VARIABLE SELECTION METHODS FOR CLASSIFICATION: APPLICATION TO METABOLOMICS DATA

Thesis submitted in accordance with the requirements of the
University of Liverpool

for the degree of

Doctor in Philosophy

in

Biostatistics

by

Nurain Binti Ibrahim

March 2020

DECLARATION

I declare that the thesis has been composed by myself and the work has not been submitted for any other degree or professional qualification. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others

Nurain Binti Ibrahim

This research was carried out in the Department of Biostatistics, in the Institute of Translational Medicine, at the University of Liverpool,
United Kingdom.

TABLE OF CONTENTS

ABSTRACT

Variable Selection Methods for Classification: Application to Metabolomics Data

by

Nurain Binti Ibrahim

Metabolomics is an emerging field, which focuses on the study of small molecules (metabolites) and their chemical processes. Metabolomics data are highly dimensional, with p>>n where p is the number of variables and n is the sample size. Variable selection is therefore a key step in metabolomics studies. There are three categories of variable selection, such as filter, wrapper and embedded methods.

Common univariate filter methods such as the t-test and ANOVA (analysis of variance) have been often used in the literature to identify important metabolites for a given clinical problem. A challenge in metabolomics research is that metabolite variables tend to be highly correlated. Multivariate approaches that take into account the correlation among variables, such as PCA (principal component analysis), have been applied to reduce the dimensionality of metabolite datasets. The correlation-sharing t-test method (corT) is a filter method that also considers the correlation among variables, but to my knowledge it has only been applied to genomic data. Penalized regression, and in particular the embedded method Lasso, has also been applied for variable selection with the aim of minimising the problem of overfitting that often affects prediction models in this area.

In this thesis I presented a literature review on variable selection methods and classification methods applied to metabolomics data. I proposed an extended version of the variable selection method corT, which I name *adjusted correlation-sharing t-test* (adjcorT). Simulation studies were carried out to compare the performance of several variable selection methods (T, corT, adjcorT and Lasso) using logistic regression for data classification. Simulations assumed a set of 200 variables of which 2 variables were discriminators. A range of sample sizes (*n*=50, 76, 100, 300, 500, 1000, 2000 and 20000) and of different correlation values among the discriminant variables (ρ=-0.8, -0.5, -0.2, 0, 0.2, 0.5, 0.8) were considered to explore the effect that sample size and correlation have on the classification accuracy of each method. These methods were also applied to metabolomics datasets, including data from patients with colorectal cancer (aimed at discriminating between non-cancer vs colorectal cancer groups, and healthy control vs adenoma groups) as well as, kidney disease and infant sepsis datasets. R code was developed to analyse the datasets. Cross validation, with data split into two sets (80% for training and 20% for validation) was used to compare the performance of the variable selection methods using classification accuracy, sensitivity, specificity and area under ROC.

Results from the simulation studies indicate that for small sample sizes (n=50, 76), T, corT, adjcorT and Lasso often failed to select the two discriminatory variables. For example, for ρ=0.5 and n=50, only 3%, 12%, 11% and 0% of the times the two

discriminatory variables were selected. Nevertheless, the detection rates for adjcorT and Lasso improved for negative strong correlations (Table 4.3). These results are consistent with the better performance in classification accuracy observed for adjcorT and Lasso for negative strong correlations ($-0.5 \leq \rho < -1.0$; Table 4.4). As the sample size increased towards n=300, all methods increased their ability to select the two discriminatory variables, with Lasso underperforming for positive strong correlations and corT underperforming for moderate and strong negative correlations. These differences can explain the dissimilarities observed across methods in classification accuracy for sample sizes n=300, 500 and 1000; with Lasso showing poorer performance than T, corT and adjcorT for positive strong correlations, and corT showing poorer performance than T, corT and adjcorT for moderate and strong negative correlations (Tables 4.5 and 4.6). As the sample size increases, T, adjcorT and Lasso offered a similar level of accuracy but corT still underperforms for moderate and strong negative correlations and larger sample sizes (Table 4.7).

In the clinical applications, corT and adjcorT show a similar level of classification accuracy, possibly due to the positive correlation that exists among most metabolites. For non-cancer and cancer discrimination, the method T showed the worst classification accuracy followed by Lasso. Methods corT and adjcorT achieved the best level of discrimination although this was still low (AUC of 0.60; Table 5.3). For healthy control and adenoma discrimination however, methods corT and adjcorT showed the lowest AUC, followed by the T method. Lasso achieved the best level of discrimination, although this remained low (AUC of 0.65; Table 5.8). For the discrimination between bacterial and non-bacterial sepsis cases, Lasso exhibited a better performance that the other variable selection methods with 83.1% classification accuracy (Table 5.13). Lasso also offered the best level of discrimination between healthy controls and kidney disease (AUC=0.90, Table 5.21), although the four methods showed a comparable performance (AUCs=0.86 and 0.87 were achieved with the T and with the corT and adjcorT methods respectively).

My work based on simulations shows that adjcorT offers a flexible approach for variable selection aimed at clinical classification, especially for datasets involving negative correlations between discriminators for medium and large samples where adjcorT consistently shows a better performance than corT. These findings were however not reproduced by the analyses on real data. I believe this is possibly due to the lack of negative correlations among metabolites in the datasets considered.

Both adjcorT and corT are filter variable selection methods. Given that adjcorT showed a better performance compared to corT for negative correlations and a similar performance for positive correlations across all sample sizes investigated, adjcorT is expected to offer advantages compared to corT as a variable selection method for the analysis of some metabolomics data.

COMMUNICATIONS

Attended conferences

Faculty Poster Day, Faculty of Health & Life Sciences, University of Liverpool, UK, 10 June 2016.

Faculty PGR Poster Day, Faculty of Health & Life Sciences, University of Liverpool, UK, 27 March 2019.

Statistical Analysis of Multi-Outcome Data (SAM), University of Liverpool, 3 – 4 July 2017.

Oral Presentations

“A Comparative Study on Filter and Wrapper Feature Selection Methods using Support Vector Machines as the Classifier”. First year PhD student presentation. Department of Biostatistics, Institute of Translational Medicine, University of Liverpool, UK, 4 May 2016.

“Variable Selection and Classification on Metabolomics and Volatile Organic Compounds (VOCs) Data”. Second year PhD student presentation. Department of Biostatistics, Institute of Translational Medicine, University of Liverpool, UK, 6 June 2017.

Poster Presentation

“A Comparative Study on Filter and Wrapper Variable Selection Methods and Application to Metabolomics Data”. Royal Statistical Society. University of Manchester, UK, 8 September 2016.

ACKNOWLEDGEMENTS

I would like to thank various people for their contributions to this thesis. After all these years, there are no proper words to convey my deep gratitude and respect for my supervisors, Professor Marta García-Fiñana, Professor Chris Probert and Professor Lu-Yun Lian for their continuous support and constructive suggestions for this thesis.

My sincere thanks must also go to Dr Kukatharmini Tharmaratnam for her help with the variable selection analysis. I would like to thank Professor Chris Probert and committee for the permission to use the Colorectal Cancer datasets, Dr Marie Phelan for her permission with the Infant sepsis dataset and Dr Baback Roshanravan for his permission with the Kidney dataset. I also thank the members of the Liverpool Multivariate Data Modelling Group from the department of Biostatistics and my fellow colleagues for advice and encouragement.

I would also like to take this opportunity to thank my viva examiners Professor Mario Cortino Borja and Professor Andrea Jorgensen for their comments and suggestions to improve the quality of this thesis.

Special thanks to Ministry of Education Malaysia and Universiti Teknologi Mara (UiTM) as the sponsors and for making me pursuing my PhD possible. I am most grateful to my husband, Muhammad Nor Azfar Bin Mohamad Gani and my daughter, Nur Aufa Aleesya Binti Muhammad Nor Azfar for their continuous supports, which there is no way to express how much it meant to me.

I also would like to thank my friends who are together with me along this PhD journey, making it even more meaningful. Finally, I thank with love my supportive family, my parents, Ibrahim Bin Halid and Rozita Binti Abdul Rahman and my siblings, Muhammad Irsyad Bin Ibrahim and Lyana Binti Ibrahim.

ABBREVIATIONS

Acc	Classification Accuracy
Sen	Sensitivity
Spe	Specificity
AUC	Area Under ROC
PCA	Principal Component Analysis
VOC	Volatile Organic Compounds
corT	Correlation Sharing T-statistics
adjcorT	Adjusted Correlation Sharing T-statistics
T	T-test feature selection
KNN	k-Nearest Neighbors
PLS-DA	Partial Least Square – Discriminant Analysis
ANOVA	Analysis of Variance
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine
RF	Random Forest
CFS	Correlation-based feature selection
DA	Discriminant Analysis
MRRMRR	Minimum Regularized Redundancy Maximum Robust Relevance

MSE	Mean Square Error
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
PC	Principal Component
Var	Variance

STATISTICAL SYMBOLS AND NOTATIONS

x	metabolites, biomarkers, variables, features
y	class, outcome
p	number of metabolites
ρ	correlation among metabolites
n	total number of samples (total sample size)
n_j	number of samples in the j-th groups (sample size)
s_i	pooled within-group standard deviation of the i^{th} variable
$\bar{x}_{ij}$	sample mean of the i-th variable $(i = 1,2, \ldots\ldots, p)$ in group j
x_i	observation of i^{th} variable (e.g., metabolite expression) where i=1, 2, ….., p
max	maximum
min	minimum
β	coefficient
λ	tuning parameter
t	upper bound of the summation of the absolute coefficients
L_1	lasso penalty
L_2	ridge penalty
μ	population mean
Σ	population covariance matrix
μ_i	population mean of the i^{th} variable
$f(x)$	multivariate density function
$E(Y\|x)$	conditional mean of Y given x

LIST OF TABLES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

LIST OF FIGURES

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Appendices

Chapter 1
Introduction

1.1 Clinical Classification

Classification techniques are often applied to allocate individuals into groups (e.g., disease/non-disease) and are widely used in biomedical and clinical applications. Examples of classification methods include logistic regression, K nearest neighbours, support vector machine, linear discriminant analysis and multivariate generalized linear mixed model among others. For example, Wah et al. [1] used logistic regression to identify patients with diabetes using the Pima Indian Diabetes dataset. The authors also used the Breast Cancer Wisconsin dataset to identify patients with malignant and benign tissue. They also applied logistic regression to a Spam base dataset in order to identify spam emails.

Classification methods are widely used in cancer research to make prediction by assigning tumours to known classes (class prediction) or investigating new cancer classes (class discovery) [2]. Chudova et al [3], for example, applied support vector machine to genomics data in order to distinguish benign from malignant thyroid nodules. Mishra et al. [4] also applied support vector machine and neural network to classify the important biomarkers into acute lymphocytic leukemia (ALL) and acute myelocytic leukemia groups.

Linear discriminant analysis was employed to distinguish the difference between Bacillus species (one of the bacterial species) by using the *lda* function (to perform linear discriminant analysis) and *predict* function (to assess prediction accuracy for linear discriminant analysis) in MASS package (R software) [5]. Discriminant analysis can be also applied to longitudinal data to classify individuals by taking into account changes over time of relevant biomarkers. In 2018, Hughes et al [6] developed a multivariate generalized linear mixed model and applied a dynamic discriminant approach to model changes of number of seizures and treatment history over time, to identify people with epilepsy who will not achieve 12 months seizure remission within 5 years of starting treatment.

Clinical datasets can be highly dimensional where the number of variables is larger than number of samples or patients (i.e., each patient/sample is characterized by hundreds/thousands of variables). One of the key challenges when modelling high dimensional datasets is how to avoid overfitting. One common approach to deal with the challenge of overfitting is reducing the dimensionality of the datasets with variable selection techniques [7].

1.2 Variable selection within the context of clinical classification

In order to accurately classify a sample into groups of interest, sometimes it is necessary to first reduce the number of variables, especially when the dataset contains a large number of potential predictors compared to the sample size (i.e., number of individuals or samples). The idea is that the reduced set of variables should still capture the most important predictor variables. Variable selection is widely used in many areas, including metabolomics. Some of the objectives of variable selection are to facilitate data understanding, reduce the storage requirement, reduce the processing time and reduce the dimensionality of the dataset while achieving a good prediction performance. There are several variable selection methods, such as correlation-based feature selection [8]–[10], principal component analysis (PCA) [11]–[16], T-test feature selection method (T method) [15]–[17], and Lasso [15], [18]–[20].

Principal component analysis (PCA) is one of the most common variable selection methods, which generates a low-dimensional representation of data that describes most of the variability in the dataset. PCA uses a mathematical procedure to do the data reduction which produces new variables as linear combinations of the original variables. It transforms the original data onto a smaller number of principal components. For example, Shahamat and Pouyan [21] used PCA in order to reduce the number of functional magnetic resonance imaging (fMRI) time points and linear discriminant analysis was used for classification of patients into schizophrenia and control groups. Rao, Sui and Zhang [22] investigated the significant walnut kernels by using PCA as the variable selection method and these kernels were used as the basis for further studies on walnut kernel metabolism.

The applications of this thesis focus on metabolomics studies. Metabolomics is an emerging field, which focuses on the study of small molecules (metabolites) and their chemical processes. Metabolomics datasets can be used to differentiate between two or more groups of outcomes such as disease or non-disease groups based on thousands of metabolites. The challenge in this area is that metabolomics data are highly dimensional, with $p>>n$ where p is the number of variables and n is the sample size. Hence, variable selection is therefore a key step in metabolomics studies. Common univariate tests such as the t-test and ANOVA (analysis of variance) have been often used in the literature to identify important metabolites for a given clinical problem. Multivariate tests, such as PCA (principal component analysis), DA (discriminant analysis) and PLS-DA (partial least square-discriminant analysis) have also been applied in this area. Another challenge in metabolomics research is that metabolite variables tend to be highly correlated. Some researchers have used penalized regression methods, such as Lasso since this method takes into account the correlation among the metabolites within the variable selection process.

1.3 Objectives of the thesis

The specific objectives of this thesis are:

Objective 1: Literature review on variable selection methods for classification applied to metabolomics data. Variable selection methods include correlation-adjusted t-scores (cat scores), forward selection and principal component analysis

(PCA), among others. Additionally, the classification methods involved in the literature review are partial least square-discriminant analysis (PLS-DA), discriminant analysis (DA), and support vector machine. Briefly, variable selection methods are characterized into three categories namely filter, wrapper and embedded methods:

a) Filter variable selection methods: These methods constitute the simplest approach. They use variable ranking and variable score methods through a univariate framework.
b) Wrapper variable selection methods: These methods use a multivariate approach as they consider variables simultaneously.
c) Embedded variable selection methods: Combination of filter and wrapper methods.

Objective 2: Development of a new approach for variable selection and comparison with existing variable selection methods in terms of classification accuracy via simulations.

Objective 3: Application of existing methods and of my proposed method to real metabolomics datasets. Three metabolomics datasets are considered, including a colorectal cancer dataset, an infant sepsis dataset and kidney disease dataset.

1.4 Datasets

I aim to apply the variable selection methods discussed in this thesis to real metabolomics datasets in order to assess their performances. The three clinical datasets used in this thesis are described below.

1.4.1 The Colorectal Cancer Dataset

The colorectal cancer dataset consists of 137 samples (samples from 60 healthy controls, 56 adenoma and 21 colorectal cancer patients) and 146 variables. Therefore, the number of samples in the non-cancer group is 116 samples and the number of samples in the colorectal cancer group is 21. I am interested in developing prediction models that allows for the classification of patients into non-cancer and cancer, and also normal and adenoma. This colorectal cancer dataset was used in a previous study with the aim of identifying the Volatile Organic Compounds (VOCs) emitted from stool that can discriminate patients between cancer and no neoplasia groups [23]. This colorectal cancer dataset was gathered by mass-spectrometry (MS) technique. A. Bond et al used Student's t-test, Man-whitey tests, Fisher's exact test, ANOVA in order to determine the significant variables. Partial least squared discriminant analysis (PLS-DA) and logistic regression were used as the classification methods.

1.4.2 The Infant Sepsis Dataset

Clinicians at the Alder Hey Hospital in Liverpool are investigating better ways to discriminate between bacterial and viral sepsis in children. They collected blood samples from patients in intensive care and transferred the samples to the University of Liverpool NMR Metabolomics Centre with the aim of acquiring 1H NMR spectra of 25 samples from infants with bacterial sepsis and 91 samples from non-bacterial sepsis infants. This data has 144 metabolites, and 116 children participated in this study. Data is publicly available in the database MetaboLights with ID MTBLS653.

1.4.3 The Kidney Disease Dataset

Chronic kidney disease (CKD) leads to a decreased sensitivity of the metabolic effects of insulin. The plasma metabolome was examined in 93 adults without diabetes in the fasted state, out of which 56 showed moderate-severe CKD and 37 a normal glomerular filtration rate. This data, which contains data on 124 metabolites, was used in the previous study [24].

1.4.4 Summary of datasets

This subsection provides a summary of datasets considered for this thesis. Table 1.1 shows the list of datasets.

Table 1.1: Summary of datasets that I considered in this thesis

No.	Datasets Name	Where the data come from?	Sample size	Number of variables
1.	Colorectal cancer	[25]	137	146
2.	Infant Sepsis	MetaboLights website https://www.ebi.ac.uk/metabolights/	91	144
3.	Kidney disease	[24]	93	124

1.5 Structure of the thesis

The structure of this thesis is as follows. In Chapter 2 I present a literature review on variable selection methods and classification methods applied to metabolomics data. In Chapter 3, I outline algorithms of existing variable selection methods (T, corT and Lasso) and I propose a new method that I name adjcorT. I present the results of a simulation study in Chapter 4 where the novel adjcorT is compared to T, corT and Lasso; and where the logistic regression is used as the classification method. In Chapter 5, I present the results of the application of T, corT, adjcorT and Lasso to three clinical datasets. Finally, in Chapter 6 I present a summary and discussion of the work I have completed for this thesis and recommendations for future work.

Chapter 2

Variable selection methods for classification in the area of metabolomics

2.1 Introduction

This thesis focuses on metabolomics which is the study of global metabolite profiles in a system (cell, tissue or organism) under a given set of conditions. Metabolomics has a number of features: 1) As the metabolome is the final downstream product of gene transcription, any changes in metabolome capture the changes in the transcriptome and the proteome, 2) Metabolome is the closest to the phenotype (physical appearance) of the biological system (e.g.: cell / organ / entire organism) compared to genome and proteome and 3) Metabolome is more diverse than genome and proteome as it contains many different biological molecules. Metabolomics datasets are often highly dimensional, with the number of metabolites being greater than the number of samples [26]. The number of sample is often limited since collection of this type of data is relatively expensive [27]. The problem of metabolite selection is complex as highlighted in the previous chapter. Metabolomics datasets commonly consist of many correlated metabolites and a small sample size (i.e., small number of samples or individuals). Several statistical methods of variable selection are available to identify important clinical predictor variables [11], [18], [28]–[30]. However, existing methods may have limitations when applied to metabolomics data due to the nature of metabolomics datasets.

This chapter highlights the importance of metabolomics and describes the range of variable selection methods and classification methods used in the area of metabolomics. The structure of this chapter is therefore as follows. I review the importance of metabolomics in Section 2.2. A literature review that focuses on variable selection methods for metabolomics is conducted in Section 2.3. The variable selection methods for metabolomics (filter, wrapper or embedded) are discussed in Section 2.4. In Section 2.5 I present a workflow of variable selection methods of metabolomics for classification application to metabolomics data. Additional variable selection methods are explained in Section 2.6, and classification methods used in metabolomics are explained in Section 2.7. The discussion is in Section 2.8.

2.2 Importance of metabolomics

Metabolomics is an emerging field which combines strategies to identify and measure quantitatively cellular metabolites from biofluids, such as blood and urine, present in organisms, cells, or tissues, from either animals or humans using advance analytical techniques with the application of statistical and multi-variant methods. Metabolomics is widely used in order to find novel biomarkers in biological systems, biofluids and for discovery of dietary and health biomarkers. Biofluids such as urine seem more advantageous as they are easy to collect. To understand the complicated biochemical systems and to uncover mechanism such as metabolic pathways related to disease, gender, diet and etc. remains a challenge. Recently, metabolomics emerged for disease diagnosis, biomarker identification, a deeper understanding of cancer metabolism and drug toxicity, the potential for improved early disease detection or therapy monitoring [19], [30]–[33]. Additionally, metabolomics has successful applications in environmental science, nutrition, characterizes biochemical systems and reveal insights into the mechanisms of pathophysiological processes [12], [34]–[36].

The two most commonly used analytical technologies are nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS). The combination of NMR and MS variables make them more attractive compared to variables of NMR or variables of MS alone. However, the majority of metabolomics studies used either NMR or MS separately. Djukovic et al. [30] combined NMR and MS data but they claimed this approach has not been well developed. They proved that the results were promising as using both NMR and MS data significantly improved predictive accuracy in all the pairwise comparisons among colorectal cancer, polyps and healthy controls.

Metabolomics data are typically high-dimensional ($p \geq n,$ where n is the sample size and p is the metabolites) and the metabolites are correlated to each other. Hence, appropriate variable selection methods and classification methods should be applied to metabolomics data in order to effectively reduce the dimensionality of the datasets and to accurately classify individuals into two or more than two different groups of datasets (for example, disease/non-disease groups). Metabolomics provides an important approach in the investigation of biological systems and the effect of internal and external perturbations through the study of changes in metabolite concentration. Yengo et al. [19] conducted a metabolomics study by developing predictive models of type 2 diabetes using logistic regression and cox regression. They also used Lasso as the variable selection methods in order to find the novel biomarkers to detect type 2 diabetes. Metabolomics showed an improvement in prediction of type 2 diabetes. Meanwhile, Everett J.R. [37] studied the metabolic profiling to predict drug efficacy and safety. The notable advantages of conducting the metabolic profiling are that it reflects the physiological status of the patient in real time and its ability to be sensitive to both genetic and environment factors such as the status of gut microbiome. Xu et al. [20] used metabolomics data as potential resources for prediction of yield in hybrid rice.

As described earlier, metabolomics is the study of small molecules. These molecules are synthesized by a diversity of enzymes. Although the association among the metabolites can be low, moderate or high, with either positive or negative correlations, metabolites in metabolomics data tends to be highly correlated due to stronger mutual control by a single enzyme [38]. There are often missing values in metabolomics data since certain compounds cannot be identified/quantified in certain samples. The missing values might be produced by random error or stochastic fluctuations during the data acquisition [39]. Additionally, other factors of missing values are: i) computational detection failure, ii) measurement error, iii) signals of low intensity which is distracted by background noise and iv) imperfection of the detection method used [40]. In terms of metabolomics data distribution, certain metabolites tend to demonstrate right-skewed distributions and certain metabolites displays a substantial proportion of zero values that may be regarded as true zero values based on biological grounds [41].

Metabolites can be negatively correlated to each other as shown in the previous studies [41],[42]. It is important to deal with negative correlation metabolomics datasets in order to capture dynamics in the correlation structure of the metabolites which can improve the classification accuracies when these metabolites are included into the classification model [42].

Since metabolomics data is highly dimensional, a variable selection process is an important step in metabolomic studies where often the goal is to find the most informative of metabolites. Variable selection methods have been applied to metabolomics datasets in order to identify a minimal set of strongest biomarkers related to a predefined research outcome; for example, to identify potential diagnostic or prognostic biomarkers of disease and non-disease. Biomarker discovery is an important goal in metabolomics. This thesis focuses on variable selection methods for classification, and therefore the interest is not merely on the discovery of metabolites that are highly associated with the outcome of interest (e.g., development of a disease) but also on their predictive ability. In the next section I will discuss the categorisation of variable selection methods for metabolomics.

2.3 Literature search

In this literature review I focus on variable selection methods that are applied to metabolomics. The literature search applies the following terms: “variable selection” AND “metabolomics” in the full text articles. The search considered research papers and publications written in English and published within the last 10 years (i.e., from 2009 to 2019).

Bramer et al. suggested to use multiple databases to search relevant references in order to conduct efficient searches [44]. Three databases were used for literature search purpose in this thesis: Public MEDLINE (PubMed), *Medical Literature Analysis and Retrieval System Online* (MEDLINE) and American Chemical Society (ACS) publications databases. PubMED and MEDLINE databases focus on biomedical and life sciences publications. Meanwhile, ACS publications had the potential to capture metabolomics data publications. Figure 2.1 shows the databases used while searching the literature, the number of articles excluded, and the number of articles selected. Using the PubMed database, 44 articles were identified. By using multi-field search in MEDLINE database, 62 articles were found. Additionally, ACS publications database identified 117 articles. In overall, 82 articles have been reviewed.

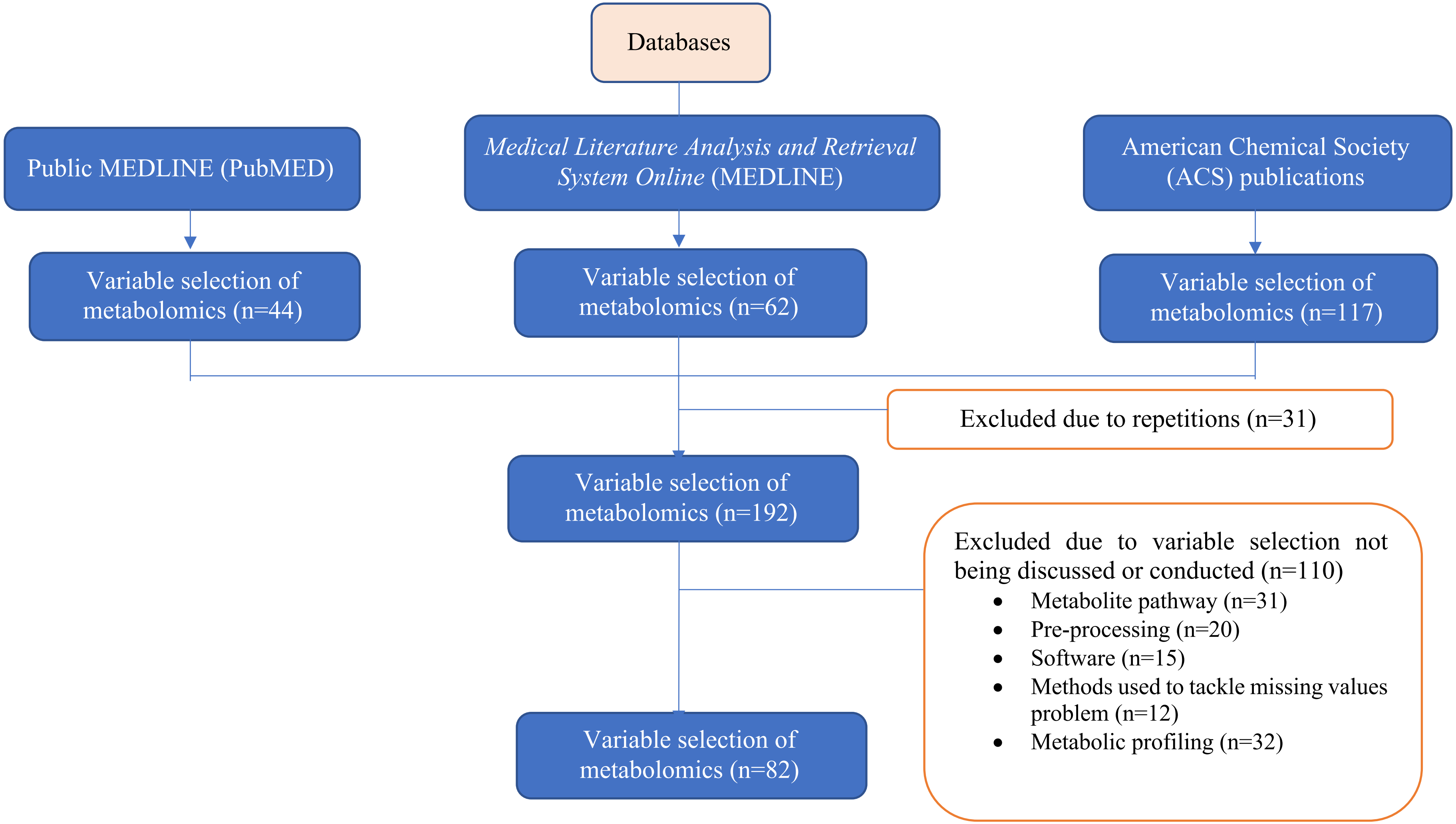

Databases
Public MEDLINE (PubMED)
Medical Literature Analysis and Retrieval System Online (MEDLINE)
American Chemical Society (ACS) publications
Variable selection of metabolomics (n=44)
Variable selection of metabolomics (n=62)
Variable selection of metabolomics (n=117)
Excluded due to repetitions (n=31)
Variable selection of metabolomics (n=192)
Excluded due to variable selection not being discussed or conducted (n=110)
Metabolite pathway (n=31)
Pre-processing (n=20)
Software (n=15)
Methods used to tackle missing values problem (n=12)
Metabolic profiling (n=32)
Variable selection of metabolomics (n=82)

Figure 2.1: Databases used to search research articles, indicating the number of papers excluded and the reasons, and the number of articles selected to be reviewed

2.4 Variable selection methods of metabolomics

Table 2.1: List of variable selection methods and classification to analyse metabolomics data

Variable selection methods			
Univariate approach	Multivariate approach		
Filter	Wrapper	Embedded	Other
• Correlation-adjusted t-scores (cat score) [44] • Analysis of Variance (ANOVA) [12], [13], [49], [50], [14]–[16], [18], [45]–[48] • Error rate p-values (ERp) [27] • Extension of ERp (XERp) [51] • T-tests [15]–[17], [52]–[55] • Rank aggregation [52] • Selectivity ratio [56], [57] • Relief algorithm[52] • Wilcoxon rank-sum [52], [53] • Correlation-based feature selection (CFS) [8]–[10] • Mutual information [58] • Signal to Noise Ratio [59] • Chi Square [50] • Sensitivity ratio [57]	• Minimum Regularized Redundancy Maximum Robust Relevance (MRRMRR) [58] • Forward selection [60] • Stepwise regression [61] • Backward selection [49]	• Least Absolute Shrinkage and Selection Operator (Lasso)[15], [18], [65]–[67], [19], [20], [33], [35], [50], [62]–[64] • Elastic Net [15], [18], [65], [68] • MUVR[69] • Sparse Group Lasso [65], [68] • Group Lasso [68] • Adaptive group-regularized ridge regression [68]	• Principal Component Analysis (PCA) [11], [12], [54], [55], [59], [60], [70]–[75], [13], [76]–[85], [14], [86], [87], [15], [16], [31], [42], [50], [53]Logit-Normal Continuous Analogue of the Spike-and-Slab Prior (LN-CASS) [65] • Horseshoe [65] • Ordinary Least Square [65] • Multi-block Variable Influence on Orthogonal Projections (MB-VIOP) [73] • Orthogonal Projections to Latent Structures (OnPLS)[73] • Wisdom of artificial crowd (WoAC) [66] • SVM-RFE [47] • RF-RFE [47] • Best linear unbiased prediction (BLUP) [20] • Boruta[88] • Kruskall-wallis non-parametric test [15] • Genetic Algorithm [89], [90] • Variance of the b regression vector[91]
Classification methods			
• Partial Least Square – Discriminant Analysis (PLS-DA) [5], [11], [48], [52]–[54], [57], [58], [60], [67], [69], [71], [12], [74], [78], [80], [89], [92]–[97], [14], [98]–[104], [15], [28]–[30], [37], [46] • Logistic regression [19], [47], [49], [55], [61] • Kernel-based PLS [105] • Sparse PLS-DA [99], [106], [107] • Discriminant Analysis (DA) [5], [26], [50], [72], [77] • Support Vector Machine (SVM) [5], [26], [47], [65], [72], [95] • Random Forest (RF) [5], [26], [35], [47], [65], [66], [69], [72], [95], [108] • K-nearest neighbour (KNN) [26] • Neural Network [65]			

Table 2.1 shows the list of variable selection and classification methods applied to metabolomics data. The variable selection methods are categorised into filter, wrapper and embedded methods in univariate or multivariate approaches. Some additional variable selection methods are listed in 'Other' column. In addition, some of the classification methods applied metabolomics data including Partial Least Square-Discriminant Analysis (PLS-DA), Discriminant Analysis (DA) and Neural Network. However, this thesis focuses on the variable selection methods of metabolomics only.

T-test and Analysis of Variance (ANOVA) can be applied for variable selection. The assumptions of these methods are normality of the data and homogeneity of variance. Havlicek, L. L. & Peterson N. L. (1979) studied the empirical effects of quantified violations of assumptions underlying t-test and ANOVA using Monte Carlo procedure. They concluded that t-test and ANOVA are remarkably robust to deviations from normality and different sample sizes [110]. Blanca et. al. (2018) investigated the robustness of t-test and ANOVA in relation to variance homogeneity using the Monte Carlo simulation. The ratio of the largest to smallest variance (variance ratio) is a measure of variance homogeneity. The results suggest that a variance ratio above 1.5 may be established as a rule of thumb for considering the robustness under homogeneity for t-test and ANOVA [111].

Filter, wrapper and embedded variable selection methods are explained in the sub sections 2.4.1, 2.4.2, and 2.4.3, respectively.

2.4.1 Filter variable selection methods

The goal of filter variable selection is to extract the most important metabolites from metabolome gathered by NMR or MS by using a variable ranking or variable score. Filter variable selection methods are the simplest methods and most widely used in metabolomics studies [29].

Most filter methods follow a univariate approach. Univariate approaches are useful for uncovering simple associations between biomarkers and responses. They offer simplicity as well as inexpensive and computational efficiency when applied to complex datasets such as metabolomics data. However, univariate approaches are unable to reveal some key features of the data such as patterns of correlation among the biomarkers since each variable is considered independently. As a result, this leads to redundancy issue. It is one of the common issues in metabolomics which there are recurrent detection of adducts that greatly inflate the number of detected peaks [109]. In addition, the classification and prediction are also affected if the key features of the data are not revealed. Some metabolites might be non-significant on their own but become significant when analysed in combination with other metabolites (as they have association among them). Hence, in order to overcome this problem, some filter methods that considers the correlation have been proposed such as correlation-based feature selection [8]–[10].

The most commonly used univariate filter method is the Analysis of Variance (ANOVA). This method calculates a p-value for each metabolite in order to find the most significant metabolites. ANOVA is often used to identify variables that significantly differ between two or more independent groups through the p-values. ANOVA partitions the total variance of the metabolomics dataset into a number of components, so that the significant contributions of identified sources of variance to the total variation in responses can be determined. As a result, the ANOVA coefficient F, is calculated to allow the differences between means of groups to be assessed. Previous research used ANOVA in their studies [12], [14]–[16], [110]. Kirpich et al. studied a software named SECIMtools, which a suite of metabolomics data analysis tools. The authors claimed that ANOVA was in this software and it was commonly used to analyse metabolomics data. ANOVA equations can be expressed as follows in Equations 2.1-2.5 [111]:

$$F = \frac{MST}{MSE} \tag{2.1}$$

$$MST = \frac{SST}{g-1} \tag{2.2}$$

$$SST = \sum_{i=1}^{g} n_i(\bar{x}_i - \bar{x})^2 \tag{2.3}$$

$$MSE = \frac{SSE}{\sum_{i=1}^{g} n_i - g} \tag{2.4}$$

$$SSE = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \tag{2.5}$$

where F is the ANOVA coefficient, where $MST = \frac{SST}{g-1}$ is the mean sum of squares due to treatment and $MSE = \frac{SSE}{\sum_{i=1}^{g} n_i - g}$ is the mean sum of squares due to error. SST, where $SST = \sum_{i=1}^{g} n_i (\bar{x}_i - \bar{x})^2$, is the sum of squares due to treatment and $SSE = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$is the sum of squares due to error. The total number of groups (treatments) is denoted as g, n_i is the total number of samples in the i-th group, $\bar{x}_i$ is the mean value for each group, x_{ij} is the data value for each sample i and $\bar{x}$ is the mean value for all data. ANOVA uses the null hypothesis that there is no difference in means between the groups versus the alternative hypothesis that means differ across groups. Hence, if more than two groups are of interest in the study, and the null hypothesis is rejected, further analysis needs to be conducted to determine which groups differ (such as Tukey test) [14], [45], [112]. ANOVA may not be appropriate when the variance across groups are not equal and/or the data is not normally distributed. Some other univariate filter methods of metabolomics are rank aggregation, selectivity ratio, relief algorithm, mutual information, signal-to-noise ratio, chi square and sensitivity ratio. Wilcoxon rank-sum test is a non-parametric test for non-normal metabolomics data.

The Chi square filter variable selection is less complex computationally than Gain Ratio as the latter requires a decision tree [113]. However, Chi Square can only be used for categorical datasets. Since metabolomics datasets also involve continuous independent variables, Chi square was not applied to our datasets [114].

The next subsection will discuss the wrapper variable selection methods.

2.4.2 Wrapper variable selection methods

Wrapper methods are not commonly used in metabolomics since this method is more complicated and costly compared to filter methods. They commonly use a multivariate approach as they consider metabolites simultaneously. Wrapper methods are wholly data-driven, do not require interpretation of variable importance score and they are independent of the chosen modelling methodology. One of the wrapper methods is forward selection which starts the procedure with a null set of variables [1], [115]–[117]. The algorithm starts with the empty variable set *S*. Then, continuously add variables selected by some evaluation function that minimizes the mean square error (MSE). At each looping, the algorithm will select among the remaining available of the variable set which has not been added to the variable set. The algorithm will stop when the maximum appropriate number of variables is reached. Marcano-Cedeno et al. [118] claimed that if the optimal subset of a number of variables is low, less computational time is needed to employ forward selection. Forward selection uses the parsimony concept and it only shows the "most important" biomarkers. This method also is less susceptible to multicollinearity when applied to metabolomics data. There is possibility of redundancy when using this method. Let say, forward selection method was selected x_1 as the first variable and x_{20} was selected as the second variables. However, actually x_1 and x_{20} are highly correlated and there is redundancy issue in the variable selection process. Other disadvantage of forward selection is that it does not include a mechanism for removing variables after these have been included in the model, even if the model is insignificant or irrelevant. The complement of forward selection is backward elimination. This method initializes the full set of variables which is opposite to forward selection. Forward selection computes faster than backward elimination because forward selection evaluates very small variables sets, compared to backward elimination that evaluates an almost full set of variables. Guyon Isabelle [119] claimed that backward elimination has the ability to remove the "worst" biomarkers early, and consequently, relatively few models are considered by leaving only "important" biomarkers. Besides that, the first model is the most complicated and it is susceptible to multicollinearity when applying this method to metabolomics data.

An additional wrapper method is stepwise regression and it is a hybrid method which combines both backward elimination and forward selection. This method checked all variables and only includes significant variables that have the most contribution to the classifier into the model. At the same time, it removes non-significant variables from the model. Stepwise regression has the ability to manage large amounts of the potential biomarkers. However, some variables (especially dummy variables) may be excluded from the model despite being truly important for the model. Nishiumi et al. [61] used stepwise regression after the pre selection process of the metabolites through the p-values obtained by the Mann-Whitney U test which is used for not normal metabolomics data. There are two limitations of wrapper methods, which may explain why these methods are not commonly used in metabolomics (see Table 2.1). One limitation is the substantial computation time that is required for a large number of variables. A second limitation is that when the number of observations is relatively small (compared to the number of variables), the risk of overfitting increases.

Another wrapper method applied in the previous study is the Minimum Regularized Redundancy Maximum Robust Relevance (MRRMRR) method, which is insensitive to the presence of outliers in the continuous measurements [58]. This method is an extension version of the Minimum Redundancy Maximum Relevance (MRMR) approach which is sensitive to outliers. MRRMRR is suitable for high-dimensional data and it combines the principle of regularization and robust statistics. Nevertheless, it is complex computationally.

2.4.3 Embedded variable selection methods

An embedded method is a hybrid method which combines both filter and wrapper methods. It takes advantage of the selection process by performing variable selection and classification simultaneously. This method is more complicated and expensive compared to filter and wrapper methods as they are model dependent, and they assess the model performances while selecting the important metabolites. The most commonly used embedded methods in metabolomics is the Least Absolute Shrinkage and Selection Operator (Lasso). The formula of Lasso is explained in Chapter 3 (Methodology) since Lasso is used in this thesis. Yengo et al. [19] used Lasso to determine the impact of this method on the prediction of type 2 diabetes using 293 non-targeted metabolomics profiling and 1172 subjects. The difference between this study and other studies is the authors used non high dimensional metabolomics datasets. In particular, area under the receiver operating characteristic curve (AUROC) was used as the predictive tool. The authors used Lasso to predict of type 2 diabetes with the aim of maximizing the out-of-sample AUROC. As a result, Lasso improves the prediction of type 2 diabetes on top of known clinical and biological markers and it achieved 90% in total AUROC. In my point of view, even though Lasso achieves high percentage of AUROC, I think, this study does not need to used Lasso as the sample sizes are larger than number of variables and this data can only use the internal variable selection in logistic regression. Xu et al. [20] claimed that Lasso is one of the best methods for prediction of metabolomics dataset compared to transcriptomics and genomics dataset. Yengo et al. [19], Xu et al. [20] and Marco-Ramell et al. [67] used Lasso as variable selection method in order to make predictions. Marco-Ramell et al. [67] used Lasso as a predictive biomarker model to identify samples with high insulin resistance and it reached a high predictive power which is 80.1% of AUROC percentage. Meanwhile, Newman et al. [18] used Lasso in order to investigate its performance for a small sample size which is less than 100. The study used a simulation study and two real datasets (including maize data and genomic expression dataset for type 1 diabetes). Other methods also used in this study including ANOVA an Elastic Net and the authors only made a conclusion that ANOVA is an excellent choice if the goal of a study is to advance a set of variables to the next round of testing for biological relevance because the Type II error rate for the ANOVA is lower than other methods.

The second most common embedded method used in metabolomics is Elastic Net, which is the weighted combination of both Lasso and ridge regression penalties. First of all, Elastic Net was introduced for prediction involving linear models. Then, it was extended for generalized linear model, such as logistic regression which can be used for classification. There is little difference between the formula of Lasso and Elastic Net. Lasso is using a L1 penalty term which is equal to the absolute value of the magnitude of the coefficients $\beta, \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$. When using this penalty Lasso selects at most p biomarkers before it saturates and if there is a group of highly correlated biomarkers, Lasso tends to select only one variable from the group. Elastic Net was used in order to overcome these limitations by adding a quadratic part to the L2 penalty ($\|\beta\|^2$) which used in ridge regression where ($\|\beta\|^2$) $= \sum_{j=1}^{p} |\beta_j|^2$. The formula of Elastic Net can be expressed as follows in Equation 2.6:

$$\hat{\beta} = \underset{argmin}{\beta} (\|y - X\beta\|^2 + \lambda_2 \|\beta\|^2 + \lambda_1 \|\beta\|_1) \tag{2.6}$$

where β represents coefficients of all parameters. λ_2 is the value of the lambdas for L2 penalty and λ_1 is the value of the lambdas for L1 penalty. A special case of Elastic Net is where $\lambda_1 = \lambda, \lambda_2 = 0$, or $\lambda_1 = 0, \lambda_2 = \lambda$. The choice of the penalty parameter affects the result. Increasing of bias and a poor prediction in the result might be happening because this method uses a two-stage procedure, where for each fixed λ_2 it finds the ridge regression coefficients followed by a Lasso type shrinkage. Lasso and Elastic Net are quite computationally demanding since they are going through the validation stage while selecting the "important" biomarkers. Lu et al. [120] compared the predictive performance of Lasso and Elastic Net using stability-based selection and it was implemented in the R package with the name BioMark. Newman et al. [18] claimed that Elastic Net has an inflated Type I error compared to ANOVA. Elastic Net lacks of makes the interpretation of the estimates based on values of the original measurement challenging. Elastic Net also can be found in SECIMTools: a suite of metabolomics data analysis tools [15].

2.5 Workflow of variable selection methods for classification

Variable selection methods for classification can be categorised into filter, wrapper and embedded methods. This is illustrated in Figure 2.2. When using filter methods, the variables are ranked, and a number of top ranked variables can be included into a prediction model as the filter methods regard these variables as the most 'important' variables. Wrapper methods evaluate subsets of variables based on their classification performance, but a feature selection algorithm is not embedded in the processes as with embedded methods. Embedded methods tend to choose variables to be included into the model via penalization methods and the classification performance is calculated simultaneously. Therefore, embedded methods, in contrast to wrapper methods, do not separate the learning and the feature selection processes, they are part of the same procedure.

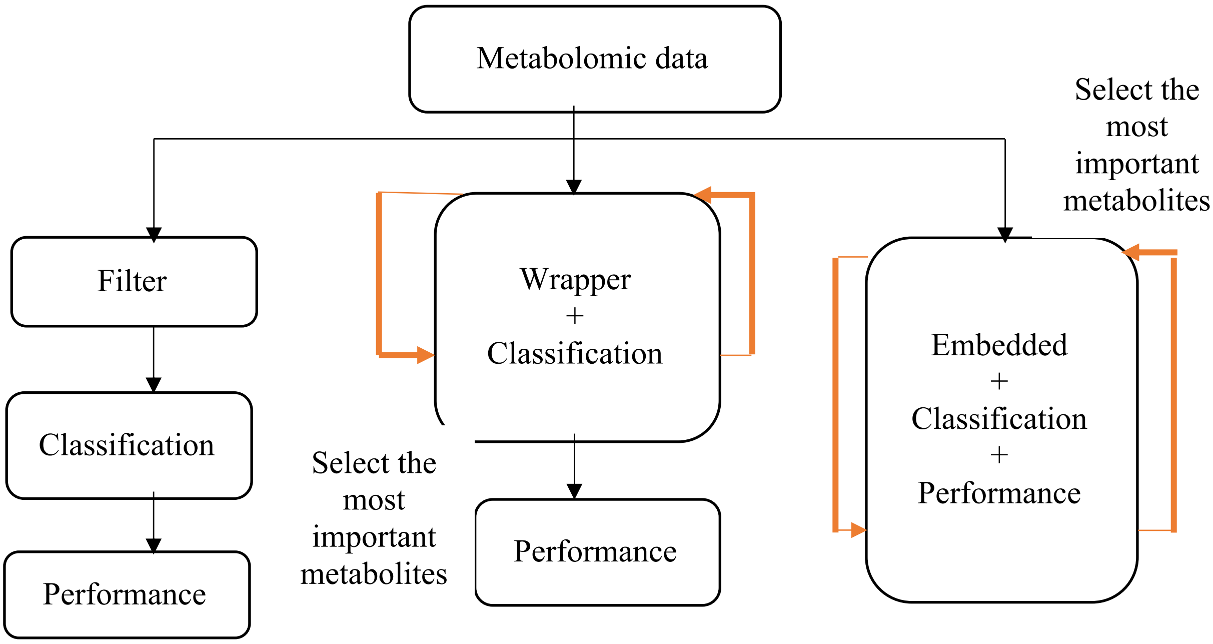

Metabolomic data
Select the most important metabolites
Filter
Wrapper + Classification
Embedded + Classification + Performance
Classification
Select the most important metabolites
Performance
Performance

Figure 2.2: Workflow of variable selection methods for classification applied to metabolomics data

In term of model performance, here is the description of different accuracy parameters used in the previous studies. Baratloo et. al. described a simple description of Accuracy, Sensitivity and Specificity in their paper [121]. They provided three simple examples in order to explain these accuracy parameters for easy understanding of the reader. In addition, Janecek et. al. [122] used classification accuracy in order to assess performances of Information Gain and Wrapper variable selection methods. As the results, the classification accuracy for Information Gain is better than Wrapper method as it offers simplicity. Additionally, Wah et. al. [1] compared Correlation-based Feature Selection, Information Gain, Sequential Forward Selection and Sequential Backward Elimination and they used six different accuracy parameters including the AIC, BIC, AUC, Accuracy, Sensitivity and Specificity. As we can see, Accuracy, Sensitivity, Specificity and AUC are commonly used in previous study. Hence, these parameters will be used in this thesis and will briefly explained in Chapter 3.

