in Section 3.2 of Chapter 3. Hence, we used various training and testing sizes as shown in Table 3.2. For each training and testing size, just as in the wine data evaluation and also in Chapter 3, we repeated the experiments 100 times with different training and testing sets. To correspond closely to Section 3.2's evaluation, we also tested for performance in the presence of outliers in the urine data. We randomly assigned one of the BC TIC in the training set to have the class value of H to make that TIC appear as a H outlier and repeated the experiments of the urine data. Next, we randomly assigned one of the H TIC in the training set to have the class value of BC to make that TIC appear as a BC outlier and similarly repeated the urine data experiments.

4.2.1 Choice of CV method

Although the stratified 10-fold CV is the commonly used CV method [160] and stratification has been substantiated to improve variance of the CV estimator by ensuring that in each fold, each class is represented with approximately equal proportions for the modelling and validation sets [160], we did not use a stratified 10-fold CV due to a sample limitation of 8 per class in the training set. Previous metabonomic classification studies involving PLS-DA [161, 162], Random forest [163], PLS-k nearest neigbhbours (kNN) [164] and Soft independent modelling of class analogies (SIMCA) [165] have previously utilized leave one out (LOO) as the only CV method studied. However, it is worth to note that there is a bias-variance trade-off [160, 166, 167] with the choice of *k* in a *k*-fold CV with a largest *k* as in LOO-CV having highest variance since LOO-CV is inherently non-stratified though LOO-CV comes with minimal bias. Hence in order to maximize the number of folds in CV while maintaining stratification, we used stratified 8-fold CV as the CV method for all the optimization variations in the sparsity approaches evaluated. For stratified 8-fold CV, in every fold, the modelling set would consist of 7 samples per class while the corresponding validation set would consist of the remaining 1 sample per class.

In general, CV involves sampling without replacement and if re-sampling is involved, it would be explicitly referred to as including bootstrapping or bagging [160]. We did not explore other CV methods such as repeated CV where the *k*-fold CV is repeated *m* number of times which is expected to reduce the variance as a previous study [168] has demonstrated that in a low sample setting, non-repeated-CV generally gives better accuracy than repeated CV methods for metabonomic data. The same previous study [168] also showed that bootstrapping or bagging did not improve the best performer and though it did improve the accuracy of a few feature selection methods, it also impeded the accuracy other methods. A disadvantage of bootstrapping or bagging is the introduction of unbalanced modelling set with the repetition of certain samples via resampling [169]. Hence we also did not study the addition of bootstrapping or bagging to the CV which involves sampling with replacement to increase the number of validation folds beyond that used for stratified 8-fold CV.

The following parameters were optimized during the cross validation for the various algorithms:

RDA: $\Delta \in \{0 : 0.11 : 0.99\} \times \alpha \in \{0 : 0.33 : 3\}$, which are the default values in the rda package.

NSC: Δ, using the default 30 values in the pamr package.

PLDA: $\lambda \in \{0 : 0.03 : 3\}$, λ_2=0.3 (default)

4.2.2 Optimizations studied for each sparsity approach

We describe the various optimizations tested for each sparsity approach below.

4.2.2.1 RDA

The availability to vary the regularization option for RDA gave rise to two versions of RDA being studied. One was the regularization via covariance which would be referred to as RDA_{cov} and the other was the regularization via correlation which would be known as RDA_{cor}. For both versions, the optimization is carried out as follows. First, we find all the parameter pairs for *alpha* and *delta* that correspond to the minimal CV error. Then, if there are parameters give the same CV error rate, the first parameter pair that gives the smallest number of variables was chosen.

4.2.2.2 NSC

Three different optimization methods were applied to NSC giving rise to three different versions that were evaluated. For NSC_{minV}, we find all the values for *delta* that correspond to the minimal CV error. From this subset of *delta*, the first *delta* that gives the corresponding smallest number of variables was chosen. Then we find the corresponding threshold for this *delta*. For NSC_{minT}, first, we find all the values for *delta* that correspond to the minimal CV error. From this subset of *delta*, we find the values of *delta* corresponding minimum threshold. The first delta that gives the corresponding smallest number of variables was chosen. NSC_{maxT} is similar to NSC_{minT} except in the second filter, instead of finding the corresponding minimum threshold; we chose the corresponding maximum threshold. A summary of each NSC version is presented in Table 4.1.

Table 4.1 Comparison of the individual steps for each NSC version.

	NSC$_{minV}$	**NSC$_{minT}$**	**NSC$_{maxT}$**
Step 1	Find all the values of *delta* that correspond to the minimal CV error		
Step 2	From this subset of *delta*, choose the first *delta* the corresponds to the smallest number of variables	From this subset of *delta*, choose the values of *delta* with the minimum threshold	From this subset of *delta*, choose the values of *delta* with the maximum threshold
Step 3	-	From this sub-subset of *delta*, choose the first *delta* with the smallest number of variables	From this sub-subset of *delta*, choose the first *delta* with the smallest number of variables

4.2.2.3 sPLS-DA

Seven different versions of sPLS-DA were studied. sPLS-DA has been evaluated only with the default maximum distance measure for prediction in previous literature [5, 170] while other distance measures such as centroids or Mahalanobis remain unstudied. In this study, we attempted to also experiment with various versions using the maximum distance, centroid distance and a combination of both.

The following version of sPLS-DA used only the maximum distance for prediction:

sPLS-DA$_{max}$ *: optimize using minimum CV errors for 50 components; use the corresponding number of variables and components and the maximum distance measure for prediction*

The following version of sPLS-DA used only the centroid distance for prediction:

sPLS-DA$_{cen}$: *optimize using minimum CV errors for 50 components; use the corresponding number of variables and components and the centroid distance measure for prediction*

The remaining five versions of sPLS-DA used dynamically chose either the maximum or the centroid distance for prediction depending on its respective optimization:

sPLS-DA$_{com}$: *get the number of variables for **sPLS-DA$_{max}$** and **sPLS-DA$_{cen}$**; if **sPLS-DA$_{max}$** has the lesser number of variables, use its corresponding number of components and the maximum distance measure for prediction, otherwise use the corresponding number of components in **sPLS-DA$_{cen}$** and the centroid distance measure for prediction*

sPLS-DA$_{var}$: *get the number of components for **sPLS-DA$_{max}$** and **sPLS-DA$_{cen}$**; if **sPLS-DA$_{max}$** has the lesser number of components, use its corresponding number of variables and the maximum distance measure for prediction, otherwise use the corresponding number of variables in **sPLS-DA$_{cen}$** and the centroid distance measure for prediction*

sPLS-DA$_{com*var}$: *Get the value of the number of components multiplied by the number of variables for **sPLS-DA$_{max}$** and **sPLS-DA$_{cen}$**; if **sPLS-DA$_{max}$** has the lesser value, use its corresponding number of variables and components and the maximum distance measure for prediction, otherwise* *use the corresponding number of variables and components in **sPLS-DA$_{cen}$** and the centroid distance measure for prediction*

***sPLS-DA$_{com/var}$**: Get the value of the number of components divided by the number of variables for **sPLS-DA$_{max}$** and **sPLS-DA$_{cen}$**; if **sPLS-DA$_{max}$** has the lesser value, use its corresponding number of variables and components and the maximum distance measure for prediction, otherwise use the corresponding number of variables and components in **sPLS-DA$_{cen}$** and the centroid distance measure for prediction*

***sPLS-DA$_{var/com}$**: Get the value of the number of variables divided by the number of components for **sPLS-DA$_{max}$** and **sPLS-DA$_{cen}$**; if **sPLS-DA$_{max}$** has the lesser value, use its corresponding number of variables and components and the maximum distance measure for prediction, otherwise use the corresponding number of variables and components in **sPLS-DA$_{cen}$** and the centroid distance measure for prediction*

4.2.2.4 PLDA

Two versions of PLDA were studied. PLDA$_s$ was the standard type where lasso penalties were used. PLDA$_o$ was the ordered type where fused lasso penalties were used. We find all the values for *lambda* that correspond to the minimal CV error. The first *lambda* that gives the corresponding smallest number of discriminant vectors was chosen.

4.2.3 Prediction accuracy indices

The overall prediction accuracy in each sampling of training and testing sets is the percentage derived from total number of test classes predicted correctly divided by the total number of test samples. This gave higher weightage to the class with more test classes. We had a fixed number of test samples which were 4, 7 and 3 for class A, B and C respectively which gave a total number of test samples to be consistently 14.

$$Overall\ Prediction\ accuracy\ = \frac{(\text{total number of test samples predicted correctly})}{14} \qquad (4.14)$$

Take for instance, the number of test classes predicted correctly for class A, B and C were 2, 5 and 2 respectively, then the prediction accuracy would be as follows

$$Overall\ Prediction\ accuracy\ = \frac{(2+5+2)}{14} \approx 0.643$$

The mean and median of the overall prediction accuracies across the 100 samplings for each approach was computed. Even though the training classes were ensured to have an equal number of samples per class, the availability of an unequal number of test samples for each class might create a bias in the overall prediction accuracy towards the classes with more test samples.

The area under the response operating characteristic curve (AUC) [125] has been widely used as a prediction metric for analysing the performance of two-class classification which includes the size of the individual class in its computation. However there has not been an established standard similar to

AUC for multi-nominal classification where number of classes is more than 2. Over the last decade, receiver operating characteristic (ROC) surface analysis have been proposed as extensions of the two-class AUC to three-class classification problems but their usage has not been widely accepted because it has still not reached a theoretically robust state [171].

In this section, we introduce the following computationally appealing measures of overall prediction accuracies.

$$
\begin{aligned}
&\mathit{Overall\ prediction\ accuracy\ A, OPA(A)}\\
&= \left.\begin{pmatrix} \mathit{Mean\ prediction\ accuracy\ for\ class\ A} \\ +\mathit{mean\ prediction\ accuracy\ for\ class\ B} \\ +\mathit{Mean\ prediction\ accuracy\ for\ class\ C} \end{pmatrix}\middle/ 3\right.
\end{aligned}
\tag{4.15}
$$

$$
\begin{aligned}
&\mathit{Overall\ prediction\ accuracy\ B, OPA(B)}\\
&= \left.\begin{pmatrix} \mathit{Median\ prediction\ accuracy\ for\ class\ A} \\ +\mathit{Median\ prediction\ accuracy\ for\ class\ B} \\ +\mathit{Median\ prediction\ accuracy\ for\ class\ C} \end{pmatrix}\middle/ 3\right.
\end{aligned}
\tag{4.16}
$$

$$
\begin{aligned}
&\mathit{Overall\ prediction\ accuracy\ C, OPA(C)}\\
&= \mathit{Median}\begin{pmatrix} \mathit{Median\ prediction\ accuracy\ for\ class\ A} \\ ,\mathit{Median\ prediction\ accuracy\ for\ class\ B} \\ ,\mathit{Median\ prediction\ accuracy\ for\ class\ C} \end{pmatrix}
\end{aligned}
\tag{4.17}
$$

4.2.4 Reason to proceed with sample size limitation

Franceschi et al. [172] were the pioneers to study the effects of sample size on metabonomic biomarker identification where the smallest sample size per class being evaluated was 3. Small experimental sample size may arise for various reasons, such as when there is limited availability of biological samples (e.g. those from rare conditions or diseases), or due to enforcement of complex protocols in the experiments. Hence, it would be meaningful to present a basis study using sparse embedded approaches on a reduced sample sized data. Furthermore, our analysis of a tri-class dataset would complement the biomarker identification study [172] which also evaluated small sample sizes per class for CV but used bi-class datasets. As the biomarker identification study [172] presented the performance based on the Receiver Operating Characteristics (ROC) [173] of the CV set without evaluating it on any separate testing set, our study would fill this important gap by describing the performance of the CV set on a separate testing set.

4.3 Results for time complexity

4.3.1 Overall

Generally, NSC is the least computationally intensive, followed by RDA, sPLS-DA and finally PLDA as shown in **Fig. 4.1**. It can also be noted that the standard deviation of the computation time is significantly correlated using Pearson's correlation to the mean (*r*=0.73, *p*<0.003) and median (*r*=0.74, *p*<0.003) computational speeds. In other words, there is a degree of proportionality between the computation time and the standard deviation of the computation time which can be extrapolated to the fact that a faster algorithm is more consistent in its time taken for processing. This would complement the attractiveness of choosing a faster algorithm to process.

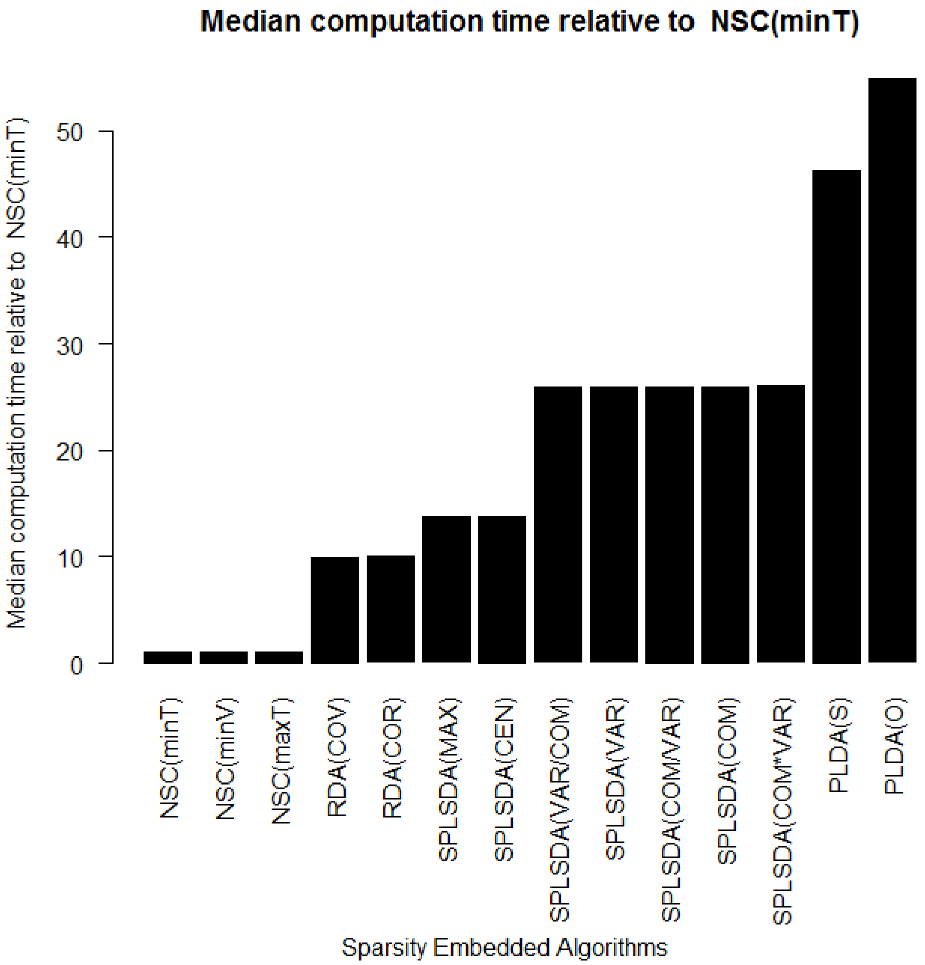


Fig. 4.1 Bar graph of median computation times relative to the fastest algorithm, NSC$_{\mathrm{minT}}$.

