# APPLIED FUNCTIONAL DATA CLASSIFICATION

by

Jonathan Babyn

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
March 2018

*Dedicated to my parents and everybody else who made this possible.*
*Thanks.*

## Table of Contents

**List of Tables** . . . . . **v**

**List of Figures** . . . . . **ix**

**Abstract** . . . . . **xi**

**List of Abbreviations and Symbols Used** . . . . . **xii**

**Acknowledgements** . . . . . **xiv**

**Chapter 1 Introduction** . . . . . **1**

**Chapter 2 Review of Methodologies** . . . . . **4**

2.1 Decision Trees . . . . . 4
2.1.1 Classification And Regression Trees . . . . . 4
2.1.2 Random Forests . . . . . 5

2.2 Functional Data Analysis . . . . . 6
2.2.1 Functional Principal Component Analysis . . . . . 7
2.2.2 Functional Linear Discriminant Analysis . . . . . 10
2.2.3 Two-Sample Tests for Functional Data . . . . . 12

2.3 $DD^G$ Plot Classifiers . . . . . 15

**Chapter 3 Data Collection Methods** . . . . . **18**

3.1 Prototype Data . . . . . 20

3.2 Advanced Data . . . . . 23

**Chapter 4 Analysis of Prototype Data** . . . . . **26**

4.1 Missing Data . . . . . 27

4.2 Two Sample Testing for $\tau_0$ Data . . . . . 33

4.3 Limitations . . . . . 34

4.4 CART on Prototype Data . . . . . 34

4.5 Random Forests . . . . . 38

4.6 FLDA on Prototype Data . . . . . 40

4.7 FPCA and Related Techniques on Prototype Data . . . . . 46

4.8 $DD^G$ Plots on Prototype Data . . . . . 48

**Chapter 5 Analysis of the Advanced Data . . . . . 57**

5.1 Missing Data . . . . . 59

5.2 Decision Trees on Advanced Data . . . . . 60
5.2.1 Random Forests on the Full Set of Advanced Data . . . . . 63
5.2.2 CART and Random Forests on Smaller Bands of Wavelengths 64

5.3 FLDA on Advanced Data . . . . . 69

5.4 FPCA and Related Techniques on Advanced Data . . . . . 73

5.5 $DD^G$ Plots on Advanced Data . . . . . 75

**Chapter 6 Conclusion . . . . . 79**

6.1 Future Directions . . . . . 80

**Bibliography . . . . . 82**

# List of Tables


3.1 The organization of the experiment in Picomole's pilot study. Picomole changed plans part way through which is why some sets of conditions only have one or two measurements. . . . . 21

3.2 The order in which subjects' samples were processed from the 200°C desorb temperature measurements. Note the clustering of Lung Cancer subjects between March 9th and April 12th. This pattern is similar for the other desorb temperatures. . . . . 22

3.3 The organization of the experiment in Picomole's real data. Picomole changed the setup of the experiment from that of the prototype data. The majority of lung cancer subjects available at this time were also all post-treatment which resulted in only post treatment subjects being used. These changes make conclusions drawn from the prototype data not generally relevant to the advanced data. . . . . 23

3.4 The best optimal age and gendered matched set of subjects from the advanced data. For this set of subjects all 8 measurements over the various desorption temperatures exist . . . . . 24

4.1 The order in which subjects' samples were processed from the 200°C desorb temperature measurements . . . . . 32

4.2 P-values of FAD test on prototype $\tau_0$ data . . . . . 34

4.3 The accuracy, sensitivity and specificity for each of the second set of pruned CART trees grown. Note that H refers to healthy subjects and LC to lung cancer subjects. . . . . 38

4.4 The confusion matrix for second desorption at 80°C for the random forest's performance on training data. Classification performance is very poor, with rates worse than random guessing and is unable to correctly deal with COPD subjects. . . . . 39

4.5 Confusion Matrix for the second desorption at 80°C for the random forest grown on the training data and tested on the testing set. Performance is as poor as on the training data. . . . . 39

4.6 The confusion matrix for the second desorption at 80°C with the COPD subjects removed for the random forest applied to the training data. Removing the COPD subjects did not have a positive impact on performance. . . . . 40

4.7 The confusion matrix for the second desorption at 80°C with the COPD subjects removed for the random forest applied to the testing data. It initially appears as though the random forest is capable of correctly classifying lung cancer subjects, although performance on the training data and the small sample size makes this unlikely. . . . . 40

4.8 The accuracy (and corresponding 95% confidence interval), sensitivity and specificity for the FLDA models constructed using 10 fold cross validation. Removing COPD subjects improves performance. Once again H refers to Healthy and LC to lung cancer. . . . . 42

4.9 The accuracy, sensitivity and specificity for each of the desorptions for FPCA clustering with all three classes included. Overall performance is not very good. . . . . 48

4.10 The accuracy, sensitivity and specificity for each of the desorptions for FPCA clustering with the COPD subjects removed. The best performing class assignments were used. Performance is not any better than random guessing. . . . . 48

4.11 The accuracy, sensitivity and specificity for the DD plot using LDA as the classifier and FM depth with the COPD subjects binned with the lung cancer subjects. . . . . 50

4.12 The accuracy, sensitivity and specificity for the DD plot using QDA as the classifier and FM depth with the COPD subjects binned with the lung cancer subjects. . . . . 50

4.13 The accuracy, sensitivity and specificity for the DD plot using K nearest neighbors as the classifier and FM depth with the COPD subjects binned with the lung cancer subjects. . . . . 50

4.14 The accuracy, sensitivity and specificity for the DD plot using the non-parametric kernel method as the classifier and FM depth with the COPD subjects binned with lung cancer subjects. . . 51

4.15 The accuracy,sensitivity and specificity for the $DD^G$ plot using the LDA classifier and FM depth on all three classes. . . . . 51

4.16 The accuracy,sensitivity and specificity for the $DD^G$ plot using the QDA classifier and FM depth on all three classes. . . . . 51

4.17 The accuracy,sensitivity and specificity for the $DD^G$ plot using the K nearest neighbors classifier with FM depth on all three classes. . . . . 51

4.18 The accuracy, sensitivity and specificity for the $DD^G$ plot using the non-parametric kernel method classifier with FM depth on all three classes. . . . . 51

5.1 The accuracy, sensitivity and specificity for each of the pruned trees. Performance is improved over the prototype trees. . . . . 60

5.2 The accuracy, sensitivity and specificity for each of random forests grown. Performance is similarly poor to that of the random forests grown on the prototype data. . . . . 63

5.3 The accuracy, sensitivity and specificity for the CART trees grown on the band of lowest wavelengths. . . . . 66

5.4 The accuracy, sensitivity and specificity for the random forests grown on the band of lowest wavelengths. . . . . 66

5.5 The accuracy, sensitivity and specificity for the CART trees grown on the band of highest wavelengths. . . . . 67

5.6 The accuracy, sensitivity and specificity for the random forests grown only on the band of highest wavelengths. . . . . 67

5.7 The accuracy, sensitivity and specificity for the CART trees grown on the seven highest and seven lowest wavelengths. . . 68

5.8 The accuracy, sensitivity and specificity for the random forests grown only on the band of the seven highest and seven lowest wavelengths. . . . . 68

5.9 The accuracy, sensitivity and specificity for the two FLDA trials conducted along with the no information rate. Performance on the three tube set of conditions is similar to that of the FLDA on the prototype data. . . . . 70

5.10 The accuracy, sensitivity and specificity for the 2 group clustering via. FPCA on the advanced data. Performance of FPCA clustering is equally poor as with the prototype data. . . . . 75

5.11 The accuracy, sensitivity and specificity for the $DD^G$ plot using the LDA classifier. . . . . 76

5.12 The accuracy, sensitivity and specificity for the $DD^G$ plot on the advanced data with the QDA classifier. . . . . 76

5.13 The accuracy, sensitivity and specificity for the $DD^G$ plot on the advanced data with the DD3 classifier. . . . . 76

5.14 The accuracy, sensitivity and specificity for the $DD^G$ plot on the advanced data with the NP classifier. . . . . 78

## List of Figures


3.1 Wavelength spacing of Picomole's CRDS analyzer. . . . . 20

4.1 Plots $\tau$, $\tau_0$ and K absorption from the 200°C prototype data desorption. . . . . 26

4.2 The percentage of missing observations per wavelength on the prototype data. . . . . 28

4.3 Aggregation Plots of the K absorption data for all of the subjects on the 200°C desorption . . . . . 29

4.4 Aggregation plots for the prototype K absorption data grouped by health status . . . . . 30

4.5 Aggregation plot for the prototype $\tau_0$ measurements desorbed at 200°C . . . . . 31

4.6 The first CART trees constructed with the age variable included 36

4.7 Pruned and cross validated trees grown on the K absorption data 37

4.8 Plots showing the separation between classes on all the data generated by FLDA . . . . . 43

4.9 The FLDA separation between classes on the prototype data with 10-fold cross validation . . . . . 44

4.10 The FLDA separation after the COPD subjects have been removed . . . . . 45

4.11 Diagnostic plots for the FPCA performed on desorption performed at 200°C . . . . . 47

4.12 The $DD^G$ plots constructed on the 200°C desorb. . . . . 52

4.13 The $DD^G$ plot for the 200°C desorb on the prototype data using the LDA classifier with all three classes . . . . . 53

4.14 The $DD^G$ plot for the 200°C desorb on the prototype data using the QDA classifier with all three classes . . . . . 54

4.15 The $DD^G$ plot for the 200°C desorb on the prototype data using the kNN classifier with all three classes . . . . . 55

4.16 The $DD^G$ plot for the 200°C desorb on the prototype data using the non-parametric kernel method classifier with all three classes 56

5.1 Plots of $\tau$, $\tau_0$ and K absorption on the 200°C Tube 1 desorption on the advanced data. . . . . 58

5.2 Aggregation plots for the advanced data grouped by health status. 59

5.3 Pruned and cross validated trees for the K absorption data for the advanced data . . . . . 61

5.4 Pruned and cross validated trees for the K absorption data for the advanced data on the four tube set of desorptions . . . . . 62

5.5 Pruned and cross validated trees for the K absorption data with a reduced number of lines . . . . . 65

5.6 The $\hat{\alpha}$ estimates for the FLDA performed on the four tube desorption. One thing that stands out is the subject whose value of $\hat{\alpha}$ is several times larger than that of all the others. The FLDA here correctly classifies only 50% of the subejcts. . . . . 71

5.7 The $\hat{\alpha}$ estimates for the FLDA conducted on the three tube set of desorptions. Performance is better than the FLDA on the four tube set, the separation between the two groups is better and it didn't run into the same outlier issue with one subject having an extreme value of $\hat{\alpha}$. . . . . 72

5.8 Diagnostic plots for the FPCA performed on the 300°C desorption of the advanced data . . . . . 74

5.9 The $DD^G$ plots for the 300°C condition on the new set of data 77

## Abstract

Picomole is a New Brunswick based company developing a lung cancer diagnosis system based on a person's breath sample. Picomole conducted two different studies in an effort to ascertain whether their breath analysis system utilizing cavity ring-down laser spectroscopy is capable of determining whether or not a subject has lung cancer. One of the resulting datasets had a very large percentage of non-random missing data which is explored in detail.

Most breath analysis systems operate by trying to determine the makeup of volatile organic compounds and see if any are known signs of cancer. By contrast, the work done here is based entirely on statistical learning methods. Spectroscopy data is naturally a curve, as the concentrations of compounds are measured over a series of infrared wavelengths. This kind of data is referred to as functional data for which there exist unique techniques for dealing with problems specific to it. This motivated the consideration of techniques including Functional Principal Component Analysis, Functional Linear Discriminant Analysis and $DD^G$ plots. Classification trees and random forests which have previously shown success on spectroscopy data were also explored.

Classification trees, $DD^G$ plots and Functional Linear Discriminant Analysis were found to be able to correctly classify subjects with accuracy greater than random guessing.

## List of Abbreviations and Symbols Used

AIC Akaike Information Criterion

BIC Bayesian Information Criterion

CART Classification And Regression Trees

COPD Chronic Obstructive Pulmonary Disease

CT Computed Tomography

CV Cross Validation

DD Classifier Depth vs. Depth Classifiers

DD plot Depth vs. Depth Plot

FAD Functional Anderson-Darling

FDA Functional Data Analysis

FLDA Functional Linear Discriminant Analysis

FM depth Fraiman and Muniz depth

FN False Negative

FP False Positives

FPC Functional Principal Components

FPCA Functional Principal Component Analysis

GCMS Gas Chromatography Mass Spectrometry

hM depth h-mode depth

IMS Ion Mobility Spectrometry

kNN k-nearest neighbors

LDA Linear Discriminant Analysis

NP non-parametric

OOB Out-of-Bag

PACE Principal Analysis by Conditional Expectation

PC Principal Component

PCA Principal Component Analysis

PTR-MS Proton-transfer-reaction Mass Spectrometry

PVE Proportion of Variance Explained

SIFT-MS Selected Ion Flow Tube Mass Spectrometry

TN True Negative

TP True Positive

VOCs Volatile Organic Compounds

## Acknowledgements

Breath sample data was provided by Picomole.

# Chapter 1

# Introduction

Lung cancer screening plays an important part role in increasing survival rates. The gold standard method for lung cancer screening at the moment is to use a low dose Computed Tomography(CT) scan[28]. CT scans can be very costly and may involve long wait times due to limited access to expensive CT scanners and radiologists. CT scans have also been found to have a 96.4% false postive rate for lung cancer screening[25]. Picomole Inc. has developed a lung cancer screening method that could potentially be cheaper and faster and more accurate for patients. Picomole's system involves taking breath samples of subjects and then measuring the concentration of Volatile Organic Compounds (VOCs) in those samples. It has been previously shown that there is a difference in the VOCs emitted in the breath of subjects with lung cancer versus those without [28][21].

VOCs are gaseous chemicals and emissions found not only in human breath but also in skin, urine and feces. Over 1800 VOCs have been found to be emitted by the human body and can be impacted by environmental factors like cigarette smoking, alcohol consumption, medications and overall diet [1].

VOC concentration levels can be measured by a variety of different methods. Methods that have, or are currently being used include Gas Chromatography Mass Spectrometry (GCMS), Proton-transfer-reaction Mass Spectrometry (PTR-MS), selected ion flow tube mass spectrometry (SIFT-MS), ion mobility spectrometry (IMS), [1] and electronic noses [10]. Some of these methods are referred to as direct or real time analysis systems since they do not require storage of breath data. These include PTR-MS, SIFT-MS,IMS, laser spectrometry and electronic noses. There are also off-line analysis systems that require storage of the breath sample to be analyzed later, GCMS being one example. Several of these systems can actually do both real time and offline analysis, PTR-MS is an example. Picomole's system uses a form of laser spectrometry called cavity ring-down spectroscopy (CRDS) and requires stored

breath samples.

CRDS can be more cost effective and accurate than mass spectrometry but has not been used before in breath sample analysis.

For a preliminary screening test the ideal situation is to not miss any patients who actually do have lung cancer. That means avoiding false negatives (FN). In medical tests there are two commonly used measures; sensitivity and specificity. Sensitivity, is also known as the true positive (TP) rate and is defined as

$$100\% \times \frac{TP}{TP + FN}. \tag{1.1}$$

So when a test has 100% sensitivity it has zero false negatives. Ideally we would want all preliminary tests to have 100% sensitivity. However having 100% sensitivity is not the only thing that matters. Otherwise having a screening test that just sent every patient to the next stage of testing would be acceptable. Diagnostic tests can often be costly, time consuming or invasive which is why preliminary tests are needed in the first place.

The other measure commonly used to evaluate medical tests is specificity. Specificity is also known as the true negative(TN) rate and is defined as

$$100\% \times \frac{TN}{TN + FP} \tag{1.2}$$

where FP is the number of false positives[12]. A good screening test would have a high sensitivity while still maintaining an acceptable specificity rate. This would ensure patients that actually have the disease are not missed from receiving treatment while minimizing the number of patients who do not have the disease from undergoing a more invasive next round of testing.

A common and simple metric used to measure the performance of machine learning classifiers is the error rate. The error rate refers to how often the classifier mis-classifies and can be more formally written as

$$\text{Error Rate} = \frac{FN + FP}{FN + FP + TN + TP}. \tag{1.3}$$

Related to the error rate is accuracy of a classifier. The accuracy can be defined as

$$\text{Accuracy} = \frac{TN + TP}{FN + FP + TN + TP} = 1 - \text{Error Rate[12]}. \tag{1.4}$$

Current classification methods based on analyzing a subject's breath sample using VOCs require a list of known VOCs found in subjects with lung cancer and try to determine whether or not those VOCs can be found in that subject's breath sample[2][10][21][13]. This requires studying what (or in what concentrations) VOCs may be emitted in a breath sample for the type of cancer of interest (e.g. lung cancer). By using classification techniques based on machine learning methods it may be possible to build classification rules to identify lung cancer patients using only breath sample data from healthy subjects and subjects known to have the disease. Once shown in lung cancer machine learning methods could be easily applied to other diseases. This could potentially reduce development time, research time, costs and increase the number of diagnostic tests that Picomole could offer.

Machine learning or statistical learning involves learning from data, such as using a dataset of handwritten characters in order to develop a system capable of recognizing handwriting through some sort of algorithm. Machine learning can be broadly split into two groups. The first group, supervised learning is the case when the data upon which all the machine learning algorithm is trained, called the training data, has the values of the response variable known for each observation[12]. The other group of machine learning is unsupervised learning where the training data does not contain the values of the response variable and instead it tries to determine differences between classes based solely off the predictor variables[12].

Related to and a common use of machine learning is statistical classification or simply classification. It is when the data contain a finite number of classes and once trained, is able to identify what class a new set of data belongs to. Picomole's desire to determine whether a subject has lung cancer or not based on their breath sample is an example of a classification task. The method for deciding to what class a new observation belongs to is known as a classification rule.

Developing a classification rule for determining which subjects have or do not have lung cancer that has good overall accuracy and high sensitivity using Picomole's breath sample based system would lead to improved lung cancer screenings for patients.

# Chapter 2

# Review of Methodologies

## 2.1 Decision Trees

Decision trees refer to a collection of related machine learning methods based on the idea of building trees using tests on predictor variables. Decision tree methods are all based on recursively partitioning the data set into a number of regions. The regions are split based on rules derived from tests performed on the predictive variables. Since the resulting rules can be easily visualized in a tree-like structure the methods have come to be known as decision trees. The tree-like structure gives decision trees an easy to interpret rule system[18]. Over the years several different decision tree algorithms have been developed with differences in the criteria used to grow the tree, the number of trees grown, the number of children each parent node is allowed to have, etc. The main decision tree algorithms are also all invariant to monotone transformations[12].

### 2.1.1 Classification And Regression Trees

One of the first developed and popularized decision tree algorithms is referred to as Classification And Regression Trees (CART). CART was developed by Breiman et. al [6]. The CART algorithm as the name implies is capable of both classification and regression trees, however only classification trees will be examined here. CART, unlike some of the other alternative decision tree algorithms, only uses binary trees in it's construction, that is all nodes can have at most two children[6].

Like all decision tree algorithms CART begins by building the tree outwards using a splitting criterion to pick the optimal split. CART uses the Gini index as it's splitting criterion for classification trees. In a dataset with K classes the Gini index is defined as

$$G(m) = \sum_{k=1}^{K} p_{\hat{m}k}(1 - p_{\hat{m}k}) \tag{2.1}$$

where $m$ is the current partition in the data set and $p_{\hat{m}k}$ is the sample proportion

of observations in the $m$th partition that are in class $k$ for $i = 1, ...,$K. The CART algorithm splits a node when there is a maximal impurity reduction. The impurity of a split at a given node A is

$$\Delta I = p(A)G(A) - p(A_L)G(A_L) - p(A_R)G(A_R) \tag{2.2}$$

where $p(m) = p_{\hat{m}k}$ and L and R refer to the left and right nodes of the split respectively. The tree stops growing when there is either no change in the impurity reduction, or the change in the impurity reduction is below a threshold deemed insignificant or if it reaches another stopping threshold such that the node under consideration for splitting does not have the minimum number of observations required[6].

The original tree constructed by CART is typically going to overfit the data for which it was built. To account for this some nodes are removed via a process called pruning. Pruning is done by calculating the cost and risks of adding a new variable to the tree and then using k-fold cross validation (CV) to get estimates of the costs and risks for subtrees with or without that variable. In k-fold cross validation the original dataset is randomly partitioned into k different equally sized subsamples. One of the "folds" is withheld for use as a testing set and the remaining k-1 folds are used as the training dataset. This is repeated k times for each fold and the average of results over all k times is used as the result. The subtree that has the smallest risk is selected to be the final tree[6].

### 2.1.2 Random Forests

Random forests were developed by Breiman as an extension to CART in an effort to overcome some shortcomings such as having high variance and making decision trees applicable to more data sets. Random Forests are built on the idea of bagged trees. Bagging involves taking the average of a group of bootstrapped training samples. Each bootstrapped training sample is created by pulling N observations from the original sample with replacement, this process is called bootstrapping [18].

In bagged trees, a tree is grown for each of the bootstrapped training samples. By using bagged trees the variance of the statistical learning method can be reduced. If the number of bootstrapped classification trees is B, the bagged estimate $\hat{f}_{bag}(x)$ is found by taking the majority vote of what class an observation belongs to among the

B trees [18]. Bagged models have an alternative approach for calculating the error rate.

Bagged models can use what is known as the Out-of-Bag (OOB) error rate. For each of the B bootstrapped trees there are observations that are not included in the construction of the tree. These observations are outside of the "bag" and are used as a testing set for their corresponding tree. Predictions for the testing set on the B trees are made and in the case of classification the majority vote among all the predictions for a given observation is used to determine that observation's predicted class. Then the error rate is calculated using those predictions[18].

Random forests are more than just bagged trees. Random forests also limit the number of predictor variables that can be chosen each time a tree splits. If there are $p$ predictor variables the random forest algorithm takes a sample of $m$ of them at each split and is only allowed to build the split from those $m$ predictors. Restricting the choice of predictor variables has the benefit of decorrelating the trees. In ordinary bagged trees if there is one very strong predictor variable for a split each tree grown is likely to pick that predictor each time. Random Forests, by limiting the choice of what predictor variables can be used, improve the performance of the classifier on advanced data[18]. Breiman, one of the creators of Random Forests, recommends trying several values of $m$ predictors, growing small forests of 20 to 30 trees with those values of $m$ and picking the ones with the smallest OOB error rate[5].

## 2.2 Functional Data Analysis

Functional Data Analysis (FDA) is at the heart of many of the methods discussed here. FDA is a field that has seen significant growth and development over the last two decades since the publication of Ramsay and Silverman's 1997 book on Functional Data Analysis of the same name [23]. A functional random variable $\boldsymbol{\chi}$ takes values in an infinite dimensional or functional space. That is, $\boldsymbol{\chi} = \{\boldsymbol{\chi}(t); t \in T\}$, where $T$ is some infinite dimensional space. If $T$ is one-dimensional, then $\boldsymbol{\chi}$ refers to curve data like growth curves or spectrometry data. FDA can be applied to far more than one-dimensional curve data such as images, handwriting data, etc. [11].

Data that takes values in an infinite dimensional space can pose a different range of issues than that of non-functional data. Functional data problems will often have

sample sizes smaller than the number of predictor variables and in addition the predictor variables can also be very highly correlated. These problems can cause traditional multivariate methods to fail[11]. FDA methods are tailored to be used with data from an infinite dimensional space and can thus be used with data sets where traditional methods would fail. In practice it is not possible for us to observe the data over the entire infinite dimensional space, instead only a finite measure can be captured. The usual first step in using FDA is to represent the finite data in such a way that they can be treated as smooth curves[23]. However if the data is considered "dense" enough then sometimes smoothing is not used[14]. The types of curve representation systems range from the B-splines basis system for very smooth curves, Fourier basis system for periodic and cyclical data[23], wavelets, and to kernel smoothing using various types of different kernels (typically for use with non-parametric functional models)[11].The choice of curve representation should be motivated by the type of data being examined.

### 2.2.1 Functional Principal Component Analysis

Functional Principal Component Analysis (FPCA) is the functional analogue of the traditional dimension reduction technique known as Principal Component Analysis (PCA). FPCA like PCA is widely used for dimension reduction on functional data sets. In a review of 84 articles applying FDA techniques, FPCA was found to be used in over 60% of them and was the main data reduction technique applied [27]. FPCA has been applied to a number of real-world applications such as modeling the curvature of the human eye and fMRI (functional magnetic resonance imaging) scans of the brain[27]. FPCA is also the backbone of a number of FDA clustering techniques. This is because the principal component scores generated by the application of FPCA are capable of being used with traditional clustering techniques. It can also be used to visualize otherwise hard to perceive data sets[11]. FPCA has also been used as a method for imputing missing data on functional datasets[8].

The classical technique PCA is the process of computing and analyzing the principal components (PC). PCA is used to create a low dimensional representation of a dataset while maximizing the amount of variation it contains. The PCs try to explain as much of the variance as they can while being uncorrelated with each other. In any

one dataset there are $\min(\mathrm{n}-1, p)$ possible principal components, where n is the number of observations and p is the number of predictor variables. The first PC, or score vector is the linear combination of predictor variables

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + ... + \phi_{p1}X_p \tag{2.3}$$

with the constraint $\sum_{j=1}^{p} \phi_{j1}^2 = 1$ and that maximizes the variance. Each element of the PC, is known as the principal component score and is written as

$$z_{ij} = \phi_{1j}x_{i1} + \phi_{2j}x_{i2} + ... + \phi_{pj}x_{ip}. \tag{2.4}$$

The $\phi_{i1}$s are the loadings or weights of the first PC and they make up the first PC loading vector, $\phi_1 = (\phi_{11}, \phi_{21}, ..., \phi_{p1})^T$. The first PC loading vector solves

$$\max_{\phi_{11},...,\phi_{p1}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( \sum_{j=1}^{p} \phi_{j1}x_{ij} \right)^2 \right\}. \tag{2.5}$$

The $\left(\sum_{j=1}^{p} \phi_{j1}x_{ij}\right)^2$ from above can be rewritten as $z_{i1}^2$. PCA is done under the assumption that the predictor variables have been centered to have mean zero. This implies that the $z_{ij}$s also have mean zero. This means that Equation 2.5 is just maximizing the sample variance of $z_{1i}$. The second PC can be obtained by finding the next vector that maximizes the variance with the extra constraint of being orthogonal to the first principal component. Being orthogonal to each other is equivalent to being uncorrelated to each other. The PCs for the remaining dimensions are found in the same way, maximizing the variation explained by the vector while ensuring they are orthogonal to the previous PCs[18]. Since the predictor variables are assumed to have mean zero the total variation present in the dataset can be calculated by

$$\sum_{j=1}^{p} \mathrm{Var}(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2, \tag{2.6}$$

and the variation explained by the *mth* PC is

$$\frac{1}{n} \sum_{i=1}^{n} z_{im}^2 \tag{2.7}$$

and the proportion of variance explained(PVE) by that component is their ratio between them

$$\frac{\frac{1}{n}\sum_{i=1}^{n} z_{im}^2}{\sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2}. \tag{2.8}$$

The sum of all the PVE of the principal components is equal to one. The total variation explained by the first $m$ principal components is simply the sum of the first $m$ principal components ordered by descending PVE[18].

The goal of FPCA is similar to that of PCA with some differences to account for the data's functional nature. FPCA involves finding the orthogonal basis of eigenfunctions that explain the maximum amount of variation in the first functional principal component. For a random function X defined on $[0, T]$ that has an unknown smooth mean function $E[X(t)] = \mu(t)$ and covariance function $\text{Cov}(X(s), X(t)) = G(s, t)$, $s, t \in [0, T]$ one can write the orthogonal expansion

$$G(s,t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s) \phi_k(t) \tag{2.9}$$

where $\lambda_k$ are the eigenvalues and $\phi_k$ are eigenfunctions that form an orthonormal basis with a unit norm. This expansion enables each functional observation to be written as

$$X_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik} \phi_k(t) \tag{2.10}$$

where $\xi_{ik} = \int (X_i(t) - \mu(t)) \phi_k(t) dt$ are Functional Principal Components. This is known as the Karhunen-Loève representation. The estimated eigenfunctions $\hat{\phi}_k$ and eigenvalues $\hat{\lambda}_k$ are obtained by applying eigen-decomposition on the estimated covariance function[8]. If the grid is fine enough and there are no missing values it is possible to get the estimated FPC $\hat{\xi}_{ik}$ by numerical integral approximation[23]. In the cases where the functional data is not dense or there is a significant amount of missing data the approximation of the integral fails. In those cases FPCA can still be accomplished using what is called Principal Analysis by Conditional Expectation (PACE)[30].