2.6 Other variable selection methods for metabolomics

A commonly used variable selection method, which is not regarded as a filter, wrapper or embedded method, is Principal Component Analysis (PCA). PCA is an unsupervised multivariate statistical technique which aims to capture most of the variation in the data in as few components as possible. It also aims to reveal the major patterns in the data. In other words, PCs focuses on data reduction and it can be used to summarize the similarities and differences between variables using the score plot which shows the amount of explained variance on each pair of PC. PCA requires the calculation of new variables, known as principal components (PCs) that are weighted linear combinations of the original variables. The computation of PCA is reduced to an eigenvalue-eigenvector problem. Firstly, an adjusted data matrix, X that consists of the data from n observations (rows) and p variables (columns) is defined. PCA deals with the covariances among the original variables, hence, means are irrelevant. The new variables or PCs are also known as factors. Their specific values on a specific row are known as the factor scores or the component scores. Equations and explanations below give a better understanding of calculation of PCA. The matrix of scores are referred as matrix Y and the basic equation of PCA is in Equation 2.7:

$$Y = W'X \tag{2.7}$$

where W is a matrix of coefficients that are determined by PCA. Equation 2.7 is also written as in Equation 2.8:

$$y_{ij} = w_{1i}x_{1j} + w_{2i}x_{2j} + +.....+w_{pi}x_{pj} \tag{2.8}$$

The factors are a weighted average of the original variables and when the weights, W, are generated, the variance of $y_1, Var(y_1)$ and $y_2, Var(y_2)$ are maximized. The implication is the correlation between y_1 and y_2 becomes zero. The remaining $y_i's$ are calculated so that their variances are maximized, with the constraint of the variance between $y_i$and y_j, for all i and j (i not equal to j), is zero. The weights, W is calculated from the variance-covariance matrix, S and it is calculated using the formula in Equation 2.9:

$$s_{ij} = \frac{\sum_{k=1}^{n}(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{n-1} \tag{2.9}$$

The singular value decomposition of S provides the solution to the PCA problem and it can be expressed as in Equation 2.10:

$$U'SU \ = L \tag{2.10}$$

where L is a diagonal matrix of the eigenvalues of S, and U is the matrix of eigenvectors of S. W is calculate from the L and U as follows in Equation 2.11:

$$W = UL^{-\frac{1}{2}} \tag{2.11}$$

PCA can be calculated using the correlation or variance-covariance. The formula for the correlation between an i^{th} factor and the j^{th} original variable is shown in Equation 2.12:

$$r_{ij} = \frac{u_{ij}\sqrt{l_i}}{s_{jj}} \tag{2.12}$$

where u_{ij} is an element of U, l_i is a diagonal element of L and s_{jj} is a diagonal element of S. The correlations are also known as the factor loadings. If the correlation is used instead of variance-covariance, the equation for Y is changed as follows in Equation 2.13:

$$Y = W'D^{-\frac{1}{2}}X \tag{2.13}$$

where D is a diagonal matrix made up of the diagonal elements of S. Hoyos Ossa et al. [77] used PCA to check the behaviour of the QC samples as a measure of technical variability. Additionally, López-Álvarez et al. [110] used PCA with three objectives such as to examine the structure of the data, to assess whether the observed groupings were consistent with the taxonomic circumscriptions proposed for three groups and to evaluate the level of covariation in variables.

One advantage of PCA is that it accounts for correlated variables. As mentioned previously, most of metabolomics data are highly correlated. Hence, it is a key step to reduce the number of variables before developing the classification tool. PCA is able to produce new uncorrelated variables, which explain most of the variability in the dataset. PCA can deal with normally distributed and non-normally distributed datasets. In addition, PCA may improve algorithm performance. With so many variables in the data, the computation of the algorithm might be slow. PCA is an efficient way to improve the efficiency of the machine learning algorithms by removing correlated variables that are irrelevant in any decision making. The risk of overfitting may be reduced with PCA by reducing the number of variables. It is difficult to visualise and understand high dimensional data. Principal components are linear combinations of the original variables. Hence, the independent PCs tend to be more difficult to interpret than the original variables, especially for high dimensional data such as metabolomics data. Although the first principal components tend to capture the maximum variance of the original variables, information loss may also occur depending on the number of principal components selected.

2.7 Classification methods of metabolomics

As we can see from Table 2.1, Partial Least Square – Discriminant Analysis (PLS-DA) is the most common classification method used in metabolomics. PLS-DA is a supervised linear classification model for the discrimination structure in the data using the VIP variable selection method. Given data with group labels it produces a set of latent variables (similar to PCA) that maximise correlation between variables and the labels. The risk of overfitting issue can be reduced imposing cross validation.

There were a number of researches that used PLS-DA in their studies. One of the studies was a study that analysed the stool samples as they believed that analysis of fecal metabolites can offer new possibilities of routine diagnostics of the helminths infections. They used a metabolomics dataset consisting of 30 stool samples and 429 variables in order to classify a sample into ‘no infection detected’ class or ‘infected’ class [53]. Kostidis et al. [53] argued that PCA failed to describe a clear pattern in the data, hence, the authors built PLS-DA and the model proved that there is a lack of association between the infection status and metabolic composition of the faeces. In addition, Djukovic et al. [30] used PLS-DA to classify colorectal cancer, polyps and healthy control groups and they proposed a backward variable elimination PLS-DA combined with Monte Carlo cross validation (MCCV-BVE-PLSDA), which are applied to a combination of NMR and MS variables. Other classification methods of metabolomics are k-nearest neighbours, discriminant analysis, support vector machine and random forest. The authors recommended to use MCCV-BVE-PLSDA as the variable selection step for biomarker discovery as it is straightforward and easy to implement compared to other methods. Even though SVM can be applied to metabolomics data, it may be computationally intensive compared to logistic regression [126]. SVM output class label is +1 and -1. It does not compute probabilities as the logistic regression does, which allow to assess ROC curves for model performance [56]. An additional limitation of SVM is the lack of test statistics, such as scores and loadings, available for easy visualization and interpretation [127]. KNN can be also computationally intensive and it is does not work well with high dimensional datasets, such as metabolomics data since it is difficult for the algorithm to calculate the distance in each dimension [125]. Random Forest can be time consuming when constructing the decision tree [129]. Neither SVM, KNN nor Random Forest were used as the classification method. I focused on logistic regression in order to produce a simpler equation, easy to use and to interpret. This will be described in Chapter 3.

2.8 Discussion

I have reviewed the literature related to variable selection methods in metabolomics. Metabolomics data often consist of a large number of correlated metabolites from a small number of samples. Hence, the variable selection process is an important step in metabolomics studies. The aim of variable selection is to determine the most significant and important metabolites that can discriminate between classes with minimum error and make an accurate prediction. Many variable selection methods used in previous studies were either filter, wrapper or embedded methods, with filter methods being the commonly used. More specifically, in metabolomics ANOVA is the most popular univariate filter method. It is easy to use and fast for selecting the most informative metabolites. However, it ignores the correlation among the metabolites and only considers each metabolite at a time. Additionally, ANOVA is unable to provide good results if the data do not meet the assumptions (i.e., same variance across groups and data normally distributed). PCA and PLS-DA are often used for variable selection methods and classifications respectively. On the other hand, wrapper methods are rarely used in metabolomics and only a few embedded methods have been applied.

This thesis considers the corT method (see Chapter 3). This method is a filter variable selection method that has been applied to genomic data and it takes into account the correlation among variables. However, this method can only be used to datasets that involve positive correlations. It can be applied to metabolomics data if the data having positive correlation. There will be a problem if the metabolomics data having negative correlation. Hence, to address this problem, an extension of this method namely adjusted corT (adjcorT) method is proposed in this thesis for the analysis of metabolomics area. None of the previous studies have applied this method to metabolomics data. As far as I know, none of the previous studies compared Lasso and corT even though both methods consider correlations among the variable. In this thesis I compare the performance of Lasso, corT and adjcorT as variable selection methods. Since corT is based on *t*-tests (T method), I also consider the T method for comparison purposes.

Chapter 3
Methodology

3.1 Introduction

In Chapter 2 I conducted a literature review of the variable selection methods commonly used in metabolomics. I identified limitations of the existing methods which motivated me to propose a new variable selection algorithm (adjcorT) and to compare the proposed method with already available variable selection methods. Hence, the goal of this chapter is to describe the rationale and application of adjcorT.

In order to fully identify the most important compounds in metabolomics datasets for clinical classification, and thus develop an appropriate model, a number of steps need to be followed. These include data pre-processing, variable selection, classification and assessment of the model performance. The performance of a model, which assess the ability of a model to provide accurate predictions, can be assessed in different ways. Model performance is in this thesis was assessed by using classification accuracy, sensitivity, specificity and Area Under ROC (AUC). Please see Section 2.5 in Chapter 2 for a description of the different accuracy parameters. The higher the model performance, the better the variable selection method is in selecting the most important compounds. Before any analysis is conducted, it is helpful to explore the structure of the dataset via a scatter plot in order to assess the correlation between variables and patterns of the data. In addition, data pre-processing is often conducted in order to deal with missing values and data scaling of the data before any further analysis is done. Methods that are used during this process (from data pre-processing to classification) are discussed in this chapter and explored in the subsequent chapters.

Section 3.2 describes the data pre-processing, which as described above, includes ways of dealing with missing values and data scaling. Several variable selection algorithms are described in Section 3.3, including adjusted correlation sharing t-statistics (adjcorT), least absolute shrinkage and selection operator (Lasso), t-statistics (T) and correlation sharing *t*-statistics (corT). Logistic regression, as a method for classification, is discussed in Section 3.4. Meanwhile, Section 3.5 describes commonly used measures to assess the classification performance of a classifier (for example, a classifier that is generated from a logistic regression model). This is followed by a discussion in Section 3.6.

3.2 Data pre-processing

Pre-processing involves the evaluation of missing values, checking for duplicate samples, assessment of noise in the data, and assessment of aspects related to the relationships within the dataset (e.g., which may indicate multicollinearity).

Data pre-processing may affect greatly the outcome of the analysis and different data pre-processing may generate different results.

3.2.1 Missing values

Having missing values is common with real datasets. Two methods known as *listwise deletion* and *pairwise deletion* are commonly used to handle missing values. In terms of clinical application, listwise deletion is a complete case analysis which simply removes those patients with the missing data and analyse the remaining data. However, this technique often produces bias in the estimation of parameters since it rarely supports the assumption of missing completely at random (MCAR). If the sample size of the data is large enough, and the assumption of MCAR is satisfied, listwise deletion may be a reasonable technique to handle missing data. Pairwise deletion, on the other hand, removes certain data points only when those data points are missing. All remaining existing information are used in the statistical analysis. This technique preserves more information as it uses all information observed. However, if most of the entire variables consist of a large number of missing values, these methods are not appropriate because the researcher may lose important information of the dataset, especially when the sample size is small. Imputation methods provide an alternative way to deal with missing values. In general, different imputation methods, especially when the rate of missing data is large, will affect the accuracy of the classifier differently. Multiple imputation and other sophisticated imputation methods are often preferable in real data application. Multiple imputation may minimise bias and increase the level of precision of the estimates of the model parameters [130]. Additional imputation methods can be used, such as k nearest neighbors (KNN). KNN can be used for datasets that are continuous, discrete, ordinal and categorical, which makes it useful for dealing with all type of missing data jointly [131]. KNN requires the selection of the number of nearest neighbours and a distance metric. KNN is easy to implement, however, this method is computationally expensive and is very sensitive to outliers. It is not often used with high dimensional data as it becomes difficult to calculate the distance in each dimension [129]. As metabolomics datasets are highly dimensional, imputation using KNN is has not been considered in this thesis [130]. Other imputation method, such as regression imputation, have also its limitations [40],[134].

Wei et al. compared eight imputation methods (zero, half minimum, mean, median, random forest, Singular Value Decomposition (SVD), k nearest neighbors and quantile regression imputation), which were applied to metabolomics data. They used four real metabolomics datasets containing different missing values scenarios in order to compare these eight imputation methods. Normalized root mean squared error (NRMSE) was used to evaluate their performance. The authors found that mean imputation was an acceptable method to use when tackling the missing values in metabolomics data [39]. The authors also displayed the results of SVD imputation applied to metabolomics data and showed that this method did not perform well. Based on the results, for the dataset with a proportion of missing values of 0.3, the NRMSE for mean imputation was 1.0 and for KNN imputation was 1.2, which do not greatly differed.

Simple imputation methods such as the mean (applied in this thesis) and median imputations allow to maintain the sample size and are easy to use. Other imputation methods could be explored in order to minimise the bias introduced by missingness.

3.2.2 Data scaling

After applying mean imputation, the datasets are scaled with the aim of limiting the range of the variables so that they can be easily compared. The scaling methods are based on the data dispersion or size measure [131]. Autoscaling, range scaling and pareto scaling are the scaling methods based on data dispersion such as standard deviation. Meanwhile, level scaling is a scaling method that uses a size measure such as the mean. This thesis uses the autoscaling method which uses the standard deviation as the scaling factor. The advantage of using this method in metabolomics is that all metabolites are given the same importance. In the clinical applications, once missing values were imputed by the mean the data were scaled using the scale function in R.

3.3 Variable selection algorithms

This thesis focuses on the performances of variable selection methods in the area of metabolomics in order to tackle the limitation of the existing variable selection methods (some methods are ignoring correlations among variables and some other methods are only considers positive correlations among variables), in this section I propose a new method for variable selection named *adjcorT*. I also explore three existing variable selection methods: Lasso, corT and T methods.

3.3.1 Variable selection algorithm: *adjcorT*

The method adjcorT is an extension of the correlation sharing t-statistics (corT), which is described in Section 3.3.4. Metabolomics data might exhibit both positive and negative correlations. The aim of using adjcorT is to identify important biomarkers in metabolomics data while allowing for both negative and positive correlation among biomarkers. This method examines the correlation between biomarkers from -1 to 1, and therefore it includes both negative and positive correlations as opposed to corT, which only considers positive correlation among predictors (corT is explained in Section 3.3.4). CorT was applied to genomic data in the previous study [136] and, to my knowledge, no study has applied corT to metabolomics data. Tibshirani R. & Wasserman L. [136] applied corT to four datasets, of which all datasets are highly dimensional. Based on the results, corT performed well for all datasets, where corT often exhibits lower false discovery rates than the simple t-test. However, corT do not consider negative correlation in the algorithm. Hence, an extension of corT is proposed by adding the algorithm that finds both positive and negative correlations while searching the important biomarkers.

Let X be a matrix that consist of p x n of expression values, for p variables and n samples. I assume that the samples fall into two groups $j = 0$ and 1 (e.g., disease, non-disease). I start with the standard (unpaired) t-statistics as shown in Equation 3.1:

$$T_i = \frac{\bar{x}_{i1} - \bar{x}_{i0}}{s_i} \tag{3.1}$$

Here $\bar{x}_{ij}$ is the mean of the i-th variable $(i = 1,2, \ldots\ldots, p)$ in group 0 or 1 and s_i is the pooled within-group standard deviation of the i-th variable. Let x_i denote the i^{th} row of X. For each variable i, I define the set $C_\rho(i) = \{k: |corr(x_i, x_k)| \geq \rho\}$ where $\rho \geq 0$, which is the set of the indices of the variables with correlation (absolute value) equal or larger than ρ with variable x_i. Additionally, w is the cardinal of the set $C_\rho(i)$. For example, if $C_\rho(1) = \{x_1, x_{40}, x_{120}\}$ then $w = 3$. I define u_i and r_i in Equation 3.2 and Equation 3.3 respectively:

$$u_i = max_{(0 \leq \rho \leq 1)} \frac{1}{w} \sum_{j \in C_\rho(i)} |T_j| \tag{3.2}$$

$$r_i = sign(T_i) \,.\, u_i \tag{3.3}$$

where, max is the maximum. In addition, each variable is assigned a score r_i which equals to the average of all t-statistics for variables having correlation (absolute) at least ρ with variable i, choosing the best value of ρ to maximize the average. AdjcorT is a filter method and it is easy to use. This method is only applicable for continuous variables as it calculates the t-scores of each variable and assesses the correlation with the other independent variables.

3.3.2 Lasso

In this thesis, the selection method Lasso was used and the results were compared with the adjcorT method as Lasso is one of the most commonly used in metabolomics area. Lasso was introduced in geophysics literature in 1986 and later rediscovered by Robert Tibshirani in 1996 [133]. Briefly, Lasso stands for Least Absolute Shrinkage and Selection Operator and it is a regression method that includes a penalty term during the variable selection process to increase the classification accuracy with minimum error. In other words, Lasso aims at reducing variance in models that contain a large number of useless variables. This makes the final model simpler and easier to interpret. Lasso penalises the absolute values of the model coefficients, known as L_1 penalty term and which can be expressed as $\sum_{j=1}^{p}|\beta_j|$. The mathematical equation of Lasso can be expressed by the Equation 3.4:

$$\hat{\beta}_{Lasso} = \underset{\beta}{minimize}\left\{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\} \tag{3.4}$$

subject to

$$\sum_{j=1}^{p}|\beta_j| \leq t \tag{3.5}$$

where t is the upper bound of the summation of the absolute coefficients, λ is the tuning parameter that controls the strength of the penalty [134]. λ is often set using cross-validation in the cv.glmnet package in R software. The larger λ, the larger the amount of shrinkage. When the λ is equal to zero, the Lasso will produce the classical the least square coeficients (the penalty term has no effect). When we increase the λ value, the coefficients of informative variables will shrink a little bit and the non-informative variables will go all the way to zero. If λ becomes ∞, all the coefficients will be equal to zero. There is reverse relationship between λ and t. That means, t becomes 0 as λ goes to ∞. Vice versa, t becomes ∞ as λ goes to 0. Lasso will be computed as a quadratic programming problem. The ten-folds cross-validation approach is used as the default approach in cv.glmnet package.

Figure 3.1 illustrates the estimation process for Lasso. It shows the contours of the error (in red) and the constraint regions (solid blue area; eg: $|\beta_1| + |\beta_2| \leq t$). Lasso finds the first point where the least square regression border touches the constraint region. Based on the figure, the constraint region for Lasso is a diamond, it has corners. It means that if the first point touches the corner, then that coefficient β_j equal to 0.

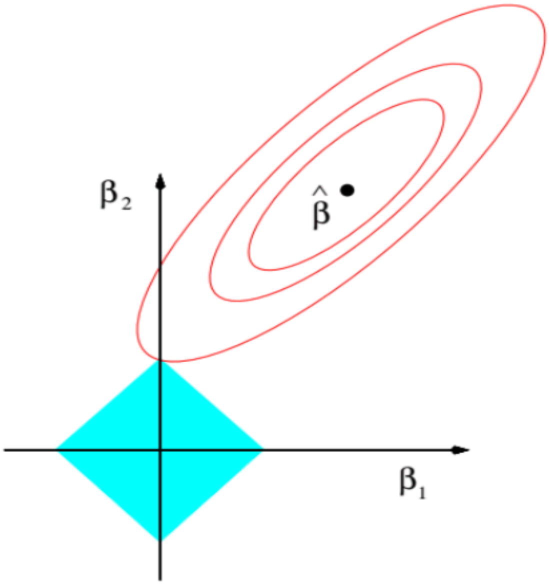


Figure 3.1: Estimation process for the Lasso approach

Source: Hastie, Tibshirani, and Friedman, The elements of statistical learning, 2009

Commonly, Lasso is used in high dimensional data where the number of variables is larger than the sample size. High dimensional data is usually costly and that is why the sample size is relatively small.

Lasso has a number of limitations. First, for high dimensional data with a large number of covariates (p) and small sample size (n), Lasso tends to select at most n variables before it saturates. Secondly, if there are correlated variables, Lasso tends to select one variable and ignore the other variables in that correlated group. One of the methods that can overcome Lasso limitation is Elastic Net method which is a hybridisation of ridge regression and Lasso. However, I used Lasso in this thesis since the computational cost of Elastic Net is expensive due to L_1 and L_2 penalty terms.

3.3.3 T-test Feature Selection Method (the T-method)

The *t*-statistics was introduced by William Sealy Gosset in 1908, a chemist working in Dublin, Ireland. Specifically, this research used the idea of independent sample *t*-test. *T*-test assumes the cases are independent of each other: an inaccurate p-values will occur if the assumption is violated. Furthermore, the next assumption is it should be random samples of the data from the population. The *t*-test assumes homogeneity of variances across groups and no outliers in the data.

The T method is a method of variable ranking that uses the *t* score. As defined earlier, let X be the p x n matrix of expression values, for p variables and n samples. I assume that the samples fall into two groups $j = 0$ and 1. I start with t standard (unpaired) t-statistics for each variable as shown in Equation 3.6:

$$T_i = \frac{\bar{x}_{i1} - \bar{x}_{i0}}{s_i} \text{ where } i = 1,2,\ldots.,p \tag{3.6}$$

Here $\bar{x}_{ij}$ is the mean of variable x_i in group 0 or 1 and s_i is the pooled within group standard deviation of variable x_i. Mathematically, $\bar{x}_{ij}$can be calculated as in Equation 3.7:

$$\bar{x}_{ij} = \sum_{i=1}^{n} \frac{x_{ij}}{n_j} \tag{3.7}$$

Meanwhile, s_i can be calculated using the following Equation 3.8:

$$s_i = \sqrt{\left(\frac{1}{n_0} + \frac{1}{n_1}\right) * s_i^2} \tag{3.8}$$

The strength of using the T method is that it can be used to discriminate between the two groups based on the t scores, simply by estimating the differences in means between the two groups divided by its standard error. No difference in means between the two groups would imply that the variable does not offer any degree of discrimination between the groups. To apply the T method, I used an adjusted version of the cst.stat function from the st package in R environment since it can calculate the *t*-scores for each variable.

3.3.4 corT method

corT stands for *correlation sharing t-statistics*. CorT method was proposed by Tibshirani and Wasserman in 2008 [132]. CorT method is a method of variable ranking that takes into account the t scores as well as the correlation among the variables.

corT uses the same procedure to adjcorT. As mentioned in section 3.3.1, the main different between the corT and adjcorT is the set $C_\rho(i) = \{k: corr(x_i, x_k) \geq \rho\}$ where $\rho \geq 0$, which is the set of the indices of the variables with correlation equal or larger than ρ with variable x_i. Additionally, this method only considers positive correlation among predictors. It does not account for negative correlations. The limitation of this method is it not design to select the correct discriminators when there is negative correlation between discriminators. The corT method uses the function cst.stat. Limitation of corT is that it was design only for continuous covariates and it has a slow computational speed for large sample sizes.

In this thesis I applied corT to metabolomics data. To my knowledge, no previous study in the area of metabolomics has applied corT to metabolomics data. CorT can be regarded as an extension of the T method. CorT was used in this thesis to compare its performance with the performance of the T method and assess whether there is any improvement when the correlation is taken into account. In addition, I was interested in comparing corT with the proposed method adjcorT to explore their performance when dealing with negative correlations among variables.

3.4 Logistic regression for classification methods

Logistic modelling can be used to develop a classification rule. Logistic regression was developed by David Cox in 1958. Logistic regression can be regarded as an extension of simple linear regression when the dependent variable is dichotomous or binary. Logistic regression was here used to model the relationship between binary outcome variables (eg: disease and non-disease groups) and predictor or explanatory variables. The predictor variables can be categorical (nominal/ordinal) or continuous (interval/ratio). Both simple linear regression and logistic regression aim to find the best fitting model and the most parsimonious model. Logistic regression generates probabilities, which can be used to classify new patients using continuous and discrete measurements.

Logistic regression was chosen as the classification method since this is a simple approach and it is commonly used in the metabolomics literature for classification. Logistic regression can be used for parameter estimation, prediction and classification. L1/L2 regularization can be also used, which means that Lasso or ridge regression can be incorporated. Logistic regression also tends to provide a simple equation compared to other classification methods. When the independent variables don't satisfy these assumptions, the estimates of the coefficients and standard errors might be large. Consequently, the confidence intervals tend to be wider and fail to detect truly statistically significant differences. Meanwhile, when the assumptions of the logistic regression classifier do not hold (such as the independent variables are not linearly related to the log of odds, the dependent variable is not binary and the observations are dependent to each other), the model interpretations might also be not valid [139].

Other assumptions also apply. Logistic regression assumes that the observations are independent of each other. Strictly speaking, the observations should not come **from** repeated measurements. It also assumes that there is little or no multicollinearity among the predictors. The predictors are assumed to have a linear association with the log odds. In addition, logistic regression requires a large sample size. However, it does not mean that the analysis is wrong if the sample size is small, but if the ratio of number of samples per number of predictors is low, the estimates might be biased.

Let Y be an outcome and x be the independent variables. Mathematically, $\pi(x) = E(Y|x)$ was used in logistic regression to represent the conditional mean of Y

given x. The mathematical equation for a logistic regression model reads as in Equation 3.9:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots\ldots\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots\ldots\beta_n x_n}} \tag{3.9}$$

where $\beta_0 + \beta_1 + \beta_2 \ldots . \beta_n$ is the linear predictors ($\beta_0, \beta_1, \beta_2, \ldots \beta_n$ to be predicted). A logit transformation represents the transformation of $\pi(x)$. The equation is in Equation 3.10:

$$\begin{aligned} g(x) &= ln\left[\frac{\pi(x)}{1-\pi(x)}\right] \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots . \beta_n x_n \end{aligned} \tag{3.10}$$

$g(x)$ shows linearity with predictors, which is one of the assumptions in logistic regression. The method of maximum likelihood yields values for the unknown parameters ($\boldsymbol{\beta} = \beta_0, \beta_1, \beta_2, \ldots \beta_n$), which maximises the probability of obtaining the observed set of data given the model. Likelihood function needs to be constructed in order to get the maximum likelihood. Then, the maximum likelihood is chosen as it is maximise the likelihood function. Generally, the likelihood function represents the probability of the observed data as a function of the unknown parameters.

Let n be number of the independent observations of the pair (x_i, y_i), $i = 1{,}2 \ldots . ., n$, where y_i is the value of a binary outcome variable and x_i is the vector of value of the predictors for the i^{th} observation. The conditional probability that Y is equal to 1 given x is denoted as $P(Y = 1|x)$. Therefore, $1 - \pi(x)$ gives the conditional probability that Y is equal to 0 given x, i.e., $P(Y = 0|x)$. The likelihood function can be expressed as follows in Equation 3.11:

$$\pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i} \tag{3.11}$$

Any pair of (x_i, y_i) where $y_i = 1$ contributes to the likelihood function of $\pi(x)$ and any pair of (x_i, y_i) where $y_i = 0$ contributes to the likelihood function of $1 -$

$\pi(x)$. As mentioned above, one of the assumptions of logistic regression is the observations need to be independent of one another, Under independence, the likelihood function can be expressed as the following product in Equation 3.12:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i} \tag{3.12}$$

The log likelihood can be defined as in Equation 3.13:

$$L(\boldsymbol{\beta}) = \ln[l(\beta)] = \sum_{i=1}^{n} \{y_i ln[\pi(x_i)] + (1 - y_i) ln[1 - \pi(x_i)]\} \tag{3.13}$$

In order to find the value of $\boldsymbol{\beta}$ that maximises $L(\boldsymbol{\beta})$, the $L(\boldsymbol{\beta})$ is differentiated with the respect to β_0 and β_1 and set the equation equal to zero as in Equation 3.14 and Equation 3.15:

$$\sum [y_i - \pi(x_i)] = 0 \tag{3.14}$$

and

$$\sum x_i[y_i - \pi(x_i)] = 0 \tag{3.15}$$

The maximum likelihood estimates, $\hat{\boldsymbol{\beta}}$ will be obtained from Equation 3.14 and Equation 3.15. "^" denotes the maximum likelihood estimate of the respective parameter. It represents the fitted value of the parameters in logistic regression model. Consequently:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i) \tag{3.16}$$

The sum of the observed values of y is equal to the sum of the fitted values. The glm function from the R environment was used to generate the logistic regression models. The decision boundary for a logistic regression is linear. Logistic regression was chosen as the classification methods to the investigate the performance of the T, corT, adjcorT and Lasso methods when this classification methods.

3.5 Classification performance of a classifier

The performance of a variable selection methods is based on the classification accuracy, Area Under Receiver operator curve (AUC), as well as sensitivity and specificity of the classifier. The accuracy parameters can be described using the two-by-two confusion matrix (Table 3.1) [92], [112], [136], [137].

Table 3.1: An outline of the two-by-two confusion matrix

		Actual Condition	
		Positive	Negative
Predicted Condition	Positive	True positive TP	False positive FP
	Negative	False negative FN	True negative TN

Definitions:

TP = Number of positive samples that are correctly classified as positive

FP = Number of negative samples that are incorrectly classified as positive

TN = Number of negative samples that are correctly classified as negative

FN = Number of positive samples that are incorrectly classified as negative

The accuracy of a test measures its ability to discriminate samples correctly into positive and negative. The proportion of true positive and true negative from all evaluated samples was calculated to estimate get the accuracy parameters. The function misClassError from the package Information Value package in the R environment was used to calculate the classification accuracy using:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{3.17}$$

Sensitivity is defined as the proportion of true positive samples that are correctly classified:

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.18}$$

Specificity is defined as the proportion of negative samples that are correctly classified:

$$Specificty = \frac{TN}{TN + FP} \tag{3.19}$$

ROC curve is a plot of the true positive fraction (sensitivity) on the y-axis and false-positive fraction (1-specificity) on the $x - axis$ used to evaluate the classifier. Figure 3.2 is showing an example of ROC curve. The area under the curve (AUC) of the ROC can take values between 0 to 1. AUC = 1, correspond to a perfect prediction, while AUC = 0.5 means a random guess [138]. The higher the AUC value, the better the classifier. Additionally, the closer the curve is to the upper left corner, the larger the area under the curve. The functions sensitivity and specificity from the package Information Value package [139] in the R environment were used to calculate sensitivity and specificity, respectively. The AUC value was calculated using the function AUROC.

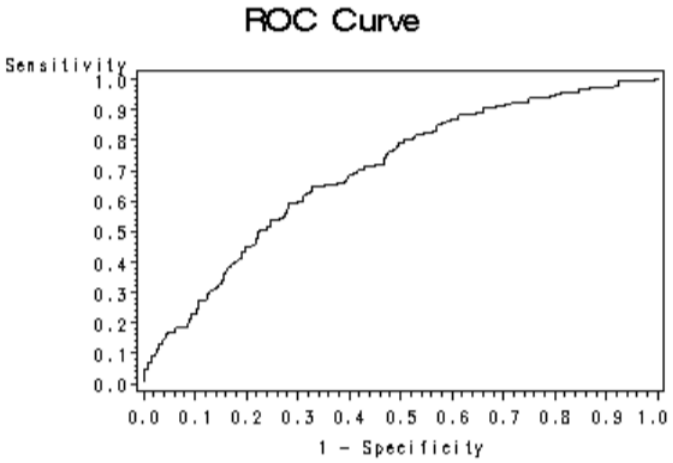


Figure 3.2: An example of ROC curve

Source: Hosmer & Lemeshow, Applied logistic regression, 2001

3.6 Discussion

I started this chapter by describing how I have handled the missing values in the metabolomics datasets and the data pre-processing that I have followed. Then I introduced the main statistical methods used in this thesis: the method that I proposed, adjcorT, and three existing methods: (T, corT and the Lasso). T, corT and adjcorT methods are based on t-statistics. The method adjcorT is a filter method, it is easy to use and to understand, takes into account both positive and negative correlations. adjcorT is however slow computationally for large sample sizes and it is only can be applied to continuous variables.

In the next chapter I will apply T, corT, adjcorT and Lasso for variable selection using simulated data that are generated from a multivariate normal distribution.

Chapter 4

Comparison the performance of T, corT, adjcorT and Lasso: A simulation study

4.1 Introduction

This chapter evaluated the classification accuracy of T, corT, adjcorT and Lasso using the simulated datasets. A range of sample sizes and several correlation values among some of the variables were applied in order to explore their performance.

The simulated datasets involved variables with some level of discriminatory ability and non-discriminators in order to assess whether the T, corT, adjcorT and Lasso methods were able to capture the discriminatory variables correctly and/or additional non-important variables. The classification performance was subsequently studied using logistic regression modelling.

This chapter is structured as follows. In Section 4.2 I explain how the simulated data were generated. The performance of T, corT, adjcorT and Lasso is reported in Section 4.3 and the discussion can be found in Section 4.4.

4.2 Simulated data

The simulation study presented in this chapter aimed to (i) assess what variables the methods T, corT, adjcorT and Lasso are able to select (i.e., discriminatory and/or non-discriminatory variables) and (ii) to investigate the effect that both the sample size and underlying correlation among variables have on the methods performances.

The choice of model in this simulation study was chosen as a simple way to assess the effect that the relationship between variables have on various sample sizes (*n*=50, 76, 100, 300, 500, 1000, 2000 and 20000) and correlations (ρ=-0.8, -0.5, -0.2, 0, 0.2, 0.5 and 0.8). I assumed that the variables were normally distributed continuous variables to follow a simple, easy to interpret and well documented distribution. The variables were simulated using the multivariate normal distribution (*mvrnorm* function in R), and were generated using the following multivariate density function as in Equation 4.1:

$$f(x) = f(x_1, x_2).f(x_3) \cdots f(x_{200}) \tag{4.1}$$

$$= \frac{1}{(2\pi)^{200/2}\left|\boldsymbol{\Sigma}_{\mathbf{1},2}\right|^{1/2}} e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_{\mathbf{z}})\,\boldsymbol{\Sigma}_{\mathbf{1},2}{}^{-1}(\mathbf{z}-\boldsymbol{\mu}_{\mathbf{z}})'}$$

$$\cdot f(x_3) \cdots f(x_{200})$$

where the corresponding parameters are defined as follows:

x=$[x_1, x_2, x_3, \dots, x_{200}]$

z=$[x_1, x_2]$

$\boldsymbol{\mu}$ =$[\mu_1, \mu_2, \mu_3, \dots, \mu_{200}]$

$\boldsymbol{\mu_z}$ =$[\mu_1, \mu_2]$

$$\boldsymbol{\Sigma}_{\mathbf{1},2} = \begin{bmatrix} \mathrm{Var}_1 & \mathrm{Cov}_{1,2} \\ \mathrm{Cov}_{1,2} & \mathrm{Var}_2 \end{bmatrix} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

Variables x_1 and x_2 were set to be the discriminators as they are having different means between the group 0 and group 1 (Table 4.1). The remaining variables

$(x_3, \dots, x_{200})$ were set to be non-discriminators and therefore have the same means across the two groups. The variances of all variables were set as 1 for both groups.

I considered the same sample size for each group (i.e., $n_0 = n_1$ and $n_0 + n_1 = n$). The sample sizes considered were n= 50, 76, 100, 300 500, 1000, 2000 and 20000 in order to cover small, medium and large sample sizes. For example, when the sample size is n= 50, the sample size for each group is $n_0 = n_1 = 25$. In Figures 4.1-4.4 I illustrated a number of scenarios applied to different sample sizes including sample sizes of 50, 100, 2000 and 20000. A total of 100 iterations (i.e., 100 multivariate random samples) were considered for the analyses. I explored 7 scenarios of correlations as described in Table 4.1.

Table 4.1: Population means and correlations considered

Population means considered for groups 0 and 1	
$\boldsymbol{\mu_1} = 0$ for group 0 and $\boldsymbol{\mu_1} = 0.5$ for group 1	
$\boldsymbol{\mu_2} = 0$ for group 0 and $\boldsymbol{\mu_2} = 1$ for group 1	
$\boldsymbol{\mu_i} = 0 \quad i = 3,4, \dots, 200$ for groups 0 and 1	
Scenario no.	Correlation value between x_1 and x_2
1	$\rho = 0.8$
2	$\rho = 0.5$
3	$\rho = 0.2$
4	$\rho = 0$
5	$\rho = -0.2$
6	$\rho = -0.5$
7	$\rho = -0.8$

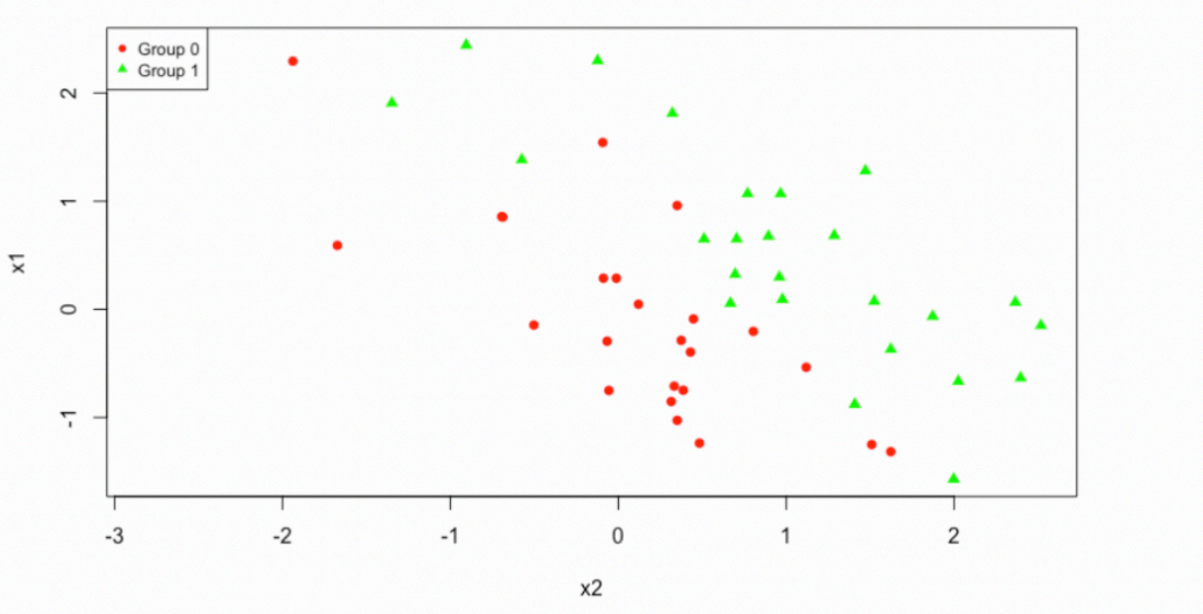


Figure 4.1: Simulated data with $n = 50$ and scenario 7, ρ = -0.8.

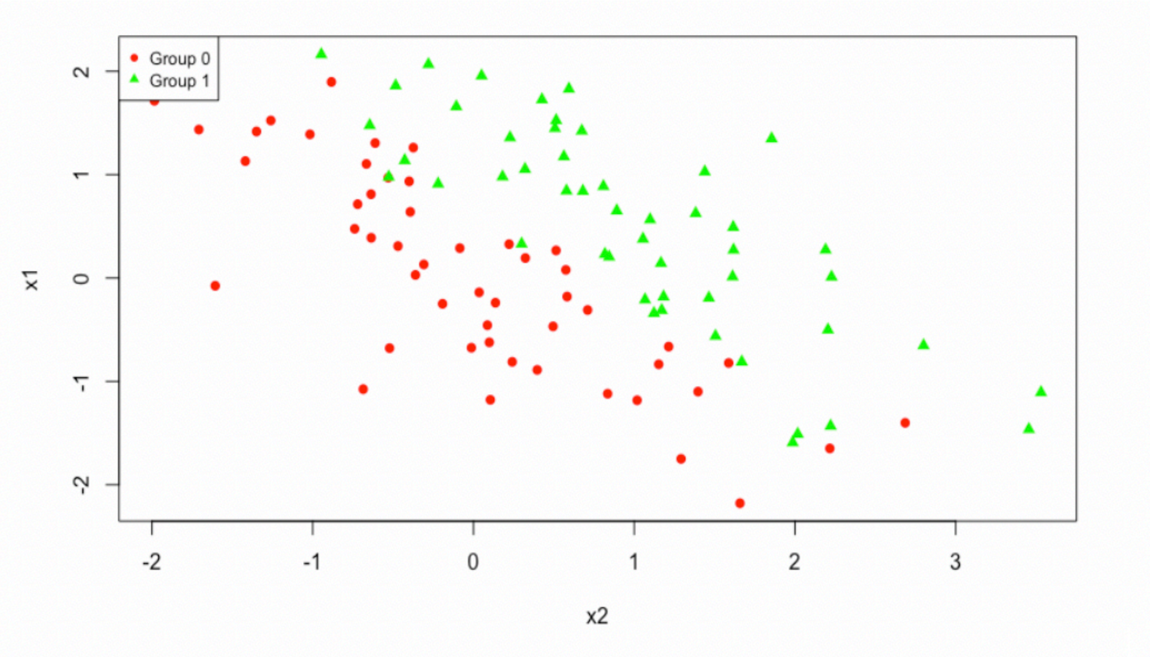


Figure 4.2: Simulated data with $n = 100$ and scenario 7, ρ = -0.8.

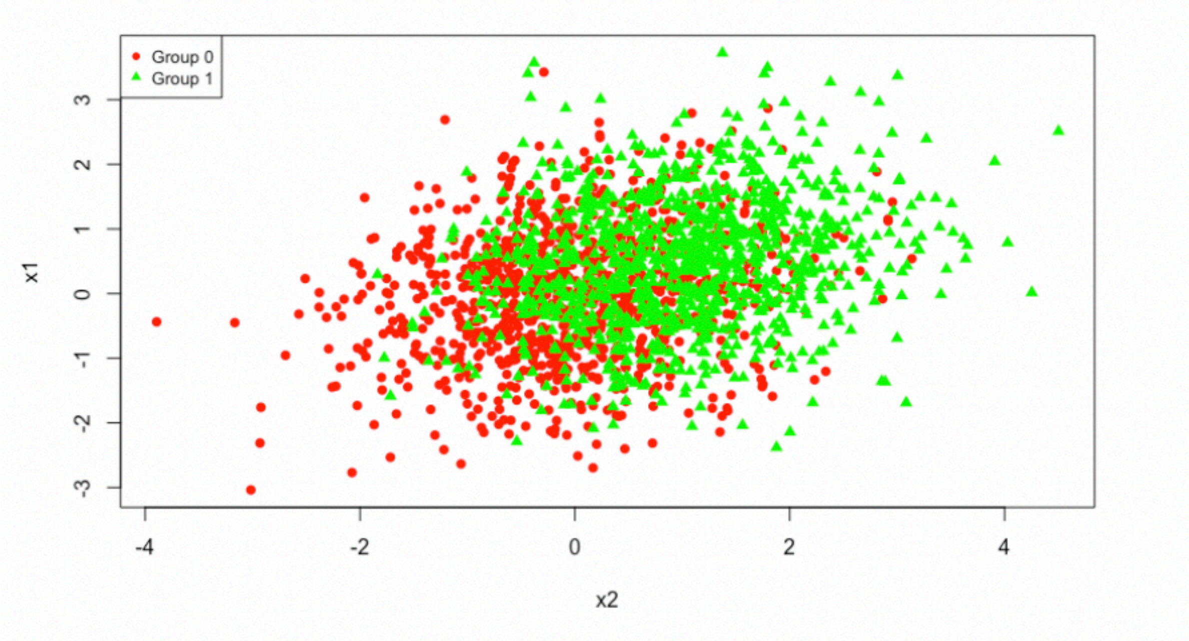


Figure 4.3: Simulated data with $n = 2000$ and scenario 3, ρ = 0.2.

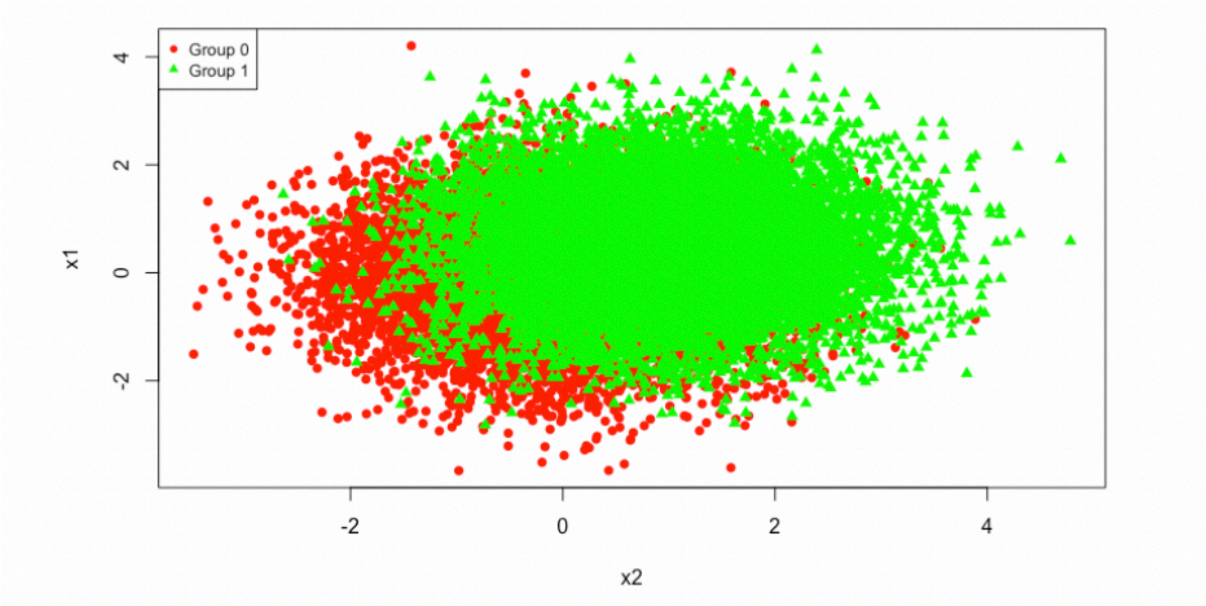


Figure 4.4 : Simulated data with $n = 20000$ and scenario 4, ρ = 0.

As mentioned in Section 4.1, this simulation study investigated whether the T, corT, adjcorT and Lasso variable selection methods are able to select the correct variables that will be subsequently used for classification. Information on how often the discriminatory variables were selected based on the 100 runs for each of the variable selection methods is presented in Tables 4.2 and 4.3. In addition, the estimates of classification accuracy, sensitivity, specificity and AUC for average performances were calculated and these are presented in Tables 4.4-4.7 (Section 4.3.2). The histograms of AUC generated for sample size n=50 are displayed in Section 4.3.3. Other histograms for the sample size of 76 and 300 were displayed in the Appendices. The sample size of 50 and 76 represent the small sample size and the sample size of 300 represents the large sample size.

Interval validation was undertaken with the aim to reducing the effect of overfitting on the estimates of the accuracy parameters. Each simulated data was partitioned into two sets namely training data (80%) for variable selection and for building of the logistic regression model and testing data (20%) for internal assessment of the accuracy of the model. For example, with n=50, the number of samples for training is 40 and the number of samples for testing is 10. However, bootstrapping can also be used to resample simulated data for small sample size. It has a number of advantages: 1) it has an equal probability of randomly drawing each original data point to be included into the resampled data, 2) it can select a data point more than once in order to resample data as long as the property of "with replacement" is being used, and 3) The same size of the original data is reproduced. Future research may consider this method [146].

4.3 Results

This section shows the performance of T, corT, adjcorT and Lasso for a range of sample sizes (n=50, 76, 100, 300, 500, 1000, 2000 and 20000) and a range of correlation values between $x_1$and x_2 ($\rho =$ 0.8, 0.5, 0.2, 0, -0.2, -0.5 and -0.8). The performances of these methods measured using classification accuracy, sensitivity, specificity and area under ROC curve (AUC) were displayed in percentages.

4.3.1 Selection of variables by each method for different sample size and correlations between the discriminatory variables.

Table 4.2 and Table 4.3 show how often the discriminatory variables were selected by each variable selection method based on 100 iterations for sample sizes 50, 76, 100, 300, 1000, 2000 and 20000. Specifically, Table 4.2 reports the results for zero and positive correlations and Table 4.3 for negative correlations. For the smallest sample size n=50, one or two discriminatory variables ($x_1$and x_2) were selected by T, corT and adjcorT in most cases (between 9% and 14% of the times none of the two variables were selected). Lasso, on the other hand, was not able to identify any of the discriminatory variables in 12%-30% of the times when the sample size was equal to 50. Lasso selected x_2 as the unique variable in a number of occasions (between 13% and 28% for sample size of 50) as opposed to T, corT and adjcorT, which selected additional variables (most of the times non-discriminatory variables) together with x_2 in the majority of the cases.

