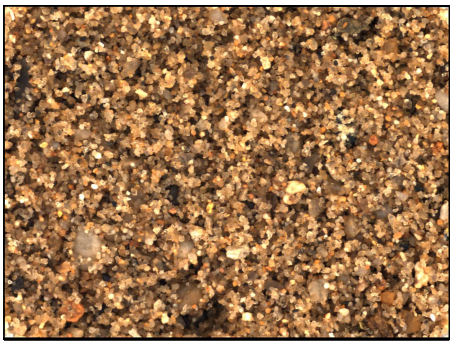
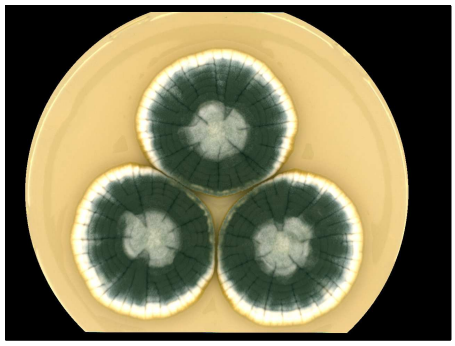
Estimation and Classification through Regression with Variable Selection amongst Features Extracted from Multi-Spectral Images

Estimation of moisture content in sand
&
Identification of *Penicillium* fungi





Line Harder Clemmensen

supervisor Bjarne K Ersbøll

IMM

IMM-Master Thesis-2006-12
Technical University of Denmark

Preface

This report documents a 30 ECTS (European Credit Transfer System) credits master thesis at the image analysis group, IMM (Informatics and Mathematical Modeling), DTU (Technical University of Denmark).

Data used in the project consists of multi-spectral images of sand samples and of Penicillium fungi. The images have 9 and 18 spectra, respectively, which run from ultra blue to infra red.

The aim is to classify three species of *Penicillium* fungi and estimate the moisture content in sand samples. For this purpose, regression methods that reduce the dimensions of data are investigated. The dimensions must be reduced, by projections or exclusion of variables, since the number of variables extracted from the multi-spectral images is much larger than the number of observations. Furthermore, model selection methods that reduce the dimensions and perform regression in one step are of interest.

The general framework of the project is multivariate statistics, pattern classification and digital image analysis. It is assumed that the reader has a basic knowledge of the three areas.

Lyngby, February 2006

Line Harder Clemmensen

Acknowledgements

I am grateful to my supervisor Associate Professor Bjarne Kjær Ersbøll for his support and advice throughout the work. I would like to thank Dr Michael Edberg Hansen for his encouragement and motivation through our many discussions about the project. I also thank Professor Jens Christian Frisvad for his inputs to the project related to the mycology.

Without the approval and support of the SCC-consortium, the analyses of the sand data would not have been possible. Four institutions have been involved in the gathering of the sand data: Danish Technological Institute, 4K-Beton A/S, Videometer A/S, and IMM, DTU.

Furthermore, I would like to thank Associate Professor Jens Michael Carstensen for his interest in the project, in particular in relation to the digital image analysis. Likewise, I thank Associate Professor Rasmus Larsen and his PhD student Karl Skoglund for their interest in the project, in particular in relation to the model selection methods. Finally, I would like to thank Dr Charlotte Bech for her giving conversations about work performance.

Abstract

This report deals with identification of three different species of *Penicillium* fungi and estimation of moisture content in sand used to make concrete. Multi-spectral images of 9 or 18 bands are used to analyze samples of sand and fungi, respectively. The project covers the image acquisition of the samples, the identification of *Regions Of Interest* (ROIs) in the images, the feature extraction from the ROIs, and classification or estimation based on the extracted features. The number of features extracted is much larger than the number of observations and the dimensionality is therefore a big issue in the analysis of the data. Traditional multivariate, statistical methods for variable selection, decomposition, classification, and regression are compared to newer methods that select variables and/or perform coefficient shrinkage within the regression. Dummy variables are constructed to use the newer methods for classification.

Chapter 1 is an introduction to problems of many variables in relation to the number of observations. The idea behind methods used in this project to solve such problems is also described. In addition to that the chapter motivates an objective identification of *Penicillium* fungi and an estimation of moisture content in sand used to make concrete. Finally, a problem formulation of the project is given, as well as a disposition of the report.

Chapter 2 gives the mathematical notation used throughout the report and briefly describes the subjects the reader is assumed to have knowledge of.

Chapter 3 describes the three species of *Penicillium* fungi, the inoculation of fungal isolates, and the design of the experiment.

Chapter 4 describes the sampling of sand, the reference measurements of moisture content, and the design of the experiment.

Chapter 5 describes the acquisition of the multi-spectral images of both fungi and sand samples.

Chapter 6 introduces the methods used in this project. The first section describes two methods for segmenting the fungal colonies in the images of the fungi samples. The second section reviews the traditional multivariate, statistical methods for regression and classification of problems with many variables in relation to the number of observations. The third section introduces the newer methods to deal with these problems. Finally, the fourth section describes additional features to these methods.

Chapter 7 states the results of the pre-processing. Here in the reproducibility of the images over time, the segmentation of the fungal colonies in the images of fungi, and the feature extraction from the ROIs of both fungi and sand images.

Chapter 8 describes and discusses the results obtained analyzing the fungi data. *Discriminant Analysis* and LARS-EN with dummy variables are compared for the classification of the three *Penicillium* species. Mahalanobi's *distance between species* and Hotelling's T^2*-test*, detecting differences in means, are calculated. Finally, several tests are calculated determining the significance of additional information provided by each medium to the discrimination.

Chapter 9 describes and discusses the results obtained analyzing the sand data. *Forward Selection* of original variables and of principal components are compared to the Ridge regression, Lasso, and LARS-EN methods.

Chapter 10 concludes upon the results obtained. The fungi are identified with low error rates using two to three variables on just one medium. The distances between species reflect the visual appearance, and all means differ significantly. The Discriminant Analysis is more robust and performs slightly better than LARS-EN with dummy variables, but LARS-EN is computationally much faster. The newer methods yield lower standard deviations than the traditional for the estimation of moisture content in sand.

Chapter 11 discusses future work in relation to this project.

Resumé

Denne rapport omhandler identifikation af tre arter af *Penicillium* svampe og estimering af fugtindholdet i sand brugt til beton. Multispektrale billeder med 9 eller 18 bånd er anvendt til at analysere prøver af henholdsvis sand eller svampe. Projektet dækker billedoptagelsen af prøverne, bestemmelse af *Regioner af Interesse* (ROIs) og konstruktionen af features fra ROIs. Antallet af variable er meget større end antallet af observationer og dimensionaliteten er derfor et vigtigt emne i dataanalysen. Traditionelle, multivariate, statistiske metoder til variabel selektion, dekomposition, klassifikation og regression sammenlignes med nyere metoder, der laver variabel selektion og/eller parameter shrinkage sammen med regression. Dummyvariable konstrueres, så de nyere metoder kan anvendes til klassifikation.

Kapitel 1 er en introduktion til problemer med mange variable i forhold til antal af observationer. Ydermere beskrvies ideen bag de metoder, som i dette projekt anvendes til at løse sådanne problemer. Kapitlet motiverer desuden identifikation af *Penicillium* svampe og estimering af fugtindhold i sand. Afslutningsvis gives en problemformulering og en disposition for rapporten.

Kapitel 2 giver den matematiske notation brugt i rapporten og beskriver kort de emner som læseren antages at have kendskab til.

Kapitel 3 beskriver tre forskellige arter af *Penicillium* svampe, podning af svampeisolater og eksperimentets design.

Kapitel 4 beskriver prøvetagning af sand, reference mål af fugtindhold og eksperimentets design.

Kapitel 5 beskriver billedoptagelserne af multispektrale billeder af både mikrobiologiske svampe og sandprøver.

Kapitel 6 introducerer metoder som benyttes i dette projekt. Første afsnit beskriver

to metoder til at segmentere svampekolonier i billeder af mikrobiologiske svampe. Andet afsnit opfrisker traditionelle multivariate, statistiske metoder til regression og klassifikation af problemer med mange variable i forhold til observationer. Tredje afsnit introducerer nyere metoder, som behandler disse problemer, og fjerde afsnit beskriver yderligere egenskaber ved disse metoder.

Kapitel 7 beskriver resultater af præprocesseringen. Herunder genskabelsen af billeder over tid, segmentering af svampekolonier og konstruktion af features fra ROIs i både svampe- og sandbilleder.

Kapitel 8 beskriver og diskuterer resultaterne fra analyserne af svampedata. *Diskriminant Analyse* og *Least Angle Regression - Elastic Net* (LARS-EN) med dummyvariable sammenlignes til klassifikation af de tre *Penicillium* arter. Mahalanobis afstand mellem arter og Hotellings T^2-test af forskel i middelværdi beregnes. Endelig udføres tests af signifikans af yderligere bidrag til diskrimination fra hvert medium.

Kapitel 9 beskriver og sammenligner resultaterne fra analyserne af sanddata. *Forward Selection* af originale variable og af principale komponenter sammenlignes med de nyere Ridge regressions-, Lasso- og LARS-EN metoder.

Kapitel 10 konkluderer på de opnåede resultater. De mikrobiologiske svampe klassificeres med en lav fejlrate for to til tre variable fra kun et medium. Afstandene mellem arter reflekterer den visuelle fremtoning af prøver og alle middelværdier er signifikant forskellige. *Diskriminant Analyse* er mere robust og giver en anelse bedre resultater end LARS-EN med dummyvariable, men LARS-EN er beregningsmæssigt hurtigere. De nyere metoder giver lavere standardafvigelser end de traditionelle ved estimering af fugtindhold i sandprøver.

Kapitel 11 diskuterer fremtidigt arbejde i forbindelse med dette projekt.

Contents

List of Abbreviations

Abbreviation	Full description
DA	Discriminant Analysis (variable from)
CV	Cross-Validation
CYA	Czapeck Yeast extract Agar
DTU	Technical University of Denmark
EN	Elastic Net (variable from)
EVD	Eigen Value Decomposition
GLM	General Linear Model
HP	Histogram Pursuit
IBT	Industrial Bio-Test Laboratories
IMM	Informatics and Mathematical Modelling
LARS	Least Angle Regression
LARS-EN	Least Angle Regression - Elastic Net
Lasso	Least Absolute Shrinkage and Selection Operator
Mel	Melanoconidium
MSE	Mean Squared Error
OAT	Oatmeal agar
OLS	Ordinary Least Squares
P.	Penicillium
PC	Principal Component
PCA	Principal Component Analysis
PP	Projection Pursuit
Pol	Polonicum
RGB	Red Green Blue
ROI	Region Of Interest
RSS	Residual Sums of Squares
SPC	Sparse Principal Component
SS	Sums of Squares
SVD	Singular Value Decomposition
Ven	Venetum

YES Yeast Extract Sucrose agar

Chapter 1

Introduction

Traditional multivariate, statistical methods are adequate in situations with few variables in relation to the number of observations. Unfortunately, the same methods are not applicable in most cases where the situation is reversed, i.e. there are more variables than observations.

This project concerns problems where the number of variables is much larger than the number of observations. Such problems often arise when digital images are analyzed. The number of pixels and the number of features extracted to characterize one observation is often large, the number increases if images of more spectra than the usual RGB are examined.

Previously such problems have been solved, successfully, by combining data compression techniques, e.g. Principal Components and Factor Analysis, with a subsequent method of analysis such as t-tests, Discriminant Analysis etc. Furthermore, cross-validation has proven advantegous in regard to variable selection, cf. [Conradsen 2002*b*], [Skettrup 2003], and [Hastie, Tibshirani & Friedman 2001].

Recently, methods have been suggested which integrate the data compression and variable selection in one step. These will be investigated and compared to the well known methods.

Two sets of data will be examined; multi-spectral images of sand samples and multi-spectral images of *Penicillium* fungi. In the first case the aim is to estimate the moisture content of the sand samples based on the images. In the second case the aim is to classify the *Penicillium* fungi into species. The two sets of data demand different approaches; a continuous dependent variable to estimate the moisture content of the sand and a nominal dependent class variable to identify the fungi. Consequently, the

two situations must be handled differently. In both cases, however, the dimensions of the feature space must be reduced, either by selecting a subset of features, or by using adequate projections.

The first sections of this chapter give a motivation for identifying *Penicillium* fungi into species and for estimating the moisture content in sand samples. The third section discusses *the curse of dimensionality* and hereby also motivates the use of dimension reductive methods. Finally, the fourth section sums up the problem formulation of this project.

1.1 Identification of fungi

Identification of fungi is of importance for several reasons; for a further phylogenetic study, to reveal new species or isolates to use in e.g. food or medical industries, and, recently, to substitute pesticides.

Traditionally, the identification has been performed by means of chemical and visual studies of the fungi. In the last decade digital image analysis has also been utilized for the classification, but till now it has been based on RGB images, as in [Hansen 2003]. This project will study classification by means of features derived by image analysis on multi-spectral images.

Since the dimension of data is increased by using multi-spectral images (eighteen spectra in stead of the traditional three for RGB images) it is important to consider methods which reduce the dimensionality of the feature space. In particular, because the number of observations in our case is smaller than the dimension of the feature space. The latter will be discussed further in Section 1.3.

1.2 Estimation of moisture content in sand

The sand samples considered here are used to make concrete. It is of great importance to know the moisture content of the sand in order to secure that the concrete obtains the right texture when it is mixed.