When FPCA is computed typically only the first k FPCs are calculated. These are either selected beforehand or because they explain a given threshold of the variance. When using a threshold of the total variance explained, the PVE for each FPC is calculated and when the sum of PVEs for the currently calculated FPC exceeds the threshold the calculation of more FPCs is stopped. The estimated FPC can be used to help interpret the data like principal components can be for multivariate data. They have also been used in two-sample hypothesis testing[7][15].

### 2.2.2 Functional Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a classical technique used for classification of data into two or more classes. In LDA we find the class conditional probabilities for each of the predictor variables and then use Bayes Theorem to get the probabilities of an observation falling into a class given the predictor variable. More formally if we define $f_k(x)$ as the density function for $x$ given a class $k$, when there are K classes and $\pi_k$ is the prior probability of each class where $\sum_{i=1}^{K} \pi_k = 1$, then we obtain

$$P(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{i=1}^{K} f_i(x)\pi_i}. \tag{2.11}$$

We can then classify $x$ according to the class with the highest resulting probability. LDA is a very simple technique that provides good results with many data sets[12]. However LDA is not designed to be used on functional datasets[17].

Functional Linear Discriminant Analysis (FLDA) is an extension of Linear Discriminant Analysis with the capability to handle functional data and in addition deal with data that might only be measured in segments of curves. FLDA was developed by Gareth James & Trevor Hastie for classifying sparse growth curve data and bone density data[17]. They define their functional model as

$$\boldsymbol{Y}_{ij} = \boldsymbol{g}_{ij} + \varepsilon_{ij}, \quad i = 1, ...K \quad j = 1, ..., m_i \tag{2.12}$$

with K classes, where j refers to the curve, $m_i$ is the number of curves in the ith class, $\boldsymbol{g}_{ij}$ is the curve itself with each measurement point at position t being $\boldsymbol{g}_{ij}(t)$ and $\varepsilon$ is the random error. In their paper they use natural cubic functions to represent curves. Each $\boldsymbol{g}_{ij}(t)$ can be represented as

$$\boldsymbol{g}_{ij}(t) = \boldsymbol{s}(t)^T \boldsymbol{\eta}_{ij} \tag{2.13}$$

where $\boldsymbol{s}(t)$ is the natural cubic spline basis with dimension q and $\boldsymbol{\eta_{ij}}$ is the qth dimensional vector of spline coefficients. This enables $\boldsymbol{Y}_{ij}$ to be rewritten in a more restricted form since

$$\boldsymbol{g}_{ij} = \boldsymbol{S}_{ij}\boldsymbol{\eta}_{ij} \quad \text{where} \quad \boldsymbol{S}_{ij} = (\boldsymbol{s}(t_{ij1}), ..., \boldsymbol{s}(t_{ijn_{ij}}))^T. \tag{2.14}$$

Substituting this expression into Equation 2.12 gives us a reduced model version where the only thing that needs to be estimated is the value of $\boldsymbol{\eta}_{ij}$. James et. al

assume $\boldsymbol{\eta}_{ij}$ and the error terms are normally distributed as

$$\boldsymbol{\eta}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\gamma}_{ij}, \quad \boldsymbol{\gamma}_{ij} \sim N(\mathbf{0}, \boldsymbol{\Gamma}) \tag{2.15}$$

Where $\boldsymbol{\mu}_i$ is the mean function for the ith class. The functional model can then again be rewritten as

$$\boldsymbol{Y}_{ij} = \boldsymbol{S}_{ij}(\boldsymbol{\mu_i} + \boldsymbol{\gamma}_{ij}) + \varepsilon_{ij}, \quad i = 1, ...K, j = 1, ..., m_i, \varepsilon \sim N(\mathbf{0}, \sigma^2 I), \quad \gamma_{ij} \sim N(\mathbf{0}, \Gamma). \tag{2.16}$$

FLDA uses the constraints

$$\boldsymbol{\mu}_i = \boldsymbol{\lambda}_0 + \boldsymbol{\Lambda}\boldsymbol{\alpha}_i, \quad \boldsymbol{\Lambda}^T \boldsymbol{S}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{S} \boldsymbol{\Lambda} = \boldsymbol{I}, \quad \sum_i \boldsymbol{\alpha}_i = 0, \tag{2.17}$$

where $\boldsymbol{\lambda}_0$ is a q dimensional vector, $\boldsymbol{\alpha}_i$ is an h-dimensional vector, $\boldsymbol{\Lambda}$ is a a $q \times h$ matrix, $h < \min(q, K)$, $\Sigma = \sigma^2 \boldsymbol{I} + \boldsymbol{S}\boldsymbol{\Gamma}\boldsymbol{S}^T$ and $\boldsymbol{S}$ is the basis matrix which has been evaluated over a grid of points.The grid used in $\boldsymbol{S}$ should contain all the measurement points that were actually observed. These constraints are necessary to prevent confounding of $\boldsymbol{\lambda_0}, \boldsymbol{\Lambda}$ and $\boldsymbol{\alpha_i}$. Using these constraints gives the final version of the FLDA model as

$$\begin{aligned} \boldsymbol{Y_{ij}} &= \boldsymbol{S}_{ij}(\boldsymbol{\lambda}_0 + \boldsymbol{\Lambda}\boldsymbol{\alpha}_i + \boldsymbol{\gamma}_{ij}) + \boldsymbol{\varepsilon}_{ij}, \quad i = 1, ...K, \quad j = 1, ..., m_i, \\ \boldsymbol{\varepsilon} &\sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}), \quad \boldsymbol{\gamma_{ij}} \sim N(\mathbf{0}, \boldsymbol{\Gamma}), \end{aligned} \tag{2.18}$$

FLDA fits model 2.18 estimating $\boldsymbol{\lambda}_0$,$\boldsymbol{\Lambda}$,$\boldsymbol{\alpha}_i$,$\boldsymbol{\Gamma}$ and $\sigma^2$. It uses a version of the EM algorithm to perform the model fit. The way it is used in FLDA is to treat the $\boldsymbol{\gamma}_{ij}$ as missing data and then maximize the joint likelihood. It assumes that all the observations from different individuals are independent. The E step is done using the equation

$$E(\boldsymbol{\gamma}_{ij} | \boldsymbol{Y_i}, \gamma_0, \Lambda, \alpha_i, \Gamma, \sigma^2) = (\sigma^2 \Gamma^{-1} + S_{ij}^T S_{ij})^{-1} S_{ij}^T (\boldsymbol{Y}_{ij} - S_{ij}\lambda_0 - S_{ij}\Lambda\alpha_i) \tag{2.19}$$

Then the M-step for maximizing the joint likelihood requires maximizing

$$\begin{aligned} Q = -\frac{1}{2} \sum_{i=1}^{K} \sum_{j=1}^{m_i} E\big\{ &\frac{(\boldsymbol{Y}_{ij} - S_{ij}[\lambda_0 - \Lambda\alpha_i - \gamma_{ij}])^T (\boldsymbol{Y}_{ij} - S_{ij}[\lambda_0 - \Lambda\alpha_i - \gamma_{ij}])}{\sigma^2} + \\ &+ n_{ij} \log(\sigma^2) + \gamma_{ij}^T \Gamma^{-1} \gamma_{ij} + \log|\Gamma| \big\} \end{aligned} \tag{2.20}$$

[17]. The choice of $q$ is important to the model selection process. James et al. recommends either calculating the cross-validated likelihoods for the choices of $q$ under

consideration and choosing the model which maximizes the likelihood, or alternatively to use another model selection technique like Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) [17].

It can be seen that classifying an observation $\boldsymbol{X}$ using reduced-rank LDA is the same as classifying

$$\arg\min_i(||\hat{\alpha}_x - \alpha_i||^2 - 2\log\pi_i) \tag{2.21}$$

where $\hat{\alpha}_x$ and $\alpha_i$ are equal to the linear discriminants of $\boldsymbol{X}$ and $\boldsymbol{\mu_i}$ up to an additive constant. A similar result can be found with FLDA. The FLDA classification reduces to

$$\arg\min_i\left(||\hat{\alpha}_{\boldsymbol{Y}} - \alpha_i||^2_{Cov(\hat{\alpha}_{\boldsymbol{Y}})^{-1}} - 2\log\pi_i\right) \tag{2.22}$$

where $\text{Cov}(\hat{\alpha_{\boldsymbol{Y}}}) = (\Lambda^T S_{\boldsymbol{Y}}^T \Sigma_{\boldsymbol{Y}}^{-1} S_{\boldsymbol{Y}} \Lambda)^{-1}$ and once again $\hat{\alpha}_{\boldsymbol{Y}}$ and $\alpha_i$ are again similar to the linear discriminants of $\boldsymbol{Y}$ and $\boldsymbol{\mu_i}$ with the only difference being an additive constant. If $Cov(\hat{\alpha_{\boldsymbol{Y}}}) = \boldsymbol{I}$ such as when $\boldsymbol{Y}$ has been measured over all grid points then Equation 2.22 reduces further to

$$\arg\min_i\left(||\hat{\alpha}_{\boldsymbol{Y}} - \alpha_i||^2 - 2\log\pi_i\right). \tag{2.23}$$

[17]

FLDA, like traditional LDA, can be used as a data dimension reduction technique enabling high dimensional curves to be reduced to a simpler lower dimensional space and make classification possible. FLDA can also be extended to Functional Quadratic Discriminant Analysis in a similar way to how LDA can be extended to Quadratic Discriminant Analysis and can also be extended to Functional Regularized Discriminant Analysis.[17]

### 2.2.3 Two-Sample Tests for Functional Data

Dealing with data that takes on values in an infinite space with possibly very correlated predictor variables can cause traditional multivariate hypothesis testing methods to fail or lead to invalid results. This has led to techniques being extended from existing multivariate methods or created with functional data in mind[15].

**Schilling's Statistic Permutation Test**

Cabaña et al. extended the multivariate Schilling procedure to deal with functional data.

In the multivariate Schilling procedure suppose we have two samples $X_1, ..., X_m$ i.i.d from a distribution F and $Y_1, ..., Y_n$ from a distribution G. In order to test $H_0 : F = G$ vs. $H_A : F \neq G$ the Schilling Procedure begins by concatenating the two samples together into $Z_1, ..., Z_N$ where $N = m + n$. Then for some fixed non-zero integer k for each $i \leq N$ it finds the k-nearest neighbors based on the Euclidean distance of each $Z_i$ to each other $Z_j$. Let $\mathrm{NN}_i(r)$ represent the $r$th nearest neighbor to $Z_i$ where $r \leq k$. If there are any ties for distance they are broken at random. Then for each $\mathrm{NN}_i(r)$ it computes the indicator variable $I_i(r)$ where $I_i(r) = 1$ if $\mathrm{NN}_i(r)$ belongs to the same original sample as $Z_i$ and $I_i(r) = 0$ otherwise. The test statistic is the computed,

$$T_{N,k} = \frac{1}{Nk} \sum_{i=1}^{N} \sum_{r=1}^{k} I_i(r). \tag{2.24}$$

$T_{N,k}$ is the proportion of all neighbor pairs in which the point and it's neighbor both belong to the same sample. So when $H_0$ does not hold $T_{N,k}$ will be high as observations from the same sample will be grouped together. It can be shown that under the null hypothesis that

$$E[T_{N,k}] = E[I_i(r)] = \frac{m(m-1) + n(n-1)}{N(N-1)}. \tag{2.25}$$

[7]

Cabaña et al. extended the Schilling procedure to handle functional data by replacing the Euclidean distance measure used in the multivariate case by a distance measure approximation designed for functional data. Cabaña et al. suggest for data measured on a common grid to use

$$d_{i,j} = \sum_{l=1}^{L} \delta_l (Z_i(t_l) - Z_j(t_l))^2 \tag{2.26}$$

where $\delta_l = t_l - t_{l-1}$ and $l = 1, 2, ..., L$ is each point of grid. Cabaña et al. used this functional version of the Schilling procedure in a permutation test. They found it was sufficient to simply permute the table containing the $\mathrm{NN}_i(r)$s instead. This makes performing the permutation test much faster than permuting the data itself. Among

the functional hypothesis tests compared by Cabaña et al. it was among the best for statistical power.[7]

**Functional Anderson-Darling Test**

The Functional Anderson-Darling two-sample test for functional data was developed by Pomann et al. based on FPCA using an appropriate mixture process that they called marginal FPCA. This test, unlike other two-sample tests for functional data, is capable of handling data measured on sparse grids, dense grids. It can compare samples of different sample sizes and samples based on different sampling designs.[22]

Given two sets of curves $X_1 = \{X_{11}(\cdot), ..., X_{1n}(\cdot)\}$ and $X_2 = \{X_{22}(\cdot), ..., X_{2n}(\cdot)\}$ defined on [0,1] it is of interest to test

$$H_0 : X_1 =^d X_2 \quad \text{vs.} \quad H_A : X_1 \neq^d X_2 \tag{2.27}$$

where $=^d$ indicates that the two samples are equal in distribution. Let $X_1(\cdot)$ and $X_2(\cdot)$ have a mixture process $X(\cdot)$ where $X_1(\cdot)$ has mixture probability $p$ and $X_2(\cdot)$ has mixture probability $1 - p$. If a binary variable Z is defined with $P(Z = 1) = p$ and $P(Z = 2) = 1 - p$ then it can be said that $X_1(\cdot)$ is the conditional process of $X(\cdot)$ given $Z = 1$ and $X_2(\cdot)$ is the conditional process of $X(\cdot)$ given $Z = 2$. Let $X(\cdot)$ have marginal mean function $\mu(t) = E[X(t)]$ and marginal covariance function $\Sigma(t, s) = \text{Cov}\{X(t), X(s)\}$. Recall from section 2.2.1 that it is possible to represent a functional process in terms of the Karhunen-Loève expansion. Thus $X(\cdot)$ can be written as

$$X(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_k \phi_k(t) \tag{2.28}$$

The Karhunen-Loève expansion is often truncated for practical or theoretical reasons with a choice of some integer K that can reasonably approximate $X(\cdot)$. The approximation can be written as

$$X^K(t) = \mu(t) + \sum_{k=1}^{K} \xi_k \phi_k(t). \tag{2.29}$$

K can be chosen by setting a threshold for how much variance is explained by the FPCs. Pomann et al. recommend using 95% of variance explained. Then for each $z$ one can write the approximation for $X_1(t) = X_1^K(t) = \mu(t) + \sum_{k=1}^{K} \xi_{1k} \phi_k(t)$ and

$X_2(t) = X_2^K(t) = \mu(t) + \sum_{k=1}^{K} \xi_{2k}\phi_k(t)$. It can be shown that the test in equation (2.27) is equivalent to the test

$$H_0^K : \{\xi_{1k}\}_{k=1}^K =^d \{\xi_{2k}\}_{k=1}^K \quad \text{vs.} \quad H_A^K : \{\xi_{1k}\}_{k=1}^K \neq^d \{\xi_{2k}\}_{k=1}^K. \tag{2.30}$$

As the name implies the Functional Anderson-Darling test is an extension of the Anderson-Darling test for functional data. As the test given in Equation 2.30 hints at, Anderson-Darling tests are performed for each of the K FPCs. Since K comparisons are performed the Bonferroni correction is used to ensure the test maintains it's nominal size. The test is rejected if $\min_{1\leq k\leq K} p_k \leq \frac{\alpha}{K}$[22].

This method of marginal FPCA can be used with more than the Anderson-Darling test. It can be used with other two sample tests including other univariate tests such as the Kolmogorov-Smirnov test using K comparisons or using a multivariate two sample test such as Schillings[22].

## 2.3 DD$^G$ Plot Classifiers

Data depths are measures that order observations in relation to how deep they are relative to a probability distribution $\boldsymbol{P}$. In the univariate case there is an obvious measure of data depth, the median is the deepest point with the points being further out from the median being considered less deep. In multivariate and functional statistics the choice of the data depth metric is not quite as obvious. Several different metrics have been developed to handle functional data. Data depth measures have been used to provide visualization[24], inference[20][7] along with classification and clustering[19][9].

There has been no consensus on what is the best measure of data depth. Due to differences between functional and multivariate datasets there have been different depth measures purposed for each. Due to the dataset under consideration here being functional in nature only functional data depth measures will be discussed[9].

The first depth measure proposed for functional data is the Fraiman and Muniz depth(FM depth), it is also sometimes called integrated depth. Given a functional sample $x_1, ..., x_N$ defined on $[0, T]$, let $S_t = \{x_1(t), x_2(t), ..., x_N(t)\}$ be the values of those functions at a given t belonging to T. If $F_{N,t}$ is the empirical distribution of $S_t$

then the univariate depth measure $D_i(t)$ for $x_i(t)$ is written as

$$D_i(t) = 1|\frac{1}{2} - F_{N,t}(x_i(t))|. \tag{2.31}$$

As suggested by the alternate name integrated depth, the FM depth for the ith piece of data is the integration of the univariate depth measure over the entire grid,

$$FM_i = \int_0^T D_i(t)dt. \tag{2.32}$$

The FM depth can be generalized to use other univariate depth measures such as the Half-space depth or Mahalanobis depth[9].

Another measure of depth for functional datasets is the h-mode depth (hM). For a given random functional sample $x_1, ..., x_N$ of X the empirical hM depth is defined as

$$\hat{f}_h(x_0) = N^{-1} \sum_{i=1}^{N} K(m(x_0, x_i)/h) \tag{2.33}$$

where $K(\cdot)$ is a kernel, h is the bandwidth parameter and m is a metric or semi-metric. [9]

Depth vs. Depth Classifiers are classifiers constructed on Depth vs. Depth(DD) plots. DD plots were created for the purpose of graphically comparing two multivariate distributions or samples. Consider two random samples $\boldsymbol{X}$ and $\boldsymbol{Y}$ from distributions F and G respectively. Then a DD plot is defined as

$$DD(F, G) = \{(D_F(x), D_G(x)), x \in \boldsymbol{X} \cup \boldsymbol{Y}\} \tag{2.34}$$

where $D(x)$ is some measure of depth. If the distributions of F and G are unknown then they can be substituted with their empirical versions. DD plots are capable of showing differences between two distributions visually since differences in things like location, scale and kurtosis show up as different patterns[19].

Since DD-plots are constructed to always be two dimensional the standard DD-classifier uses majority voting among every combination of groups. The original DD-classifiers use a polynomial function up to some order $k$ that includes the origin as their classification rule. This means that for g groups in a sample size of N, $2\binom{g}{2}\binom{N}{k}$ polynomials must be calculated to determine the classification rule[9]. For some problems this can lead to very long calculation times. Recently Cuesta-Albertos et.

al extended the DD-classifier to work with more than two groups along with extending it to work with more than just polynomial function classifiers and functional data. Cuesta-Albertos et. al call this the $DD^G$-classifier as it is capable of handling G groups natively. The $DD^G$-classifier begins by selecting some measure of depth D such as FM depth and it computes the map

$$x \to \boldsymbol{d} = (D_1(x), ..., D_g(x)) \in \mathbb{R}^G. \tag{2.35}$$

With this mapping to g-dimensional space it is now possible to use any classifier that can classify G groups such as LDA, logistic regression models, k nearest neighbors, classification trees, etc.

For DD and $DD^G$ classifiers the choice of depth can have an effect on the classification result. Li et al. recommend finding the best depth using cross validation for their DD-classifier[19]. Cuesta-Albertos et al. instead recommend computing the bias-corrected distance correlation between the vector of depths $\boldsymbol{d}$ and the indicator of the classes,

$$Y = (\mathbf{1}_{\{x \in C_1\}}, \mathbf{1}_{\{x \in C_2\}}, ..., \mathbf{1}_{\{x \in C_g\}}) \tag{2.36}$$

and then selecting the depth that maximizes the distance.[9]

# Chapter 3

# Data Collection Methods

Data were collected by Picomole to study the capacity of their technology to discriminate between healthy subjects and those with lung cancer (or Chronic Obstructive Pulmonary Disease (COPD)). Picomole created a system wherein they capture the VOCs from the breath of subjects in a sorbent tube and then analyze the concentration of the captured compounds. They conjecture that it is possible to differentiate between healthy individuals and sick individuals based on differing concentrations of VOCs. This idea has been studied before and a variety of VOCs have been previously identified as being found in the breath of subjects with lung cancer but either not (or at a different levels) in that of healthy subjects[21][2].

As previously mentioned, Picomole's system uses CRDS to determine the concentration of VOCs in a breath sample. Picomole's CRDS system measures 77 different wavelengths. The wavelengths range from 9.2295$\mu$m to 11.3099$\mu$m. The 77 measures are not evenly spaced and they are separated into six different groups with some overlap between two of these groups. The wavelengths measured by Picomole and their spacing are shown in Figure 3.1. These wavelengths were chosen because they are the ones believed to best be able to capture the concentrations of the VOCs of interest and because they are captured by Picomole's $CO_2$ and Carbon isotope 13 lasers. The 77 wavelengths measured by Picomole's CRDS system are also referred to by an ID number known as the line ID.

CRDS involves shooting a laser pulse into an optical cavity with mirrors at each end and measuring the time it takes for the laser pulse to decay. This is called a ring-down time. Picomole's system tries to capture 500 ring-down times for each laser wavelength measurement. Sometimes however a line is unable to achieve the tuning frequency and it may end up with less or more than 500 measurements recorded. The system will keep trying to capture 500 ring-down times, but sometimes it cannot get tuned quickly and will take more than 500 tries and other times it cannot get tuned at

all which results in less than 500 ring-down times. The average of these measurements for the data based on the breath samples is referred to as $\tau$. Sometimes when a line ID would only get a handful of raw measurements then the $\tau$ value for that observation is considered missing.

By measuring an empty cavity and one filled with a gas capable of absorbing light it is possible to get measurements as small as one part in $10^7$[31]. Picomole also takes measurements of a cavity filled with nitrogen before every breath sample is measured as a baseline. The average of the ring-down times based on the measurement of nitrogen tube is known as $\tau_0$. Using $\tau$ and $\tau_0$ along with the speed of light (c) it is possible to get the absorption coefficient K in parts per million per centimeter where K is

$$K = \frac{\tau_0 - \tau}{c\tau_0\tau}. \tag{3.1}$$

In CRDS the temperatures at which a sample is desorbed affects what compounds or VOC concentrations are measured. However, Picomole is unsure of what temperature (or temperatures) is best for capturing the VOCs of interest. Therefore Picomole originally ran each sample through three different desorption temperatures: 80°C, 200°C and 300°C in the case of prototype data. More detail about the data collection methods used for the prototype data are provided in Section 3.1. The advanced data has a different sequence of desorptions which are explained in section 3.2.

Picomole's preliminary tests have shown that certain substances can have a serious impact on the measurements recorded by their CRDS system. Ethanol, found in alcoholic drinks, is known to completely overwhelm measurements of other VOCs in a person's breath sample. Due to these issues, subjects were asked to refrain from drinking alcohol 24 hours prior to having their breath sampled. Breath samples also naturally decay and must be used within a few days if stored at room temperature or a few weeks if frozen. Meaning that they must be analyzed within these time frames to be useful.

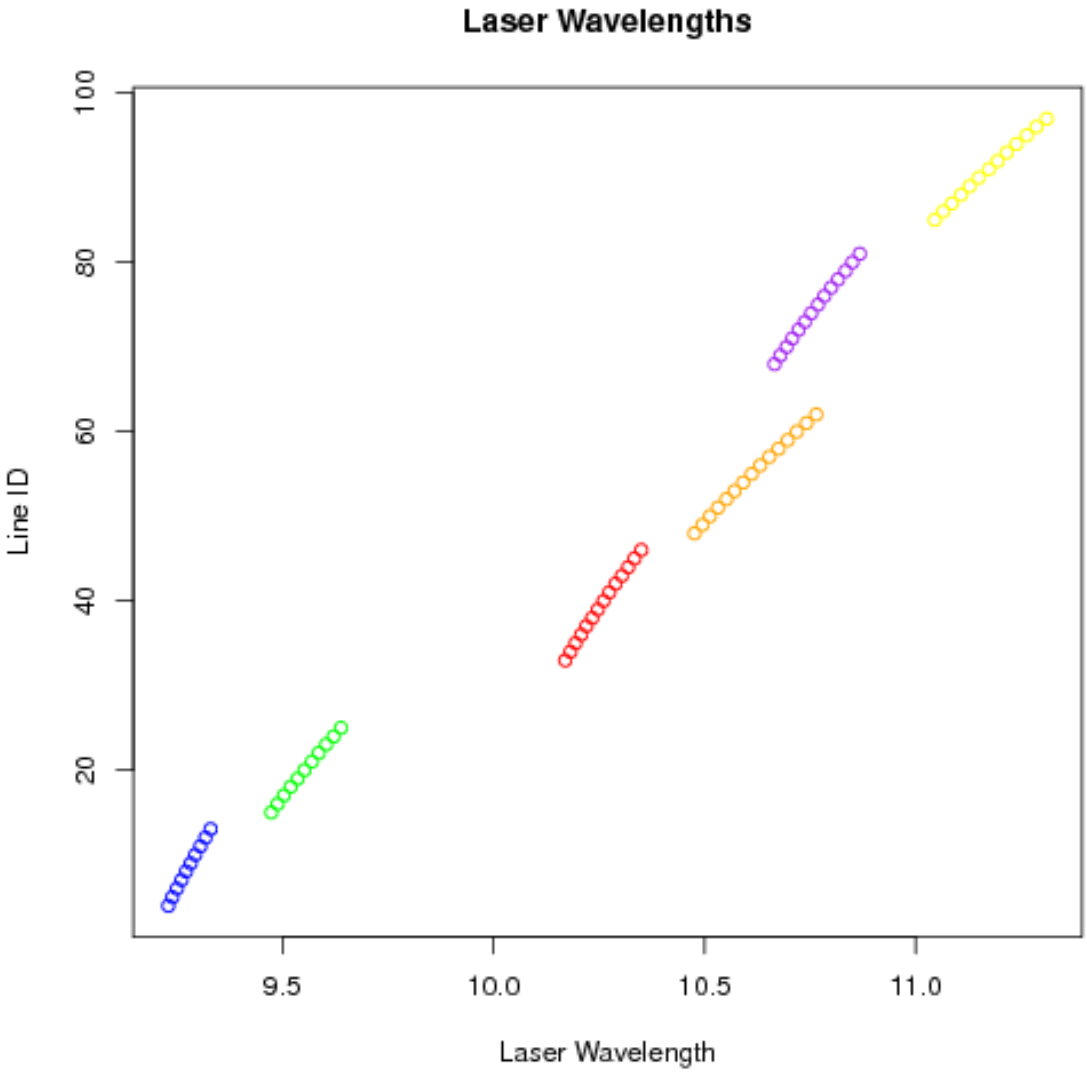


Figure 3.1: The spacing of the wavelengths measured by Picomole's CRDS analyzer. The overlapping wavelengths in the purple and orange groups are clearly visible. It is also possible to see the unequal numbers of measurements that occur in each group.

### 3.1 Prototype Data

Picomole collected data from February 18, 2012 to July 16, 2012 from 47 subjects, 20 of them healthy, 20 with lung cancer and 7 with COPD. There was no age matching done on this sample and all subjects were non-smokers at time of collection.

Picomole originally planned to have multiple breath samples collected from each

person and run each of them through the three different temperatures, 80°C, 200°C, 300°C. However for various reasons this was changed part way through the study and instead each sample was run through a sequence of temperatures. Most of the data comes from a group of breath samples that went through three desorb cycles using the sorbent material Tenax TA(Sorbent Type 2) ,one at 80°C, then another cycle at 80°C and finally one at 300°C. By running the sample at the 80°C temperature twice, it could be ensured that any VOCs recorded at that temperature would be fully captured and not have impact on the higher temperatures. Most of the remaining usable measurements come from samples that only went through one desorb cycle using the sorbent material Chromosorb 106 (Sorbent Type 1) at 200°C. The choice of sorbent material depends on the temperature at which the sample is being desorbed. There are also differences in the compounds favored by each of the sorbent materials. The Tenax material is better at capturing heavier compounds, while Chromosorb 106 is better at smaller compounds. The remaining conditions were not used in any analysis due to having only one or two measurements. The full breakdown of what conditions were used and how many measurements were obtained are shown in Table 3.1.

| | Total Desorbs | Current Desorb | Desorb Temp.(°C) | Sorbent Type | # of Measurements |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 200 | 1 | 41 |
| 2 | 1 | 1 | 300 | 2 | 1 |
| 3 | 2 | 1 | 80 | 2 | 2 |
| 4 | 2 | 2 | 300 | 2 | 2 |
| 5 | 3 | 1 | 80 | 2 | 43 |
| 6 | 3 | 2 | 80 | 2 | 43 |
| 7 | 3 | 3 | 300 | 2 | 43 |

Table 3.1: The organization of the experiment in Picomole's pilot study. Picomole changed plans part way through which is why some sets of conditions only have one or two measurements.