Table 4.2 Comparison of computation time for the various classification techniques.

Classification Technique	**Mean time in seconds (SD)**	**Median computation time relative to NSC$_{minT}$**
NSC$_{minT}$	2.95 (0.102)	1.00
NSC$_{minV}$	3.02 (0.242)	1.01
NSC$_{maxT}$	3.16 (0.296)	1.05
RDA$_{cov}$	29.28 (0.289)	9.95
RDA$_{cor}$	29.61 (0.590)	10.03
sPLS-DA$_{max}$	40.23 (0.354)	13.66
sPLS-DA$_{cen}$	40.61 (0.452)	13.79
sPLS-DA$_{var/com}$	76.07 (0.534)	25.86
sPLS-DA$_{var}$	76.08 (0.472)	25.88
sPLS-DA$_{com/var}$	76.28 (0.639)	25.92
sPLS-DA$_{com}$	76.31 (0.575)	25.92
sPLS-DA$_{com*var}$	76.53 (0.625)	26.00
PLDA$_s$	137.52 (1.261)	46.32
PLDA$_o$	159.47 (5.783)	54.88

SD = standard deviation

4.3.2 NSC

Within the NSC versions, NSC$_{minT}$ is the fastest as shown in **Table 4.2**. Despite NSC$_{minT}$ having an additional step compared to NSC$_{minV}$ as shown in **Table 4.1**, its faster speed compared to NSC$_{minV}$ suggests that the filtering of the smaller thresholds is a negligible overhead which not only improves the computational complexity of choosing the minimum number of variables but also the computation time of the entire algorithm.

4.3.3 RDA

Amongst the RDA versions, RDA_{cov} has the highest computational efficiency. This is due to the fact that RDA_{cor} has an inherent computational overhead since it requires an additional computation involving the standardization of the covariance.

4.3.4 sPLS-DA

For the various sPLS-DA versions, $\text{sPLS-DA}_{\text{max}}$ is the fastest. $\text{sPLS-DA}_{\text{max}}$ is on the average 0.3s faster than $\text{sPLS-DA}_{\text{cen}}$ possibly due to a simpler computation using the maximum distance for prediction as compared to the centroid distance. $\text{sPLS-DA}_{\text{max}}$ and $\text{sPLS-DA}_{\text{cen}}$ are almost twice as fast as the other five versions which is expected since these five versions involved the computation of both the centroid and maximum distance based predictions following by selecting the prediction from the predictor with the better respective metric for a corresponding version as described in 4.2.2.3.

4.3.5 PLDA

For PLDA, PLDA_{s} has the faster computational time. Hence the use of lasso penalties is computationally quicker as compared to the use of fused lasso penalties.

4.4 Results for the prediction accuracy

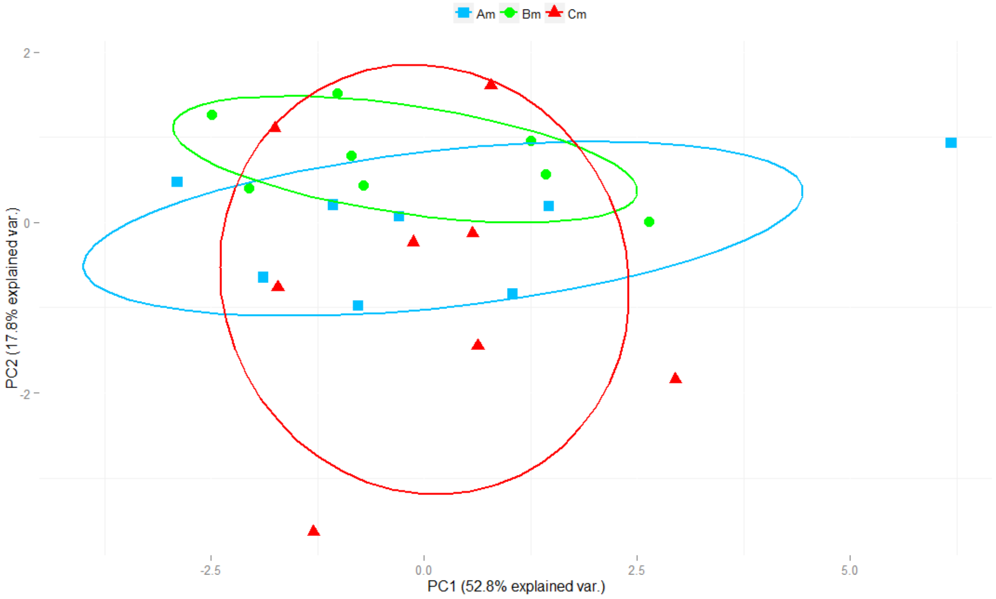
Table 4.3 Comparison of prediction accuracy indices for the various algorithms. Cells highlighted in **blue**, **light blue** and **turquoise** indicate the column wise 1st, 2nd and 3rd algorithms with the highest prediction accuracies. Columns with more than 1 cell highlighted for a colour denotes ties.

	Mean (SD)	**OPA(A)**	**OPA(B)**	**OPA(C)**
RDA$_{cov}$	0.739 (0.142)	0.675	0.722	0.667
RDA$_{cor}$	0.844 (0.100)	0.813	0.889	1.000
NSC$_{minV}$	0.809 (0.107)	0.757	0.806	0.750
NSC$_{minT}$	0.794 (0.120)	0.748	0.782	0.750
NSC$_{maxT}$	0.814 (0.0908)	0.762	0.806	0.750
sPLS-DA$_{max}$	0.796 (0.113)	0.756	0.806	0.750
sPLS-DA$_{cen}$	0.787 (0.114)	0.749	0.782	0.750
sPLS-DA$_{com}$	0.789 (0.119)	0.756	0.758	0.750
sPLS-DA$_{var}$	0.791 (0.113)	0.756	0.758	0.750
sPLS-DA$_{com*var}$	0.791 (0.107)	0.756	0.758	0.750
sPLS-DA$_{com/var}$	0.781 (0.107)	0.743	0.758	0.750
sPLS-DA$_{var/com}$	0.801 (0.107)	0.762	0.806	0.750
PLDA$_{s}$	0.653 (0.166)	0.637	0.710	0.714
PLDA$_{o}$	0.614 (0.132)	0.583	0.599	0.714

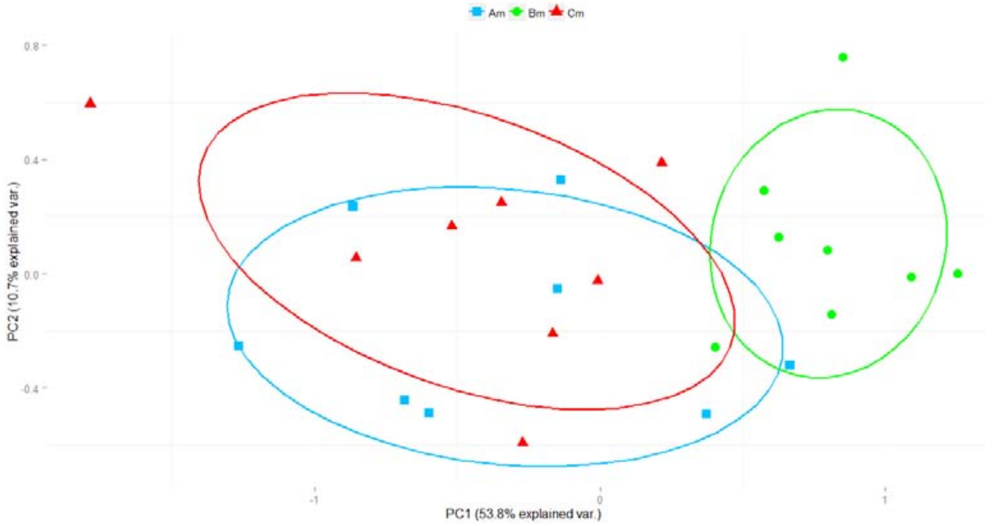
SD = Standard deviation

From **Table 4.3**, we can see that RDA_{cor} is consistently the best performer using every prediction accuracy indices followed NSC_{maxT} although prediction accuracies: OPA(A), OPA(B) and OPA(C) had multiple 2nd best performers. NSC_{minT} was ranked third using the mean and OPA(A) while OPA(B) and OPA(C) displayed multiple 3rd best performers.

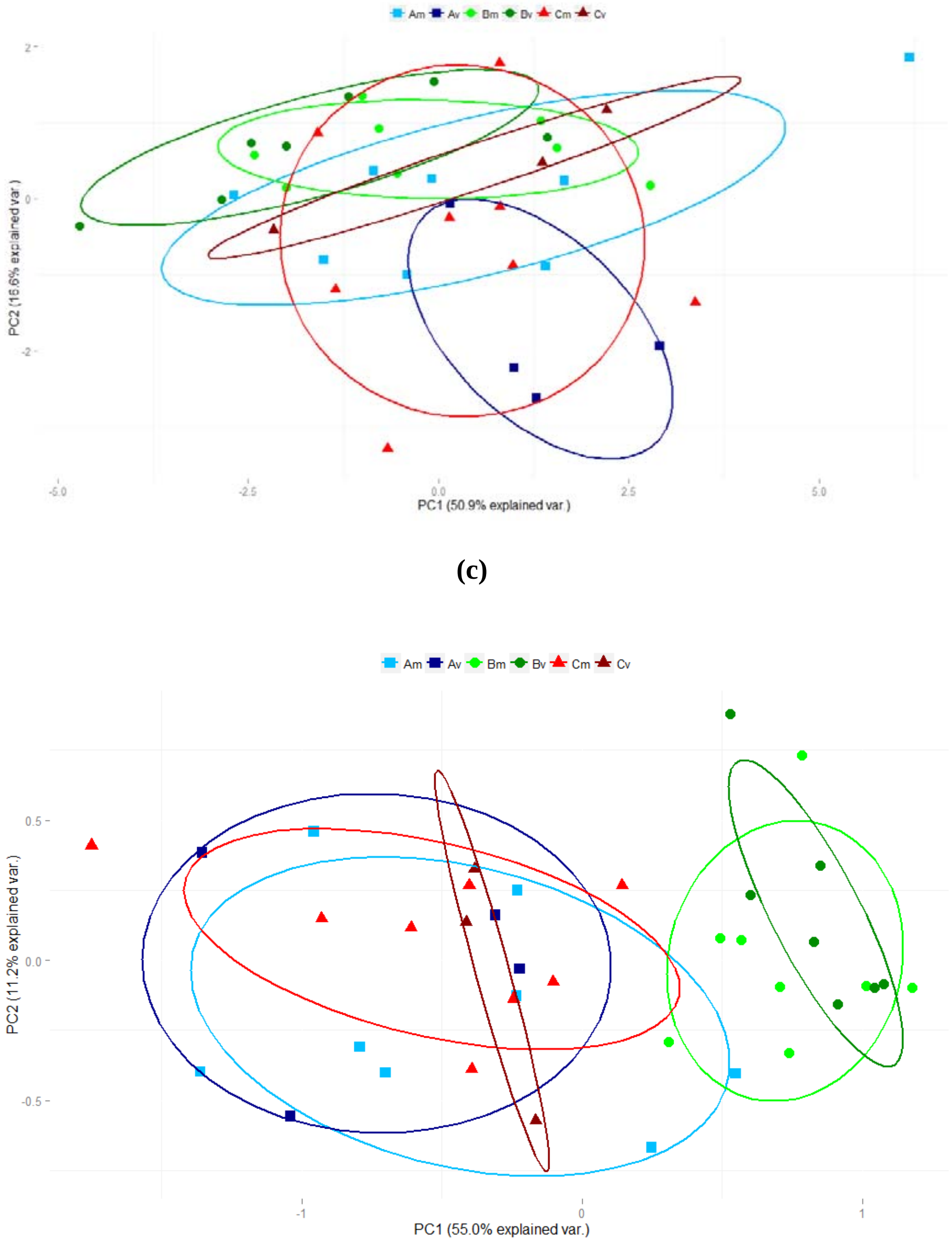
Fig. 4.2 shows the 1st two PCA components of RDA_{cor} before and after variable selection using the modelling set alone and also including both the modelling and validation sets for a single experiment. A more prominent separation between the three class types can be observed after variable selection in **Fig. 4.2 (b)** compared to **Fig. 4.2 (a)** using only the modelling set and in **Fig. 4.2 (d)** compared to **Fig. 4.2 (c)** using both the modelling and validation sets.


Am
Bm
Cm
2
0
-2
PC2 (17.8% explained var.)
-2.5
0.0
2.5
5.0
PC1 (52.8% explained var.)

(a)


Am
Bm
Cm
0.8
0.4
0.0
-0.4
PC2 (10.7% explained var.)
-1
0
1
PC1 (53.8% explained var.)

(b)



(c)

(d)

Fig. 4.2 PCA Plots with 1st two principal components of RDA$_{cor}$: **(a)** the modelling set with all the variables **(b)** the modelling set after variable selection **(c)** both the modelling and validation sets with all the variables **(d)** both the modelling and validation sets after variable selection

Am: modelling set for class A
Av: validation set for class A
Bm: modelling set for class B
Bv: validation set for class B
Cm: modelling set for class C
Cv: validation set for class C

The mean prediction accuracies across the 100 experiments using the urine data for the different versions of both NSC and RDA are shown in **Fig. 4.3 (a).** It can be seen that all 3 NSC versions had better mean prediction accuracy than the 2 RDA versions across all the training sample sizes including the outlier testing experiments. **Fig. 4.3 (b)** shows the prediction accuracies of APC3 and the top 2 sparsity embedded algorithms from **Fig. 4.3 (a)** which are NSC_{minV} and $NSC_{maxT.}$ A one tailed paired t-test with a $p < 0.05$ to denote significance was used to compare the accuracy between every pair of algorithms for the various training sample sizes including the outlier experiments in **Fig. 4.3 (b)**. For the non-outlier experiments, both NSC_{minV} and NSC_{maxT} had significantly better accuracy than APC3 for all training sample sizes. For the outlier experiments when a bladder cancer patient was made an outlier by misclassifying it as a healthy subject in the training set, both NSC_{minV} and NSC_{maxT} had significant better accuracy than APC3 for all the training sample sizes except for the largest training sample size of 36. For the outlier experiments when a healthy subject was made an outlier by misclassifying it as a bladder cancer patient, both NSC_{minV} and NSC_{maxT} had significantly better accuracy than APC3 only when the training samples sizes were either 32 or 28 while it was vice-versa with APC3 significantly outperforming both NSC_{minV} and NSC_{maxT} for the smallest the training sample size of 16. Additionally for the outlier experiments when a healthy subject was made an outlier by misclassifying it as a bladder cancer patient, NSC_{minV} also had significantly better accuracy than APC3 for a training sample size of 28.