For the sample size 76, as the correlation changed towards -0.8 Lasso selected $x_1$and x_2 more often (for correlation -0.8, $x_1$and x_2 were selected 57% of the times). T, corT and adjcorT mainly selected x_2 together with other non-discriminatory variables, reaching 90% for correlation 0.2, 0, -0.2, -0.5 and -0.8 for T and corT.

For sample sizes of 100 and 300, all methods were able to select at least one of the discriminatory variables in all iterations (either “x1 and x2” or “x2 and others “or “x2 only”). For sample size 100, corT and adjcorT selected both $x_1$and x_2 in most iterations (89% and 90%) when the correlation was highly positive (0.8, 0.5). For low correlations (between 0.2 and -0.2) both discriminatory variables were not often selected. For highly negative correlations (-0.5 and -0.8) however, only corT showed an improvement. Lasso selected both discriminatory variables in the majority of the cases for negative correlation.

For $n = 300$, the methods T, corT and adjcorT outperformed Lasso for moderate and highly positive correlation. Lasso selected both discriminatory variables in 98% and 100% of the times when the correlation was zero and negative, respectively. For non-positive correlations, corT became the worst method with only 37% to 45% of the times selecting both discriminatory variables ($x_1$and x_2). Lasso selected both discriminatory variables ($x_1$and x_2) in 100% of the times for negative correlation datasets.

Table 4.2: Selected variables by each method (out of 100 iterations) for sample sizes n =50, 76, 100 and 300, and $\rho \geq 0$.

Sample sizes (n)		$\boldsymbol{n}$ = 50				$\boldsymbol{n}$ = 76				$\boldsymbol{n}$ = 100				$\boldsymbol{n}$ = 300			
Correlation between x1 and x2	**Selected variables**	**Variable selection methods**															
		T	**corT**	**adjcorT**	**Lasso**	**T**	**corT**	**adjcorT**	**Lasso**	**T**	**corT**	**adjcorT**	**Lasso**	**T**	**corT**	**adjcorT**	**Lasso**
0.8	x2 only				28				33				39				34
	x1 and x2	3	17	17		11	47	48		28	90	89		98	100	100	1
	x1 and others																
	x2 and others	87	73	73	49	89	53	52	63	72	10	11	61	2			65
	Neither x1 nor x2	10	10	10	7												
	No selected variable				16				4								
0.5	x2 only				21				35				39				33
	x1 and x2	3	12	11		10	46	47		27	90	90	2	98	100	100	
	x1 and others																
	x2 and others	84	75	76	50	90	54	53	63	73	10	10	59	2			67
	Neither x1 nor x2	13	13	13	7												
	No selected variable				22				2								
0.2	x2 only				16				27				39				16
	x1 and x2	3	3	3	6	9	9	7	6	30	25	18	11	98	92	87	73
	x1 and others																
	x2 and others	83	83	83	48	91	91	93	59	70	75	82	50	2	8	13	11
	Neither x1 nor x2	14	14	14	7												
	No selected variable				23				8								
0	x2 only				28				30				34				2
	x1 and x2	3	3	3	4	8	8	7	17	29	19	19	30	99	45	71	98
	x1 and others	1	1	1													
	x2 and others	87	87	87	44	92	92	93	51	71	81	81	36	1	55	29	
	Neither x1 nor x2	9	9	9	8												
	No selected variable				16				2								

Table 4.3: Selected variables by each method (out of 100 iterations) for sample sizes n=50, 76, 100 and 300, and $\rho < 0$

Sample sizes (n)		$\boldsymbol{n}$ = 50				$\boldsymbol{n}$ = 76				$\boldsymbol{n}$ = 100				$\boldsymbol{n}$ = 300			
Correlation between x1 and x2	**Selected variables**	**Variable selection methods**															
		T	**corT**	**adjcorT**	**Lasso**	**T**	**corT**	**adjcorT**	**Lasso**	**T**	**corT**	**adjcorT**	**Lasso**	**T**	**corT**	**adjcorT**	**Lasso**
-0.2	x2 only				23				29				20				
	x1 and x2	5	5	5	12	7	6	6	29	27	17	19	70	97	43	89	100
	x1 and others																
	x2 and others	87	87	87	40	93	94	94	39	73	83	81	10	3	57	11	
	Neither x1 nor x2	8	8	8	7												
	No selected variable				18				3								
-0.5	x2 only				24				20				2				
	x1 and x2	4	4	12	19	7	7	16	64	27	15	89	98	96	42	100	100
	x1 and others																
	x2 and others	86	86	78	34	93	93	84	16	73	85	11		4	58		
	Neither x1 nor x2	10	10	10	7												
	No selected variable				16												
-0.8	x2 only				13				1								
	x1 and x2	2	2	14	57	7	7	65	99	25	16	90	100	96	37	100	100
	x1 and others	2	2	2													
	x2 and others	89	89	77	18	93	93	35		75	84	10		4	63		
	Neither x1 nor x2	7	7	7	4												
	No selected variable				8												

4.3.2 Methods performance: Overall accuracy, sensitivity, specificity and AUC.

Tables 4.4-4.7 report the average performances of T, corT, adjcorT and Lasso in terms of overall accuracy, sensitivity, specificity (reported as a %) and AUC (reported as a decimal point) for different sample sizes and different correlation values of x_1 and x_2.

For the lowest sample size n=50 (which, given that the dataset contains 200 variables, it corresponds to a number of samples per variable ratio of 0.25) the T, corT and adjcorT methods show a similar performance and Lasso consistency reached lower accuracy level. However, Lasso outperformed the other three methods only when $\rho = -0.8$. A slight increment in sample size (n=76) showed a similar picture with the exception that for negative correlation both Lasso and adjcorT seemed to outperformed the methods corT and T.

For sample sizes n=100 and 300 (which correspond to a ratio of number of samples per variable of 0.5 and 1.5, respectively) and no negative correlations the T, corT and adjcorT show similar performances (Table 4.5). For negative correlations, however, adjcorT and Lasso methods outperformed corT. The fact that corT becomes the worst variable selection method in terms of accuracy for n=300 is consistent with the results of Table 4.3, where corT was not able to select the discriminatory variables as often as the other three methods. For example, a reduction of 13.4% and 13.9% in classification accuracy and AUC, respectively, is observed with corT when compared with adjcorT for $\rho = -0.8$ and n=300. While the method T seems to offer similar levels of accuracy as adjcorT and Lasso for n=300, for n =100 its performance is poor. In my point of view, T method's results for n =300 are not valid as logistic regression used its internal variable selection. This is again consistent with the ability to detect the discriminatory variables as reported in Table 4.2.

For larger sample sizes (n =500, 1000, 2000 and 20000), which relate to ratios of 2.5, 5, 10 and 100 for the number of samples per variable) the behaviour is similar to what has been observed for n=300. For non-negative correlations T, corT and adjcorT showed similar performances while for negative correlations corT underperformed adjcorT, Lasso and the T method (Tables 4.6, 4.7). For example, the overall classification accuracy when using corT is reduced by 11.85%, 12.95%, 12.88% and 5.11% when the sample sizes are 500, 1000, 2000 and 20000 respectively when compared to adjcorT.

These analyses indicate that corT is not able to capture the correct discriminators when the discriminators are highly and negatively correlated for small, medium and large sample sizes and that adjcorT achieves a better performance consistently.

Table 4.4: Average performances of T, corT, adjcorT and Lasso sample (n =50 and n =76)

	Average performance of T, corT, adjcorT and Lasso															
	n =50								n =76							
Corr	T		corT		adjcorT		Lasso		T		corT		adjcorT		Lasso	
0.8	Acc	56.30	Acc	57.00	Acc	56.60	Acc	49.50	Acc	62.00	Acc	65.56	Acc	65.50	Acc	53.62
	Sen	57.70	Sen	58.97	Sen	58.97	Sen	54.68	Sen	63.78	Sen	67.89	Sen	67.79	Sen	64.47
	Spe	57.58	Spe	57.49	Spe	56.66	Spe	55.72	Spe	62.16	Spe	64.81	Spe	64.81	Spe	50.86
	AUC	64.80	AUC	66.08	AUC	65.66	AUC	59.36	AUC	68.39	AUC	71.78	AUC	71.80	AUC	61.93
0.5	Acc	55.80	Acc	56.40	Acc	56.10	Acc	50.30	Acc	60.62	Acc	62.94	Acc	63.19	Acc	54.06
	Sen	58.08	Sen	58.39	Sen	58.39	Sen	58.96	Sen	61.98	Sen	64.20	Sen	64.80	Sen	66.51
	Spe	55.76	Spe	56.45	Spe	55.78	Spe	53.88	Spe	60.82	Spe	63.29	Spe	63.34	Spe	49.77
	AUC	64.73	AUC	65.43	AUC	65.34	AUC	60.25	AUC	67.75	AUC	69.03	AUC	69.36	AUC	62.90
0.2	Acc	57.80	Acc	58.00	Acc	57.70	Acc	51.30	Acc	60.25	Acc	61.06	Acc	60.88	Acc	53.94
	Sen	60.44	Sen	60.64	Sen	60.48	Sen	59.70	Sen	62.37	Sen	63.29	Sen	63.13	Sen	65.33
	Spe	57.21	Spe	57.28	Spe	56.78	Spe	54.83	Spe	59.62	Spe	60.37	Spe	60.04	Spe	50.56
	AUC	64.65	AUC	64.79	AUC	64.67	AUC	60.58	AUC	67.86	AUC	67.14	AUC	67.47	AUC	62.65
0	Acc	58.70	Acc	58.70	Acc	58.70	Acc	50.20	Acc	60.94	Acc	60.50	Acc	60.75	Acc	53.88
	Sen	60.12	Sen	59.73	Sen	59.40	Sen	56.18	Sen	61.08	Sen	60.20	Sen	60.59	Sen	62.04
	Spe	59.63	Spe	60.58	Spe	60.68	Spe	55.89	Spe	61.73	Spe	61.67	Spe	61.99	Spe	52.36
	AUC	63.96	AUC	64.29	AUC	64.47	AUC	57.97	AUC	67.62	AUC	67.35	AUC	67.52	AUC	62.09
-0.2	Acc	60.90	Acc	61.10	Acc	60.60	Acc	52.00	Acc	61.62	Acc	61.38	Acc	61.00	Acc	55.75
	Sen	63.78	Sen	64.13	Sen	63.73	Sen	56.97	Sen	63.28	Sen	62.97	Sen	62.17	Sen	64.10
	Spe	60.16	Spe	60.36	Spe	59.76	Spe	58.19	Spe	60.71	Spe	60.65	Spe	60.44	Spe	53.28
	AUC	62.75	AUC	62.36	AUC	62.52	AUC	57.72	AUC	66.07	AUC	65.63	AUC	65.67	AUC	61.56
-0.5	Acc	60.10	Acc	60.00	Acc	60.20	Acc	53.80	Acc	61.69	Acc	61.62	Acc	67.81	Acc	63.87
	Sen	62.95	Sen	62.78	Sen	62.64	Sen	56.97	Sen	62.49	Sen	62.37	Sen	69.73	Sen	70.41
	Spe	58.99	Spe	59.49	Spe	59.75	Spe	59.41	Spe	61.36	Spe	61.45	Spe	65.95	Spe	59.76
	AUC	62.63	AUC	62.30	AUC	63.26	AUC	59.41	AUC	66.72	AUC	66.45	AUC	76.28	AUC	71.57
-0.8	Acc	60.00	Acc	60.00	Acc	61.80	Acc	66.90	Acc	61.88	Acc	62.06	Acc	75.94	Acc	84.06
	Sen	60.76	Sen	60.58	Sen	61.51	Sen	67.97	Sen	61.60	Sen	61.48	Sen	73.92	Sen	79.60
	Spe	61.96	Spe	62.66	Spe	64.13	Spe	70.60	Spe	62.56	Spe	63.19	Spe	78.11	Spe	88.32
	AUC	64.34	AUC	64.49	AUC	66.69	AUC	74.71	AUC	67.73	AUC	67.60	AUC	83.37	AUC	92.49

Table 4.5: Average performance of T, corT, adjcorT and Lasso sample (n =100 and n =300)

	Average performance of T, corT, adjcorT and Lasso															
	n =100								n =300							
Corr	T		corT		adjcorT		Lasso		T		corT		adjcorT		Lasso	
0.8	Acc	61.75	Acc	66.75	Acc	66.70	Acc	53.05	Acc	68.78	Acc	68.77	Acc	68.77	Acc	59.30
	Sen	62.82	Sen	68.41	Sen	68.41	Sen	57.18	Sen	70.69	Sen	70.58	Sen	70.58	Sen	61.76
	Spe	62.50	Spe	66.29	Spe	66.19	Spe	55.92	Spe	67.48	Spe	67.59	Spe	67.59	Spe	60.55
	AUC	69.59	AUC	76.21	AUC	76.14	AUC	61.59	AUC	78.40	AUC	78.42	AUC	78.42	AUC	66.72
0.5	Acc	59.75	Acc	63.40	Acc	63.40	Acc	53.80	Acc	67.43	Acc	67.43	Acc	67.43	Acc	59.45
	Sen	60.21	Sen	65.38	Sen	65.38	Sen	56.66	Sen	70.02	Sen	69.97	Sen	69.97	Sen	63.61
	Spe	61.32	Spe	62.85	Spe	62.85	Spe	58.17	Spe	65.47	Spe	65.55	Spe	65.55	Spe	59.18
	AUC	67.63	AUC	72.62	AUC	72.62	AUC	61.72	AUC	75.70	AUC	75.76	AUC	75.76	AUC	66.49
0.2	Acc	59.60	Acc	60.45	Acc	61.35	Acc	54.50	Acc	68.50	Acc	68.42	Acc	68.18	Acc	64.98
	Sen	60.25	Sen	61.23	Sen	62.04	Sen	56.84	Sen	69.44	Sen	69.50	Sen	69.48	Sen	66.47
	Spe	60.80	Spe	61.43	Spe	62.18	Spe	59.41	Spe	68.09	Spe	67.93	Spe	67.52	Spe	65.44
	AUC	67.69	AUC	68.49	AUC	69.44	AUC	62.26	AUC	76.51	AUC	76.62	AUC	76.53	AUC	71.97
0	Acc	61.80	Acc	63.55	Acc	63.50	Acc	56.05	Acc	71.48	Acc	68.98	Acc	70.03	Acc	71.10
	Sen	61.83	Sen	63.60	Sen	63.85	Sen	60.22	Sen	73.34	Sen	70.56	Sen	71.63	Sen	73.15
	Spe	63.89	Spe	65.53	Spe	65.11	Spe	58.97	Spe	69.84	Spe	67.92	Spe	68.79	Spe	69.36
	AUC	68.21	AUC	69.45	AUC	69.75	AUC	63.54	AUC	78.33	AUC	76.57	AUC	77.00	AUC	77.86
-0.2	Acc	61.05	Acc	64.50	Acc	64.05	Acc	58.95	Acc	72.17	Acc	69.12	Acc	71.87	Acc	72.43
	Sen	60.99	Sen	65.87	Sen	64.48	Sen	59.68	Sen	71.48	Sen	69.78	Sen	71.47	Sen	71.79
	Spe	63.28	Spe	65.15	Spe	64.99	Spe	63.11	Spe	73.00	Spe	68.73	Spe	72.43	Spe	73.21
	AUC	69.16	AUC	71.47	AUC	72.05	AUC	71.05	AUC	80.17	AUC	76.71	AUC	79.61	AUC	80.51
-0.5	Acc	62.80	Acc	65.55	Acc	70.55	Acc	71.75	Acc	76.86	Acc	71.15	Acc	77.33	Acc	77.33
	Sen	62.78	Sen	65.87	Sen	71.79	Sen	72.49	Sen	76.67	Sen	71.90	Sen	77.23	Sen	77.23
	Spe	64.95	Spe	67.43	Spe	71.19	Spe	73.17	Spe	76.98	Spe	70.58	Spe	77.41	Spe	77.41
	AUC	70.10	AUC	71.56	AUC	82.13	AUC	83.97	AUC	85.04	AUC	78.68	AUC	85.79	AUC	85.79
-0.8	Acc	65.75	Acc	67.85	Acc	84.15	Acc	87.10	Acc	86.98	Acc	74.48	Acc	87.88	Acc	87.88
	Sen	66.56	Sen	68.60	Sen	85.05	Sen	87.50	Sen	86.89	Sen	75.68	Sen	87.92	Sen	87.92
	Spe	66.52	Spe	68.72	Spe	83.68	Spe	86.69	Spe	87.07	Spe	73.61	Spe	87.87	Spe	87.87
	AUC	72.36	AUC	72.65	AUC	92.20	AUC	95.42	AUC	94.23	AUC	81.48	AUC	95.38	AUC	95.38

Table 4.6: Average performance of T, corT, adjcorT and Lasso sample (n =500 and n =1000)

	Average performance of T, corT, adjcorT and Lasso															
	n =500								n =1000							
Corr	T		corT		adjcorT		Lasso		T		corT		adjcorT		Lasso	
0.8	Acc	68.09	Acc	68.09	Acc	68.09	Acc	63.46	Acc	72.10	Acc	72.10	Acc	72.10	Acc	68.19
	Sen	68.92	Sen	68.92	Sen	68.92	Sen	60.93	Sen	73.08	Sen	73.08	Sen	73.08	Sen	68.96
	Spe	67.48	Spe	67.48	Spe	67.48	Spe	67.72	Spe	71.18	Spe	71.18	Spe	71.18	Spe	68.10
	AUC	77.81	AUC	77.81	AUC	77.81	AUC	69.59	AUC	78.59	AUC	78.59	AUC	78.59	AUC	74.11
0.5	Acc	66.70	Acc	66.70	Acc	66.70	Acc	61.54	Acc	68.54	Acc	68.54	Acc	68.54	Acc	60.61
	Sen	66.53	Sen	66.53	Sen	66.53	Sen	60.07	Sen	69.26	Sen	69.26	Sen	69.26	Sen	61.39
	Spe	67.22	Spe	67.22	Spe	67.22	Spe	65.07	Spe	67.86	Spe	67.86	Spe	67.86	Spe	62.02
	AUC	75.24	AUC	75.24	AUC	75.24	AUC	68.33	AUC	75.91	AUC	75.91	AUC	75.91	AUC	66.09
0.2	Acc	68.45	Acc	68.45	Acc	68.45	Acc	68.27	Acc	69.83	Acc	69.83	Acc	69.83	Acc	69.83
	Sen	68.03	Sen	68.03	Sen	68.03	Sen	67.35	Sen	70.69	Sen	70.69	Sen	70.69	Sen	70.69
	Spe	69.27	Spe	69.27	Spe	69.27	Spe	69.65	Spe	69.01	Spe	69.01	Spe	69.01	Spe	69.01
	AUC	76.35	AUC	76.35	AUC	76.35	AUC	76.12	AUC	76.87	AUC	76.87	AUC	76.87	AUC	76.87
0	Acc	70.93	Acc	69.64	Acc	69.37	Acc	70.93	Acc	71.28	Acc	70.21	Acc	70.09	Acc	71.28
	Sen	69.28	Sen	68.55	Sen	68.43	Sen	69.28	Sen	71.83	Sen	71.47	Sen	71.48	Sen	71.83
	Spe	72.87	Spe	71.10	Spe	70.64	Spe	72.87	Spe	70.90	Spe	69.08	Spe	68.80	Spe	70.90
	AUC	78.06	AUC	76.25	AUC	75.49	AUC	78.06	AUC	78.12	AUC	76.11	AUC	75.91	AUC	78.12
-0.2	Acc	73.38	Acc	70.10	Acc	73.01	Acc	73.38	Acc	72.12	Acc	70.11	Acc	72.12	Acc	72.12
	Sen	72.83	Sen	69.40	Sen	72.31	Sen	72.83	Sen	72.11	Sen	69.86	Sen	72.11	Sen	72.11
	Spe	74.09	Spe	70.95	Spe	73.87	Spe	74.09	Spe	72.32	Spe	70.53	Spe	72.32	Spe	72.32
	AUC	80.23	AUC	77.01	AUC	79.82	AUC	80.23	AUC	80.73	AUC	77.12	AUC	80.73	AUC	80.73
-0.5	Acc	76.83	Acc	71.74	Acc	76.83	Acc	76.83	Acc	76.75	Acc	71.78	Acc	76.75	Acc	76.75
	Sen	75.57	Sen	70.94	Sen	75.57	Sen	75.57	Sen	75.94	Sen	71.67	Sen	75.94	Sen	75.94
	Spe	78.20	Spe	72.79	Spe	78.20	Spe	78.20	Spe	77.74	Spe	72.08	Spe	77.74	Spe	77.74
	AUC	85.47	AUC	78.88	AUC	85.47	AUC	85.47	AUC	85.90	AUC	79.09	AUC	85.90	AUC	85.90
-0.8	Acc	86.26	Acc	74.41	Acc	86.26	Acc	86.26	Acc	87.78	Acc	74.83	Acc	87.78	Acc	87.78
	Sen	84.71	Sen	73.86	Sen	84.71	Sen	84.71	Sen	87.58	Sen	75.48	Sen	87.58	Sen	87.58
	Spe	87.93	Spe	75.28	Spe	87.93	Spe	87.93	Spe	88.02	Spe	74.27	Spe	88.02	Spe	88.02
	AUC	94.96	AUC	81.49	AUC	94.96	AUC	94.96	AUC	95.27	AUC	81.35	AUC	95.27	AUC	95.27

Table 4.7: Average performance of T, corT, adjcorT and Lasso sample (n =2000 and n =20000)

	Average performance of T, corT, adjcorT and Lasso															
	n =2000								n =20000							
Corr	T		corT		adjcorT		Lasso		T		corT		adjcorT		Lasso	
0.8	Acc	71.34	Acc	71.34	Acc	71.34	Acc	71.34	Acc	71.41	Acc	71.41	Acc	71.41	Acc	71.41
	Sen	70.74	Sen	70.74	Sen	70.74	Sen	70.74	Sen	71.43	Sen	71.43	Sen	71.43	Sen	71.43
	Spe	71.98	Spe	71.98	Spe	71.98	Spe	71.98	Spe	71.40	Spe	71.40	Spe	71.40	Spe	71.40
	AUC	78.19	AUC	78.19	AUC	78.19	AUC	78.19	AUC	78.63	AUC	78.63	AUC	78.63	AUC	78.63
0.5	Acc	68.21	Acc	68.21	Acc	68.21	Acc	68.34	Acc	69.18	Acc	69.18	Acc	69.18	Acc	69.18
	Sen	67.38	Sen	67.38	Sen	67.38	Sen	67.83	Sen	69.20	Sen	69.20	Sen	69.20	Sen	69.20
	Spe	69.06	Spe	69.06	Spe	69.06	Spe	68.89	Spe	69.18	Spe	69.18	Spe	69.18	Spe	69.18
	AUC	75.67	AUC	75.67	AUC	75.67	AUC	75.55	AUC	76.12	AUC	76.12	AUC	76.12	AUC	76.12
0.2	Acc	69.55	Acc	69.55	Acc	69.55	Acc	69.55	Acc	69.95	Acc	69.95	Acc	69.95	Acc	69.95
	Sen	69.16	Sen	69.16	Sen	69.16	Sen	69.16	Sen	70.04	Sen	70.04	Sen	70.04	Sen	70.04
	Spe	69.98	Spe	69.98	Spe	69.98	Spe	69.98	Spe	69.87	Spe	69.87	Spe	69.87	Spe	69.87
	AUC	76.73	AUC	76.73	AUC	76.73	AUC	76.73	AUC	77.11	AUC	77.11	AUC	77.11	AUC	77.11
0	Acc	71.29	Acc	68.61	Acc	68.55	Acc	71.29	Acc	71.00	Acc	70.82	Acc	70.57	Acc	71.00
	Sen	70.70	Sen	68.10	Sen	68.19	Sen	70.70	Sen	71.46	Sen	71.23	Sen	70.93	Sen	71.46
	Spe	71.91	Spe	69.13	Spe	68.92	Spe	71.91	Spe	70.56	Spe	70.42	Spe	70.22	Spe	70.56
	AUC	78.19	AUC	76.03	AUC	75.90	AUC	78.19	AUC	78.48	AUC	78.17	AUC	77.77	AUC	78.48
-0.2	Acc	72.78	Acc	70.78	Acc	72.78	Acc	72.78	Acc	73.00	Acc	71.71	Acc	73.00	Acc	73.00
	Sen	72.71	Sen	70.98	Sen	72.71	Sen	72.71	Sen	72.79	Sen	71.61	Sen	72.79	Sen	72.79
	Spe	72.87	Spe	70.61	Spe	72.87	Spe	72.87	Spe	73.21	Spe	71.82	Spe	73.21	Spe	73.21
	AUC	80.60	AUC	77.61	AUC	80.60	AUC	80.60	AUC	80.71	AUC	79.19	AUC	80.71	AUC	80.71
-0.5	Acc	76.96	Acc	72.56	Acc	76.96	Acc	76.96	Acc	77.72	Acc	75.31	Acc	77.72	Acc	77.72
	Sen	77.08	Sen	72.43	Sen	77.08	Sen	77.08	Sen	77.51	Sen	75.19	Sen	77.51	Sen	77.51
	Spe	76.84	Spe	72.71	Spe	76.84	Spe	76.84	Spe	77.93	Spe	75.43	Spe	77.93	Spe	77.93
	AUC	85.87	AUC	79.85	AUC	85.87	AUC	85.87	AUC	85.96	AUC	83.19	AUC	85.96	AUC	85.96
-0.8	Acc	88.22	Acc	75.34	Acc	88.22	Acc	88.22	Acc	88.31	Acc	83.20	Acc	88.31	Acc	88.31
	Sen	88.11	Sen	75.26	Sen	88.11	Sen	88.11	Sen	88.19	Sen	83.11	Sen	88.19	Sen	88.19
	Spe	88.33	Spe	75.45	Spe	88.33	Spe	88.33	Spe	88.43	Spe	83.29	Spe	88.43	Spe	88.43
	AUC	95.39	AUC	82.16	AUC	95.39	AUC	95.39	AUC	95.43	AUC	90.27	AUC	95.43	AUC	95.43

4.3.3 Variability of the estimators of accuracy parameters.

4.3.3.1 Distributions of the estimates of accuracy, sensitivity, specificity and AUC (based on100 iterations).

In this subsection I examine the distributions of the estimates of the accuracy parameters, namely overall accuracy, sensitivity, specificity and AUC based on 100 iterations. Figures 4.5, 4.6, 4.7 and 4.8 display the histograms for T, corT, adjcorT and Lasso respectively, for sample size equal to 50 and correlation between $x_1$and x_2 equal to -0.8. Histograms for sample sizes 76 and 300 were displayed in Appendix I.

As expected, the distributions showed a large level of variability when the sample sizes are small, especially for n =50, and were less spread for large sample sizes. For sample size 50 and correlation between $x_1$and x_2 equal to -0.8, Lasso showed the lowest variability compared to the other three methods. The distributions for the overall accuracy, specificity and AUC seemed more skewed towards 1 when Lasso was used. The distributions of the accuracy parameters showed similar features for T, corT and adjcorT (Figure 4.5-4.8).

As the sample size increases to 76, the distributions of accuracy, sensitivity, specificity and AUC still show a large level of variability for T and corT (Figures S4.1-2, Appendix II). The distributions for adjcorT and Lasso on the other hand, seem to be more confined, with a more clear reduction in variability for Lasso (Figures S4.3-4, Appendix II).

For sample size 300, T, adjcorT and Lasso showed a lower level of variability compared to corT method (Figures S4.5-8). AdjcorT and Lasso exhibits very similar distributions, with accuracy estimates between 80% and 100%, sensitivity and specificity estimates between 75% and 100% and AUC estimates between 90% and 100%

The level of variability observed from the histograms have an effect on the mean values of of accuracy, sensitivity, specificity and AUC displayed in

Tables 4.4 and 4.5, and on their precision. As expected, larger values of the sample size imply that the estimates of the accuracy parameters are less spread and clustered around the true values of the parameters. In addition, I have identified that corT requires a larger sample size (compared for example to adjcorT) to achieve similar acceptable levels of variability and precision.

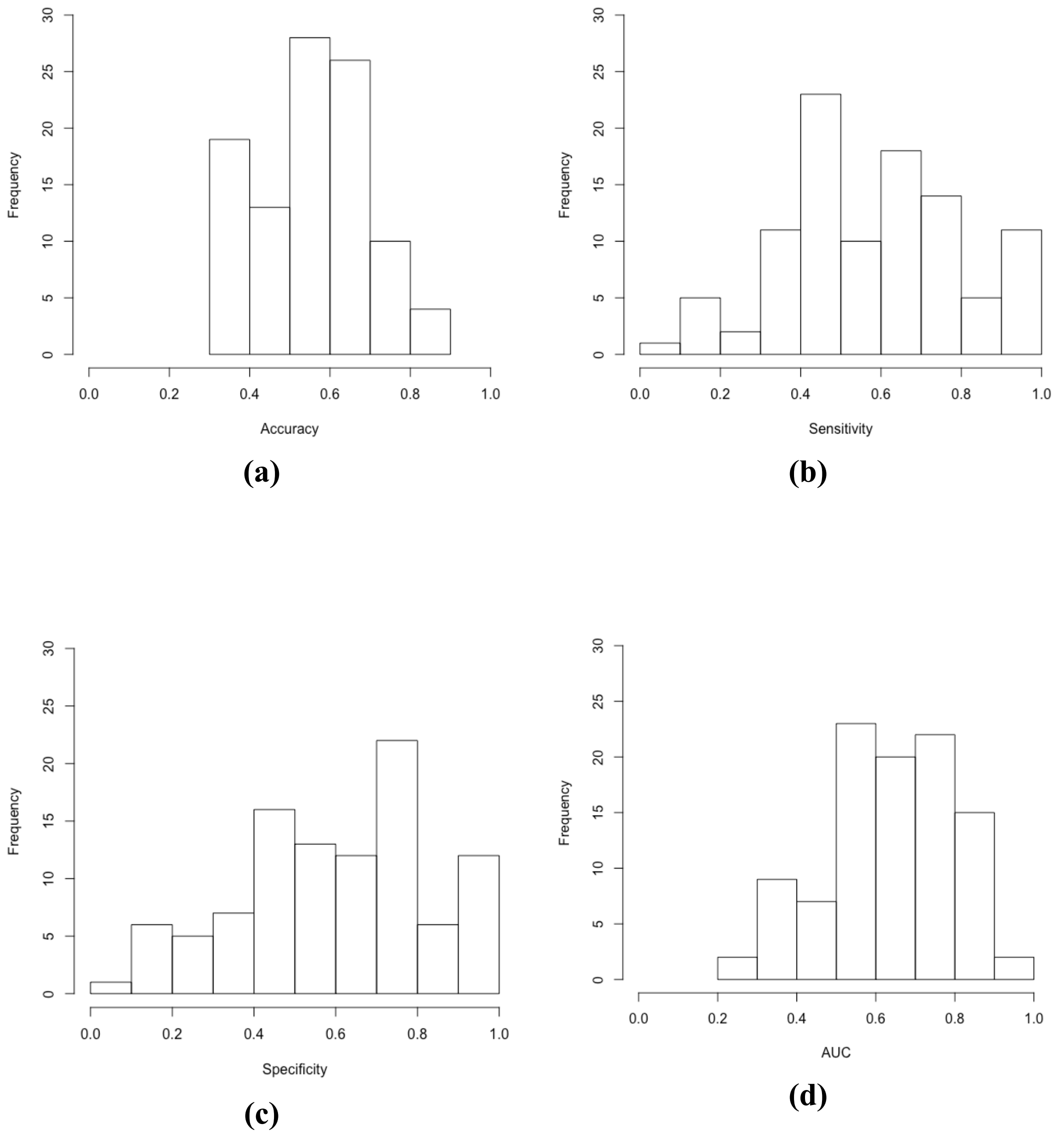
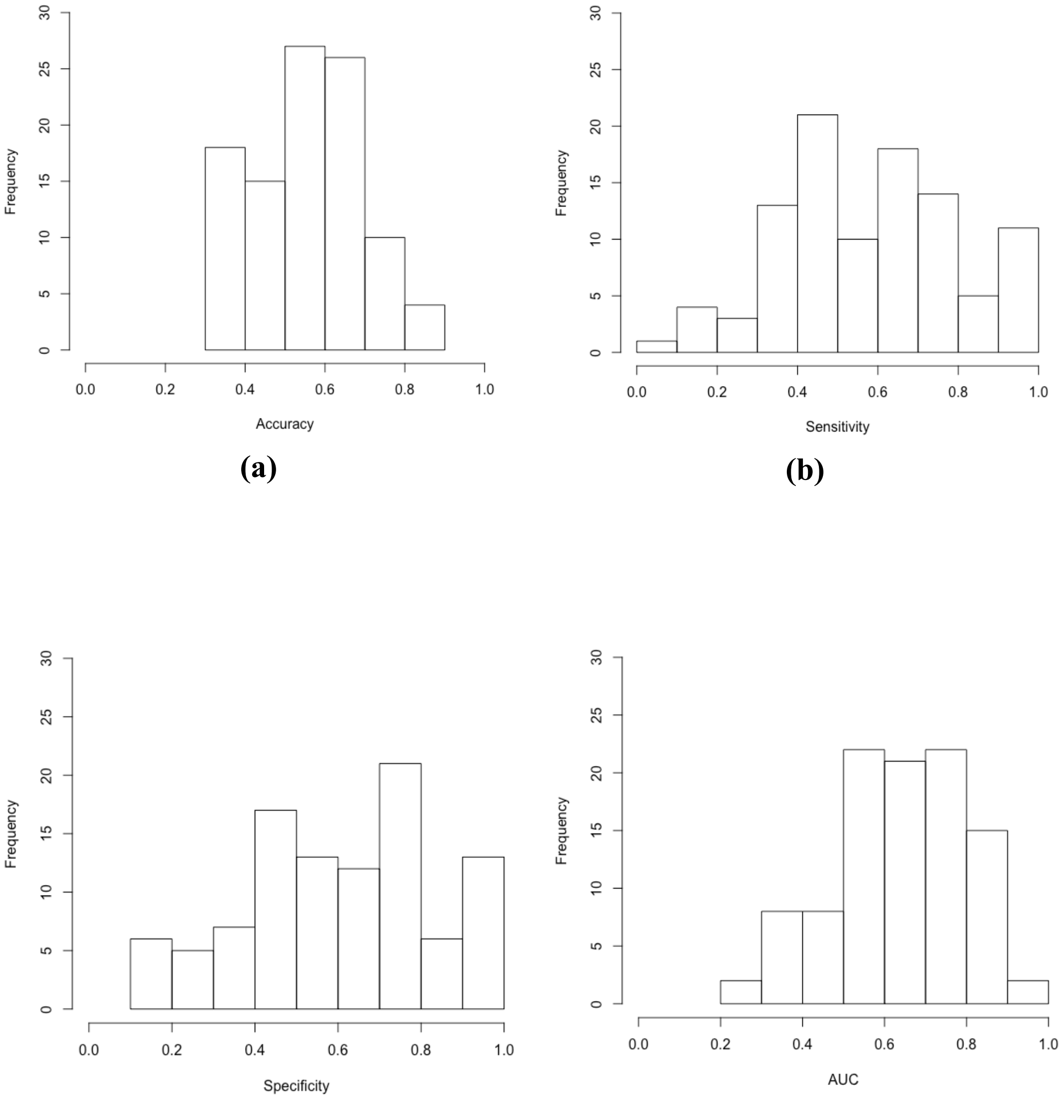


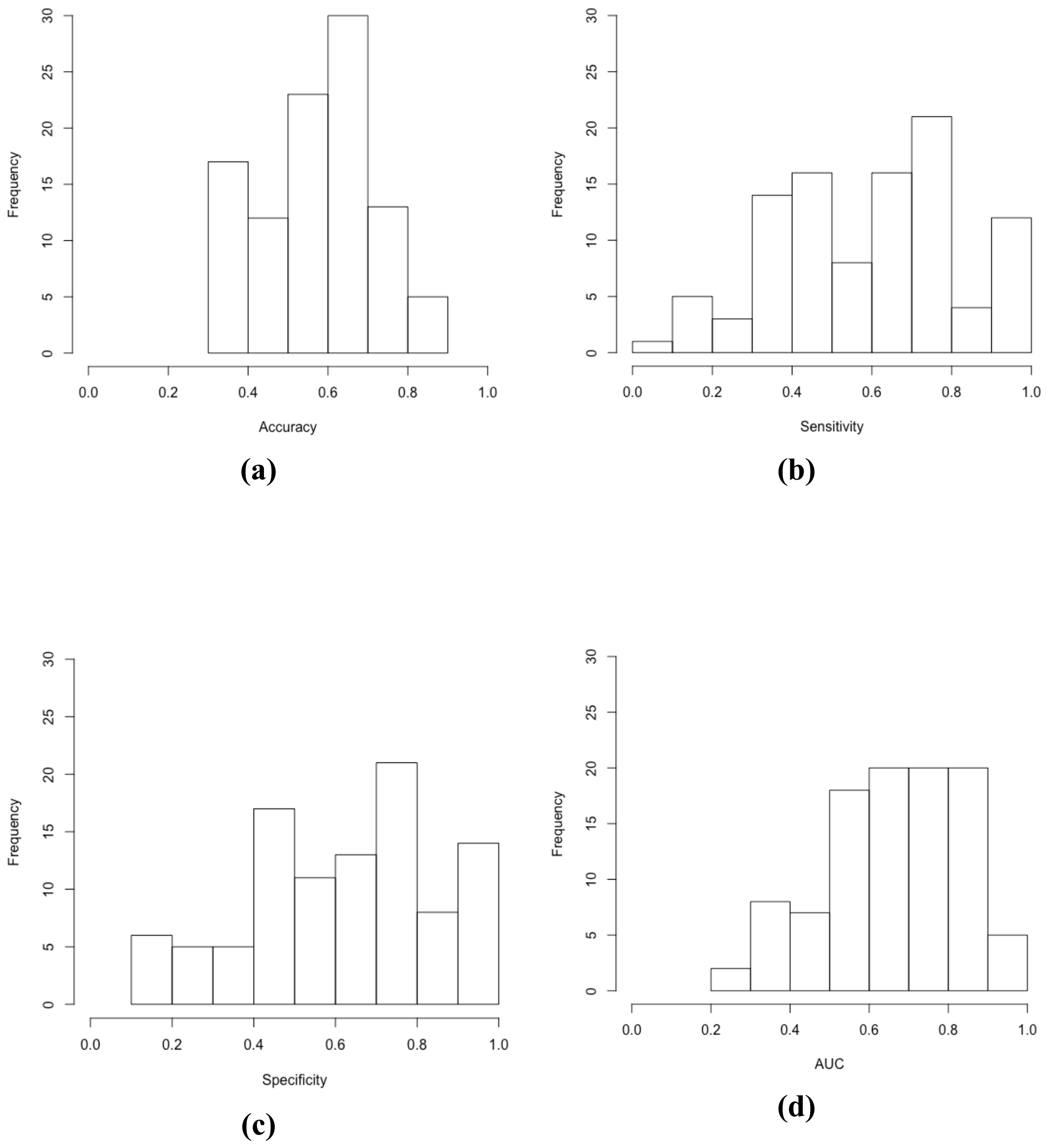
Figure 4.5: Distribution of the estimates of **(a)** Accuracy, **(b)** Sensitivity, **(c)** Specificity and **(d)** AUC for sample size n =50 and ρ= -0.8 for the T method (based on 100 iterations).



(a) **(b)**

(c) **(d)**

Figure 4.6: Distribution of the estimates of **(a)** Accuracy, **(b)** Sensitivity, **(c)** Specificity and **(d)** AUC for sample size n =50 and ρ= -0.8 for the corT method (based on 100 iterations).



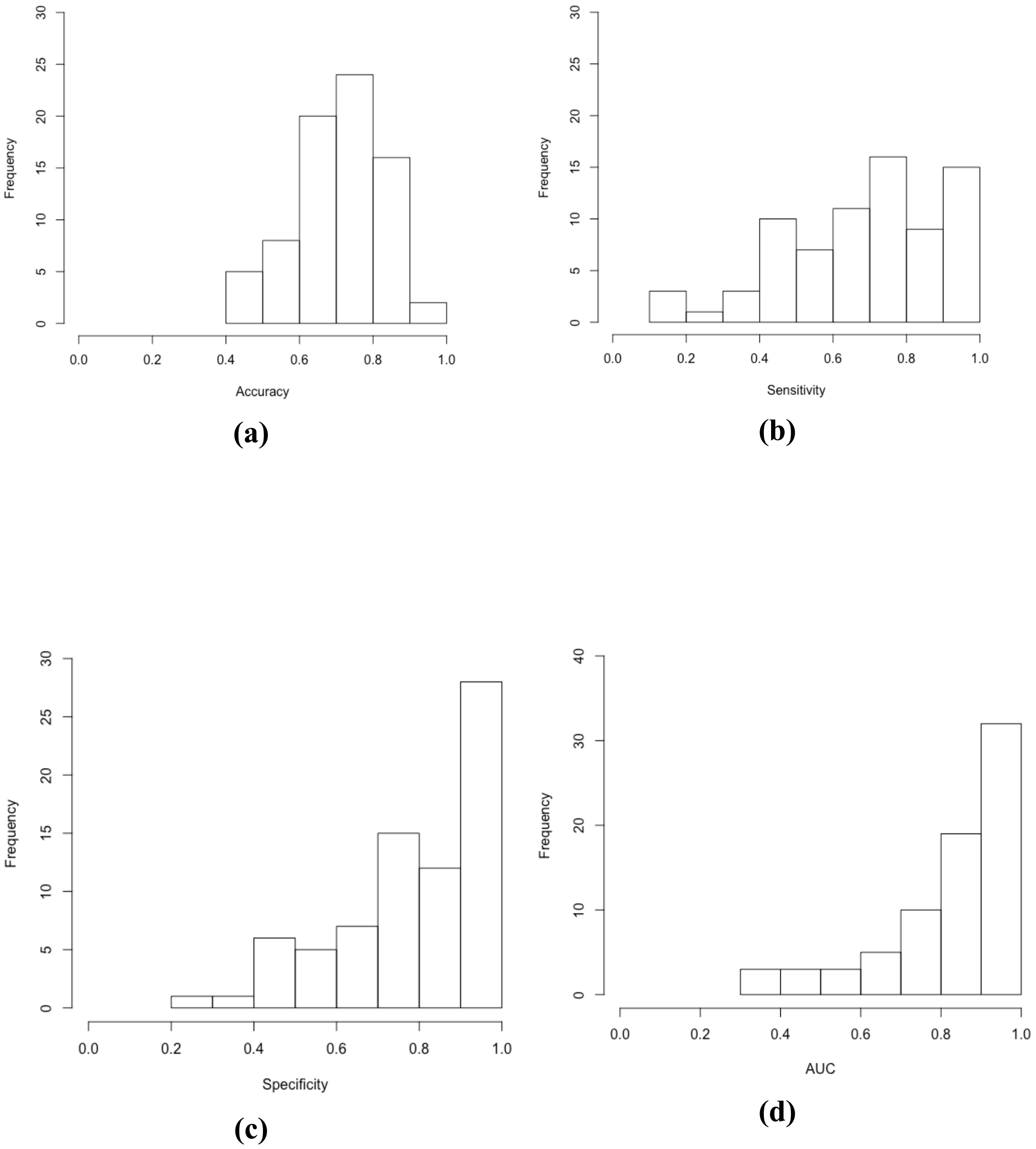
(a)

(b)

(c)

(d)

Figure 4.7: Distribution of the estimates of **(a)** Accuracy, **(b)** Sensitivity, **(c)** Specificity and **(d)** AUC for sample size n =50 and ρ= -0.8 for the adjcorT method (based on 100 iterations).



(a)

(b)

(c)

(d)

Figure 4.8: Distribution of the estimates of **(a)** Accuracy, **(b)** Sensitivity, **(c)** Specificity and **(d)** AUC for sample size n =50 and ρ= -0.8 for the Lasso method (based on 100 iterations).