Picomole did not randomize the order in which breath samples were desorbed. Nor did they do any sort of pairing between healthy subjects and those with COPD or Lung Cancer. They instead gave priority to lung cancer or COPD subject breath samples. This led to an unintentional batching of most of the lung cancer subject samples being analyzed together around the same period of time. The consequences of this batching will be shown in depth in section 4.1.

| | Measurement Time | Health Status |
|---|---|---|
| 1 | 2012-02-18 17:12:50 | Healthy |
| 3 | 2012-02-20 16:05:38 | Healthy |
| 6 | 2012-02-23 14:41:32 | Healthy |
| 13 | 2012-02-27 12:45:26 | Healthy |
| 17 | 2012-02-28 10:15:00 | Healthy |
| 21 | 2012-02-28 14:43:58 | Healthy |
| 22 | 2012-03-02 10:35:01 | Healthy |
| 26 | 2012-03-01 07:22:46 | COPD |
| 30 | 2012-03-01 15:42:06 | Healthy |
| 34 | 2012-03-03 14:42:42 | Healthy |
| 38 | 2012-03-09 12:28:10 | Lung Cancer |
| 45 | 2012-03-12 10:08:13 | Lung Cancer |
| 46 | 2012-03-13 14:12:11 | Healthy |
| 50 | 2012-03-14 12:16:48 | Lung Cancer |
| 60 | 2012-03-18 17:28:23 | Lung Cancer |
| 61 | 2012-03-19 13:01:04 | COPD |
| 71 | 2012-03-23 14:34:22 | Lung Cancer |
| 75 | 2012-03-26 14:41:00 | Healthy |
| 79 | 2012-03-26 10:08:15 | Lung Cancer |
| 83 | 2012-03-29 14:22:53 | Lung Cancer |
| 87 | 2012-03-30 13:23:21 | Lung Cancer |
| 91 | 2012-03-30 21:52:45 | Lung Cancer |
| 95 | 2012-04-02 15:21:54 | COPD |
| 102 | 2012-04-03 16:12:53 | Lung Cancer |
| 103 | 2012-04-04 14:52:21 | COPD |
| 107 | 2012-04-05 09:37:00 | Lung Cancer |
| 114 | 2012-04-12 11:12:57 | Lung Cancer |
| 118 | 2012-04-13 13:00:05 | COPD |
| 119 | 2012-04-30 13:12:58 | Healthy |
| 123 | 2012-05-02 14:03:59 | Healthy |
| 127 | 2012-05-03 14:14:01 | Healthy |
| 134 | 2012-05-09 16:01:10 | Healthy |
| 138 | 2012-05-07 14:51:18 | Healthy |
| 141 | 2012-05-08 12:16:43 | Healthy |
| 142 | 2012-05-10 14:01:42 | Healthy |
| 146 | 2012-05-16 14:20:54 | Healthy |
| 153 | 2012-05-17 15:03:46 | Lung Cancer |
| 157 | 2012-05-22 15:38:18 | Lung Cancer |
| 161 | 2012-06-13 16:13:39 | COPD |
| 168 | 2012-06-20 14:58:06 | Lung Cancer |
| 172 | 2012-06-20 14:17:14 | Lung Cancer |

Table 3.2: The order in which subjects' samples were processed from the 200°C desorb temperature measurements. Note the clustering of Lung Cancer subjects between March 9th and April 12th. This pattern is similar for the other desorb temperatures.

### 3.2 Advanced Data

Picomole applied to obtain additional breath sample data from patients with lung cancer. Using the method of sample size planning for classification models given in Beleites et. al[4] it was determined that for a 95% confidence interval width of 5% sensitivity for a classifier then a sample size of 500 is needed to reach the 88.9% sensitivity of CT scans for detecting lung cancer[26]. Beleites et. al's method treat the sensitivity of a binary classifier as a Bernoulli process. They then use a binomial distribution with a uniform Beta distribution prior to find the minimum sample size at which the desired confidence interval width occurs[4]. They also outline a way to find the sample size needed for a given power better than some other classifier at a specific sensitivity level. It was found that in order to show an improvement over the CT scan sensitivity of 88.9% at 80% power a sample size of around 300 are needed for an expected test sensitivity of 93% and for an expected test sensitivity of 95% around 200 subjects would be needed. Picomole applied for and received approval to collect breath samples from 200 lung cancer subjects and 200 control samples.

| | Total Desorbs | Current Desorb | Desorb Temperature | Sorbent Type | # of Measurements |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 200 | 1 | 16 |
| 2 | 3 | 1 | 75 | 1 | 16 |
| 3 | 3 | 2 | 150 | 1 | 16 |
| 4 | 3 | 3 | 200 | 1 | 16 |
| 5 | 4 | 1 | 75 | 2 | 16 |
| 6 | 4 | 2 | 150 | 2 | 16 |
| 7 | 4 | 3 | 225 | 2 | 16 |
| 8 | 4 | 4 | 300 | 2 | 16 |

Table 3.3: The organization of the experiment in Picomole's real data. Picomole changed the setup of the experiment from that of the prototype data. The majority of lung cancer subjects available at this time were also all post-treatment which resulted in only post treatment subjects being used. These changes make conclusions drawn from the prototype data not generally relevant to the advanced data.

Picomole started to collect breath samples on the new sample in late 2017 with collection continuing on into 2018. Picomole made changes to their experimental design. The new design setup is shown in Table 3.3. Since Picomole is still in the process of collecting breath samples at the time of this writing, only a fraction of the planned number of samples have been collected. Unlike the prototype study, Picomole is also collecting breath samples from lung cancer subjects who are post-treatment. Of

the eleven lung cancer subjects from which they have collected breath samples, nine of them are post-treatment and the remaining two are pretreatment. Due to the small number of pretreatment subjects they were removed from the analysis. One of the post-treatment lung cancer subjects also did not have breath sample measurements from every set of conditions and was also removed. This resulted in 8 post-treatment lung cancer subjects being left, all of whom are male. These 8 subjects were age and gender matched with the healthy subjects. All of the subjects are over the age of fifty. The matched pairs are shown in Table 3.4.

| Health | Age | Sex |
|---|---|---|
| Healthy | 77 | Male |
| Lung Cancer | 77 | Male |
| Healthy | 71 | Male |
| Lung Cancer | 67 | Male |
| Healthy | 57 | Male |
| Lung Cancer | 60 | Male |
| Healthy | 54 | Male |
| Lung Cancer | 61 | Male |
| Healthy | 61 | Male |
| Lung Cancer | 62 | Male |
| Healthy | 56 | Male |
| Lung Cancer | 59 | Male |
| Healthy | 66 | Male |
| Lung Cancer | 65 | Male |
| Healthy | 80 | Male |
| Lung Cancer | 77 | Male |

Table 3.4: The best optimal age and gendered matched set of subjects from the advanced data. For this set of subjects all 8 measurements over the various desorption temperatures exist

Picomole's advanced data having had the conditions of the experiment changed from the prototype data makes comparisons between the two datasets not possible. Even though some of desorption temperatures are the same, such as both datasets containing a desorption at 300°C on the same Tenax TA sorbent material, the concentration of compounds seen at that temperature depends on the preceding desorption's temperatures. The preceding temperatures on the 300°C changed from two 80°C desorptions to successive desorptions at 75°C, 150°C, 225°C. Even in the case where a 200°C desorption was performed on a single tube as in the prototype data, the amount

of sorbent material in each tube was reduced due to a change in supplier. The reduction in sorbent material has the effect of reducing the amount of signal captured. This along with the issue that the majority of lung cancer subjects in the advanced data are post-treatment poses a problem as it makes direct comparisons between the two datasets impossible.

# Chapter 4

## Analysis of Prototype Data

Plots of the $\tau$, $\tau_0$ and K concentration data against the measured wavelengths for the breath sample data that was desorbed at 200°C can be seen in figure 4.1. The curve like nature of the data motivates using FDA techniques. These plots make evident that there are missing data.

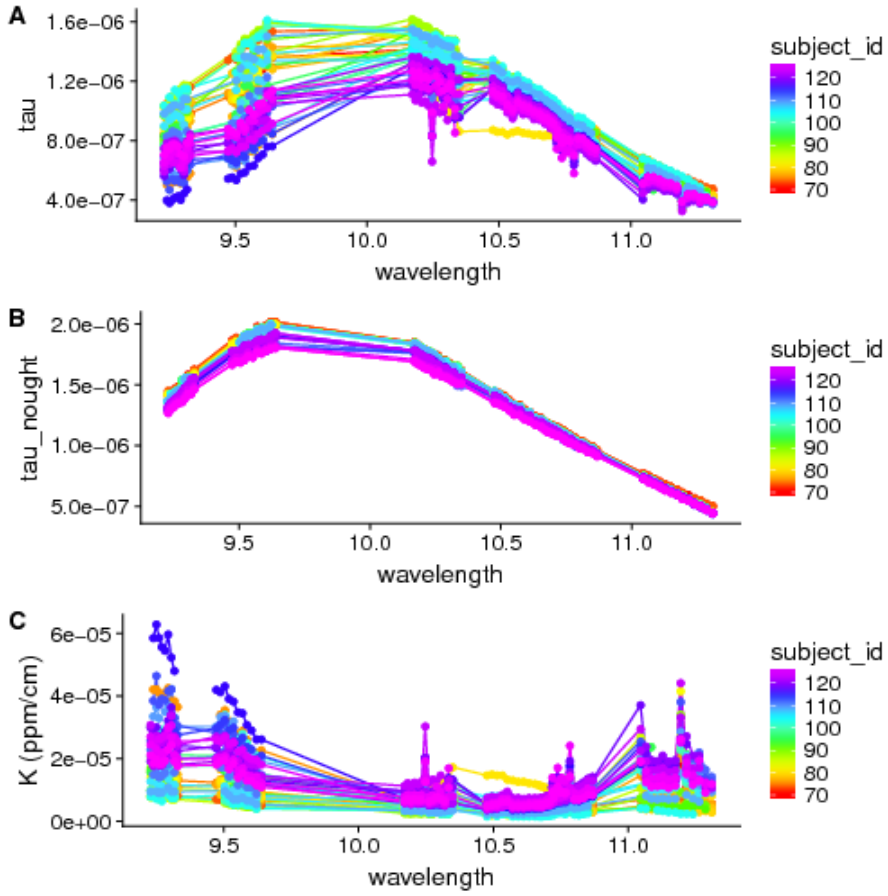


Figure 4.1: A: The $\tau$ data vs. the infrared light wavelengths measured. B: The $\tau_0$ data versus wavelengths measured. C: The K absorption data against the wavelengths. These plots refer only to measurements that were desorbed on 200°C. Each line represents a single breath or nitrogen sample.

## 4.1 Missing Data

The prototype data contains 12,876 rows with each row representing one measurement from Picomole's CRDS system on one laser wavelength. Since there are 77 laser wavelengths measured by Picomole on 175 total measurements for 47 subjects, there should have been $77 \times 175 = 13475$ rows. There are therefore 599 missing rows of data. This was confirmed with Picomole who reported that laser wavelength measurements for which $\tau$ and $\tau_0$ were missing were not included in the dataset. In addition to these missing line measurements there is also explicitly missing data in both the $\tau$ and $\tau_0$ data. Since the K absorption is calculated from both $\tau$ and $\tau_0$ if either is missing then K can not be calculated and is therefore missing.

Examining the missing data reveals that some wavelengths have a much greater number of missing observations than others. In figure 4.2 the percentage of missing observations per wavelength is reported. Certain wavelengths, particularly those on the edges of the six bands have a much higher percentage of missing observations than those in the middle. The wavelengths with the most missing observations are 9.2295$\mu$m, 9.2396$\mu$m, 9.3294$\mu$m, 9.4884$\mu$m, 9.6392$\mu$m and 10.867$\mu$m. These wavelengths have over 40% of their measurements missing.

Aggregation plots were used to examine the patterns of missing data. Aggregation plots involve plotting combinations of missing observations. The left hand plot in figure 4.3 is a bar chart displaying the counts of the number of missing observations per line for all the subjects that underwent 200°C desorption. The plot's right hand side shows the observed patterns of missing observations. Each row of the plot represents a combination of missing observations. Red boxes represent a missing observation and blue those that were observed. On the extreme right hand side is a bar chart showing the frequency with which each pattern has occurred.

This reveals a trend that when 9.2295$\mu$m, 9.2396$\mu$m, 9.3294$\mu$m, 9.4884$\mu$m, 9.6392$\mu$m and 10.867$\mu$m(also known as lines 4, 5, 13, 16, 25, 81 and 97) are missing they are usually all missing together. The same pattern can be seen in the other sets of measurements at the desorption temperatures.

Aggregation plots were also created to compare the patterns of missing observations between subjects that were healthy versus unhealthy. These can be seen in figure 4.4. Interestingly the pattern of having lines 4, 5, 13, 16, 25, 81 and 97 all

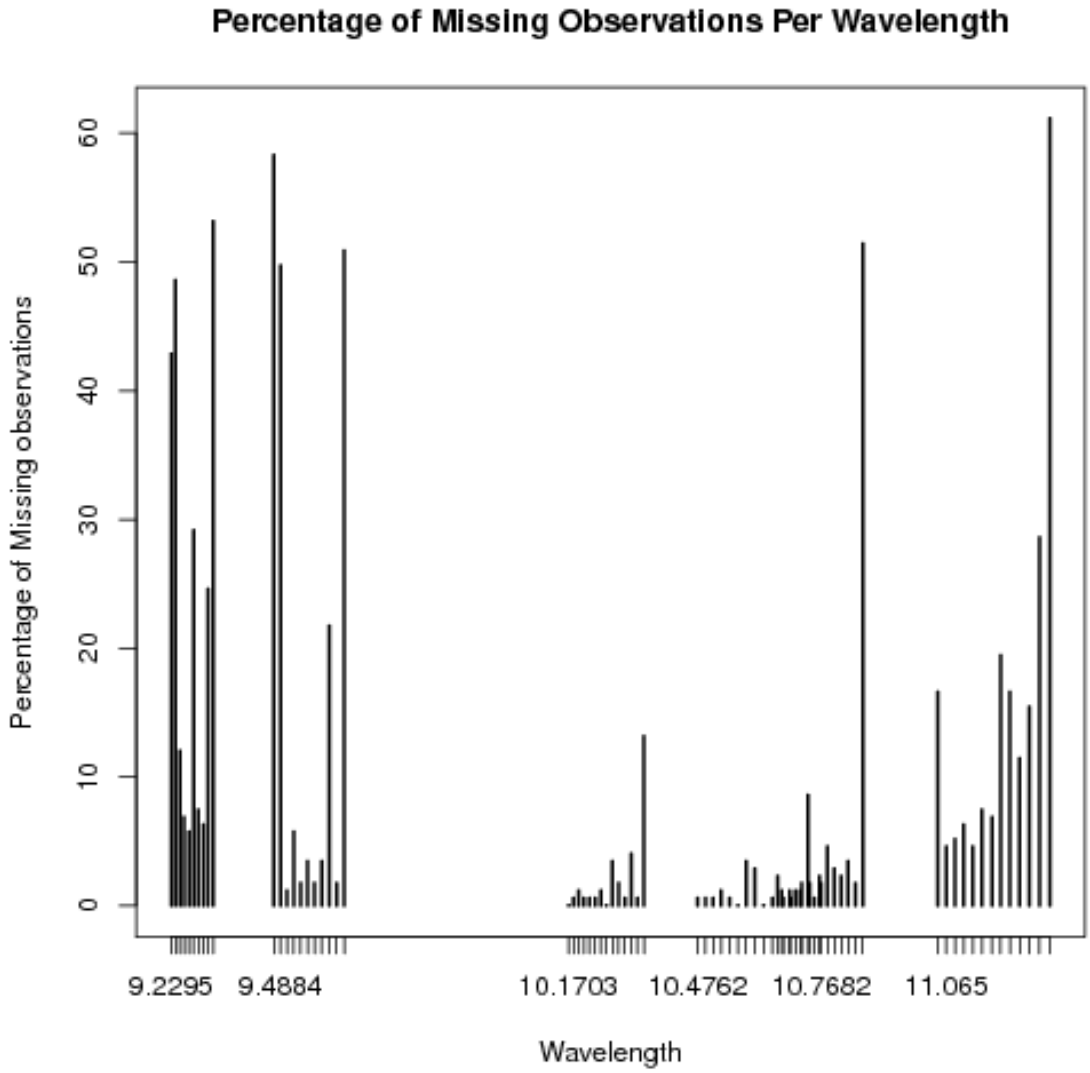


Figure 4.2: The percentage of missing observations per wavelength. Some wavelengths have a much greater percentage of missing observations than others. Wavelengths on the edges of each of the 6 wavelength bands tend to have much higher percentages of missing observations.

missing at once are much more prevalent in breath samples of subjects with lung cancer or COPD. Initially it was suggested by Picomole staff that unhealthy subjects may be "blowing out" the measurements of those wavelengths, that is giving a much higher concentration than the CRDS system can register.

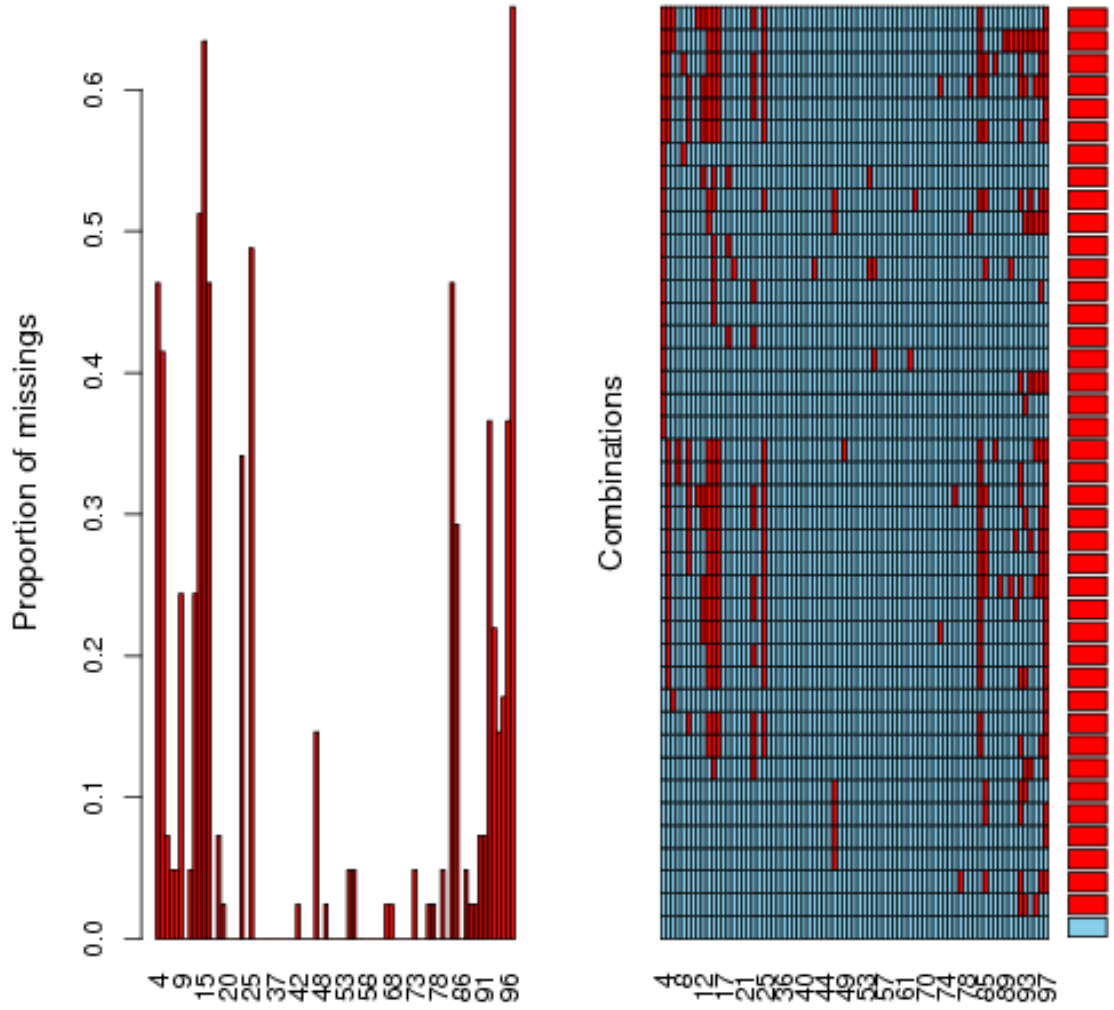


Figure 4.3: Aggregation Plots of the K absorption data for all of the subjects on the 200°C desorption. The plot on the right side represents the combinations of missing observations that occurred. There is a clear pattern of missing data involving lines 4 or 5 and 13, 16, 25 and 87 are usually all missing together.

However when looking at only $\tau_0$ patterns the same lines are missing. The aggregation plot for $\tau_0$ is shown in figure 4.5. This suggests that the missing line issues likely stem from problems with Picomole's CRDS system and not from differences between healthy and unhealthy subjects. This led to a request to Picomole for information on what dates the subjects' samples were processed. When this information was added

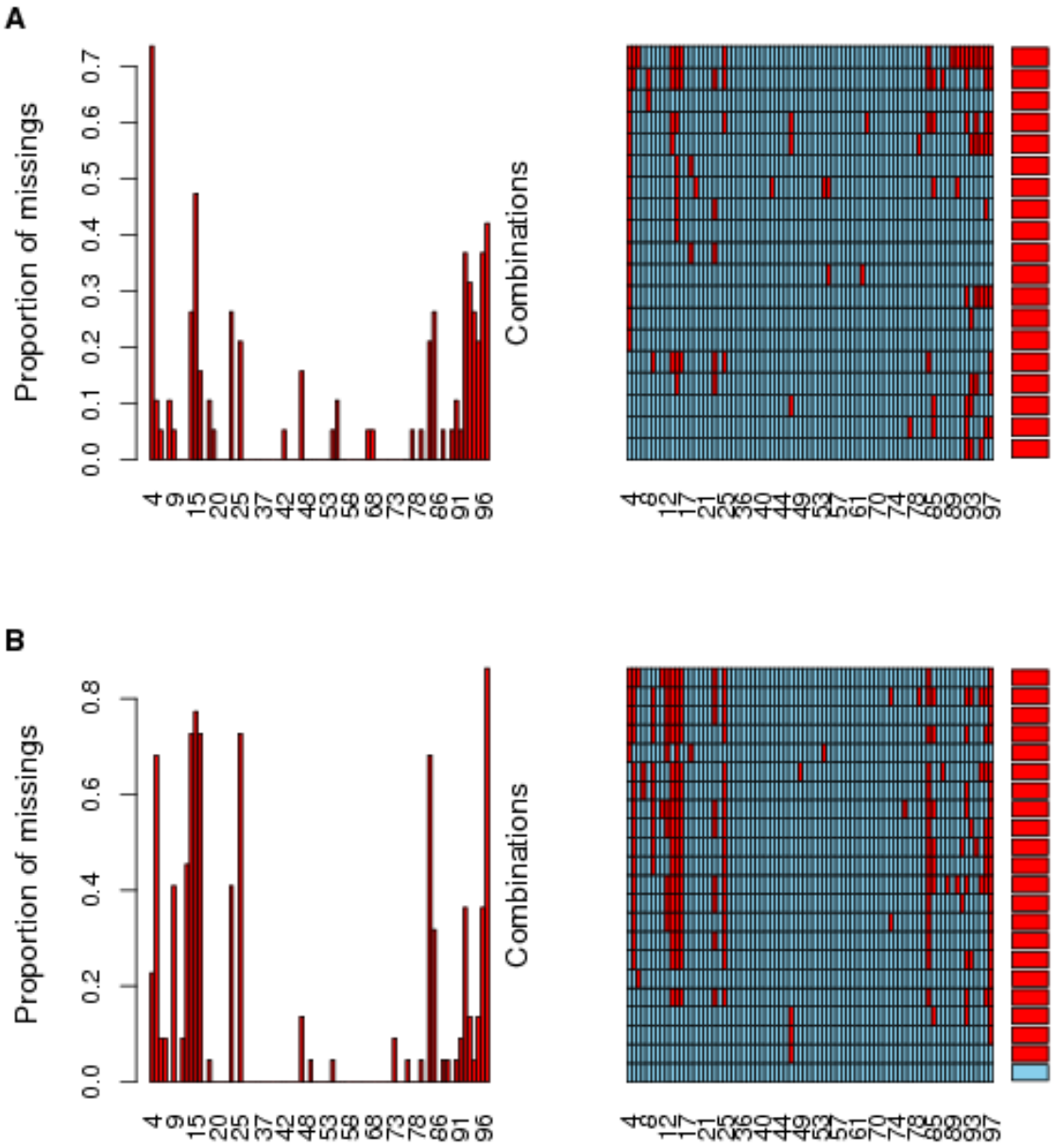


Figure 4.4: Aggregation plots for the K absorption data separated by health status. A: Healthy subjects. B: Unhealthy subjects.

and the samples were ordered by date it became clear that there was an issue with Picomole's system in March and April of 2012. The pattern of missing wavelengths is only observed within that period. Unfortunately most of the lung cancer subjects

were also measured during that period giving the impression that the missingness was somehow related to the subjects' health status. There is no further K absorption data available during this period to determine whether there was an impact on the measurements outside of the missing values. The best we can do is to examine what affect this period of time had on the $\tau_0$ data. To do so, two sample testing for the period where the pattern of missing observations occurred versus where it did not occur on the $\tau_0$ data as is described in the next section.