The aim of measuring the moisture content through imaging is to obtain inline registration in the mixing process. Hence, calculation issues are important and the fewer variables involved, the fewer calculations are necessary. Furthermore, there is a tendency that fewer dimensions give more robust results.

The methods presently used to measure the moisture content are fairly uncertain. Exact standard deviations are not available, as the construction companies consider this information confident.

1.3 The curse of dimensionality

When working with data in high dimensions there are several issues to consider. Briefly, these are:

Computational issues: Solutions to this problem can be increasing computational power or reducing computational complexity of the algorithms; e.g. by approximations with fewer computations. This, however, is not of major interest in this project, and will only be commented on briefly.

Sparse sampling in high dimensions: Sample size must grow exponentially with the dimension of the feature space in order to preserve the sampling density. In particular, this is a problem if the joint probability function is desired. Solutions to this can be either clustering or reduction of dimensionality. The first mentioned is particularly useful if data has high probability density in small regions, the clusters, and if the density is small elsewhere. Reduction of dimensionality can be obtained either by decomposition of data or by variable selection.

Such issues are related to as *the curse of dimensionality*, and are often seen in relation to multivariate, digital images, as in [Hilger 2001], [Conradsen 2002*b*], [Skettrup 2003], and [Windfeld 1992]. This project aims at providing regression and classification methods to model the high dimensional data obtained and, simultaneously, reduce the dimensionality.

The consequences of a sparse sampling in high dimensions are the following. One, that all observations are close the boundaries of the data set, making prediction difficult. Two, that in order to analyze a small percentage of data, we will have to cover a large percentage of the range of the variables, making local analyses practically impossible. These two consequences will in the following be quantified.

Given n uniformly distributed observations in a p-dimensional unit sphere centered at origin, according to Hastie[1], the median distance from the origin of the feature space to the closest data point in data sets of these dimensions is given by

$$d_{median}(p,n) = (1 - \frac{1}{2}^{1/n})^{1/p} \quad . \tag{1.1}$$

[1] [Hastie et al. 2001, Sec. 2.5]

As will be described later, the data sets examined in this project consist of 36 observations, or from 9 to 59 observations ($n = 36 \vee n = 9, ..., 59$), and 3754 or 2016 features ($p = 3754 \vee p = 2016$). For the fungi, we have $d_{median}(36, 3754) = 0.999$, and for the sand samples the distances are $d_{median}(9, 2016) = 0.999$ to $d_{median}(59, 2016) = 0.998$. Consequently, the median of the distance to the nearest point will cover all but 0.1-0.2% of the distance the boundary. Hence, the majority of data points is closer to the boundary of the sample space than to any other data point, making prediction much more difficult. It is necessary to extrapolate from the neighbor samples rather than interpolate to obtain predictions.

In the following we suppose that data is enclosed in a p-dimensional hypercube. When we want to analyze a fraction f of the observations, which corresponds to a fraction f of the unit volume, the expected edge length of a hypercube that encloses that fraction of the observations will be

$$e_p(f) = f^{1/p} \quad . \tag{1.2}$$

In our case we have that $e_{3754}(0.01) = 0.999$ and $e_{2016}(0.01) = 0.998$. So, in order to analyze 1% of data in any of the data sets we must cover more than 99% of the range of each of the input variables. An analysis of 1% of data is meant to be local, but a neighborhood covering 99% of the range of the input variables cannot be be considered local.

1.4 Problem formulation and disposition

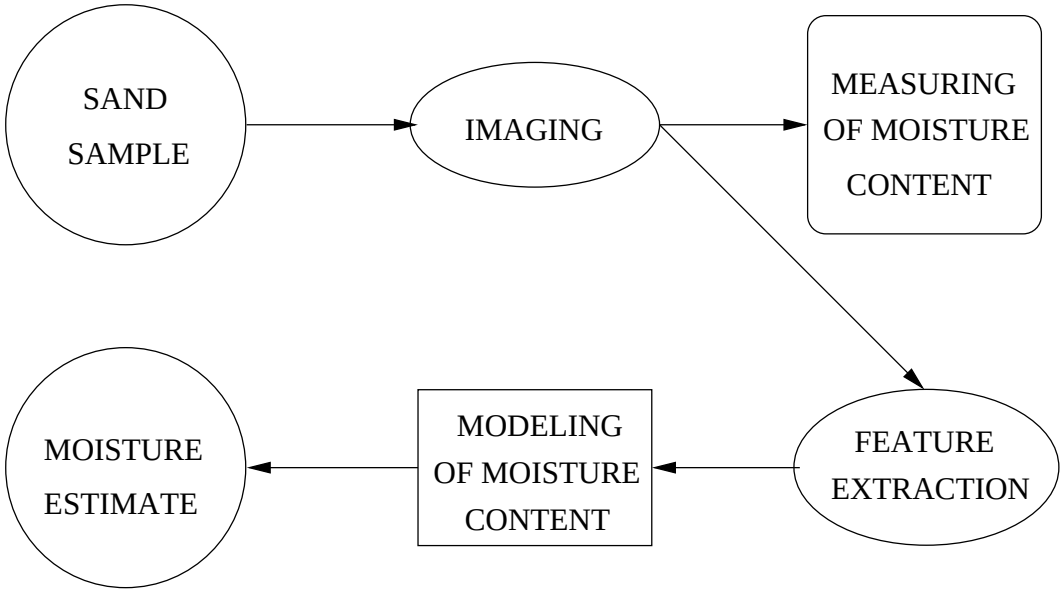
The aim of this project is to examine newer model selection methods to model high dimensional data with few observations relative to the number of variables.

Two problems are desired solved:

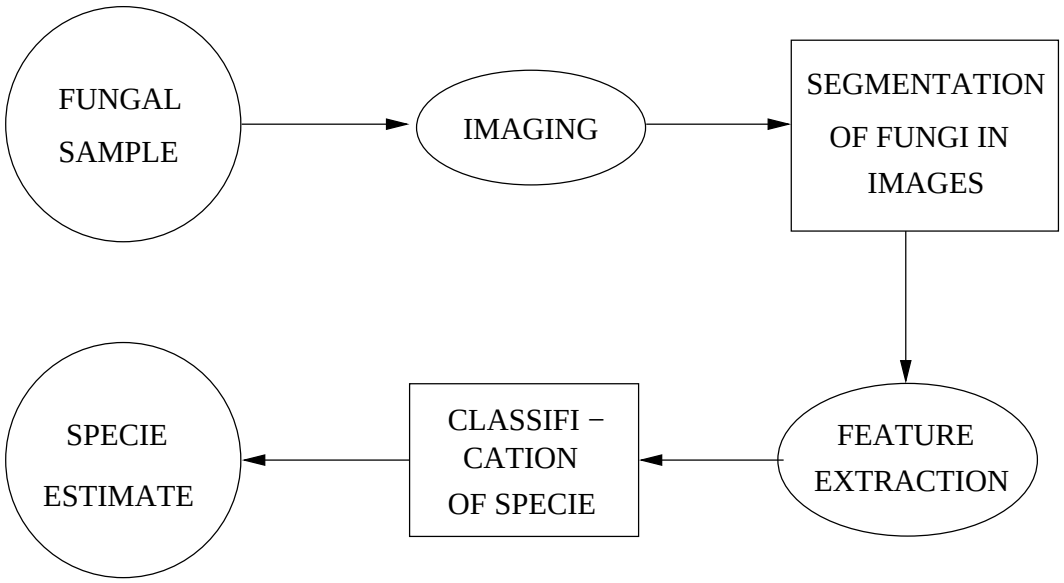
(a) A regression problem where it is of interest to estimate the moisture content in sand samples used for mixing concrete.

(b) A classification problem where it is of interest to find an objective method to classify three fungal species of the *Penicillium* genus.

In order to obtain an inline approach for the concrete mixing, and an objective method for classifying the fungi, image analysis is used. Multi-spectral images of samples are acquired and features are extracted from these images. In the images of the fungi it is necessary to first segment the fungal colonies before features are extracted from the



(a) Sand



(b) Fungi

Figure 1.1: Diagrams of the flow of the data in the two problems; estimation of the moisture content in the sand samples and classification of the fungi samples. Squares indicate that methods explained in Chapter 6 are used. Ellipses either indicate the digitalization of the samples by imaging, or feature extraction from the images. The circles are the input samples in petri dishes and output estimates related to the samples.

images. The features are then used as data sets in the regression and classification, respectively. Flow diagrams of these processes are illustrated in Figure 1.1.

The sampling steps are explained in Chapter 3 and 4. The digitalization of the samples to multi-spectral images are explained in Chapter 5. The segmentation of fungi in the images and the extraction of features from the regions of interest in the images is explained in Chapter 7. Results of the analyses, modeling, and classification of data are given in Chapter 8 and 9.

Chapter 2

Reading This Report

It is assumed that the reader has a basic knowledge of the three areas: multivariate statistics, pattern classification, and digital image analysis. The flow of data illustrated in Section 1.4, Figure 1.1, can be helpful to keep in mind while reading the report.

In next section the notation used throughout this report is listed.

2.1 Mathematical Notation

Scalars are lower case italic letters, as:

$$a \in \mathbb{R} \quad .$$

Vectors are denoted by italic lower case letters in bold, and are by default column vectors

$$\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T,$$

where ^{T} indicates transposed and n is used to denote the number of observations.

Matrices are denoted by italic upper case letters in bold, such as

$$\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p] \quad ,$$

where $\boldsymbol{X}_i$ is the ith column of the matrix $\boldsymbol{X}$, and p is used to denote the number of variables.

The 2-norm is notated and defined by

$$\|\boldsymbol{x}\|_2 = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2} \quad ,$$

and the 1-norm by

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i| \, ,$$

where $|x_i|$ is the absolute value of x_i.

The determinant of a matrix is denoted

$$\det(\boldsymbol{X}) \quad .$$

The covariance between two vectors is defined as

$$\mathrm{Cov}(\boldsymbol{X}_i, \boldsymbol{X}_j) = (\boldsymbol{X}_i - \mu_i)^T (\boldsymbol{X}_j - \mu_j) \quad ,$$

where the estimate of the mean μ_i is $\hat{\mu}_i = \frac{1}{n} \sum_{k=1}^{n} X_{ki}$, the mean of the ith variable with X_{ki} as the kth element in vector $\boldsymbol{X}_i$. The mean is also denoted $\bar{\boldsymbol{X}}_i$. The covariance matrix is

$$\mathrm{Cov}(\boldsymbol{X}) = \begin{bmatrix} \mathrm{Cov}(\boldsymbol{X}_1, \boldsymbol{X}_1) & \mathrm{Cov}(\boldsymbol{X}_1, \boldsymbol{X}_2) & \dots & \mathrm{Cov}(\boldsymbol{X}_1, \boldsymbol{X}_n) \\ \mathrm{Cov}(\boldsymbol{X}_2, \boldsymbol{X}_1) & \mathrm{Cov}(\boldsymbol{X}_2, \boldsymbol{X}_2) & \dots & \mathrm{Cov}(\boldsymbol{X}_2, \boldsymbol{X}_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(\boldsymbol{X}_n, \boldsymbol{X}_1) & \mathrm{Cov}(\boldsymbol{X}_n, \boldsymbol{X}_2) & \dots & \mathrm{Cov}(\boldsymbol{X}_n, \boldsymbol{X}_n) \end{bmatrix} \quad .$$

The correlation between two vectors is defined as

$$\mathrm{Corr}(\boldsymbol{X}_i, \boldsymbol{X}_j) = \frac{\mathrm{Cov}(\boldsymbol{X}_i, \boldsymbol{X}_j)}{\sqrt{\mathrm{Cov}(\boldsymbol{X}_i, \boldsymbol{X}_i)\,\mathrm{Cov}(\boldsymbol{X}_j, \boldsymbol{X}_j)}} \quad ,$$

and the correlation matrix denoted

$$\mathrm{Corr}(\boldsymbol{X}) \quad .$$

Chapter 3
Fungi Data

3.1 Genus

The genus *Penicillium* is a filamentous fungus also known as mold. *Penicillium* is one of the most important fungal genera, as some of its species produce important drugs (e.g. penicillin and compactin) and other species are used in food fermentation (e.g. white cheeses, P. *camemberti*; blue cheeses, P. *roqueforti* and mold fermented salami, P. *nalgiovense*) [Samson, Seifert, Kuijper, Houbraken & Frisvad 2004]. However, there also exist species that deteriorate foods and other materials. Hence, in order to prevent this, accurate identification is very important [Pitt 1979, Frisvad & Samson 2004]. Unfortunately, identification to species level in the genus *Penicillium* is very difficult because of minute differences in conidium (spore) colors, diffusible pigments, exudates, droplets and texture [Frisvad 2006]. The recording of these features are rather subjective [Samson & Frisvad 1993, Christensen, Miller & Tuthill 1994] and objective methods are needed [Dorge, Carstensen & Frisvad 2000]. Due to the large interest in the *Penicillium* genus the knowledge of the species is large and well identified isolates exist which gives an accurate ground truth for the classification in this project.

3.2 Species

Three species of the *Penicillium* genus are investigated here: P. *polonicum* (pol), P. *venetum* (ven), and P. *melanoconidium* (mel). The three species are all in the section *Viridicata* [Frisvad & Samson 2004] but belong to different series.