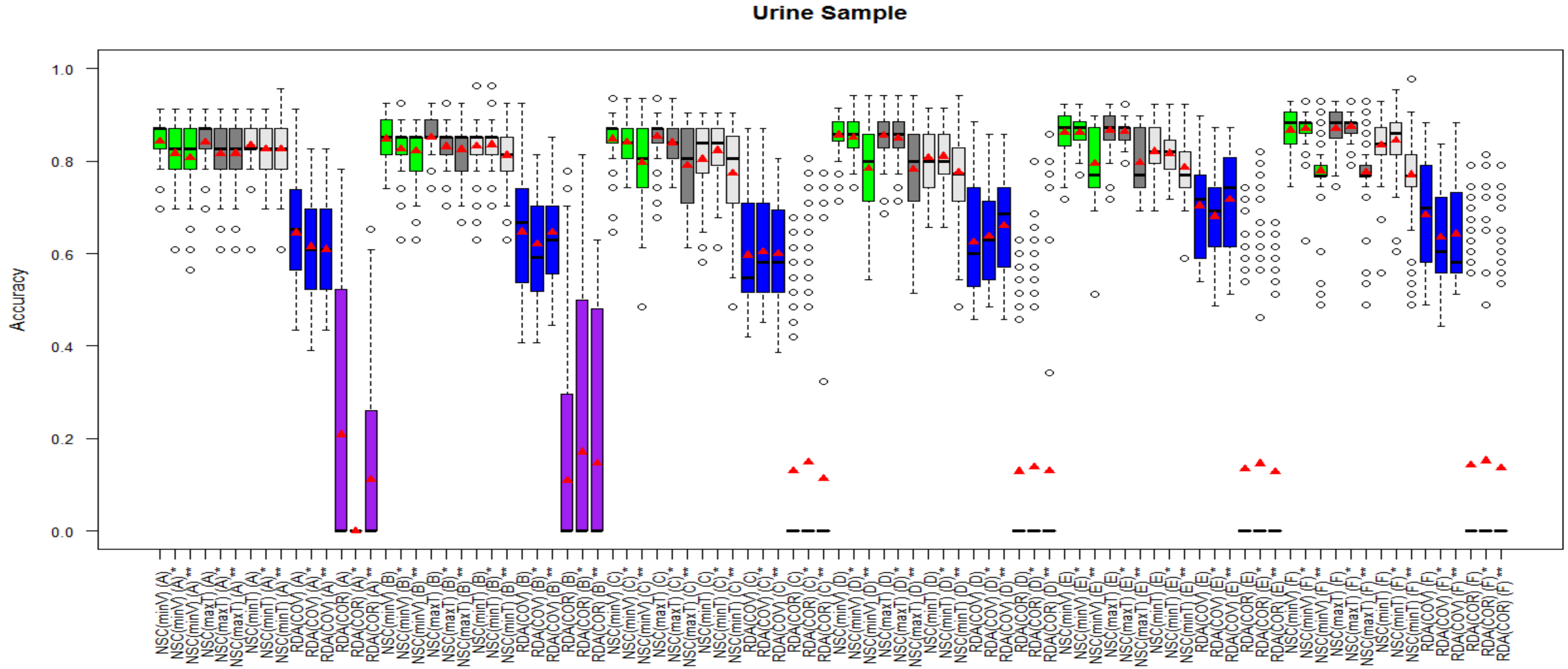


Fig. 4.3 a Accuracy boxplot of the average accuracy for the urine data set splits for the different versions of both NSC and RDA. The red triangle represents the mean value of each boxplot.

* one of the BC TIC in the training set assigned the class value of H.** one of the H TIC in the training set assigned the class value of BC.

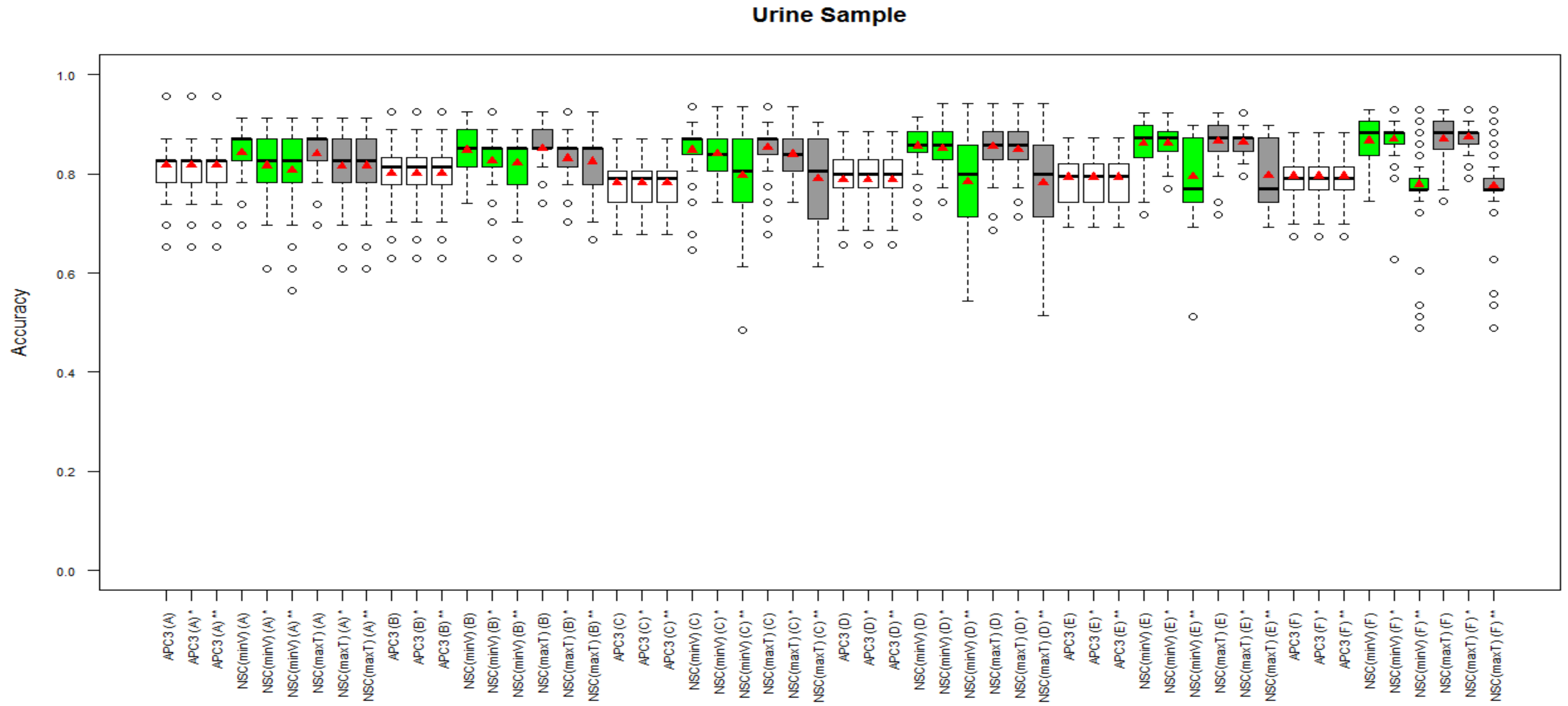


Fig. 4.3 b Accuracy boxplot of the average accuracy for the urine data set splits for the different versions of NSC with APC3. The red triangle represents the mean value of each boxplot.

* one of the BC TIC in the training set assigned the class value of H.** one of the H TIC in the training set assigned the class value of BC.

4.5 Conclusion

For the wine dataset, RDA_{cor} emerged as the most accurate approach whereas NSC_{minT} is the fastest technique. RDA_{cor} is around 10 times slower the $NSC_{minT.}$ Considering the fact that NSC_{maxT} is only 1.05 times slower than NSC_{minT} but NSC_{maxT} is the most accurate amongst the NSC versions tested and NSC_{maxT} second to RDA_{cor} in terms of accuracy, it appears to be a choice between accuracy and speed to select between RDA_{cor} and NSC_{maxT} respectively for the wine dataset.

Although PLSR is the preferred approach in metabonomic data modelling and classification [98, 99], it is interesting to note that a sparsity embedded variant of PLSR, sPLS-DA, has worse accuracy compared to the RDA and NSC approaches. We hypothesize the requirement of large training sample size and large number of indicators of each latent variable for PLSR [100, 101], could have been inherited to its sparsity embedded variant, sPLS-DA. However, this could only be evaluated against a larger dataset to verify if there is any change in performance accuracy ranking in comparison to the RDA and NSC versions.

With respect to the urine dataset, NSC_{minV} and NSC_{maxT} were the best performers amongst the sparsity embedded approaches and hence we would propose to use these two for further evaluations. Although NSC_{minV} and NSC_{maxT} outperformed APC3 generally for the non-outlier experiments with a maximum of 7.5% difference in mean prediction accuracy, APC3 was able to perform significantly better for the smallest training sample size for one of the outlier experiments. In comparison to APC3, the reduction of training sample size and outlier inclusion in the training sample is more deteriorating to both NSC_{minV} and NSC_{maxT} as clearly shown **Fig. 4.3 (b).** Hence in terms of consistency in prediction accuracy even when using a small training set and with possibility of having outliers in the training set, APC3 would be preferred. It would be interesting to evaluate NSC_{minV}, NSC_{maxT} and APC3 with more datasets to further establish the consistency of their performance.

The urine dataset has been evaluated with conventional dimensional reduction approaches in Chapter 3 with APC3 ranking similarly or better to two of the top performing conventional dimensional reduction approaches tested. The better performance of both NSC_{minV} and NSC_{maxT} in comparison to APC3 particularly for the non-outlier experiments can serve as evidence that noise eradication in addition to dimensional reduction does improve prediction accuracies. However, this combination of noise eradication with dimensional reduction may not have the stability to withstand a reducing training sample size with the possibility of inherent outliers.

Chapter 5: Major Contributions, Limitations and Future Recommendations

5.1 Major Contributions

In this thesis, three separate studies were performed to address the current limitations that exist within the metabonomic workflow as shown in Fig 1.4. AIMA [115] fits into the pre-processing phase of the metabonomic workflow by addressing the limitation of baseline artefacts. Both APC3 [174] and the investigation of the sparsity embedded classification algorithms fit into the data analysis phase of the metabonomic workflow by dealing with the limitations of existing metabonomic classification algorithms such as the need for large training datasets and large number of features.

5.1.1 Parameter-less algorithms

Both AIMA [115] and APC3 [174] are fully automated and hence would be computationally efficient and fast to prepare for the emergence of Big Data [109]. The fully automated attribute of AIMA [115] and APC3 [174] provides a parameter-less feature that exempts these two algorithms from the need to tune any parameter. This gives AIMA [115] and APC3 [174] a computational edge over other non-fully automated algorithms. A two-fold benefit can be observed from both AIMA [115]'s and APC3 [174]'s fully automated feature.

First is the notion that there is no single perfect baseline correction method that performs well for all regions of a spectra and the best approach would be to test several baseline correction techniques to achieve the best prediction [111, 112]. With reference to this notion and when testing a wide range of baseline approaches to achieve the best prediction, a fully automated and fast baseline correction approach, such as AIMA [115], would be an ideal choice in such a suite of baseline approaches since it would not add much effort nor time to the tests. In contrast, non-automated algorithms often require parameter optimization via a grid search through a large combination of parameters, which is very time-consuming. Hence, it would not be practical in most cases to include many non-automated algorithms in the suite of baseline approaches to be tested. For the data analysis phase, APC3 [174] has similar advantages as AIMA due its parameter-less feature. Hence, it would be useful and convenient to add APC3 to the suite of classification algorithms to be tested.

Next is with the rise of Big Data [109], a fully automated baseline correction algorithm and a fully automated classification algorithm would be easier and faster to run as it requires minimal user interactivity.

By doing away with a computational overhead to optimize parameters, parameter-less algorithms may improve the speed of the analysis in clinical diagnosis if the analytical workflow requires the concatenation of the training dataset with the diagnostic data to redo the pre-processing for steps such as retention time alignment. Retention time alignment may require selection the common peaks from all the samples to be used 'Retention Reference' [175]. This may also require remodelling with the re-processed training dataset. This speed up in the clinical diagnosis can be crucial for conditions with high mortality rate such as sepsis [176].

The speed of analysis would aid to provide a close to real time analytics in a specific subset of metabonomics known flux-analysis or fluxomics for experiments such as *in-vitro* toxicity studies. These studies require sampling of data to be performed at high speed and via multiple sequential time points [177]. The availability of fast analytics would minimize the delay in deciding on the proceeding experiments to perform or whether experiments need to be repeated.

Real time metabonomics is also taking centre stage with real time profiling techniques developed for uses such as screening of microbial metabolic output [178]. Micro-organisms may produce metabolites that are vital for agriculture, biological research, and drug discovery [178]. For users of advanced rapid profiling techniques where the bottleneck no longer lies in the domain of the acquisition, fast analytical algorithms would make the right fit.

5.1.2 Addressing the need for dimensional reduction

Since metabonomic data tend to be very large [102], dimensional reduction is highly recommended. APC3 [174] is a computationally efficient classification algorithm that combines both dimensional reduction and classification into a single algorithm via prediction based on correlation based feature selection.

Sparsity embedded classification approaches has the ability to perform simultaneous dimensional and noise reduction via the use of embedded sparse structures [7]. In metabonomics, the use of sparsity embedded classification algorithms has been limited to only metabolic profiles. This work is the first to evaluate its efficacy directly on minimally pre-processed metabonomic TICs. The results from this work will serve as a baseline study for future metabonomic TIC analytics using these algorithms. The evaluation of sparsity embedded classification approaches on three-class and two-class GC/MS TIC datasets showed that $\mathrm{NSC}_{\mathrm{maxT}}$ showed consistently promising results. However, in terms of robustness against the presence outliers or mislabelled samples with a small training set, APC3 would take the lead. The choice between $\mathrm{NSC}_{\mathrm{maxT}}$ and APC3 is more clear-cut when the number of classes is more than two which makes $\mathrm{NSC}_{\mathrm{maxT}}$ the only possible option amongst the two.