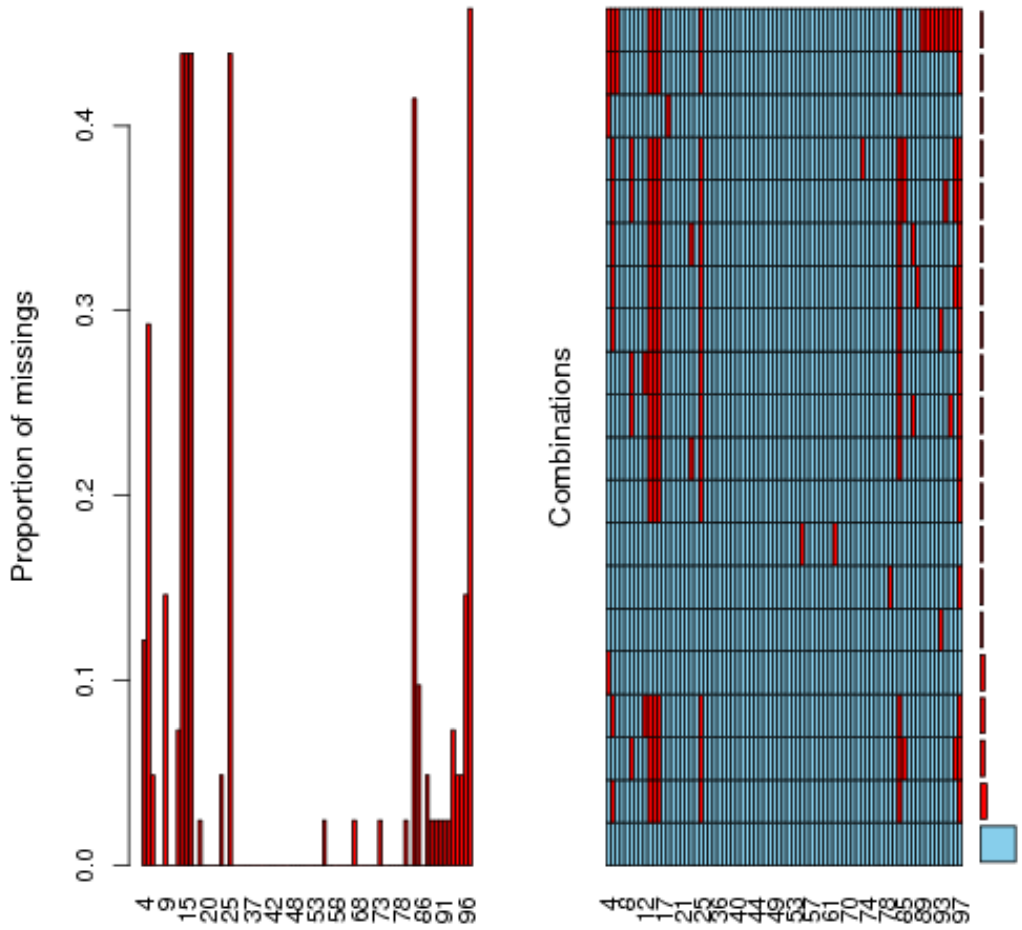


Figure 4.5: Aggregation plot for the prototype $\tau_0$ measurements desorbed at 200°C. The combination of lines 4, 5, 13, 16, 25, 81 and 97 all being missing at the same time is still visible. Since the $\tau_0$ measurements are from samples of nitrogen run before each subject's measurement they should be independent of any subject's health status.

| Measurement Time | Health Status | Pattern of Missing Lines |
|---|---|---|
| 2012-02-18 17:12:50 | Healthy | FALSE |
| 2012-02-20 16:05:38 | Healthy | FALSE |
| 2012-02-23 14:41:32 | Healthy | FALSE |
| 2012-02-27 12:45:26 | Healthy | FALSE |
| 2012-02-28 10:15:00 | Healthy | FALSE |
| 2012-02-28 14:43:58 | Healthy | FALSE |
| 2012-03-02 10:35:01 | Healthy | FALSE |
| 2012-03-01 07:22:46 | COPD | FALSE |
| 2012-03-01 15:42:06 | Healthy | FALSE |
| 2012-03-03 14:42:42 | Healthy | FALSE |
| 2012-03-09 12:28:10 | Lung cancer | FALSE |
| 2012-03-12 10:08:13 | Lung cancer | TRUE |
| 2012-03-13 14:12:11 | Healthy | TRUE |
| 2012-03-14 12:16:48 | Lung cancer | TRUE |
| 2012-03-18 17:28:23 | Lung cancer | TRUE |
| 2012-03-19 13:01:04 | COPD | TRUE |
| 2012-03-23 14:34:22 | Lung cancer | TRUE |
| 2012-03-26 14:41:00 | Healthy | TRUE |
| 2012-03-26 10:08:15 | Lung cancer | TRUE |
| 2012-03-29 14:22:53 | Lung cancer | TRUE |
| 2012-03-30 13:23:21 | Lung cancer | TRUE |
| 2012-03-30 21:52:45 | Lung cancer | TRUE |
| 2012-04-02 15:21:54 | COPD | FALSE |
| 2012-04-03 16:12:53 | Lung cancer | TRUE |
| 2012-04-04 14:52:21 | COPD | TRUE |
| 2012-04-05 09:37:00 | Lung cancer | TRUE |
| 2012-04-12 11:12:57 | Lung cancer | TRUE |
| 2012-04-13 13:00:05 | COPD | TRUE |
| 2012-04-30 13:12:58 | Healthy | FALSE |
| 2012-05-02 14:03:59 | Healthy | FALSE |
| 2012-05-03 14:14:01 | Healthy | FALSE |
| 2012-05-09 16:01:10 | Healthy | FALSE |
| 2012-05-07 14:51:18 | Healthy | FALSE |
| 2012-05-08 12:16:43 | Healthy | FALSE |
| 2012-05-10 14:01:42 | Healthy | FALSE |
| 2012-05-16 14:20:54 | Healthy | FALSE |
| 2012-05-17 15:03:46 | Lung cancer | FALSE |
| 2012-05-22 15:38:18 | Lung cancer | FALSE |
| 2012-06-13 16:13:39 | COPD | FALSE |
| 2012-06-20 14:58:06 | Lung cancer | FALSE |
| 2012-06-20 14:17:14 | Lung cancer | FALSE |

Table 4.1: The order in which subjects' samples were processed from the 200°C desorb temperature measurements. Note the clustering of Lung Cancer subjects between March 9th and April 12th. This pattern is similar for the other desorb temperatures.

## 4.2 Two Sample Testing for $\tau_0$ Data

There is no suitable K absorption data that can be used to determine whether there was an impact on the distribution of the data between the period of March and April 2012 when the combination of missing lines occurred. Instead the $\tau_0$ data was examined to see if there was any change in the measurement of the data besides the missing values during that period. $\tau_0$ was used because of the fact that it is independent of any subjects' health status. We let $\boldsymbol{X_1}$ be the set of $\tau_0$ curves that were measured between March 9th 2012 and April 12th 2012 and $\boldsymbol{X_2}$ be the set of remaining $\tau_0$ curves that were measured outside of that time period.

First $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ were compared using the Functional Anderson-Darling test as described in section 2.2.3. This test is well suited to this dataset due to it's high number of missing measurements. It is a test of

$$H_0 : \boldsymbol{X_1} =^d \boldsymbol{X_2} \quad \text{versus} \quad H_A : \boldsymbol{X_1} \neq^d \boldsymbol{X_2}. \tag{4.1}$$

As noted by Equation 2.30 this boils down to a test between the FPCs that explain a certain percentage of the variance of the two samples using the Anderson-Darling test. As recommended by Pomann et al. a threshold of 95% of variation explained by the FPCs was used. The FPCs were calculated using `fdapace`, an implementation of the PACE method of FPCA for the open source software package `R`. Then the Anderson-Darling test was performed pairwise on the FPC using the `ad.test` function from the `R` package `kSamples`.

The results of the FAD test on the different sets of $\tau_0$ curves can be seen in Table 4.2. All of the tests for each set of desorption were found to be highly significant at the $\alpha = 0.05$ level suggesting that we reject $H_0$ and that the two sets of curves come from two different distributions.

Another two sample hypothesis method was attempted as well: The functional Schilling procedure described in section 2.2.3. The test was implemented in `R`. Since this procedure requires that the data contain no missing values imputation was required. This was done using FPCA and the Karhunen-Loève representation as discussed previously to fill in the missing data. Since for any choice of k nearest neighbors there would be over 40 factorial permutations an exact test was not conducted but instead a test with 100,000 permutations and $k = 4$ nearest neighbors.

| | Trial | FAD p-value |
|---|---|---|
| 1 | 200 | 0.00721 |
| 2 | 80 Desorb 1 | 0.00228 |
| 3 | 80 Desorb 2 | 0.00298 |
| 4 | 300 Desorb 3 | 0.00298 |

Table 4.2: P-values of the FAD test conducted to compare the distribution of $\tau_0$ data measured between March 9th 2012 and April 12 2012 and those measured outside that date range. For each set of measurements the FAD test was found to be highly significant suggesting the two sets of curves come from different distributions

A p-value of 0 was obtained for each of four trials. The expected value was 52% while the calculated test statistics were all over 90%. The outcome of the Schilling procedure also suggests that the that the two sets of curves do not come from the same distribution.

### 4.3 Limitations

As suggested by the previous sections this prototype data has some serious limitations. In addition to being a small dataset which makes using both a training and testing set less than ideal, it has a high percentage of missing values. However the issue that has the most affect on the analysis is the issue of the bunched lung cancer subjects that occurred during a period when the CRDS system operated differently. As there is no other K absorption data from that period it is not possible to conclude definitively that the differences between healthy and unhealthy subjects is due to actual differences between the subjects or if it's just an artifact of the system not operating correctly. Assuming the mis-calibration affected the $\tau$ data in the same way as the $\tau_0$ data then any inference based on the K absorption data which uses $\tau_0$ as a baseline measurement should be valid.

### 4.4 CART on Prototype Data

Classification trees using the CART method are described in section 2.1.1. Classification trees were tried first due to their speed, ability to handle missing values, and the potential to highlight any predictor variables that may be important in classifying healthy and unhealthy subjects. The classification trees were constructed using the

`rpart` package. Trees were originally constructed with the goal of being able to classify all three different health statuses: healthy, lung cancer and COPD. They were also constructed with each set of conditions having their own classification tree due to the fact that classification trees are incapable of handling repeated measures designs.

Originally the trees included the subjects' age as a predictor variable. They were also constructed on the $\tau$ data as that is all that was provided initially. Since the subjects were not age matched and most of the people with lung cancer and COPD are over the age of 60 each tree split on age. These trees are provided in figure 4.6. This is not particularly useful in helping to identity new patients with lung cancer, especially since the incidence of lung cancer in the subjects tested over sixty is over 40 percent which is much greater than the actual rate. However this did help point out how skewed the dataset is towards older subjects. These trees were also built using the default settings of `rpart` which require 20 observations to be in a node before a split occurs. Since the number of observations in each of the trials is around 40 the trees are very limited in how large they could grow. Further trees were grown with the number of observations required for a split to occur reduced from 20 to 5.

More trees were grown on the $\tau$ data but as pointed out in section 4.3 there are limitations with the $\tau$. When the K absorption data was received new trees were grown on that data using leave one out cross validation. These trees were also pruned to help compensate for overfitting. The trees in figure 4.7 are all radically different from one another. The tree grown on the data that was desorbed at 80°C the first time resulted in a one node tree that simply classifies all subjects as having lung cancer. This is of course not very helpful. The other trees all identify line IDs towards the higher end of spectrum.

Performance with CART trees varies greatly. From an accuracy rate of 44% and sensitivity of 0% for the healthy and COPD subjects on the first 80°C desorption to a high of an 87.8% accuracy on the 200°desorption. The tree with the best overall accuracy the 200°C desorption has not so great lung cancer sensitivity of 75%, the second best performing tree, the 300°desorption has a much improved but still not ideal lung cancer sensitivity of 94.74% but at the expense of the the COPD subjects, correctly classifying none of them while also incorrectly labeling them as both healthy and having lung cancer questioning its value as a preliminary identifier of

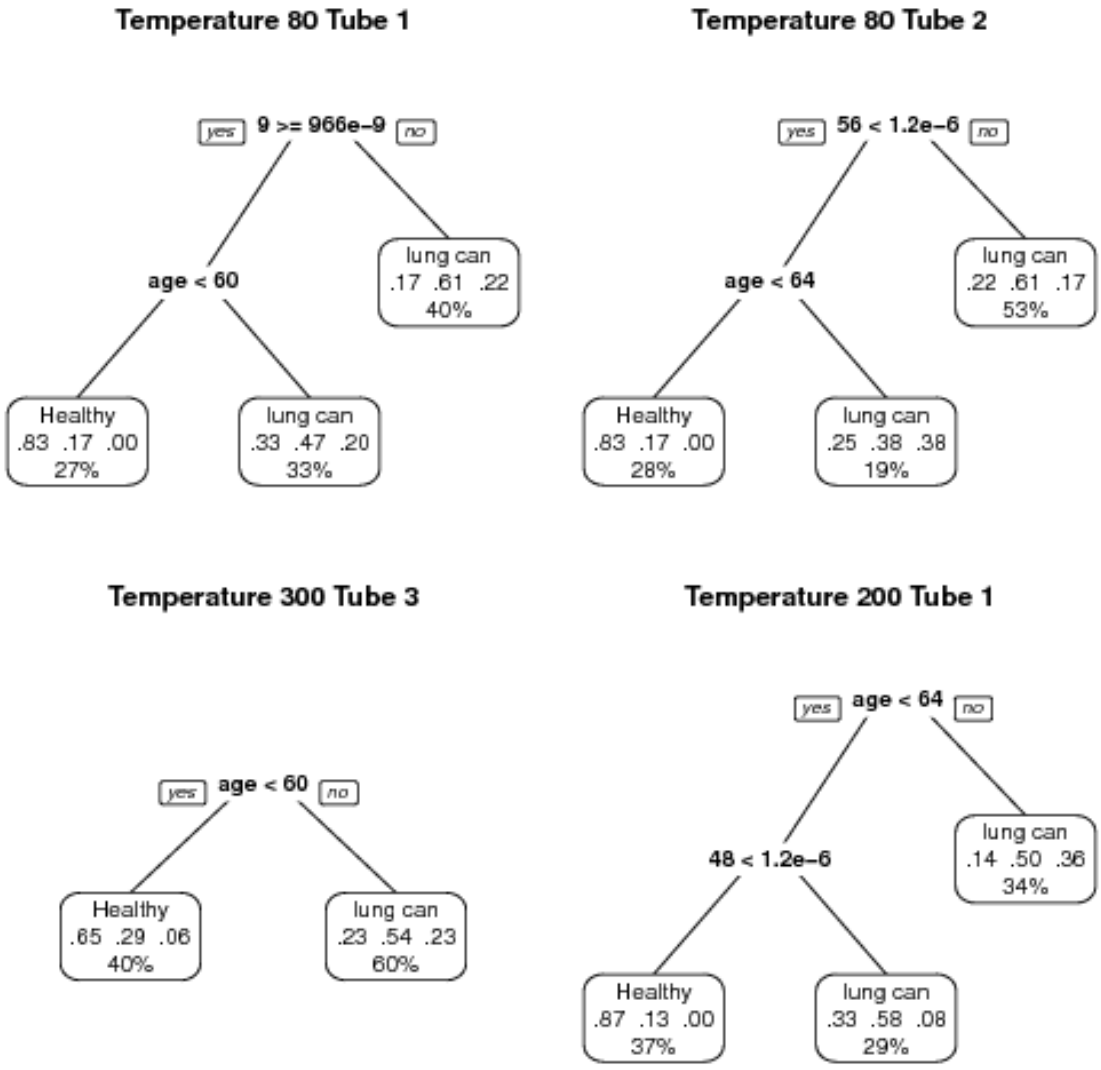


Figure 4.6: The first CART trees constructed with the age variable included. Since subjects were not age matched and the majority of lung cancer subjects are older it becomes the best variable to split on. The percentage at the bottom of a node refers to the percentage of total observations that fall in that node. The three numbers in the middle of a node refer to proportions of each class in that node in the order, healthy, lung cancer, COPD.

COPD. Table 4.3 summarizes the performance of the CART trees grown on the K concentration data. One thing to note is that since the trees grown are so small the choice in pruning either removes all nodes as in the case of first 80°C desorption or do not get pruned at all as in the case of the 200°C desorption which raises questions

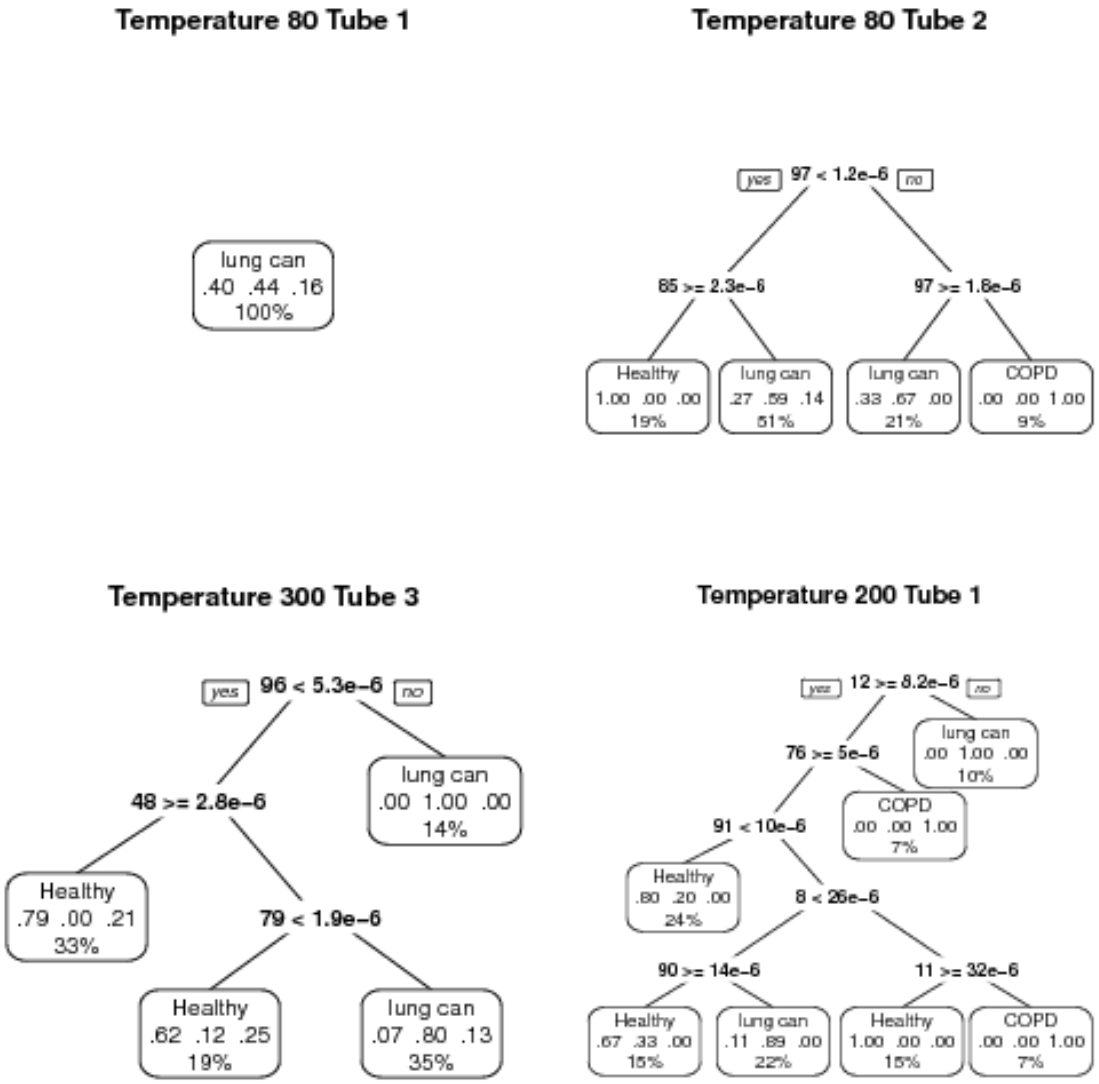


Figure 4.7: Pruned and cross validated trees grown on the K absorption data. The tree grown on the data from temp. 80 tube 1 is not very helpful, classifying all subjects as having lung cancer. The other trees do better and tend to suggest that lines higher in the higher band(lines 90 and above).

about the possibility of overfitting.

| | Accuracy(95% CI) | Sensitivity | Specificity |
|---|---|---|---|
| 80°C Tube 1 | 0.44(0.30,0.60) | H:0.0,LC:1.0,COPD:0.0 | H:1.0,LC:0.0,COPD:1.0 |
| 80°C Tube 2 | 0.72(0.56,0.85) | H:0.47,LC:1.0,COPD:0.57 | H:1.0,LC:0.5,COPD:1.0 |
| 300°C Tube 3 | 0.79(0.64,0.90) | H:0.94,LC:0.95,COPD:0.0 | H:0.77,LC:0.88,COPD:1.0 |
| 200°C Tube 1 | 0.88(0.74,0.96) | H:0.95,LC:0.75,COPD:1.0 | H:0.82,LC:0.96,COPD:1.0 |

Table 4.3: The accuracy, sensitivity and specificity for each of the second set of pruned CART trees grown. Note that H refers to healthy subjects and LC to lung cancer subjects.

## 4.5 Random Forests

Random Forests were also constructed on the prototype data. The Random Forests were grown using the `randomForest` package available in R. Despite the capability of trees grown using CART to handle missing values, `randomForest` is unable to deal with them and requires a dataset with no missing values. The data was imputed using a method which utilizes FPCA[8]. A grid search was performed in order to determine the best parameters for the number of trees to grow in each forest and the number of predictor variables that the random forest algorithm considers at each split. The number of trees tested was 50, 100 and 200 and the number of variables tried was 5, 10, 200, 30, 50 and 77(The total number of predictor variables). The dataset was split into an 80% training set and 20% testing set. This was done using the `createDataPartition` from the R package `caret`.

Ten fold cross-validation was used during the grid search on the training data to help select the parameters of the best performing random forest for each set of conditions. This was done to avoid data leakage. The set of parameters with the lowest mean OOB error rate over the ten fold cross validation were used as the final parameters to assess the performance of random forests on this data. For the 200°C desorb, the parameters with the lowest OOB error rate was $n_{tree} = 100$ trees grown in the forest with $M$ =77 predictor variables allowed to be selected from at each split, for the first time the breath sample was desorbed at 80°C $n_{tree} = 50$ and $M = 50$, for the second 80°C desorption $n_{tree} = 100$ and $M = 20$ and finally for the 300°C desorption $n_{tree} = 100$ and $M = 5$. Random forests were grown using these parameters on the training data and tested on the testing set.

Performance using the random forests is poor. Even on the training data the

| | Healthy | Lung cancer | COPD | Class Error |
|---|---|---|---|---|
| Healthy | 5 | 7 | 2 | 0.6429 |
| Lung cancer | 7 | 7 | 2 | 0.5625 |
| COPD | 3 | 3 | 0 | 1.0000 |

Table 4.4: The confusion matrix for second desorption at 80°C for the random forest's performance on training data. Classification performance is very poor, with rates worse than random guessing and is unable to correctly deal with COPD subjects.

| | Healthy | Lung cancer | COPD | Class Error |
|---|---|---|---|---|
| Healthy | 1 | 1 | 1 | 0.6666 |
| Lung cancer | 0 | 2 | 1 | 0.3333 |
| COPD | 0 | 1 | 0 | 1.0000 |

Table 4.5: Confusion Matrix for the second desorption at 80°C for the random forest grown on the training data and tested on the testing set. Performance is as poor as on the training data.

random forests struggle at classifying COPD subjects correctly. Only one of the random forests grown, the one on the 200°C desorption, classified a single COPD correctly on the training data. As an example for the random forests performance on this dataset, confusion matrices for the random forest grown using the second desorption at 80°C on the training and testing data can be seen in Tables 4.4 and 4.5 respectively. As is shown by the two confusion matrices, the random forest classifier is unable to classify COPD subjects at all and is not very effective with the remaining two classes either.

Random forests were tried one more time with the COPD subjects removed. The small number of COPD subjects likely has an impact and just end up confusing the classifier on the other two classes. The random forests were grown in the same fashion as before just with the COPD subjects removed. The data were once again separated into a testing a training dataset and then the parameters used were determined by cross validation on the training dataset and then verified on the test dataset using those parameters. The best parameters for the random forest with no COPD subjects done on the 200°C desorption was $n_{tree} = 50$, $M = 50$, for the first desorption at 80°C $n_{tree} = 200$, $M = 5$, the second desorption at 80°C $n_{tree} = 200$, $M = 77$ and the desorption at 300°C $n_{tree} = 50$ and $M = 77$.

An example of the performance of the random forests with COPD subjects confusion matrices for the random forest grown on the second 80°C desorb are shown

in Tables 4.6 and 4.7. There is no real improvement in performance using random forests with the COPD subjects removed. Performance on the training data appears to be the random forest guessing between the two classes. The performance on the testing dataset appears better but with such a small dataset that result could simply be pure chance.

Despite good performance of these classifiers on other types of spectrometry data[6][29] the tree based classifiers are more of a mixed bag on Picomole's prototype dataset. The fact that the random forests performed so poorly while the CART trees performed well raises questions about their performance as well as random forests were designed to deal with the shortcomings of CART trees, in particular overfitting. Perhaps with a larger dataset the tree-based methods could be more effective.

| | Healthy | Lung cancer | Class error |
|---|---|---|---|
| Healthy | 6 | 8 | 0.5714 |
| Lung cancer | 7 | 9 | 0.4375 |

Table 4.6: The confusion matrix for the second desorption at 80°C with the COPD subjects removed for the random forest applied to the training data. Removing the COPD subjects did not have a positive impact on performance.

| | Healthy | Lung cancer | Class error |
|---|---|---|---|
| Healthy | 2 | 1 | 0.3333 |
| Lung cancer | 0 | 3 | 0.0000 |

Table 4.7: The confusion matrix for the second desorption at 80°C with the COPD subjects removed for the random forest applied to the testing data. It initially appears as though the random forest is capable of correctly classifying lung cancer subjects, although performance on the training data and the small sample size makes this unlikely.

## 4.6 FLDA on Prototype Data

FLDA was one of the first techniques applied to the prototype data because of it's ability to handle data that is sparse or has a high percentage of missing values as is the case here. For the full details on the theory behind FLDA refer to section 2.2.2. The version of FLDA used here is G. James' original FLDA implementation in the `S-plus` programming language available from his research web-page[16] with a very minor tweak to enable it to run on `R`.

With FLDA all the prototype data was used at the same time. All of the measurements were used on a grid going from 1 to 308. The first 77 points represented the measurements from the 200°C desorption, 78 to 155 were the measurements from the first 80°C desorption, 156 to 223 were the measurements from the second 80°C desorption and 224 to 308 are the measurements from the 300°C desorption. As recommended by James et al. several dimensions of $q$ were tried and the dimension with the highest likelihood during cross validation was selected. The dimensions of $q$ tried were 5 (the default), 10, 25, 50, 100, 150, 200 and 300. Ten fold cross validation was used. The dimension of q with the highest likelihood over the CV was found to be 300. Next the rank constraint p on the $\gamma$'s was selected. This was done by finding the value of p that minimized classifier error rate over 10 fold cross validation. The values of p tried were 1 (the default), 5, 10, 20, 30 and 40. The maximum choice of p is limited by the number of subjects, as otherwise a non-singular matrix occurs. The value of p that was found to result in the best performance on the CV data was 30.

Originally all the data was used to build the classification rule using FLDA as it was believed at the time that the advanced data would be comparable to the prototype data and could be used as a validation set. Performance of FLDA on all the data is quite good. While it is incapable of separating subjects with COPD or lung cancer it is able to separate between healthy and sick subjects perfectly. Plots of the $\hat{\alpha}$ with points colored to correspond to the predicted and actual classes of subjects are seen in Figure 4.8. Considering that FLDA is a linear classifier it is no surprise that it is only able to correctly separate the classes into two groups.

Unfortunately the performance of FLDA when it was applied to the prototype dataset using 10-fold cross validation isn't quite as good. The FLDA classifier here is no longer able to separate between healthy and sick subjects as seen in Figure 4.9. Performance of the FLDA classifier reduced to 70.21% when COPD subjects are grouped together with the those that have lung cancer. FLDA was also tried with the COPD subjects removed using the same 10 fold cross-validation as used previously. In addition the same parameters were used as beforehand with $q = 300$ and p= 30. With the COPD subjects removed the FLDA total classification is 77.5%. It incorrectly classfies 5 of the healthy subjects as having lung cancer and 4 of the lung cancer subjects as being healthy out of 40 total subjects without COPD. This

| | Accuracy(95% CI) | Sensitivity | Specificity |
|---|---|---|---|
| FLDA COPD included | 0.49(0.34,0.64) | H:0.79,LC:0.22,COPD:0.0 | H:0.61,LC:0.93,COPD:0.69 |
| FLDA COPD removed | 0.78(0.62,0.8916) | 0.80 | 0.75 |

Table 4.8: The accuracy (and corresponding 95% confidence interval), sensitivity and specificity for the FLDA models constructed using 10 fold cross validation. Removing COPD subjects improves performance. Once again H refers to Healthy and LC to lung cancer.

leads to a sensitivity rate of 80% and a specificity rate of 75%. The plot showing the seperation between the two classes is Figure 4.10. When trying to classify all three groups only 48.94% are correctly classfied and it classifed most lung cancer subject's as having COPD. Table 4.8 summarizes the accuracy, sensitivity and specificity of the FLDA models used.

The FLDA classifier shows performance that suggests it is possible to predict a subjects health status based on their breath samples but not with the 100% sensitivity that is desired in a preliminary medical test.