P. *melanoconidium* habits grains such as wheat, rye, oat, rice, and barley. Hence, it is most commonly found in cereals. It may produce penicillic acid, verrucosidin, xanthomegnin and viomellein vioxanthin [Samson & Frisvad 2005*a*]. It is one of the *Penicillium* species that has the most pure green colors *en masse* in the genus and is of the series *Viridicatum* [Frisvad & Samson 2004].

P. *polonicum* is a common mold on dry-cured meat products. Also, it habits wheat, barley, rice, rye, oat, rice, corn, peanuts, onions, and vegetable field soil [Samson & Frisvad 2005*b*]. It is able to produce verrucosidin a potent neurotoxin [Nunez, Diaz, Rodriguez, Aranda, Martin & Asensio 2000]. Furthermore, it may produce penicillic acid and nephrotoxic glycopeptides. It is typically the *Penicillium* specie with the largest amount of blue in the conidium color *en masse* and is of the series *Cyclopium* [Frisvad & Samson 2004].

P. *venetum* is commonly found in soil decaying vegetation as onions and flower bulbs and is therefore ecologically different from the cereal-borne members of the *Viridicata* section. It is rare on foods, but is known to produce the mycotoxin Roquefortine C. [Samson & Frisvad 2005*c*]. It has blue green conidia *en masse* and is of the series *Corymbifera* [Frisvad, Smedsgaard, Larsen & Samson 2004].

The striking color difference between P. *melanoconidium* and P. *polonicum* is illustrated in [Raper & Thom 1949, page 428a] in one of the few color pictures in their 1949 monograph on *Penicillium*. Superficially P. *polonicum* and P. *venetum* could look like they were the most closely related, but it is in fact P. *polonicum* and *P. melanoconidium* that are the most closely related. Any data that can show this fact would be of interest, though, as the images used in this project mainly capture the appearance in color this is not likely.

Furthermore, all species produce different mycotoxins, a list of the natural products produced by the three species examined here can be found in Appendix B. Hence, an objective method that can separate these three important species, and allow identification based on objective image analysis, is highly desirable.

3.3 Samples

Three species of the *Penicillium* genus were chosen. Two with similar appearance (P. *polonicum* and P. *venetum*) and a third (P. *melanoconidium*) with visually distinct appearance from the other two. This is done to investigate the performance of the image based classification, both when the differences should be obvious and when they should not. For each specie 4 isolates were chosen that represent a wide geographical range. The fungal isolates were obtained from the IBT Culture Collection held at

BioCentrum-DTU, Technical University of Denmark. The IBT numbers of the species are listed in Table 3.1.

Isolate/Specie	P. *melanoconidium*	P. *polonicum*	P. *venetum*
a	IBT 3445	IBT 22439	IBT 23039
b	IBT 21534	IBT 15982	IBT 21549
c	IBT 3443	IBT 14320	IBT 16215
d	IBT 10031	IBT 11383	IBT 16308

Table 3.1: IBT numbers of the *Penicillium* isolates.

The isolates were inoculated on three different media: CYA (Czapeck Yeast extract Agar), YES (Yeast Extract Sucrose Agar), and OAT (Oatmeal agar) and with three replica on each medium. In total this results in $3\ species \times 4\ isolates \times 3\ media \times 3\ replica = 108\ samples$. An overview of the experimental design is seen in Table 3.2.

Specie	P. *polonicum*				P. *venetum*				P. *melanoconidium*			
Medium/ isolate	a	b	c	d	a	b	c	d	a	b	c	d
CYA	×3	×3	×3	×3	×3	×3	×3	×3	×3	×3	×3	×3
YES	×3	×3	×3	×3	×3	×3	×3	×3	×3	×3	×3	×3
OAT	×3	×3	×3	×3	×3	×3	×3	×3	×3	×3	×3	×3

Table 3.2: Overview of the experimental design.

3.4 Inoculation

The inoculation has been conducted at BioCentrum at the Technical University of Denmark. The 12 isolates have been grown beforehand in order to produce the necessary spores. The isolates have been inoculated as three point cultures, i.e. the aim has been to grow the individuals in three well separated colonies. The inoculation has been performed in 9cm petri dishes containing one of the three growth substrates: YES, CYA or OAT, also referred to as media.

First step is to scrape out spores from an isolate, remembering to sterilize the scraper each time. The scraping is illustrated in Figure 3.1. During the inoculation, it is of great importance to keep the tools sterilized as the spores spread and grow easily. The scrape is then placed in a small container with water and shaken to spread the spores

(a) Sterilizing (b) Scraping

Figure 3.1: Small pieces of the grown mold are scraped and put into small containers with water. The scraping tool is sterilized using a burner.

in the water, cf. Figure 3.2. Finally, a needle is dipped in the water and pricked into the medium at three spots which will become the centers of the colonies, cf. Figure 3.3. The needle is dipped once for each isolate, and that is enough to inoculate three repetitions on each medium. The needle is, as the scraper, sterilized between each isolate.

After incubation in complete darkness for 7 days at 25°C, the cultures reach their stationary phase and are able to produce secondary metabolites. At this stage the colonies have grown into three circular objects within the petri dish and the fungal colonies can be digitized.

(a) Water with scrape before shaking

(b) Shaking

(c) Water with scrape after shaking

Figure 3.2: The water with sample scrape is shaken to spread the spores in the water.

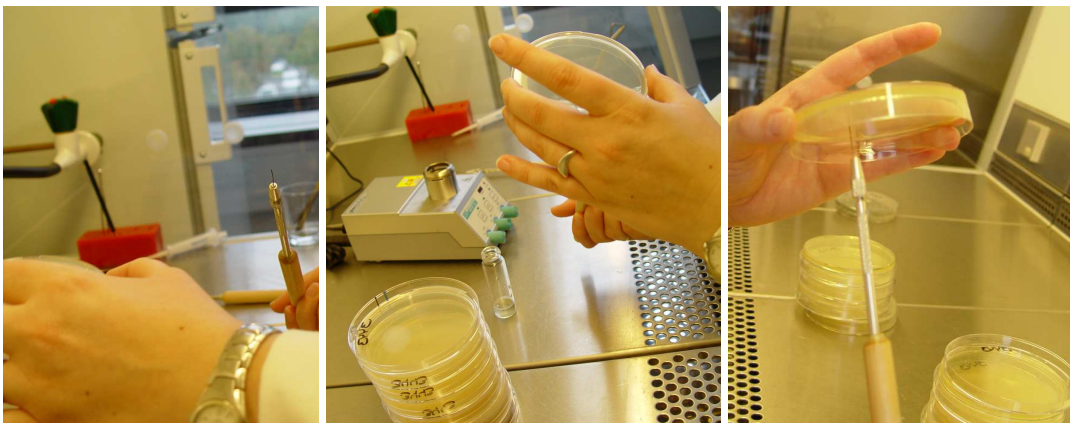


Figure 3.3: The media are inoculated using a needle that is first dipped in the water with spores and then pricked into the medium in three spots. In the three images the inoculation is seen from different angles.

Chapter 4

Sand Data

Five types of sand with different geographical origins have been examined in this experiment. A further description of the origin of the five sand types is listed in Table 4.1. The sand types vary in distribution of grains. Consequently, the sand is furthermore classified by grain curves reflecting the distributions of grains. A grain curve is the curve that describes the amount of sand in percent that falls through a sieving as a function of the size of the mesh in the sieve. Typically, the mesh size runs from 0 to 32mm. There are three different grain curves: fine (F), medium (M) and large (L). When the sand belongs to the fine grain curve the sand grains are small, and larger percentages of sand than the medium fall through the sieves with large meshes. When the sand belongs to the large grain curve the sand grains are large, and smaller percentages of the sand than the medium fall through the sieves with large meshes.

Type	Description	Origin
1	hill sand	Tarup Grusgrav, Nymølle Stenindustrier
2	hill material	Brejning Grusgrav
3	sea sand	Starnholmen, RN Sten & Grus
4	dry screened hill sand	Års, Hornum Murer- & Entreprenørforretning
5	dry screened hill sand	Løgstrup, Jorbomølle Grus og Sandgrav

Table 4.1: Description of the five sand types. All types are 0-4mm washed sand.

Buckets of 10L with sand and water are mixed with the aim of reaching one of eight endeavored nominal moisture levels. Three samples of small amounts of sand is then taken from each bucket and placed in petri dishes. The content of each petri dish is then imaged by a multi-spectral camera. The moisture content in each sample is measured after the imaging by placing each sample in a special oven that dries out the sample

and measures the amount of vaporized water in relation to the amount of dry sand.

The sampling is conducted so that:

- For sand type 1, 3 and 5 there are three grain curves.
- For sand type 2 and 4 there is only one grain curve, the medium.
- The experiments have been conducted with up to eight different levels of moisture content. The endeavored nominal moisture levels are 0%, 1.25%, 2.5%, 3.75%, 5%, 6.25%, 7.5% and 8.75%.
- Three to twelve repetitions were performed for each set of parameters.

An overview of the experimental design is seen in Table 4.2.

	Type	1			2			3			4			5		
	Curve	F	M	L	F	M	L	F	M	L	F	M	L	F	M	L
Moisture Level	0.00%	3	3	3	-	3	-	3	9	3	-	3	-	3	3	3
	1.25%	-	3	-	-	3	-	-	3	-	-	3	-	-	3	-
	2.50%	3	3	3	-	3	-	3	9	3	-	3	-	3	3	3
	3.75%	-	3	-	-	3	-	-	3	-	-	3	-	-	3	-
	5.00%	3	6	3	-	3	-	3	12	3	-	3	-	3	6	3
	6.25%	-	3	-	-	3	-	-	3	-	-	3	-	-	3	-
	7.50%	3	3	3	-	3	-	3	9	3	-	3	-	3	3	3
	8.75%	-	3	-	-	3	-	-	3	-	-	3	-	-	3	-
	TOTAL	12	27	12	0	24	0	12	51	12	0	24	0	12	27	12

Table 4.2: Obsevations in each group. F: fine grain curve, M: medium grain curve, and L: large grain curve.

There are 7 missing observations where the moisture content has not been measured adequately, these are listed in Table 4.3.

Type	Grain Curve	Moisture Level	Number of NaNs
3	F	0%	3
3	F	5%	1
3	M	0%	3

Table 4.3: The seven missing observations.

The samples with a moisture content of 0% are dried at over 100°C. This gives an abrupt change in appearance of the sample. Since this is not a realistic situation the samples are not included in the analyses.

To illustrate the analyzed data, the measured moisture content for each of the sand types is plotted as a function of the grain curve in Figure 4.1.