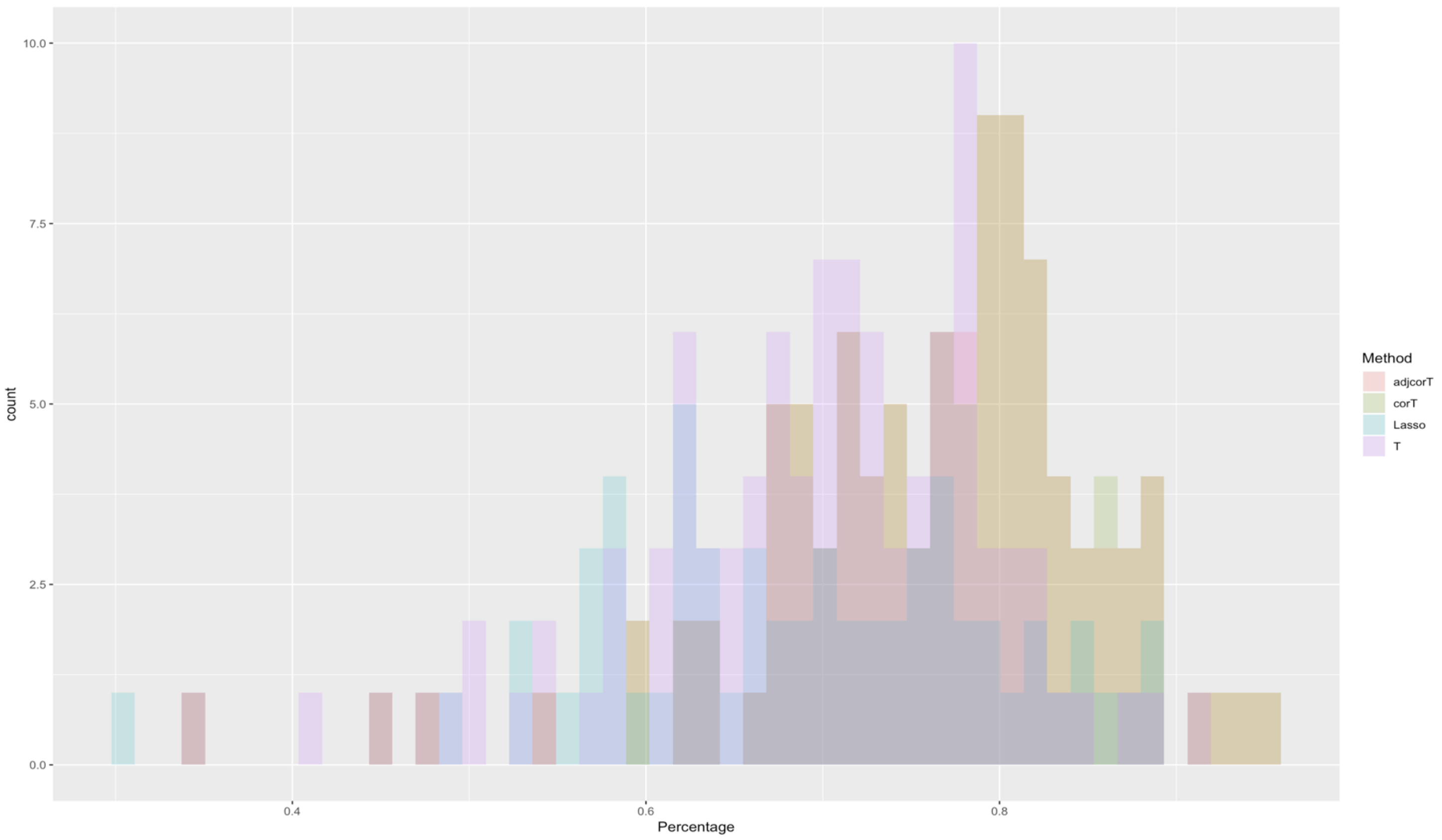
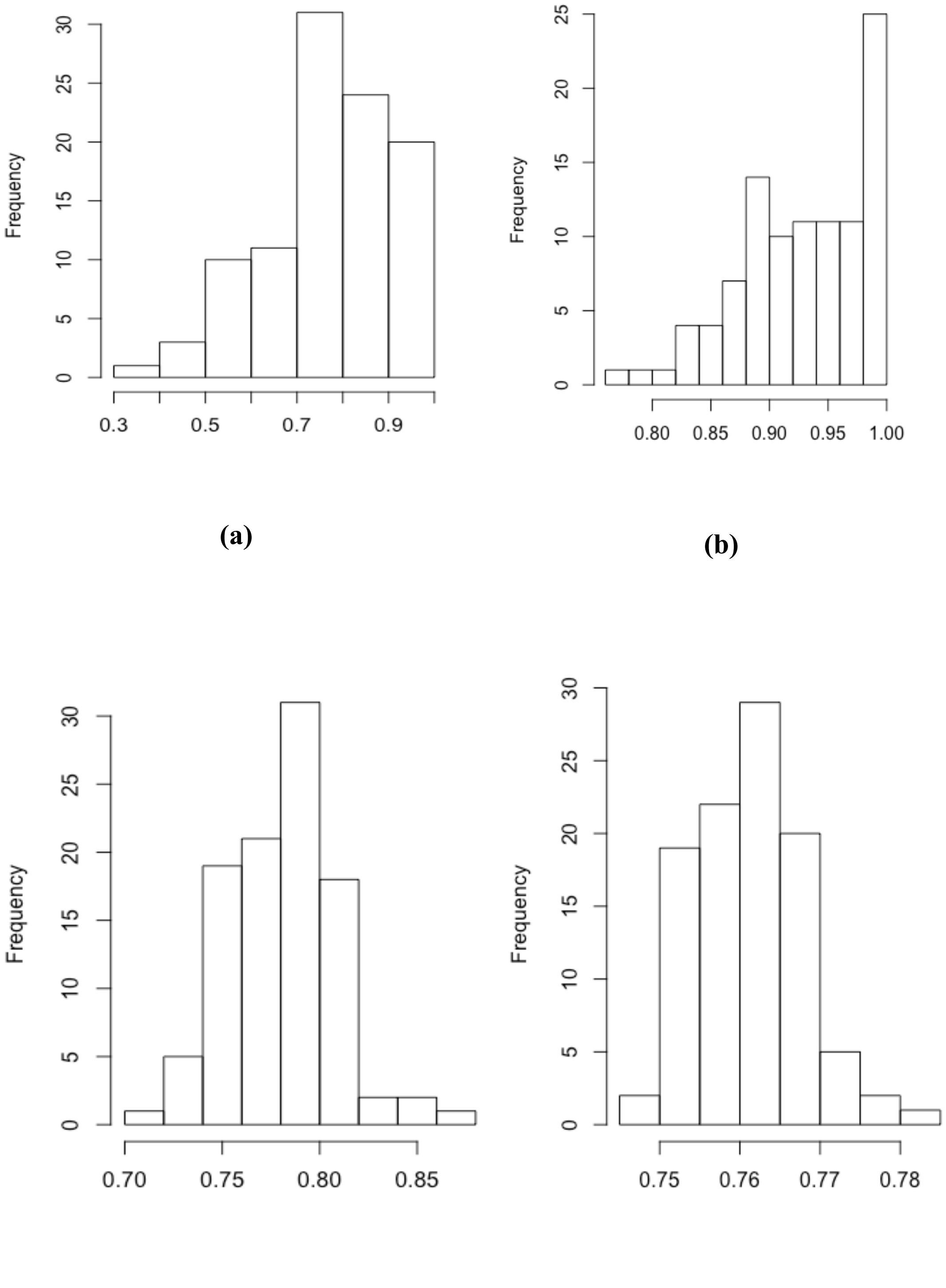


Figure 4.9: Distribution of the AUCs for sample size *n*=50 and *ρ*=0.8 across the methods

4.3.3.2 Distributions of the estimates of AUC (based on 1000 iterations)

Figure 4.10 shows the histogram of AUC (based on 1000 iterations) for a model which contains x_1 and x_2 for sample sizes 50, 76, 1000 and 20000. The aim of this analysis was to assess the effect the number of iterations has on the distribution of the accuracy parameters, and in particular, of the AUC. As described earlier, simulated data was split into two partitions namely a training set and a test set. The accuracy measure AUC was used to assess the performance of T, corT, adjcorT and Lasso methods.

Figures 4.10a and 4.10b show that the distributions of AUC are quite spread out for smaller sample sizes, especially for sample sizes 50 and 76. In contrast, the distributions of AUC showed a low degree of variability for larger sample size such as 1000 and 20000. The shape of the distributions of AUC affected the average of estimate of AUC that displayed in Tables 4.4, 4.5, 4.6 and 4.7. For larger sample sizes such as 1000 and 20000, T, corT, adjcorT and Lasso were selected the discriminators (x_1 and x_2) most of the times. Hence, these discriminators are included into logistic models in order to calculate the accuracy, sensitivity, specificity and AUC. By looking at Figure 4.10c and Figure 4.10d, the distributions of AUC are having lower variability (values are between 0.70 and 0.90) when the discriminators included into the logistic model for large sample sizes. Based on Figure 4.10, as the sample size increases, the distributions of AUC are getting smaller level of variability.



(a)

(b)

(c)

(d)

Figure 4.10: Histogram of AUC **(a)** n = 50, scenario 4 (ρ= 0), **(b)** n = 76, scenario 7 (ρ= -0.8), **(c)** n = 1000, scenario 4 (ρ= 0), and **(d)** n = 20000, scenario 2 (ρ = 0.5).

4.3.4 Estimation of accuracy parameters based on 1000 iterations compared to 100 iterations

As described at the beginning of this chapter, I used 100 iterations to conduct the analyses of the simulation study. In this subsection I considered 1000 iterations to investigate whether increasing the number of iterations has an effect on the calculation of the performance measures for each variable selection method. Table 4.8 shows the performance measures for the four variable selection methods based on 1000 iterations for three different sample sizes (n = 50, 76, 300) and for fixed correlations between $x_1$and x_2 (-0.8, 0 and 0.5). Overall, I observed that the estimates of the overall accuracy, sensitivity, specificity and AUC based on 100 iterations were similar to the estimates based on 1000 iterations. The two approaches differed by small amounts, mainly within 4% as shown in Tables 4.8-4.10. The exception is given by adjcorT for sample size 50 and correlation 0.5, where the difference in specificity and AUC reached values around 4% and 8%, respectively.

In terms of the distributions of performance measures, the corresponding histograms show a considerable spread even when 1000 iterations were used, especially for small sample sizes. This result suggests that the degree of variability observed in accuracy measures for a given sample size is mainly due to the stochastic nature of the process (random training sets show some level of variability in sensitivity, specificity, overall accuracy and AUC) and the refinement that could come from increasing the number of iterations would not be substantial enough to reduce the intrinsic level of variance captured by the histograms. Figures 4.11-4.14 displayed the histograms for sample size 50. Histograms for sample sizes 76 and 300 and for correlation equal to -0.8 are displayed in Appendix III (Figures S4.10-S4.17).

Table 4.8: Results from 100 and 1000 iterations and differences in performance (ρ = 0.5)

Sample size (n)	**Performance Measures**	**Variable selection methods**								**Differences (%)**			
		100 iterations				**1000 iterations**							
		T	**corT**	**adjcorT**	**Lasso**	**T**	**corT**	**adjcorT**	**Lasso**	**T**	**corT**	**adjcorT**	**Lasso**
50	Acc	55.80	56.40	63.40	50.30	57.62	57.32	57.44	52.34	1.82	0.92	5.96	2.04
	Sen	58.08	58.39	65.38	58.96	58.10	58.29	58.32	58.24	0.02	0.10	7.06	0.72
	Spe	55.76	56.45	62.85	53.88	59.34	58.52	58.82	56.43	3.58	2.07	4.03	2.55
	AUC	64.73	65.43	72.62	60.25	64.95	64.59	64.67	61.52	0.22	0.84	7.95	1.27
76	Acc	60.62	62.94	63.19	54.06	59.93	63.37	63.23	53.61	0.69	0.43	0.04	0.45
	Sen	61.98	64.20	64.80	66.51	60.30	65.17	64.97	60.54	1.68	0.97	0.17	5.97
	Spe	60.82	63.29	63.34	49.77	61.03	62.94	62.88	53.64	0.21	0.35	0.46	3.87
	AUC	67.75	69.03	69.36	62.90	66.94	69.69	69.48	61.05	0.81	0.66	0.12	1.85
300	Acc	67.43	67.43	67.43	59.45	67.16	67.19	67.19	60.02	0.27	0.24	0.24	0.57
	Sen	70.02	69.97	69.97	63.61	68.42	68.49	68.49	61.83	1.60	1.48	1.48	1.78
	Spe	65.47	65.55	65.55	59.18	66.27	66.28	66.28	61.22	0.80	0.73	0.73	2.04
	AUC	75.70	75.76	75.76	66.49	75.18	75.26	75.26	66.78	0.52	0.50	0.50	0.29

Table 4.9: Results from 100 and 1000 iterations and differences in performance (ρ = 0)

Sample size (n)	**Performance Measures**	**Variable selection methods**								**Differences (%)**			
		100 iterations				**1000 iterations**							
		T	**corT**	**adjcorT**	**Lasso**	**T**	**corT**	**adjcorT**	**Lasso**	**T**	**corT**	**adjcorT**	**Lasso**
50	Acc	58.70	58.70	58.70	50.20	58.25	58.31	58.20	51.67	0.45	0.39	0.50	1.47
	Sen	60.12	59.73	59.40	56.18	59.03	59.14	58.90	57.93	1.09	0.59	0.50	1.75
	Spe	59.63	60.58	60.68	55.89	59.76	59.74	59.81	55.59	0.13	0.84	0.87	0.30
	AUC	63.96	64.29	64.47	57.97	63.48	63.36	63.34	59.81	0.48	0.93	1.13	1.84
76	Acc	60.94	60.50	60.75	53.88	59.79	60.00	59.98	52.51	1.15	0.50	0.77	1.37
	Sen	61.08	60.20	60.59	62.04	59.01	59.16	59.00	58.78	2.07	1.04	1.59	3.26
	Spe	61.73	61.67	61.99	52.36	61.75	62.00	62.10	53.10	0.02	0.33	0.11	0.74
	AUC	67.62	67.35	67.52	62.09	66.24	66.61	66.48	59.94	1.38	0.74	1.04	2.15
300	Acc	71.48	68.98	70.03	71.10	71.21	69.04	69.99	71.13	0.27	0.06	0.04	0.03
	Sen	73.34	70.56	71.63	73.15	72.63	70.11	71.21	72.70	0.71	0.45	0.42	0.45
	Spe	69.84	67.92	68.79	69.36	70.09	69.35	69.09	69.94	0.25	1.43	0.30	0.58
	AUC	78.33	76.57	77.00	77.86	78.22	76.19	76.93	78.17	0.11	0.38	0.07	0.31

Table 4.10: Results from 100 and 1000 iterations and differences in performance (ρ = -0.8)

Sample Size (n)	Performance Measures	Variable selection methods								Differences (%)			
		100 iterations				1000 iterations							
		T	corT	adjcorT	Lasso	T	corT	adjcorT	Lasso	T	corT	adjcorT	Lasso
50	Acc	60.00	60.00	61.80	66.90	60.48	60.32	61.07	65.84	0.48	0.32	0.73	1.06
	Sen	60.76	60.58	61.51	67.97	61.90	61.80	62.35	68.35	1.11	1.22	0.84	0.38
	Spe	61.96	62.66	64.13	70.60	61.45	61.17	61.84	69.75	0.51	1.49	2.29	0.85
	AUC	64.34	64.49	66.69	74.71	63.79	63.42	64.33	74.19	0.55	1.07	2.36	0.52
76	Acc	61.88	62.06	75.94	84.06	61.16	61.14	74.13	85.12	0.72	0.92	1.81	1.06
	Sen	61.60	61.48	73.92	79.60	60.49	60.27	72.79	82.83	1.11	1.21	1.13	3.23
	Spe	62.56	63.19	78.11	88.32	63.32	63.43	76.34	87.79	0.76	0.24	1.77	0.53
	AUC	67.73	67.60	83.37	92.49	67.21	67.29	81.80	93.91	0.52	0.31	1.57	1.42
300	Acc	86.98	74.48	87.88	87.88	87.14	76.03	87.87	87.87	0.16	1.55	0.01	0.01
	Sen	86.89	75.68	87.92	87.92	87.09	76.77	87.85	87.85	0.20	1.09	0.07	0.07
	Spe	87.07	73.61	87.87	87.87	87.40	75.58	88.11	88.11	0.33	1.97	0.24	0.24
	AUC	94.23	81.48	95.38	95.38	94.71	82.90	95.55	95.55	0.48	1.42	0.17	0.17

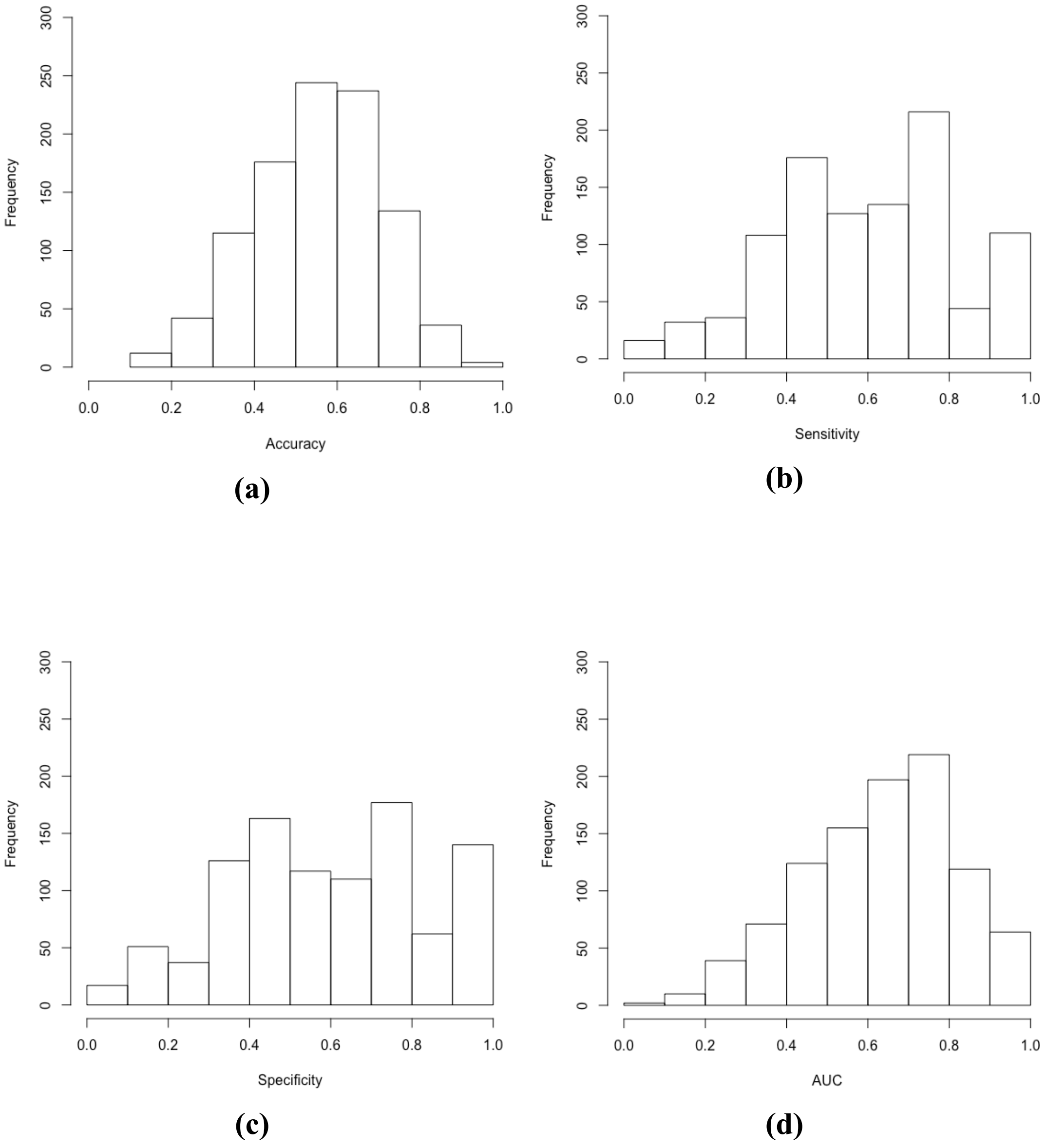
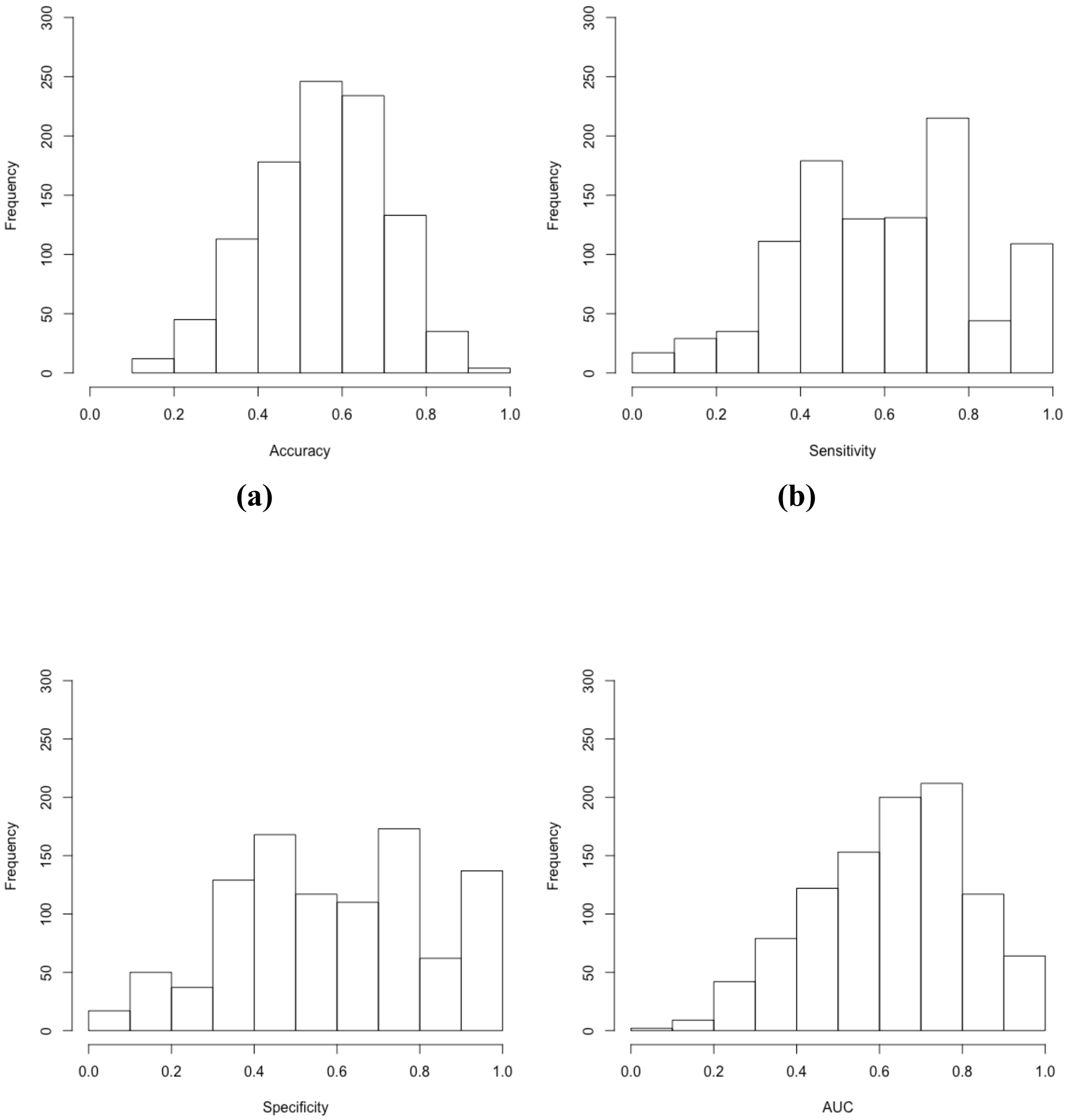


Figure 4.11: Distribution of the estimates of **(a)** Accuracy, **(b)** Sensitivity, **(c)** Specificity and **(d)** AUC for sample size *n*=50 and *ρ*= -0.8 for the T method (based on 1000 iterations).



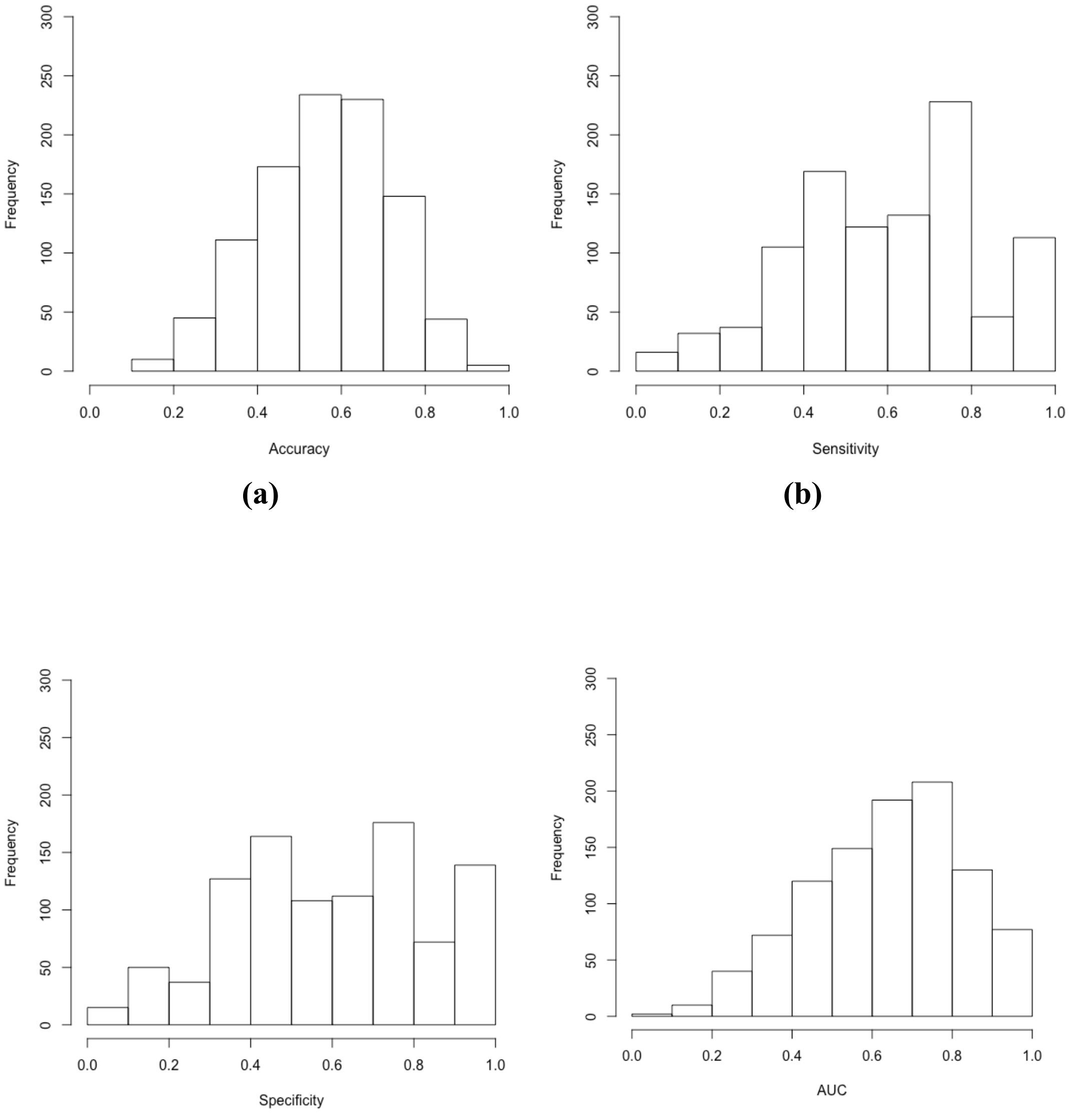
(a)

(b)

(c)

(d)

Figure 4.12: Distribution of the estimates of **(a)** Accuracy, **(b)** Sensitivity, **(c)** Specificity and **(d)** AUC for sample size *n*=50 and *ρ*= -0.8 for the corT method (based on 1000 iterations).



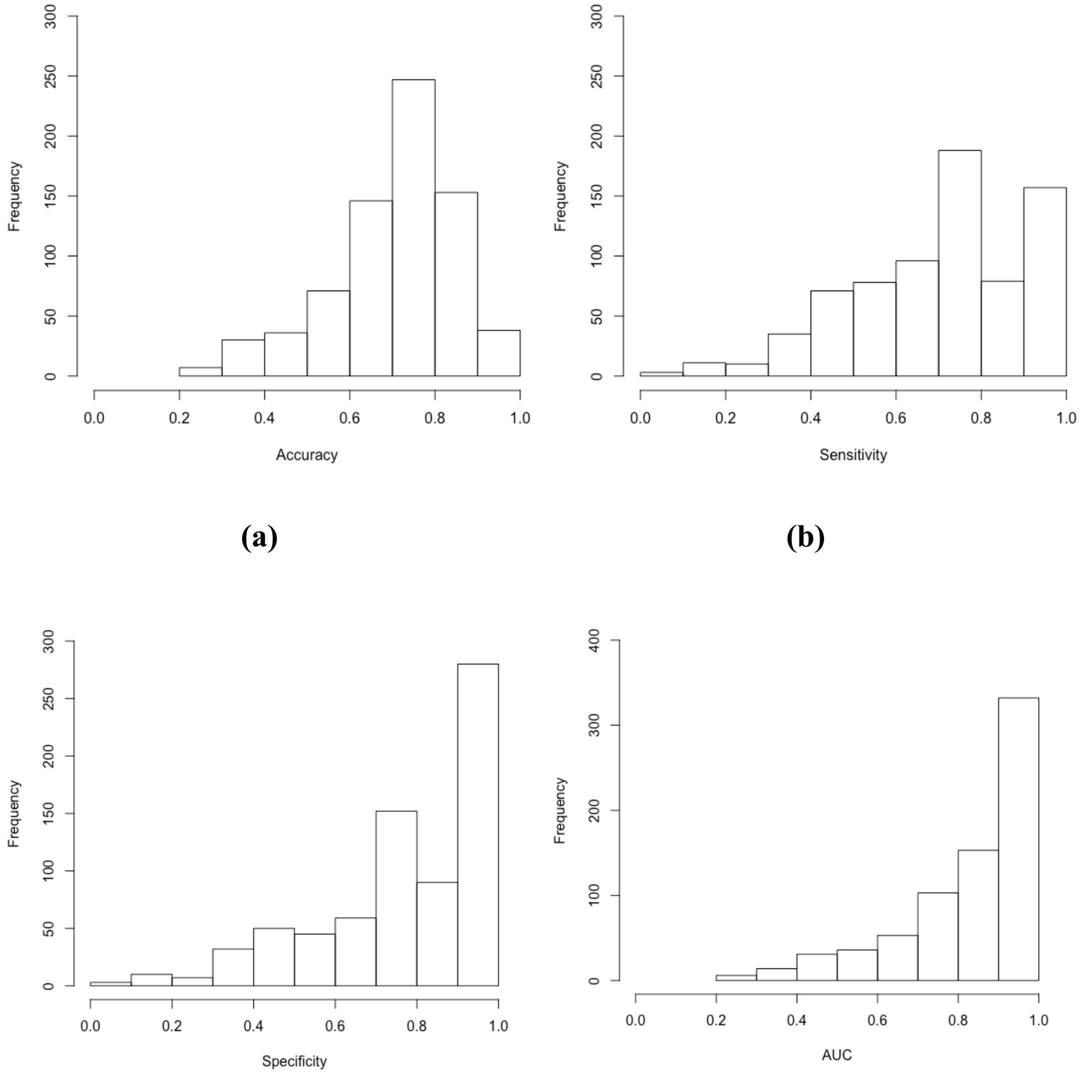
(a)

(b)

(c)

(d)

Figure 4.13: Distribution of the estimates of **(a)** Accuracy, **(b)** Sensitivity, **(c)** Specificity and **(d)** AUC for sample size *n*=50 and *ρ*= -0.8 for the adjcorT method (based on 1000 iterations).



(a) **(b)**

(c) **(d)**

Figure 4.14: Distribution of the estimates of **(a)** Accuracy, **(b)** Sensitivity, **(c)** Specificity and **(d)** AUC for sample size *n*=50 and *ρ*= -0.8 for the Lasso (based on 1000 iterations).

4.4 Discussion

In this chapter, I compared the performance of T, corT, adjcorT and Lasso methods in a simulation study. Various sample sizes and correlation values among the discriminators have been considered as both elements may impact the performance of these variable selection methods

T, corT and adjcorT show a similar performance for sample sizes n=50, 76, 100, 300, 500, 1000, 2000 and 20000 with non-negative correlation datasets as shown in Tables 4.4-4.7. Lasso outperformed all other three methods for highly negative (-0.8) correlation for sample size n =50 (Table 4.4). Additionally, Lasso and adjcorT outperformed corT for highly negative correlation for sample sizes n=76 and n=100 (Tables 4.4, 4.5). As the sample size increases, T seems to offer similar level of accuracy as adjcorT and Lasso for non-negative correlation datasets (Tables 4.5, 4.6 and 4.7).

In a nutshell, these simulation results demonstrate that different sample sizes and different correlations among discriminators have an impact on the performance of T, corT, adjcorT and Lasso. Based on the simulation study, I also noticed that corT requires a larger sample size to achieve similar acceptable performance (for example, when compared to adjcorT). AdjcorT achieves a better performance consistently. Both adjcorT and corT are filter variable selection methods. Given that adjcorT showed a better performance compared to corT for negative correlations and a similar performance for positive correlations across all sample sizes investigated, adjcorT may offer advantages compared to corT as a variable selection method for the analysis of metabolomics data.

In Chapter 5 I investigate the application of T, corT, adjcorT and Lasso to clinical datasets.

Chapter 5
Application to Real Data

5.1 Introduction

Simulations were conducted in Chapter 4 to explore the performance of the adjcorT variable selection method and to compare it to a set of existing variable selection methods. The advantage of simulations is that one can create scenarios where the association between the variables, their discriminatory ability and the sample size are specified. Such framework is useful to assess the effect that each factor separately has on the classification accuracy. However, the main limitation of a simulation study is its generalisability to real settings. In this chapter, I apply the variable selection algorithms to several clinical datasets and I assess their performance.

There are several challenges when dealing with clinical datasets. The true joint distribution of the variables under study is often unknown and assumptions are made based on the distribution of the data available. This contrasts with simulated datasets, where the distribution of the data is known. Real datasets, and in particular metabolomics data, may contain a high percentage of missing values and one of the challenges is to find the best way to deal with missingness. Multicollinearity among the variables in metabolomics datasets is another aspect that needs considerations.

This chapter will explore the performance of the T, corT, adjcorT and Lasso methods when applied to three clinical datasets: colorectal cancer, infant sepsis and kidney datasets (Sections 5.2, 5.3 and 5.4 respectively). The discussion about this chapter are presented in Section 5.5.

Each clinical dataset was split into two sets, 80% of the dataset was used for training and the remaining 20% for testing. For each approach, the variable selection process was run 100 times (100 iterations). Hence, there were 100 training sets that were used for selecting the variables and for building the logistic model, and 100 testing sets were used for estimation of accuracy parameters. Consequently, I generated 100 values for each accuracy parameters: overall classification accuracy, sensitivity, specificity and AUC and their averages were calculated and displayed in the results section. All the analyses used VOCs information only during the variable selection process, except for the colorectal cancer datasets. The colorectal cancer datasets were used age and VOCs information during the variable selection process since it was provided by the authors. Hence, in other words, kidney and infant sepsis datasets used VOCs only as the starting set of variables for selecting the most important variables.

5.2 Discrimination of colorectal cancer using volatile organic compounds

5.2.1 Colorectal cancer dataset and aims

The purpose of this analysis was to develop a diagnostic model that could accurately discriminate between colorectal cancer cases and non-cancer cases using the VOCs data. Non-cancer cases included patients with adenoma and healthy controls. Within the non-cancer cases, I was also interested in discriminating between healthy controls and adenoma patients.

The colorectal dataset consists of 137 samples (samples from 60 healthy controls, 56 adenoma and 21 colorectal cancer patients) and 146 variables [23]. Therefore, the number of samples in the non-cancer group is 116 and the number of samples in the colorectal cancer group is 21. There are 27 variables having more than 90% of zero values. The proportion of variables having more than 90% of zero values is 18.6%. After removing those variables, the number of variables left is 119 variables. Table 5.1 shows the summary of this dataset. There are a number of measures collected for colonoscopy, which includes Bowel Cancer Screening Programme (BCSP), Iron-deficiency anaemia (IDA), change in bowel habit diarrhoea, surveillance previous neoplasia/family history (FH), inflammatory bowel disease (IBD) assessment/surveillance, gastrointestinal (GI) bleeding and unknown.

Table 5.1: Description of the colorectal cancer dataset

	Total	Healthy control	Adenoma	Colorectal Cancer
Number, *n* (%)	**137 (100)**	**60 (100)**	**56 (100)**	**21 (100)**
Age, mean (SD)	**64.3 (16.2)**	**61.9 (12.4)**	**65.6 (17.5)**	**72.7 (20.6)**
Indication for colonoscopy (binary variables)				
Bowel Cancer Screening Programme (BCSP), *n* (%)	38 (27.7)	13 (21.6)	22 (39.3)	3 (14.3)
Iron-deficiency anaemia (IDA), *n* (%)	23 (16.8)	16 (26.0)	6 (10.7)	1 (4.8)
Change in bowel habit diarrhoea, *n* (%)	16 (11.7)	11 (18.3)	4 (7.1)	1 (4.8)
Surveillance previous neoplasia/ family history (FH), *n* (%)	35 (25.5)	10 (16.0)	24 (42.9)	1 (4.8)
Inflammatory bowel disease (IBD) assessment/surveillance , *n* (%)	9 (6.6)	9 (15.0)	0 (0)	0 (0)
Gastrointestinal (GI) bleeding, *n* (%)	1 (0.7)	1 (1.6)	0 (0)	0 (0)
Unknown, *n* (%)	15 (10.9)	0 (0)	0 (0)	15 (71.3)

Volatile organic compounds (VOCs) were gathered by mass-spectrometry (MS). Autoscaling was subsequently applied in order to give all variables the same weight (i.e., initially regarded as equally important). The datasets were randomly partitioned into a training set and a test set (see Results section). The T, corT, adjcorT and Lasso variable selection methods were applied to the training datasets and the top 10 discriminatory variables were identified (Table 5.2). Lasso was used as a variable selection method and the top 10 variables selected by Lasso were included into the classification model. To classify the samples, a logistic regression model was fitted using the top 10 important variables selected by T, corT, adjcorT and Lasso.

5.2.2 Volatile organic compounds

Volatile organic compounds are a large group of carbon-based molecules. Most vapours emitted from biological samples such as breath, sweat, blood, urine and faeces contain VOCs which may have a potential link to a specific disease [141]. For example, 3-methyhexane, decane, caryophyllene naphthalene have been detected at significantly lower level in the breath of breast cancer patients [142].

Faecal samples contain VOCs which may be used to identify gastrointestinal (GI) disease. Distinctive VOCs are generated from faeces of patients suffering from GI diseases such as Crohn's diseases, chronic pancreatitis or intestinal infections [141]. Other example is the identification of the VOCs that are being detected by canine olfaction which has the potential of improving the detection of melanoma in contemporary clinical practice [143]. This study reported that VOCs obtained in urine can be used as biomarkers of bladder cancer.

Rossi et. al used VOCs from faeces in order to investigate its association with response to dietary interventions in patients with irritable bowel syndrome [144]. Aggio et. al. suggested that VOCs profiling are able to differentiate patients with irritable bowel syndrome (IBS), inflammatory bowel disease and healthy controls with a minimum errors [145].

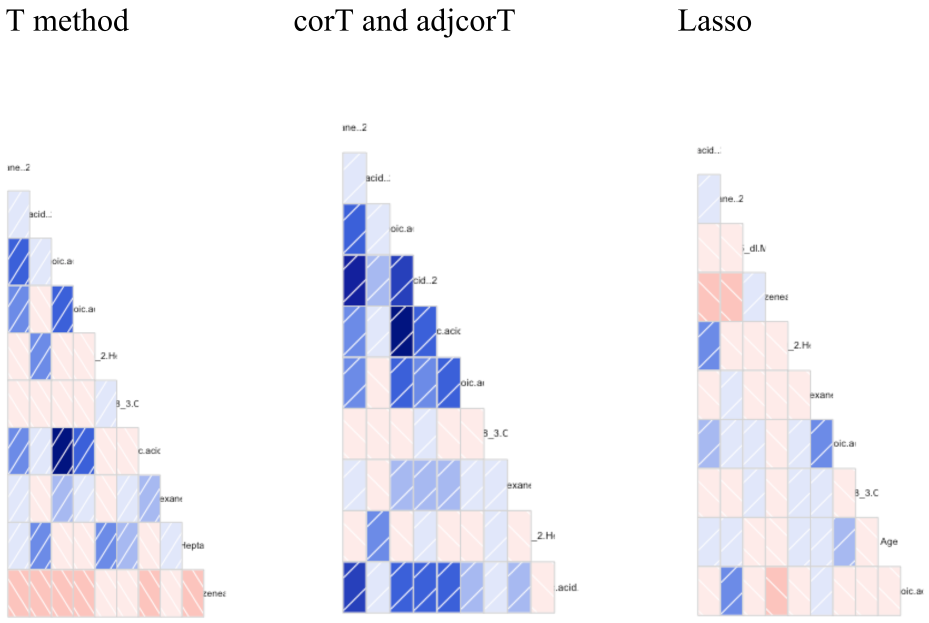
A solvent-free extraction technique that is used for metabolite extraction is the solid phase micro-extraction fibre (SPME) technique. SPME minimises contact with possible infectious agents from biological samples (blood, stool and urine samples). SPME can be coupled to gas chromatography-mass spectrometry (GC-MS) and it is one of the most popular methods for the analysis of VOCs emitted from stool samples [141].

5.2.3 Results

In terms of data partition, the data were partitioned into a training set (n =109) and a test set (n =28). Table 5.2 shows the top 10 important metabolites selected by the methods T, corT, adjcorT and Lasso. Methods T, corT and adjcorT selected X27.19_Pentane..2.3.4.trimethyl as the most important metabolite to discriminate non-cancer and colorectal cancer patients. The methods corT and adjcorT selected the same ten metabolites but in different order, giving them a different level of importance. In addition, eight variables selected by the T method were also selected by corT and adjcorT. In contrast to other methods, Lasso selected X33.44_Hexanoic.acid..2.methylbutyl as the most important variable. In addition, Lasso selected six common variables that were also selected by T, corT and adjcorT (these are represented in bold in Table 5.2). The coefficients of the logistic model based on 100 iterations are very similar (Table 5.2). Hence, the model coefficients displayed in Table 5.2 are based on one iteration.

Table 5.3 shows the performances of the T, corT, adjcorT and Lasso methods in discriminatory accuracy when applied to the colorectal cancer dataset. Method T is the worst method followed by Lasso. The methods corT and adjcorT showed a similar performance; they achieved the best level of discrimination although this is still low, with an AUC of 0.60. The correlation that exists among the selected VOCs is displayed in Table S5.1 for T, Table S5.2 is for corT and adjcorT, and in Table S5.3 for Lasso in Appendix IV. Given that the sample size is 137 and the number of variables is 119 in this application, the ratio number of samples per variables is 1.15. The simulation study conducted in Chapter 4 with 1.5 ratio (i.e, n =300 and 200 variables) indicated that for positive correlations between the discriminatory variables, the methods corT and adjcorT achieved similar level of discrimination, and that these were better than with Lasso (Table 4.6). With the colorectal cancer dataset, a similar behaviour was observed, although the difference in accuracy between the methods here is not substantial. A direct comparison is nevertheless not possible given that the set of selected discriminatory variables is different across methods with a different correlation structure.

Most of the discriminators selected by T, corT and adjcorT show a positive correlation (low, moderate and high; Table S5.1 and Table S5.2) and most discriminators selected by Lasso show low negative and low positive correlations (Table S5.3). In order to visualise the correlation matrix, I displayed the correlogram of discriminators obtained by all variable selection methods. High correlations are associated with a dark colour in the correlogram and low correlations with a light colour. Therefore, colour intensity is proportional to the absolute value of the correlation coefficients. Positive correlations are displayed in blue and negative correlations in peach colour. For negative correlations, as the correlation increases towards -1, the colour changes from peach to red. Based on Figure 5.1, the correlograms of corT and adjcorT are darker than the correlograms of T and Lasso. Figure 5.2 shows one of the ROC curves (selected from the 100 iterations).



T method

corT and adjcorT

Lasso

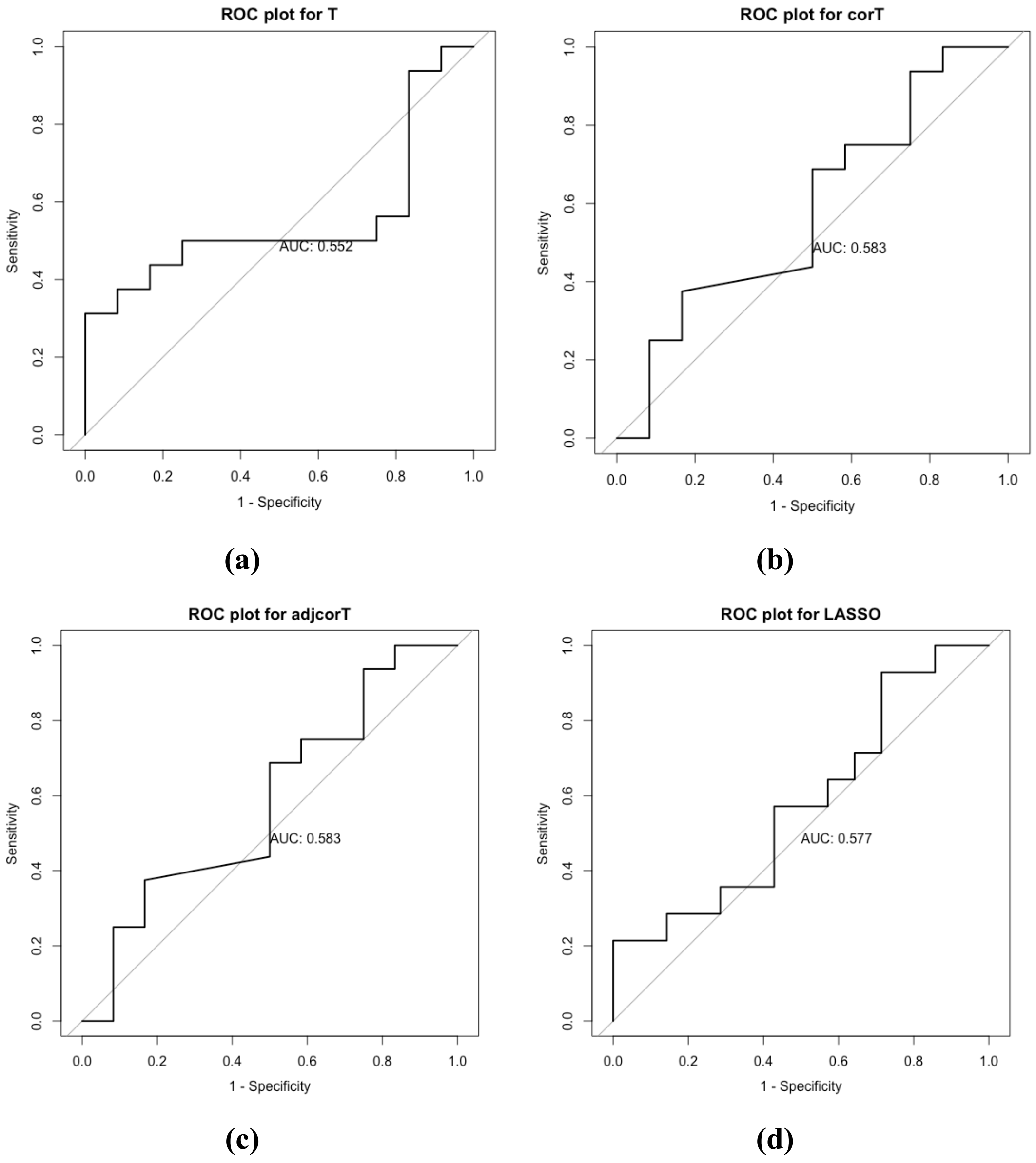
Figure 5.1: Correlogram of the top 10 VOCs selected by T (left), corT and adjcorT (middle) and Lasso (right)

Table 5.2: Top 10 selected VOCs and model coefficients by each variable selection methods for colorectal cancer and non-cancer discrimination

Variables	Coefficients			
	T, coefficients(SE)	corT, coefficients(SE)	adjcorT, coefficients(SE)	Lasso, coefficients(SE)
Intercept	0.45 (0.17)	0.49 (0.20)	0.49 (0.20)	1.51 (0.25)
X27.19_Pentane..2.3.4.trimeth	0.29 (0.05)	0.55 (0.25)	0.55 (0.25)	0.43 (0.09)
X33.44_Hexanoic.acid..2.meth	0.59 (0.03)	0.75 (0.33)	0.75 (0.33)	0.74 (0.21)
X22.19_2.Heptanol	0.43 (0.10)	0.49 (0.21)	0.49 (0.21)	0.51 (0.03)
X17.93_Propanoic.acid..propy	0.38 (0.08)	0.35 (0.15)	0.35 (0.15)	0.65 (0.21)
X29.18_3.Carene	-0.42 (0.14)	-0.40 (0.17)	-0.40 (0.17)	-0.41 (0.10)
X31.48_Cyclohexanecarboxylic.	-0.55 (0.15)	-0.51 (0.29)	-0.51 (0.29)	-0.50 (0.12)
X25.32_Propanoic.acid..pentyl	0.18 (0.01)	0.31 (0.10)	0.31 (0.10)	
X22.01_Acetic.acid..pentyl.	0.44 (0.06)	0.64 (0.20)	0.64 (0.20)	
X29.47_Heptanoic.acid	0.22 (0.02)			
X28.53_Benzeneacetaldehyde	0.02 (0.00)			-0.04 (0.00)
X23.49_Butanoic.acid..2.		-0.25 (0.09)	-0.25 (0.09)	
X27.52_Butanoic.acid..4.pentenyl		-0.25 (0.08)	-0.25 (0.08)	
X32.25_dl.Menthol				0.14 (0.01)
X24.00_Propanoic.acid..pentyl.ester				0.29 (0.05)
Age				-0.02 (0.00)

Table 5.3: Performances of T, corT, adjcorT and Lasso variable selection methods to discriminate between colorectal cancer and non-cancer cases

Measure of performances	T	corT	adjcorT	Lasso
Acc (%)	54.93	58.68	58.68	57.54
Sen (%)	51.50	57.23	57.23	54.96
Spe (%)	59.83	64.19	64.19	62.23
AUC	0.56	0.60	0.60	0.58



(a) (b)

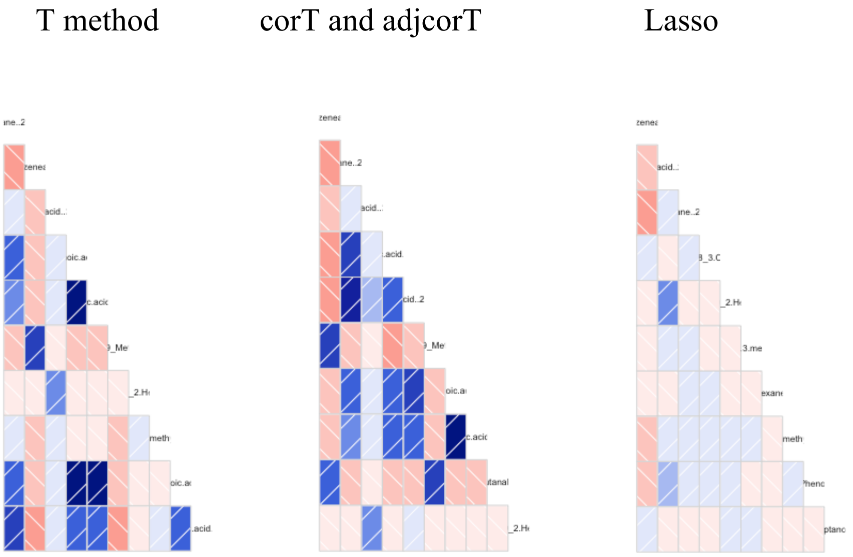
(c) (d)

Figure 5.2: Representative ROC curves for each of the variable selection approaches: **(a)** T method **(b)** corT **(c)** adjcorT **(d)** Lasso

I conducted a second analysis to identify VOCs that may be used to differentiate between healthy (n =60) and adenoma patients (n =56). Table 5.4 shows the top ten selected variables by T, corT, adjcorT and Lasso. Eight of the VOCs selected by the T method were also selected by corT and adjcorT. In addition, the top ten variables selected by the corT and adjcorT were the same and consequently, the measures of performances of both methods were equal. The methods corT, adjcorT and Lasso selected X28.53_Benzeneacetaldehyde as the most important discriminatory variable.