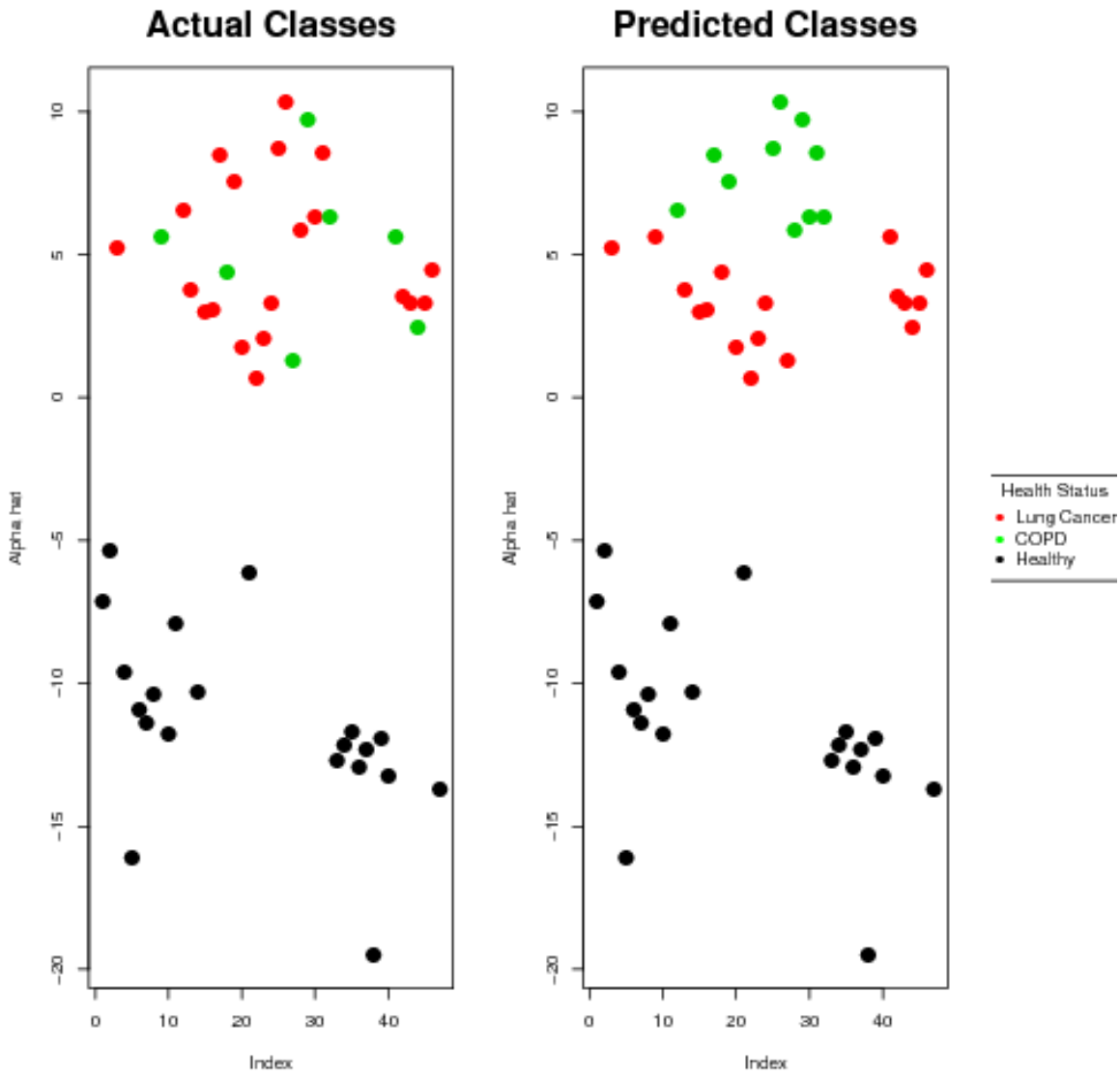


Figure 4.8: Plots showing the separation between classes on all the data generated by FLDA. It's unable to distinguish between COPD and lung cancer patients but can separate healthy and sick subjects.

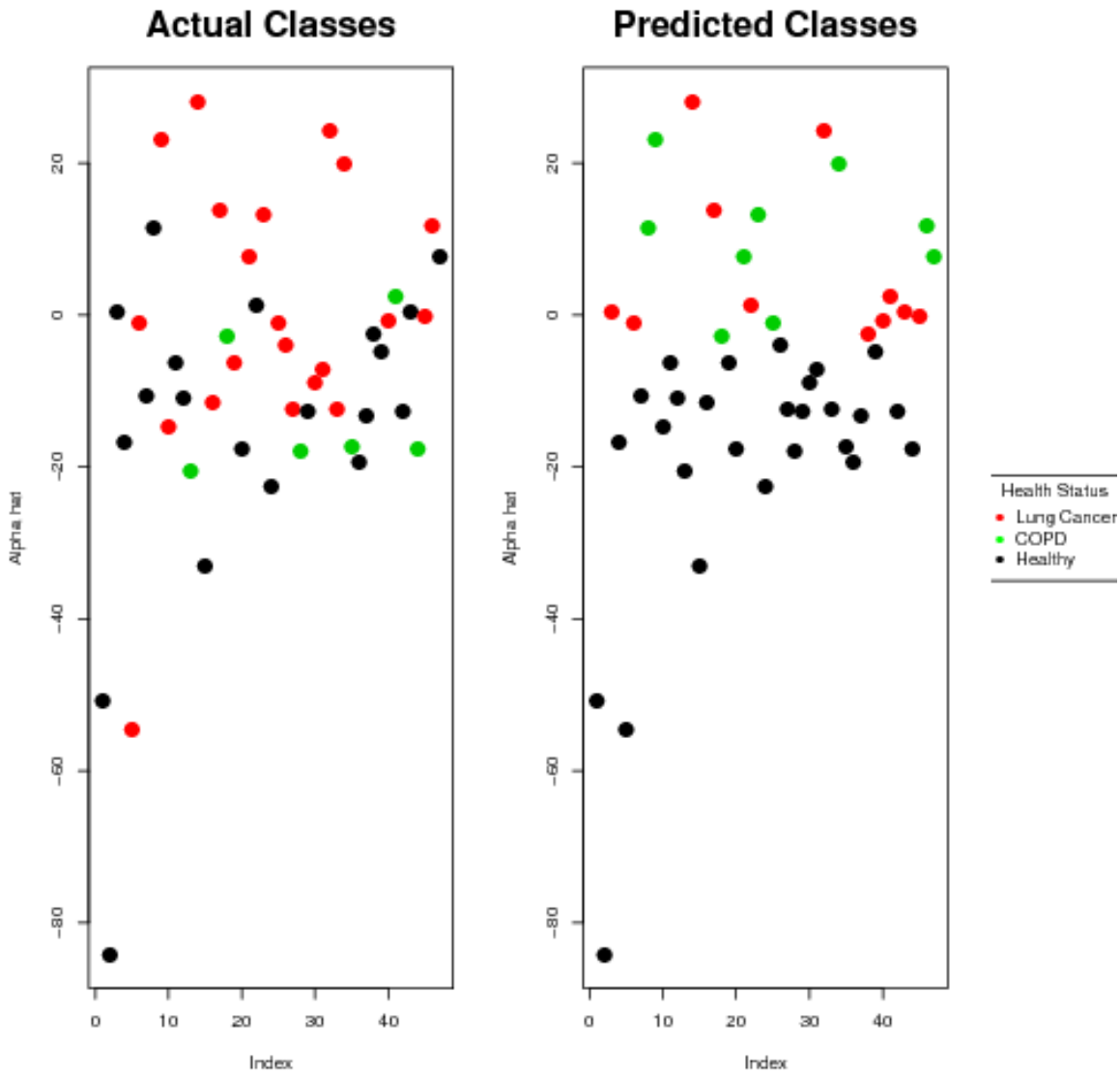


Figure 4.9: The FLDA separation between classes on the prototype data with 10-fold cross validation. The FLDA classifier is no longer able to reliably separate healthy and sick subjects suggesting the previous results were likely the result of overfitting.

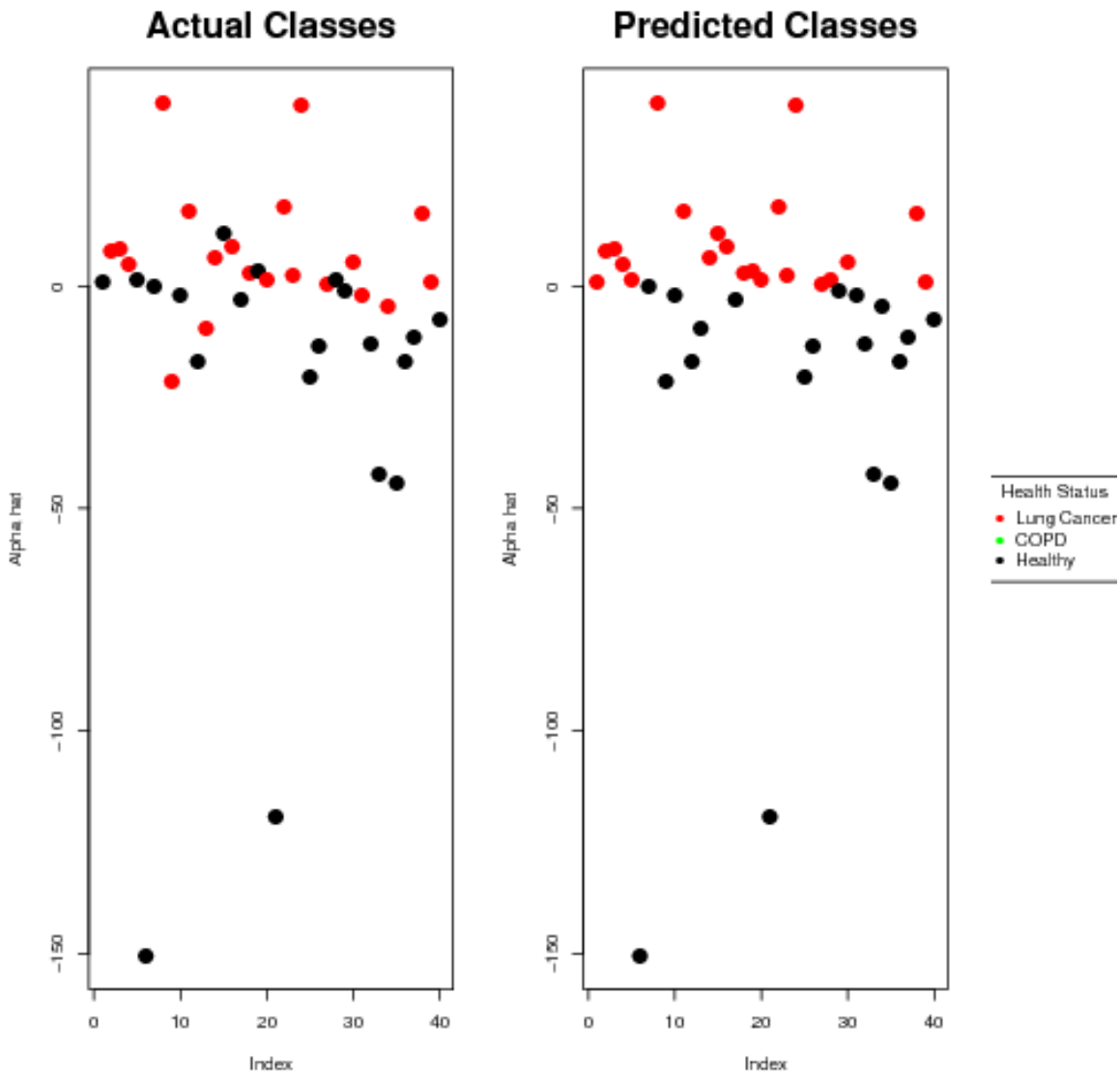


Figure 4.10: The FLDA separation after the COPD subjects have been removed.

## 4.7 FPCA and Related Techniques on Prototype Data

Effort was also made to try and glean information about the prototype data by performing a FPCA on the dataset. In addition an FPCA clustering method was applied. As with the imputation done by FPCA, the FPCA applied to the prototype data were performed by the PACE method. The prototype data is considered to be dense functional data with missing values by `fdapace`.

FPCA was applied to each of the desorbs separately. Diagnostic plots for each of the FPCA conducted were examined. An example of the diagnostic plots for the 200°C condition are in Figure 4.11. The 200°C and 300°C conditions have over 80% of their variation explained by the first FPC, the other two 80°C conditions have over 67% of their variation explained by the first FPC. All of the different sets of conditions have 95% of their variation explained by the first four FPCs. The FPCs do not appear to have any obvious interpretation in seperating the three groups of subjects.

FPCA clustering is an unsupervised learning technique which tries to place the measurements into groups using the FPC scores. Originally the FPCA clustering was performed with the `Rmixmod` algorithm trying to cluster the data into three different clusters, one for each health status. Then, as with the random forest, COPD subjects were removed to see if that had a positive impact on correctly determining a subject's health status.

Performance of FPCA clustering on this prototype dataset is poor. The best correct classification rate on the FPCA clustering with three groups is 53.65% on the desorbtion that was performed at 200°C. Removing the COPD subjects does not greatly improve things. With the COPD subjects removed the best classification performance is 63.88%.

FPCA while doing a better job at imputing the data than the other methods that were attempted, did not reveal anything particularly illuminating on the relationship between a subject's health status and their measurement.

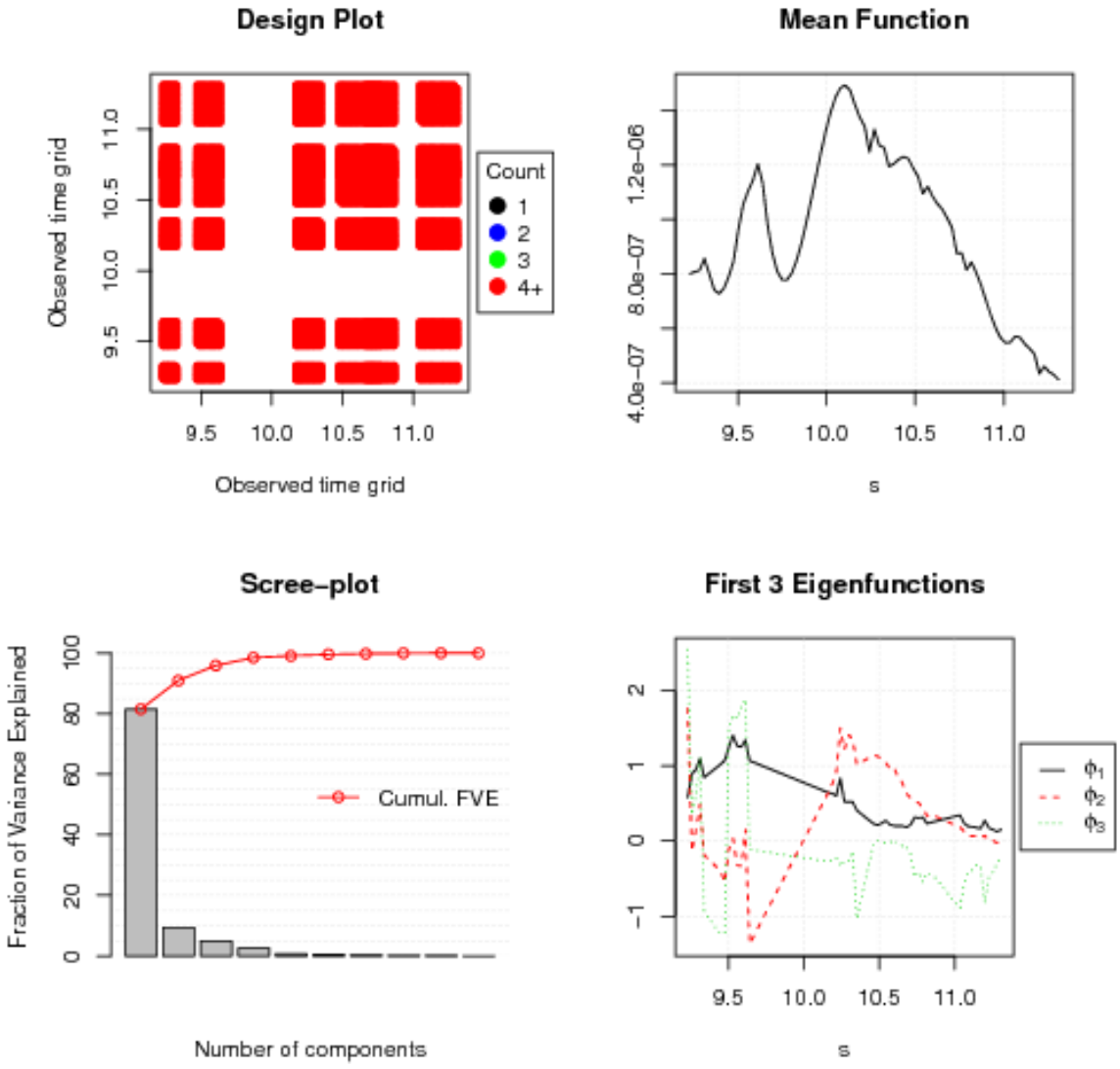


Figure 4.11: Diagnostic plots for the FPCA performed on desorption performed at 200°C. The top left plot show spacing and density of the grid used. This data is considered dense with missing data. The top right plot is the path of the mean curve. The lower left plot shows the fraction of variance explained by each FPC, the first principal component in this case explains over 80% of the variance. 95% of the variance is explained by the first four FPC. The final plot, the bottom right shows the path of the first three eigenfuction curves.

| | Accuracy(95% CI) | Sensitivity | Specificity |
|---|---|---|---|
| First 80 C Desorption | 0.44(0.30,0.60) | H:0.11,LC:0.85,COPD:0.14 | H:0.96,LC:0.20,COPD:0.89 |
| Second 80 C Desorption | 0.51(0.35,0.67) | H:0.41,LC:0.68,COPD:0.29 | H:0.69,LC:0.58,COPD:0.92 |
| 300 C Desorption | 0.49(0.33,0.65) | H:0.35,LC:0.79,COPD:0.0 | H:0.85,LC:0.42,COPD:0.89 |
| 200 C Desorption | 0.44(0.28,0.60) | H:0.47,LC:0.50,COPD:0.17 | H:0.73,LC:0.60,COPD:0.80 |

Table 4.9: The accuracy, sensitivity and specificity for each of the desorptions for FPCA clustering with all three classes included. Overall performance is not very good.

| | Accuracy(95% CI) | Sensitivity | Specificity |
|---|---|---|---|
| First 80 C Desorption | 0.55(0.38,0.71) | 0.11 | 0.95 |
| Second 80 C Desorption | 0.64(0.46,0.79) | 0.59 | 0.68 |
| 300 C Desorption | 0.53(0.35,0.70) | 0.94 | 0.16 |
| 200 C Desorption | 0.63(0.45,0.79) | 0.47 | 0.81 |

Table 4.10: The accuracy, sensitivity and specificity for each of the desorptions for FPCA clustering with the COPD subjects removed. The best performing class assignments were used. Performance is not any better than random guessing.

### 4.8 $DD^G$ Plots on Prototype Data

The final technique applied to the prototype data was the $DD^G$ plot classifiers as outlined in section 2.3. Since current measures of depth available for functional data requires that there be no missing values the data was imputed using the same FPCA method used for the random forests (as described in section 4.5).

With the imputed data, $DD^G$ plots were generated using the `fda.usc` `R` package which was developed by the creators of the $DD^G$ plot. The default FM depth was used. For the first set of plots the subjects with COPD were binned with the lung cancer subjects, making G = 2. $DD^G$ plots are easier to interpret when the dimension is two, as it is easier to see the classification rule. Four different classifiers were tried on the $DD^G$ plots, LDA, QDA, k-nearest neighbors (kNN) and a non-parametric kernel method (NP).

The $DD^G$ plots for the prototype data trial conducted at 200°C are displayed in Figure 4.12. The LDA classifier applied to the plot is able to correctly classify 47% of healthy subjects and 73% of sick subjects with a total correct classification rate of 61%. Based on the placement of points in the $DD^G$ plot it is pretty clear that a linear classifier was never going to do a satisfactory job. The majority of the sick

patients appear to be nestled between two separate groups of healthy subjects. The plot that used the QDA classifier performed better than LDA as it was able to better capture the two groups. QDA was able to correctly classify 81.81% of the sick subjects correctly and 57.89% of the healthy subjects correctly for an overall total of 70.73%.

The next two classifiers applied to the $DD^G$ plots are able to have arbitrary decision boundaries and are not restricted to polynomials. This allows for better classification. When looking at the the $DD^G$ plots with the kNN classifier applied for the 200°C desorption performance is again improved over the LDA and QDA methods. The kNN classifier finds three different groups, one going from the bottom left to top right diagonally classifying subjects falling into that group as sick and the other two flanking it classifying them as healthy. The kNN $DD^G$ plot classifier is able to correctly classify 70.58% of the healthy subjects and 84.62% of the sick subjects. This gives a a total rate of correct classification of 79.07%, an almost 9% improvement over the QDA method. The non-parametric kernel method has even better performance than the kNN classifier, however some of the groups are specific to individual points which raises the question of whether the non-parametric kernel method is overfitting the data or not.

The $DD^G$ plot classifiers were also tried with the all three classes. Unsurprisingly, performance of the $DD^G$ plot using the LDA classifier is worse. The single line of separation does not allow for any real way to classify three groups. Figure 4.13 has the $DD^G$ for the 200°C set of conditions with the LDA classifier used. The plots are harder to interpret with all three classes included as the groups are compared pairwise with only shaded shapes as indication of correct classification or not unlike the easy to distinguish two class case. Looking at the plot reveals it is much better at correctly classifying healthy subjects than those with COPD or lung cancer.

Using QDA as the classifier on the three class $DD^G$ plot results in similar performance to the two group case using QDA. An example of QDA applied to the 200°C breath samples is seen in Figure 4.14.

As with the two class case, the best performing classifiers on the $DD^G$ plots with all three classes included are still the kNN and non-parametric kernel method classifiers. They only see small decreases in performance compared to when COPD subjects were binned with lung cancer subjects. The plots using the kNN and non-parametric

| | Accuracy(95% CI) | Sensitivity | Specificity |
|---|---|---|---|
| First 80 C Desorption | 0.73(0.58,0.85) | 0.5 | 0.89 |
| Second 80 C Desorption | 0.72(0.56,0.85) | 0.53 | 0.85 |
| 300 C Desorption | 0.65(0.49,0.79) | 0.41 | 0.81 |
| 200 C Desorption | 0.61(0.45,0.76) | 0.47 | 0.73 |

Table 4.11: The accuracy, sensitivity and specificity for the DD plot using LDA as the classifier and FM depth with the COPD subjects binned with the lung cancer subjects.

| | Accuracy(95% CI) | Sensitivity | Specificity |
|---|---|---|---|
| First 80 C Desorption | 0.76(0.60,0.87) | 0.56 | 0.89 |
| Second 80 C Desorption | 0.67(0.51,0.81) | 0.41 | 0.85 |
| 300 C Desorption | 0.74(0.59,0.86) | 0.53 | 0.88 |
| 200 C Desorption | 0.71(0.54,0.84) | 0.58 | 0.81 |

Table 4.12: The accuracy, sensitivity and specificity for the DD plot using QDA as the classifier and FM depth with the COPD subjects binned with the lung cancer subjects.

kernel method are in Figures 4.15 and 4.16 respectively.

Tables 4.11 through 4.14 summarize the accuracy, sensitivity and specificity for the DD plots where the COPD subjects have been grouped with the lung cancer subjects for the four different classifiers used, LDA, QDA, K-nearest neighbors and non-parametric kernel method. Similarly Tables 4.15 through 4.18 contain the same information for the $DD^G$ plots where all three classes were used.

| | Accuracy(95% CI) | Sensitivity | Specificity |
|---|---|---|---|
| First 80 C Desorption | 0.82(0.68,0.92) | 0.72 | 0.89 |
| Second 80 C Desorption | 0.79(0.64,0.90) | 0.71 | 0.85 |
| 300 C Desorption | 0.77(0.61,0.88) | 0.65 | 0.85 |
| 200 C Desorption | 0.80(0.54,0.84) | 0.74 | 0.86 |

Table 4.13: The accuracy, sensitivity and specificity for the DD plot using K nearest neighbors as the classifier and FM depth with the COPD subjects binned with the lung cancer subjects.

| | Accuracy(95% CI) | Sensitivity | Specificity |
|---|---|---|---|
| First 80 C Desorption | 0.96(0.85,0.99) | 0.94 | 0.96 |
| Second 80 C Desorption | 0.93(0.81,0.99) | 0.94 | 0.92 |
| 300 C Desorption | 0.91(0.78,0.97) | 0.82 | 0.96 |
| 200 C Desorption | 0.95(0.83,0.99) | 0.89 | 1.00 |

Table 4.14: The accuracy, sensitivity and specificity for the DD plot using the non-parametric kernel method as the classifier and FM depth with the COPD subjects binned with lung cancer subjects.

| | Accuracy(95% CI) | Sensitivity | Specificity |
|---|---|---|---|
| First 80 C Desorption | 0.56(0.40,0.70) | H:0.61,LC:0.65,COPD:0.14 | H:0.70,LC:0.52,COPD:1.00 |
| Second 80 C Desorption | 0.63(0.47,0.77) | H:0.71,LC:0.74,COPD:0.14 | H:0.73,LC:0.63,COPD:1.00 |
| 300 C Desorption | 0.58(0.42,0.73) | H:0.65,LC:0.74,COPD:0.00 | H:0.73,LC:0.54,COPD:1.00 |
| 200 C Desorption | 0.59(0.42,0.74) | H:0.84,LC:0.44,COPD:0.17 | H:0.50,LC:0.76,COPD:1.00 |

Table 4.15: The accuracy,sensitivity and specificity for the $DD^G$ plot using the LDA classifier and FM depth on all three classes.

| | Accuracy(95% CI) | Sensitivity | Specificity |
|---|---|---|---|
| First 80 C Desorption | 0.80(0.65,0.90) | H:0.78,LC:0.75,COPD:1.00 | H:0.85,LC:0.84,COPD:0.97 |
| Second 80 C Desorption | 0.77(0.61,0.88) | H:0.65,LC:0.79,COPD:1.00 | H:0.85,LC:0.79,COPD:0.97 |
| 300 C Desorption | 0.67(0.51,0.81) | H:0.76,LC:0.74,COPD:0.29 | H:0.81,LC:0.63,COPD:1.00 |
| 200 C Desorption | 0.76(0.60,0.88) | H:0.84,LC:0.63,COPD:0.83 | H:0.82,LC:0.84,COPD:0.94 |

Table 4.16: The accuracy,sensitivity and specificity for the $DD^G$ plot using the QDA classifier and FM depth on all three classes.

| | Accuracy(95% CI) | Sensitivity | Specificity |
|---|---|---|---|
| First 80 C Desorption | 0.80(0.65,0.90) | H:0.89,LC:0.70,COPD:0.86 | H:0.78,LC:0.92,COPD:0.97 |
| Second 80 C Desorption | 0.79(0.64,0.90) | H:0.76,LC:0.84,COPD:0.71 | H:0.81,LC:0.83,COPD:1.00 |
| 300 C Desorption | 0.79(0.64,0.90) | H:0.82,LC:0.89,COPD:0.43 | H:0.84,LC:0.79,COPD:1.00 |
| 200 C Desorption | 0.80(0.65,0.91) | H:0.94,LC:0.75,COPD:0.50 | H:0.73,LC:0.92,COPD:1.00 |

Table 4.17: The accuracy,sensitivity and specificity for the $DD^G$ plot using the K nearest neighbors classifier with FM depth on all three classes.

| | Accuracy(95% CI) | Sensitivity | Specificity |
|---|---|---|---|
| First 80 C Desorption | 0.91(0.79,0.98) | H:0.94,LC:0.85,COPD:1.00 | H:0.93,LC:0.96,COPD:0.97 |
| Second 80 C Desorption | 0.91(0.78,0.97) | H:0.94,LC:0.89,COPD:0.86 | H:0.88,LC:0.96,COPD:1.00 |
| 300 C Desorption | 0.81(0.67,0.92) | H:0.82,LC:1.00,COPD:0.29 | H:0.96,LC:0.71,COPD:1.00 |
| 200 C Desorption | 0.85(0.71,0.94) | H:0.89,LC:0.88,COPD:0.67 | H:0.91,LC:0.84,COPD:1.00 |

Table 4.18: The accuracy, sensitivity and specificity for the $DD^G$ plot using the non-parametric kernel method classifier with FM depth on all three classes.