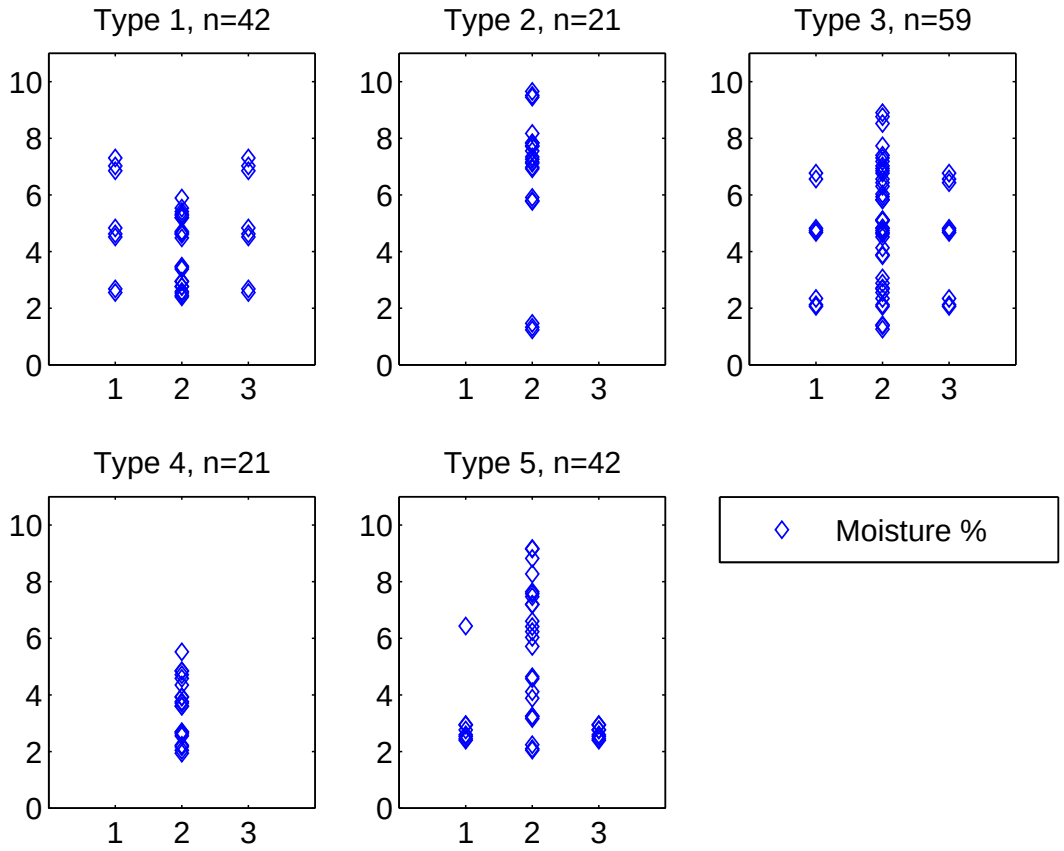
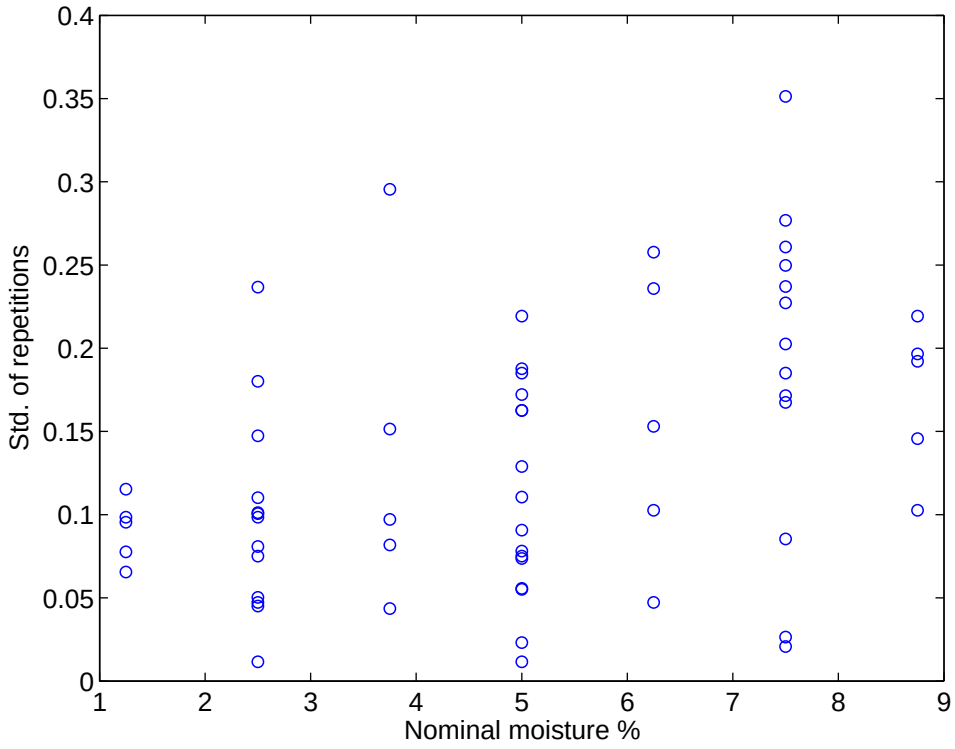
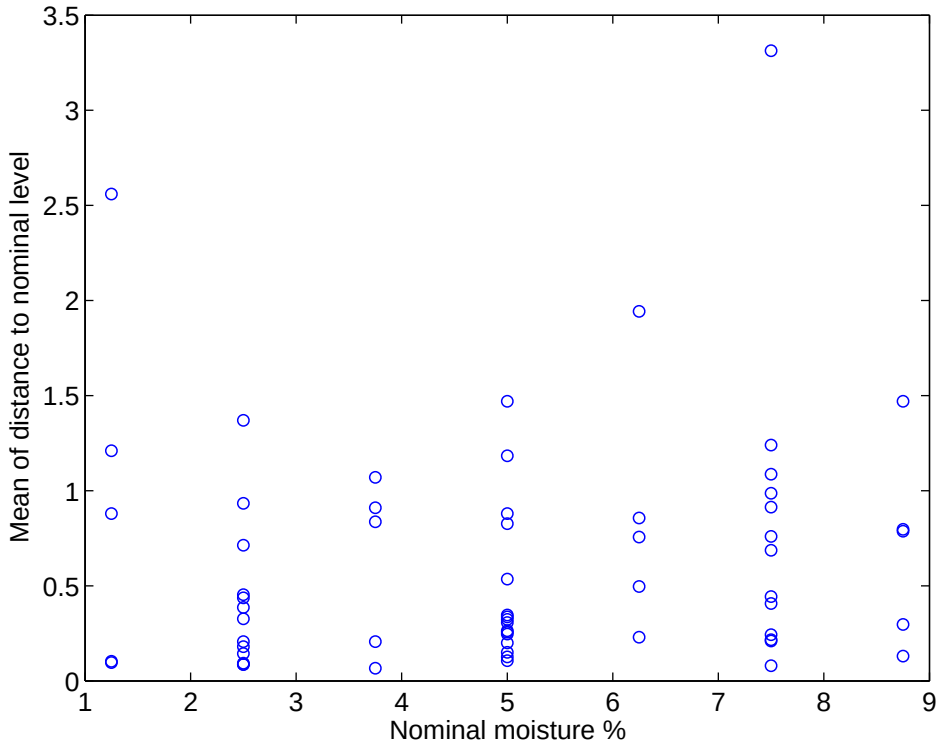


Figure 4.1: Illustration of the moisture content observations divided into groups for each grain curve and sand type. For the grain curves 1=Fine, 2=Medium, and 3=Large. Observations of 0% moisture content are left out.

There is a rather large difference, up to 3%, between the nominal moisture content levels and the measured moisture contents, cf. Figure 4.2. Furthermore, the standard deviation of the three repetitions of sand samples taken from the same bucket is up to 0.3%, cf. Figure 4.2. This indicates that the sample variation is large and that it is difficult to reach the nominal moisture contents in the buckets.



(a) Standard deviation of repetitions



(b) Mean of distance to nominal level

Figure 4.2: Standard deviation of repetitions and mean distance of repetitions to nominal level as functions of the nominal moisture level.

Chapter 5

Image Acquisition

In this chapter the digitizing of the samples is described and a conversion from the multi-spectral bands to an RGB representation is performed. The conversion to RGB is made to illustrate the appearance of the samples.

5.1 The Image system

The samples have been digitized using a multi-spectral digital camera system as seen in Figure 5.1, provided by Videometer A/S[1].

Figure 5.1: Illustration of the camera system.

[1]URL http://www.videometer.com

The camera system consists of an integrating sphere illumination (an Ulbricht sphere) combined with a two step calibration procedure, which provides a high precision and reproducibility, based on a multi-spectral camera. The inside of the sphere is covered with a matte titanium paint that ensures a diffuse and homogenous illumination of the sample. The illumination of the sample should be diffuse to avoid shadows and reflections. Light diodes are placed inside the sphere as illustrated in Figure 5.2.

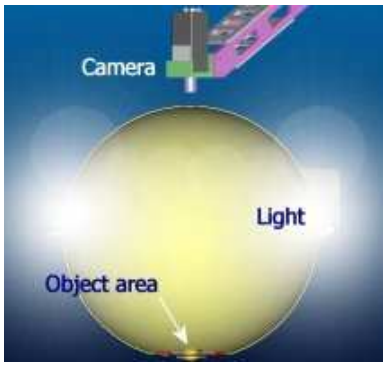


Figure 5.2: Cross section of sphere illustrating the illumination.

In order to adjust the geometric and chromatic set-ups the camera sirst calibrated. The geometric and chromatic representations in the camera may change over time due to differences in temperature, humidity etc., and the calibration should then redefine these representations. This is done by imaging of two predefined chromatic intensities (light gray and dark gray) and of a predefined geometric grid. Then, by means of numerical algorithms the images are adjusted to these conditions.

5.2 Fungi

The next step is to assure that the dynamic range is fully exploited. This is done by adjusting the light set-up through imaging of the lightest of the samples (the background should represent the lowest value in the dynamic range). The images are taken on a standard 1000 NCS sheet as background. The lid of the petri dish is removed to avoid reflections during the process, and the sphere is lowered to avoid illumination from the bottom of the sphere, as seen in Figure 5.3. Both sides of the fungi have been imaged, as illustrated in Figure 5.4. In the images of the backside the lens of the camera is reflected in the petri dish and dark shaded circles appear in these images. The information obtained from the back side could be used as additional information for classification. When samples are classified visually it is normal procedure to look

(a) Sphere & sample (b) Sphere lowered

Figure 5.3: The sphere in the process of image acquisition of a sample.

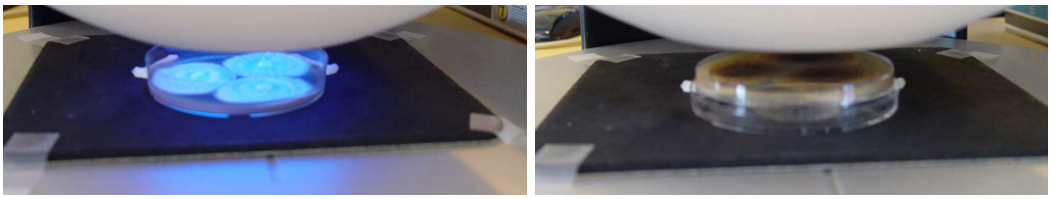
at the back side as well since the color information here is relevant. However, in this project focus has been put on the front side images.

The multi-spectral camera has constructed color intensity images for 18 different wavelengths. Hence, a multi-spectral image has 18 frames of color intensity images, each with a resolution of 960×1280 pixels. For each sample this amounts to $18 \times 960 \times 1280 \simeq 2 \cdot 10^7$ pixels in total for the 18 frames.

The 18 wavelengths used are: 430, 450, 470, 505, 565, 590, 630, 645, 660, 700, 850, 870, 890, 910, 920, 940, 950, and 970nm. The spectra represent the colors from ultra blue to infra red, see Table 5.1.

To represent the images in RGB the color-matching functions from Wyszecki[2], illustrated in Figure 5.5, have been used. The weights for R, G and B of each spectral band are chosen to represent the approximated area under the color-matching functions, this is illustrated in Figure 5.7. The weights for R, G and B are scaled to sum to one. Appendix C contains RGB images of all the samples, one of them is seen in Figure 5.6.

[2][Wyszecki & Stiles 1982]



(a) Front of sample (b) Back of sample

Figure 5.4: Imaging of front and back side of a sample.

Range (nm)	Color	Human eye
400-430	ultra violet-blue	Visible
430-460	blue	Visible
460-510	cyan	Visible
510-540	green	Visible
540-560	yellow	Visible
560-630	amber-orange	Visible
630-700	red	Visible
700-970	NIR	Not visible

Table 5.1: Description of the colors of the wavelengths. The wavelength ranges of the colors are approximate.

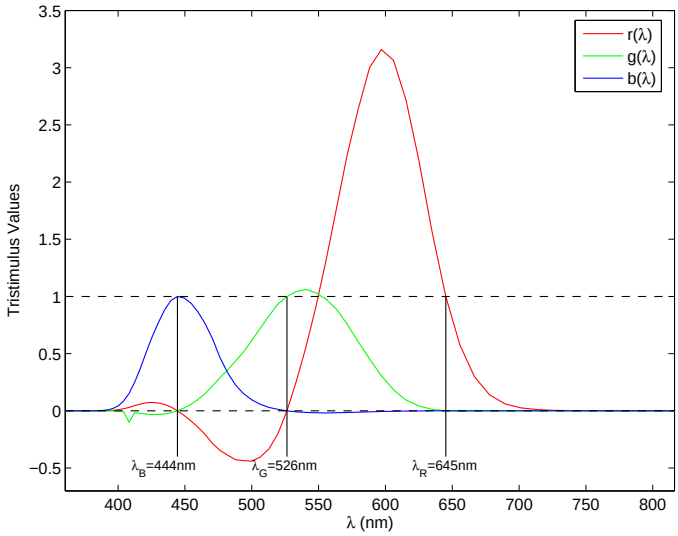


Figure 5.5: Color-matching functions of the CIE 1964 supplementary standard colorimetric observer in the system of real primary stimuli R(645.2nm), G(526.3nm) and B(444.4nm). The units of the primary stimuli are of unit radiant power.