The evaluation of APC3 [174] and the sparsity embedded approaches in processing GC/MS data suggests a natural extension of their usage to other types of biological chromatographic data such as LC/MS TICs without much modifications. APC3 will be applicable for other two-class spectra data, whereas sparsity embedded approaches will be applicable for two-class and multi-class spectra data.

5.2 Limitations and Future Studies

The major limitation in the evaluation of AIMA [115] is the datasets which it was evaluated on. Although baseline correction serves to remove artefacts, the ultimate goal would be to better differentiate spectra originating from different classes. The HPLC chromatograms and Raman spectra were both single classed spectra and hence it would be useful to have further evaluations on two or more-class datasets of such spectra. The LC-MS dataset was a two-class dataset but it was limited in the sample size and hence a useful future evaluation would be on a multi-classed larger sample size LC-MS dataset.
A possible future extension of AIMA [115], other than to use a more extensive dataset space, would be to evaluate AIMA [115] as a standalone algorithm and compare it against an approach that includes it as part of a suite of several baseline correction techniques [111, 112].

Another limitation of AIMA [115] could the choice of MZmine [130] as the platform for implementation. There are 28 citations to MZmine [130] in literature to date which is very limited. There are almost 20 citations to AIMA [115], most of which cited AIMA [115] as an example in their literature review, while some cited it as an additional reference for airPLS [96]. It could be that airPLS [96] being available in Matlab is more convenient to evaluate. Krier et al [111] was the only paper to have evaluated AIMA [115] and even in that paper, the authors re-implemented AIMA [115] in Matlab instead of using its availabile plugin for MZmine [130].

As APC3 [174] was tested on the GC/MS TICs without correcting the baseline, it would be interesting to evaluate the performance of APC3 on different spectra that has been corrected via various baseline correction algorithms including AIMA [115], and also using the proposed baseline correction approach that involves testing several baseline correction techniques to achieve the best prediction [111, 112]. This would provide a combined performance evaluation of AIMA [115] and APC3 [174] on the same datasets.

The main limitation in the sparsity embedded approaches study is the use of a single three-class GC/MS TIC dataset due to the limited availability of three or more class GC/MS TIC datasets in the public domain. Although the biasness of the small number of sample data of the GC/MS TIC was reduced by splitting the dataset into 100 training and testing sets and deriving average accuracy and average AUC from the 100 sets, it would be useful to perform further testing on more TIC datasets with at least three classes. The evaluation of the sparsity embedded approaches with a two-class GC/MS TIC dataset was another attempt to delimitate the initial evaluation using a single three-class GC/MS TIC dataset. The two-class GC/MS TIC dataset was useful in providing consensus that the NSC algorithms were more superior to the others although they had similar performance to RDA algorithms for the three-class dataset. A possible extension would be to complement the use of baseline correction methods including AIMA [115] to evaluate if the performance truly improves with baseline correction.

.

References

1. Barton RH: A decade of advances in metabonomics. *Expert Opinion on Drug Metabolism & Toxicology* 2011, 7(2):129-136.
2. Lindon J, Holmes E, Nicholson J: Metabonomics Techniques and Applications to Pharmaceutical Research &amp; Development. *Pharmaceutical Research* 2006, 23(6):1075-1088.
3. Guo Y, Hastie T, Tibshirani R: Regularized linear discriminant analysis and its application in microarrays. *Biostatistics* 2007, 8(1):86-100.
4. Tibshirani R, Hastie T, Narasimhan B, Chu G: Class Prediction by Nearest Shrunken Centroids, with Applications to DNA Microarrays. *Statistical Science* 2003, 18(1):104-117.
5. Le Cao K-A, Boitard S, Besse P: Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics* 2011, 12(1):253.
6. Witten DM, Tibshirani R: Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2011, 73(5):753-772.
7. Pardo M, Sberveglieri G: Random forests and nearest shrunken centroids for the classification of sensor array data. *Sensors and Actuators B: Chemical* 2008, 131(1):93-99.
8. Horgan RP, Kenny LC: 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics. *The Obstetrician & Gynaecologist* 2011, 13(3):189-195.
9. Debnath M, Prasad GKS, Bisen P: Omics Technology. In: *Molecular Diagnostics: Promises and Possibilities.* Springer Netherlands; 2010: 11-31.
10. Bjerrum JT: Metabonomics: Analytical Techniques

and Associated Chemometrics at a Glance. In: *Metabonomics: Methods and Protocols.* Edited by Bjerrum JT, vol. 1277: Springer New York; 2015: 1-14.

11. Wenk MR: The emerging field of lipidomics. *Nat Rev Drug Discov* 2005, 4(7):594-610.
12. Uzman A: Molecular biology of the cell (4th ed.): Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. *Biochemistry and Molecular Biology Education* 2003, 31(4):212-214.
13. Baltimore D: Our genome unveiled. *Nature* 2001, 409(6822):814-816.
14. Pegram M, Slamon D: Biological rationale for HER2/neu (c-erbB2) as a target for monoclonal antibody therapy. *Semin Oncol* 2000, 27(5 Suppl 9):13-19.
15. Patti GJ, Yanes O, Siuzdak G: Innovation: Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol* 2012, 13(4):263-269.
16. Nicholson JK, Lindon JC: Systems biology: Metabonomics. *Nature* 2008, 455(7216):1054-1056.
17. Bomprezzi R, Kovanen PE, Martin R: New approaches to investigating heterogeneity in complex traits. *Journal of Medical Genetics* 2003, 40(8):553-559.
18. Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M: Global quantification of mammalian gene expression control. *Nature* 2011, 473(7347):337-342.
19. Vogel C, de Sousa Abreu R, Ko D, Le SY, Shapiro BA, Burns SC, Sandhu D, Boutz DR, Marcotte EM, Penalva LO: Sequence signatures and mRNA

concentration can explain two - thirds of protein abundance variation in a human cell line, vol. 6; 2010.

20. Lundberg E, Fagerberg L, Klevebring D, Matic I, Geiger T, Cox J, Älgenäs C, Lundeberg J, Mann M, Uhlen M: Defining the transcriptome and proteome in three functionally different human cell lines, vol. 6; 2010.

21. Tian Q, Stepaniants SB, Mao M, Weng L, Feetham MC, Doyle MJ, Yi EC, Dai H, Thorsson V, Eng J *et al*: Integrated Genomic and Proteomic Analyses of Gene Expression in Mammalian Cells. *Molecular & Cellular Proteomics* 2004, 3(10):960-969.

22. Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H *et al*: Mass-spectrometry-based draft of the human proteome. *Nature* 2014, 509(7502):582-587.

23. De Preter V: Metabonomics and Systems Biology. In: *Metabonomics: Methods and Protocols.* Edited by Bjerrum JT, vol. 1277: Springer New York; 2015: 245-253.

24. Jaenisch R, Bird A: Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet* 2003, 33:245 - 254.

25. Vlahou A, Fountoulakis M: Proteomic approaches in the search for disease biomarkers. *Journal of Chromatography B* 2005, 814(1):11-19.

26. Petricoin EF, Zoon KC, Kohn EC, Barrett JC, Liotta LA: Clinical proteomics: translating benchside promise into bedside reality. *Nat Rev Drug Discov* 2002, 1(9):683-695.

27. Theodorescu D, Mischak H: Mass spectrometry based proteomics in urine biomarker discovery. *World Journal of Urology* 2007, 25(5):435-443.

28. Lu C-M, Wu Y-J, Chen C-C, Hsu J-L, Chen J-C, Chen J, Huang C-H, Ko Y-C: Identification of low-abundance proteins via fractionation of the urine proteome with weak anion exchange chromatography. *Proteome Science* 2011, 9(1):17.

29. Granger J, Siddiqui J, Copeland S, Remick D: Albumin depletion of human plasma also removes low abundance proteins including the cytokines. *Proteomics* 2005, 5(18):4713-4718.

30. Merrell K, Southwick K, Graves SW, Esplin MS, Lewis NE, Thulin CD: Analysis of Low-Abundance, Low-Molecular-Weight Serum Proteins Using Mass Spectrometry. *Journal of Biomolecular Techniques : JBT* 2004, 15(4):238-248.

31. Lindon JC, Nicholson JK, Holmes E, Everett JR: Metabonomics: Metabolic processes studied by NMR spectroscopy of biofluids. *Concepts in Magnetic Resonance* 2000, 12(5):289-320.

32. Nicholson JK, Lindon JC, Holmes E: 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* 1999, 29(11):1181-1189.

33. Fiehn O: Metabolomics--the link between genotypes and phenotypes. *Plant Mol Biol* 2002, 48(1-2):155-171.

34. Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y, Djoumbou Y, Mandal R, Aziat F, Dong E *et al*: HMDB 3.0—The Human Metabolome Database in 2013. *Nucleic Acids Research* 2012.

35. Nicholson JK, Holmes E, Lindon JC: Chapter 1 - Metabonomics and Metabolomics Techniques and Their Applications in Mammalian Systems.

In: *The Handbook of Metabonomics and Metabolomics.* Edited by Lindon JC, Nicholson JK, Holmes E. Amsterdam: Elsevier Science B.V.; 2007: 1-33.

36. Vinayavekhin N, Homan EA, Saghatelian A: Exploring Disease through Metabolomics. *ACS Chemical Biology* 2009, 5(1):91-103.

37. Urbanczyk‐Wochniak E, Luedemann A, Kopka J, Selbig J, Roessner‐Tunali U, Willmitzer L, Fernie AR: Parallel analysis of transcript and metabolic profiles: a new approach in systems biology, vol. 4; 2003.

38. Chang Y, Lee GH, Kim TJ, Chae KS: Toxicity of magnetic resonance imaging agents: small molecule and nanoparticle. *Curr Top Med Chem* 2013, 13(4):434-445.

39. Lindon JC, Keun HC, Ebbels TMD, Pearce JMT, Holmes E, Nicholson JK: The Consortium for Metabonomic Toxicology (COMET): aims, activities and achievements. *Pharmacogenomics* 2005, 6(7):691-699.

40. Lenz EM, Bright J, Knight R, Westwood FR, Davies D, Major H, Wilson ID: Metabonomics with 1H-NMR spectroscopy and liquid chromatography-mass spectrometry applied to the investigation of metabolic changes caused by gentamicin-induced nephrotoxicity in the rat. *Biomarkers* 2005, 10(2-3):173-187.

41. Keun HC, Ebbels TMD, Bollard ME, Beckonert O, Antti H, Holmes E, Lindon JC, Nicholson JK: Geometric Trajectory Analysis of Metabolic Responses To Toxicity Can Define Treatment Specific Profiles. *Chemical Research in Toxicology* 2004, 17(5):579-587.

42. Lindon JC, Nicholson JK, Holmes E, Antti H, Bollard ME, Keun H, Beckonert O, Ebbels TM, Reily MD, Robertson D *et al*: Contemporary issues in toxicology the role of metabonomics in toxicology and its evaluation by the COMET project. *Toxicology and Applied Pharmacology* 2003, 187(3):137-146.

43. Ebbels T, Keun H, Beckonert O, Antti H, Bollard M, Holmes E, Lindon J, Nicholson J: Toxicity classification from metabonomic data using a density superposition approach: 'CLOUDS'. *Analytica Chimica Acta* 2003, 490(1–2):109-122.

44. Keun HC, Ebbels TMD, Antti H, Bollard ME, Beckonert O, Schlotterbeck G, Senn H, Niederhauser U, Holmes E, Lindon JC *et al*: Analytical Reproducibility in 1H NMR-Based Metabonomic Urinalysis. *Chemical Research in Toxicology* 2002, 15(11):1380-1386.

45. Nicholson JK, Wilson ID: High resolution proton magnetic resonance spectroscopy of biological fluids. *Progress in Nuclear Magnetic Resonance Spectroscopy* 1989, 21(4–5):449-501.

46. Bollard ME, Keun HC, Beckonert O, Ebbels TMD, Antti H, Nicholls AW, Shockcor JP, Cantor GH, Stevens G, Lindon JC *et al*: Comparative metabonomics of differential hydrazine toxicity in the rat and mouse. *Toxicology and Applied Pharmacology* 2005, 204(2):135-151.

47. Robertson DG, Reily MD, Cantor GH: Chapter 9 - Metabonomics in Preclinical Pharmaceutical Discovery and Development. In: *The Handbook of Metabonomics and Metabolomics.* Edited by Lindon JC, Nicholson JK, Holmes E. Amsterdam: Elsevier Science B.V.; 2007: 241-277.

48. Robertson DG, Reily MD, Sigler RE, Wells DF, Paterson DA, Braden TK: Metabonomics: Evaluation of Nuclear Magnetic Resonance (NMR) and Pattern Recognition Technology for Rapid in Vivo Screening of Liver and Kidney Toxicants. *Toxicological Sciences* 2000, 57(2):326-337.

49. Odunsi K, Wollman RM, Ambrosone CB, Hutson A, McCann SE, Tammela J, Geisler JP, Miller G, Sellers T, Cliby W *et al*: Detection of epithelial ovarian cancer using 1H-NMR-based metabonomics. *International Journal of Cancer* 2005, 113(5):782-788.

50. Coen M, O'Sullivan M, Bubb WA, Kuchel PW, Sorrell T: Proton Nuclear Magnetic Resonance—Based Metabonomics for Rapid Diagnosis of Meningitis and Ventriculitis. *Clinical Infectious Diseases* 2005, 41(11):1582-1590.

51. Swanson MG, Vigneron DB, Tabatabai ZL, Males RG, Schmitt L, Carroll PR, James JK, Hurd RE, Kurhanewicz J: Proton HR-MAS spectroscopy and quantitative pathologic analysis of MRI/3D-MRSI-targeted postsurgical prostate tissues. *Magnetic Resonance in Medicine* 2003, 50(5):944-954.

52. Moolenaar SH, Engelke UFH, Wevers RA: Proton nuclear magnetic resonance spectroscopy of body fluids in the field of inborn errors of metabolism. *Annals of Clinical Biochemistry* 2003, 40(1):16-24.

53. Brindle JT, Antti H, Holmes E, Tranter G, Nicholson JK, Bethell HWL, Clarke S, Schofield PM, McKilligin E, Mosedale DE *et al*: Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1H-NMR-based metabonomics. *Nat Med* 2002, 8(12):1439-1445.