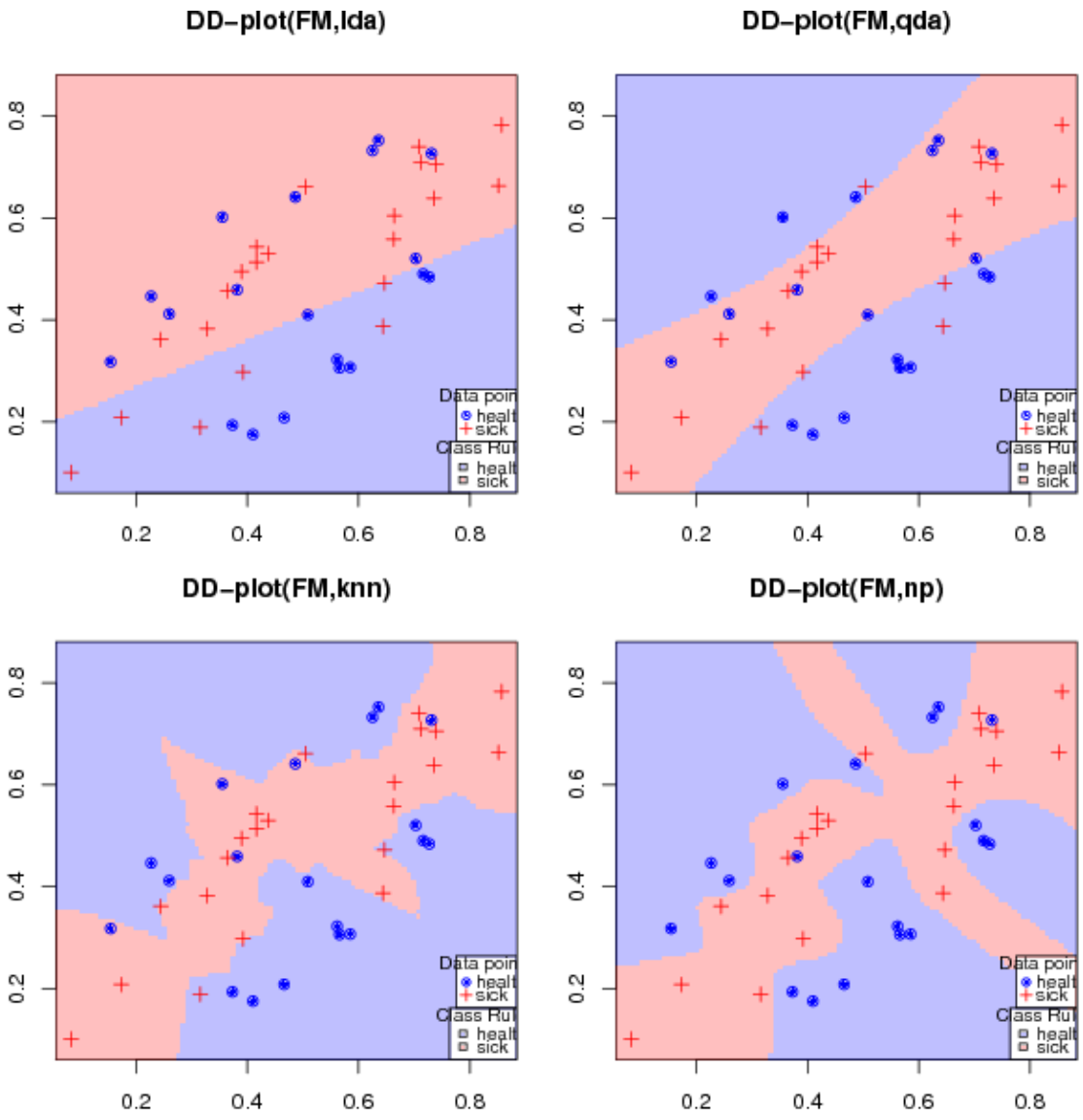


Figure 4.12: The $DD^G$ plots constructed on the 200°C desorb. The COPD subjects have been grouped with the lung cancer subjects to create the red group of sick subjects. The red shaded sections correspond to the decision rule for classifying new subjects as sick, any new points of data that have data depths in a red shaded area would be classified as sick, whereas new points landing the blue section would be classified as healthy.

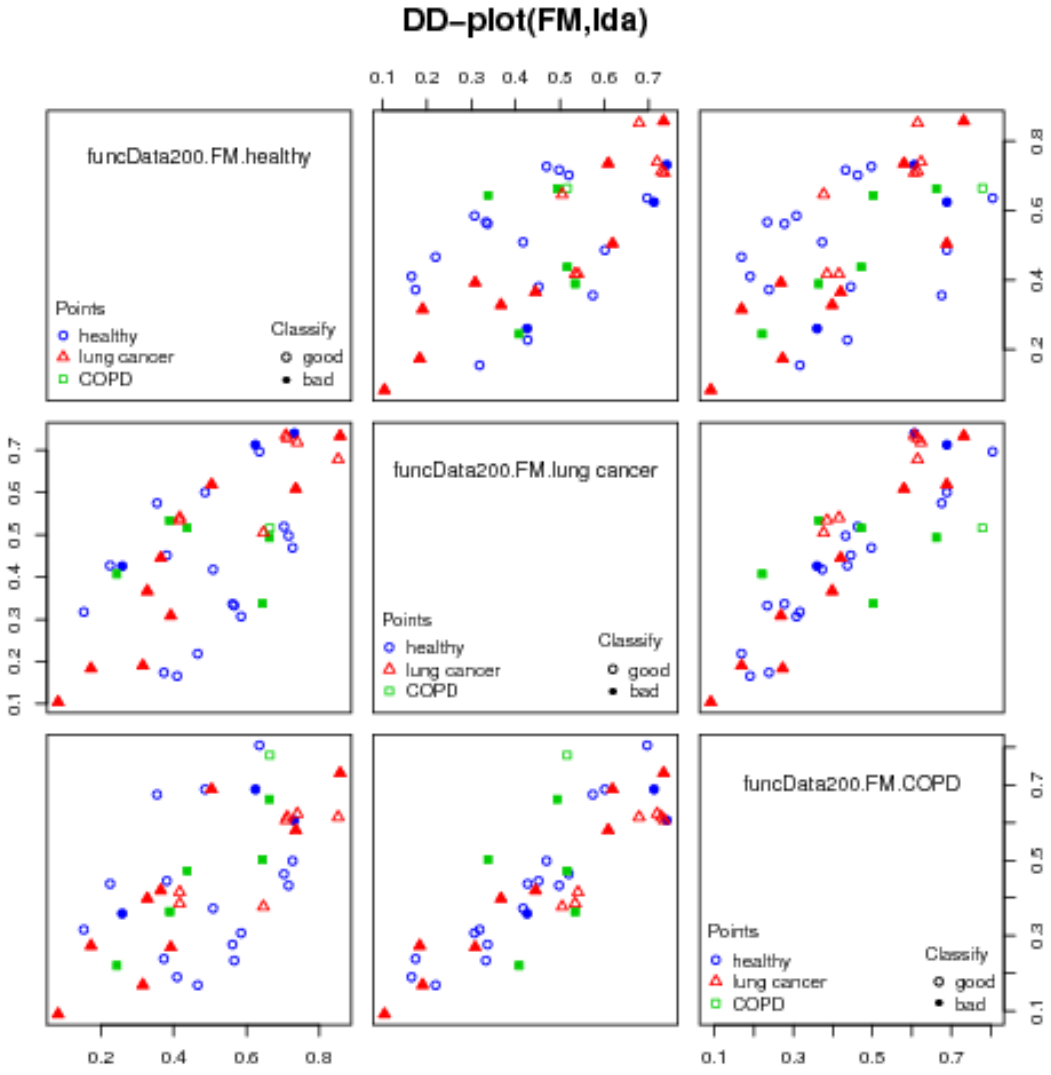


Figure 4.13: The $DD^G$ plot for the 200°C desorb on the prototype data using the LDA classifier with all three classes. Since there is now more than two groups, the groups are compared pairwise. This makes interpretation more difficult. Each non-diagonal panel compares the depths of the observations on two of the groups. Shaded points represent subjects that have been mis-classified. It is still clear from the plot that the LDA classified $DD^G$ plot is poor at distinguishing COPD and lung cancer subjects.

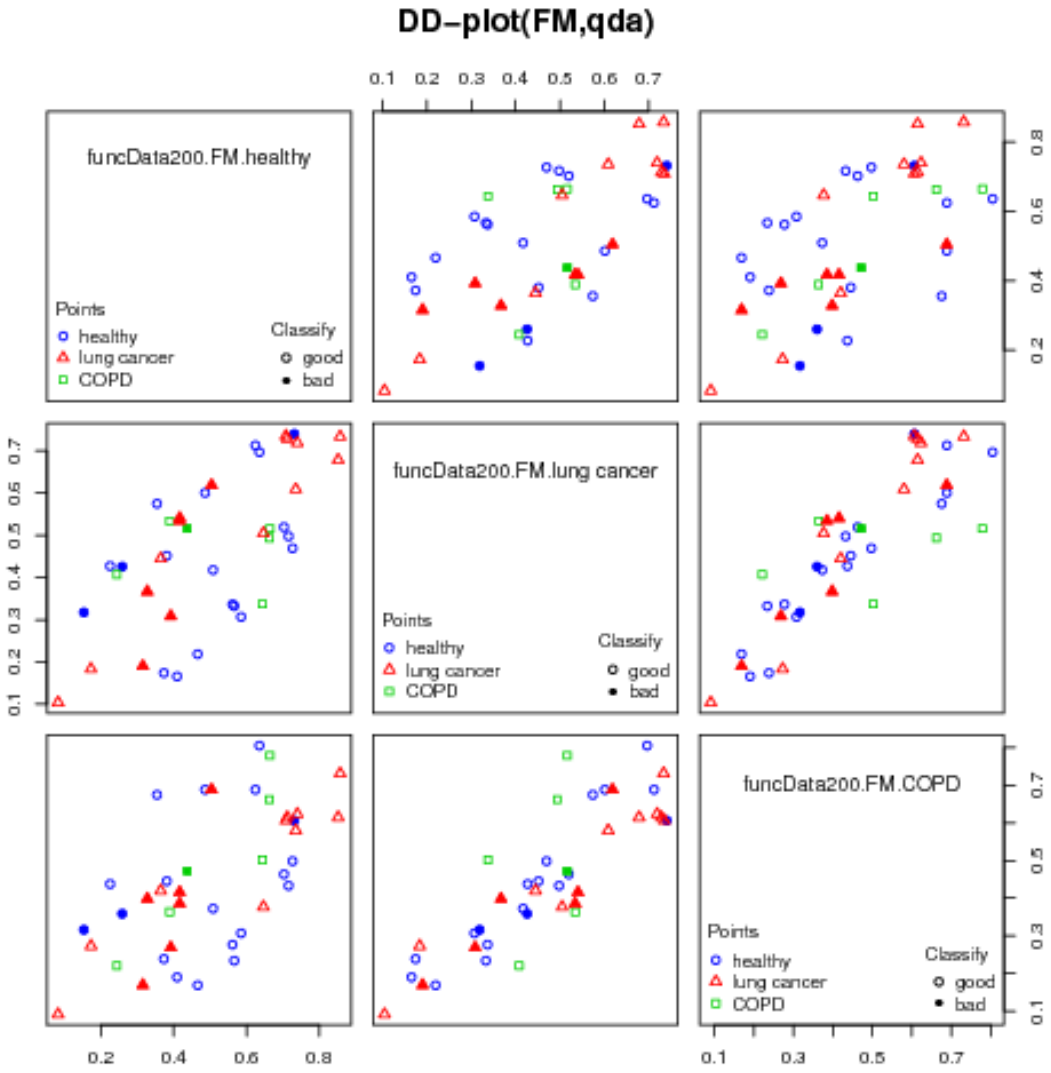


Figure 4.14: The $DD^G$ plot for the 200°C desorb on the prototype data using the QDA classifier with all three classes. Performance is better than with the LDA classifier.

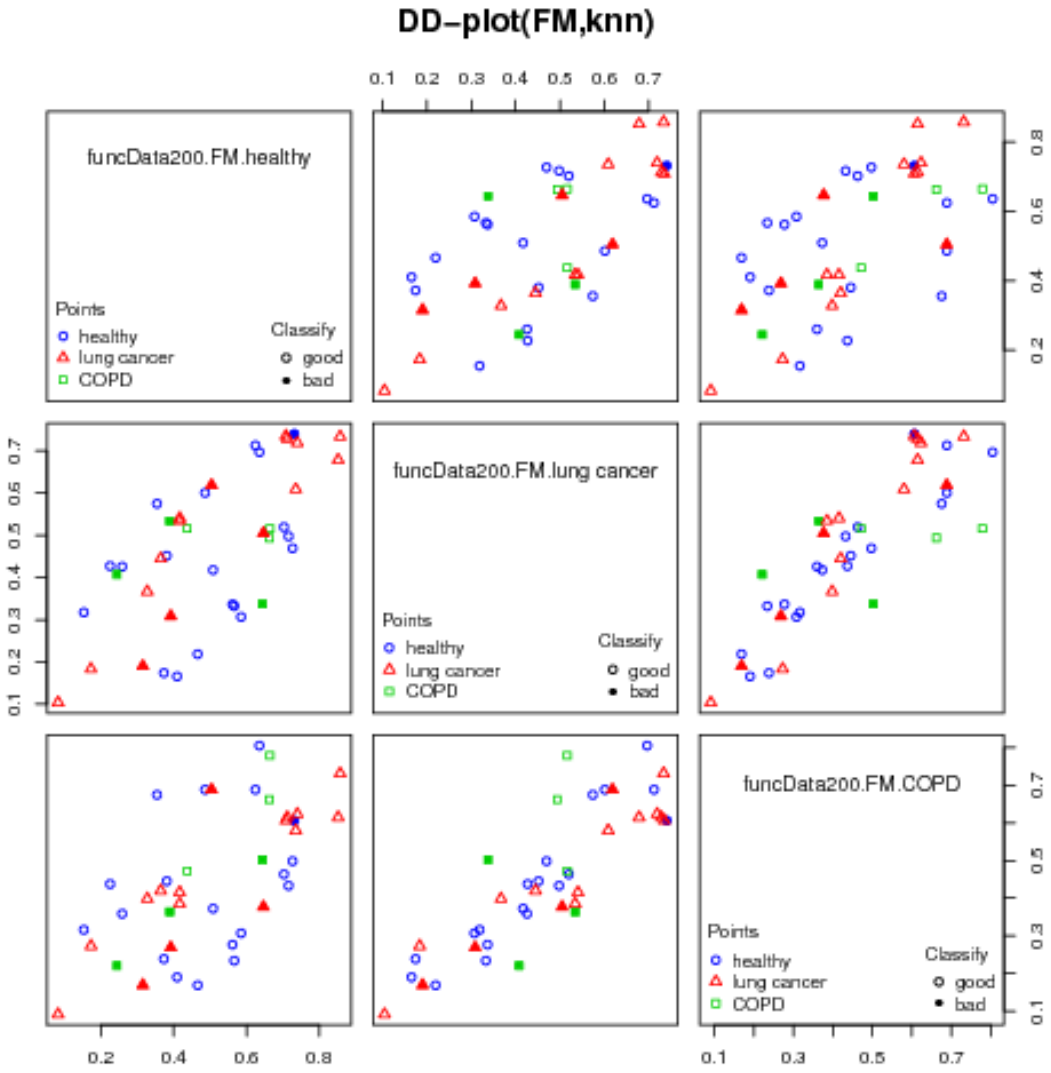


Figure 4.15: The $DD^G$ plot for the 200°C desorb on the prototype data using the kNN classifier with all three classes.

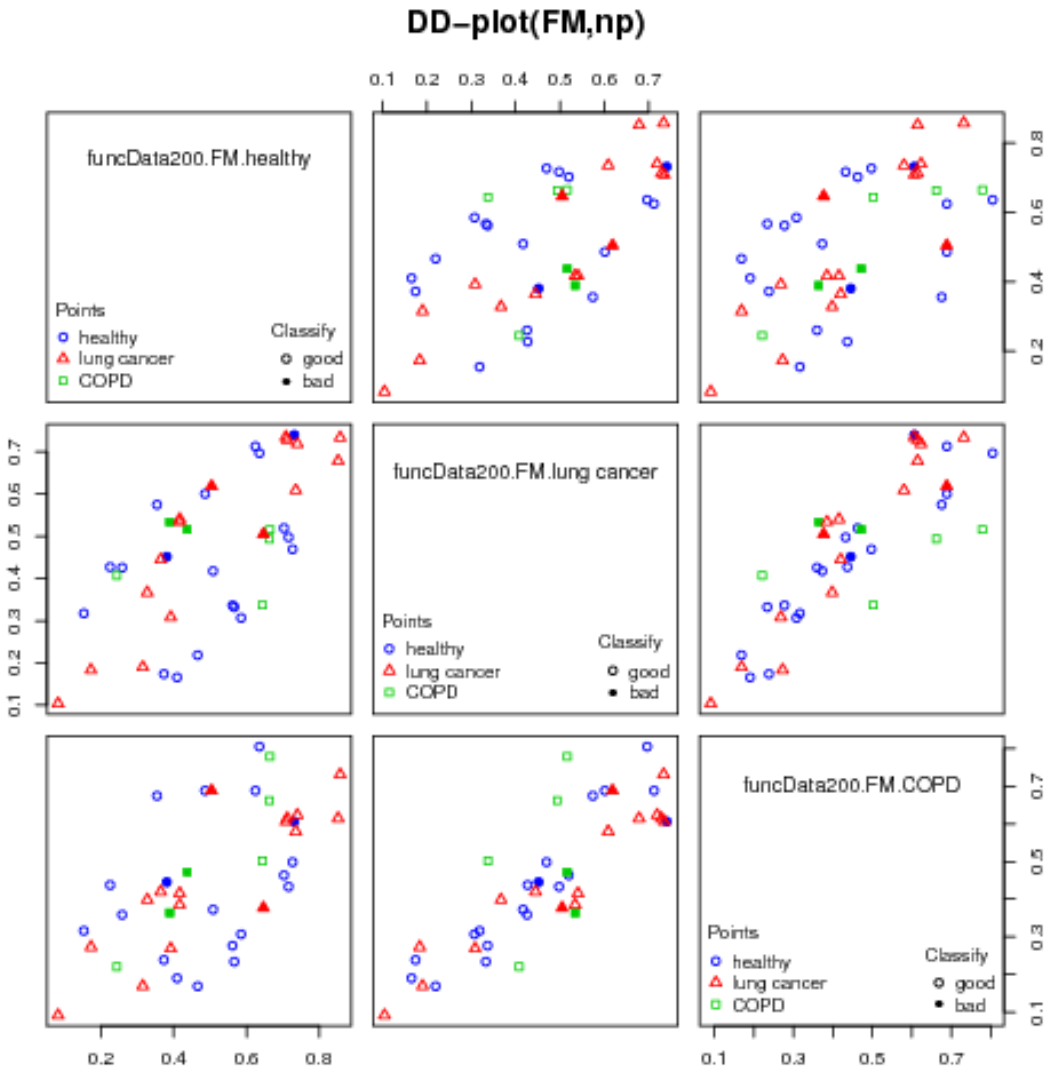


Figure 4.16: The $DD^G$ plot for the 200°C desorb on the prototype data using the non-parametric kernel method classifier with all three classes.

# Chapter 5

# Analysis of the Advanced Data

The same methods applied to the prototype dataset were applied to the advanced data. When taking a look at the advanced data a new issue cropped up. 75 out of 9856 measurements of K had a value of either -33.36 or 33.36. The positive values of 33.36 occur when the value of $\tau$ recorded is $1 \times 10^{-12}$. Similarly the values of -33.36 occur when $\tau_0$ is recorded as $1 \times 10^{-12}$. The Picomole engineering team suggested that these values were anomalous and should be treated as though they were missing. Taking their recommendation these values of K were set as missing along with the very small values of $\tau$ and $\tau_0$.

After the recommended changes were applied, the data were plotted in a similar fashion to the prototype data. The plot for the data on the 200°C desorption on one tube is in Figure 5.1. It reveals plots with a similar shape to those seen in the plot of prototype of Figure 4.1. One immediate difference is that the $\tau_0$ they appear more precisely measured with all of the $\tau_0$'s lying closer together than those seen in fig. 4.1. There is still no immediately obvious difference separating the healthy subjects from those with lung cancer revealed by plotting the data.

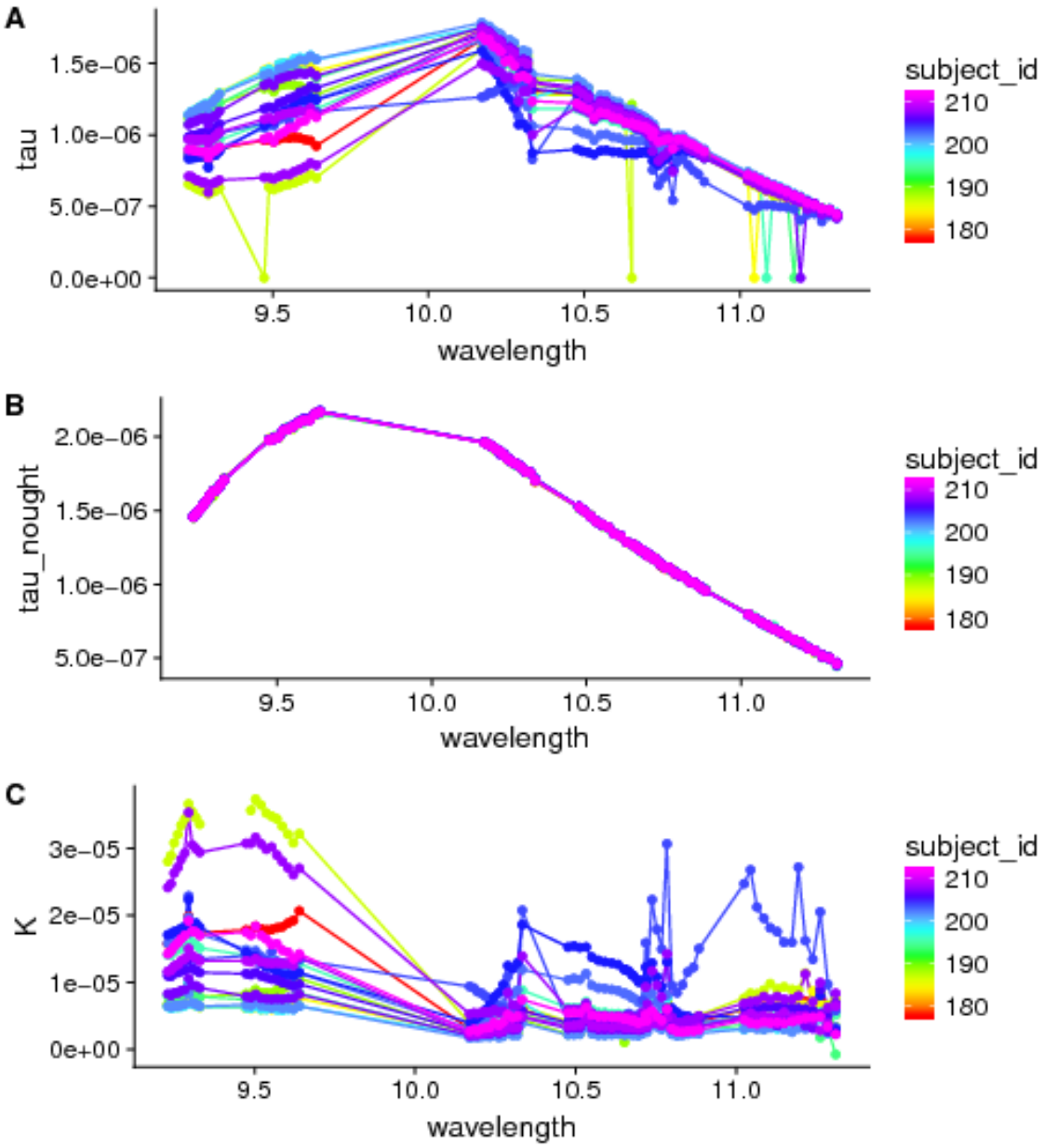


Figure 5.1: A: The $\tau$ data vs. the infrared light wavelengths measured. Points near zero are values of $\tau$ at $1 \times 10^{-12}$ before they were set to be missing. B: The $\tau_0$ data versus wavelengths measured. C: The K absorption data against the wavelengths. These plots refer only to measurements that were desorbed on 200°C tube 1 desorb on the advanced data.

## 5.1 Missing Data

Even with the problem where K values that were recorded as over 33 or under -33 are to be treated as missing, the percentage of missing observations are greatly reduced from the prototype data. In the prototype data 9.86% of the K observations are missing. By contrast, the advanced data has 0.761% of the K values missing. Not only has the number of missing values in the advanced data decreased sharply from the amount seen in the prototype data, there is also no longer any obvious pattern that can be seen among the missing values. This can be seen in the aggregation plots created for the new data in Figure 5.2 which has two aggregation plots separated by health status. It does appear however that missing values occur more likely at higher wavelengths. There is one measurement from a healthy subject on the 75°C desorption from the four tube set that is missing 23 wavelengths in a row going from 9.2295$\mu$m to 10.1823$\mu$m.

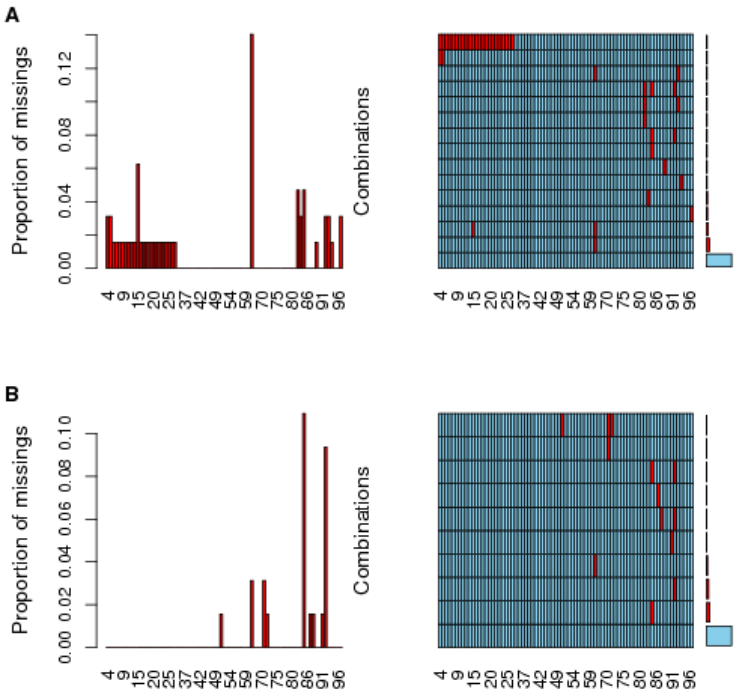


Figure 5.2: Aggregation plots for the advanced data grouped by health status. A: Healthy subjects. One desorb resulted in a string of missing measurements in the lowest wavelengths. B: Lung Cancer subjects.

| | Accuracy(95% CI) | Sensitivity | Specificity |
|---|---|---|---|
| 200°C Tube 1 | 0.81(0.54,0.96) | 1 | 0.62 |
| 75°C Tube 1 Tubes 3 | 0.94(0.7,1) | 0.88 | 1 |
| 150°C Tube 2 Tubes 3 | 1(0.79,1) | 1 | 1 |
| 200°C Tube 3 Tubes 3 | 0.94(0.7,1) | 1 | 0.88 |
| 75°C Tube 1 Tubes 4 | 0.75(0.48,0.93) | 0.62 | 0.88 |
| 150°C Tube 2 Tubes 4 | 1(0.79,1) | 1 | 1 |
| 225°C Tube 3 Tubes 4 | 0.94(0.7,1) | 1 | 0.88 |
| 300°C Tube 4 Tubes 4 | 0.88(0.62,0.98) | 0.75 | 1 |

Table 5.1: The accuracy, sensitivity and specificity for each of the pruned trees. Performance is improved over the prototype trees.

## 5.2 Decision Trees on Advanced Data

CARTs were grown in the same fashion as for the prototype data. For the CART trees these were grown with the same reduced requirement that only 5 measurements are needed in a node for a split to occur. These trees were also pruned in the same manner as the trees used on the prototype data. The trees for the three tube set of desorptions(75°C, 150°C, 200°C) along with the solitary 200°C tube desorption are shown in Figure 5.3. The trees for the remaining four trees from the same set of four tubes are in Figure 5.4. Like the prototype data the trees for each set of desorption on the advanced data all look different from one another and select different line IDs from one another to split on.

The performance of the CART trees on the advanced data is improved over that of the prototype. Table 5.1 summarizes the performance. The two trees based on the 150°C desorptions were able to perfectly classify every subject. Most of the other trees have accuracy above 90%, with the worst performing tree having 75% accuracy. This performance is quite good and is better than the CART tree performance on the prototype data.

However the performance of random forests in the next section raises some questions about the validity of CART performance.