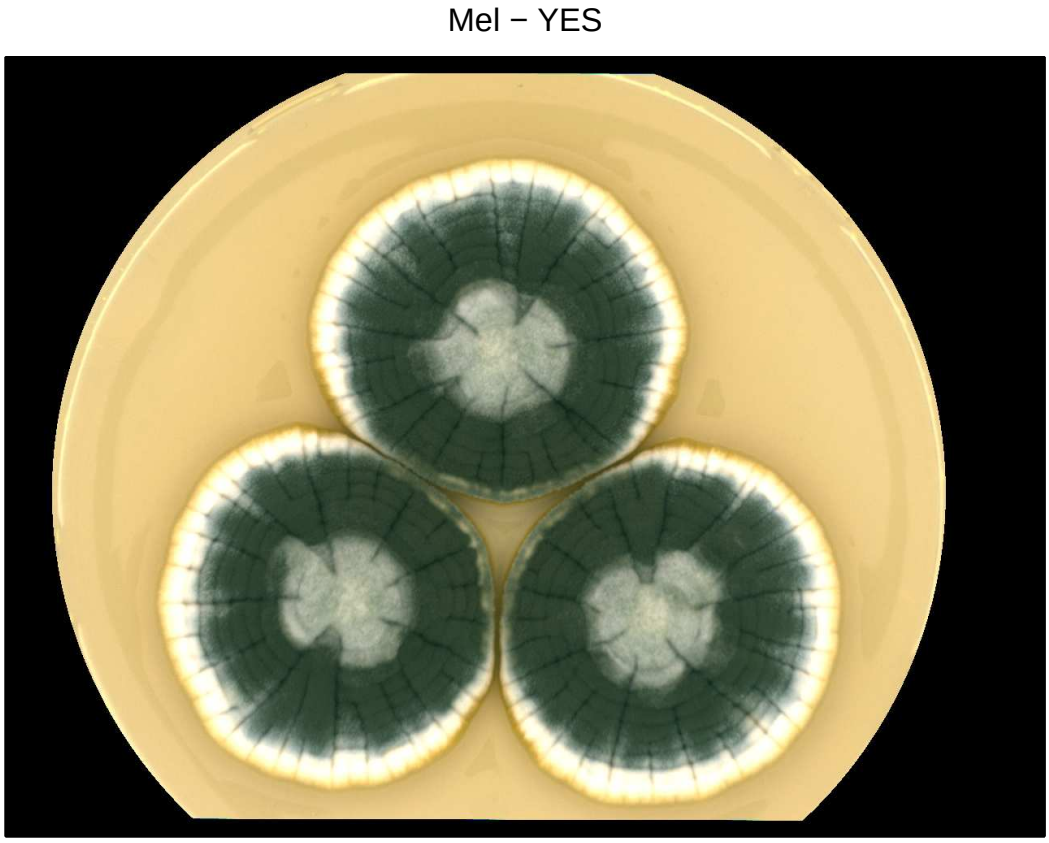
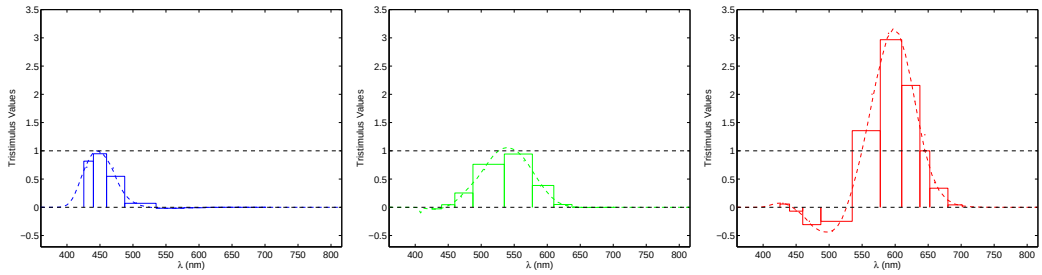


Figure 5.6: An example of one of the P. *melanoconidium* isolates on YES represented in RGB.



(a) B (b) G (c) R

Figure 5.7: Weights for the 10 spectral bands in the visual area represented by the area under the color-matching functions and later scaled to sum to one.

The intensity images of the 18 spectra are shown in Figure 5.8. Note that the wavelengths of 470nm and 505nm (cyan) are better reflected than other wavelengths in the visual area, i.e. the pixel values in the areas with fungal colonies are larger in these bands. This is in accordance with the visual appearance of the colonies, recall, that the species have green/blue conidia *en masse*.

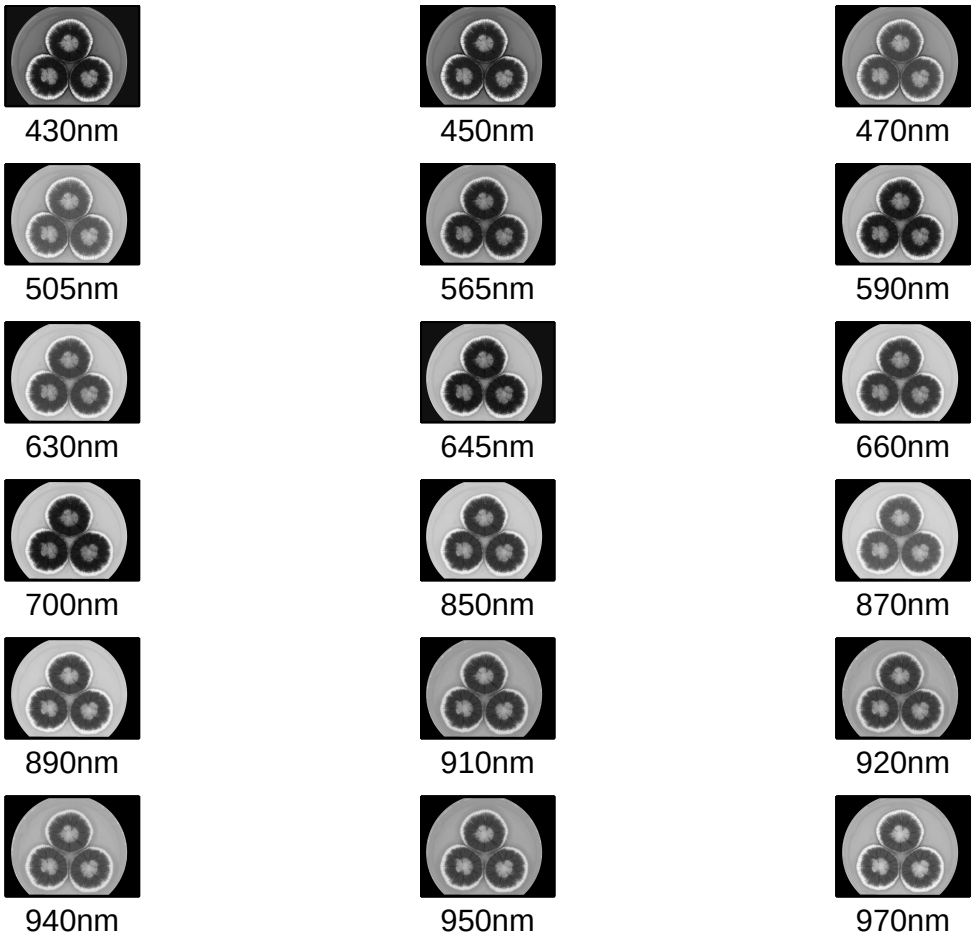
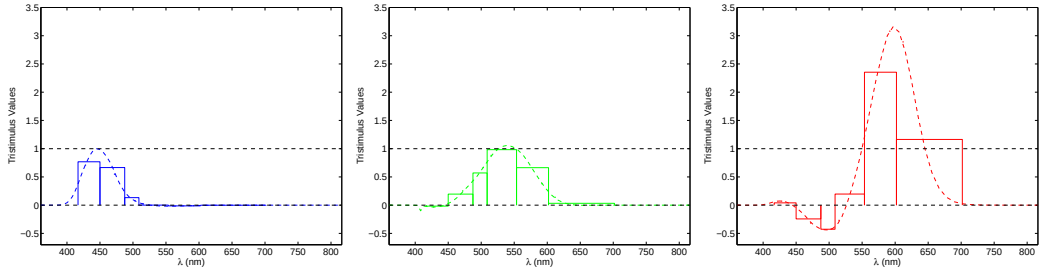


Figure 5.8: The 18 spectral bands of one of the P. *melanoconidium* isolates on YES. All images are displayed with same scale on the gray color mapping.

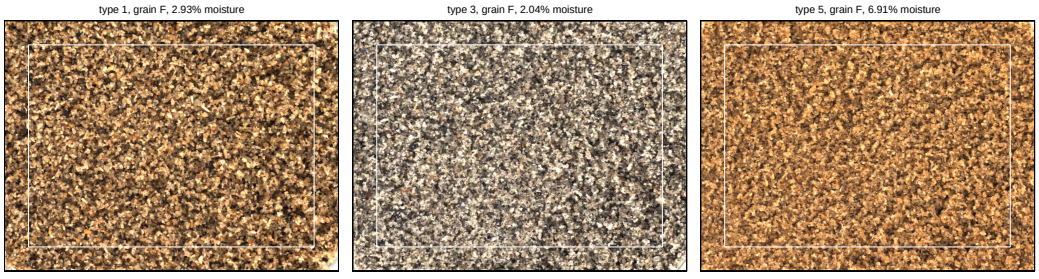
5.3 Sand

The sand samples have been imaged in the same way as the fungi samples, but only nine spectral bands have been captured. The spectra are: 428, 472, 503, 515, 592, 612, 630, 875, and 940nm. The weights of the 6 spectra in the visible area in a RGB representation are illustrated in Figure 5.9. Examples of RGB images of the sand samples for different sand types and grain curves are seen in Figure 5.10 to 5.12. In some of the sand images the background appears in the corners. *Region of Interest* (ROI) is therefore chosen to avoid including information from the background. ROI is marked with a white square.



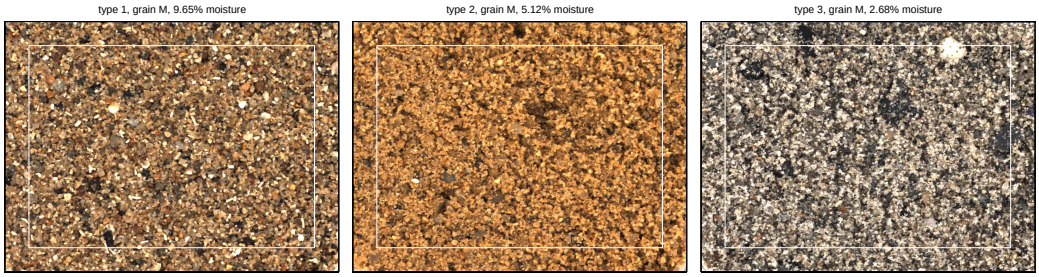
(a) B (b) G (c) R

Figure 5.9: Weights for the 6 spectral bands in the visual area represented by the area under the color-matching functions and later scaled to sum to one.

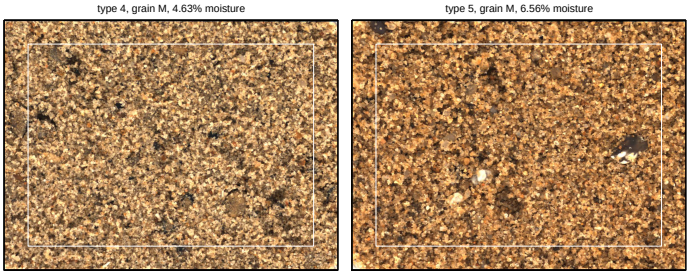


(a) Type 1 (b) Type 3 (c) Type 5

Figure 5.10: Examples of sand samples with fine grain curve. ROI is marked with a white square.

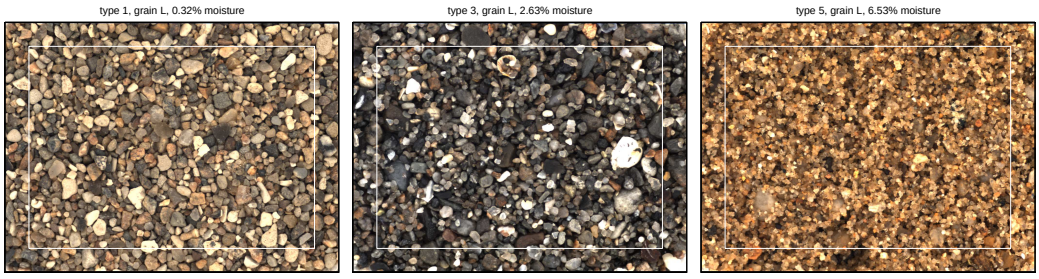


(a) Type 1 (b) Type 2 (c) Type 3



(d) Type 4 (e) Type 5

Figure 5.11: Examples of sand samples with medium grain curve. ROI is marked with a white square.



(a) Type 1 (b) Type 3 (c) Type 5

Figure 5.12: Examples of sand samples with large grain curve. ROI is marked with a white square.

Chapter 6

Methods

The first section describes two segmentation methods to segment *Regions Of Interest* (ROIs) in the images of *Penicillium* fungi. One that takes use of the geometrical shape of the fungal colonies, and another that uses information from histograms of projections of the entire multi-spectral image.

The second section walks through the traditional regression, classification, model selection, and decomposition techniques. The regression method described is *Ordinary Least Squares* (OLS). The classification method described is *Discriminant Analysis*. The model selection method described is *Forward Selection*. The decomposition method described is *Principal Component Analysis* (PCA). This section is meant as a review of these methods.

The third section introduces newer methods that join regression and model selection in one. The methods described here are: *Ridge regression*, *Least Absolute Shrinkage and Selection Operator* (Lasso), *Least Angle Regression* (LARS), LARS - *Elastic Net* (LARS-EN) and *Sparse* PCA. The description of Ridge regression and Lasso is an introduction to regression with constraints and the state of the art methods: LARS and LARS-EN. This section is meant as an introduction to these methods.

Finally, section four provides additions to the newer techniques, here in examines shrinkage problems and the use of dummy variables in order to classify via regression methods.

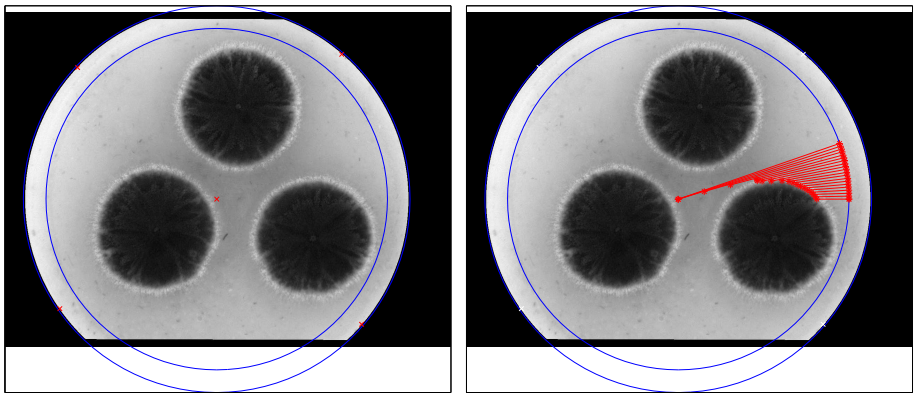
6.1 Segmentation methods

Two methods for segmenting the fungal colonies in the images are described: A method previously used to segment fungal colonies in images, and a newly developed method that previously has been used to segment lesions in images of psoriasis.