Table 5.5 shows the accuracy measures of the classifier when using the methods T, corT, adjcorT and Lasso to discriminate between healthy and adenoma cases. Methods corT and adjcorT showed the lowest AUC, followed by the T method. Lasso achieved the best level of discrimination, although this is still relatively low, with the AUC of 0.65 (see also Figure 5.4). One of the simulation studies carried out in Chapter 4 with 1.5 ratio (i.e, n =300 and 200 variables) claimed that for positive correlations between the discriminatory variables, the methods corT and adjcorT performed similar level of discrimination (Table 4.6). With the colorectal cancer dataset to discriminate healthy control and adenoma cases, a similar behaviour was observed.

Tables S5.4-S5.6 (in Appendix IV) show the correlation between the discriminators selected by T, corT or adjcorT and Lasso respectively. Positive correlations and negative correlations (either low, moderate or high) were observed between discriminators selected by corT or adjcorT, while the discriminators selected by Lasso showed low positive and low negative correlations.



T method corT and adjcorT Lasso

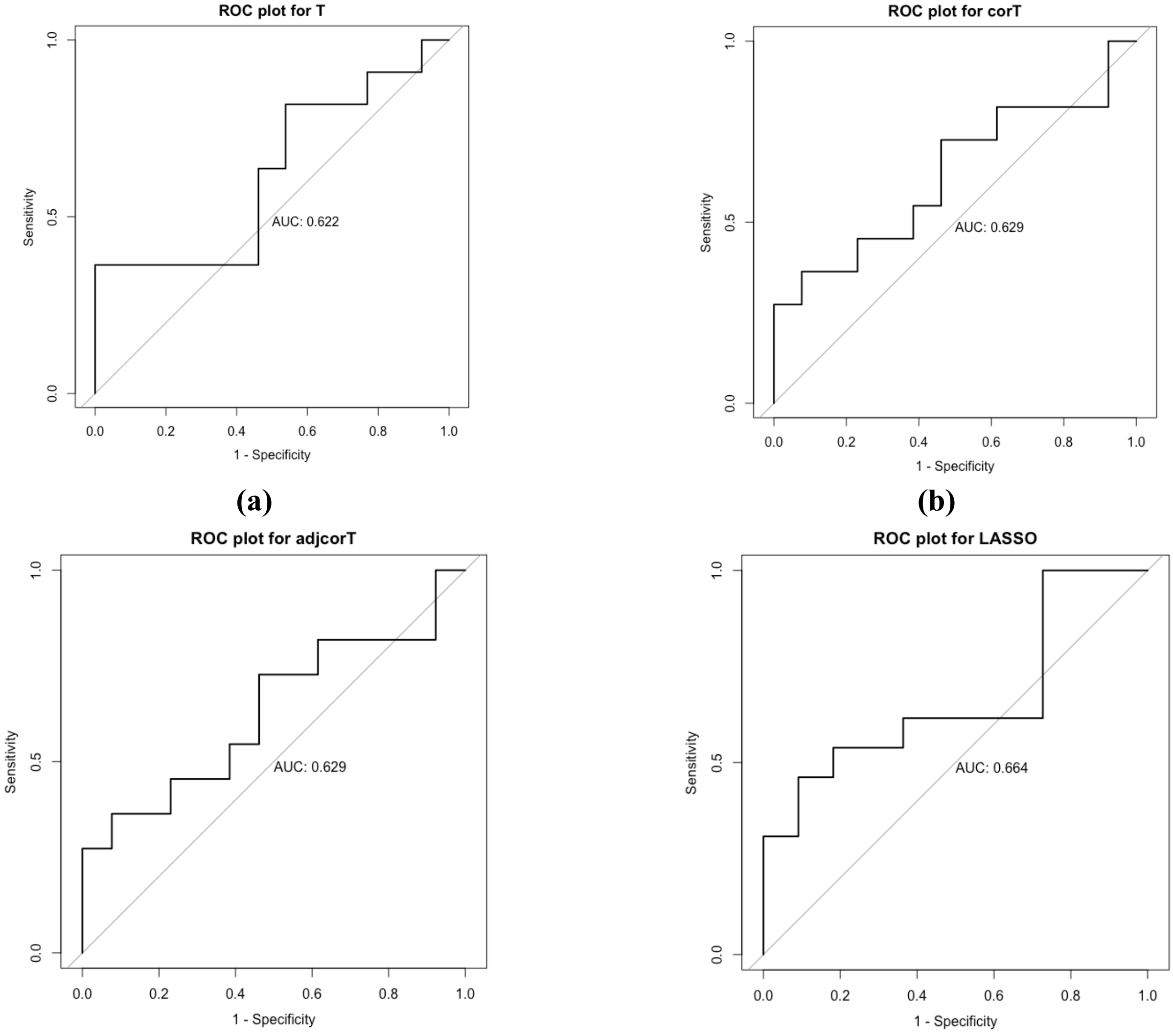
Figure 5.3: Correlogram of the top 10 VOCs selected by T (left), corT and adjcorT (middle) and Lasso (right)

Table 5.4: Top 10 selected VOCs, model coefficients and standard error (SE) of each coefficient by each variable selection methods for healthy control and adenoma discrimination

Variables	Coefficients			
	T, coefficients (SE)	corT, coefficients (SE)	adjcorT, coefficients(SE)	Lasso, coefficients(SE)
Intercept	0.14 (0.05)	0.11 (0.02)	0.11 (0.02)	0.27 (0.01)
X27.19_Pentane..2.3.4.trimeth	0.42 (0.11)	0.84 (0.11)	0.84 (0.11)	0.57 (0.12)
X28.53_Benzeneacetaldehyde	-0.46 (0.15)	-0.59 (0.13)	-0.59 (0.13)	-0.52 (0.11)
X33.44_Hexanoic.acid..2.methylbutyl	0.79 (0.21)	0.75 (0.24)	0.75 (0.24)	0.62 (0.17)
X25.32_Propanoic.acid..pentyl.ester	0.60 (0.18)			
X22.01_Acetic.acid..pentyl.ester	0.19 (0.03)	0.24 (0.05)	0.24 (0.05)	
X23.39_Methional	-0.06 (0.00)	-0.16 (0.05)	-0.16 (0.05)	
X22.19_2.Heptanol	0.64 (0.15)	0.77 (0.21)	0.77 (0.21)	0.54 (0.12)
X25.22_Dimethyl.trisulfide	0.39 (0.06)	0.45 (0.21)	0.45 (0.21)	
X24.97_Pentanoic.acid..propyl.ester	0.12 (0.01)			
X27.52_Butanoic.acid..4.pentenyl.ester	-0.67 (0.02)	-0.54 (0.14)	-0.54 (0.14)	
X23.49_Butanoic.acid..2.methylpropyl		-0.72 (0.23)	-0.72 (0.23)	
X12.47_Butanal..3.methyl.		0.09 (0.00)	0.09 (0.00)	
X29.18_3.Carene				-0.28 (0.00)
X23.27_S.Methyl.3.methylbutane				-1.98 (0.70)
X31.48_Cyclohexanecarboxylic.acid				0.64 (0.18)
X25.22_Dimethyl.trisulfide				0.28 (0.00)
X33.63_Phenol..4.ethyl.				0.15 (0.04)
X24.26_2.Heptanone..6.methyl.				0.41 (0.19)

Table 5.5: Performances of T, corT, adjcorT and Lasso variable selection methods to discriminate between healthy control and adenoma cases

Measure of performances	T	corT	adjcorT	Lasso
Acc (%)	62.83	60.46	60.46	64.79
Sen (%)	60.46	59.04	59.04	63.95
Spe (%)	68.60	65.67	65.67	66.72
AUC	0.63	0.61	0.61	0.65



(a) **(b)** **(c)** **(d)**

Figure 5.4: Representative ROC curves for each of the variable selection approaches: **(a)** T method **(b)** corT **(c)** adjcorT **(d)** Lasso

5.3 Discrimination between bacterial and non-bacterial sepsis in infants

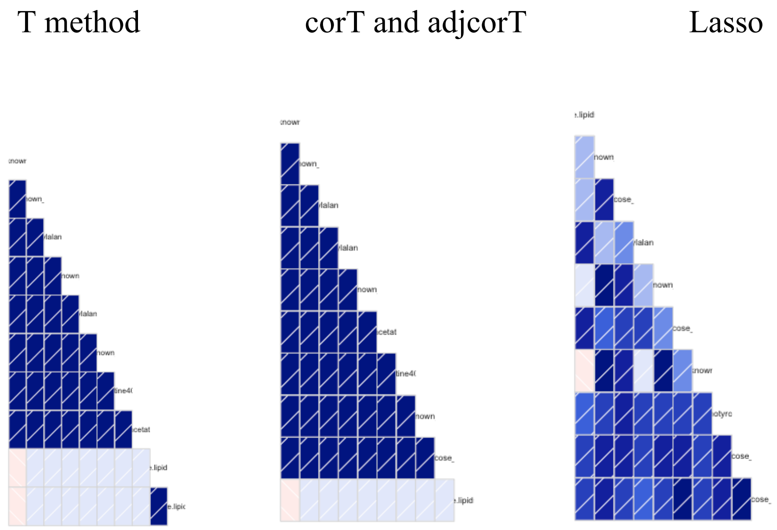
Clinicians at the Alder Hey Hospital in Liverpool are investigating better ways to discriminate between bacterial and viral sepsis in children. They collected blood samples from patients in intensive care and transferred the samples to the University of Liverpool NMR Metabolomics Centre with the aim of acquiring 1H NMR spectra of 25 samples from infants with bacterial sepsis and 91 samples from non-bacterial sepsis infants. This data has 144 metabolites and 116 children participated in this study (which gives a ratio of number of samples/variable equal to 0.81). Data is publicly available in the database MetaboLights with ID MTBLS653. For this analysis, the data were partitioned into a training set (n =92) and a test set (n =24). Table 5.6 shows the top 10 selected variables by each variable selection method (T, corT, adjcorT and Lasso).

Table 5.7 shows the performances of the classifier when using T, corT, adjcorT and Lasso as a variable selection method to discriminate bacterial and non-bacterial sepsis. Method T showed the worst performance followed by corT and adjcorT. CorT and adjcorT had equal performance with a classification accuracy of 75.83%. The simulation study conducted in Chapter 4 showed that corT and adjcorT achieved similar performances for sample size 300 and number of variables 200 for positive correlation datasets, similarly to what I observed in the application to infant sepsis dataset. Lasso exhibited a better performance than the other variable selection methods with 83.08% classification accuracy, which is approximately 7% higher than the classification accuracy of corT, adjcorT and the T-method (Table 5.7).

This dataset has known VOCs (which are named) and unknown VOCs. An unknown VOC is a metabolite that is repeatedly detected but whose chemical identity has not been identified yet. An example of unknown VOC is unknown_7. All methods except Lasso selected unknown_7 as the most important variable in discriminating bacterial and non-bacterial sepsis. Lasso selected unknown_7 as the seventh important variable. Methods corT and adjcorT selected the same 10 top variables (Table 5.6). The discriminators selected by the T method showed highly positive correlation values, except for mobile.lipids_132 and mobile.lipids_18 as shown in Table S5.7 and mobile.lipids_132 only in Table S5.8, for corT and adjcorT. These high positive correlations among discriminators were proved by the VIF values of each VOC. A VIF greater than 10 are suggest multicollinearity issue (Table 5.8 and Table 5.9). Table S5.9 shows the correlation values among the selected variables by Lasso, which involved low negative correlations as well as low, moderate positive and high positive correlation.

As can be seen from Tables S5.7 and S5.8, most of the correlations between the selected variables by T, corT and adjcorT demonstrated very strong correlations, which are 0.99 or 1.00. In metabolomics area, these values are quite common and acceptable and it can be supported by the previous study [146]. The authors studied on the approach to stool sample acquisition (home collected, or endoscopy collected) and its impact on VOC metabolome that emitted from the stool. Based on the findings, there are 39% of the total VOCs had strong correlation (correlation > 0.9). Camacho, Fuente and Mendes were investigated the origin of correlation in metabolomics data through simulation study [38]. It suggested that the highly correlation between metabolites are due to chemical equilibrium (the metabolites having near chemical equilibrium and their concentration ratio reaching the equilibrium constant).

Figure 5.5 shows the correlogram of the top 10 selected variables by T, corT, adjcorT and Lasso. The colour of correlogram is dominated by dark colours, which indicate that the correlation among discriminators tend to be high. Figure 5.6 shows a representative ROC curves generated from one of the 100 iterations.


T method

corT and adjcorT

Lasso

Figure 5.5: Correlogram of the top 10 selected variables by T (left), corT and adjcorT (middle) and Lasso (right).

Table 5.6: Top 10 selected VOCs, model coefficients and standard error (SE) of each coefficient by each variable selection methods bacterial and non-bacterial sepsis discrimination

Variables	Coefficients			
	T, coefficients(SE)	corT, coefficients(SE)	adjcorT, coefficients(SE)	Lasso, coefficients(SE)
Intercept	-0.01(0.00)	-0.02 (0.00)	-0.02 (0.00)	0.19 (0.02)
unknown_7	-0.38 (0.08)	-0.41 (0.12)	-0.41 (0.12)	-0.59 (0.15)
unknown_129	-0.16 (0.02)	-0.19 (0.01)	-0.19 (0.01)	
phenylalanine_8	-0.18 (0.05)	-0.21 (0.02)	-0.21 (0.02)	
unknown_34	0.13 (0.04)	0.11 (0.02)	0.11 (0.02)	
phenylalanine_6	-0.06 (0.00)	-0.07 (0.00)	-0.07 (0.00)	
unknown_10	-0.54 (0.12)	-0.58 (0.12)	-0.58 (0.12)	-0.57 (0.17)
creatine40_33	0.08 (0.00)	0.06 (0.01)	0.06 (0.01)	
acetoacetate_111	-0.09 (0.00)	-0.11 (0.00)	-0.11 (0.00)	
mobile.lipids_132	1.68 (0.82)	1.44 (0.25)	1.44 (0.25)	0.89 (0.25)
mobile.lipids_18	-0.21 (0.07)			
glucose_35		0.21 (0.03)	0.21 (0.03)	
unknown_94				-0.78 (0.20)
glucose_45				-0.43 (0.14)
phenylalanine_4				1.86 (0.29)
glucose_65				0.66 (0.21)
desaminotyrosine_16				-2.58 (0.30)
glucose_58				1.15 (0.25)
glucose_62				1.16 (0.26)

Table 5.7: Performances of T, corT, adjcorT and Lasso variable selection methods to discriminate between bacterial and non-bacterial sepsis cases

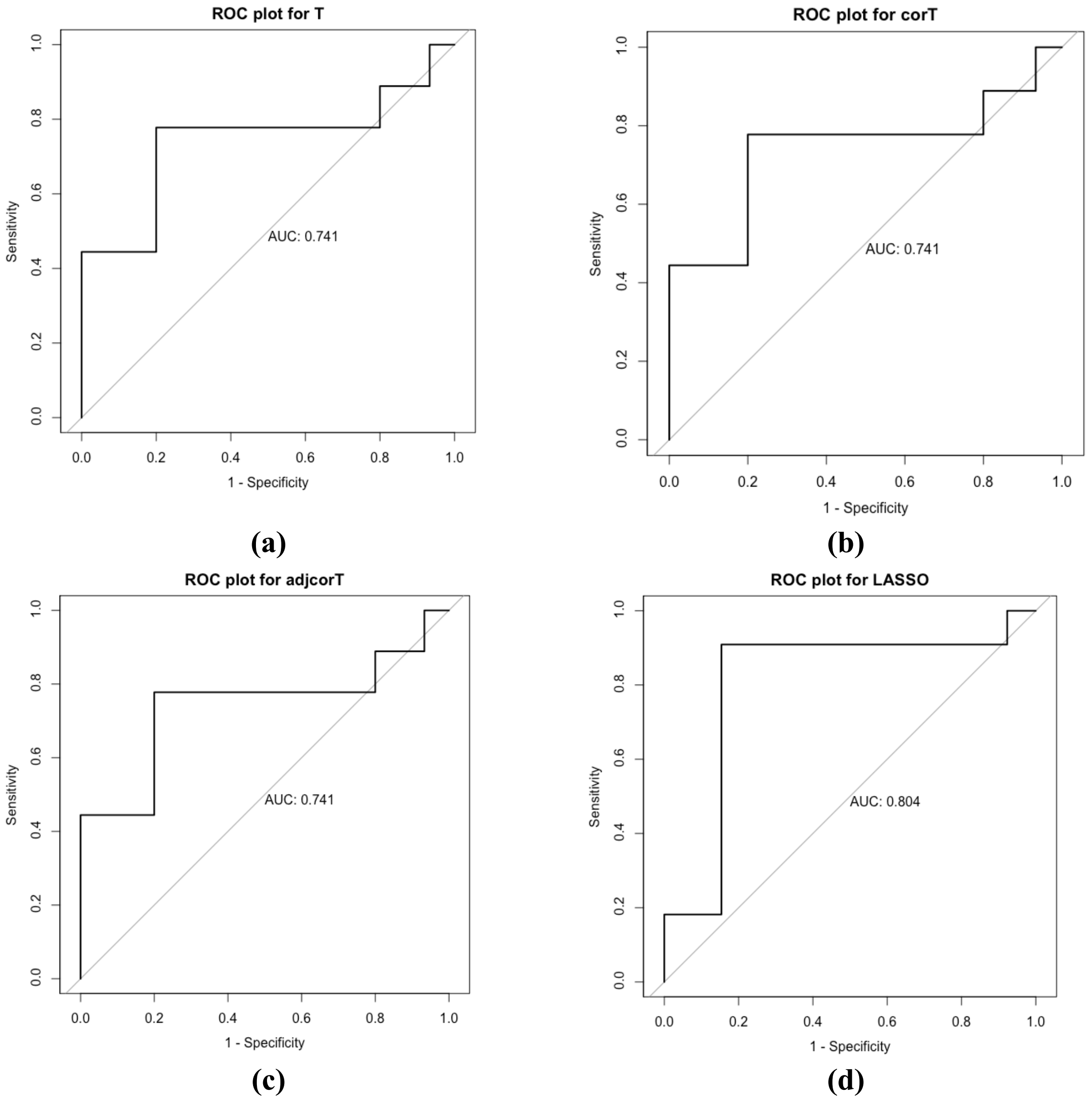
Measure of performances	T	corT	adjcorT	Lasso
Acc (%)	75.79	75.83	75.83	83.08
Sen (%)	66.45	66.70	66.70	74.43
Spe (%)	85.60	85.51	85.51	91.35
AUC	0.76	0.76	0.76	0.83

Table 5.8: VIF values for each VOC selected by T method

VOCs	VIF
unknown_7	15.17
unknown_129	14.40
phenylalanine_8	14.67
unknown_34	14.24
phenylalanine_6	13.80
unknown_10	15.77
creatine40_33	14.30
acetoacetate_111	14.61
mobile.lipids_132	4.55
mobile.lipids_18	4.52

Table 5.9: VIF values for each VOC selected by corT and ajdcorT

VOCs	VIF
unknown_7	15.72
unknown_129	14.91
phenylalanine_6	14.09
phenylalanine_8	15.16
unknown_34	14.74
acetoacetate_111	15.02
creatine40_33	14.81
unknown_10	16.41
glucose_35	14.96
mobile.lipids_132	1.14



(a) (b)

(c) (d)

Figure 5.6: Representative ROC curves for each of the variable selection approaches: **(a)** T method **(b)** corT **(c)** adjcorT **(d)** Lasso

5.4 Discrimination between healthy control and kidney disease

Chronic kidney disease (CKD) leads to a decreased sensitivity of the metabolic effects of insulin. The plasma metabolome was examined in 93 adults without diabetes in the fasted state, out of which 56 showed moderate-severe CKD and 37 a normal glomerular filtration rate. This data, which contains data on 124 metabolites, was used in the previous study [24]. Table 5.10 shows the descriptive statistics of this kidney dataset, and which includes demographic characteristics (such as age, sex, ethnicity) and medical history and lifestyle, medication use and physical characteristics.

Table 5.10: Descriptive statistics for the kidney dataset

	Healthy control	**Kidney disease**
Number	37	56
Demographics characteristics:		
Age, mean (sd)	60.6 (12.5)	63.4 (13.9)
Sex: Female, $\boldsymbol{n}$ (%)	17 (46)	30 (52)
Ethnicity, $\boldsymbol{n}$ (%)		
European descent, $\boldsymbol{n}$ (%)	32 (86)	40 (69)
Black, $\boldsymbol{n}$ (%)	4 (11)	13 (22)
Asian/ Pacific Islander, $\boldsymbol{n}$ (%)	1 (3)	5 (9)
Medical history and lifestyle (binary variables)		
History of Kidney Disease, $\boldsymbol{n}$ (%)	1 (3)	19 (33)
Current smoking, $\boldsymbol{n}$ (%)	2 (5)	10 (17)
Medication use (binary variables)		
Any antihypertensive medications, $\boldsymbol{n}$ (%)	12 (32)	52 (90)
Diuretics, $\boldsymbol{n}$ (%)	2 (5)	26 (45)
$\boldsymbol{\beta}$ Blockers, $\boldsymbol{n}$ (%)	2 (5)	22 (38)
CCBSs, $\boldsymbol{n}$ (%)	3 (8)	26 (45)
RAASi, $\boldsymbol{n}$ (%)	7 (19)	37 (64)
Physical characteristics (continuous variables)		
Height (cm), mean (sd)	172.7 (10.9)	170.4 (10.4)
Weight (kg), mean (sd)	82.9 (21.1)	87.5 (19.6)
Fat-free mass (kg), mean (sd)	55.7 (13.4)	53.3 (11.5)
Fat mass (kg), mean (sd)	28.4 (14.0)	31.6 (11.6)
Systolic blood pressure (mmHg), mean (sd)	123.5 (13.1)	134.6 (15.3)
Diastolic blood pressure (mmHg), mean (sd)	77.0 (10.2)	80.6 (9.5)

For the analysis, data were partitioned into a training set (n =74) and a test set (n =19). The variable selection approach was applied on the VOCs only (without demographic characteristics, medical history and lifestyle, medication use and physical characteristics) with the aim of findings the informative VOCs for healthy control and kidney disease discrimination. Table 5.11 shows the top 10 selected VOCs by each variable selection method (T, corT, adjcorT and Lasso) and the common selected VOCs are presented in bold. All methods selected creatinine as the most informative variable and methods corT and adjcorT selected the same top 10 variables (Table 5.11). Lasso offered the best level of discrimination (AUC=0.90; Table 5.12), although the four methods showed a comparable performance (and AUCs equal to 0.86 and 0.87 were achieved with the T and with the corT and adjcorT methods respectively). One of the simulation studies conducted in Chapter 4 was generated by using 100 samples and 200 variables, which gives the ratio of number of samples per number of variables equal to 0.5 and there is a similar performance for corT and adjcorT for positive correlations (Table 4.5). CorT and adjcorT achieved similar accuracy either in simulation study or application to real datasets. For example, in simulation study, both accuracy for corT and adjcorT are 84.30%. In real data applications, both accuracy for corT and adjcorT are 88.21%.

The correlation that exists among the selected VOCs is displayed in Table S5.10 for T and Table S5.11 for corT and adjcorT, and in Table S5.12 for Lasso. Tables S5.10 and S5.11 report the correlation structure for the top 10 important variables selected by T, corT or adjcorT, showing that the discriminators selected by adjcorT are highly correlated to each other, which is consistent with the dark colours observed in the corresponding correlogram (Figure 5.7). On the other hand, the correlation among the discriminators selected by Lasso showed low negative as well as low, moderate and high positive correlations, showing a wider range of associations in the correlation structure (Table S5.12). A representative ROC curve is plotted for each method in Figure 5.8.

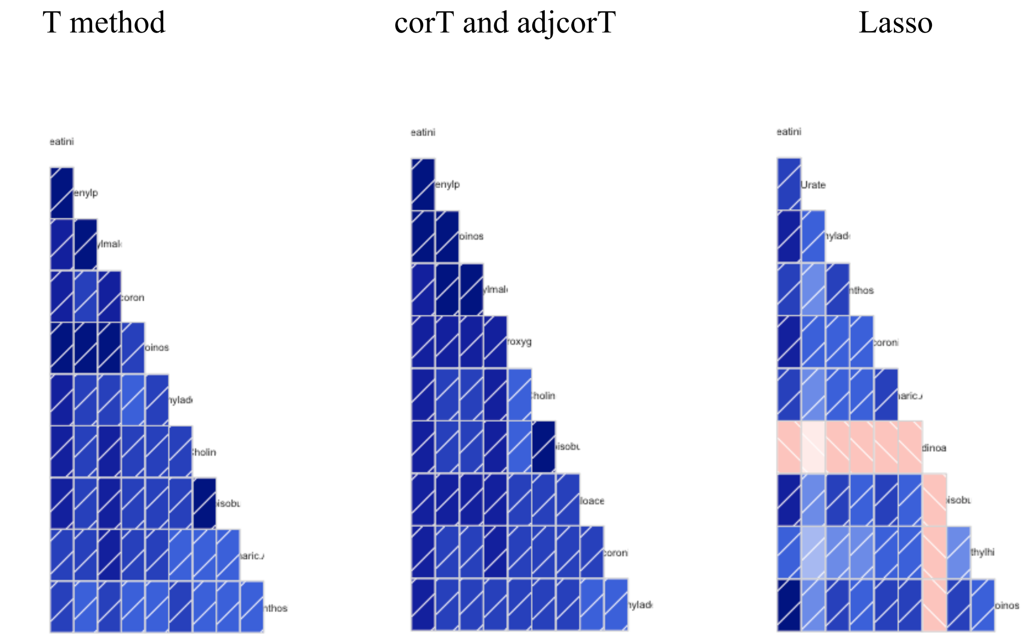


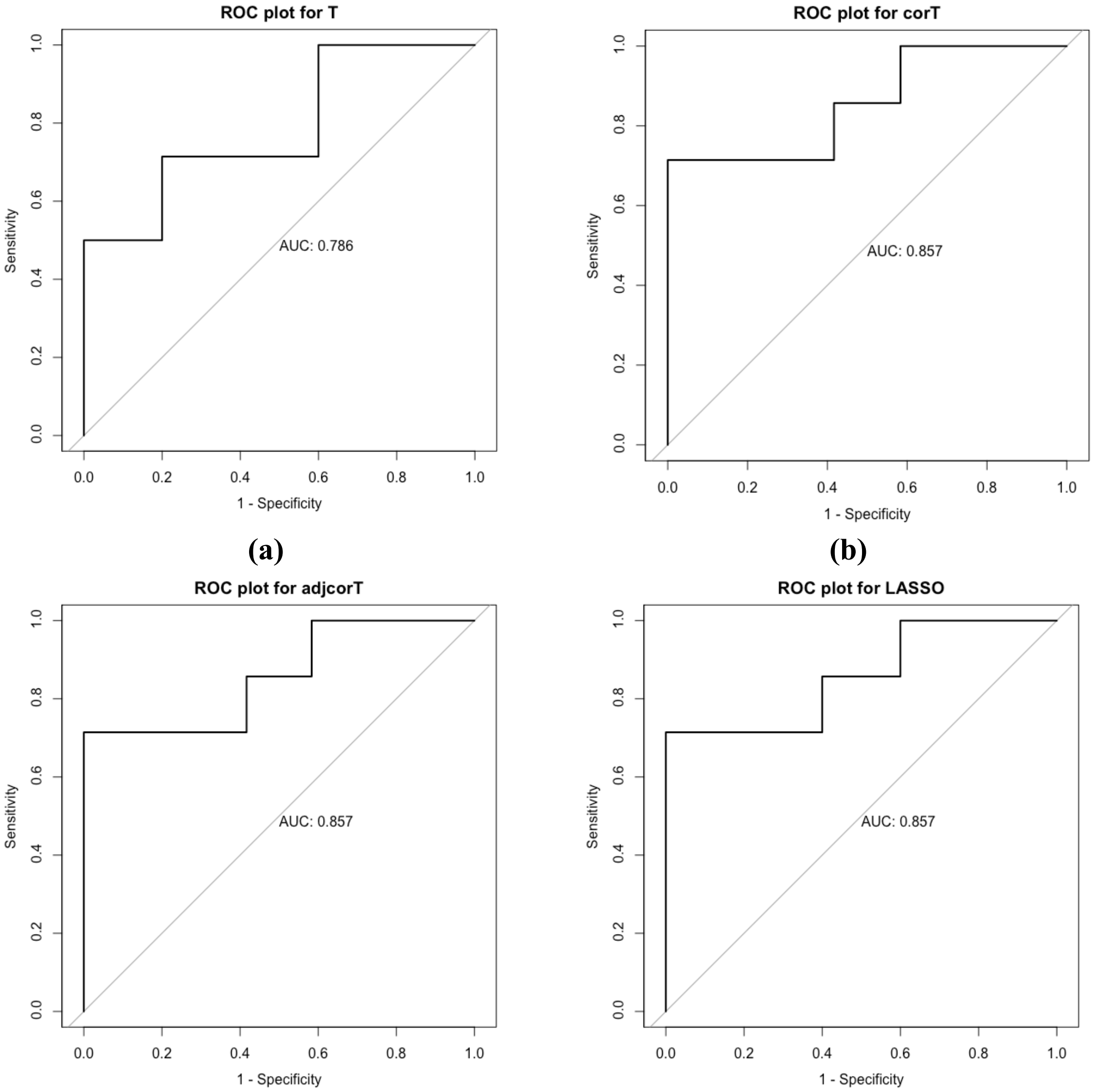
Figure 5.7: Correlogram of the top 10 VOCs selected by T (left), corT and adjcorT (middle) and Lasso (right).

Table 5.11: Top 10 selected VOCs, model coefficients and standard error (SE) of each coefficient by each variable selection methods for healthy control and kidney disease discrimination

Variables	Coefficients			
	T, coefficients(SE)	corT, coefficients(SE)	adjcorT, coefficients(SE)	Lasso, coefficients(SE)
Intercept	2.34 (0.30)	2.24 (0.40)	2.24 (0.40)	2.05 (0.35)
Creatinine	0.21 (0.05)	0.16 (0.02)	0.16 (0.02)	0.59 (0.27)
Hydroxyphenylpyruvic.acid	0.88 (0.18)	0.81 (0.31)	0.81 (0.31)	■
Methylmalonate	0.12 (0.01)	0.34 (0.14)	0.34 (0.14)	■
D.Glucoronic.acid	2.39 (0.27)	3.45 (0.52)	3.45 (0.52)	0.79 (0.29)
Myoinositol	0.73 (0.30)	0.72 (0.28)	0.72 (0.28)	1.70 (0.31)
X1.Methyladenosine	1.17 (0.03)	1.24 (0.30)	1.24 (0.30)	1.07 (0.22)
Choline	-0.12 (0.02)	0.01 (0.00)	0.01 (0.00)	■
X2.Aminoisobutyric.acid	0.40 (0.06)	0.56 (0.21)	0.56 (0.21)	0.35 (0.02)
Fumaric.Acid	1.20 (0.02)	■	■	1.01 (0.12)
Xanthosine	0.79 (0.11)	■	■	0.86 (0.22)
X2.Hydroxyglutarate	■	-0.05 (0.00)	-0.05 (0.00)	■
Oxaloacetate	■	0.23 (0.11)	0.23 (0.11)	■
Urate	■	■	■	0.82 (0.11)
Guanidinoacetate	■	■	■	-0.96 (0.34)
X1.Methylhistidine	■	■	■	0.87 (0.23)

Table 5.12: Performances of T, corT, adjcorT and Lasso variable selection methods to discriminate healthy control and kidney disease cases

Measure of performances	T	corT	adjcorT	Lasso
Acc (%)	87.63	88.21	88.21	90.89
Sen (%)	90.70	91.36	91.36	93.12
Spe (%)	83.22	83.52	83.52	87.09
AUC	0.86	0.87	0.87	0.90



(a) **(b)**

(c) **(d)**

Figure 5.8: Representative ROC curves for each of the variable selection approaches: **(a)** T method **(b)** corT **(c)** adjcorT **(d)** Lasso

5.5 Discussion

The superiority of adjcorT to select discriminatory features, when compared to corT, T and Lasso methods, has not been reproduced in the clinical applications conducted in this chapter. I believe that this might be partly due to the lack of negative correlations among the discriminant variables.

Lasso consistently produced better results although the difference in accuracy between the methods was not substantial. The fact that Lasso was able to identify discriminatory variables with low levels of correlation may have been a relevant factor. Highly correlated variables are often expected to capture similar discriminatory information, making the addition of highly correlated discriminatory variables unimportant. In the simulation study of Chapter 4 the situation was different; there were only two discriminatory variables (x_1, x_2) and the ability to select these two variables, regardless of their correlation, was key to improve the discriminatory accuracy of the model. Even when the correlation was high, such as 0.8 or -0.8, being able to select the second discriminator enhanced the level of accuracy because some additional discriminatory information was added (this would not have happened for correlations 1 or -1). In the real applications analysed in this chapter however, there was potentially larger sets of discriminatory variables with a wider range of correlations, and capturing uncorrelated features or features with low level of correlation may have contributed to a higher discrimination.

The ratio of the number of samples per variable might have also played a role in the results. Lasso showed a slightly better performance than T, corT and adjcort in two of the clinical applications when the number of samples per variable ratio was below 1 (0.81 in the infant sepsis dataset and 0.75 in the kidney disease dataset). This behaviour is nevertheless non-consistent with the results of the simulation studies, which showed that Lasso tended to underperformed adjcorT, corT and the T methods when the ratio was 0.5 (Table 4.5).

It is important to acknowledge the importance of external validation and their role in confirming the accuracy values generated by the test datasets. The *true* accuracy parameters of the models here generated may in fact be worse than the estimates derived from the test dataset, and the small improvement in terms of accuracy (e.g., AUC) achieved by the Lasso method may evaporate when the accuracy parameters are generated from a separate new set of samples.

Chapter 6

Discussion

6.1 Topics covered and main results

This thesis focuses on variable selection for classification applied to metabolomics datasets. It contains four pieces of original research: literature review on variable selection methods for classification in the area of metabolomics (Chapter 2), development of a new method, named adjcorT, as a variable selection method for selecting the most informative metabolites (Chapter 3), comparison of the performance of adjcorT and the existing methods T, corT and Lasso via a simulation study (Chapter 4) and application to real data (Chapter 5).

There were three objectives set out in Chapter 1 which were to be investigated in this thesis. The first objective was to conduct a literature review on variable selection methods for classification applied to metabolomics data. The second objective was the development of a new approach for variable selection and comparison with existing variable selection methods in terms of classification accuracy via simulations. The third objective was the application of existing methods and of the proposed method to real metabolomics datasets.

In Chapter 2, the literature related to variable selection methods in metabolomics is reviewed. ANOVA [13], [18], [45], [46] and *t*-tests [17], [53], [54], [147] are the most popular univariate filter methods applied to metabolomics datasets as they are easy to use and fast for identifying the most important metabolites. However, these methods are not optimal for the analysis of metabolite data as metabolite variables tend to be highly correlated. Multivariate techniques, such as PCA, is one of the variable reduction methods often used in this area. In the previous study, Kostidis et al. [53] argued that PCA failed to describe a clear pattern in the data. The correlation sharing t-test method (corT) is a filter method that considers the correlation among variables, but it has only been applied to genomic data [132]. The limitation of corT is that it considers positive correlations only. Wrapper methods are rarely used in metabolomics and only a few embedded methods have been applied. Lasso is one of these embedded methods. However, Lasso has a number of limitations: for high dimensional data with a large number of covariates (p) and small sample size (n), Lasso tends to select at most n variables before it saturates and if there are correlated variables, Lasso tends to select one variable and ignore the other variables in that correlated group. In this thesis I consider Lasso and corT as variable selection methods for comparison. As far as I know, none of the previous studies compared Lasso and corT even though both methods consider correlations among the variables. Since corT is based on t-tests (T method), I also consider the T method in the comparative analyses. This thesis compared these three variable selection methods in a simulation study and real dataset applications.

In terms of classification, PLS-DA are often used for classification in metabolomics studies [11], [28]–[30], [53]. Other classification methods used in metabolomics area are logistic regression [19], [47], [49], discriminant analysis [50], [72], [77], support vector machine [47], [65], [72] and random forest [35], [65], [66]. Logistic regression was chosen as the classification methods as the independent variables do not have to be normally distributed or the variances homogenous. Logistic regression also tends to generate simpler and easier to interpret equations when compared to other classification methods. For these reasons logistic regression is used in this thesis.

In Chapter 3, the data pre-processing for missing values and data scaling are briefly explained. A new variable selection algorithm, adjcorT is developed following similar conceptual ideas as with the development of the algorithm corT. The aim of using adjcorT is to identify important biomarkers in metabolomics data, while allowing for both negative and positive correlation among biomarkers, which tackles the limitation of the existing variable selection method, corT. AdjcorT considers, for each variable, the set of the indices of the variables with correlation (absolute value) equal or larger than a given threshold.

In Chapter 4, adjcorT is compared to T, corT and adjcorT. The simulation results demonstrate that different sample sizes and different correlations among discriminators have an impact on the performance of T, corT, adjcorT and Lasso (Tables 4.4-4.7). Based on the simulation study, corT requires a larger sample size in order to achieve an acceptable performance. The variability of the estimators of the accuracy parameters were also discussed in Chapter 4. The distributions showed a large level of variability when the sample size is small, especially for n=50. The distributions were less spread for large sample sizes. Given that adjcorT showed a better performance compared to corT for negative correlations and a similar performance for positive correlations across all sample sizes investigated, it is expected that adjcorT offers advantages compared to corT as a variable selection method for the analysis of metabolomics data. Additionally, the distributions of the estimates of AUC (based on 1000 iterations) were explored in order to assess the effect that the number of iterations has on the distributions. As the sample size increases, the distributions of AUC show smaller level of variability. Furthermore, the differences in accuracy parameters estimates based on 1000 iterations were compared to 100 iterations. I demonstrated that increasing the number of iterations from 100 to 1000 did not have a significant effect on the estimates of the overall accuracy, sensitivity, specificity and AUC.

In Chapter 5, T, corT, adjcorT and Lasso were applied to colorectal cancer, infant sepsis and kidney datasets. The superiority of adjcorT to select discriminatory features, when compared to corT, T and Lasso methods, has not been reproduced in the clinical applications conducted in this chapter, possibly due to the lack of negative correlations among the discriminant variables. Lasso consistently produced better results although the difference in accuracy between the methods was not substantial. The real datasets analysed in this chapter may involve several discriminatory variables with a wider range of correlations and capturing uncorrelated features or features with low level of correlation may have contributed to a higher discrimination by the Lasso method.

6.2 Limitations and further work

Simulations assumed a set of 200 variables of which 2 variables were discriminators. However, in practice more than two variables can add an important level of discrimination in real datasets and selecting only the top two variables is not desirable. Simulation studies where more than two variables are discriminatory and where different correlations structures exist (for example, correlation between the non-discriminatory variables) is an area of future research. In the clinical applications, I only considered the top ten variables as discriminatory variables, and these top ten variables were used for building the logistic model.

Each of the simulated and real datasets applications were split into two sets, 80% of the data was used for training and the remaining 20% for testing. For each approach, the variable selection process was run 100 times (100 iterations). Hence, there were 100 training sets that were used for selecting the variables and for building the logistic model, and 100 testing sets were used for estimation of accuracy parameters. In future research, the researcher may be able to use different partitions of the data for internal validation (depending on the sample size) and different number of iterations. Additionally, future research may consider bootstrapping sampling to resample the simulated dataset.

In terms of the imputation method, mean imputation was applied in this thesis, However, multiple imputation can be considered as a data pre-processing step as this method may reduce bias, improve the validity in the results and increase precision.

Future research involving datasets with negative correlations can be considered in order to explore whether the adjcorT results in the simulation studies are reproducible. It is also relevant to acknowledge the importance of external validation and their role in confirming the accuracy values generated by the test datasets, in order to validate the performance of T, corT, adjcorT and Lasso.

The proposed variable selection methods, adjcorT was employed to analyse metabolomics datasets. Future researcher may focus on applying this method to other areas in order to investigate whether this method can be successfully used to other fields (e.g., genetics, transcriptomics and proteomics), and in particular for correlation structures where negative correlations are common.

As mentioned in Chapter 5, the analysis used VOCs information only during the variable selection process, except for the colorectal cancer datasets. One may be interested in adding or combining additional clinical information with VOCs data in the analysis in order to investigate whether is there any improvement in the classification accuracy.

REFERENCES

[1] Y. B. Wah, N. Ibrahim, H. A. Hamid, S. Abdul-Rahman, and S. Fong, "Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy," *Pertanika J. Sci. Technol.*, vol. 26, no. 1, pp. 329–340, 2018.

[2] T. R. Golub *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[3] D. Chudova *et al.*, "Molecular classification of thyroid nodules using high-dimensionality genomic data," *J. Clin. Endocrinol. Metab.*, vol. 95, no. 12, pp. 5296–5304, 2010.

[4] D. Mishra, "A Signal-to-noise Classification Model for Identification of Differentially Expressed Genes from Gene Expression Data," in *2011 3rd International Conference on Electronics Computer Technology*, 2011, pp. 204–208.

[5] P. S. Gromski, Y. Xu, E. Correa, D. I. Ellis, M. L. Turner, and R. Goodacre, "A comparative investigation of modern feature selection and classification approaches for the analysis of mass spectrometry data," *Elsevier B.V.*, vol. 829, pp. 1–8, 2014.

[6] D. M. Hughes, L. J. Bonnett, G. Czanner, A. Komárek, A. G. Marson, and M. García-Fiñana, "Identification of patients who will not achieve seizure remission within 5 years on AEDs," *Neurology*, vol. 91, no. 22, pp. E2035–E2044, 2018.

[7] V. Pappu and P. M. Pardalos, *Clusters, Orders, and Trees: Methods and Applications*, vol. 92, no. December. 2014.

[8] M. Doshi, "Correlation Based Feature Selection (Cfs) Technique To Predict Student Perfromance," *Int. J. Comput. Networks Commun.*, vol. 6, no. 3, p. 197, 2014.

[9] R. Wald, T. M. Khoshgoftaar, and A. Napolitano, "Using correlation-based feature selection for a diverse collection of bioinformatics datasets," *Proc. - IEEE 14th Int. Conf. Bioinforma. Bioeng. BIBE 2014*, pp. 156–162, 2014.

[10] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, "Comparative Study of Attribute Selection Using Gain Ratio and Correlation Based Feature Selection," *Int. J. Inf. Technol. Knowl. Manag.*, vol. 2, no. 2, pp. 271–277, 2010.

[11] R. O. Bahado-Singh *et al.*, "Metabolomic analysis for first-trimester Down syndrome prediction," *YMOB*, vol. 208, p. 371.e1-371.e8, 2013.

[12] A. Hines, F. J. Staff, J. Widdows, R. M. Compton, F. Falciani, and M. R. Viant, "Discovery of metabolic signatures for predicting whole organism toxicology," *Toxicol. Sci.*, vol. 115, no. 2, pp. 369–378, 2010.

[13] P. Hsu *et al.*, "Feasibility of identifying the tobacco-related global metabolome in blood by UPLC-QTOF-MS Grant Support : Deputy Director , Comprehensive Cancer Center," *J. Proteome*, vol. 12, no. 2, pp. 679–691, 2012.

[14] E. Sherman, J. F. Harbertson, D. R. Greenwood, S. G. Villas-Bôas, O. Fiehn, and H. Heymann, "Reference samples guide variable selection for correlation of wine sensory and volatile profiling data," *Food Chem.*, vol. 267, no. March 2017, pp. 344–354, 2018.

[15] A. S. Kirpich *et al.*, "SECIMTools: A suite of metabolomics data analysis tools," *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–11, 2018.

[16] E. Marengo and E. Robotti, "Biomarkers for pancreatic cancer: Recent achievements in proteomics and genomics through classical and multivariate statistical methods," *World J. Gastroenterol.*, vol. 20, no. 37, pp. 13325–13342, 2014.

[17] J. R. Anderson, M. M. Phelan, P. D. Clegg, M. J. Peffers, and L. M. Rubio-Martinez, "Synovial Fluid Metabolites Differentiate between Septic and Nonseptic Joint Pathologies," *J. Proteome Res.*, vol. 17, no. 8, pp. 2735–2743, 2018.

[18] J. R. B. Newman, A. Kirpich, L. M. McIntyre, E. A. Ainsworth, J. M. Wedow, and G. Michailidis, "Variable selection in omics data: A practical evaluation of small sample sizes," *PLoS One*, vol. 13, no. 6, p. e0197910, 2018.

[19] L. Yengo *et al.*, "Impact of statistical models on the prediction of type 2 diabetes using non-targeted metabolomics profiling," *Mol. Metab.*, vol. 5, no. 10, pp. 918–925, 2016.

[20] S. Xu, Y. Xu, L. Gong, and Q. Zhang, "Metabolomic prediction of yield in hybrid rice," *Plant J.*, vol. 88, no. 2, pp. 219–227, 2016.

[21] H. Shahamat and A. A. Pouyan, "Feature selection using genetic algorithm for classification of schizophrenia using fMRI data," *J. Artif. Intell. Data Min.*, vol. 3, no. 1, pp. 30–37, 2015.

[22] G. Rao, J. Sui, and J. Zhang, " Metabolomics reveals significant variations in metabolites and correlations regarding the maturation of walnuts (Juglans regia L.) ," *Biol. Open*, vol. 5, no. 6, pp. 829–836, 2016.

[23] A. Bond *et al.*, "OC-048 The Use of Volatile Organic Compounds Emitted from Stool as a Biomarker for Colonic Neoplasia," *Gut*, vol. 65, no. Suppl 1, p. A28 LP-A28, Jun. 2016.

[24] B. Roshanravan *et al.*, "Chronic kidney disease attenuates the plasma metabolome response to insulin," *JCI insight*, vol. 3, no. 16, pp. 1–13, 2018.

[25] A. Bond *et al.*, "Volatile organic compounds emitted from faeces as a biomarker for colorectal cancer," *Aliment. Pharmacol. Ther.*, vol. 49, no. 8, pp. 1005–1012, 2019.

[26] F. M. Alakwaa, K. Chaudhary, and L. X. Garmire, "Deep Learning Accurately Predicts Estrogen Receptor Status in Breast Cancer Metabolomics Data," *J. Proteome Res.*, vol. 17, no. 1, pp. 337–347, 2018.

[27] M. van Reenen, C. J. Reinecke, J. A. Westerhuis, and J. H. Venter, "Variable selection for binary classification using error rate p-values applied to metabolomics data.," *BMC Bioinformatics*, vol. 17, no. 1, p. 33, Jan. 2016.