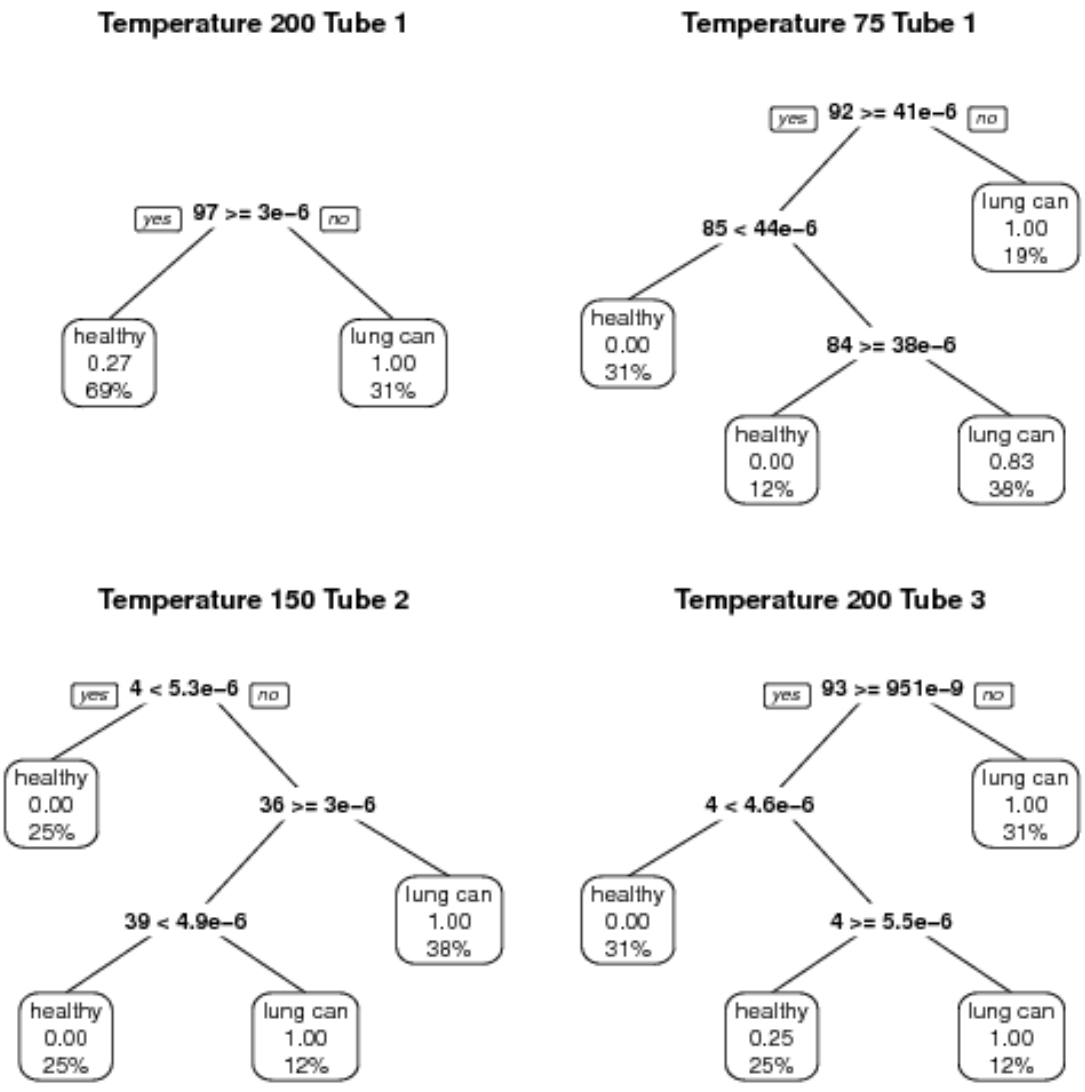


Figure 5.3: Pruned and cross validated trees for the K absorption data for the advanced data. These include the trees from the three tube desorption and the single tube 200°C desorption.

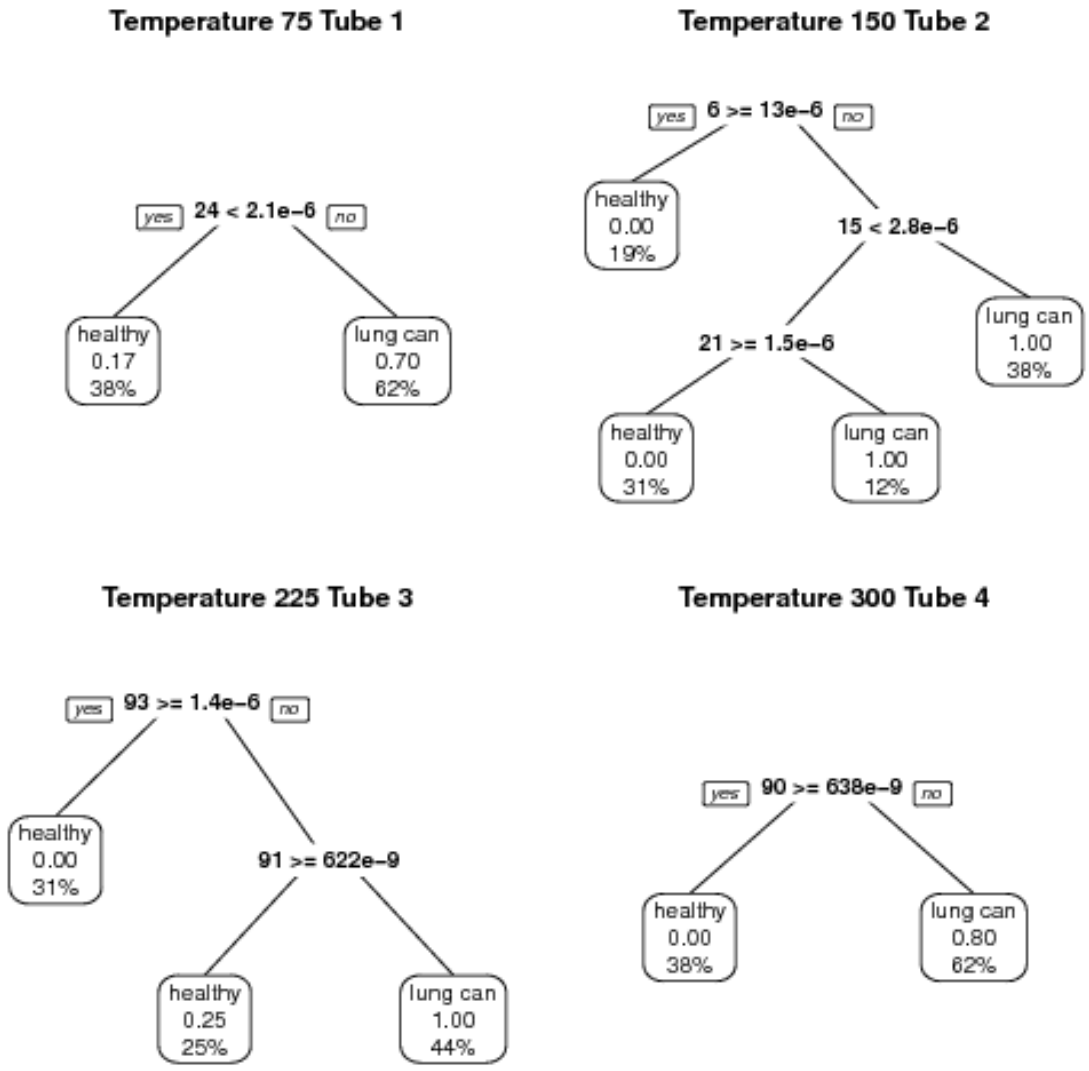


Figure 5.4: Pruned and cross validated trees for the K absorption data for the advanced data on the four tube set of desorptions. These desorptions go from 75°C to 150°C to 225°C finally to 300°C

| | Accuracy(95% CI) | Sensitivity | Specificity |
|---|---|---|---|
| 200°C Tube 1 | 0.5(0.25,0.75) | 0.5 | 0.5 |
| 75°C Tube 1 Tubes 3 | 0.94(0.7,1) | 0.88 | 1 |
| 150°C Tube 2 Tubes 3 | 0.44(0.2,0.7) | 0.38 | 0.5 |
| 200°C Tube 3 Tubes 3 | 0.5(0.25,0.75) | 0.38 | 0.62 |
| 75°C Tube 1 Tubes 4 | 0.56(0.3,0.8) | 0.62 | 0.5 |
| 150°C Tube 2 Tubes 4 | 0.062(0.0016,0.3) | 0 | 0.12 |
| 225°C Tube 3 Tubes 4 | 0.44(0.2,0.7) | 0.25 | 0.62 |
| 300°C Tube 4 Tubes 4 | 0.62(0.35,0.85) | 0.5 | 0.75 |

Table 5.2: The accuracy, sensitivity and specificity for each of random forests grown. Performance is similarly poor to that of the random forests grown on the prototype data.

### 5.2.1 Random Forests on the Full Set of Advanced Data

Random forests on the advanced data were handled differently to the random forests grown on the prototype data. The data was once again imputed using FPCA. Since the advanced dataset only has 16 subjects, a separate testing and training set was not used, and instead leave one out cross validation was applied. The same values of $M$, the number of predictor variables considered at each split were used. 500 trees were grown in each random forest.

The random forests on the advanced data perform like they did on the prototype data, which is not well. Table 5.2 summarizes the performance. With the exception of the random forest grown on the 75°C desorption which managed to have an accuracy, sensitivity and specificity of $\geq 88\%$ for each of them, the other random forests constructed perform quite poorly with second best accuracy rate being 56% and all the remaining random forests also have their no information rate fall within the 95% confidence intervals for their accuracy. The sets of conditions that were the best performing in the case of CART trees, the two 150°C desorptions did not perform well either. In fact, the 150°C on the four tube set of desorptions only managed to correctly classify a single subject.

### 5.2.2 CART and Random Forests on Smaller Bands of Wavelengths

Due to the discrepancy in performance between CART trees and Random Forests despite their common methodology it was speculated that the good performance of CART trees may be caused by the fact that there are more predictor variables than breath samples measurements for classification trees and random forests grown. In an attempt to account for this problem the classification trees and random forests were regrown on a reduced selection of wavelengths. Conversations with Picomole had previously hinted that the VOCs which only appear in people with lung cancer are best captured at the edges of the wavelengths. Therefore, for the regrown random forests and CART trees the wavelengths were restricted to the edges.

The methods used were the same as those used for the full dataset with the exception of the number of variables that the random forests are allowed to select during a split that is used in a grid search. The grid search was from two to the number of predictor variables in the reduced datasets.

They were first regrown on the band of the lowermost wavelengths, those from 9.2295$\mu$m-9.3294$\mu$m or the band colored "blue" in fig 3.1. The "blue" band contains 10 wavelengths. Since one of the 75°C measurements on the four tube set of desorptions has the entire "blue" band missing it was removed from the dataset. The other conditions still have all 16 measurements. Tables summarizing the performance of the CART trees and random forests on the "blue" band are 5.3 and 5.4 respectively. The best performing CART tree is in Figure 5.5. As with the classification trees and random forests grown on the full dataset the performance of the trees grown by CART exceeds that of the random forests. However, three of the CART trees here have their 95% confidence interval contain the 0.50 accuracy rate that could be achieved by random guessing.

Next, the band of the highest wavelengths from 11.0447$\mu$m-11.3099$\mu$m or the band colored "yellow" in 3.1 had classification trees and random forests grown on it. Since there was no measurement that had the entire band as missing all 16 measurements for each set of conditions were used. Table 5.5 has the performance of CART trees on the "yellow" band and 5.6 has it's random forest counterpart. The trend of low random forest performance and good CART tree performance on the 13 highest wavelengths continues. The CART trees perform as well as they did previously and

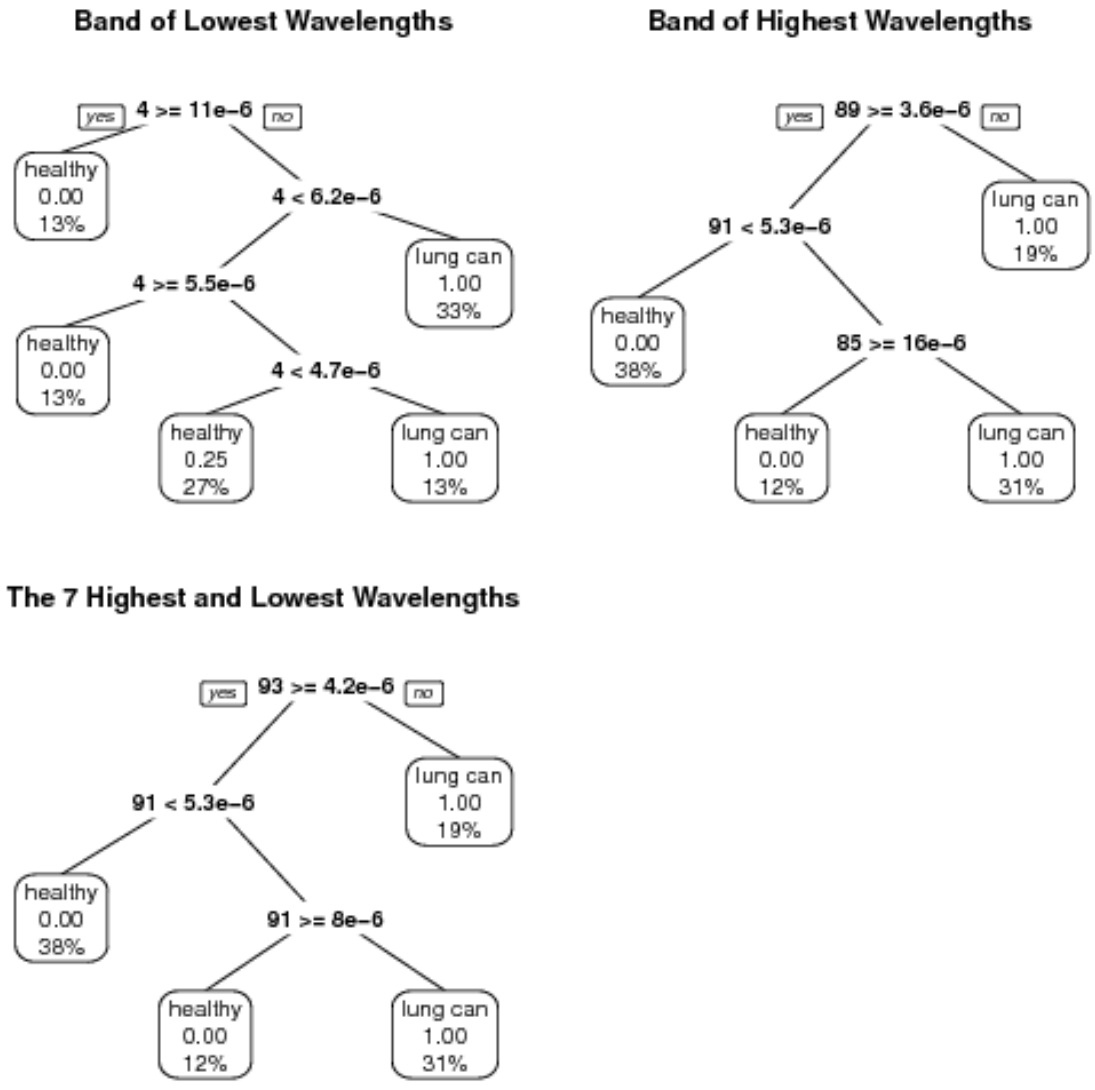


Figure 5.5: Pruned and cross validated trees for the K absorption data with a reduced number of lines. These trees are all from 75°C desorption from the four tube set. These are the better performing CART trees of the reduced datasets.

| | Accuracy(95% CI) | Sensitivity | Specificity |
|---|---|---|---|
| 200°C Tube 1 | 0.75(0.48,0.93) | 0.62 | 0.88 |
| 75°C Tube 1 Tubes 3 | 0.88(0.62,0.98) | 0.75 | 1 |
| 150°C Tube 2 Tubes 3 | 0.75(0.48,0.93) | 0.5 | 1 |
| 200°C Tube 3 Tubes 3 | 0.81(0.54,0.96) | 0.75 | 0.88 |
| 75°C Tube 1 Tubes 4 | 0.93(0.68,1) | 1 | 0.88 |
| 150°C Tube 2 Tubes 4 | 0.69(0.41,0.89) | 0.38 | 1 |
| 225°C Tube 3 Tubes 4 | 0.88(0.62,0.98) | 0.88 | 0.88 |
| 300°C Tube 4 Tubes 4 | 0.94(0.7,1) | 0.88 | 1 |

Table 5.3: The accuracy, sensitivity and specificity for the CART trees grown on the band of lowest wavelengths.

| | Accuracy(95% CI) | Sensitivity | Specificity |
|---|---|---|---|
| 200°C Tube 1 | 0.56(0.3,0.8) | 0.62 | 0.5 |
| 75°C Tube 1 Tubes 3 | 0.69(0.41,0.89) | 0.5 | 0.88 |
| 150°C Tube 2 Tubes 3 | 0.38(0.15,0.65) | 0.38 | 0.38 |
| 200°C Tube 3 Tubes 3 | 0.56(0.3,0.8) | 0.5 | 0.62 |
| 75°C Tube 1 Tubes 4 | 0.27(0.078,0.55) | 0.29 | 0.25 |
| 150°C Tube 2 Tubes 4 | 0.31(0.11,0.59) | 0.38 | 0.25 |
| 225°C Tube 3 Tubes 4 | 0.38(0.15,0.65) | 0.38 | 0.38 |
| 300°C Tube 4 Tubes 4 | 0.5(0.25,0.75) | 0.38 | 0.62 |

Table 5.4: The accuracy, sensitivity and specificity for the random forests grown on the band of lowest wavelengths.

| | Accuracy(95% CI) | Sensitivity | Specificity |
|---|---|---|---|
| 200°C Tube 1 | 0.81(0.54,0.96) | 1 | 0.62 |
| 75°C Tube 1 Tubes 3 | 0.94(0.7,1) | 0.88 | 1 |
| 150°C Tube 2 Tubes 3 | 0.88(0.62,0.98) | 1 | 0.75 |
| 200°C Tube 3 Tubes 3 | 0.81(0.54,0.96) | 1 | 0.62 |
| 75°C Tube 1 Tubes 4 | 1(0.79,1) | 1 | 1 |
| 150°C Tube 2 Tubes 4 | 0.81(0.54,0.96) | 0.75 | 0.88 |
| 225°C Tube 3 Tubes 4 | 0.94(0.7,1) | 1 | 0.88 |
| 300°C Tube 4 Tubes 4 | 0.88(0.62,0.98) | 0.75 | 1 |

Table 5.5: The accuracy, sensitivity and specificity for the CART trees grown on the band of highest wavelengths.

| | Accuracy(95% CI) | Sensitivity | Specificity |
|---|---|---|---|
| 200°C Tube 1 | 0.56(0.3,0.8) | 0.62 | 0.5 |
| 75°C Tube 1 Tubes 3 | 0.88(0.62,0.98) | 0.75 | 1 |
| 150°C Tube 2 Tubes 3 | 0.5(0.25,0.75) | 0.62 | 0.38 |
| 200°C Tube 3 Tubes 3 | 0.5(0.25,0.75) | 0.5 | 0.5 |
| 75°C Tube 1 Tubes 4 | 0.56(0.3,0.8) | 0.5 | 0.62 |
| 150°C Tube 2 Tubes 4 | 0.38(0.15,0.65) | 0.5 | 0.25 |
| 225°C Tube 3 Tubes 4 | 0.56(0.3,0.8) | 0.38 | 0.75 |
| 300°C Tube 4 Tubes 4 | 0.62(0.35,0.85) | 0.5 | 0.75 |

Table 5.6: The accuracy, sensitivity and specificity for the random forests grown only on the band of highest wavelengths.

random forests are the same including the 75°C from the three tube set random forest which has performance similar to the CART trees.

Finally the seven highest wavelengths and the seven lowest wavelengths were used as a reduced dataset for the random forest and CART classifiers. The wavelengths included are 9.2295-9.3294$\mu$m and 11.1735-11.3099$\mu$m. As with the other reduced datasets Table 5.7 summarizes the performance of the CART trees on these wavelengths and Table 5.8 for the random forests with similar results to the previous two reduced data sets.

In Beleites & Salzer[3] they found that applying aggregation to classifiers on very small sample sizes of chemometric data improved classifier stability. That is, the variance of model is reduced and predictive quality improved, poor models perform

| | Accuracy(95% CI) | Sensitivity | Specificity |
|---|---|---|---|
| 200°C Tube 1 | 0.81(0.54,0.96) | 1 | 0.62 |
| 75°C Tube 1 Tubes 3 | 0.94(0.7,1) | 1 | 0.88 |
| 150°C Tube 2 Tubes 3 | 0.88(0.62,0.98) | 0.75 | 1 |
| 200°C Tube 3 Tubes 3 | 0.81(0.54,0.96) | 1 | 0.62 |
| 75°C Tube 1 Tubes 4 | 1(0.79,1) | 1 | 1 |
| 150°C Tube 2 Tubes 4 | 0.94(0.7,1) | 1 | 0.88 |
| 225°C Tube 3 Tubes 4 | 0.94(0.7,1) | 1 | 0.88 |
| 300°C Tube 4 Tubes 4 | 0.81(0.54,0.96) | 0.62 | 1 |

Table 5.7: The accuracy, sensitivity and specificity for the CART trees grown on the seven highest and seven lowest wavelengths.

| | Accuracy(95% CI) | Sensitivity | Specificity |
|---|---|---|---|
| 200°C Tube 1 | 0.5(0.25,0.75) | 0.5 | 0.5 |
| 75°C Tube 1 Tubes 3 | 0.94(0.7,1) | 0.88 | 1 |
| 150°C Tube 2 Tubes 3 | 0.62(0.35,0.85) | 0.62 | 0.62 |
| 200°C Tube 3 Tubes 3 | 0.5(0.25,0.75) | 0.5 | 0.5 |
| 75°C Tube 1 Tubes 4 | 0.44(0.2,0.7) | 0.38 | 0.5 |
| 150°C Tube 2 Tubes 4 | 0.19(0.04,0.46) | 0.25 | 0.12 |
| 225°C Tube 3 Tubes 4 | 0.5(0.25,0.75) | 0.5 | 0.5 |
| 300°C Tube 4 Tubes 4 | 0.5(0.25,0.75) | 0.5 | 0.5 |

Table 5.8: The accuracy, sensitivity and specificity for the random forests grown only on the band of the seven highest and seven lowest wavelengths.

worse and good models better[3]. As random forests are a form of aggregated CART trees, this could account for the difference in performance. The CART trees may be picking up noise in the datasets which is then filtered out by the bagging performed by the random forests. And the 75°C Tube 3 random forest may actually perform well.

### 5.3 FLDA on Advanced Data

FLDA was applied twice to the advanced data, once for the desorption over three tubes with desorption temperatures 75°C, then 150°C then finally 200°desorption. Secondly, the other four tube desorption used goes through 75°C, 150°C, 225°C then finally 300°C.

When FLDA was applied to the advanced data 10 fold cross validation was used once again to select the best value of $q$ where the dimension of $q$ with the same dimensions tried with the prototype data, 5, 10, 25, 50, 100, 150, 200 and 300 for the four tube set and from 5 to 200 for the three tube set. This is because as with the FLDA performed on the prototype data, the data from desorptions performed on the three tube set were concatenated together to form a grid. Since there are only three tubes concatenated here the grid has $77 \times 3 = 231$ grid points. As $q$ is limited by the number of grid points, a smaller $q$ had to be used. Similarly the same method was applied for the four tube set of desorbs.

Then, as with the prototype data, the rank constraint p chosen was the one that minimized the error rate over cross validation. Due to the smaller sample size the values of p that could be used were limited. The values of p tried were 1,5,10 and 13 which was the largest that could be used successfully.

The optimal dimensions of $q$ were found to be 300 for the four tube set and 200 for the three tube set. The values of the rank constraint p chosen were 10 for the four tube set and 5 for the three tube set.

Using the selected parameters the FLDA was run using 10 fold cross validation. Plots of the $\hat{\alpha}$s for the two times FLDA was run using the set of parameters selected are in Figures 5.6 and 5.7. The FLDA on the four tube set of desorptions has an immediately obvious outlier of a healthy subject with an $\hat{\alpha}$ of over 297. The accuracy of the FLDA on the four tube desorption set is only 50%, the three tube desorption

| | Accuracy(95% CI) | Sensitivity | Specificity | No Information Rate |
|---|---|---|---|---|
| FLDA on 4 tube set | 0.5(0.25,0.75) | 0.38 | 0.62 | 0.5 |
| FLDA on 3 tube set | 0.75(0.48,0.93) | 0.62 | 0.88 | 0.5 |

Table 5.9: The accuracy, sensitivity and specificity for the two FLDA trials conducted along with the no information rate. Performance on the three tube set of conditions is similar to that of the FLDA on the prototype data.

set does a little better getting 75% of the subjects correctly classified. However the sensitivity of even three tube set FLDA is not that great at 62%.

The very small sample size here limited the values of p that could be chosen. The FLDA on the prototype data saw its performance increase greatly with the higher values of p, while the FLDA on the advanced data did not chose the highest available option of p, the performances of the p's available were all very similar.

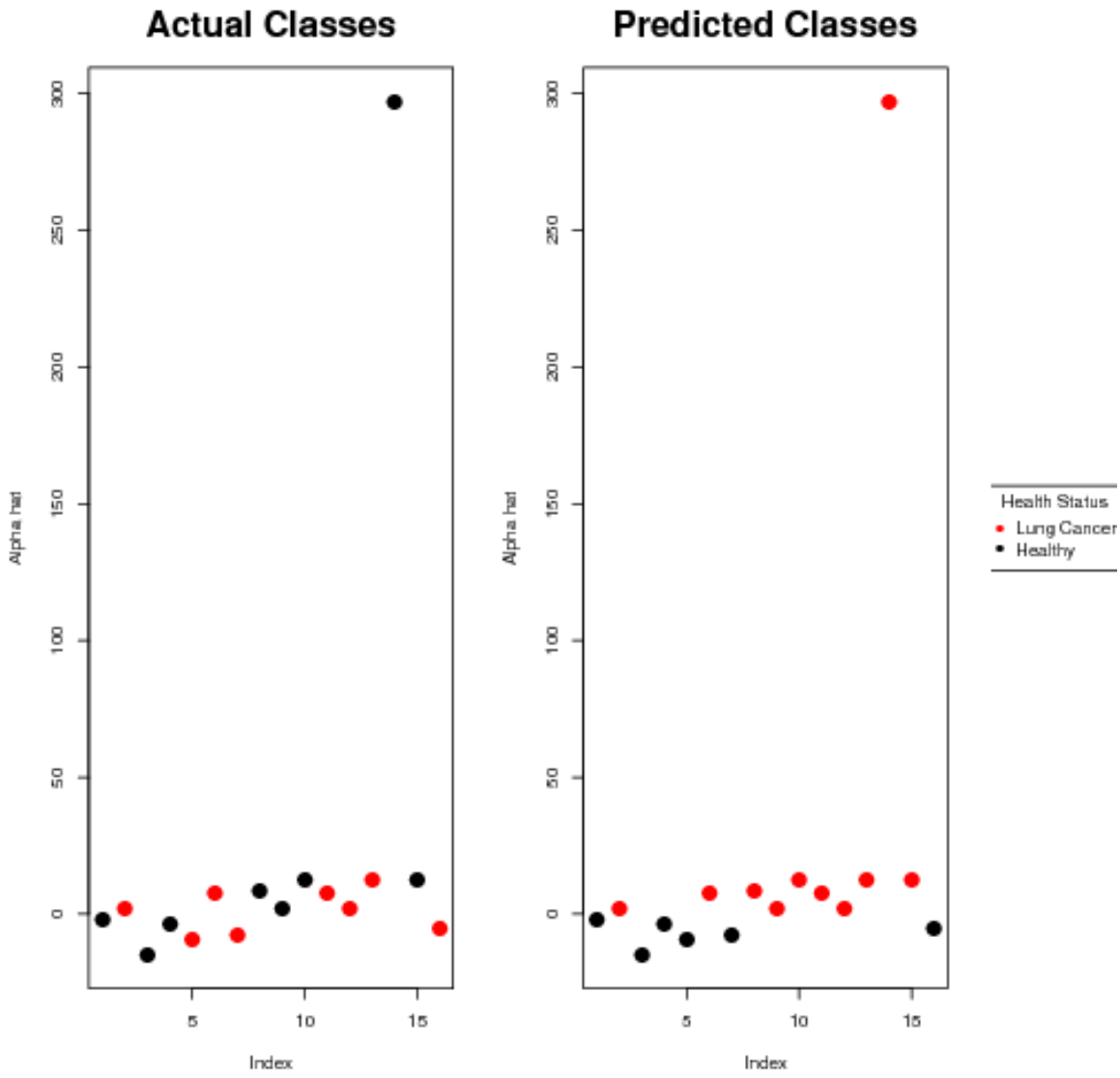


Figure 5.6: The $\hat{\alpha}$ estimates for the FLDA performed on the four tube desorption. One thing that stands out is the subject whose value of $\hat{\alpha}$ is several times larger than that of all the others. The FLDA here correctly classifies only 50% of the subejcts.