54. Moka D, Vorreuther R, Schicha H, Spraul M, Humpfer E, Lipinski M, Foxall PJD, Nicholson JK, Lindon JC: Biochemical classification of kidney carcinoma biopsy samples using magic-angle-spinning 1H nuclear magnetic resonance spectroscopy. *Journal of Pharmaceutical and Biomedical Analysis* 1998, 17(1):125-132.

55. Cheng LL, Chang I-W, Louis DN, Gonzalez RG: Correlation of High-Resolution Magic Angle Spinning Proton Magnetic Resonance Spectroscopy with Histopathology of Intact Human Brain Tumor Specimens. *Cancer Research* 1998, 58(9):1825-1832.

56. Jones GLAH, Sang E, Goddard C, Mortishire-Smith RJ, Sweatman BC, Haselden JN, Davies K, Grace AA, Clarke K, Griffin JL: A Functional Analysis of Mouse Models of Cardiac Disease through Metabolic Profiling. *Journal of Biological Chemistry* 2005, 280(9):7530-7539.

57. Watkins SM, Reifsnyder PR, Pan H-j, German JB, Leiter EH: Lipid metabolome-wide effects of the PPARγ agonist rosiglitazone. *Journal of Lipid Research* 2002, 43(11):1809-1817.

58. Dunne VG, Bhattachayya S, Besser M, Rae C, Griffin JL: Metabolites from cerebrospinal fluid in aneurysmal subarachnoid haemorrhage correlate with vasospasm and clinical outcome: a pattern-recognition 1H NMR study. *NMR in Biomedicine* 2005, 18(1):24-33.

59. Sinclair TR, Purcell LC, Sneller CH: Crop transformation and the challenge to increase yield potential. *Trends in Plant Science* 2004, 9(2):70-75.

60. Morris CR, Scott JT, Chang H-m, Sederoff RR, O'Malley D, Kadla JF: Metabolic Profiling: A New Tool in the Study of Wood Formation. *Journal of Agricultural and Food Chemistry* 2004, 52(6):1427-1434.

61. Kirk H, Choi YH, Kim HK, Verpoorte R, Van Der Meijden E: Comparing metabolomes: the chemical consequences of hybridization in plants. *New Phytologist* 2005, 167(2):613-622.

62. Andrew Clayton T, Lindon JC, Cloarec O, Antti H, Charuel C, Hanton G, Provost J-P, Le Net J-L, Baker D, Walley RJ *et al*: Pharmaco-metabonomic

phenotyping and personalized drug treatment. *Nature* 2006, 440(7087):1073-1077.

63. Zhao Y-Y, Chen H, Tian T, Chen D-Q, Bai X, Wei F: A Pharmaco-Metabonomic Study on Chronic Kidney Disease and Therapeutic Effect of Ergone by UPLC-QTOF/HDMS. *PLoS One* 2014, 9(12):e115467.

64. Ye X, Fitzgerald EF, Gomez MI, Lambert GH, Longnecker MP: The ratio of specific polychlorinated biphenyls as a surrogate biomarker of cytochrome P4501A2 activity: a pharmaco-metabonomic study in humans. *Cancer Epidemiol Biomarkers Prev* 2008, 17(4):1013-1015.

65. Bollard ME, Stanley EG, Lindon JC, Nicholson JK, Holmes E: NMR-based metabonomic approaches for evaluating physiological influences on biofluid composition. *NMR in Biomedicine* 2005, 18(3):143-162.

66. Plumb RS, Granger JH, Stumpf CL, Johnson KA, Smith BW, Gaulitz S, Wilson ID, Castro-Perez J: A rapid screening approach to metabonomics using UPLC and oa-TOF mass spectrometry: application to age, gender and diurnal variation in normal/Zucker obese rats and black, white and nude mice. *Analyst* 2005, 130(6):844-849.

67. Coen M, Lenz EM, Nicholson JK, Wilson ID, Pognan F, Lindon JC: An Integrated Metabonomic Investigation of Acetaminophen Toxicity in the Mouse Using NMR Spectroscopy. *Chemical Research in Toxicology* 2003, 16(3):295-303.

68. Shockcor JP, Unger SE, Wilson ID, Foxall PJ, Nicholson JK, Lindon JC: Combined HPLC, NMR spectroscopy, and ion-trap mass spectrometry with application to the detection and characterization of xenobiotic and endogenous metabolites in human urine. *Anal Chem* 1996, 68(24):4431-4435.

69. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: The large-scale organization of metabolic networks. *Nature* 2000, 407(6804):651-654.

70. Lehtimäki KK, Valonen PK, Griffin JL, Väisänen TH, Gröhn OHJ, Kettunen MI, Vepsäläinen J, Ylä-Herttuala S, Nicholson J, Kauppinen RA: Metabolite Changes in BT4C Rat Gliomas Undergoing Ganciclovir-Thymidine Kinase Gene Therapy-induced Programmed Cell Death as Studied by 1H NMR Spectroscopy in Vivo, ex Vivo, and in Vitro. *Journal of Biological Chemistry* 2003, 278(46):45915-45923.

71. Grossmann K: What it takes to get a herbicide's mode of action. Physionomics, a classical approach in a new complexion. *Pest Management Science* 2005, 61(5):423-431.

72. Hubert S, Manfred L, Hansjörg F: Metabolic Profiling of Plants. In: *Synthesis and Chemistry of Agrochemicals II.* vol. 443: American Chemical Society; 1991: 288-299.

73. Ott K-H, Aranı, amp, x, bar N, Singh B, Stockton GW: Metabonomics classifies pathways affected by bioactive compounds. Artificial neural network classification of NMR spectra of plant extracts. *Phytochemistry* 2003, 62(6):971-985.

74. Lange BM, Ketchum REB, Croteau RB: Isoprenoid Biosynthesis. Metabolite Profiling of Peppermint Oil Gland Secretory Cells and Application to Herbicide Target Analysis. *Plant Physiology* 2001, 127(1):305-314.

75. Wang W, Vinocur B, Altman A: Plant responses to drought, salinity and extreme temperatures: towards genetic engineering for stress tolerance. *Planta* 2003, 218(1):1-14.

76. Lui LH, Vikram A, Abu-Nada Y, Kushalappa AC, Raghavan GSV, Al-Mughrabi K: Volatile metabolic profiling for discrimination of potato tubers inoculated with dry and soft rot pathogens. *Amer J of Potato Res* 2005, 82(1):1-8.

77. Lee D, Smallbone K, Dunn W, Murabito E, Winder C, Kell D, Mendes P, Swainston N: Improving metabolic flux predictions using absolute gene expression data. *BMC Systems Biology* 2012, 6(1):73.

78. Fu X, Fu N, Guo S, Yan Z, Xu Y, Hu H, Menzel C, Chen W, Li Y, Zeng R *et al*: Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics* 2009, 10(1):161.

79. Wang Z, Gerstein M, Snyder M: RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009, 10(1):57-63.

80. Van Loo P, Campbell PJ: ABSOLUTE cancer genomics. *Nat Biotech* 2012, 30(7):620-621.

81. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA *et al*: Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotech* 2012, 30(5):413-421.

82. Elliott MH, Smith DS, Parker CE, Borchers C: Current trends in quantitative proteomics. *Journal of Mass Spectrometry* 2009, 44(12):1637-1660.

83. Wasinger VC, Zeng M, Yau Y: Current Status and Advances in Quantitative Proteomic Mass Spectrometry. *International Journal of Proteomics* 2013, 2013:12.

84. Trudgian DC, Ridlova G, Fischer R, Mackeen MM, Ternette N, Acuto O, Kessler BM, Thomas B: Comparative evaluation of label-free SINQ normalized spectral index quantitation in the central proteomics facilities pipeline. *Proteomics* 2011, 11(14):2790-2797.

85. Koek MM, Jellema RH, van der Greef J, Tas AC, Hankemeier T: Quantitative metabolomics based on gas chromatography mass spectrometry: status and perspectives. *Metabolomics* 2011, 7(3):307-328.

86. Dunn WB, Bailey NJC, Johnson HE: Measuring the metabolome: current analytical technologies. *Analyst* 2005, 130(5):606-625.

87. Roberts LD, Souza AL, Gerszten RE, Clish CB: Targeted Metabolomics. *Current Protocols in Molecular Biology* 2012, CHAPTER:Unit30.32-Unit30.32.

88. Vettukattil R: Preprocessing of Raw Metabonomic Data. In: *Metabonomics: Methods and Protocols.* Edited by Bjerrum JT, vol. 1277: Springer New York; 2015: 123-136.

89. Amigo JM, Skov T, Bro R: ChroMATHography: Solving Chromatographic Issues with Mathematical Models and Intuitive Graphics. *Chemical Reviews* 2010, 110(8):4582-4605.

90. Ortiz MC, Sarabia L: Quantitative determination in chromatographic analysis based on n-way calibration strategies. *Journal of Chromatography A* 2007, 1158(1-2):94-110.

91. Jirasek A, Schulze G, Yu MML, Blades MW, Turner RFB: Accuracy and Precision of Manual Baseline Determination. *Appl Spectrosc* 2004, 58(12):1488-1499.

92. Z. M.Zhang SC, Y. Z. Liang, Z. X. Liu, Q. M.Zhang, L. X. Ding,, Zhou FYaH: An intelligent background-correction algorithm for highly fluorescent samples in Raman spectroscopy. *J Raman Spectrosc* 2009.

93. Zhao J, Lui H, McLean DI, Zeng H: Automated Autofluorescence Background Subtraction Algorithm for Biomedical Raman Spectroscopy. *Appl Spectrosc* 2007, 61(11):1225-1232.

94. Caligiani A, Acquotti D, Palla G, Bocchi V: Identification and quantification of the main organic components of vinegars by high resolution1H NMR spectroscopy. *Analytica Chimica Acta* 2007, 585(1):110-119.

95. Brown DE: Fully Automated Baseline Correction of 1D and 2D NMR Spectra Using Bernstein Polynomials. *Journal of Magnetic Resonance, Series A* 1995, 114(2):268-270.

96. Zhang ZM, Chen S, Liang YZ: Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst* 2010, 135(5):1138-1146.

97. Hu Y, Jiang T, Shen A, Li W, Wang X, Hu J: A background elimination method based on wavelet transform for Raman spectra. *Chemometrics and Intelligent Laboratory Systems* 2007, 85(1):94-101.

98. Fonville JM, Richards SE, Barton RH, Boulange CL, Ebbels TMD, Nicholson JK, Holmes E, Dumas M-E: The evolution of partial least squares models and related chemometric approaches in metabonomics and metabolic phenotyping. *Journal of Chemometrics* 2010, 24(11-12):636-649.

99. Chapter 10 Multiple and polynomial regression. In: *Data Handling in Science and Technology.* Edited by D.L. Massart BGMVLMCBSDJPJL, Smeyers-Verbeke J, vol. Volume 20, Part A: Elsevier; 1998: 263-303.

100. McDonald R: Path Analysis with Composite Variables. *Multivariate Behavioral Research* 1996, 31(2):239-270.

101. Lohmller JB: Latent variable path modeling with partial least squares: Physica-Verlag Heidelberg; 1989.

102. Jiang W, Qiu Y, Ni Y, Su M, Jia W, Du X: An Automated Data Analysis Pipeline for GC−TOF−MS Metabonomics Studies. *Journal of Proteome Research* 2010, 9(11):5974-5981.

103. Stone CJ: Optimal Global Rates of Convergence for Nonparametric Regression. 1982(4):1040-1053.

104. Beyer KS, Goldstein J, Ramakrishnan R, Shaft U: When Is ''Nearest Neighbor'' Meaningful? In: *Proceedings of the 7th International Conference on Database Theory.* Springer-Verlag; 1999: 217-235.

105. Elad M: Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing: Springer Publishing Company, Incorporated; 2010.

106. Yu CD, Huang J, Austin W, Xiao B, Biros G: Performance optimization for the k-nearest neighbors kernel on x86 architectures. In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis.* Austin, Texas: ACM; 2015: 1-12.

107. Gonnet GH, Scholl R: Scientific Computation: Cambridge University Press; 2009.

108. Bermúdez Chacón R: Automatic problem-specific hyperparameter optimization and model selection for supervised machine learning. Zurich: ETH Zurich; 2014.

109. Berman JJ: Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information: Elsevier Science; 2013.

110. Wang T, Shao K, Chu Q, Ren Y, Mu Y, Qu L, He J, Jin C, Xia B: Automics: an integrated platform for NMR-based metabonomics spectral processing and data analysis. *BMC Bioinformatics* 2009, 10:83.

111. Krier F, Mantanus J, Sacré P-Y, Chavez P-F, Thiry J, Pestieau A, Rozet E, Ziemons E, Hubert P, Evrard B: PAT tools for the control of co-extrusion implants manufacturing process. *International Journal of Pharmaceutics* 2013, 458(1):15-24.

112. Liland KH, Almøy T, Mevik B-H: Optimal choice of baseline correction for multivariate calibration of spectra. *Applied spectroscopy* 2010, 64(9):1007-1016.

113. Kettenring JR: Coping with high dimensionality in massive datasets. *Wiley Interdisciplinary Reviews: Computational Statistics* 2011, 3(2):95-103.

114. Katajamaa M, Miettinen J, Oresic M: MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* 2006, 22(5):634-636.

115. Prakash BD, Wei YC: A fully automated iterative moving averaging (AIMA) technique for baseline correction. *Analyst* 2011, 136(15):3130-3135.

116. Prakash BD, Esuvaranathan K, Ho PC, Pasikanti KK, Yong Chan EC, Yap CW: An automated Pearson's correlation change classification (APC3) approach for GC/MS metabonomic data using total ion chromatograms (TICs). *Analyst* 2013.

117. Brereton RG: CHEMOMETRICS AND STATISTICS | Spectral Deconvolution and Filtering. In: *Encyclopedia of Analytical Science.* Edited by Paul W, Alan T, Colin P. Oxford: Elsevier; 2005: 51-60.

118. Mantini D, Petrucci F, Pieragostino D, Del Boccio P, Di Nicola M, Di Ilio C, Federici G, Sacchetta P, Comani S, Urbani A: LIMPIC: a computational method for the separation of protein MALDI-TOF-MS signals from noise. *BMC Bioinformatics* 2007, 8(1):101.

119. P. H. C. Eilers HFMB: Baseline Correction with Asymmetric Least Squares Smoothing. *http://wwwscienceuvanl/~hboelens/publications/draftpub/Eilers_2005pdf* 2005.

120. Eilers PHC: Parametric Time Warping. *Analytical Chemistry* 2003, 76(2):404-411.

121. Xi Y, Rocke D: Baseline Correction for NMR Spectroscopic Metabolomics Data Analysis. *BMC Bioinformatics* 2008, 9(1):324.