[28] T. Janvilisri, "Omics-based identification of biomarkers for nasopharyngeal carcinoma," *Dis. Markers*, vol. 2015, 2015.

[29] J. Xia, D. I. Broadhurst, M. Wilson, and D. S. Wishart, "Translational biomarker discovery in clinical metabolomics: An introductory tutorial," *Metabolomics*, vol. 9, no. 2, pp. 280–299, 2013.

[30] D. Djukovic *et al.*, "Combining NMR and LC/MS Using Backward Variable Elimination: Metabolomics Analysis of Colorectal Cancer, Polyps, and Healthy Controls," *Anal. Chem.*, vol. 88, no. 16, pp. 7975–7983, 2016.

[31] C. Pizarro, I. Esteban-Díez, I. Arenzana-Rámila, and J. M. González-Sáiz, "Discrimination of patients with different serological evolution of HIV and co-infection with HCV using metabolic fingerprinting based on Fourier

transform infrared," *J. Biophotonics*, vol. 11, no. 3, p. 10, 2018.

[32] A. Alonso, S. Marsal, and A. Julià, "Analytical methods in untargeted metabolomics: State of the art in 2015," *Front. Bioeng. Biotechnol.*, vol. 3, no. 23, 2015.

[33] R. Liu, X. Lin, Z. Li, Q. Li, and K. Bi, "Quantitative metabolomics for investigating the value of polyamines in the early diagnosis and therapy of colorectal cancer," *Oncotarget*, vol. 9, no. 4, pp. 4583–4592, 2018.

[34] I. Barba *et al.*, "High-fat diet induces metabolic changes and reduces oxidative stress in female mouse hearts," *J. Nutr. Biochem.*, vol. 40, pp. 187–193, 2017.

[35] K. Perttula *et al.*, "Untargeted lipidomic features associated with colorectal cancer in a prospective cohort," *BMC Cancer*, vol. 18, no. 1, pp. 1–10, 2018.

[36] G. Rao, X. Liu, W. Zha, W. Wu, and J. Zhang, "Metabolomics reveals variation and correlation among different tissues of olive (Olea europaea L.)," *Biol. Open*, vol. 6, no. 9, pp. 1317–1323, 2017.

[37] J. R. Everett, "From metabonomics to pharmacometabonomics: The role of metabolic profiling in personalized medicine," *Front. Pharmacol.*, vol. 7, no. 297, 2016.

[38] D. Camacho, A. de la Fuente, and P. Mendes, "The origin of correlations in metabolomics data," *Metabolomics*, vol. 1, no. 1, pp. 53–63, 2005.

[39] R. Wei *et al.*, "Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data Runmin," *Sci. Rep.*, vol. 8, no. 1, pp. 1–10, 2018.

[40] Z. Jin, J. Kang, and T. Yu, "Missing value imputation for LC-MS metabolomics data by incorporating metabolic network and adduct ion relations," *Bioinformatics*, vol. 34, no. 9, pp. 1555–1561, 2018.

[41] B. Madhu *et al.*, "Metabolomic changes during cellular transformation monitored by metabolite–metabolite correlation analysis and correlated with gene expression," *Metabolomics*, vol. 11, no. 6, pp. 1848–1863, 2015.

[42] V. Hoerr, L. Zbytnuik, C. Leger, P. P. C. Tam, P. Kubes, and H. J. Vogel, "Gram-negative and gram-positive bacterial infections give rise to a different metabolic response in a mouse model," *J. Proteome Res.*, vol. 11, no. 6, pp. 3231–3245, 2012.

[43] W. M. Bramer, M. L. Rethlefsen, J. Kleijnen, and O. H. Franco, "Optimal database combinations for literature searches in systematic reviews: A prospective exploratory study," *Syst. Rev.*, vol. 6, no. 1, pp. 1–12, 2017.

[44] V. Zuber and K. Strimmer, "Gene ranking and biomarker discovery under correlation," *Bioinformatics*, vol. 25, no. 20, pp. 2700–2707, 2009.

[45] V. Sée, P. D. Losty, M. Phelan, C. Corbishley, A. Herrmann, and Y. K. Al-Mutawa, "Effects of hypoxic preconditioning on neuroblastoma tumour oxygenation and metabolic signature in a chick embryo model," *Biosci. Rep.*, vol. 38, no. 4, p. BSR20180185, 2018.

[46] M. A. Kamleh *et al.*, "LC-MS metabolomics of psoriasis patients reveals disease severity-dependent increases in circulating amino acids that are ameliorated by anti-TNFα treatment," *J. Proteome Res.*, vol. 14, no. 1, pp. 557–566, 2015.

[47] D. Grissa, M. Pétéra, M. Brandolini, A. Napoli, B. Comte, and E. Pujos-Guillot, "Feature Selection Methods for Early Predictive Biomarker Discovery Using Untargeted Metabolomic Data," *Front. Mol. Biosci.*, vol. 3, 2016.

[48] G. Manley, "Metabolomics of aerobic metabolism in mice selected for

increased maximal metabolic rate Bernard," *Comp Biochem Physiol Part D Genomics Proteomics*, vol. 71, no. 2, pp. 233–236, 2013.

[49] Y. Zhou *et al.*, "Noninvasive Detection of Nonalcoholic Steatohepatitis Using Clinical Markers and Circulating Levels of Lipids and Metabolites," *Clin. Gastroenterol. Hepatol.*, vol. 14, no. 10, p. 1463–1472.e6, 2016.

[50] L. L. J. L.-S. Wang *et al.*, "DNA damage and oxidative stress in human liver cell L-02 caused by surface water extracts during drinking water treatment in a waterworks in China.," *Mutat. Res. - Genet. Toxicol. Environ. Mutagen.*, vol. 51, no. 1, pp. 229–235, 2014.

[51] M. Van Reenen, J. A. Westerhuis, C. J. Reinecke, and J. H. Venter, "Metabolomics variable selection and classification in the presence of observations below the detection limit using an extension of ERp," *BMC Bioinformatics*, vol. 18, no. 1, 2017.

[52] Y.-H. Yun, B.-C. Deng, D.-S. Cao, W.-T. Wang, and Y.-Z. Liang, "Variable importance analysis based on rank aggregation with applications in metabolomics for biomarker discovery *," *Anal. Chim. Acta*, vol. 911, pp. 27–34, 2016.

[53] S. Kostidis *et al.*, "1H-NMR analysis of feces: New possibilities in the helminthes infections research," *BMC Infect. Dis.*, vol. 17, no. 1, pp. 1–8, 2017.

[54] J. Ji *et al.*, "The antagonistic effect of mycotoxins deoxynivalenol and zearalenone on metabolic profiling in serum and liver of mice," *Toxins (Basel).*, vol. 9, no. 1, pp. 1–13, 2017.

[55] B. Xi, H. Gu, H. Baniasadi, and D. Raftery, "Statistical analysis and modeling of mass spectrometry-based metabolomics data," *Methods Mol. Biol.*, vol. 1198, pp. 333–353, 2014.

[56] M. Farrés, S. Platikanov, S. Tsakovski, and R. Tauler, "Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation," *J. Chemom.*, vol. 29, no. 10, pp. 528–536, 2015.

[57] T. Rajalahti, R. Arneberg, A. C. Kroksveen, M. Berle, K. M. Myhr, and O. M. Kvalheim, "Discriminating variable test and selectivity ratio plot: Quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles," *Anal. Chem.*, vol. 81, no. 7, pp. 2581–2590, 2009.

[58] J. Kalina and A. Schlenker, "A Robust Supervised Variable Selection for Noisy High-Dimensional Data," *Biomed Res. Int.*, vol. 2015, pp. 1–10, 2015.

[59] K. Brink-Jensen, S. Bak, K. Jørgensen, and C. T. Ekstrøm, "Integrative Analysis of Metabolomics and Transcriptomics Data: A Unified Model Framework to Identify Underlying System Pathways," *PLoS One*, vol. 8, no. 9, 2013.

[60] S. P. Young *et al.*, "Metabolomic analysis of human vitreous humor differentiates ocular inflammatory disease," *Mol. Vis.*, vol. 15, pp. 1210–1217, 2009.

[61] S. Nishiumi *et al.*, "A novel serum metabolomics-based diagnostic approach for colorectal cancer," *PLoS One*, vol. 7, no. 7, pp. 1–10, 2012.

[62] J. Liu *et al.*, "Metabolomics based markers predict type 2 diabetes in a 14-year follow-up study.," *Metabolomics*, vol. 13, no. 9, p. 104, 2017.

[63] C. Menni *et al.*, "Metabolomic Profiling of Long-Term Weight Change: Role of Oxidative Stress and Urate Levels in Weight Gain," *Obesity*, vol. 25, no. 9,

pp. 1618–1624, 2017.

[64] K. Tharmaratnam, M. Sperrin, T. Jaki, S. Reppe, and A. Frigessi, “Tilting the lasso by knowledge-based post-processing,” *BMC Bioinformatics*, vol. 17, no. 1, pp. 1–9, 2016.

[65] W. Thomson, S. Jabbari, A. E. Taylor, W. Arlt, and D. J. Smith, “Simultaneous parameter estimation and variable selection via the logit-normal continuous analogue of the spike-and-slab prior,” *J. R. Soc. Interface*, vol. 16, no. 150, 2019.

[66] P. J. Trainor, R. V. Yampolskiy, and A. P. DeFilippis, “Wisdom of artificial crowds feature selection in untargeted metabolomics: An application to the development of a blood-based diagnostic test for thrombotic myocardial infarction,” *J. Biomed. Inform.*, vol. 81, pp. 53–60, 2018.

[67] A. Marco-Ramell *et al.*, “Untargeted Profiling of Concordant/Discordant Phenotypes of High Insulin Resistance and Obesity to Predict the Risk of Developing Diabetes,” *J. Proteome Res.*, vol. 17, no. 7, pp. 2307–2317, 2018.

[68] R. Tissier, J. Houwing-Duistermaat, and M. Rodríguez-Girondo, “Improving stability of prediction models based on correlated omics data by using network approaches,” *PLoS One*, vol. 13, no. 2, pp. 1–23, 2018.

[69] J. A. Westerhuis, C. Brunius, J. Rosén, L. Shi, and R. Landberg, “Variable selection and validation in multivariate modelling,” *Bioinformatics*, no. August, pp. 1–9, 2018.

[70] J. J. Kellogg *et al.*, “Comparison of Metabolomics Approaches for Evaluating the Variability of Complex Botanical Preparations: Green Tea (Camellia sinensis) as a Case Study,” *J. Nat. Prod.*, vol. 80, no. 5, pp. 1457–1466, 2017.

[71] S. Dai *et al.*, “Metabolomics data fusion between near infrared spectroscopy and high-resolution mass spectrometry: A synergetic approach to boost performance or induce confusion,” *Talanta*, vol. 189, pp. 641–648, 2018.

[72] T. Chen *et al.*, “Random forest in clinical metabolomics for phenotypic discrimination and biomarker selection,” *Evidence-based Complement. Altern. Med.*, vol. 2013, 2013.

[73] S. N. Reinke *et al.*, “OnPLS-Based Multi-Block Data Integration: A Multivariate Approach to Interrogating Biological Interactions in Asthma,” *Anal. Chem.*, vol. 90, no. 22, pp. 13400–13408, 2018.

[74] M. Farrés, B. Piña, and R. Tauler, “Chemometric evaluation of Saccharomyces cerevisiae metabolic profiles using LC–MS,” *Metabolomics*, vol. 11, no. 1, pp. 210–224, 2014.

[75] L. K. Reed *et al.*, “Systems genomics of metabolic phenotypes in wild-type Drosophila melanogaster,” *Genetics*, vol. 197, no. 2, pp. 781–783, 2014.

[76] T. Matsuo, H. Tsugawa, H. Miyagawa, and E. Fukusaki, “Integrated Strategy for Unknown EI-MS Identification Using Quality Control Calibration Curve, Multivariate Analysis, EI-MS Spectral Database, and Retention Index Prediction,” *Anal. Chem.*, vol. 89, no. 12, pp. 6766–6773, 2017.

[77] D. E. Hoyos Ossa, R. Gil-Solsona, G. A. Peñuela, J. V. Sancho, and F. J. Hernández, “Assessment of protected designation of origin for Colombian coffees based on HRMS-based metabolomics,” *Food Chem.*, vol. 250, no. September 2017, pp. 89–97, 2018.

[78] U. Lutz, R. W. Lutz, and W. K. Lutz, “Metabolic profiling of glucuronides in human urine by LC-MS/MS and partial least-squares discriminant analysis for classification and prediction of gender,” *Anal. Chem.*, vol. 78, no. 13, pp. 4564–4571, 2006.

[79] A. P. de la Mata, R. H. McQueen, S. L. Nam, and J. J. Harynuk, "Comprehensive two-dimensional gas chromatographic profiling and chemometric interpretation of the volatile profiles of sweat in knit fabrics," *Anal. Bioanal. Chem.*, vol. 409, no. 7, pp. 1905–1913, 2017.

[80] J. Pinto *et al.*, "Impact of fetal chromosomal disorders on maternal blood metabolome: toward new biomarkers?," *Am. J. Obstet. Gynecol.*, vol. 213, p. 841.e1-841.e15, 2015.

[81] B. Worley and R. Powers, "MVAPACK: A complete data handling package for NMR metabolomics," *ACS Chem. Biol.*, vol. 9, no. 5, pp. 1138–1144, 2014.

[82] E. G. Armitage *et al.*, "Metabolic Clustering Analysis as a Strategy for Compound Selection in the Drug Discovery Pipeline for Leishmaniasis," *ACS Chem. Biol.*, vol. 13, no. 5, pp. 1361–1369, 2018.

[83] H. Wen *et al.*, "Urinary metabolite profiling combined with computational analysis predicts interstitial cystitis-associated candidate biomarkers," *J. Proteome Res.*, vol. 14, no. 1, pp. 541–548, 2015.

[84] J. He *et al.*, "Ambient mass spectrometry imaging metabolomics method provides novel insights into the action mechanism of drug candidates," *Anal. Chem.*, vol. 87, no. 10, pp. 5372–5379, 2015.

[85] C. Díaz Navarro *et al.*, "Comparative Metabolomics between Mycobacterium tuberculosis and the MTBVAC Vaccine Candidate," *ACS Infect. Dis.*, vol. 5, no. 8, p. 317-1326, 2019.

[86] H. Gowda *et al.*, "Interactive XCMS online: Simplifying advanced metabolomic data processing and subsequent statistical analyses," *Anal. Chem.*, vol. 86, no. 14, pp. 6931–6939, 2014.

[87] M. C. B. Ammons *et al.*, "Quantitative NMR metabolite profiling of methicillin-resistant and methicillin-susceptible staphylococcus aureus discriminates between biofilm and planktonic phenotypes," *J. Proteome Res.*, vol. 13, no. 6, pp. 2973–2985, 2014.

[88] T. Lange *et al.*, "Comprehensive Metabolic Profiling Reveals a Lipid-Rich Fingerprint of Free Thyroxine Far beyond Classic Parameters," *J. Clin. Endocrinol. Metab.*, vol. 103, no. 5, pp. 2050–2060, 2018.

[89] E. Shokry, A. E. de Oliveira, M. A. G. Avelino, M. M. de Deus, and N. R. A. Filho, "Earwax: A neglected body secretion or a step ahead in clinical diagnosis? A pilot study," *J. Proteomics*, vol. 159, pp. 92–101, 2017.

[90] R. Cavill, H. C. Keun, E. Holmes, J. C. Lindon, J. K. Nicholson, and T. M. D. Ebbels, "Genetic algorithms for simultaneous variable and sample selection in metabonomics," *Bioinformatics*, vol. 25, no. 1, pp. 112–118, 2009.

[91] J. Kuligowski *et al.*, "Evaluation of the effect of chance correlations on variable selection using Partial Least Squares-Discriminant Analysis," *Talanta*, vol. 116, pp. 835–840, 2013.

[92] E. Szymańska, E. Saccenti, A. K. Smilde, and J. A. Westerhuis, "Double-check: Validation of diagnostic statistics for PLS-DA models in metabolomics studies," *Metabolomics*, vol. 8, pp. 3–16, 2012.

[93] L. Wang *et al.*, "Ion-Pair Selection Method for Pseudotargeted Metabolomics Based on SWATH MS Acquisition and Its Application in Differential Metabolite Discovery of Type 2 Diabetes," *Anal. Chem.*, vol. 90, no. 19, pp. 11401–11408, 2018.

[94] M. S. Monteiro *et al.*, "Nuclear Magnetic Resonance metabolomics reveals an excretory metabolic signature of renal cell carcinoma," *Sci. Rep.*, vol. 6, no.

November, 2016.

[95] P. Rinaudo, S. Boudah, C. Junot, and E. A. Thévenot, “biosigner: A New Method for the Discovery of Significant Molecular Signatures from Omics Data,” *Front. Mol. Biosci.*, vol. 3, no. June, pp. 1–14, 2016.

[96] K. A. Lê Cao, I. González, and S. Déjean, “IntegrOmics: An R package to unravel relationships between two omics datasets,” *Bioinformatics*, vol. 25, no. 21, pp. 2855–2856, 2009.

[97] X. Zhou *et al.*, “A potential tool for diagnosis of male infertility: Plasma metabolomics based on GC-MS,” *Talanta*, vol. 147, pp. 82–89, 2016.

[98] B. Liquet, K. A. L. Cao, H. Hocini, and R. Thiébaut, “A novel approach for biomarker selection and the integration of repeated measures experiments from two assays,” *BMC Bioinformatics*, vol. 13, no. 1, pp. 1–14, 2012.

[99] L. C. Kim-Anh, R. Debra, R.-G. Christèle, and B. Philippe, “A Sparse PLS for Variable Selection when Integrating Omics Data,” *Stat. Appl. Genet. Mol. Biol.*, vol. 7, p. Article 35, 2008.

[100] L. A. Adutwum and J. J. Harynuk, “Unique ion filter: A data reduction tool for GC/MS data preprocessing prior to chemometric analysis,” *Anal. Chem.*, vol. 86, no. 15, pp. 7726–7733, 2014.

[101] D. Friston, H. Laycock, I. Nagy, and E. J. Want, “Microdialysis Workflow for Metabotyping Superficial Pathologies: Application to Burn Injury,” *Anal. Chem.*, vol. 91, no. 10, pp. 6541–6548, 2019.

[102] S. O. Diaz *et al.*, “Second trimester maternal urine for the diagnosis of trisomy 21 and prediction of poor pregnancy outcomes,” *J. Proteome Res.*, vol. 12, no. 6, pp. 2946–2957, 2013.

[103] G. Xie *et al.*, “Plasma metabolite biomarkers for the detection of pancreatic cancer,” *J. Proteome Res.*, vol. 14, no. 2, pp. 1195–1202, 2015.

[104] P. J. Trainor, R. V. Yampolskiy, and A. P. DeFilippis, “Wisdom of artificial crowds feature selection in untargeted metabolomics: An application to the development of a blood-based diagnostic test for thrombotic myocardial infarction,” *J. Biomed. Inform.*, vol. 81, pp. 53–60, 2018.

[105] F. Falchi *et al.*, “Kernel-Based, Partial Least Squares Quantitative Structure-Retention Relationship Model for UPLC Retention Time Prediction: A Useful Tool for Metabolite Identification,” *Anal. Chem.*, vol. 88, no. 19, pp. 9510–9517, 2016.

[106] M. Jiang, C. Wang, Y. Zhang, Y. Feng, Y. Wang, and Y. Zhu, “Sparse partial-least-squares discriminant analysis for different geographical origins of salvia miltiorrhiza by 1h-nmr-based metabolomics,” *Phytochem. Anal.*, vol. 25, no. 1, pp. 50–58, 2014.

[107] K. A. Lê Cao, P. G. P. Martin, C. Robert-Granié, and P. Besse, “Sparse canonical methods for biological data integration: Application to a cross-platform study,” *BMC Bioinformatics*, vol. 10, pp. 1–17, 2009.

[108] J. K. Haukka, N. Sandholm, C. Forsblom, J. E. Cobb, P. H. Groop, and E. Ferrannini, “Metabolomic Profile Predicts Development of Microalbuminuria in Individuals with Type 1 Diabetes,” *Sci. Rep.*, vol. 8, no. 1, pp. 1–10, 2018.

[109] M. Vinaixa, S. Samino, I. Saez, J. Duran, J. J. Guinovart, and O. Yanes, “A guideline to univariate statistical analysis for LC/MS-based untargeted metabolomics-derived data,” *Metabolites*, vol. 2, no. 4. pp. 775–795, 2012.

[110] D. López-Álvarez, H. Zubair, M. Beckmann, J. Draper, and P. Catalán, “Diversity and association of phenotypic and metabolomic traits in the close model grasses Brachypodium distachyon, B. stacei and B. hybridum,” *Ann.*

Bot., vol. 119, no. 4, pp. 545–561, 2017.

[111] R. A. Johnson, *AppliedMultivariateStatistics*. 2007.

[112] D. I. Broadhurst and D. B. Kell, “Statistical strategies for avoiding false discoveries in metabolomics and related experiments,” *Metabolomics*, vol. 2, no. 4, pp. 171–196, 2006.

[113] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, “Comparative Study of Attribute Selection Using Gain Ratio and Correlation Based Feature Selection,” *Int. J. Inf. Technol. Knowl. Manag.*, vol. 2, no. 2, pp. 271–277, 2010.

[114] M. Pirooznia, J. Y. Yang, M. Q. Yang, and Y. Deng, “A comparative study of different machine learning methods on microarray gene expression data.,” *BMC Genomics*, vol. 9 Suppl 1, p. S13, 2008.

[115] A. Jain and D. Zongker, “Feature selection: Evaluation, application, and small sample performance,” *Pattern Anal. Mach. Intell. …*, pp. 1–18, 1997.

[116] M. Tan, J. Pu, and B. Zheng, “Optimization of breast mass classification using sequential forward floating selection (SFFS) and a support vector machine (SVM) model,” *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, pp. 1005–1020, 2014.

[117] L. Ladha and T. Deepa, “Feature selection methods and algorithms,” *Int. J. Comput. Sci. Eng.*, vol. 3, no. 5, pp. 1787–1797, 2011.

[118] A. Marcano-Cedeno, J. Quintanilla-Dominguez, M. G. Cortina-Januchs, and D. Andina, “Feature selection using Sequential Forward Selection and classification applying Artificial Metaplasticity Neural Network,” *IECON Proc. (Industrial Electron. Conf.*, pp. 2845–2850, 2010.

[119] Isabelle Guyon, *Feature Extraction - Foundations and Applications*. 1937.

[120] D. Lu, A. Weljie, A. R. De Leon, Y. Mcconnell, O. F. Bathe, and K. Kopciuk, “Performance of variable selection methods using stability-based selection,” *BMC Res. Notes*, vol. 10, p. 143, 2017.

[121] A. Baratloo, M. Hosseini, A. Negida, G. El Ashal, and G. El Ashal, “Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity,” *Emergency; 2015 Press*, vol. 3, pp. 2–3, 2015.

[122] A. Janecek, W. N. W. Gansterer, M. Demel, and G. Ecker, “On the Relationship Between Feature Selection and Classification Accuracy.,” *Fsdm*, vol. 4, pp. 90–105, 2008.

[123] S. Mahadevan, S. L. Shah, T. J. Marrie, and C. M. Slupsky, “Analysis of metabolomic data using support vector machines,” *Anal. Chem.*, vol. 80, no. 19, pp. 7562–7570, 2008.

[124] P. S. Gromski *et al.*, “A tutorial review: Metabolomics and partial least squares-discriminant analysis - a marriage of convenience or a shotgun wedding,” *Anal. Chim. Acta*, vol. 879, pp. 10–23, 2015.

[125] T. K. Paul, “Extraction of Informative Genes from Microarray Data,” *Evaluation*, pp. 453–460, 2005.

[126] S. Dinakaran and P. Ranjit Jeba Thangaiah, “Comparative analysis of filter-wrapper approach for random forest performance on multivariate data,” *Proc. - 2014 Int. Conf. Intell. Comput. Appl. ICICA 2014*, pp. 174–178, 2014.

[127] C. H. Siddique, J., de Chavez, P. J., Howe, G., Cruden, G., & Brown, “Limitations in using multiple imputation to harmonize individual participant data for meta-analysis,” *Physiol. Behav.*, vol. 176, no. 3, pp. 139–148, 2017.

[128] K. T. Do *et al.*, “Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies,”

Metabolomics, vol. 14, no. 10, pp. 1–18, 2018.

[129] L. Beretta and A. Santaniello, "Nearest neighbor imputation algorithms: A critical evaluation," *BMC Med. Inform. Decis. Mak.*, vol. 16, no. Suppl 3, 2016.

[130] M. P. Lee. J. Y., & Styczynski, "NS-kNN: A modified k-nearest neighbors approach for imputing metabolomics data," *Physiol. Behav.*, vol. 176, no. 1, pp. 139–148, 2016.

[131] R. A. van den Berg, H. C. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf, "Centering, scaling, and transformations: Improving the biological information content of metabolomics data," *BMC Genomics*, vol. 7, no. 142, 2006.

[132] R. Tibshirani and L. Wasserman, "Correlation-sharing for Detection of Differential Gene Expression," *Carnegie Mellon Univ.*, 2006.

[133] F. Santosa and W. W. Symes, "Linear Inversion of Band-Limited Reflection Seismograms," *SIAM J. Sci. Stat. Comput.*, vol. 7, no. 4, pp. 1307–1330, Oct. 1986.

[134] J. B. Moore, "Regression Shrinkage and Selection via the Lasso," *Proc. Am. Soc. Int. Law its Annu. Meet.*, vol. 9, no. 1, pp. 11–23, 1915.

[135] A. Agresti, *An Introduction to Categorical Data Analysis _ Second Edition*. 2007.

[136] S. Abe, *Support Vector Machines for Pattern Classification*, vol. 26. 2010.

[137] C. Gao *et al.*, "Model-based and Model-free Machine Learning Techniques for Diagnostic Prediction and Classification of Clinical Outcomes in Parkinson's Disease," *Sci. Rep.*, vol. 8, no. 1, p. 7129, 2018.

[138] X. Liang, J. A. Bradley, D. Zheng, M. Rutenberg, D. Yeung, and N. Mendenhall, "Prognostic factors of radiation dermatitis following passive-scattering proton therapy for breast cancer," *Radiat. Oncol.*, vol. 13, no. 72, pp. 1–8, 2018.

[139] S. Prabhakaran, "InformationValue: Performance analysis and companion functions for binary classification models," *R Packag. version 1.2.3*, 2016.

[140] J. A. Hageman, R. A. van den Berg, J. A. Westerhuis, H. C. J. Hoefsloot, and A. K. Smilde, "Bagged K-means clustering of metabolome data," *Crit. Rev. Anal. Chem.*, vol. 36, no. 3–4, pp. 211–220, 2006.

[141] R. S. Mayor A, "Optimisation of Sample Preparation for Direct SPME-GC-MS Analysis of Murine and Human Faecal Volatile Organic Compounds for Metabolomic Studies," *J. Anal. Bioanal. Tech.*, vol. 5, no. 2, 2014.

[142] T. Abaffy, M. G. Möller, D. D. Riemer, C. Milikowski, and R. A. DeFazio, "Comparative analysis of volatile metabolomics signals from melanoma and benign skin: A pilot study," *Metabolomics*, vol. 9, no. 5, pp. 998–1008, 2013.

[143] L. F. Campbell, L. Farmery, S. M. C. George, and P. B. J. Farrant, "Canine olfactory detection of malignant melanoma," *BMJ Case Rep.*, pp. 1–3, 2013.

[144] M. Rossi *et al.*, "Volatile Organic Compounds in Feces Associate With Response to Dietary Intervention in Patients With Irritable Bowel Syndrome," *Clin. Gastroenterol. Hepatol.*, vol. 16, no. 3, p. 385–391.e1, 2018.

[145] R. B. M. Aggio, P. White, H. Jayasena, B. de Lacy Costello, N. M. Ratcliffe, and C. S. J. Probert, "Irritable bowel syndrome and active inflammatory bowel disease diagnosed by faecal gas analysis," *Aliment. Pharmacol. Ther.*, vol. 45, no. 1, pp. 82–90, 2017.

[146] R. D. Couch *et al.*, "The approach to sample acquisition and its impact on the derived human fecal microbiome and VOC metabolome," *PLoS One*, vol. 8,

no. 11, pp. 1–13, 2013.

[147] Y. H. Yun, B. C. Deng, D. S. Cao, W. T. Wang, and Y. Z. Liang, “Variable importance analysis based on rank aggregation with applications in metabolomics for biomarker discovery,” *Anal. Chim. Acta*, vol. 911, no. June 2015, pp. 27–34, 2016.

APPENDICES

Appendix I: R codes

#Simulated data (for example, 200 variables with x_1 and x_2 was set as 2 discriminatory variables and 1 binary outcome (group 0 and 1), n=100 and ρ = 0.8.

```
#Call the libraries
library(MASS)
library(InformationValue)
library(plyr)
library(st)
library(sda)
library(pROC)
library(glmnet)
library(stringi)
library(corrgram)

#Set the seed as 110 for the reproducible results
set.seed(110)
meanx1group0 = 0
meanx2group0 = 0
meanx1group1 = 0.5
meanx2group1 = 1
stddev = c(1,1)

#Set the correlation matrix
corMat = matrix(c(1,0.8,0.8,1),ncol = 2)
corVar = stddev %*% t(stddev) * corMat

#Generate a bivariate data which there is association between
x1 and x2 (Set them as discriminatory variables)
```

```
x1x2_group0 =
mvrnorm(50,c(meanx1group0,meanx2group0),Sigma = corVar,
empirical = TRUE)
x1x2_group1 =
mvrnorm(50,c(meanx1group1,meanx2group1),Sigma = corVar,
empirical = TRUE)
x1x2 = rbind(x1x2_group0,x1x2_group1)

#Set x3 until x200 as the not discriminatory variables
x3to200 = matrix(rnorm(100*198,0,1), 100, 198);
x <- cbind(x1x2,x3to200)
colnames(x) <- paste0("x",     1:ncol(x));

#Set the outcome groups
y = data.frame(rep(100))
y[1:50,1] = 0
y[51:100,1] = 1
colnames(y) <- paste0("y")
mode(x) = "numeric"
data <- data.frame(y,x)

#Variable selection methods
#T method function
Tfunction=function(trainset,data){
tmp = centroids(as.matrix(trainset[,2:ncol(data)]),
as.matrix(trainset[,1]), var.groups = FALSE,
centered.data = TRUE,
                lambda.var = 0, lambda.freqs = 0, verbose
= TRUE )
diff = tmp$means[, 1] - tmp$means[, 2]
n1 = tmp$samples[1]
n2 = tmp$samples[2]
v = tmp$variances[, 1]
sd = sqrt((1/n1 + 1/n2) * v)
t = abs(diff/sd)
```

```
idx = order((t),decreasing = TRUE)
return(idx)}

#corT method function
    corTfunction=function(trainset,data){
    tmp = centroids(as.matrix(trainset[,2:ncol(data)]),
as.matrix(trainset[,1]), var.groups = FALSE,
centered.data = TRUE,
                    lambda.var = 0, lambda.freqs = 0,
verbose = TRUE )
    diff = tmp$means[, 1] - tmp$means[, 2]
    n1 = tmp$samples[1]
    n2 = tmp$samples[2]
    v = tmp$variances[, 1]
    sd = sqrt((1/n1 + 1/n2) * v)
    R = cor(tmp$centered.data)
    t = diff/sd
    p = length(t)
    cst.vec = numeric(p)
    for (i in 1:p) {
      idx = order(R[i, ], decreasing = TRUE)
      nonneg = sum(R[i, ] >= 0)
      z = cumsum(abs(t[idx[1:nonneg]]))/1:nonneg
      cst.vec[i] = max(z) * sign(t[i])
    }
    idx = order(abs(cst.vec), decreasing=TRUE)
        return(idx)
    }

#adjcorT method function
adjcorTfunction=function(trainset,data){
  tmp = centroids(as.matrix(trainset[,2:ncol(data)]),
as.matrix(trainset[,1]), var.groups = FALSE,
centered.data = TRUE,
               lambda.var = 0, lambda.freqs = 0, verbose
= TRUE )
diff = tmp$means[, 1] - tmp$means[, 2]
n1 = tmp$samples[1]
n2 = tmp$samples[2]
v = tmp$variances[, 1]
sd = sqrt((1/n1 + 1/n2) * v)
R = cor(tmp$centered.data)
t = diff/sd
p = length(t)
cst.vec = numeric(p)
for (i in 1:p) {
  idx = order(abs(R[i, ]), decreasing = TRUE)
  nonneg = sum(abs(R[i, ])>= 0)
  z = cumsum(abs(t[idx[1:nonneg]]))/1:nonneg
  cst.vec[i] = max(z) * sign(t[i])
}
idx = order(abs(cst.vec), decreasing=TRUE)
return(idx)
}

#T, corT, adjcorT, Lasso and logistic model
#Create matrixes that store the results
T_results_select=matrix(NA,ncol=2,nrow=100)
corT_results_select=matrix(NA,ncol=2,nrow=100)
adjcorT_results_select=matrix(NA,ncol=2,nrow=100)
select=matrix(NA,ncol = 2,nrow = 100)
select1=matrix(NA,ncol = 2,nrow = 100)
dataCorr =matrix(NA,ncol=2,nrow = nrow(trainset))
error_T=sen_T=spe_T=auroc_T=error_corT=sen_corT=spe_cor
T=auroc_corT=error_adjcorT=sen_adjcorT=spe_adjcorT=auro
c_adjcorT=error_lasso=sen_lasso=spe_lasso=auroc_lasso=m
atrix(NA,ncol=1,nrow=100)
i=1
```

```
#Run the variable selection process in 100 iterations
for(i in 1:100){
set.seed(110+i)
smp_size <- floor(0.80 * nrow(data))
train_ind <- sample(seq_len(nrow(data)), size = smp_size)
trainset <- data[train_ind, ]
testset <- data[-train_ind, ]
dim(trainset)
dim(testset)
#trainset[complete.cases(trainset)]
testset=data.frame(testset)
trainset=data.frame(trainset)

#Top two variables selected by T
T_results=Tfunction(trainset,data)
T_results_select[i,]=T_results[1:2]

#Top two variables selected by corT
corT_results=corTfunction(trainset,data)
corT_results_select[i,]=corT_results[1:2]

#Top two variables selected by adjcorT
adjcorT_results=adjcorTfunction(trainset,data)
adjcorT_results_select[i,]=adjcorT_results[1:2]

#Logistic regression that includes the top two selected
variables by T
model_T<-glm(trainset[,1]~
trainset[,(1+c(T_results_select[i,][1]))]+trainset[,(1+
c(T_results_select[i,][2]))],
family=binomial,data=trainset)

#Logistic regression that includes the top two selected
variables by corT
```

```
cormodel_T <- glm(trainset[,1]~
trainset[,(1+c(corT_results_select[i,][1]))]+trainset[,
(1+c(corT_results_select[i,][2]))],
family=binomial,data=trainset)

#Logistic regression that includes the top two selected
variables by adjcorT
adjcormodel_T <- glm(trainset[,1]~
trainset[,(1+c(adjcorT_results_select[i,][1]))]+trainse
t[,(1+c(adjcorT_results_select[i,][2]))],
family=binomial,data=trainset)

xbeta_T=exp(model_T$coefficients[1]+model_T$coefficient
s[2]*testset[,(1+c(T_results_select[i,][1]))]+model_T$c
oefficients[3]*testset[,(1+c(T_results_select[i,][2]))]
)
predicted_T=xbeta_T/(1+xbeta_T)
model_pred_y_T <- rep("0", nrow(testset))
model_pred_y_T[predicted_T > 0.5] = "1"

error_T[i,]                  =misClassError(testset[,1],
predicted_T,threshold = 0.5)
sen_T[i,]   =   sensitivity(testset[,1],   predicted_T,
threshold = 0.5)
spe_T[i,]   =   specificity(testset[,1],   predicted_T,
threshold = 0.5)
auroc_T[i,]= AUROC(testset[,1], predicted_T)

xbeta_corT=exp(cormodel_T$coefficients[1]+cormodel_T$co
efficients[2]*testset[,(1+c(corT_results_select[i,][1])
)]+cormodel_T$coefficients[3]*testset[,(1+c(corT_result
s_select[i,][2]))])
predicted_corT=xbeta_corT/(1+xbeta_corT)
model_pred_y_corT <- rep("0", nrow(testset))
model_pred_y_corT[predicted_corT > 0.5] = "1"
```

```
error_corT[i,]              =misClassError(testset[,1],
predicted_corT,threshold = 0.5)
sen_corT[i,] = sensitivity(testset[,1], predicted_corT,
threshold = 0.5)
spe_corT[i,] = specificity(testset[,1], predicted_corT,
threshold = 0.5)
auroc_corT[i,]= AUROC(testset[,1], predicted_corT)
xbeta_adjcorT=exp(adjcormodel_T$coefficients[1]+adjcorm
odel_T$coefficients[2]*testset[,(1+c(adjcorT_results_se
lect[i,][1]))]+adjcormodel_T$coefficients[3]*testset[,(
1+c(adjcorT_results_select[i,][2]))])
predicted_adjcorT=xbeta_adjcorT/(1+xbeta_adjcorT)
model_pred_y_adjcorT <- rep("0", nrow(testset))
model_pred_y_adjcorT[predicted_adjcorT > 0.5] = "1"


error_adjcorT[i,]           =misClassError(testset[,1],
predicted_adjcorT,threshold = 0.5)
sen_adjcorT[i,]       =       sensitivity(testset[,1],
predicted_adjcorT, threshold = 0.5)
spe_adjcorT[i,]       =       specificity(testset[,1],
predicted_adjcorT, threshold = 0.5)
auroc_adjcorT[i,]= AUROC(testset[,1], predicted_adjcorT)

#Lasso model
TRfit<-
cv.glmnet(as.matrix(trainset[,2:ncol(trainset)]),as.mat
rix(trainset[,1]),nfolds=10) #initial fit
lassofit=glmnet(as.matrix(trainset[,2:ncol(trainset)]),
as.matrix(trainset[,1]),family="binomial",alpha  =  1,
lambda=TRfit$lambda.min,standardize=T)
coef=as.vector(lassofit$beta)
nonzerocoef=coef[coef!=0][1:lassofit$df]
coef1=cbind(colnames(trainset[,2:ncol(trainset)],coef))
```

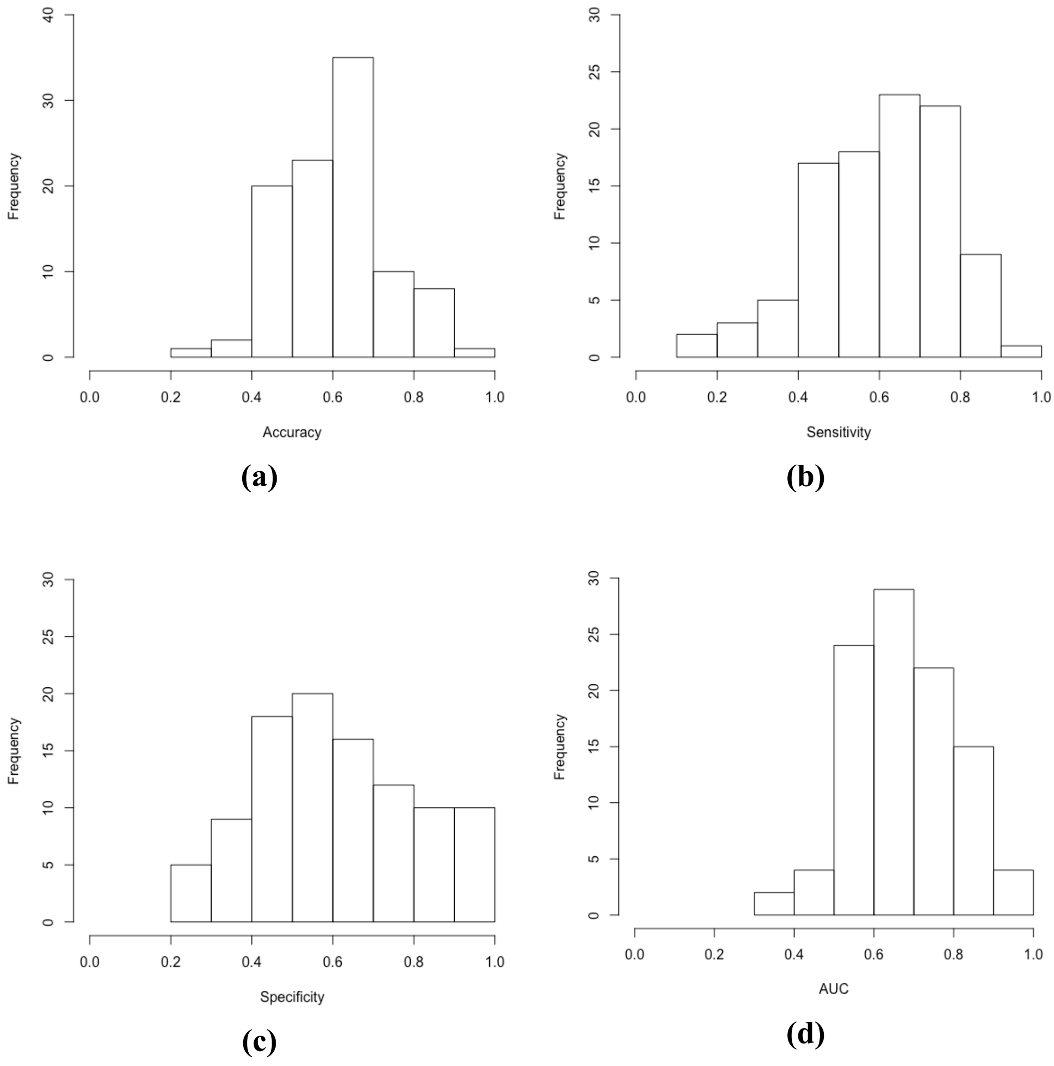
```
coef2=cbind(1:(ncol(trainset)-1),coef)
nonzerocoef2=coef2[coef!=0][1:lassofit$df]
nonzeroX=coef1[coef!=0][1:lassofit$df]

#The first variable selected by Lasso
select[i,1] = nonzeroX[1]

#The second variable selected by Lasso
select[i,2] = nonzeroX[2]

if (!is.na(select[i,1]) && (!is.na(select[i,2])))
{
 select1[i,] =  stri_sub(select[i,],2)
}



if (is.na(select1[i,1]) && (is.na(select1[i,2])))
{
  error_lasso[i,] = NA
  sen_lasso[i,] = NA
  spe_lasso[i,] = NA
  auroc_lasso[i,] = NA
}
else
{
if (is.na(select1[i,2]) && (!is.na(select1[i,1])))
{
  model_lasso            <-            glm(trainset[,1]~
trainset[,(1+c(as.numeric(as.character(select1[i,][1]))
))] , family=binomial,data=trainset)

xbeta_lasso=exp(model_lasso$coefficients[1]+model_lasso
$coefficients[2]*testset[,(1+c(as.numeric(as.character(
select1[i,][1]))))])
  predicted_lasso=xbeta_lasso/(1+xbeta_lasso)
  model_pred_y_lasso <- rep("0", nrow(testset))
  model_pred_y_lasso[predicted_lasso > 0.5] = "1"

  error_lasso[i,]            =misClassError(testset[,1],
predicted_lasso,threshold = 0.5)
  sen_lasso[i,]        =        sensitivity(testset[,1],
predicted_lasso, threshold = 0.5)
  spe_lasso[i,]        =        specificity(testset[,1],
predicted_lasso, threshold = 0.5)
  auroc_lasso[i,]= AUROC(testset[,1], predicted_lasso)
}
else
{
  model_lasso            <-            glm(trainset[,1]~
trainset[,(1+c(as.numeric(as.character(select1[i,][1]))
))]
+trainset[,(1+c(as.numeric(as.character(select1[i,][2])
)))], family=binomial,data=trainset)

xbeta_lasso=exp(model_lasso$coefficients[1]+model_lasso
$coefficients[2]*testset[,(1+c(as.numeric(as.character(
select1[i,][1]))))]+model_lasso$coefficients[3]*testset
[,(1+c(as.numeric(as.character(select1[i,][2]))))])
  predicted_lasso=xbeta_lasso/(1+xbeta_lasso)
  model_pred_y_lasso <- rep("0", nrow(testset))
  model_pred_y_lasso[predicted_lasso > 0.5] = "1"

  error_lasso[i,]            =misClassError(testset[,1],
predicted_lasso,threshold = 0.5)
  sen_lasso[i,]        =        sensitivity(testset[,1],
predicted_lasso, threshold = 0.5)
  spe_lasso[i,]        =        specificity(testset[,1],
predicted_lasso, threshold = 0.5)
  auroc_lasso[i,]= AUROC(testset[,1], predicted_lasso)
```

```
}
}

}

#Average of classification accuracy, sensitivity, specificity and AUROC for T
mean_error_T=round((1-mean(error_T))*100,2)
mean_sen_T=round(mean(sen_T)*100,2)
mean_spe_T=round(mean(spe_T)*100,2)
mean_auroc_T=round(mean(auroc_T)*100,2)

mean_error_T
mean_sen_T
mean_spe_T
mean_auroc_T

#Average of classification accuracy, sensitivity, specificity and AUROC for corT
mean_error_corT=round((1-mean(error_corT))*100,2)
mean_sen_corT=round(mean(sen_corT)*100,2)
mean_spe_corT=round(mean(spe_corT)*100,2)
mean_auroc_corT=round(mean(auroc_corT)*100,2)

mean_error_corT
mean_sen_corT
mean_spe_corT
mean_auroc_corT

#Average of classification accuracy, sensitivity, specificity and AUROC for adjcorT
mean_error_adjcorT=round((1-mean(error_adjcorT))*100,2)
mean_sen_adjcorT=round(mean(sen_adjcorT)*100,2)
mean_spe_adjcorT=round(mean(spe_adjcorT)*100,2)
mean_auroc_adjcorT=round(mean(auroc_adjcorT)*100,2)
```

```
mean_error_adjcorT
mean_sen_adjcorT
mean_spe_adjcorT
mean_auroc_adjcorT

#Average of classification accuracy, sensitivity, specificity and AUROC for Lasso
error_lasso1 = na.omit(error_lasso)
sen_lasso1 = na.omit(sen_lasso)
spe_lasso1 =na.omit(spe_lasso)
auroc_lasso1 = na.omit(auroc_lasso)

mean_error_lasso=round((1-mean(error_lasso1))*100,2)
mean_sen_lasso=round(mean(sen_lasso1)*100,2)
mean_spe_lasso=round(mean(spe_lasso1)*100,2)
mean_auroc_lasso=round(mean(auroc_lasso1)*100,2)

mean_error_lasso
mean_sen_lasso
mean_spe_lasso
mean_auroc_lasso
```

Appendix II: Histograms of performance measures for *n* =76 and 300 based on 100 iterations

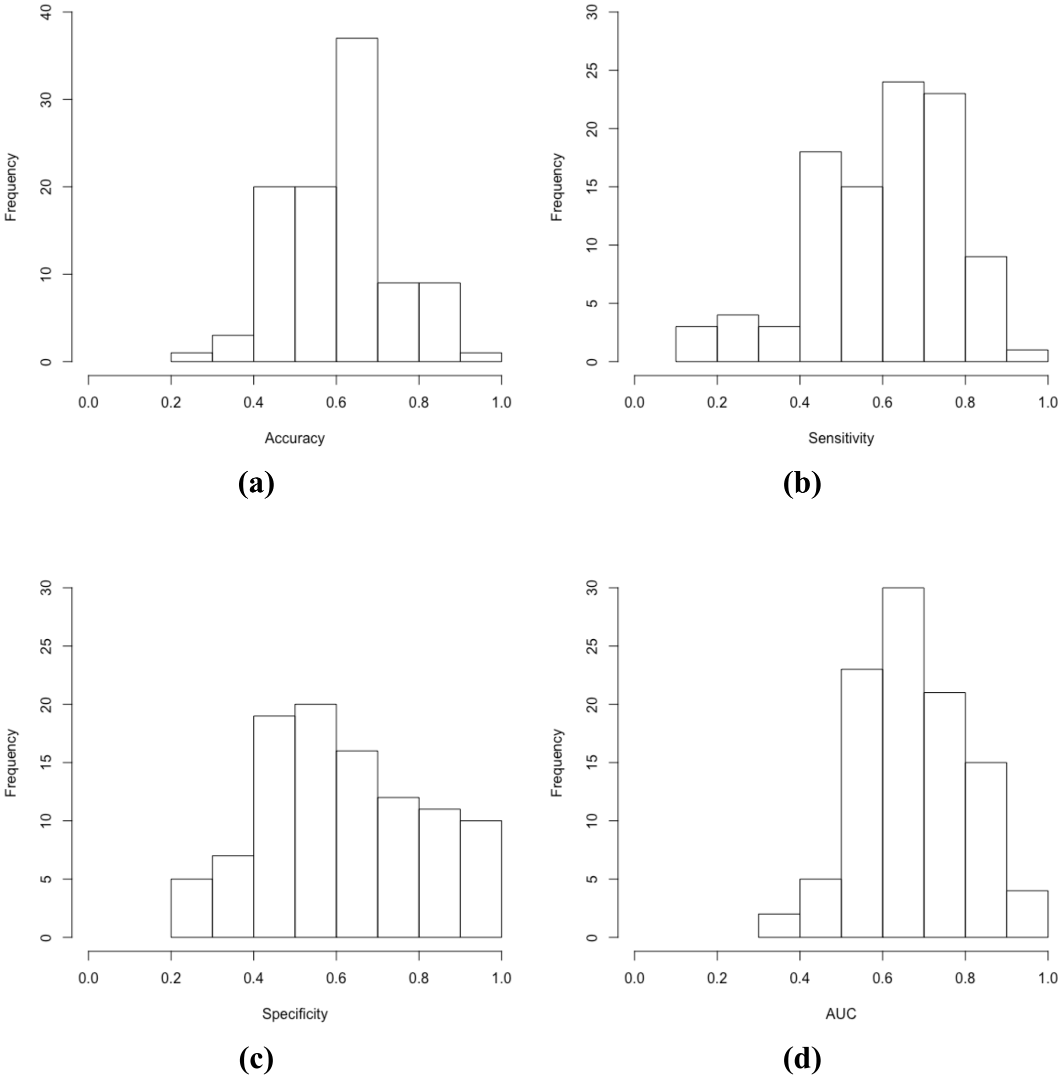
a) Histograms of performance measures for n=76 based on **100 iterations**



(a) **(b)**

(c) **(d)**

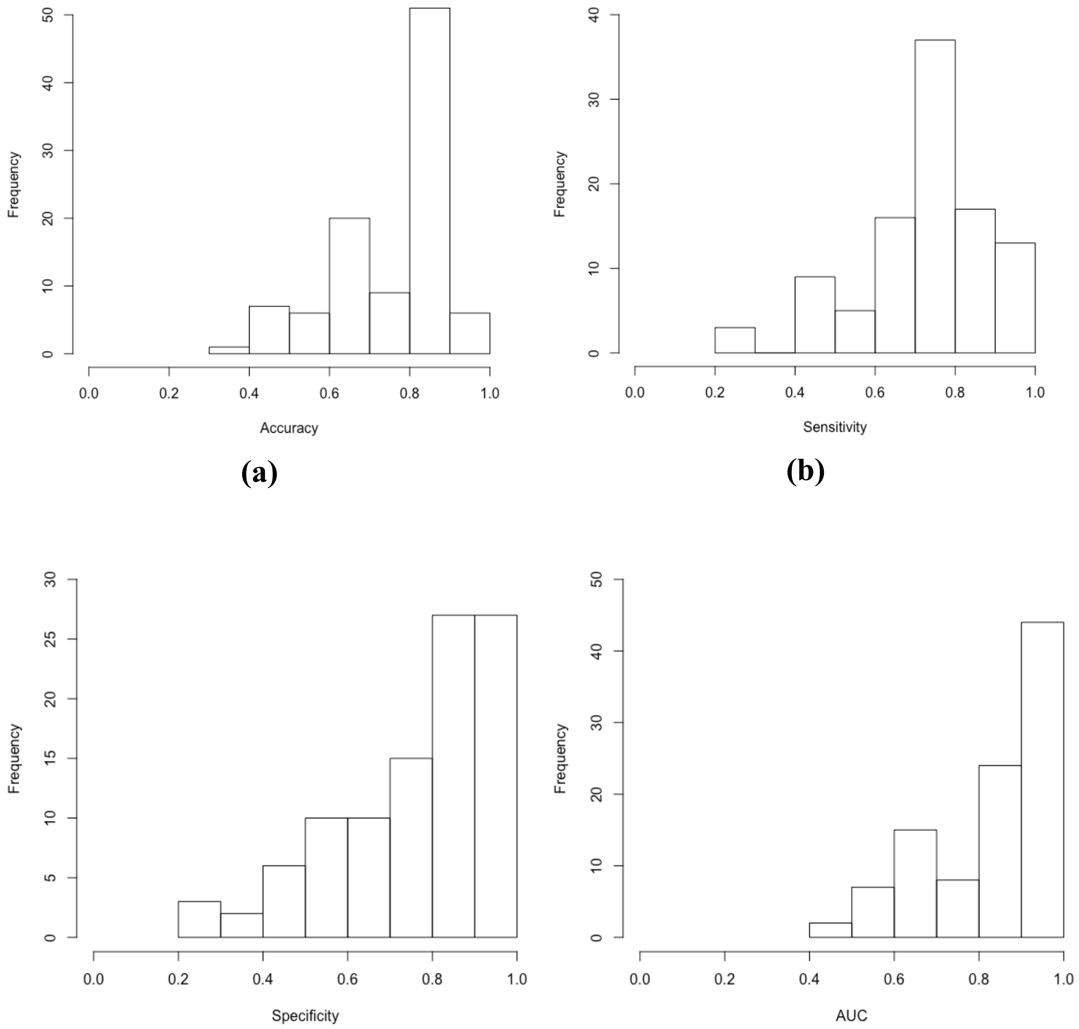
Figure S4. 1: Distribution of the estimates of **(a)** Accuracy, **(b)** Sensitivity, **(c)** Specificity and **(d)** AUC for sample size *n*=76 and ρ= -0.8, for the T method.