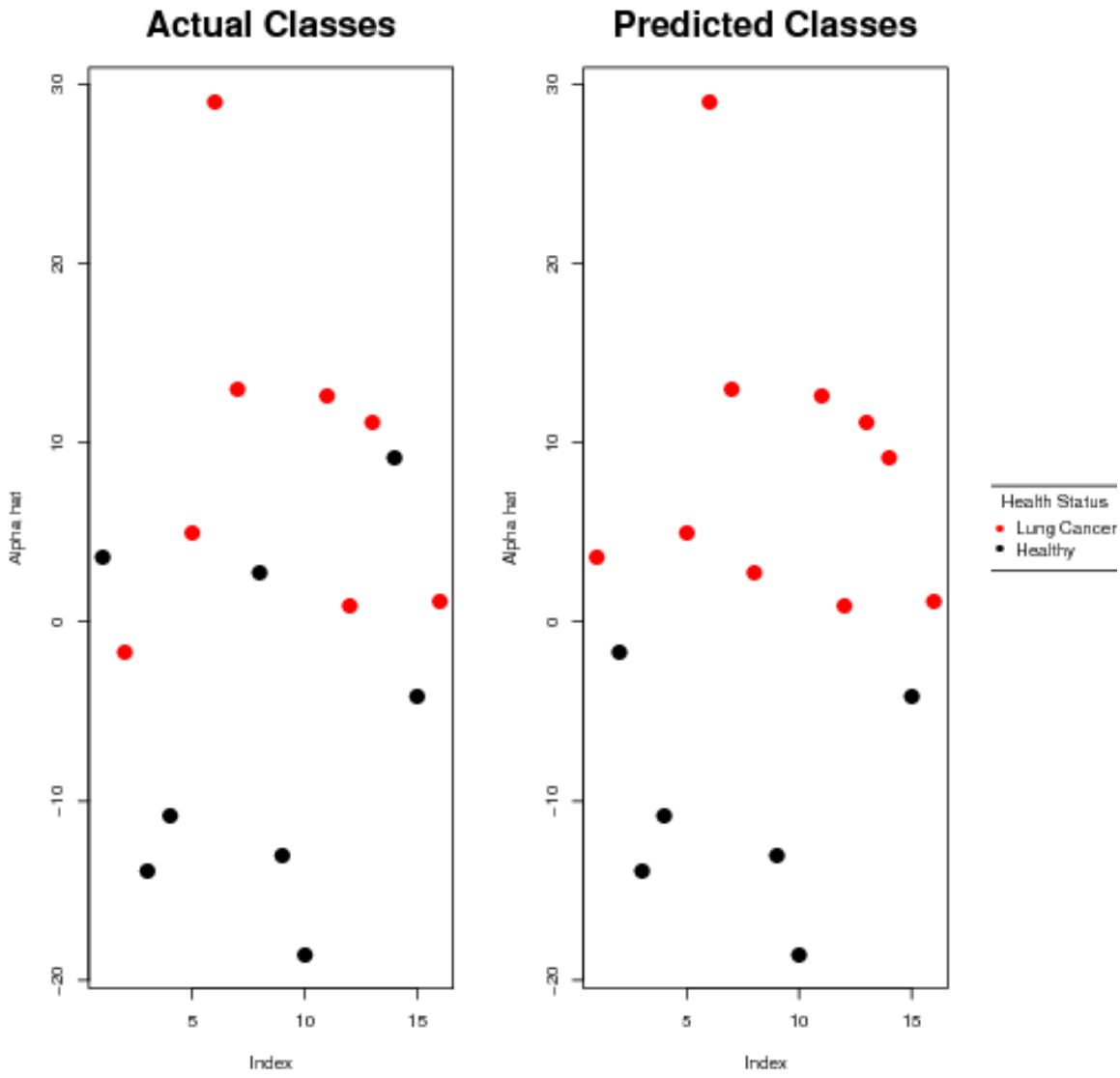


Figure 5.7: The $\hat{\alpha}$ estimates for the FLDA conducted on the three tube set of desorptions. Performance is better than the FLDA on the four tube set, the separation between the two groups is better and it didn't run into the same outlier issue with one subject having an extreme value of $\hat{\alpha}$.

## 5.4 FPCA and Related Techniques on Advanced Data

As with the prototype data, FPCA analysis was conducted on advanced data and FPCA clustering was used in attempt to see if an unsupervised machine learning method is capable of finding good separation between the two classes in addition to being used to impute any missing data which is required to use random forests and $DD^G$ plots on the advanced data.

All of the sets of conditions have over 95% of their variance explained by their first three FPCs. As with the FPCA performed on the prototype data, the FPCs have no obvious interpretation.

As with the prototype data, the performance of the FPCA clustering is very poor and the cluster is essentially guessing. Table 5.10 summarizes the performance of FPCA clustering. It struggles to achieve accuracy rates better than 50% and all the 95% confidence intervals for the accuracy rate all contain the no information rate. This suggests that the FPCA clustering is unable to find any sort of difference between the two classes.

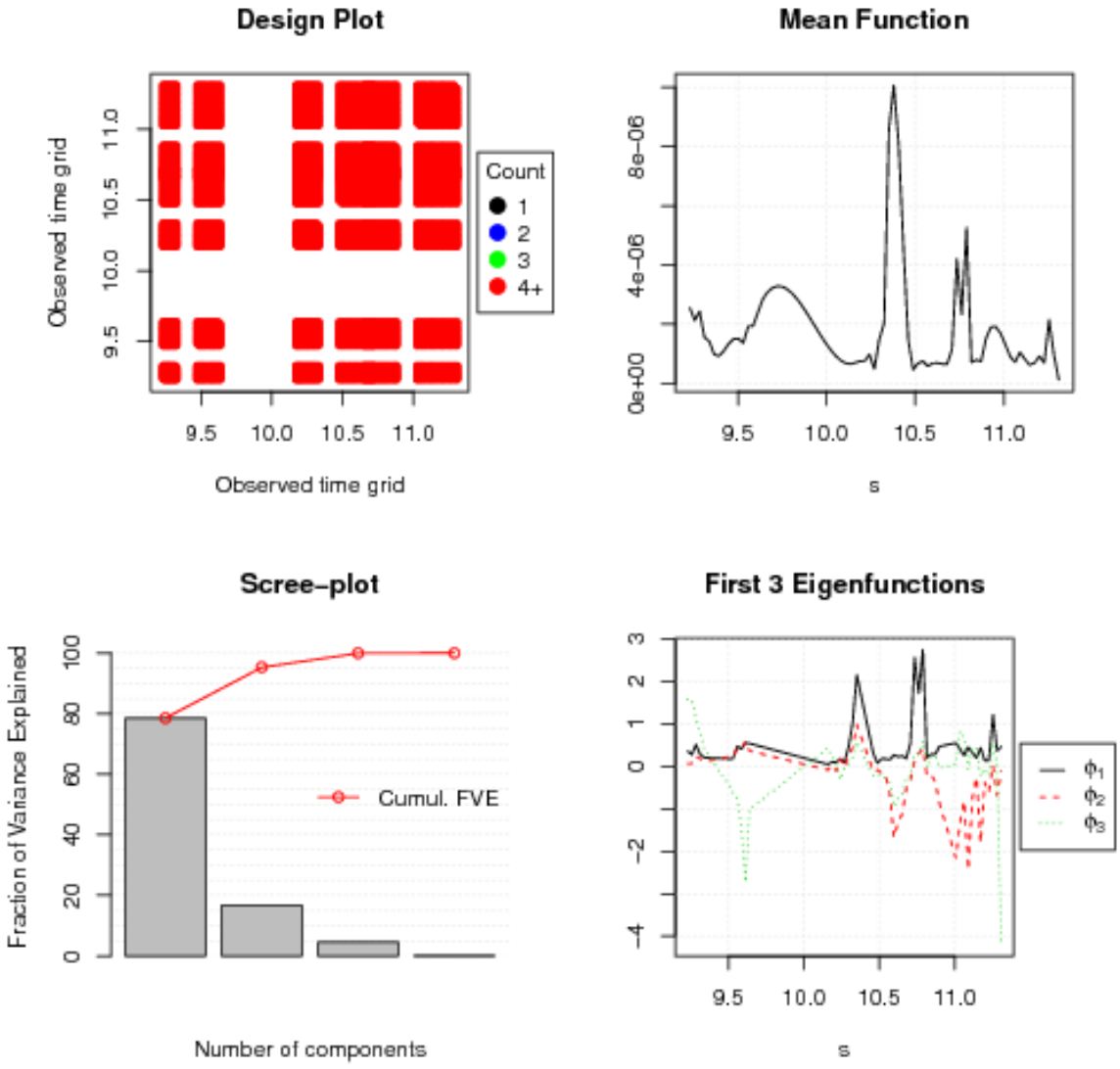


Figure 5.8: Diagnostic plots for the FPCA performed on the 300°C desorption of the advanced data. The top left plot show spacing and density of the grid used, this data is considered dense with missing data. The top right plot is the path of the mean curve. The lower left plot shows the fraction of variance explained by each FPC, the first principal component in this case explains 80% of the variance. 95% of the variance is explained by the first three FPC. The final plot, the bottom right shows the path of the first three eigenfuction curves.

| | Accuracy(95% CI) | Sensitivity | Specificity |
|---|---|---|---|
| 200°C Tube 1 | 0.31(0.11,0.59) | 0.62 | 0 |
| 75°C Tube 1 Tubes 3 | 0.44(0.2,0.7) | 0.12 | 0.75 |
| 150°C Tube 2 Tubes 3 | 0.5(0.25,0.75) | 0.75 | 0.25 |
| 200°C Tube 3 Tubes 3 | 0.38(0.15,0.65) | 0.62 | 0.12 |
| 75°C Tube 1 Tubes 4 | 0.5(0.25,0.75) | 0.88 | 0.12 |
| 150°C Tube 2 Tubes 4 | 0.44(0.2,0.7) | 0.88 | 0 |
| 225°C Tube 3 Tubes 4 | 0.44(0.2,0.7) | 0.88 | 0 |
| 300°C Tube 4 Tubes 4 | 0.69(0.41,0.89) | 0.38 | 1 |

Table 5.10: The accuracy, sensitivity and specificity for the 2 group clustering via. FPCA on the advanced data. Performance of FPCA clustering is equally poor as with the prototype data.

## 5.5 $DD^G$ Plots on Advanced Data

The $DD^G$ plots were constructed in exactly the same fashion as was done with the prototype data. The data were imputed using FPCA and FM depths were used for all plots. Originally the same four classifiers were planned to be used. However, the k nearest neighbor $DD^G$ classifier was unable to run on the smaller advanced dataset. Instead the DD3 classifier was used. DD3 just tries to fit the best possible third degree polynomial.

Tables 5.11 through 5.14 have the accuracy, sensitivity and specificity for the four different $DD^G$ plots built for each set of conditions on the advanced data. As with the prototype data, the more freedom the classifiers have in drawing the classification rules, the better they perform, with the non-parametric kernel method performing the best once again. Of course these classifiers may just be overfitting the data.

| | Accuracy(95% CI) | Sensitivity | Specificity |
|---|---|---|---|
| 200°C Tube 1 | 0.69(0.41,0.89) | 0.75 | 0.62 |
| 75°C Tube 1 Tubes 3 | 0.56(0.3,0.8) | 0.5 | 0.62 |
| 150°C Tube 2 Tubes 3 | 0.62(0.35,0.85) | 0.62 | 0.62 |
| 200°C Tube 3 Tubes 3 | 0.81(0.54,0.96) | 0.75 | 0.88 |
| 75°C Tube 1 Tubes 4 | 0.62(0.35,0.85) | 0.62 | 0.62 |
| 150°C Tube 2 Tubes 4 | 0.69(0.41,0.89) | 0.75 | 0.62 |
| 225°C Tube 3 Tubes 4 | 0.69(0.41,0.89) | 0.75 | 0.62 |
| 300°C Tube 4 Tubes 4 | 0.81(0.54,0.96) | 0.75 | 0.88 |

Table 5.11: The accuracy, sensitivity and specificity for the $DD^G$ plot using the LDA classifier.

| | Accuracy(95% CI) | Sensitivity | Specificity |
|---|---|---|---|
| 200°C Tube 1 | 0.81(0.54,0.96) | 0.88 | 0.75 |
| 75°C Tube 1 Tubes 3 | 0.69(0.41,0.89) | 0.62 | 0.75 |
| 150°C Tube 2 Tubes 3 | 0.94(0.7,1) | 1 | 0.88 |
| 200°C Tube 3 Tubes 3 | 0.88(0.62,0.98) | 0.75 | 1 |
| 75°C Tube 1 Tubes 4 | 0.69(0.41,0.89) | 0.75 | 0.62 |
| 150°C Tube 2 Tubes 4 | 0.75(0.48,0.93) | 0.88 | 0.62 |
| 225°C Tube 3 Tubes 4 | 0.88(0.62,0.98) | 1 | 0.75 |
| 300°C Tube 4 Tubes 4 | 0.88(0.62,0.98) | 0.88 | 0.88 |

Table 5.12: The accuracy, sensitivity and specificity for the $DD^G$ plot on the advanced data with the QDA classifier.

| | Accuracy(95% CI) | Sensitivity | Specificity |
|---|---|---|---|
| 200°C Tube 1 | 0.81(0.54,0.96) | 1 | 0.62 |
| 75°C Tube 1 Tubes 3 | 0.81(0.54,0.96) | 1 | 0.62 |
| 150°C Tube 2 Tubes 3 | 0.81(0.54,0.96) | 1 | 0.62 |
| 200°C Tube 3 Tubes 3 | 0.94(0.7,1) | 0.88 | 1 |
| 75°C Tube 1 Tubes 4 | 0.81(0.54,0.96) | 1 | 0.62 |
| 150°C Tube 2 Tubes 4 | 0.81(0.54,0.96) | 0.62 | 1 |
| 225°C Tube 3 Tubes 4 | 0.75(0.48,0.93) | 1 | 0.5 |
| 300°C Tube 4 Tubes 4 | 0.94(0.7,1) | 1 | 0.88 |

Table 5.13: The accuracy, sensitivity and specificity for the $DD^G$ plot on the advanced data with the DD3 classifier.

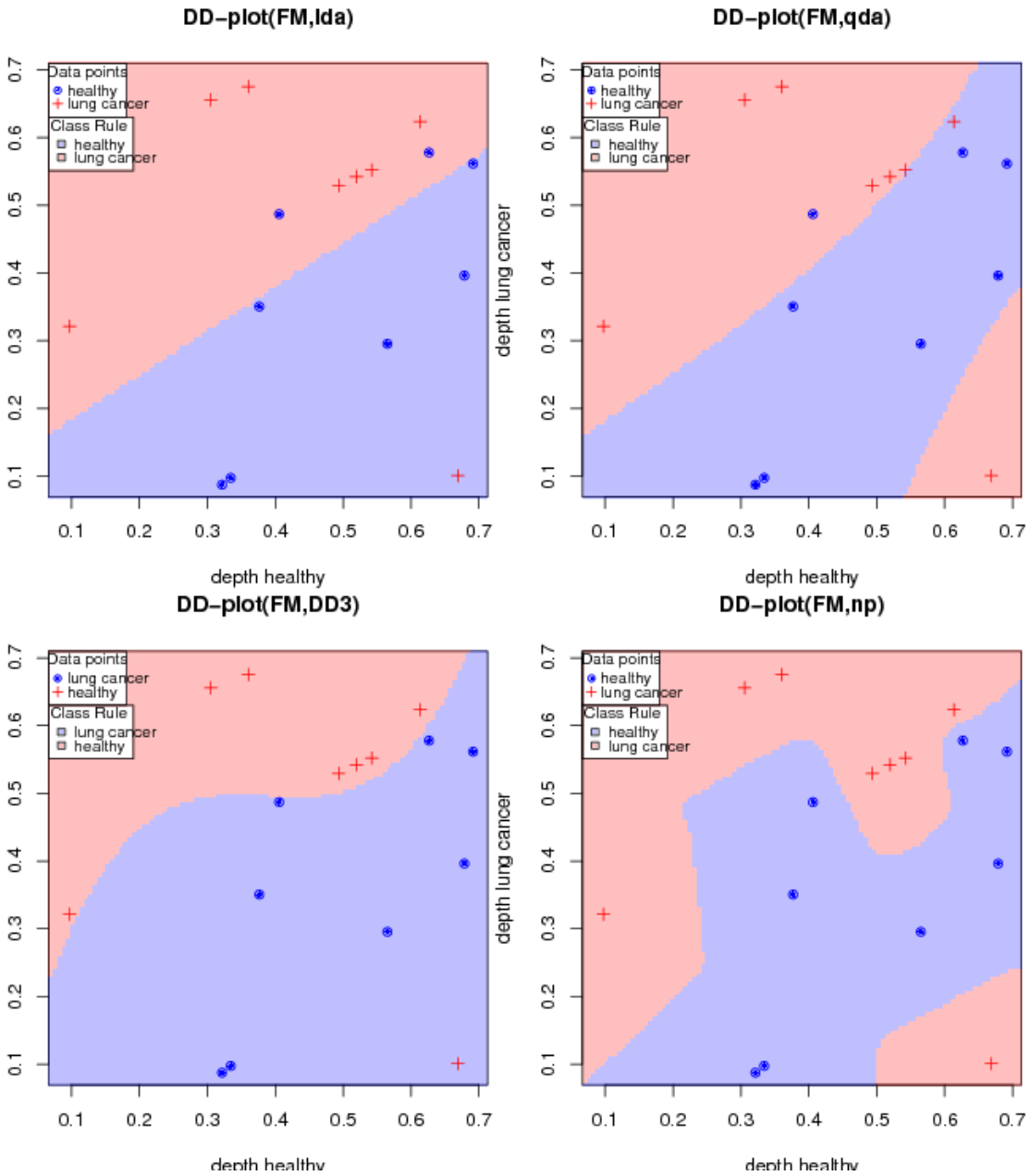


Figure 5.9: The DD$^G$ plots for the 300°C condition on the new set of data. This is Representative of the application of DD$^G$ plots on the advanced data.

| | Accuracy(95% CI) | Sensitivity | Specificity |
|---|---|---|---|
| 200°C Tube 1 | 0.94(0.7,1) | 1 | 0.88 |
| 75°C Tube 1 Tubes 3 | 0.94(0.7,1) | 0.88 | 1 |
| 150°C Tube 2 Tubes 3 | 0.88(0.62,0.98) | 1 | 0.75 |
| 200°C Tube 3 Tubes 3 | 0.94(0.7,1) | 0.88 | 1 |
| 75°C Tube 1 Tubes 4 | 0.94(0.7,1) | 0.88 | 1 |
| 150°C Tube 2 Tubes 4 | 1(0.79,1) | 1 | 1 |
| 225°C Tube 3 Tubes 4 | 0.94(0.7,1) | 1 | 0.88 |
| 300°C Tube 4 Tubes 4 | 1(0.79,1) | 1 | 1 |

Table 5.14: The accuracy, sensitivity and specificity for the $DD^G$ plot on the advanced data with the NP classifier.

# Chapter 6

# Conclusion

The high rate and unique pattern of missing data visible in the prototype data warranted the further investigation conducted in Section 4.1. While it had appeared at first as though the missing data was the result of differences between the unhealthy and healthy subjects, it was actually revealed to be the result of issues with Picomole's breath analysis system during March and April of 2012. Using the $\tau_0$ data and a couple of two sample tests designed for functional data, the FAD test and functional Schilling procedure revealed that the distributions of the data during that period and the data outside that period are not the same. With this caveat in place, several classification methods were examined on the prototype data in Chapter 4.

The first of the classifiers applied to the prototype were classification trees using the CART algorithm. Several of the trees were capable of classifying the subjects better than random guessing as shown in Table 4.3. The random forests grown on the prototype data are poorly performing along with FPCA clustering. Neither method was able to do a better job than random guessing. FLDA also achieved good results when classifying only two of the three groups. $DD^G$ plots with four different classifiers, LDA, QDA, kNN and a non-parametric kernel method were also applied. These results were good. The results from the prototype data are promising at times while also being haunted by the specter of large amounts of missing data, and the fact the majority of lung cancer subjects were processed during the period with which most of the missing data occurred.

Data from Picomole's ongoing data collection efforts was looked at with similar methods to what was done on the prototype data in Chapter 5. Unlike the prototype data the advanced data was age and sex matched and no samples from COPD subjects were collected. More crucially, breath samples from lung cancer subjects were from post-treatment subjects rather than pre-treatment as had been the case with the prototype data. This change, along with changes in experimental design made using

the advanced dataset as a testing set for classifiers trained on the prototype data impossible and led to all methods having to be retrained on the advanced data. The methods that performed well on the prototype data also saw good results on the real data. Those methods were classification trees, FLDA and $DD^G$ plots. Several of the CART trees such as the two 150°C desorbs seen in Table 5.1, FPCA clustering and random forests still did poorly on the advanced data.

The two samples here were both quite small and are not perfect datasets yet both suggest that it is possible to separate lung cancer subjects from those that are healthy using only breath sample data. The limitations of the small sample sizes make it difficult to tell how accurate the classifiers would be able to perform in real world applications. Hopefully with the complete version of the advanced data being collected it will be possible to get a better picture of how things might perform in practice.

## 6.1 Future Directions

The receipt of a NSERC Engage Plus grant has allowed for work on this problem to continue into the future. More data from the currently ongoing collection will allow for further validation of the methods outlined here while also hopefully opening new avenues that could not have been explored before.

Since the datasets provided here were both small and suffer from data quality issues, one obvious direction would be to retry the same methods especially the better performers, and see how they fare on a larger better quality dataset such as the one that Picomole will have when their most recent data collection is complete. Having the larger sample would allow for using a significantly larger training and testing dataset which could lead to more accurate classification rules. The larger data set might also give a better indication of how the models might perform if allowed to run on data from patients in real world applications.

Another possible direction would be to take the concept of ensemble classifiers further than the random forests that were grown. For instance classification trees were all grown on separate sets of conditions. All of those trees could be combined together into one classifier using a committee method such as stacked generalization. Other learning methods could also be combined in the same fashion such as combining

FLDA or $DD^G$.

In addition, methods that could not be considered with the data at hand could be attempted with a larger dataset of higher quality. For example a common technique in FDA is to examine the first and second derivative estimates of the data curves[23], with the high level of missingness found in the prototype data such estimates are numerically unstable and could not be used. Other machine learning tools that require a greater number of samples to be practical such as neural networks could be investigated as well.

With the collection of post-treatment lung cancer subjects in addition to the pre-treatment subjects on the data collection that is currently ongoing, when more pre-treatment subjects are available it should be possible to determine whether or not there are observable differences between the two groups of lung cancer subjects. If it turns out that there is no difference between the two different groups then it should be possible to apply analysis from one group to another.

While there appears to a detectable difference between lung cancer subjects and healthy subjects with breath samples based on some of the classifiers applied, more work needs to be done in creating a classification system that could be used as a good preliminary test.

# Bibliography


[1] Anton Amann et al. "The human volatilome: volatile organic compounds (VOCs) in exhaled breath, skin emanations, urine, feces and saliva". *Journal of breath research* 8.3 (2014), p. 034001.

[2] Amel Bajtarevic et al. "Noninvasive detection of lung cancer by analysis of exhaled breath". *BMC cancer* 9.1 (2009), p. 348.

[3] Claudia Beleites and Reiner Salzer. "Assessing and improving the stability of chemometric models in small sample size situations". *Analytical and bioanalytical chemistry* 390.5 (2008), pp. 1261–1271.

[4] Claudia Beleites et al. "Sample size planning for classification models". *Analytica chimica acta* 760 (2013), pp. 25–33.

[5] Leo Breiman and Adele Cutler. *Random Forests*. Oct. 31, 2017. URL: `https://www.stat.berkeley.edu/~breiman/RandomForests/cc_manual.htm#l2`.

[6] Leo Breiman et al. *Classification and regression trees*. CRC press, 1984.

[7] Alejandra Cabaña et al. "Permutation tests in the two-sample problem for functional data". *arXiv preprint arXiv:1610.06960* (2016).

[8] Jeng-Min Chiou et al. "A functional data approach to missing value imputation and outlier detection for traffic flow data". *Transportmetrica B: Transport Dynamics* 2.2 (2014), pp. 106–129.

[9] Juan A Cuesta-Albertos, Manuel Febrero-Bande, and M Oviedo de la Fuente. "The $DD^G$-classifier in the functional setting". *Test* 26.1 (2017), pp. 119–142.

[10] Silvano Dragonieri et al. "An electronic nose in the discrimination of patients with non-small cell lung cancer and COPD". *Lung cancer* 64.2 (2009), pp. 166–170.

[11] Frédéric Ferraty and Philippe Vieu. *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media, 2006.

[12] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. 2nd ed. Vol. 1. Springer series in statistics New York, 2009.

[13] SM Gordon et al. "Volatile organic compounds in exhaled air from patients with lung cancer." *Clinical chemistry* 31.8 (1985), pp. 1278–1282.

[14] Hadjipantelis et al. *Functional PCA in R. A software primer using fdapace*. Jan. 24, 2017. URL: `https://cran.r-project.org/web/packages/fdapace/vignettes/fdapaceVignetteKnitr.pdf`.

[15] Lajos Horváth and Piotr Kokoszka. *Inference for functional data with applications*. Vol. 200. Springer Science & Business Media, 2012.

[16] Gareth James. "Gareth James" (July 10, 2017). URL: `http://www-bcf.usc.edu/~gareth/research/Research.html`.

[17] Gareth M James and Trevor J Hastie. "Functional linear discriminant analysis for irregularly sampled curves". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.3 (2001), pp. 533–550.

[18] Gareth James et al. *An introduction to statistical learning*. Vol. 112. Springer, 2013.

[19] Jun Li, Juan A Cuesta-Albertos, and Regina Y Liu. "DD-classifier: nonparametric classification procedure based on DD-plot". *Journal of the American Statistical Association* 107.498 (2012), pp. 737–753.

[20] Regina Y Liu, Jesse M Parelius, Kesar Singh, et al. "Multivariate analysis by data depth: descriptive statistics, graphics and inference,(with discussion and a rejoinder by Liu and Singh)". *The annals of statistics* 27.3 (1999), pp. 783–858.

[21] Michael Phillips et al. "Volatile organic compounds in breath as markers of lung cancer: a cross-sectional study". *The Lancet* 353.9168 (1999), pp. 1930–1933.

[22] Gina-Maria Pomann, Ana-Maria Staicu, and Sujit Ghosh. "A two-sample distribution-free test for functional data with application to a diffusion tensor imaging study of multiple sclerosis". *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 65.3 (2016), pp. 395–414.

[23] J. O. (James O.) Ramsay and B. W. Silverman. *Functional data analysis*. eng. 2nd ed.. Springer series in statistics. New York: Springer, 2005. ISBN: 9780387227511.

[24] Ying Sun and Marc G Genton. "Functional boxplots". *Journal of Computational and Graphical Statistics* 20.2 (2011), pp. 316–334.

[25] National Lung Screening Trial Research Team. "Reduced lung-cancer mortality with low-dose computed tomographic screening". *New England Journal of Medicine* 365.5 (2011), pp. 395–409.

[26] Y Toyoda et al. "Sensitivity and specificity of lung cancer screening using chest low-dose computed tomography". *British journal of cancer* 98.10 (2008), p. 1602.

[27] Shahid Ullah and Caroline F Finch. "Applications of functional data analysis: A systematic review". *BMC medical research methodology* 13.1 (2013), p. 43.

[28] DO Wilson et al. "Reduced lung-cancer mortality with CT screening". *N Engl J Med* 2011.365 (2011), pp. 2035–2038.

[29] Baolin Wu et al. "Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data". *Bioinformatics* 19.13 (2003), pp. 1636–1643.

[30] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. "Functional data analysis for sparse longitudinal data". *Journal of the American Statistical Association* 100.470 (2005), pp. 577–590.

[31] Piotr Zalicki and Richard N Zare. "Cavity ring-down spectroscopy for quantitative absorption measurements". *The Journal of chemical physics* 102.7 (1995), pp. 2708–2717.