122. Carlos Cobas J, Bernstein MA, Martín-Pastor M, Tahoces PG: A new general-purpose fully automatic baseline-correction procedure for 1D and 2D NMR data. *Journal of Magnetic Resonance* 2006, 183(1):145-151.

123. Farashi SA, M.D.; Salimpour, Y.; Alirezaie, J.: Combination of PCA and undecimated wavelet transform for neural data processing. In: *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE 2010*; 2010: 6666 - 6669.

124. Mevik BH, Cederkvist HR: {Mean squared error of prediction(MSEP) estimates for principal component regression(PCR) and partial least squares regression(PLSR)}. *Journal of Chemometrics* 2004, 18(9):422-429.

125. Mason SJ, Graham NE: Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical signififcance and interpretation. *Q J R Meteorol Soc* 2002, 128:2145-2166.

126. Saghatelian A, Trauger SA, Want EJ, Hawkins EG, Siuzdak G, Cravatt BF: Assignment of endogenous substrates to enzymes by global metabolite profiling. *Biochemistry* 2004, 43(45):14332-14339.

127. Hall M: Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. In: *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning: 2000*: Morgan Kaufmann Publishers Inc.; 2000: 359-366.

128. Pasikanti KK, Esuvaranathan K, Ho PC, Mahendran R, Kamaraj R, Wu QH, Chiong E, Chan ECY: Noninvasive Urinary Metabonomic Diagnosis of Human Bladder Cancer. *Journal of Proteome Research* 2010, 9(6):2988-2995.

129. Skov T, Ballabio D, Bro R: Multiblock variance partitioning: a new approach for comparing variation in multiple data blocks. *Anal Chim Acta* 2008, 615(1):18-29.

130. Pluskal T, Castillo S, Villar-Briones A, Oresic M: MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* 2010, 11(1):395.

131. Pearson K: On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 1901, 2(6):559-572.

132. Benzécri JP: L'Analyse des données: L'Analyse des correspondances: Dunod; 1976.

133. Hill MO, Gauch HG: Detrended Correspondence Analysis: an improved ordination technique. *Vegetatio* 1980, 42:47-58.

134. Jari Oksanen FGB, Roeland Kindt, Pierre, Legendre PRM, R. B. O’Hara, Gavin L. Simpson,Peter Solymos, M. Henry H. Stevens, Helene Wagner: Ordination methods, diversity analysis and other functions

for community and vegetation ecologists. [online] http://cran.r-project.org/web/packages/vegan/vegan.pdf. 2011.

135. Romanski P: Selecting attributes, [online] http://cran.r-project.org/web/packages/FSelector/FSelector.pdf. 2009.

136. Serrano AJ, Soria E, Marti, x, n JD, Magdalena R, Go, mez J: Feature selection using ROC curves on classification problems. In: *Neural Networks (IJCNN), The 2010 International Joint Conference on: 18-23 July 2010 2010*; 2010: 1-6.

137. Ahdesmäki M, Strimmer K: Feature selection in omics prediction problems using cat scores and false non-discovery rate control; 2009.

138. Miika Ahdesmaki VZ, and Korbinian Strimmer: Shrinkage Discriminant Analysis and CAT Score Variable Selection [online] http://cran.r-project.org/web/packages/sda/sda.pdf. 2011.

139. Kira K, Rendell L: A Practical Approach to Feature Selection. In: *ML '92: Proceedings of the Ninth International Workshop on Machine Learning: 1992*: Morgan Kaufmann Publishers Inc.; 1992: 249-256.

140. Kononenko I: On biases in estimating multi-valued attributes. In: *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2.* Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc.; 1995: 1034-1040.

141. Dietterich T, Kearns M, Mansour Y: Applying the weak learning framework to understand and improve {C}4.5. In: *Proc 13th International Conference on Machine Learning: 1996*: Morgan Kaufmann; 1996: 96-104.

142. CORElearn - classification, regression, feature evaluation and ordinal evaluation [http://cran.r-project.org/web/packages/CORElearn/CORElearn.pdf]

143. Christian Roever NR, Karsten Luebke, Uwe Ligges, Gero Szepannek, Marc Zentgra: Classification and visualization [online] http://cran.r-project.org/web/packages/klaR/klaR.pdf. 2011.

144. Wehrens R, Mevi B-H: [online] http://cran.r-project.org/web/packages/pls/pls.pdf. 2007.

145. Tutz G, Binder H: Localized classification. *Statistics and Computing* 2005, 15(3):155-166.

146. Dayal B, Macgregor J: Improved PLS algorithms. *Journal of Chemometrics* 1997, 11(1):73-85.

147. Andersson M: A comparison of nine PLS1 algorithms. *Journal of Chemometrics* 2009, 23(10):518-529.

148. Szymańska E, Saccenti E, Smilde A, Westerhuis J: Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics* 2012, 8(1):3-16.

149. R Development Core Team: R: A language and environment for statistical computing.R Foundation for Statistical Computing, Vienna, Austria., 2011, . In.; 2011.

150. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH: The WEKA data mining software: an update. *SIGKDD Explor Newsl* 2009, 11(1):10--18.

151. KENDALL MG: A NEW MEASURE OF RANK CORRELATION. *Biometrika* 1938, 30(1-2):81-93.

152. Shannon CE: A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review* 2001, 5(1):3-55.

153. Rantalainen M, Holmes CC: Accounting for control mislabelling in case-control biomarker studies. *Journal of Proteome Research* 2011, 10(12):5562-5567.

154. Szatmari P, Jones MB: Effects of misclassification on estimates of relative risk in family history studies. *Genet Epidemiol* 1999, 16(4):368-381.

155. Yasui Y, Pepe M, Hsu L, Adam BL, Feng Z: Partially supervised learning using an EM-boosting algorithm. *Biometrics* 2004, 60(1):199-206.

156. Chen C, Zhang Z-M, Ouyang M-L, Liu X, Yi L, Liang Y-Z, Zhang C-P: Shrunken centroids regularized discriminant analysis as a promising strategy for metabolomics data exploration. *Journal of Chemometrics* 2014:n/a-n/a.

157. Le Cao K, Martin P, Robert-Granie C, Besse P: ofw: Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics* 2009, 10(34).

158. Le Cao K, Rossouw D, Robert-Granie C, Besse P: Sparse PLS: Variable Selection when Integrating Omics data. *Statistical Application and Molecular Biology* 2008, 7(1):37.

159. Shen H, Huang J: Sparse Principal Component Analysis via Regularized Low Rank Matrix Approximation. *Journal of Multivariate Analysis* 2008, 99:1015 - 1034.

160. Witten IH, Frank E: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations: Morgan Kaufmann; 2000.

161. Miyamoto S, Taylor S, Barupal D, Taguchi A, Wohlgemuth G, Wikoff W, Yoneda K, Gandara D, Hanash S, Kim K *et al*: Systemic Metabolomic Changes in Blood Samples of Lung Cancer Patients Identified by Gas Chromatography Time-of-Flight Mass Spectrometry. *Metabolites* 2015, 5(2):192-210.

162. Lau SKP, Lam C-W, Curreem SOT, Lee K-C, Lau CCY, Chow W-N, Ngan AHY, To KKW, Chan JFW, Hung IFN *et al*: Identification of specific metabolites in culture supernatant of Mycobacterium tuberculosis using metabolomics: exploration of potential biomarkers. *Emerg Microbes Infect* 2015, 4:e6.

163. Xiao Y-p, Wu T-x, Hong Q-h, Sun J-m, Chen A-g, Yang C-m, Li X-y: Response to weaning and dietary L-glutamine supplementation: metabolomic analysis in piglets by gas chromatography/mass spectrometry. *Journal of Zhejiang University Science B* 2012, 13(7):567-578.

164. Wang X, Yan S-K, Dai W-X, Liu X-R, Zhang W-D, Wang J-J: A metabonomic approach to chemosensitivity prediction of cisplatin plus 5-fluorouracil in a human xenograft model of gastric cancer. *International Journal of Cancer* 2010, 127(12):2841-2850.

165. Sieber M, Hoffmann D, Adler M, Vaidya VS, Clement M, Bonventre JV, Zidek N, Rached E, Amberg A, Callanan JJ *et al*: Comparative Analysis of Novel Noninvasive Renal Biomarkers and Metabonomic Changes in a Rat Model of Gentamicin Nephrotoxicity. *Toxicological Sciences* 2009, 109(2):336-349.

166. Hastie T, Tibshirani R, Friedman J: The Elements of Statistical Learning. New York, NY, USA: Springer New York Inc.; 2001.

167. Kohavi R: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2.* Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc.; 1995: 1137-1143.

168. Guan W, Zhou M, Hampton CY, Benigno BB, Walker LD, Gray A, McDonald JF, Fernández FM: Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines. *BMC Bioinformatics* 2009, 10:259-259.

169. Meyer MA, Booker JM: Eliciting and Analyzing Expert Judgment: A Practical Guide: Society for Industrial and Applied Mathematics; 2001.

170. Liquet B, Cao K-AL, Hocini H, Thiebaut R: A novel approach for biomarker selection and the integration of repeated measures experiments from two assays. *BMC Bioinformatics* 2012, 13(1):325.

171. Nakas CT: DEVELOPMENTS IN ROC SURFACE ANALYSIS AND ASSESSMENT OF DIAGNOSTIC MARKERS IN THREE-CLASS CLASSIFICATION PROBLEMS. *REVSTAT–Statistical Journal* 2014, 12(1):43-65.

172. Franceschi P, Mattivi F, Wehrens R, Vrhovsek U: Metabolic Biomarker Identification with Few Samples: INTECH Open Access Publisher; 2012.

173. Brown CD, Davis HT: Receiver operating characteristics curves and related decision measures: A tutorial. *Chemometrics and Intelligent Laboratory Systems* 2006, 80(1):24-38.

174. Prakash BD, Esuvaranathan K, Ho PC, Pasikanti KK, Chan EC, Yap CW: An automated Pearson's correlation change classification (APC3) approach for GC/MS metabonomic data using total ion chromatograms (TICs). *Analyst* 2013, 138(10):2883-2889.

175. Chan ECY, Pasikanti KK, Nicholson JK: Global urinary metabolic profiling procedures using gas chromatography-mass spectrometry. *Nat Protocols* 2011, 6(10):1483-1499.

176. Coelho FR, Martins JO: Diagnostic methods in sepsis: the need of speed. *Revista da Associação Médica Brasileira* 2012, 58:498-504.

177. Ramirez T, Daneshian M, Kamp H, Bois FY, Clench MR, Coen M, Donley B, Fischer SM, Ekman DR, Fabian E *et al*: Metabolomics in Toxicology and Preclinical Research. *ALTEX* 2013, 30(2):209-225.

178. Hsu C-C, ElNaggar MS, Peng Y, Fang J, Sanchez LM, Mascuch SJ, Møller KA, Alazzeh EK, Pikula J, Quinn RA *et al*: Real-Time Metabolomics on Living Microorganisms Using Ambient Electrospray Ionization Flow-Probe. *Analytical Chemistry* 2013, 85(15):7014-7018.

Appendix

This section contains the tables and figures taken from the supplementary materials of both AIMA [115] and APC3 [174].

Table I. Mean, standard deviation (SD) and p-values in the accuracy for the various PLSR dimension reduction combinations for latent variable optimization using 10-fold cross validation(CV) versus double cross validation (CV) for (a)wine set A, (b)wine set B and (c) wine set C. Significant higher mean and SD are highlighted in bold italics.

(a)

Dimension Reduction	Average accuracy (SD) using 10-fold CV	Average accuracy (SD) using double CV	P-value
LC	***0.88(0.11)***	0.08(0.24)	1.53E-48
colAUC	***0.85(0.14)***	0.072(0.22)	3.98E-49
DDA	***0.87(0.11)***	0.076(0.23)	2.52E-50
LDA	***0.84(0.15)***	0.077(0.24)	1.60E-46
ReliefFexpRank	***0.81(0.13)***	0.078(0.24)	1.22E-47
ReliefFequalK	***0.81(0.13)***	0.078(0.24)	1.63E-47
ReliefFbestK	***0.69(0.14)***	0.063(0.19)	2.88E-47
Relief	***0.7(0.14)***	0.062(0.19)	1.36E-47
MDL	***0.84(0.12)***	0.072(0.22)	6.66E-49
Gini	***0.84(0.12)***	0.072(0.22)	6.66E-49
MyopicReliefF	***0.84(0.12)***	0.072(0.22)	6.66E-49
DKM	***0.84(0.12)***	0.072(0.22)	6.66E-49
NCA	***0.93(0.066)***	0.091(0.27)	3.91E-54
PCA1	***0.64(0.12)***	0.049(0.15)	4.05E-56
PCA2	***0.72(0.12)***	0.03(0.098)	4.45E-67
CCA1	***0.36(0)***	0.036(0.11)	2.66E-51
CCA2	***0.36(0)***	0.036(0.11)	2.66E-51
DCA	***0.88(0.066)***	0.085(0.26)	2.95E-55

(b)

Dimension Reduction	Average accuracy (SD) using 10-fold CV	Average accuracy (SD) using double CV	P-value
LC	0.79(0.16)	0.79(0.15)	0.441
colAUC	0.73(0.17)	0.72(0.17)	0.326
DDA	0.78(0.17)	0.79(0.16)	0.456
LDA	0.78(0.16)	0.77(0.16)	0.429
ReliefFexpRank	0.67(0.16)	0.66(0.17)	0.565
ReliefFequalK	0.67(0.16)	0.67(0.17)	0.828
ReliefFbestK	0.59(0.14)	0.62(0.16)	0.138
Relief	0.6(0.14)	0.62(0.15)	0.156
MDL	0.73(0.18)	0.72(0.17)	0.374
Gini	0.73(0.18)	0.72(0.17)	0.374
MyopicReliefF	0.74(0.17)	0.74(0.16)	0.779
DKM	0.73(0.18)	0.72(0.17)	0.374
NCA	***0.55(0.15)***	0.048(0.15)	1.84E-42
PCA1	***0.31(0.12)***	0.06(0.18)	6.21E-21
PCA2	***0.25(0.099)***	0.022(0.075)	7.53E-32
CCA1	***0.52(0.041)***	0.045(0.14)	7.80E-58
CCA2	***0.52(0.041)***	0.045(0.14)	7.80E-58
DCA	***0.53(0.13)***	0.06(0.19)	8.12E-37

(c)

Dimension Reduction	Average accuracy (SD) using 10-fold CV	Average accuracy (SD) using double CV	P-value
LC	***1(0.026)***	0.97(0.047)	3.14E-05
colAUC	***0.93(0.095)***	0.9(0.1)	0.0011
DDA	***1(0.026)***	0.97(0.047)	3.14E-05
LDA	***1(0.022)***	0.97(0.051)	3.74E-07
ReliefFexpRank	***0.95(0.1)***	0.93(0.11)	0.0436
ReliefFequalK	0.95(0.1)	0.93(0.11)	0.102
ReliefFbestK	***0.82(0.2)***	0.72(0.21)	8.14E-05
Relief	***0.79(0.21)***	0.71(0.21)	0.000207
MDL	***0.93(0.095)***	0.9(0.1)	0.0011
Gini	***0.93(0.095)***	0.9(0.1)	0.0011
MyopicReliefF	***0.93(0.095)***	0.9(0.1)	0.0011
DKM	***0.93(0.095)***	0.9(0.1)	0.0011
NCA	***0.94(0.073)***	0.094(0.28)	1.29E-46
PCA1	***0.4(0.14)***	0.031(0.096)	3.25E-39
PCA2	***0.68(0.15)***	0.058(0.18)	1.83E-50
CCA1	***0.3(0)***	0.03(0.09)	2.66E-51
CCA2	***0.3(0)***	0.03(0.09)	2.66E-51
DCA	***0.95(0.059)***	0.094(0.28)	1.53E-50

Table II. Mean, standard deviation (SD) and p-values in the AUC for the various PLSR dimension reduction combinations for latent variable optimization using 10-fold cross validation(CV) versus double cross validation (CV) for (a)wine set A, (b)wine set B and (c) wine set C. Significant higher mean and SD are highlighted in bold italics.