6.1.1 Identification of circular colonies

The method described in this section has previously been used in [Dorge et al. 2000] and [Hansen 2003] to segment fungal colonies in RGB images. The method assumes that the fungi have grown into three circular colonies and is based on information from one spectral band.

The intensity, separating colony from petri dish, is used directly to locate the colonies. Hence, the intensity difference between dish and colony in the band chosen should be as big as possible. First, the petri dish is found by simple edge detection from the corners of the image along the diagonals. The edge is detected in four points, as illustrated in Figure 6.1 (a), and a circle is fitted to the petri dish. A circle with same center as the petri dish but smaller radius is used for further analyses of the colonies. The smaller radius is used to avoid light reflections near the edge of the petri dish.



(a) Identification of petri dish (b) Scans to detect fungal colonies

Figure 6.1: (a): The detected edge of the petri dish is marked with four red xs. The circle fitted to the petri dish and the circle with analyzing radius are likewise plotted in red. (b): The scan lines, from the circle of analyzing radius towards the center of the petri dish detecting the fungal colonies, are marked in red.

Next, scans from the analyzing circle to the center of the petri dish are performed going counter clockwise from 0° to 360°, with one scan line for each degree. The scan is stopped when there is a change in the intensity separating dish from colony as illustrated in Figure 6.1 (b). Local minima of the distance from the detected colony to the center of the petri dish as a function of the scan angle are identified and two points on each side of a minimum are chosen to identify the edge of the colony. The four points for each colony are used to fit a circle to that colony. The center and the radius of the circle are used as identification. This process is illustrated in Figure 6.2.

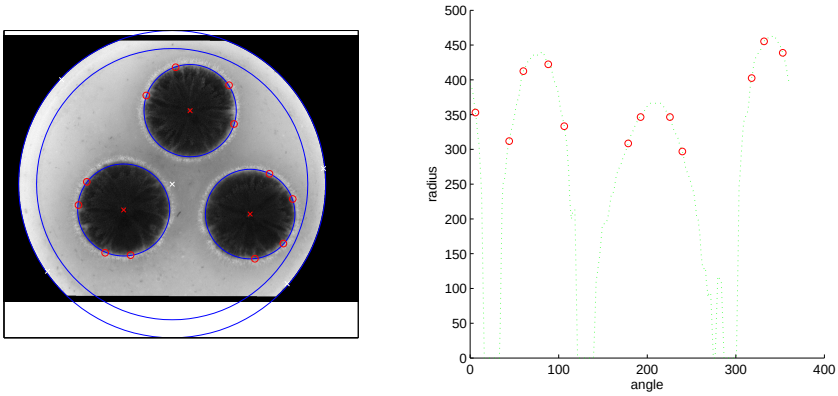


Figure 6.2: Identification of circular colonies. Left: The 6th spectral band with the circles, the centers of the fungal colonies, and the points on the edge of the colonies marked. Right: The distance from the detected colony to the center of the petri dish versus the angle of the scans.

Only segments of the colonies are used to extract features from, as the colonies are known to interact chemically when they are situated closely. The *Regions Of Interest* (ROIs) are illustrated in Figure 6.3.

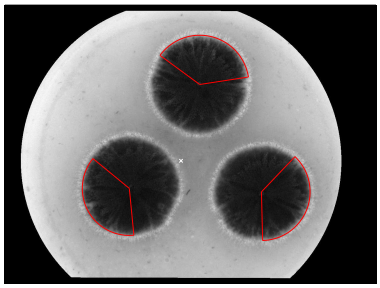


Figure 6.3: ROIs from where the features should be extracted. An angle of 135° ($\frac{3}{4}\pi$ radians) pointing away from the center of the petri dish is used.

Pros and Cons

Disadvantages: This method assumes that the colonies are circular and have a good distinction in pixel value between medium and colony. It is rare that all colonies are exactly circular of shape. The approach only makes use of one band and therefore all available information is not exploited.

Advantages: The method identifies the center of the fungal colonies and it is therefore possible to extract features according to growth direction. As the colonies grow from the center and outwards and produce different mycotoxins according to the aging, this can be useful. The aging difference can be seen from the differences between the light edges of the colonies compared to the blue/green centers of the colonies. Hence, spatial information can be included in the features. Additionally, a segment of each colony can be chosen as ROI according to geometric placement so the parts of the fungi that are almost in contact and known to be chemically interacting can be excluded.

6.1.2 Histogram Pursuit

The *Histogram Pursuit* (HP) [Gomez 2005] is an algorithm striving for bi- or multimodality in data in order to segment interesting features in data. It is built on Friedman's statistical approach to find interesting structured projections of a multivariate data set, the *Projection Pursuit* (PP) algorithm [Friedman 1987].

Projection Pursuit finds interesting structures via linear projections where the projected data differs as much as possible from the Gaussian distribution. Friedman gives four heuristic arguments for the normal distribution being the least interesting:

- The normal distribution is totally specified by mean and covariance, and we are seeking projections that can discover additional information to those captured by the correlation structure of the data.
- All projections of a multivariate normal distribution are normally distributed.
- Most linear combinations of variables will be approximately normally distributed, as indicated by the central limit theorem; sums tend to be normally distributed.
- For fixed variance, the normal distribution has the least information (Fisher, negative entropy).

In one dimension Projection Pursuit looks for a linear combination $\boldsymbol{X} = \boldsymbol{\alpha}^T \boldsymbol{Z}$, such

that the index

$$I(\boldsymbol{\alpha}) = \frac{1}{2}\sum_{j=1}^{J}(2j+1)\left[\frac{1}{N}\sum_{i=1}^{n}P_j(2\Phi(\boldsymbol{\alpha}^T\boldsymbol{z}_i)-1)\right]^2 \tag{6.1}$$

is maximized. This is the sample version of Friedman's projection index, where P_j is the Legendre polynomium of order j and $\Phi(\boldsymbol{X})$ is the standard normal density function. The PP method has previously proved to be a useful supplement to classical linear projection methods such as *Principal Component Analysis* in finding interesting views of multivariate images, cf. [Windfeld 1992].

Once an interesting projection has been found, the algorithm looks for the next informative view by removing the structure that makes the projection just found interesting and then remaximizing the projection index.

In data sets with more than two classes, or data sets with one or more non-Gaussian variables the first projection of PP may not be optimal, in the sense that the classes in the data set are not separated, and therefore require more than one projection to separate the classes. This is illustrated in the article added in Appendix A.

The Histogram Pursuit (HP) algorithm uses the same approach as PP for projecting the data, but only projections that separates the data in n classes are considered. The method takes into account the assumed number of classes in the image, and maximizes the index corresponding to the $n-1$ largest areas between consecutive modes in the histogram of the projected data. This index is given by:

$$I(H) = \sum_{j=1}^{n-1}\left(\sum_{i=x_j}^{x_{j+1}}\{\min(H_i, \min(M_j, M_{j+1}))\} - \min(M_j, M_{j+1})\cdot n_{bins}(j)\right) \quad , \tag{6.2}$$

where M_j is the j^{th} local maximum located at x_j. n_{bins} is the number of bins between the j^{th} and the $(j+1)^{\text{th}}$ maxima and H_i is the frequency of the i^{th} bin. The index is illustrated in Figure 6.4.

In order to force the algorithm to provide only projections with n modes, the algorithm gives an index of zero to all projections with a different number of modes.

Pros and Cons

Disadvantages: The centers of the fungal colonies are not identified, and hence, spatial features cannot be provided. Computationally, it is slower than the method described in Section 6.1.1.

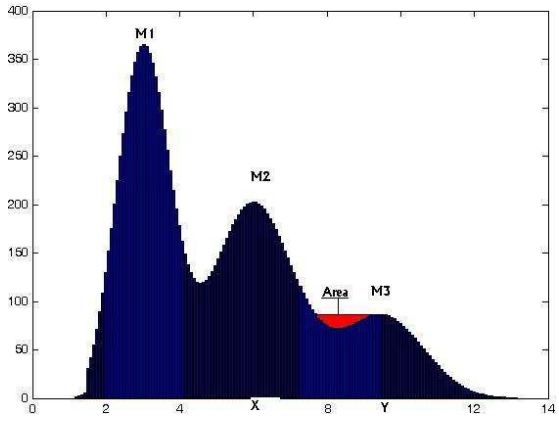


Figure 6.4: Region where HP calculates the index. Here $x = x_2$ and $y = x_3$.

Advantages: The method does not use assumptions of the shape of the colonies. This is an even larger advantage if the fungi have not grown into three colonies. Information provided by all 18 bands is utilized. Structures, such as the lighter edge of the colonies can be segmented separately, and this might give additional information in relation to the classification.

6.2 Traditional regression and classification methods

In this section regression by *Ordinary Least Squares* is discussed, *Discriminant Analysis*, and the orthonormal projection method of *Principal Components* are reviewed. Additionally, the traditional variable selection method *Forward Selection* is explained. The projection method and the variable selection method can be combined with regression and Discriminant Analysis to analyze a problem of reduced dimensions. In an inline production the variable selection can be preferred to the projection method, as only a subset of features is required. On the other hand, the projection method can include more features in reduced dimensions and can therefore contain more information which might yield better results.

6.2.1 Ordinary Least Squares

Consider the *General Linear Model* (GLM)

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad , \boldsymbol{\epsilon} \in N(0, \sigma^2) \quad . \tag{6.3}$$

The *Ordinary Least Squares* (OLS) estimates are obtained by minimizing the *Residual Sums of Squares* (RSS), i.e.

$$\hat{\boldsymbol{\beta}}_{OLS} = \text{argmin}_{\beta} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \quad . \tag{6.4}$$

For a full rank matrix $\boldsymbol{X}$ this can be solved by use of the normal equations as

$$\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \quad . \tag{6.5}$$

For normally distributed and independent residuals ϵ_i this is also known as the *Maximum Likelihood* estimator. However, this is often not good enough for two reasons:

Prediction accuracy: The OLS estimate often suffers from having a large variance, and therefore predicts poorly even though the estimate is unbiased.

Interpretation: With a large number of variables the solution can be difficult to interpret, and hence, we would like to reduce the number of variables to a subset characterizing only the strongest effects.

Traditionally, the latter problem is reduced using *Forward Selection* or *Principal Component Analysis*. The solution is often a trade off between over fitting data and including enough information to model data well.

6.2.2 Discriminant Analysis

This section briefly reviews *Discriminant Analysis* for classifying data, if more information is desired then see [Conradsen 2002*a*, Chapt. 7], [Rencher 2002, Chapt. 8] or [Hastie et al. 2001, Sec. 4.3].

The discrimination between two normally distributed populations $\pi_1 \leftrightarrow, \text{N}(\boldsymbol{\mu_1}, \boldsymbol{\Sigma})$ and $\pi_2 \leftrightarrow \text{N}(\boldsymbol{\mu_2}, \boldsymbol{\Sigma})$ is performed using the Bayes solution, i.e. minimizing the expected losses, and with equal loss the discriminant function between the two classes is given by

$$s_1 - s_2 = \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 = 0 \quad . \tag{6.6}$$

If $s_1 - s_2 > 0$ we classify the observation as belonging to π_1, and otherwise as π_2. The $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}$ are replaced by estimates based on the training data as in [Conradsen 2002*a*, Sec. 7.1.3]. A pooled estimate of the *within group sums of squares deviation* matrix $\boldsymbol{W}$, described in the next section, is used as an estimate of the dispersion matrix.

For more than two classes we can expand the two class situation from before so that class i has a discriminant scoring function of

$$s_i = \boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i - \frac{1}{2} \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i \quad . \tag{6.7}$$

We then classify an observation to be from the class with the highest score. As in the two class situation, the classes are assumed to be normally distributed and with equal dispersion.

The classification by means of Discriminant Analysis can be performed with the SAS program `proc discrim`.