(a) (b)

(c) (d)

Figure S4. 2: Distribution of the estimates of **(a)** Accuracy, **(b)** Sensitivity, **(c)** Specificity and **(d)** AUC for sample size *n*=76 and ρ= -0.8, for the corT method.



(a) **(b)**

(c) **(d)**

Figure S4. 3: Distribution of the estimates of **(a)** Accuracy, **(b)** Sensitivity, **(c)** Specificity and **(d)** AUC for sample size *n*=76 and *ρ*= -0.8, for the adjcorT method.

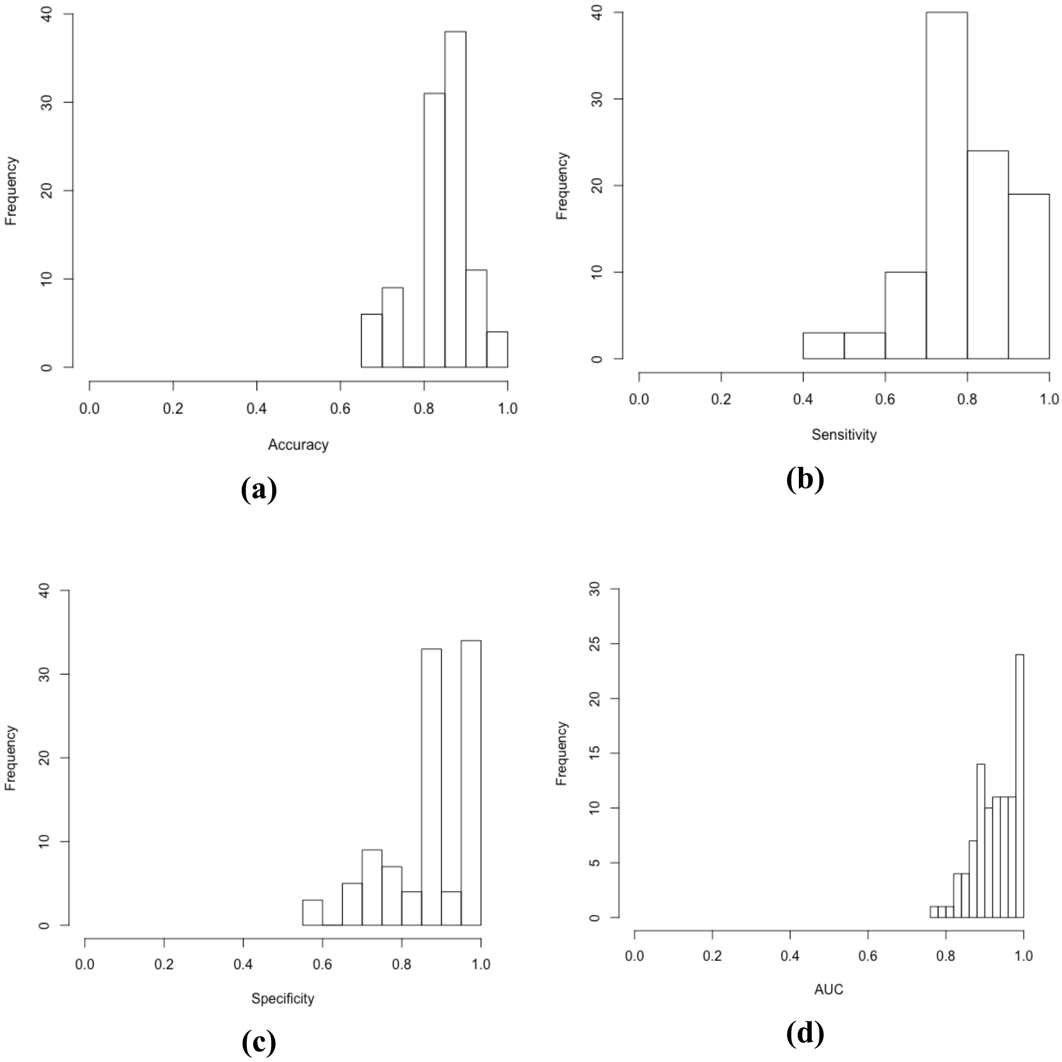


Figure S4. 4: Distribution of the estimates of **(a)** Accuracy, **(b)** Sensitivity, **(c)** Specificity and **(d)** AUC for sample size *n*=76 and *ρ*= -0.8, for the Lasso method.

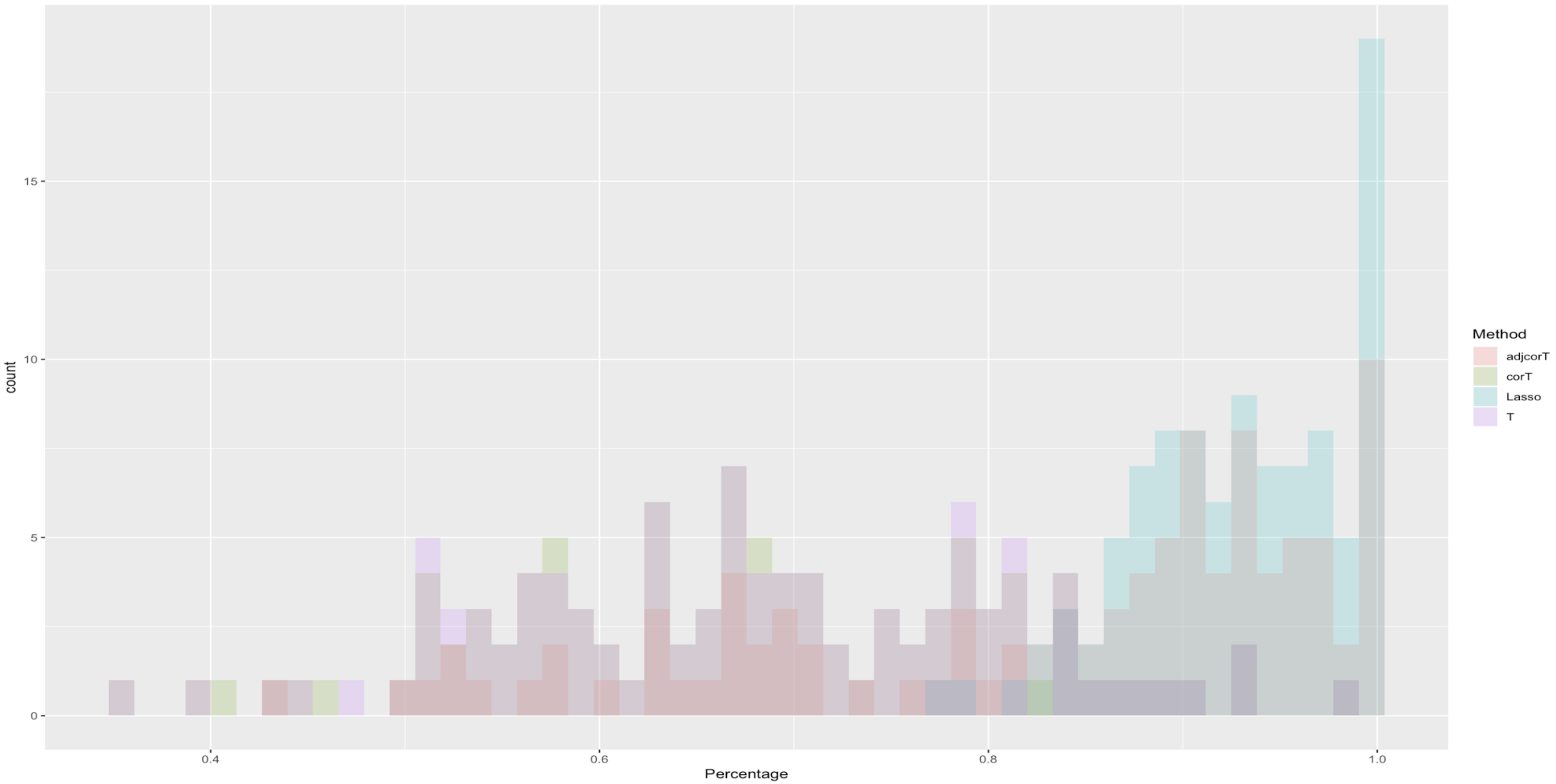
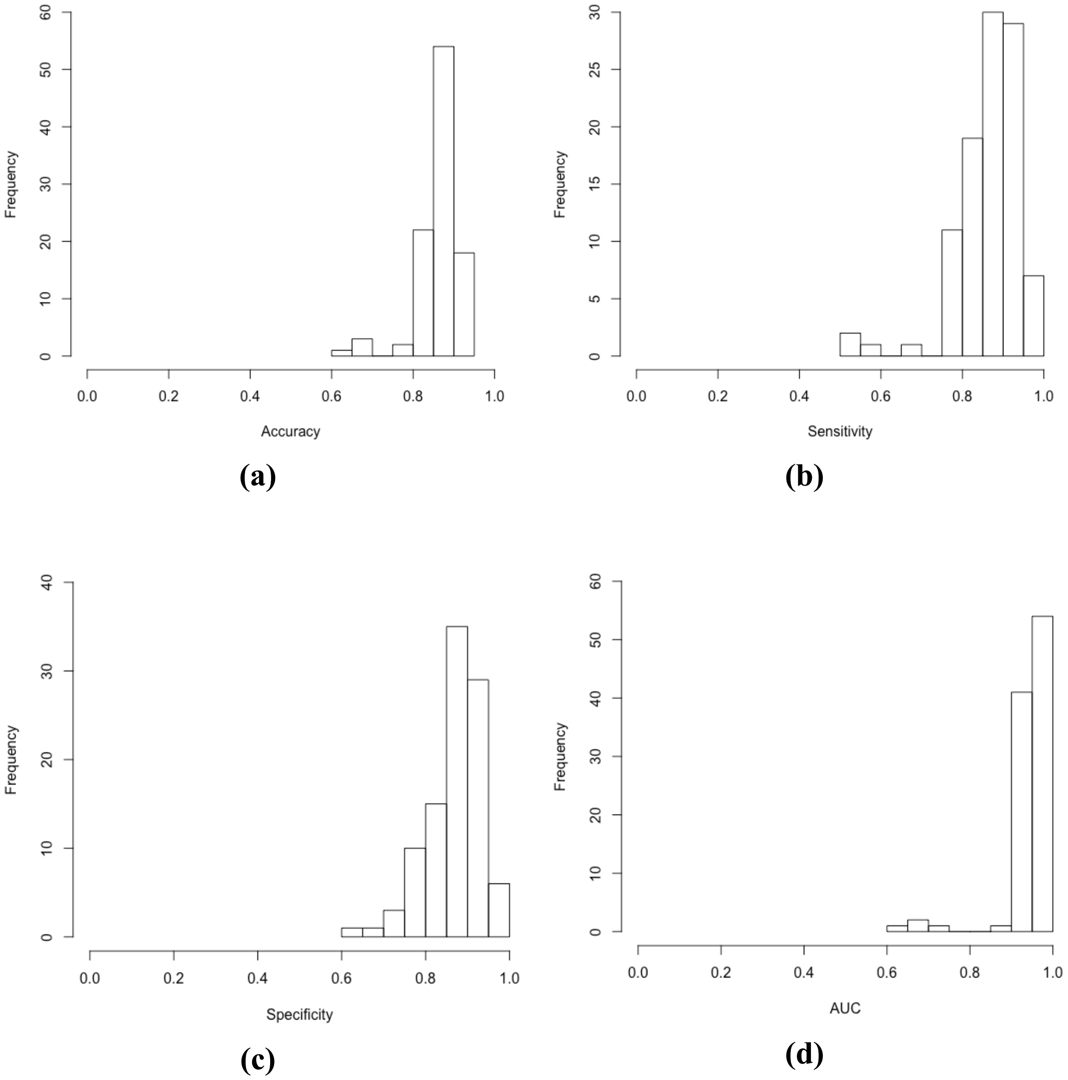


Figure S4. 5: Distributions of AUCs for sample size *n*=76 and *ρ* =-0.8 across methods (100 iterations)

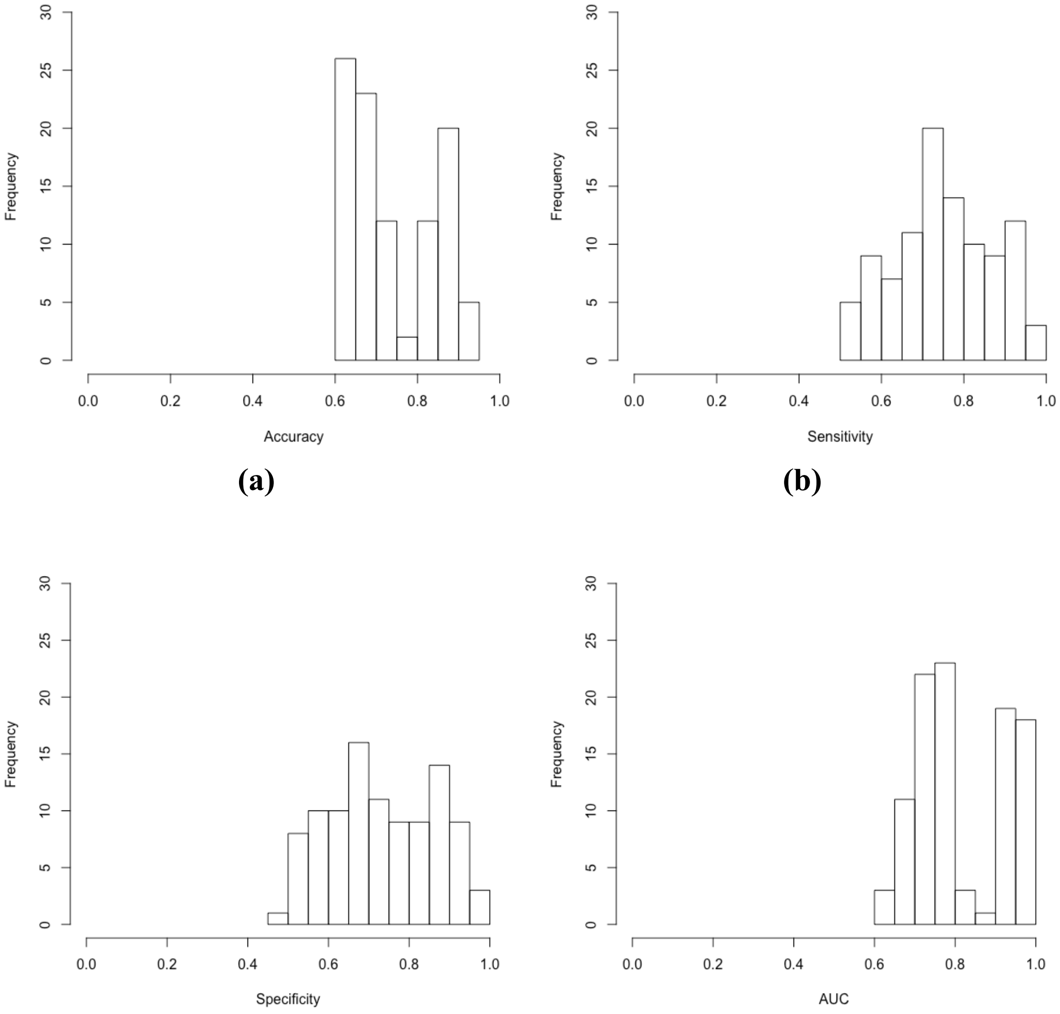
b) Histograms of performance measures for *n* =300 based on **100 iterations**



(a) **(b)**

(c) **(d)**

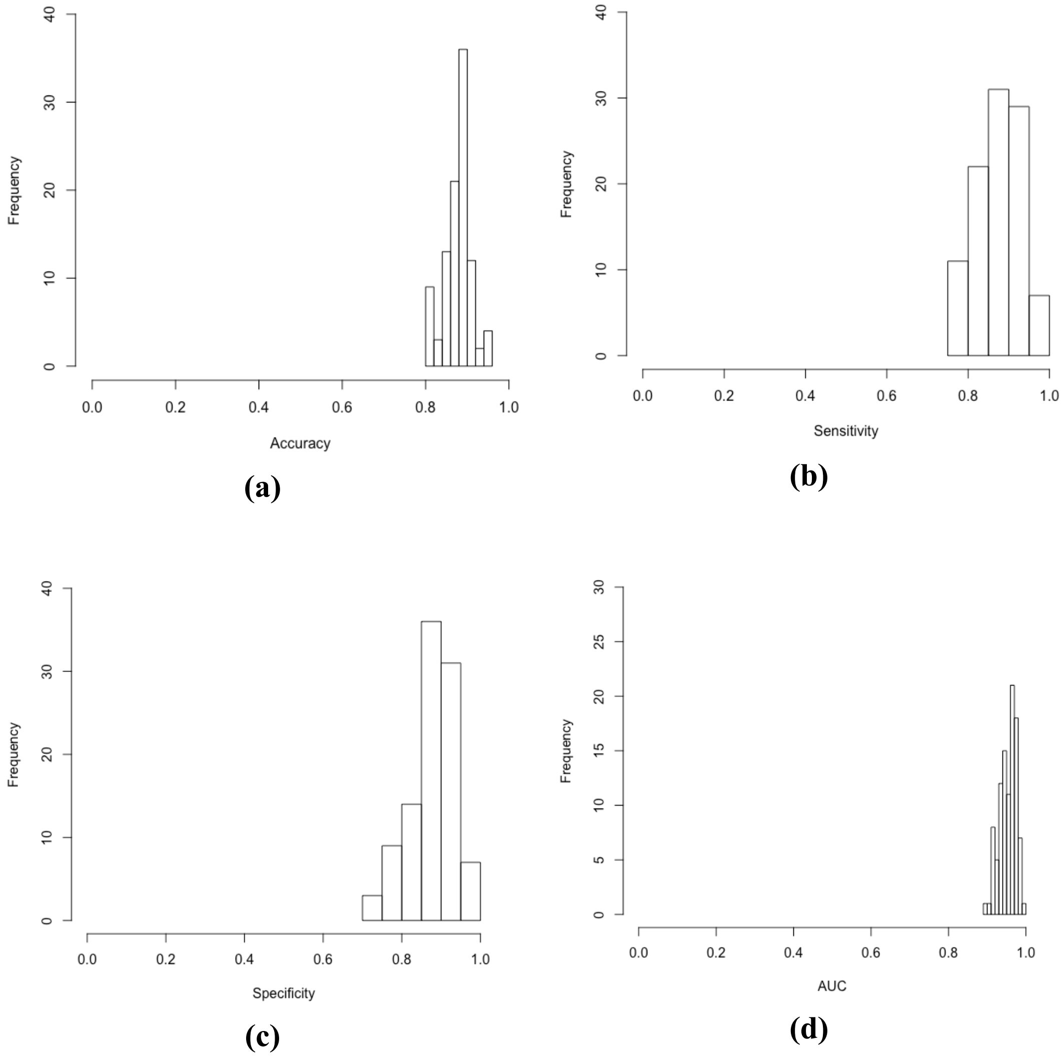
Figure S4. 6: Distribution of the estimates of **(a)** Accuracy, **(b)** Sensitivity, **(c)** Specificity and **(d)** AUC for sample size *n*=300 and *ρ*= -0.8, for the T method.



(a) (b)

(c) (d)

Figure S4. 7: Distribution of the estimates of **(a)** Accuracy, **(b)** Sensitivity, **(c)** Specificity and **(d)** AUC for sample size *n*=300 and *ρ*= -0.8, for the corT method.



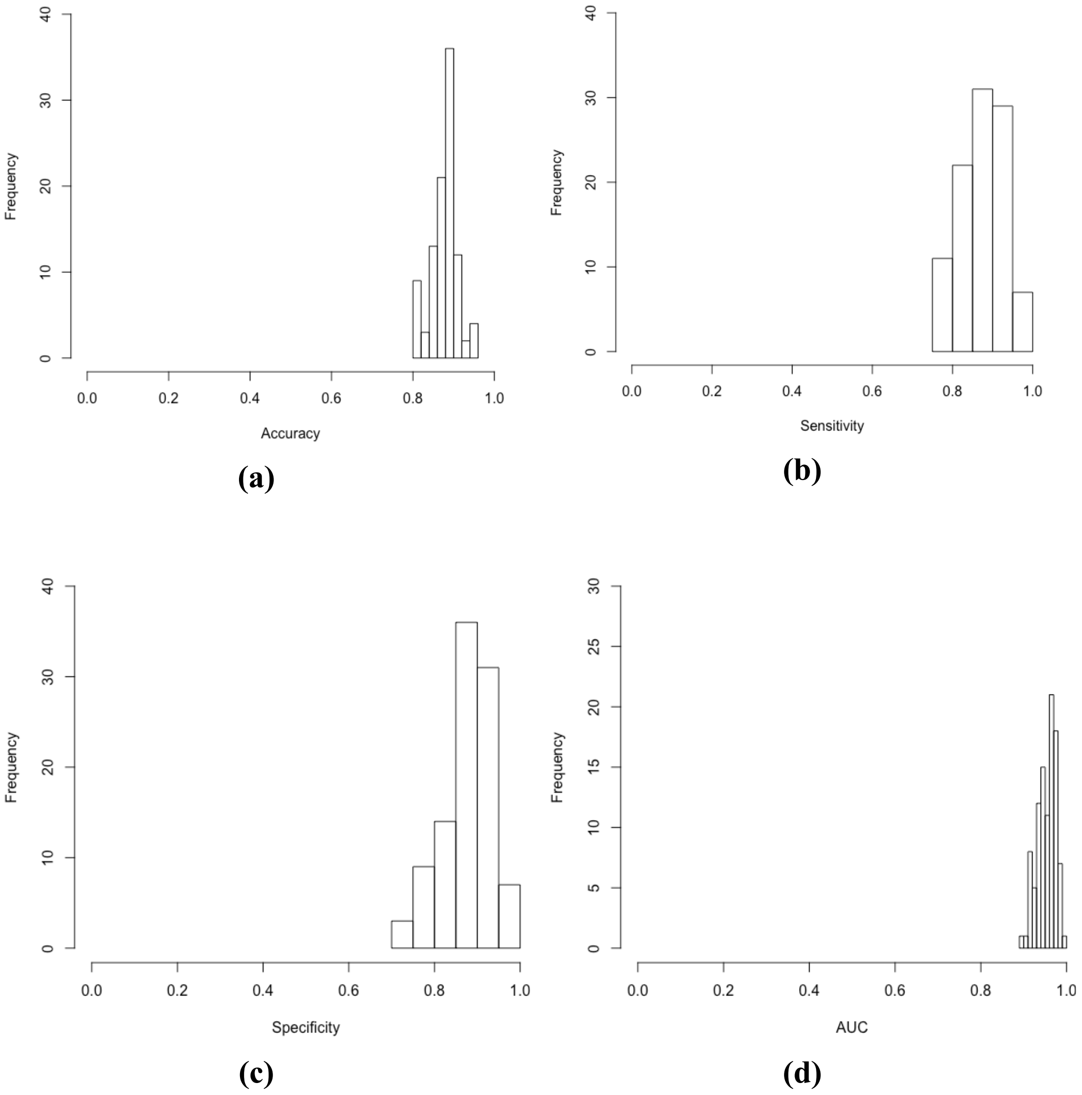
(a)

(b)

(c)

(d)

Figure S4. 8: Distribution of the estimates of **(a)** Accuracy, **(b)** Sensitivity, **(c)** Specificity and **(d)** AUC for sample size *n*=300 and *ρ*= -0.8, for the adjcorT method.



(a)

(b)

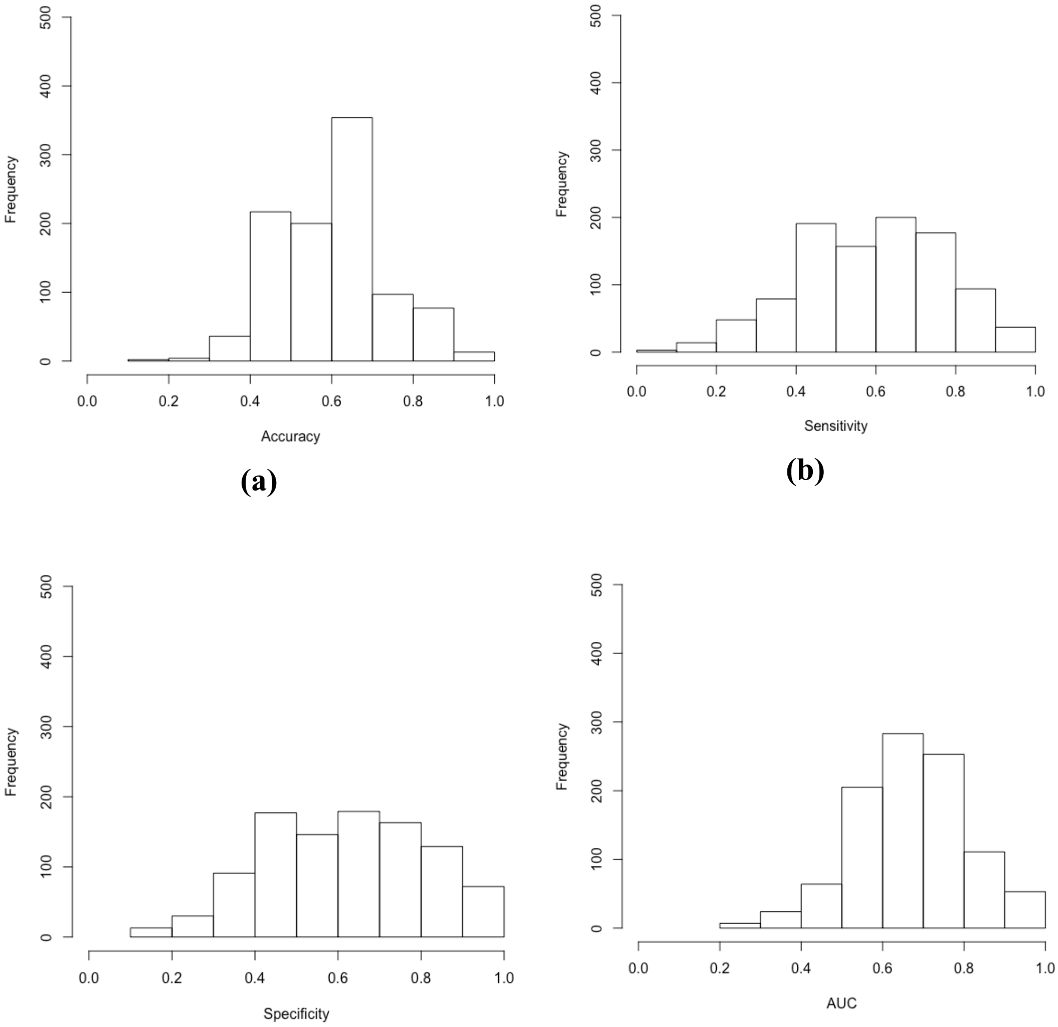
(c)

(d)

Figure S4. 9: Distribution of the estimates of **(a)** Accuracy, **(b)** Sensitivity, **(c)** Specificity and **(d)** AUC for sample size *n*=300 and *ρ*= -0.8, for the Lasso method

Appendix III: Histograms of performance measures for *n* =76 and 300 based on 1000 iterations

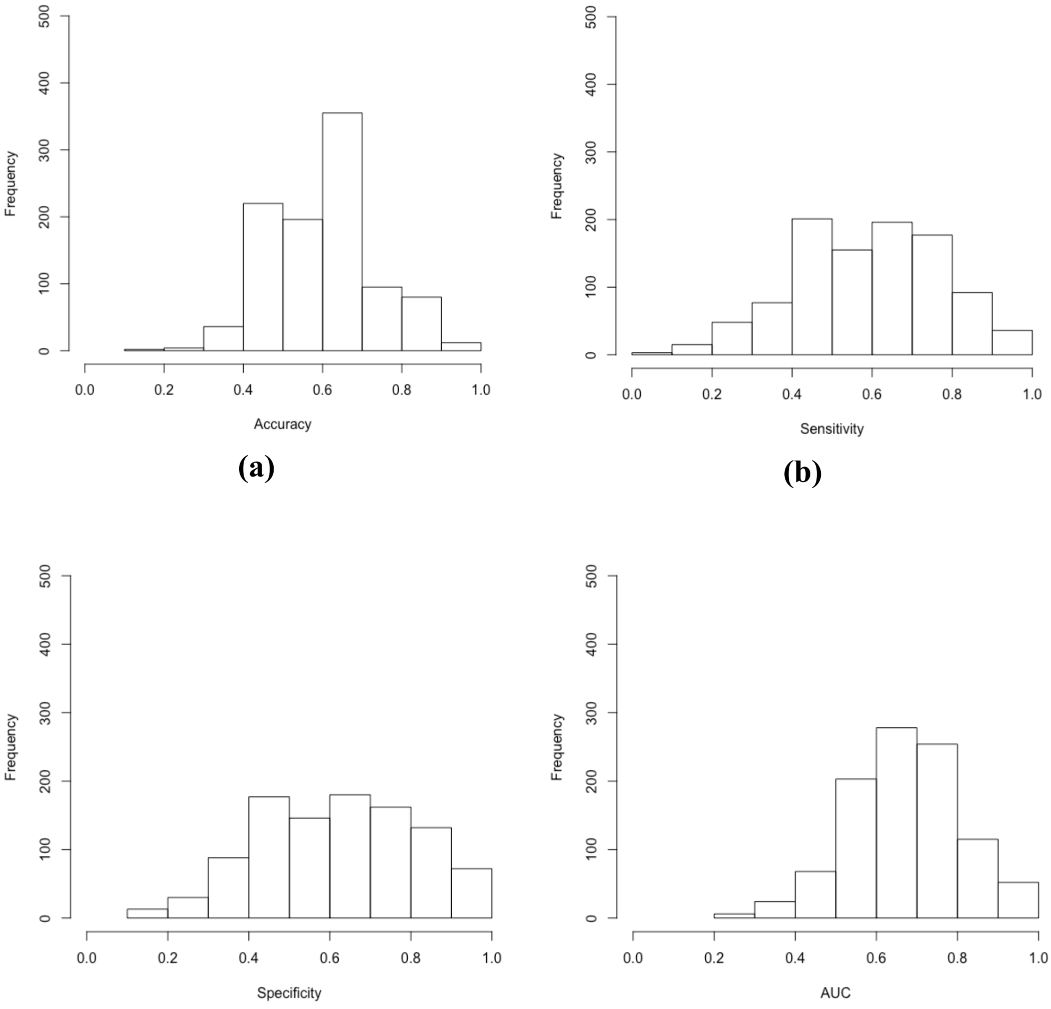
a) Histograms of performance measures for *n* =76 based on **1000 iterations**



(a) **(b)**

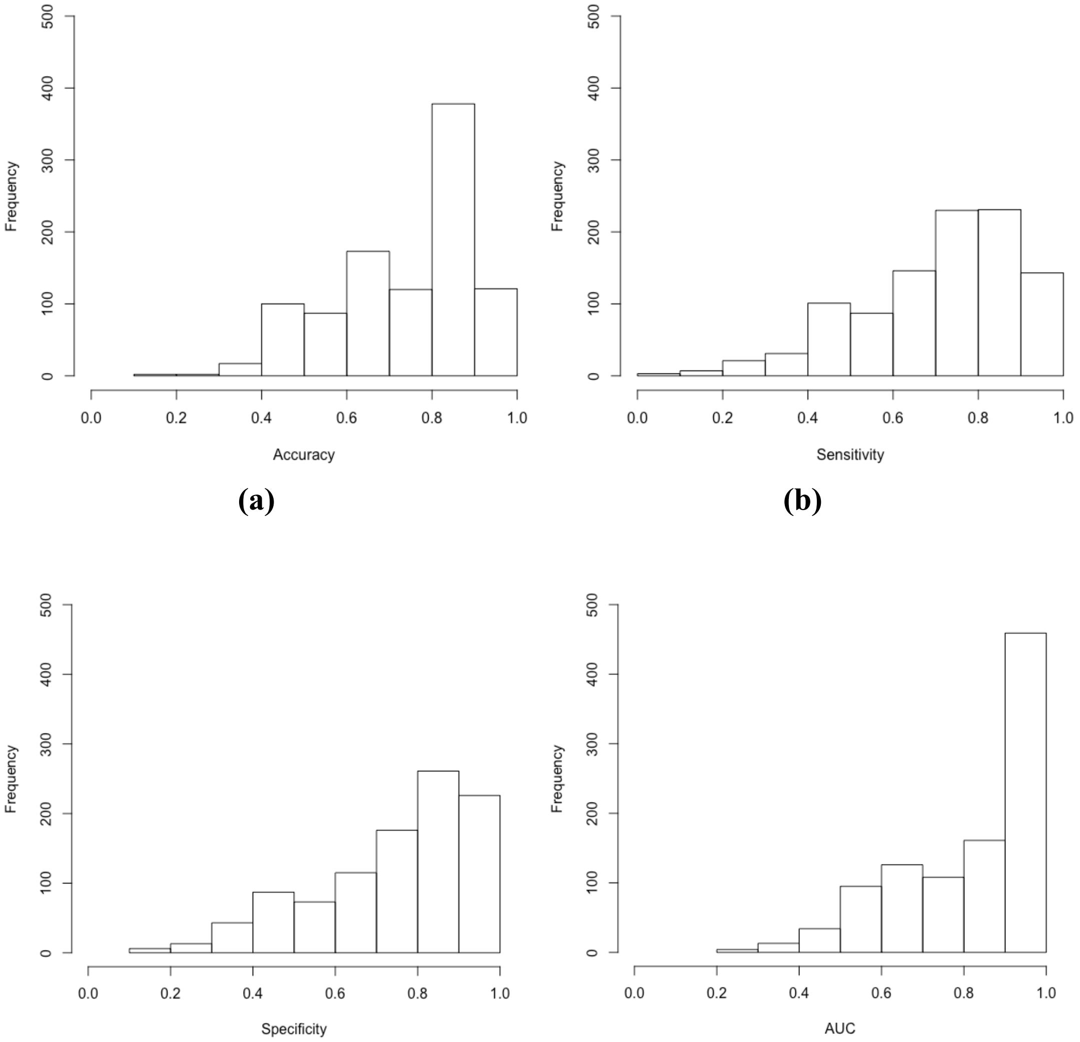
(c) **(d)**

Figure S4. 10: Distribution of the estimates of **(a)** Accuracy, **(b)** Sensitivity, **(c)** Specificity and **(d)** AUC for sample size *n* =76 and *ρ*= -0.8, for the T method.



(a) (b) (c) (d)

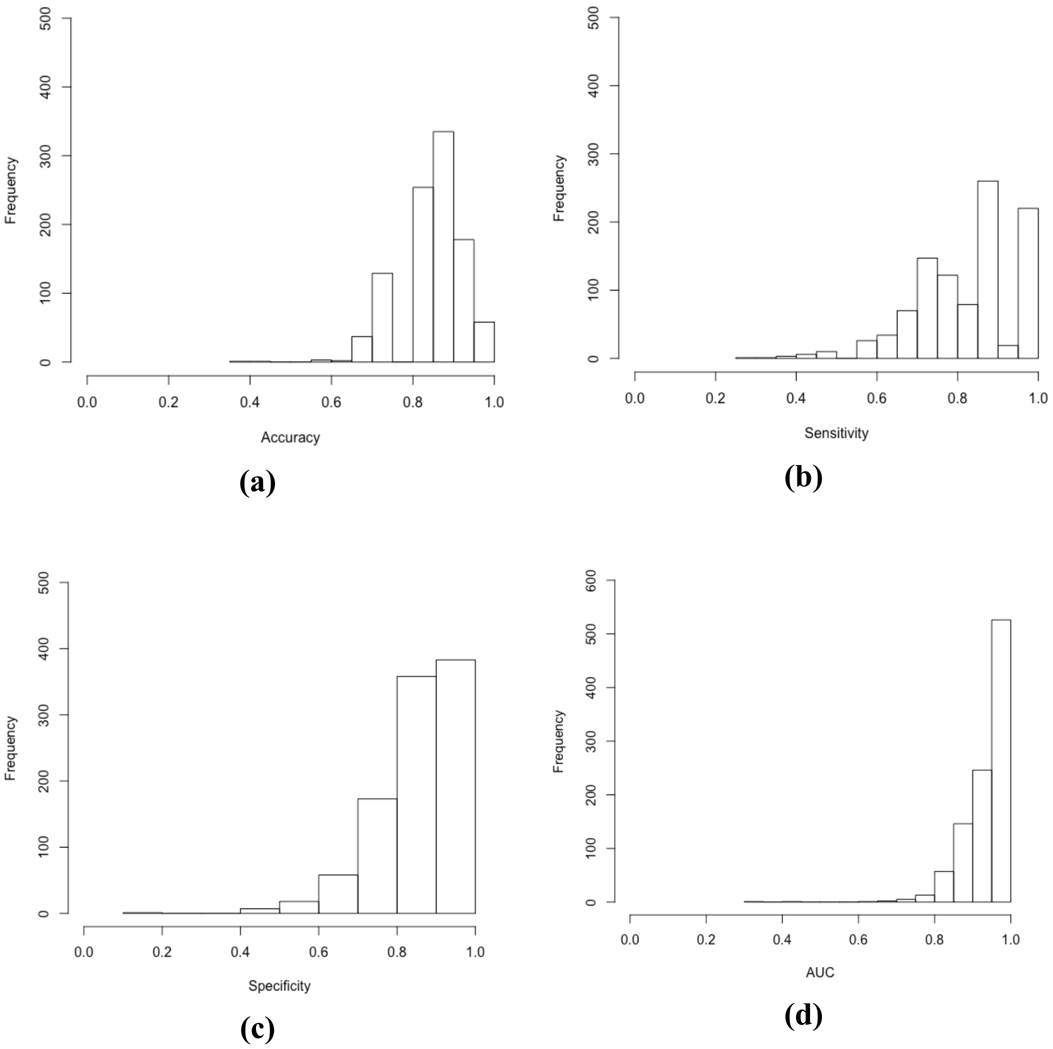
Figure S4. 11: Distribution of the estimates of **(a)** Accuracy, **(b)** Sensitivity, **(c)** Specificity and **(d)** AUC for sample size *n*=76 and *ρ*= -0.8, for the corT method.



(a) **(b)**

(c) **(d)**

Figure S4. 12: Distribution of the estimates of **(a)** Accuracy, **(b)** Sensitivity, **(c)** Specificity and **(d)** AUC for sample size *n* =76 and *ρ*= -0.8, for the adjcorT method.



(a)

(b)

(c)

(d)

Figure S4. 13: Distribution of the estimates of **(a)** Accuracy, **(b)** Sensitivity, **(c)** Specificity and **(d)** AUC for sample size *n*=76 and *ρ*= -0.8, for the Lasso method.