(a)

Dimension Reduction	Average AUC (SD) using 10-fold CV	Average AUC (SD) using double CV	P-value
LC	***0.86(0.11)***	0.081(0.25)	1.18E-47
colAUC	***0.84(0.14)***	0.071(0.22)	6.66E-49
DDA	***0.85(0.11)***	0.078(0.24)	4.18E-49
LDA	***0.83(0.15)***	0.077(0.24)	3.81E-46
ReliefFexpRank	***0.81(0.13)***	0.079(0.24)	1.73E-47
ReliefFequalK	***0.8(0.13)***	0.079(0.24)	2.16E-47
ReliefFbestK	***0.68(0.14)***	0.063(0.2)	4.40E-46
Relief	***0.69(0.14)***	0.063(0.19)	8.80E-47
MDL	***0.82(0.13)***	0.071(0.22)	1.48E-47
Gini	***0.82(0.13)***	0.071(0.22)	1.48E-47
MyopicReliefF	***0.82(0.13)***	0.071(0.22)	1.48E-47
DKM	***0.82(0.13)***	0.071(0.22)	1.48E-47
NCA	***0.91(0.083)***	0.089(0.27)	1.83E-54
PCA1	***0.62(0.13)***	0.039(0.12)	8.82E-60
PCA2	***0.71(0.12)***	0.033(0.11)	8.19E-67
CCA1	***0.5(0)***	0.05(0.15)	2.66E-51
CCA2	***0.5(0)***	0.05(0.15)	2.66E-51
DCA	***0.85(0.082)***	0.084(0.25)	8.86E-54

(b)

Dimension Reduction	Average AUC (SD) using 10-fold CV	Average AUC (SD) using double CV	P-value
LC	0.78(0.15)	0.79(0.14)	0.344
colAUC	0.72(0.16)	0.71(0.16)	0.255
DDA	0.78(0.16)	0.79(0.15)	0.457
LDA	0.77(0.15)	0.77(0.15)	0.434
ReliefFexpRank	0.68(0.16)	0.66(0.17)	0.373
ReliefFequalK	0.67(0.16)	0.66(0.18)	0.611
ReliefFbestK	0.6(0.14)	0.62(0.16)	0.197
Relief	0.6(0.14)	0.62(0.15)	0.223
MDL	0.72(0.17)	0.71(0.17)	0.348
Gini	0.72(0.17)	0.71(0.17)	0.348
MyopicReliefF	0.74(0.17)	0.73(0.16)	0.677
DKM	0.72(0.17)	0.71(0.17)	0.348
NCA	***0.55(0.15)***	0.047(0.15)	3.57E-44
PCA1	***0.3(0.12)***	0.057(0.17)	8.58E-22
PCA2	***0.25(0.098)***	0.021(0.071)	7.58E-32
CCA1	***0.5(0)***	0.05(0.15)	2.66E-51
CCA2	***0.5(0)***	0.05(0.15)	2.66E-51
DCA	***0.54(0.13)***	0.06(0.18)	1.46E-37

(c)

Dimension Reduction	Average AUC (SD) using 10-fold CV	Average AUC (SD) using double CV	P-value
LC	***1(0.019)***	0.96(0.077)	1.30E-06
colAUC	***0.92(0.11)***	0.89(0.13)	0.000188
DDA	***1(0.019)***	0.96(0.077)	1.30E-06
LDA	***1(0.016)***	0.95(0.08)	9.45E-08
ReliefFexpRank	***0.94(0.12)***	0.91(0.14)	0.00825
ReliefFequalK	***0.94(0.12)***	0.91(0.14)	0.0213
ReliefFbestK	***0.8(0.22)***	0.68(0.21)	1.03E-06
Relief	***0.78(0.21)***	0.68(0.21)	1.11E-06
MDL	***0.92(0.11)***	0.89(0.13)	0.000188
Gini	***0.92(0.11)***	0.89(0.13)	0.000188
MyopicReliefF	***0.92(0.11)***	0.89(0.13)	0.000188
DKM	***0.92(0.11)***	0.89(0.13)	0.000188
NCA	***0.9(0.11)***	0.09(0.27)	4.31E-44
PCA1	***0.41(0.15)***	0.024(0.076)	3.56E-39
PCA2	***0.71(0.14)***	0.062(0.19)	9.89E-49
CCA1	***0.5(0)***	0.05(0.15)	2.66E-51
CCA2	***0.5(0)***	0.05(0.15)	2.66E-51
DCA	***0.93(0.077)***	0.09(0.27)	4.51E-50

Table III. Overall average accuracy and average AUC for each transformation/variable and classification combination using the wine testing sets. Values are expressed in (average accuracy ± standard deviation, average AUC ± standard deviation). Top 2 approaches values for average accuracy and average AUC highlighted in bold italics.

		Transformation					
		NCA	**PCA$_1$**	**PCA$_2$**	**CCA$_1$**	**CCA$_2$**	**DCA**
Classification	**PLSR**	0.81±0.22 0.79±0.10	0.45±0.17 0.44±0.16	0.55±0.26 0.56±0.27	0.39±0.11 0.50±0.00	0.39±0.11 0.50±0.00	0.78±0.23 0.77±0.20
	NB	0.69±0.08 0.71±0.06	0.62±0.16 0.60±0.14	0.59±0.09 0.59±0.13	0.68±0.14 0.67±0.13	0.68±0.14 0.67±0.13	0.78±0.16 0.78±0.15
	locLDA	0.57±0.10 0.57±0.10	0.47±0.07 0.47±0.05	0.51±0.08 0.51±0.08	0.47±0.05 0.51±0.01	0.46±0.05 0.49±0.01	0.77±0.20 0.77±0.19

		Variable Selection					
		LC	**colAUC**	**DDA**	**LDA**	**ReliefFexpRank**	**ReliefFequalK**
Classification	**PLSR**	0.89±0.10	0.84±0.10	0.88±0.11	0.87±0.11	0.81±0.14	0.81±0.14
		0.88±0.11	0.83±0.10	0.88±0.11	0.87±0.12	0.81±0.13	0.80±0.13
	NB	***0.89±0.10****	0.84±0.09	***0.90±0.09***	0.88±0.10	0.81±0.12	0.81±0.12
		0.89±0.10	0.83±0.09	***0.89±0.10***	0.88±0.10	0.80±0.11	0.80±0.11
	locLDA	0.88±0.10	0.84±0.10	0.88±0.10	0.87±0.11	0.81±0.14	0.81±0.14
		0.88±0.10	0.83±0.09	0.89±0.10	0.88±0.10	0.80±0.11	0.80±0.11

		Variable Selection					
		ReliefFbestK	Relief	MDL	Gini	MyopicReliefF	DKM
Classification	PLSR	0.70±0.11	0.70±0.10	0.83±0.10	0.83±0.10	0.84±0.09	0.83±0.10
		0.69±0.10	0.69±0.09	0.82±0.10	0.82±0.10	0.83±0.09	0.82±0.10
	NB	0.76±0.07	0.75±0.06	0.85±0.09	0.85±0.09	0.85±0.08	0.85±0.09
		0.75±0.06	0.75±0.05	0.84±0.09	0.85±0.09	0.85±0.09	0.85±0.09
	locLDA	0.70±0.10	0.69±0.09	0.83±0.11	0.83±0.11	0.84±0.09	0.83±0.11
		0.75±0.06	0.75±0.05	0.85±0.09	0.85±0.09	0.85±0.09	0.85±0.09

*LC-NB had an average accuracy of 0.892 which was higher than that of LC-PLSR's 0.886

Table IV. Variable length after transformation

		Wine Set A	Wine Set B	Wine Set C
Transformation	**NCA**	2700		
	***PCA$_1$**	26	22	25
	***PCA$_2$**			
	***CCA$_1$**			
	***CCA$_2$**			
	****DCA**	4		

* (Number of samples) – 1

** Only finds 4 axes according to the original implementation in M. O. Hill and H. G. Gauch, Vegetatio, 1980, 42, 47-58.

Table V. Mean number of latent variable (SD) after transformation and PLSR latent variable optimization

		Wine Set A	Wine Set B	Wine Set C
Transformation	**NCA**	2.75(0.60)	3.01(0.89)	2.63(0.56)
	PCA$_1$	1(0)	1(0)	1(0)
	PCA$_2$	1(0)	1(0)	1(0)
	CCA$_1$	1(0)	1(0)	1(0)
	CCA$_2$	1(0)	1(0)	1(0)
	DCA	1.56(0.77)	1.76(0.77)	1.14(0.495)

Table VI. Mean number of latent variable (SD) for training model using PLSR after variable selection

		Wine Set A	Wine Set B	Wine Set C
Variable Selection	**LC**	4.63(2.63)	2.60(1.38)	2(0)
	colAUC	4.05(2.18)	3.11(1.63)	2.77(0.77)
	DDA	4.73(2.61)	2.61(1.46)	2(0)
	LDA	4.05(1.89)	2.74(1.54)	2.01(0.10)
	ReliefFexpRank	6.12(2.88)	5.38(2.13)	2.57(0.96)
	ReliefFequalK	6.16(2.91)	5.41(2.11)	2.56(0.96)
	ReliefFbestK	7.00(2.70)	6.07(2.43)	5.71(2.48)
	Relief	7.10(2.76)	6.04(2.39)	5.75(2.55)
	MDL	5.46(2.60)	3.57(2.34)	2.77(0.77)
	Gini	5.46(2.60)	3.56(2.31)	2.77(0.77)
	MyopicReliefF	5.46(2.60)	3.21(1.73)	2.77(0.77)
	DKM	5.46(2.60)	3.56(2.31)	2.77(0.77)

Table VII. Mean variable length (SD) for training model using NB after variable selection

		Wine Set A	Wine Set B	Wine Set C
Variable Selection	**LC**	10.60(9.75)	3.13(3.50)	2(0)
	colAUC	6.97(7.22)	5.05(8.16)	2.84(1.08)
	DDA	11.70(10.60)	3.15(3.44)	2(0)
	LDA	6.56(5.32)	4.42(5.58)	2.01(0.10)
	ReliefFexpRank	9.67(9.34)	6.67(7.51)	2.98(1.90)
	ReliefFequalK	9.83(10.20)	6.58(7.34)	2.92(1.86)
	ReliefFbestK	15.90(15.10)	8.58(11.20)	11.70(10.90)
	Relief	16.30(15.60)	8.16(10.80)	11.60(11.30)
	MDL	14.10(11.30)	5.48(8.93)	2.84(1.08)
	Gini	14.10(11.30)	5.48(8.93)	2.84(1.08)
	MyopicReliefF	13.80(11.00)	4.29(7.47)	2.84(1.08)
	DKM	14.10(11.30)	5.48(8.93)	2.84(1.08)

Table VIII. Mean variable length (SD) for training model using locLDA after variable selection

		Wine Set A	**Wine Set B**	**Wine Set C**
Variable Selection	**LC**	3.66(1.69)	2.46(1.18)	2(0)
	colAUC	3.87(2.06)	2.80(1.29)	2.92(0.99)
	DDA	3.66(1.69)	2.49(1.29)	2(0)
	LDA	3.49(1.40)	2.59(1.40)	2.01(0.10)
	ReliefFexpRank	5.06(2.42)	4.48(1.77)	2.27(0.53)
	ReliefFequalK	5.05(2.49)	4.45(1.79)	2.25(0.52)
	ReliefFbestK	6.08(2.51)	5.16(2.25)	5.18(2.34)
	Relief	6.16(2.57)	5.16(2.20)	5.36(2.44)
	MDL	4.75(2.10)	3.13(1.74)	2.92(0.99)
	Gini	4.75(2.10)	3.13(1.74)	2.92(0.99)
	MyopicReliefF	4.75(2.10)	2.83(1.32)	2.92(0.99)
	DKM	4.75(2.10)	3.13(1.74)	2.92(0.99)

Table IX. Mean variable length (SD) for training model using urine sample splits after variable selection for both DDA-NB and LC-NB

	Urine Sample Split					
	A	**B**	**C**	**D**	**E**	**F**
DDA-NB	19.4(12.8)	12.5(11.9)	7.42(6.49)	5.49(4.17)	5.75(6.47)	5.01(3.23)
LC-NB	18.5(12.6)	11.2(11.8)	6.39(7.54)	3.83(4.17)	3.83(6.35)	2.81(3.17)

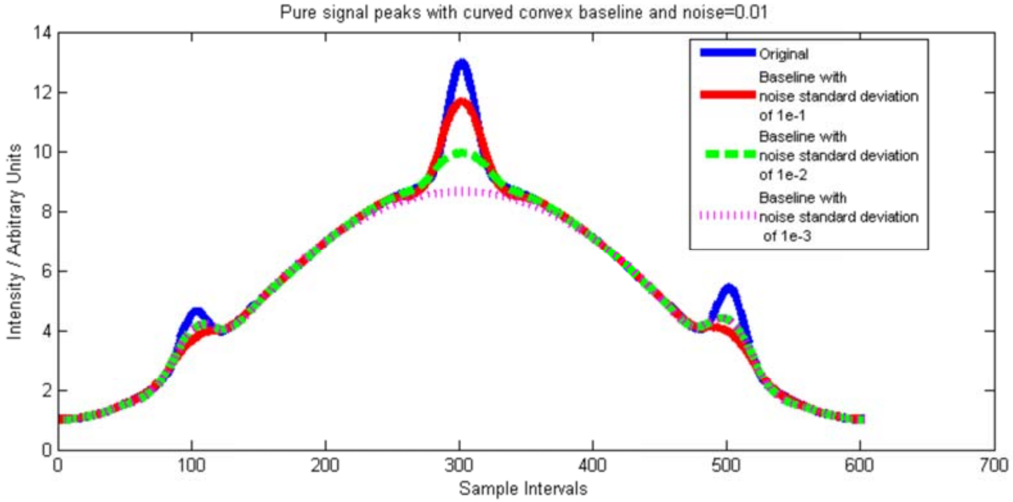


Fig. S1**.** A curve convex baseline simulated spectrum with a noise factor of 0.01 along with baseline estimations using parametric method via different noise standard deviation values.