Wilks' Lambda

Consider the following three *sums of squares deviation* measures for stochastic independent variables $\boldsymbol{X}_{ij} \in N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, ..., c$ and $j = 1, ..., n_i$ of c classes with $n_1, ..., n_c$ observations, respectively. The group means are denoted by $\bar{\boldsymbol{X}}_1, ..., \bar{\boldsymbol{X}}_c$. The between group sums of squares deviation matrix is defined as

$$\boldsymbol{B} = \sum_{i=1}^{c} n_i (\bar{\boldsymbol{X}}_i - \bar{\boldsymbol{X}})(\bar{\boldsymbol{X}}_i - \bar{\boldsymbol{X}})^T \quad , \tag{6.8}$$

the within group sums of squares deviation matrix as

$$\boldsymbol{W} = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)^T \quad , \tag{6.9}$$

and the total sums of squares deviation matrix as

$$\boldsymbol{T} = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}})(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}})^T \quad . \tag{6.10}$$

It is given that we have $\boldsymbol{T} = \boldsymbol{B} + \boldsymbol{W}$. To discriminate between the classes, we want the within group deviation to be small compared to that between groups. One way of accomplishing this is to maximize

$$\Lambda = \frac{\det(\boldsymbol{W})}{\det(\boldsymbol{T})} \quad , \tag{6.11}$$

which is also called Wilks' Λ. The test of the hypothesis

$$H_0 : \boldsymbol{\mu}_1 = ... = \boldsymbol{\mu}_c \quad \text{vs.} \quad H_1 : \exists i, j | i \neq j (\boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j) \quad , \tag{6.12}$$

is given by $\Lambda \leq U(p, c-1, n-c)$, see further in [Conradsen 2002*a*, Chapt. 7]. We consider this test useful to see if the classes statistically can be discriminated. In the SAS programs `proc discrim` and `proc stepdisc` this test is calculated. Wilk's Λ-test can be further extended to two-sided or three-sided analysis of variance. In this case we have

$$\Lambda = \frac{\det(\boldsymbol{Q_1})}{\det(\boldsymbol{Q_1} + \boldsymbol{Q_2})} \quad , \tag{6.13}$$

where the null-hypothesis is that the effect of $\boldsymbol{Q_2}$ is insignificant and $\boldsymbol{Q_1}$ denotes the error effect. $c-1$ is substituted with the degrees of freedom for the examined effect, $\boldsymbol{Q_2}$, and $n-c$ substituted with the degrees of freedom of the error effect, $\boldsymbol{Q_1}$, cf. [Conradsen 2002*a*, Chapt. 6] and [Rencher 2002, Chapt. 6].

6.2.3 Forward Selection

Forward Selection starts by evaluating a model containing only a constant. We then choose the variable with the largest partial correlation with the response variable. We find the F-value for the coefficient of this variable being significantly different from zero at an α-level. If it is, we include it in the model and start over, if not, we stop. The null hypothesis that the coefficient is zero is equivalent to the null hypothesis that *the partial correlation coefficient between the dependent variable and the independent variable, conditioned on all the independent variables not included in the model* is zero, cf. [Conradsen 2002*a*, Chapt. 4]. A flow diagram of the variable selection is seen in Figure 6.5.

Forward Selection of variables to Discriminant Analysis uses Wilk's Λ-test described in Section 6.2.2 instead of a test of the partial correlation.

Backward Selection uses as starting point the full model, i.e. the model including all the effects that are desired examined. When the number of variables is larger than the number of observations Backward Selection is therefore not adequate since the system of the full model is underdetermined.

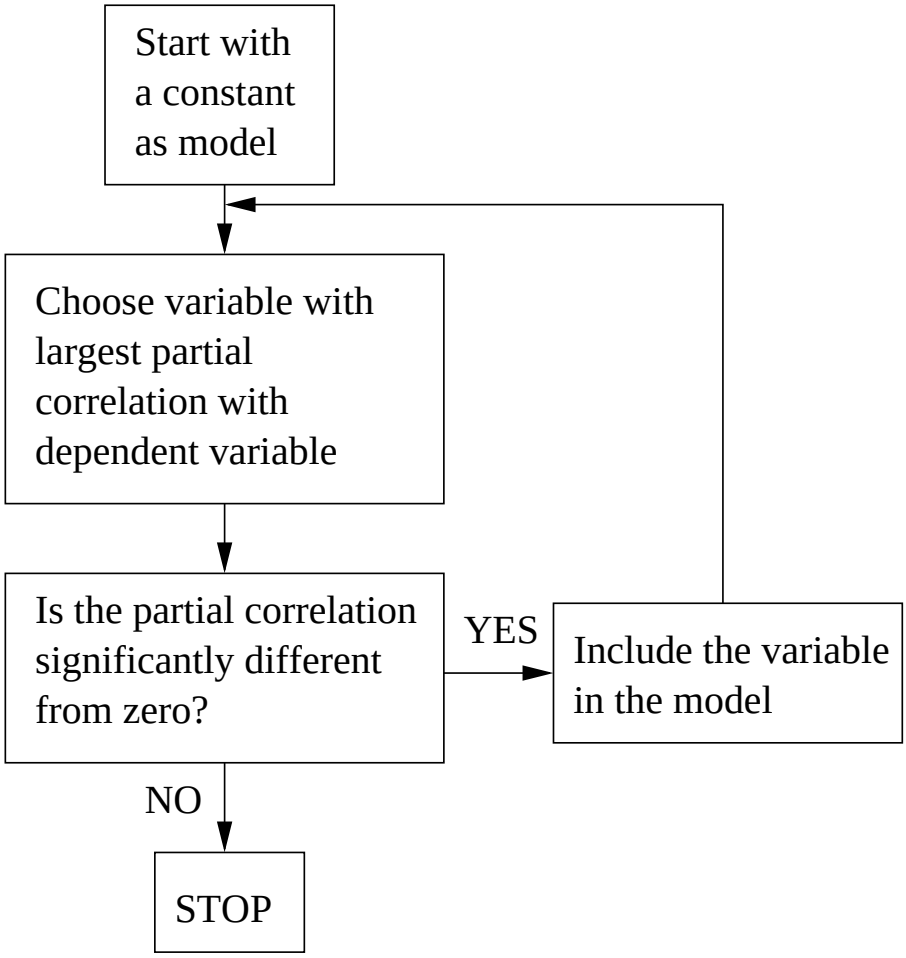


Figure 6.5: Flow diagram of Forward Selection. The partial correlation is the partial correlation between the the dependent variable and the independent variable chosen, conditioned on all the independent variables not included in the model.

6.2.4 Principal Component Analysis

Principal Component Analysis (PCA) decomposes data by means of the *Eigen Value Decomposition* (EVD) of the dispersion matrix. It transforms data by producing linear combinations of the original variables. Each linear combination is composed in order to describe as much of the variance in data as possible. Data is standardized and the correlation matrix considered for reasons described later in this section. The EVD of the dispersion matrix, $\boldsymbol{\Sigma} = \boldsymbol{X}^T\boldsymbol{X}$ is defined as

$$\boldsymbol{\Sigma} = \boldsymbol{P}^T\boldsymbol{\Lambda}\boldsymbol{P} \quad . \tag{6.14}$$

Where $\boldsymbol{\Lambda}$ is a diagonal matrix consisting of the eigenvalues of $\boldsymbol{\Sigma}$ ordered in decreasing order ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$), and $\boldsymbol{P}$ is an orthonormal matrix with the corresponding eigenvectors. The directions or loadings of the principal components are the eigenvectors $\boldsymbol{p}_i$. The principal components are the projections of the data onto the directions of the principal components, i.e. ith principal component (PC) is given by

$$\boldsymbol{y}_i = \boldsymbol{p}_i^T\boldsymbol{X} \quad . \tag{6.15}$$

The amount of variance explained by the m first principal components is, cf. [Conradsen 2002*a*, Sec. 8.1],

$$\frac{\lambda_1 + \ldots + \lambda_m}{\lambda_1 + \ldots + \lambda_p} \quad . \tag{6.16}$$

Choosing the m first principal components reduces the dimensions of the data set, but it has the disadvantage that each principle component (new dimension) is described using all of the original variables. In an inline production, as for example mixing of concrete, time is an issue and it is an advantage to calculate as few variables as possible.

PCA can also be done via the *Singular Value Decomposition* (SVD) of the data matrix $\boldsymbol{X}$. We have the SVD of the data

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T \quad , \tag{6.17}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthonormal and $\boldsymbol{D}$ is a diagonal matrix of the singular values, cf. [Hansen 1998, Sec. 2.1.1]. Remembering the symmetry of $\boldsymbol{\Sigma}$, from (6.14) and (6.17) we have $\boldsymbol{X}^T\boldsymbol{X} = (\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T)^T(\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T) = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^T = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^T$. $\boldsymbol{U}$ are the PCs of unit length, and the columns of $\boldsymbol{V}$ are the corresponding loadings of the principal components, cf. [Zou, Hastie & Tibshirani 2004*b*]. The variance of the ith PC is $\boldsymbol{D}_i^2$, the ith diagonal element in $\boldsymbol{D}^2$.

The SVD is used theoretically in Section 6.3.5 and the singular values are calculated for the data sets to see if the data matrices have a numerical rank, i.e. there is a

gap in the spectrum of the singular values, as described in [Hansen 1998, Chapter 3]. Furthermore, if the singular values decay gradually to zero, with no particular gap in the spectrum, the problem is likely to be *ill posed* [Hansen 1998, Chapter 2]. These features cause us to expect that it is difficult to select an exact number of features to include in the solutions and that it is necessary to regularize the solution since the system of linear equations is ill conditioned.

The Principal Component Analysis can be performed on both the covariance as well as the correlation matrix. Typically, the correlation matrix is preferred as all variables are weighted equally because they are transformed to have equal variance. In this project the correlation matrix is used.

6.2.5 Cross-Validation

To avoid over fitting in regression and supervised classification problems, as those considered in this project, cross-validation (CV) is a useful tool, cf. [Conradsen 2002*b*], [Skettrup 2003], [Hastie et al. 2001], and [Duda, Hart & Stork 2001].

Simple validation is when the dataset is randomly split into two: A training set and a validation or test set. The parameter adjustment for the training set is stopped when the error of the validation set reaches a minimum.

In k-fold cross-validation the dataset is split into k equally sized parts, each containing approximately $\frac{n}{k}$ observations. For the kth part of the data, a model is fitted to the remaining $k-1$ parts, and the prediction error on the kth part is calculated. The parameters for the model are chosen where the mean prediction error of the k parts is minimal. The case where $k = n$ is known as *leave-one-out* cross-validation.

With $k = n$, CV is approximately unbiased for the true prediction error, but the variance might be high if the n training sets are very similar to each other. With lower values of k the prediction error has lower variance, but bias could be a problem.

Hence, for a dataset of few observations, or at least few similar observations, leave-one-out CV would be sufficient. As example, for the fungi dataset which have three repetitions of each isolate it seems reasonable to also examine a lower value of k.

6.3 State of the art methods

Reconsider the problems with OLS:

Prediction accuracy: The OLS estimate often suffers from having a large variance, although it is unbiased. By sacrificing some bias to reduce the variance one might obtain a better prediction accuracy. This can be done by e.g. coefficient shrinkage or by forcing some of the coefficients to zero. Coefficient shrinkage is also useful when $p \gg n$ since the coefficients tend to become very large in this case. This issue is also known as over fitting.

Interpretation: With a large number of variables the solution can be difficult to interpret, and hence, we would like to reduce the number of variables to a subset characterizing only the strongest effects.

The traditional methods combine variable selection or decomposition of data with OLS in order to obtain fewer dimensions. The methods introduced in this section join variable selection and/or coefficient shrinkage with regression analysis.

First Ridge and Lasso regression are described as an introduction to regression with constraints and to the state of the art methods introduced following. These two methods perform coefficient shrinkage. The newer methods perform both regression analysis and variable reduction and/or shrinkage. These methods are LARS and LARS-EN. LARS performs regression with variable selection and LARS-EN combines the variable selection of LARS with coefficient shrinkage by use of both the Ridge and the Lasso constraints. Finally, Sparse PCA is introduced. Not all variables are included in the sparse principal components, hence, solving the time consuming issue of calculating all variables in an inline production.

6.3.1 Ridge Regression

Ridge regression was introduced by Hoerl[1] in 1970 to achieve better prediction accuracy than OLS while sacrificing some bias. The smaller variance of the prediction error is obtained by shrinking the coefficients towards zero by solving the regularization problem

$$\hat{\boldsymbol{\beta}}_{Ridge} = \text{argmin}_{\beta}\{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2\} \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_2^2 \leq t \quad . \tag{6.18}$$

[1] [Hoerl & Kennard 1970]

This is equivalent to solving

$$\hat{\boldsymbol{\beta}}_{Ridge} = \text{argmin}_{\beta}\{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2\} \quad , \tag{6.19}$$

in the sense that there is a one to one relation between t and λ. That is, for any $t \geq 0$ there exists a $\lambda \in [0, \infty[$ such that the two problems have the same solution, and vice versa. The latter can be solved by the normal equations

$$\hat{\boldsymbol{\beta}}_{Ridge} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} \quad . \tag{6.20}$$

The degrees of freedom for the linear Ridge smoother $\boldsymbol{S} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{X}$ is $df(\boldsymbol{S}) = \text{tr}(\boldsymbol{S})$ which in most cases is close to the number of variables [Zou, Hastie & Tibshirani 2004*a*].

Hence, the Ridge shrinkage does not solve the problem of reducing the dimensionality of the feature space. The following three sections describe methods which additionally set some of the coefficients to zero or perform variable selection.

6.3.2 Lasso

The Least Absolute Shrinkage and Selection Operator (Lasso) method was proposed by Tibshirani[2] in 1996, and it minimizes the RSS subject to the 1-norm of the coefficients being less than a constant. Using the 1-norm in the constraint instead of, as in Ridge regression, the 2-norm, causes the method to produce a number of coefficients that are exactly zero.

Problem solved

The Lasso estimate of the coefficients $\boldsymbol{\beta}$ is defined as

$$\hat{\boldsymbol{\beta}}_{Lasso} = \text{argmin}_{\beta}\{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2\} \quad \text{s.t.} \|\boldsymbol{\beta}\|_1 \leq t \quad . \tag{6.21}$$

The constraints of Ridge and Lasso are graphically illustrated for two dimensions in Figure 6.6. The OLS solution to a linear problem is also marked in the figure, and the contours of the quadratic function[3] $(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$ are sketched. The Ridge and Lasso solutions are obtained where the contours first touch the respective constraint. For the Lasso method this is likely to occur at or near a corner (as illustrated in the figure) where one of the coefficients is zero, while for the Ridge regression it is not very likely that one of the coefficients is zero.