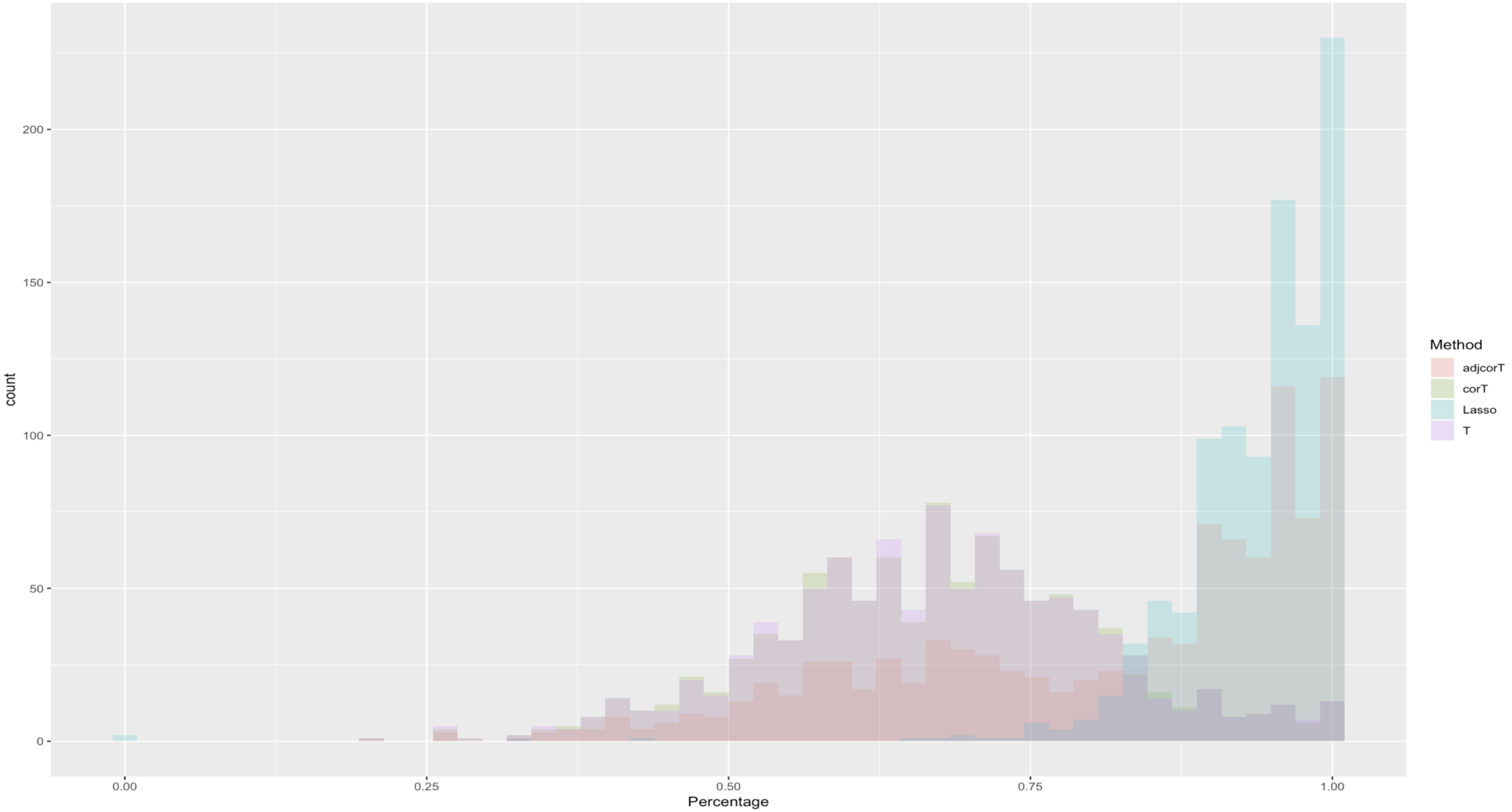
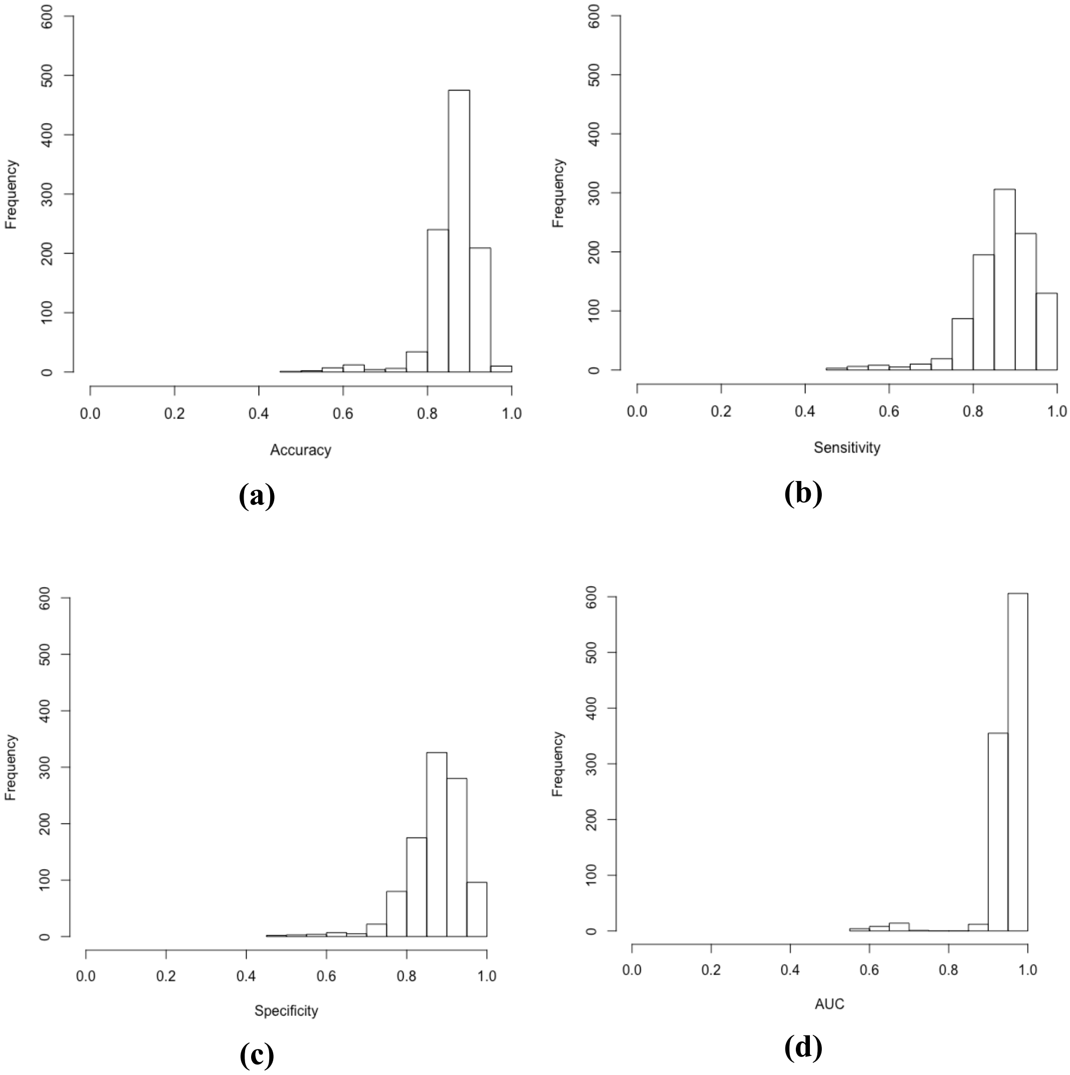


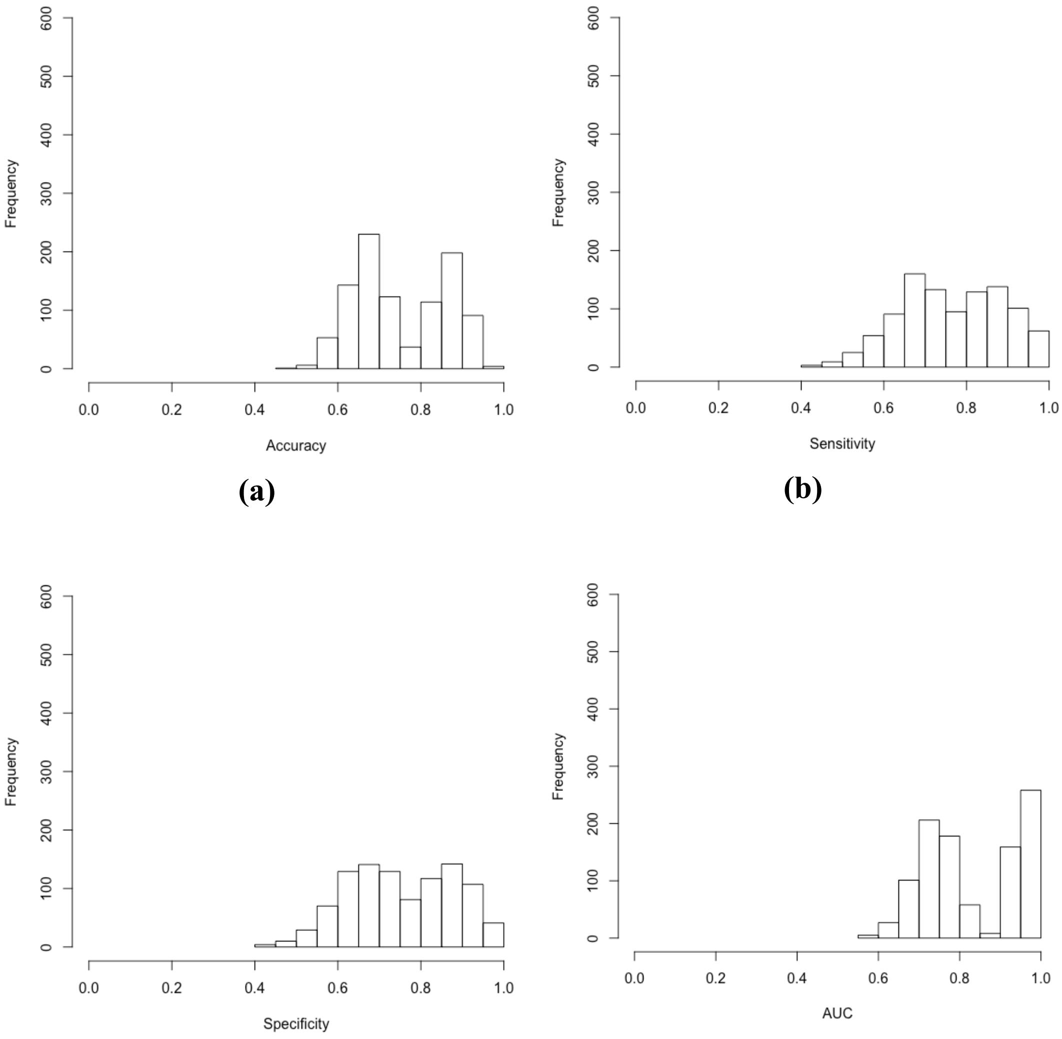
Figure S4.13: Distributions of AUCs for sample size *n*=76 and ρ =-0.8 across methods (1000 iterations)

b) Histograms of performance measures for n=300 based on **1000 iterations**



(a) **(b)** **(c)** **(d)**

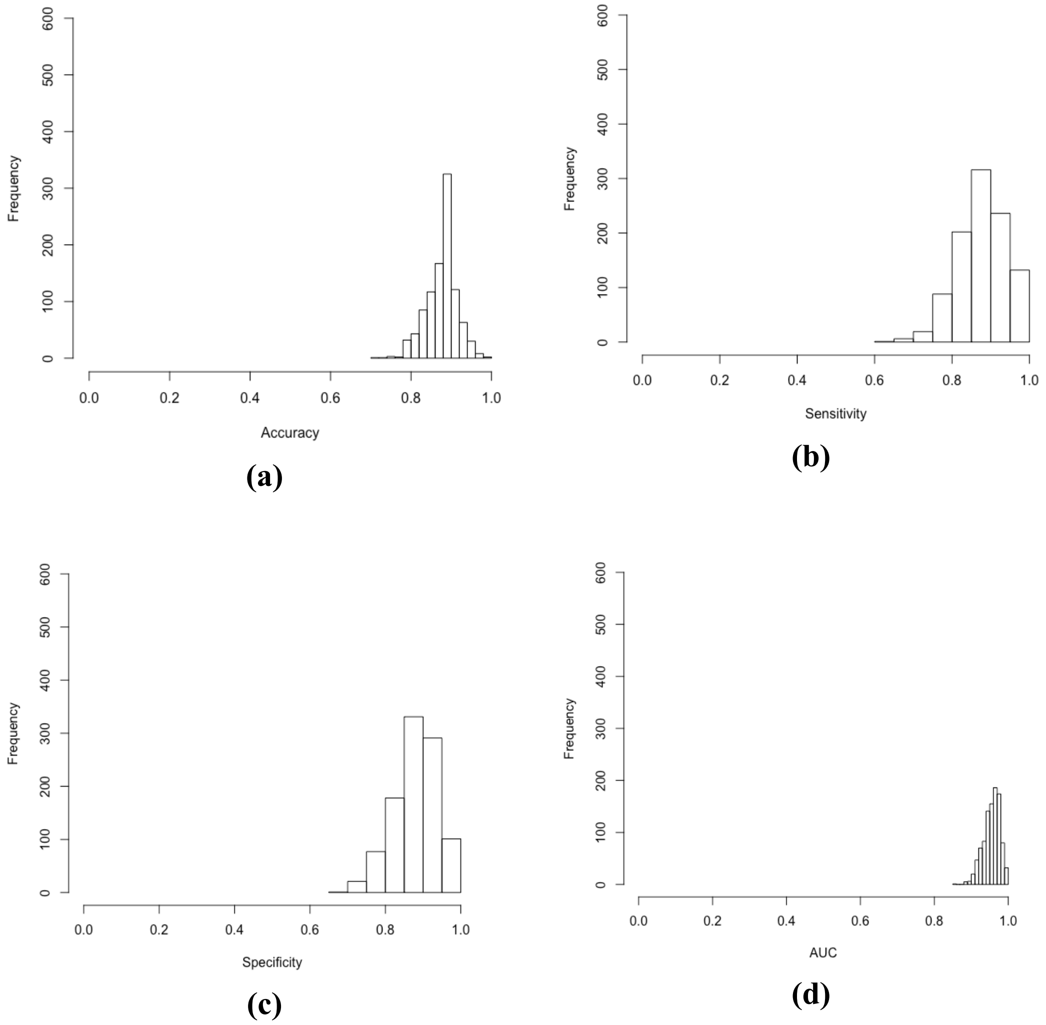
Figure S4. 14: Distribution of the estimates of **(a)** Accuracy, **(b)** Sensitivity, **(c)** Specificity and **(d)** AUC for sample size *n*=300 and *ρ*= -0.8, for the T method.



(a) **(b)**

(c) **(d)**

Figure S4. 15: Distribution of the estimates of **(a)** Accuracy, **(b)** Sensitivity, **(c)** Specificity and **(d)** AUC for sample size *n*=300 and *ρ*= -0.8, for the corT method.



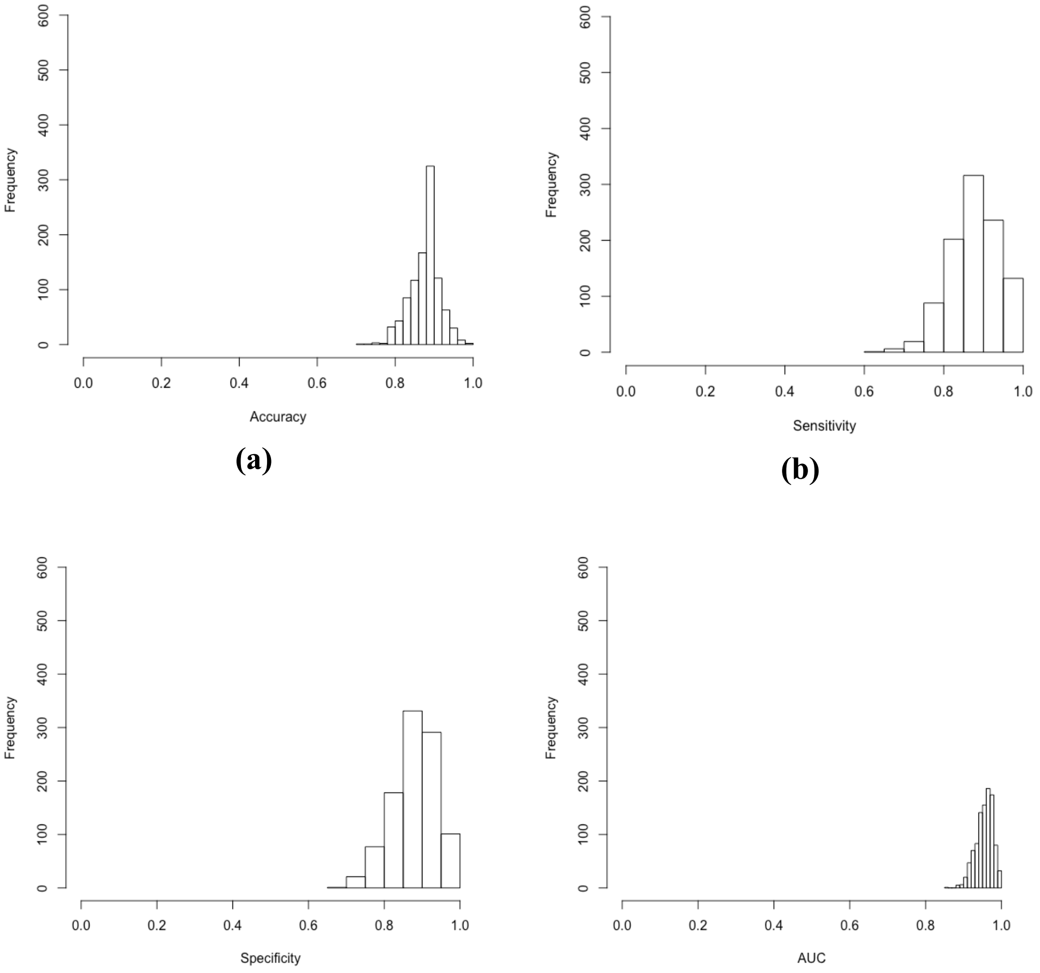
(a)

(b)

(c)

(d)

Figure S4. 16: Distribution of the estimates of **(a)** Accuracy, **(b)** Sensitivity, **(c)** Specificity and **(d)** AUC for sample size *n*=300 and *ρ*= -0.8, for the adjcorT method.



(a)

(b)

(c)

(d)

Figure S4. 17: Distribution of the estimates of **(a)** Accuracy, **(b)** Sensitivity, **(c)** Specificity and **(d)** AUC for sample size *n*=300 and *ρ*= -0.8, for the Lasso method.

Appendix IV: Classification Accuracy, Sensitivity, Specificity and AUC of each method for all sample sizes and at different correlations

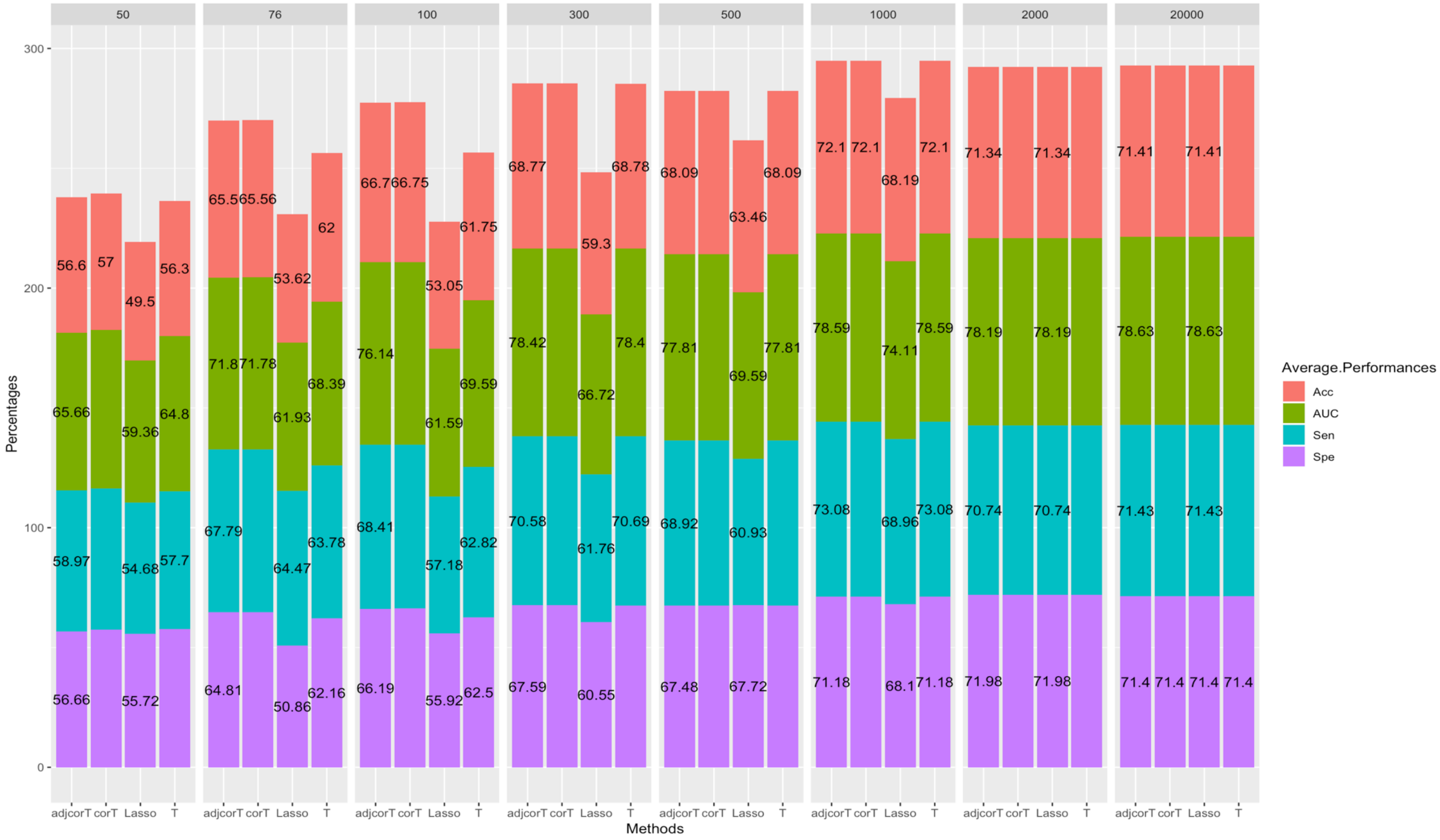


Figure S4. 18: Classification Accuracy, Sensitivity, Specificity and AUC of each method for all sample sizes and ρ=0.8

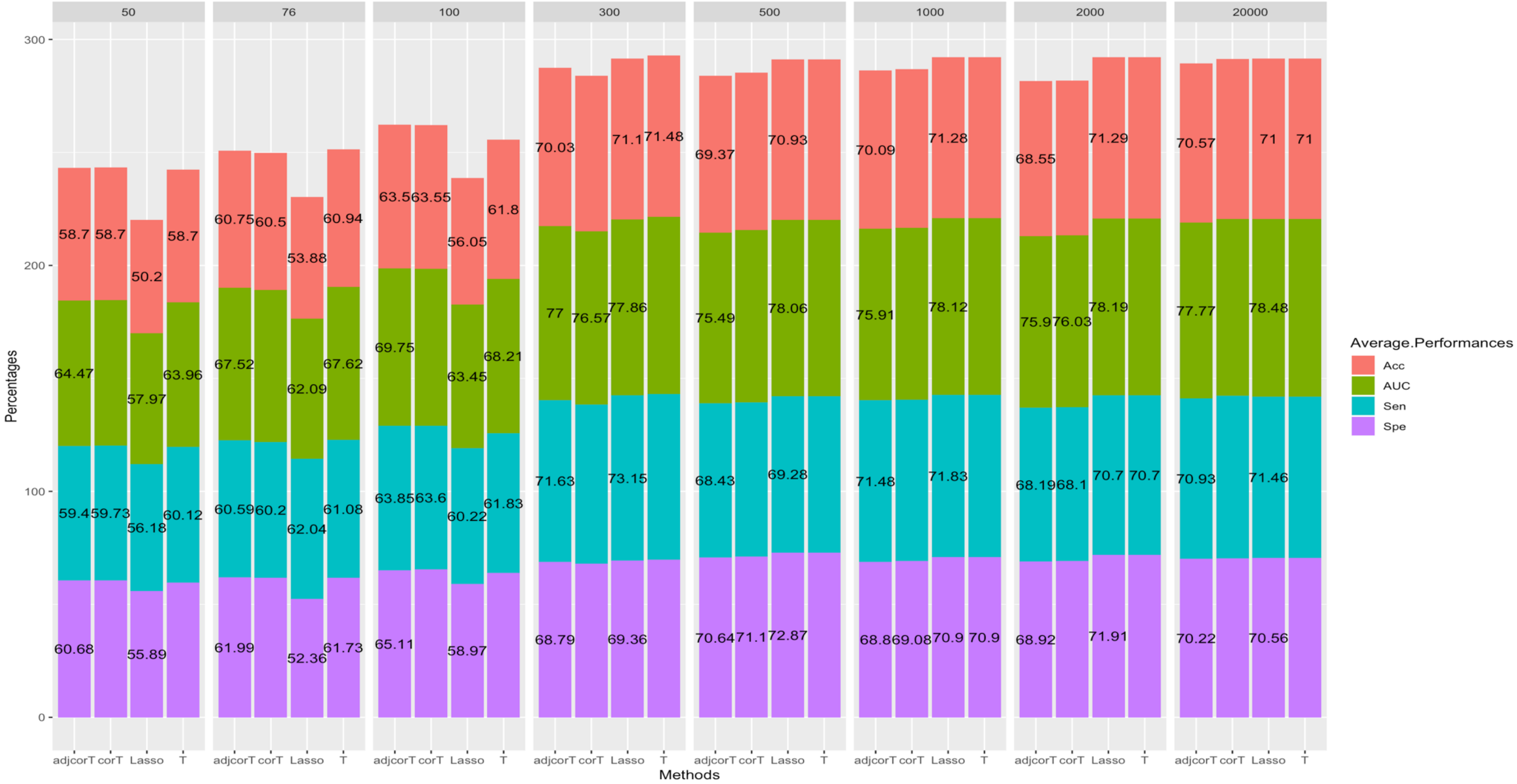


Figure S4. 19: Classification Accuracy, Sensitivity, Specificity and AUC of each method for all sample sizes and ρ=0

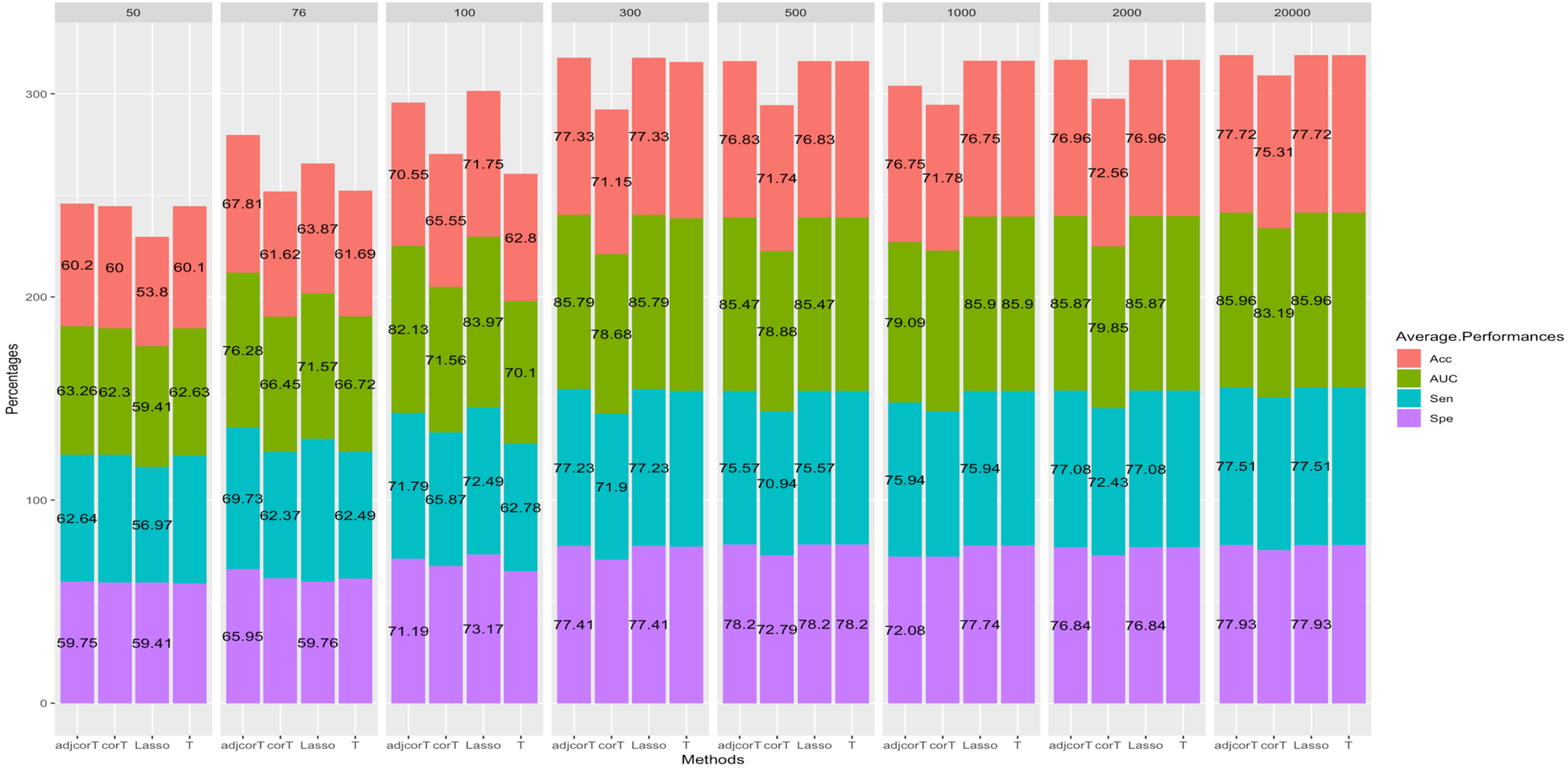


Figure S4. 20: Classification Accuracy, Sensitivity, Specificity and AUC of each method for all sample sizes and ρ=-0.5

Appendix V: Correlation tables of the top 10 VOCs selected by each variable selection method for each dataset

i) Colorectal cancer and non-cancer discrimination

Table S5. 1: Correlation of the top 10 VOCs selected by T method

	X27.19_Pentane..2.3.4.trimethyl.	X33.44_Hexanoic.acid..2.methylbutyl.ester	X25.32_Propanoic.acid..pentyl.ester	X17.93_Propanoic.acid..propyl.ester	X22.19_2.Heptanol	X29.18_3.Carene	X22.01_Acetic.acid..pentyl.ester	X31.48_Cyclohexanecarboxylic.acid	X29.47_Heptanoic.acid	X28.53_Benzeneacetaldehyde
X27.19_Pentane..2.3.4.trimethyl.	1	0.16	0.49	0.34	0.00	-0.03	0.33	0.06	0.16	-0.26
X33.44_Hexanoic.acid..2.methylbutyl.ester	0.16	1	0.04	-0.06	0.29	-0.11	0.12	-0.03	0.36	-0.18
X25.32_Propanoic.acid..pentyl.ester	0.49	0.04	1	0.56	-0.06	-0.07	0.88	0.13	-0.02	-0.17
X17.93_Propanoic.acid..propyl.ester	0.34	-0.06	0.56	1	-0.06	-0.06	0.52	0.17	-0.10	-0.18
X22.19_2.Heptanol	0.00	0.29	-0.06	-0.06	1	0.00	-0.06	-0.04	0.42	-0.12
X29.18_3.Carene	-0.03	-0.11	-0.07	-0.06	0.00	1	-0.04	0.17	0.18	-0.04
X22.01_Acetic.acid..pentyl.ester	0.33	0.12	0.88	0.52	-0.06	-0.04	1	0.11	-0.02	-0.21
X31.48_Cyclohexanecarboxylic.acid	0.06	-0.03	0.13	0.17	-0.04	0.17	0.11	1	0.00	-0.02
X29.47_Heptanoic.acid	0.16	0.36	-0.02	-0.10	0.42	0.18	-0.02	0.00	1	-0.22
X28.53_Benzeneacetaldehyde	-0.26	-0.18	-0.17	-0.18	-0.12	-0.04	-0.21	-0.02	-0.22	1

Table S5. 2: Correlation of the top 10 VOCs selected by corT and adjcorT

	X27.19_Pentane..2.3.4.trimethyl.	X33.44_Hexanoic.acid..2.methylb	X25.32_Propanoic.acid..pentyl.ester	X23.49_Butanoic.acid..2.methylpropyl.ester	X22.01_Acetic.acid..pentyl.ester	X17.93_Propanoic.acid..p	X29.18_3.Carene	X31.48_Cyclohexanecarboxylic.acid	X22.19_2.Heptanol	X27.52_Butanoic.acid..4.pentenyl.ester
X27.19_Pentane..2.3.4.trimethyl.	1	0.12	0.55	0.72	0.42	0.33	-0.01	0.12	-0.02	0.64
X33.44_Hexanoic.acid..2.methylbutyl.ester	0.12	1	0.01	0.21	0.09	-0.06	-0.11	-0.04	0.29	0.13
X25.32_Propanoic.acid..pentyl.ester	0.55	0.01	1	0.59	0.89	0.56	-0.04	0.15	-0.07	0.53
X23.49_Butanoic.acid..2.methylpropyl.ester	0.72	0.21	0.59	1	0.50	0.31	0.04	0.23	0.01	0.50
X22.01_Acetic.acid..pentyl.ester	0.42	0.09	0.89	0.50	1	0.53	-0.02	0.16	-0.07	0.50
X17.93_Propanoic.acid..propyl.ester	0.33	-0.06	0.56	0.31	0.53	1	-0.05	0.04	-0.06	0.18
X29.18_3.Carene	-0.01	-0.11	-0.04	0.04	-0.02	-0.05	1	0.07	0.00	0.02
X31.48_Cyclohexanecarboxylic.acid	0.12	-0.04	0.15	0.23	0.16	0.04	0.07	1	-0.03	0.25
X22.19_2.Heptanol	-0.02	0.29	-0.07	0.01	-0.07	-0.06	0.00	-0.03	1	-0.03
X27.52_Butanoic.acid..4.pentenyl.ester	0.64	0.13	0.53	0.50	0.50	0.18	0.02	0.25	-0.03	1

Table S5. 3: Correlation of the top 10 VOCs selected by Lasso

	X33.44_Hexanoi	X27.19_ Pentane..2.3.4.	X32.25_dl.M	X28.53_Benzenea	X22.19_2.Heptanol	X31.48_ Cyclohexanecarbo	X24.00_ Propanoic.acid..p	X29.18_3.Ca	Age	X17.93_Propanoi
X33.44_Hexanoic.acid..2.meth	1	0.11	-0.04	-0.17	0.29	-0.04	0.18	-0.11	0.04	-0.06
X27.19_Pentane..2.3.4.	0.11	1	-0.06	-0.25	-0.02	0.12	0.10	-0.01	0.01	0.33
X32.25_dl.Menthol	-0.04	-0.06	1	0.05	-0.03	-0.02	0.03	0.06	-0.11	-0.07
X28.53_Benzeneacetaldehyde	-0.17	-0.25	0.05	1	-0.11	-0.05	-0.00	-0.03	-0.06	-0.19
X22.19_2.Heptanol	0.29	-0.02	-0.03	-0.11	1	-0.03	0.00	0.00	0.11	-0.06
X31.48_Cyclohexanecarboxylic.acid	-0.04	0.12	-0.02	-0.05	-0.03	1	0.37	0.07	0.12	0.04
X24.00_Propanoic.acid..pentyl.ester	0.18	0.10	0.03	-0.00	0.00	0.37	1	-0.02	0.15	-0.01
X29.18_3.Carene	-0.11	-0.01	0.06	-0.03	0.00	0.07	-0.02	1	-0.02	-0.05
Age	0.04	0.01	-0.11	-0.06	0.11	0.12	0.15	-0.02	1	-0.05
X17.93_Propanoic.acid..propyl.ester	-0.06	0.33	-0.07	-0.19	-0.06	0.04	-0.01	-0.05	-0.05	1

ii) Healthy control and Adenoma discrimination

Table S5. 4: Correlation of the top 10 VOCs selected by T method

	X27.19_Pentane..2.3.4.trimethyl.	X28.53_Benzeneacetaldehyde	X33.44_Hexanoic.acid..2.methylbutyl.ester	X25.32_Propanoic.acid..pentyl.ester	X22.01_Acetic.acid..pentyl.ester	X23.39_Methional	X22.19_2.Heptanol	X25.22_Dimethyl.trisulfide	X24.97_Pentanoic.acid..propyl.ester	X27.52_Butanoic.acid..4.pentenyl.ester
X27.19_Pentane..2.3.4.trimethyl.	1	-0.32	0.12	0.56	0.44	-0.26	0.01	0.02	0.47	0.57
X28.53_Benzeneacetaldehyde	-0.32	1	-0.22	-0.24	-0.27	0.72	-0.10	-0.22	-0.26	-0.36
X33.44_Hexanoic.acid..2.methylbutyl.ester	0.12	-0.22	1	0.07	0.19	-0.13	0.45	0.10	0.07	0.22
X25.32_Propanoic.acid..pentyl.ester	0.56	-0.24	0.07	1	0.90	-0.27	-0.07	-0.13	0.95	0.56
X22.01_Acetic.acid..pentyl.ester	0.44	-0.27	0.19	0.90	1	-0.28	-0.07	-0.12	0.90	0.58
X23.39_Methional	-0.26	0.72	-0.13	-0.27	-0.28	1	-0.12	-0.19	-0.26	-0.38
X22.19_2.Heptanol	0.01	-0.10	0.45	-0.07	-0.07	-0.12	1	0.04	-0.07	-0.03
X25.22_Dimethyl.trisulfide	0.02	-0.22	0.10	-0.13	-0.12	-0.19	0.04	1	-0.12	-0.02
X24.97_Pentanoic.acid..propyl.ester	0.47	-0.26	0.07	0.95	0.90	-0.26	-0.07	-0.12	1	0.54
X27.52_Butanoic.acid..4.pentenyl.ester	0.57	-0.36	0.22	0.56	0.58	-0.38	-0.03	-0.02	0.54	1

Table S5. 5: Correlation of the top 10 VOCs selected by corT and adjcorT

	X28.53_Benzeneacetaldehyde	X27.19_Pentane..2.3.4.trimethyl.	X33.44_Hexanoic.acid..2.methylbutyl.ester	X27.52_Butanoic.acid..4.pentenyl.ester	X23.49_Butanoic.acid..2.methylpropyl.ester	X23.39_Methional	X25.32_Propanoic.acid..pentyl.ester	X22.01_Acetic.acid..pentyl.ester	X12.47_Butanal..3.methyl.	X22.19_2.Heptanol
X28.53_Benzeneacetaldehyde	1	-0.33	-0.24	-0.35	-0.36	0.69	-0.24	-0.27	0.56	-0.10
X27.19_Pentane..2.3.4.trimethyl.	-0.33	1	0.12	0.70	0.75	-0.24	0.57	0.42	-0.22	-0.01
X33.44_Hexanoic.acid..2.methylbutyl.ester	-0.24	0.12	1	0.12	0.20	-0.10	0.06	0.08	-0.11	0.30
X27.52_Butanoic.acid..4.pentenyl.ester	-0.35	0.70	0.12	1	0.49	-0.34	0.54	0.51	-0.23	-0.03
X23.49_Butanoic.acid..2.methylpropyl.ester	-0.36	0.75	0.20	0.49	1	-0.26	0.59	0.50	-0.22	0.00
X23.39_Methional	0.69	-0.24	-0.10	-0.34	-0.26	1	-0.26	-0.27	0.67	-0.12
X25.32_Propanoic.acid..pentyl.ester	-0.24	0.57	0.06	0.54	0.59	-0.26	1	0.89	-0.14	-0.07
X22.01_Acetic.acid..pentyl.ester	-0.27	0.42	0.08	0.51	0.50	-0.27	0.89	1	-0.19	-0.06
X12.47_Butanal..3.methyl.	0.56	-0.22	-0.11	-0.23	-0.22	0.67	-0.14	-0.19	1	-0.08
X22.19_2.Heptanol	-0.10	-0.01	0.30	-0.03	0.00	-0.12	-0.07	-0.06	-0.08	1

Table S5. 6: Correlation of the top 10 VOCs selected by Lasso

	X28.53_Benzeneacetaldehyde	X33.44_Hexanoic.acid..2.methylbutyl.ester	X27.19_Pentane..2.3.4.trimethyl.	X29.18_3.Carene	X22.19_2.Heptanol	X23.27_S.Methyl.3.methylbutanethioate	X31.48_Cyclohexanecarboxylic.acid	X25.22_Dimethyl.trisulfide	X33.63_Phenol..4.ethyl.	X24.26_2.Heptanone..6.methyl.
X28.53_Benzeneacetaldehyde	1	-0.24	-0.33	0.03	-0.10	-0.06	-0.04	-0.21	-0.16	0.03
X33.44_Hexanoic.acid..2.methylbutyl.ester	-0.24	1	0.12	-0.11	0.30	0.12	-0.05	0.14	0.20	-0.04
X27.19_Pentane..2.3.4.trimethyl.	-0.33	0.12	1	0.01	-0.01	0.03	0.14	0.07	0.9	-0.13
X29.18_3.Carene	0.03	-0.11	0.01	1	0.00	-0.05	0.06	0.01	0.05	-0.08
X22.19_2.Heptanol	-0.10	0.30	-0.01	0.00	1	-0.05	-0.03	0.04	0.06	0.09
X23.27_S.Methyl.3.methylbutanethioate	-0.06	0.12	0.03	-0.05	-0.05	1	-0.06	0.03	-0.01	0.05
X31.48_Cyclohexanecarboxylic.acid	-0.04	-0.05	0.14	0.06	-0.03	-0.06	1	-0.08	-0.05	-0.05
X25.22_Dimethyl.trisulfide	-0.21	0.14	0.07	0.01	0.04	0.03	-0.08	1	0.10	-0.06
X33.63_Phenol..4.ethyl.	-0.16	0.20	0.09	0.05	0.06	-0.01	-0.05	0.10	1	-0.11
X24.26_2.Heptanone..6.methyl.	0.03	-0.04	-0.13	-0.08	0.09	0.05	-0.05	-0.06	-0.11	1

iii) Bacterial and non-bacterial sepsis discrimination

Table S5. 7: Correlation of the top 10 VOCs selected by T method

	unknown _7	unknown_1 29	phenylalanine _8	unknown_ 34	phenylalanine _6	unknown_ 10	creatine40_ 33	acetoacetate_ 111	mobile.lipids_ 132	mobile.lipids_ 18
unknown_7	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-0.09	-0.07
unknown_129	1.00	1	1.00	1.00	1.00	1.00	1.00	1.00	-0.08	-0.06
phenylalanine_ 8	1.00	1.00	1	1.00	1.00	1.00	1.00	1.00	-0.06	-0.05
unknown_34	1.00	1.00	1.00	1	1.00	1.00	1.00	1.00	-0.07	-0.06
phenylalanine_ 6	1.00	1.00	1.00	1.00	1	1.00	1.00	1.00	-0.07	-0.05
unknown_10	1.00	1.00	1.00	1.00	1.00	1	1.00	1.00	-0.03	-0.02
creatine40_33	1.00	1.00	1.00	1.00	1.00	1.00	1	1.00	-0.06	-0.05
acetoacetate_1 11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	-0.04	-0.03
mobile.lipids_ 132	-0.09	-0.08	-0.06	-0.07	-0.07	-0.03	-0.06	-0.04	1	0.94
mobile.lipids_ 18	-0.07	-0.06	-0.05	-0.06	-0.05	-0.02	-0.05	-0.03	0.94	1

Table S5. 8: Correlation of the top 10 VOCs selected by corT and adjcorT

	unknown_7	unknown_129	phenylalanine_6	phenylalanine_8	unknown_34	acetoacetate_111	creatine40_33	unknown_10	glucose_35	mobile.lipids_132
unknown_7	1	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	-0.06
unknown_129	0.99	1	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.06
phenylalanine_6	0.99	0.99	1	0.99	0.99	0.99	0.99	0.99	0.99	0.01
phenylalanine_8	0.99	0.99	0.99	1	0.99	0.99	0.99	0.99	0.99	0.02
unknown_34	0.99	0.99	0.99	0.99	1	0.99	1.00	0.99	0.99	0.01
acetoacetate_111	0.99	0.99	0.99	0.99	0.99	1	0.99	0.99	0.99	0.04
creatine40_33	0.99	0.99	0.99	0.99	1.00	0.99	1	0.99	1.00	0.02
unknown_10	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1	0.99	0.05
glucose_35	0.99	0.99	0.99	0.99	0.99	0.99	1.00	0.99	1	0.05
mobile.lipids_132	-0.06	0.06	0.01	0.02	0.01	0.04	0.02	0.05	0.05	1

Table S5. 9: Correlation of the top 10 VOCs selected by Lasso

	mobile.lipids_132	unknown_94	glucose_45	phenylalanine_4	unknown_10	glucose_65	unknown_7	desaminotyrosine_16	glucose_58	glucose_62
mobile.lipids_132	1	0.21	0.26	0.73	0.05	0.76	0.00	0.51	0.62	0.65
unknown_94	0.21	1	0.83	0.19	0.91	0.52	0.89	0.69	0.76	0.72
glucose_45	0.26	0.83	1	0.32	0.80	0.59	0.78	0.79	0.74	0.68
phenylalanine_4	0.73	0.19	0.32	1	0.16	0.69	0.13	0.61	0.58	0.57
unknown_10	0.05	0.91	0.80	0.16	1	0.43	0.99	0.65	0.72	0.68
glucose_65	0.76	0.52	0.59	0.69	0.43	1	0.38	0.71	0.83	0.90
unknown_7	0.00	0.89	0.78	0.13	0.99	0.38	1	0.63	0.69	0.64
desaminotyrosine_16	0.51	0.69	0.79	0.61	0.65	0.71	0.63	1	0.83	0.74
glucose_58	0.62	0.76	0.74	0.58	0.72	0.83	0.69	0.83	1	0.95
glucose_62	0.65	0.72	0.68	0.57	0.68	0.90	0.64	0.74	0.95	1

iv) Healthy control and Kidney Disease discrimination

Table S5. 10: Correlation of the top 10 VOCs selected by T method

	Creatinine	Hydroxyphenylpyruvic.acid	Methylmalonate	D.Glucoronic.acid	myoinositol	X1.Methyladenosine	Choline	X2.Aminoisobutyric.acid	Fumaric. Acid	Xanthosine
Creatinine	1	0.89	0.85	0.72	0.88	0.74	0.78	0.74	0.78	0.68
Hydroxyphenylpyruvic.acid	0.89	1	0.88	0.72	0.98	0.60	0.74	0.71	0.74	0.56
Methylmalonate	0.85	0.88	1	0.74	0.87	0.64	0.79	0.76	0.80	0.64
D.Glucoronic.acid	0.72	0.72	0.74	1	0.69	0.55	0.66	0.63	0.68	0.58
myoinositol	0.88	0.98	0.87	0.69	1	0.59	0.73	0.70	0.75	0.56
X1.Methyladenosine	0.74	0.60	0.64	0.55	0.59	1	0.66	0.66	0.55	0.63
Choline	0.78	0.74	0.79	0.66	0.73	0.66	1	0.97	0.62	0.61
X2.Aminoisobutyric.acid	0.74	0.71	0.76	0.63	0.70	0.66	0.97	1	0.58	0.59
Fumaric.Acid	0.78	0.74	0.80	0.68	0.75	0.55	0.62	0.58	1	0.54
Xanthosine	0.68	0.56	0.64	0.58	0.56	0.63	0.61	0.59	0.54	1

Table S5. 11: Correlation of the top 10 VOCs selected by corT and adjcorT

	Creatinine	Hydroxyphenylpyruvic.acid	myoinositol	Methylmalonate	X2.Hydroxyglutarate	Choline	X2.Aminoisobutyric.acid	Oxaloacetate	D.Glucoronic.acid	X1.Methyladenosine
Creatinine	1	0.87	0.87	0.84	0.80	0.76	0.72	0.80	0.72	0.74
Hydroxyphenylpyruvic.acid	0.87	1	0.98	0.87	0.75	0.70	0.67	0.82	0.71	0.59
myoinositol	0.87	0.98	1	0.86	0.74	0.70	0.67	0.83	0.69	0.57
Methylmalonate	0.84	0.87	0.86	1	0.73	0.74	0.72	0.80	0.74	0.63
X2.Hydroxyglutarate	0.80	0.75	0.74	0.73	1	0.55	0.53	0.71	0.60	0.70
Choline	0.76	0.70	0.70	0.74	0.55	1	0.97	0.65	0.63	0.66
X2.Aminoisobutyric.acid	0.72	0.67	0.67	0.72	0.53	0.97	1	0.64	0.61	0.66
Oxaloacetate	0.80	0.82	0.83	0.80	0.71	0.65	0.64	1	0.67	0.52
D.Glucoronic.acid	0.72	0.72	0.69	0.74	0.60	0.63	0.61	0.68	1	0.57
X1.Methyladenosine	0.74	0.59	0.57	0.63	0.70	0.66	0.66	0.52	0.57	1

Table S5. 12: Correlation of the top 10 VOCs selected by Lasso

	Creatinine	Urate	X1.Methyladenosine	Xanthosine	D.Glucoronic.acid	Fumaric.Acid	Guanidinoacetate	X2.Aminoisobutyric.acid	X1.Methylhistidine	myoinositol
Creatinine	1	0.58	0.74	0.67	0.72	0.71	-0.26	0.72	0.47	0.87
Urate	0.58	1	0.45	0.38	0.46	0.38	-0.11	0.38	0.27	0.42
X1.Methyladenosine	0.74	0.44	1	0.63	0.57	0.53	-0.25	0.66	0.38	0.57
Xanthosine	0.67	0.38	0.63	1	0.55	0.55	-0.25	0.54	0.37	0.51
D.Glucoronic.acid	0.72	0.46	0.57	0.55	1	0.68	-0.18	0.61	0.45	0.69
Fumaric.Acid	0.71	0.38	0.53	0.55	0.68	1	-0.17	0.54	0.47	0.66
Guanidinoacetate	-0.26	-0.11	-0.25	-0.25	-0.18	-0.17	1	-0.21	-0.20	-0.27
X2.Aminoisobutyric.acid	0.72	0.38	0.66	0.54	0.61	0.54	-0.21	1	0.29	0.67
X1.Methylhistidine	0.46	0.27	0.38	0.37	0.45	0.47	-0.20	0.29	1	0.49
myoinositol	0.87	0.42	0.57	0.51	0.69	0.66	-0.27	0.67	0.49	1