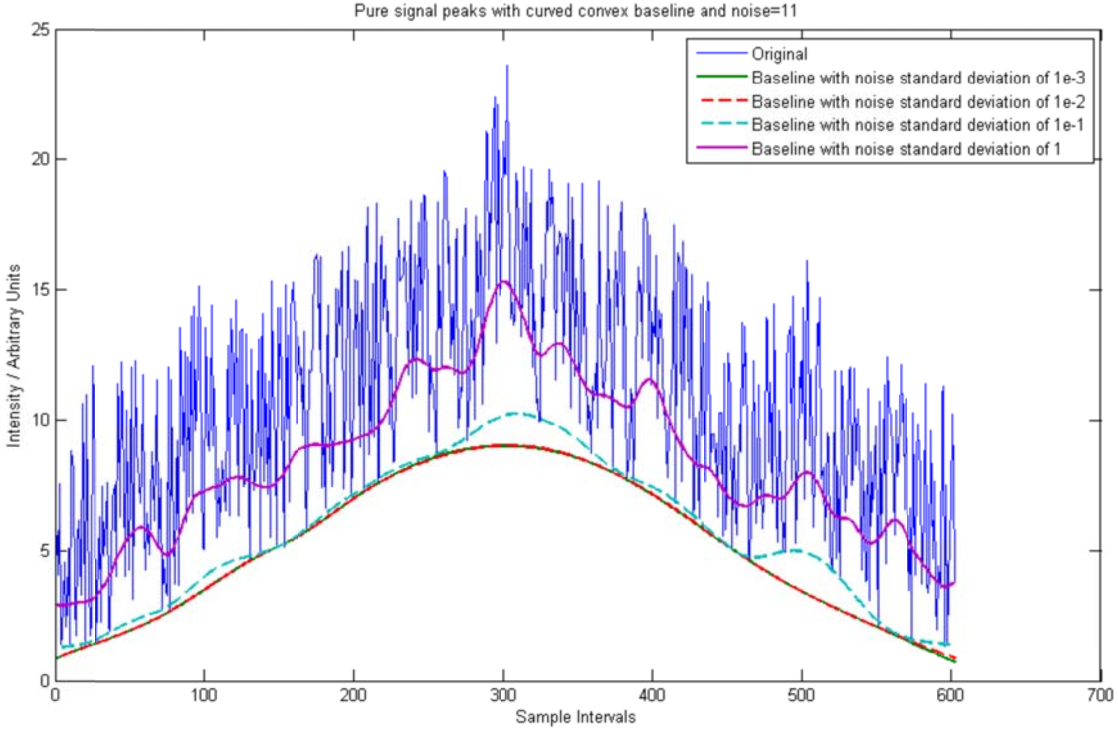


Fig. S2. A curve convex baseline simulated spectrum with a noise factor of 11 along with baseline estimations using parametric method via different noise standard deviation values.

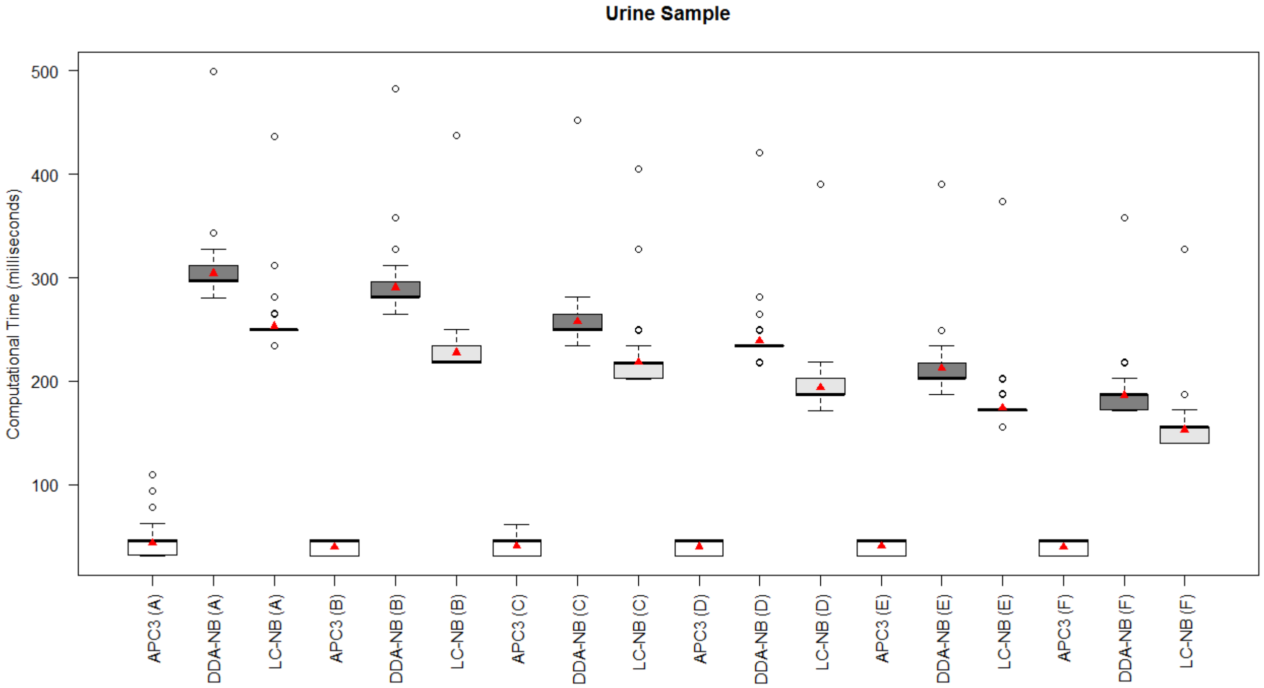


Fig Ia. Boxplot of computational time (both training and testing phases) across the 100 sampling for urine sample splits using APC3, DDA-NB and LC-NB. The red triangle represents the mean value of each boxplot.

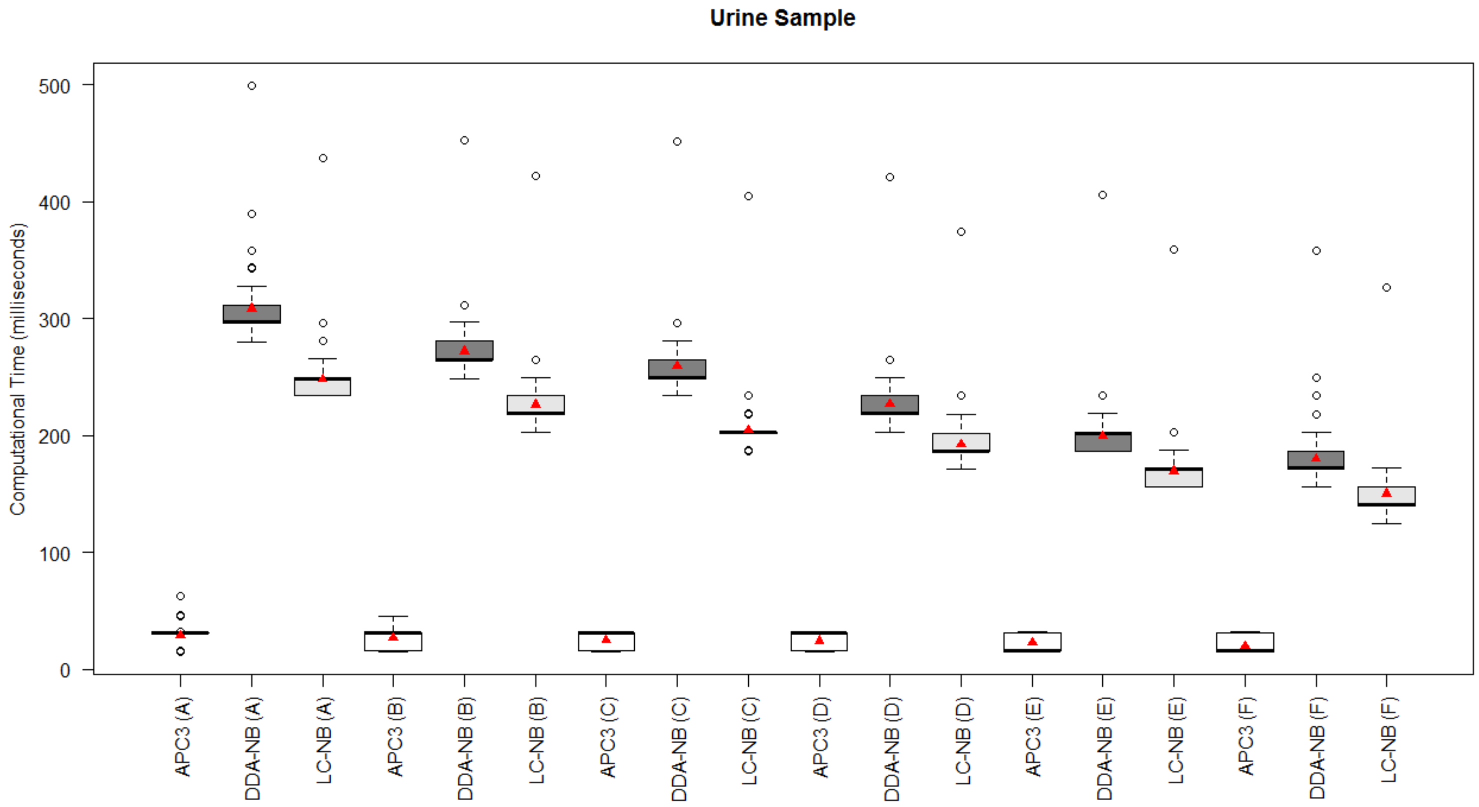


Fig IIb. Boxplot of computational time (only training phase) across the 100 sampling for urine sample splits using APC3, DDA-NB and LC-NB. The red triangle represents the mean value of each boxplot.

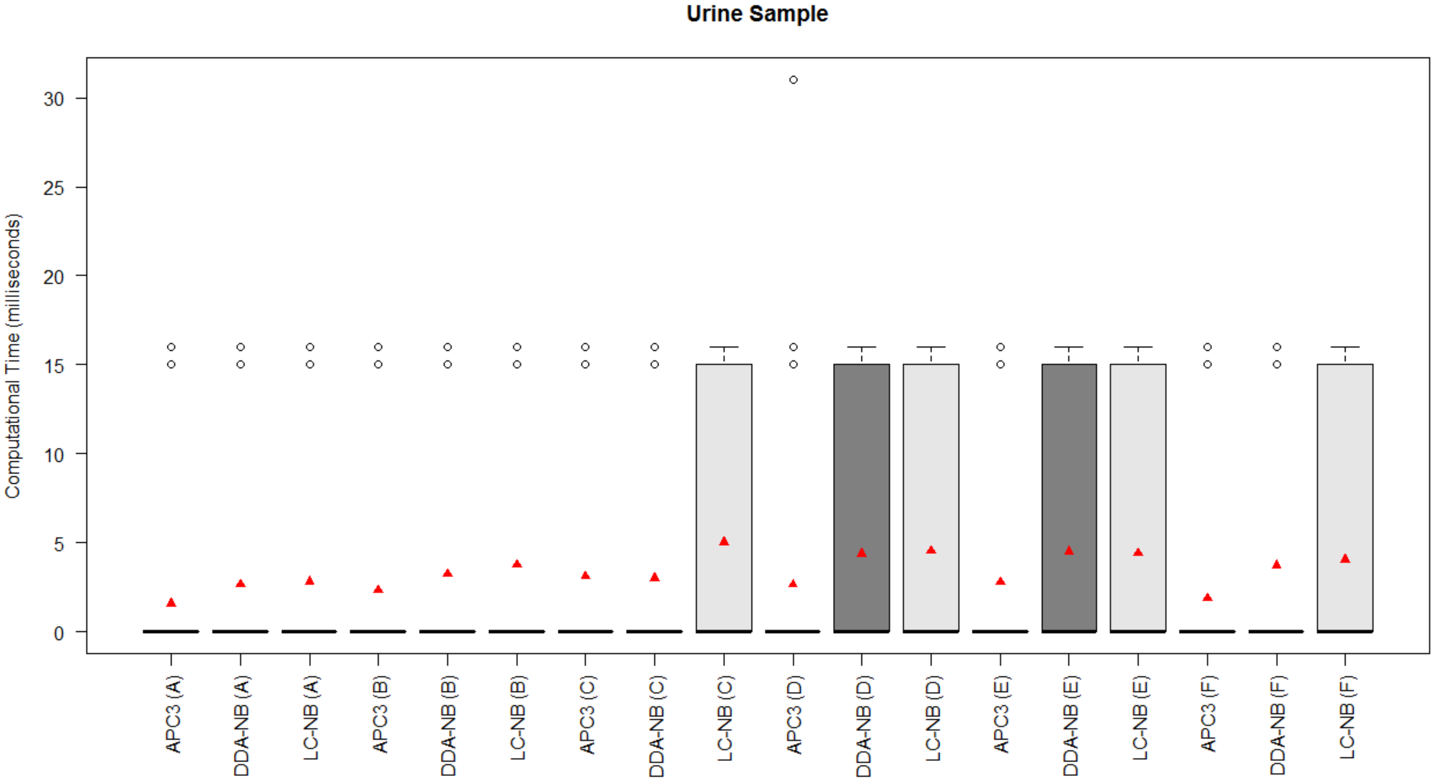


Fig IIIc. Boxplot of computational time (only testing phase) across the 100 sampling for urine sample splits using APC3, DDA-NB and LC-NB. The red triangle represents the mean value of each boxplot.