[2] [Tibshirani 1996]

[3] This function equals the RSS criterion plus a constant, and the contours are centered at the OLS solution [Tibshirani 1996].

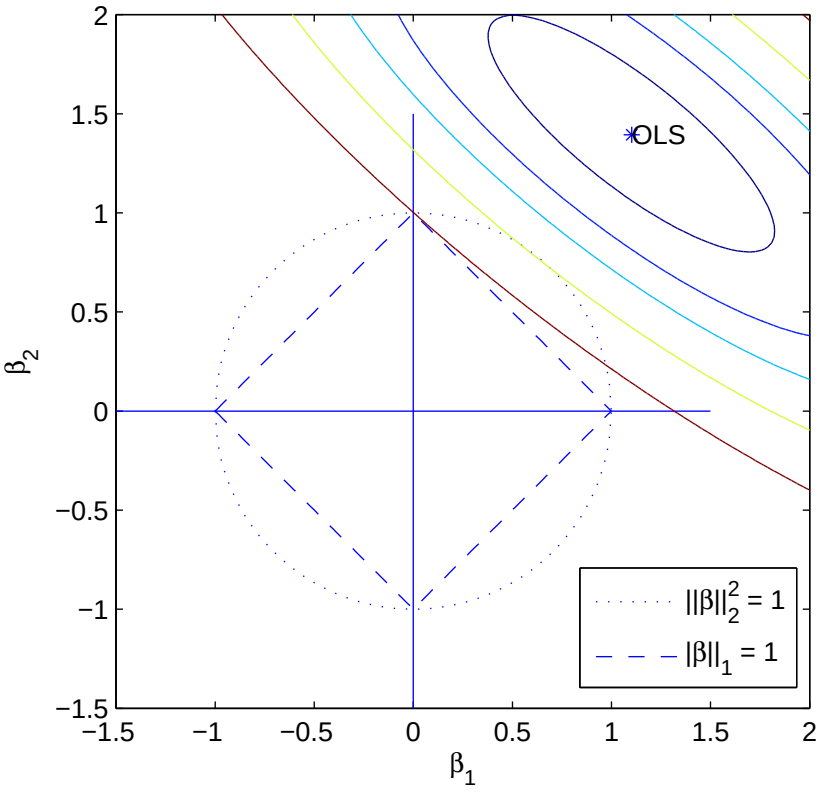


Figure 6.6: Illustration of the estimation with Ridge ($\|\boldsymbol{\beta}\|_2^2 \leq 1$) and Lasso ($\|\boldsymbol{\beta}\|_1 \leq 1$).

Algorithm

Instead of solving (6.21) the algorithm solves the equivalent problem[4]

$$\hat{\boldsymbol{\beta}}_{Lasso} = \text{argmin}_{\beta}\{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1\} \quad . \tag{6.22}$$

The problems in (6.21) and (6.22) are equivalent in the sense that for any $t \geq 0$ there exists a $\lambda \in [0, \infty[$ such that the two problems have the same solution, and vice versa, cf. [Leng, Lin & Wahba 2004]. Consequently, introducing the regularization parameter λ instead of t. A threshold sets coefficients of size less than 10^{-5} to zero, and the method is, hence, operating on an active set of coefficients. The coefficients are updated using a second derivative method, as described in [Gill, Murray & Wright 1981, Chapt. 5]. The following update is used for the coefficients: $-\boldsymbol{H}_A^{-1}\boldsymbol{g}_A$. Where $\boldsymbol{H}$ is the second derivative and $\boldsymbol{g}$ the first derivative of the object function in (6.22). The algorithm is stopped once the change in the coefficients is less than 10^{-9}. The Matlab implementation used here is implemented by PhD Henrik Øjelund, IMM, DTU.

Choice of Parameter

The choice of parameter, λ, can be chosen by cross-validation, as described in Section 6.2.5. CV has proved useful in selecting the shrinkage parameter λ for both Ridge regression and Lasso in [Fu 1998]. A suitable choice has few nonzero parameters, but not so few that the prediction errors become too large.

Limitations

A limitation to Lasso, in particular for $p > n$, is that it selects at most n variables before it saturates, cf. [Zou & Hastie 2005] and [Tibshirani 1996]. Furthermore, the solution is not well defined unless the bound on the 1-norm of the coefficients is smaller than a certain value.

Advantages

The *effective degrees of freedom*, which is an informative measurement of the model complexity described in [Hastie & Tibshirani 1990], for the Lasso as a function of λ corresponds to the number of active parameters, cf. [Zou et al. 2004*a*]. That is, the global trend of the effective degrees of freedom is monotonically decreasing, implying that the dimensions of the problem can be reduced using Lasso.

[4] [Tibshirani 1996]

6.3.3 LARS

The Least Angle Regression (LARS) model selection algorithm suggested in [Efron, Hastie, Johnstore & Tibshirani 2003] is computationally simpler than Lasso or Forward Selection. LARS provides an alternative way of including nonzero coefficients to the regression problem. As in Forward Selection, the method starts with all coefficients equal to zero and proceed by including one nonzero coefficient at each iteration till all coefficients are nonzero and the OLS solution is reached. It is like Forward Selection an iterative method.

The algorithm can be modified to either calculate all possible Lasso solutions, or all possible Forward Selection solutions. The latter is the case only for an idealized Forward Selection where the step size goes to zero, cf. [Efron et al. 2003]. Since the algorithm yields all Lasso or Forward Selection solutions, respectively, the stopping criterion is of great importance.

Algorithm

LARS finds the predictor most correlated with the response, takes a step in this direction until the correlation is equal to another predictor, then it takes the equiangular direction between the predictors of equal correlation (*the least angle direction*). An example with 2 independent variables is illustrated in Figure 6.7.

It is assumed that

$$\sum_{i=0}^{n} y_i = 0 \ , \quad \sum_{i=0}^{n} x_{ij} = 0 \ , \quad \text{and} \quad \sum_{i=0}^{n} x_{ij}^2 = 1 \quad \text{for} \quad j = 1, 2, ..., p \ , \tag{6.23}$$

so that $\boldsymbol{X}^T\boldsymbol{X} = \text{Corr}(\boldsymbol{X})$. And $\boldsymbol{y}$ is centered so that a constant term should be redundant.

Equiangular direction
The equiangular vector for a set of observations A is given by

$$\boldsymbol{u}_A = \boldsymbol{X}_A \boldsymbol{w}_A \tag{6.24}$$

where

$$\boldsymbol{w}_A = A_A(\boldsymbol{X}_A^T\boldsymbol{X}_A)^{-1}\mathbf{1}_A \quad \text{and} \quad A_A = (\mathbf{1}_A^T(\boldsymbol{X}_A^T\boldsymbol{X}_A\mathbf{1}_A)^{-1})^{-\frac{1}{2}} \quad , \tag{6.25}$$

where $\mathbf{1}_A$ is a vector of A ones. The equiangular vector if multiplied to $\boldsymbol{X}_A$, i.e. $\boldsymbol{a} = \boldsymbol{X}_A^T\boldsymbol{u}_A$, yields angles between the columns of $\boldsymbol{a}$ and the columns of $\boldsymbol{X}_A$ that are 90 degrees. Furthermore the vector $\boldsymbol{u}_A$ is of unit length.

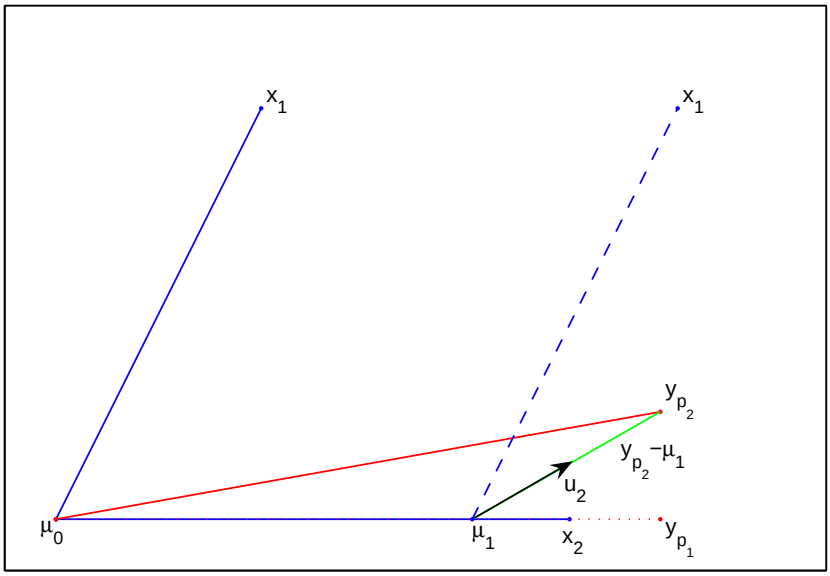


Figure 6.7: Illustration of LARS iterations in the case with 2 independent variables [Efron et al. 2003]. $\boldsymbol{y_{p_2}}$ is the projection of $\boldsymbol{y}$ into the space spanned by $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$. The initial guess is $\boldsymbol{\mu_0} = 0$. The residual vector $\boldsymbol{y_{p_2}} - \boldsymbol{\mu_0}$ has greater correlation with $\boldsymbol{x_2}$ than $\boldsymbol{x_1}$, hence, the next LARS estimate is $\boldsymbol{\mu_1} = \boldsymbol{\mu_0} + \gamma_1 \boldsymbol{x_2}$, where γ_1 is chosen such that $\boldsymbol{y_{p_2}} - \boldsymbol{\mu_1}$ bisects the angle between $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ (the equiangular direction $\boldsymbol{u_2}$) hence making the correlations between $\boldsymbol{y_{p_2}}$ and $\boldsymbol{\mu_1}$, and $\boldsymbol{y_{p_2}}$ and $\boldsymbol{x_1}$ equal. Then $\boldsymbol{\mu_2} = \boldsymbol{\mu_1} + \gamma_2 \boldsymbol{u_2}$, where in the case with two independent variables $\gamma_2 = \|\boldsymbol{y_{p_2}} - \boldsymbol{\mu_1}\|_2$, leaving $\boldsymbol{\mu_2} = \boldsymbol{y_{p_2}}$.

Length of step in the equiangular direction
The length of the step taken in this direction is exactly such that a new variable has the same correlation with the response, as the ones in the active set of variables, A, and the variables of equal variance becomes active. The length of the step is given by

$$\gamma = \min_{j \in A}\{\frac{C - c_j}{A_A - a_j}, \frac{C + c_j}{A_A + a_j}\} \quad , \tag{6.26}$$

the minimum distance for one of the inactive variables to become active when progressing in direction $\boldsymbol{u}_A$. a_j is the j^{th} value in $\boldsymbol{a}$ and c_j is the j^{th} value in the vector of current correlations $\boldsymbol{c} = \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{\mu}_A)$, where $\boldsymbol{\mu}_A$ is the current LARS estimate. C is the maximum current correlation, i.e. $C = \max_j c_j$.

Iterative update
The updating of the LARS estimate is then:

$$\boldsymbol{\mu}_{A_+} = \boldsymbol{\mu}_A + \gamma \boldsymbol{u}_A. \tag{6.27}$$

Lasso and Forward Selection
Modifications of this algorithm gives the Lasso and Forward Selection solutions, these modifications are given in [Efron et al. 2003]. The Lasso modification is invoked when the sign of a non-zero parameter does not agree with that of the current correlation, as they must be of same sign to be a Lasso solution. The parameter is then removed from the active set and excluded from the calculations of the equiangular direction. For more details see [Zou & Hastie 2005, Sec. 3.1]. The Matlab implementation used is made by PhD student Karl Skoglund, IMM, DTU.

Stopping Criterion

As before CV can be used to find appropriate choices of the regularization parameters and the number of iterations. The algorithm yields all Lasso/Forward Selection/LARS solutions and one stopping criterion is therefore the number of iterations. The number of iterations can be chosen from CV or from a restriction on the number of desired variables. CV based on RSS(ite), the RSS as a function of the number of iterations, can be performed over the different criteria to find a good solution.

6.3.4 LARS-EN

The Elastic Net LARS (LARS-EN) uses two constraints, both the Ridge and the Lasso constraints. The Lasso chooses at most n variables before it sets all coefficients to nonzero. Since we are interested in variable selection this might be limiting, and therefore LARS-EN is considered. Furthermore, groups of variables can enter at the same

time with the LARS-EN algorithm, unlike previously mentioned methods. LARS-EN performs the variable selection of the LARS algorithm with Lasso modification and the shrinkage of Ridge.

Problem solved

The naive elastic net estimator is defined by Zou[5] as

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\beta}\{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2\} \quad . \tag{6.28}$$

Choosing $\lambda_1 = 0$ yields Ridge solutions, and likewise choosing $\lambda_2 = 0$ yields Lasso solutions. Let $\alpha = \frac{\lambda_2}{\lambda_1+\lambda_2}$, then, according to Zou[6], solving (6.28) is equivalent to solving the optimization problem

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\beta}\{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2 \quad s.t. \ (1-\alpha)\|\boldsymbol{\beta}\|_1+\alpha\|\boldsymbol{\beta}\|_2^2 \leq t \quad \text{for some } t \quad . \tag{6.29}$$

The function $(1-\alpha)\|\boldsymbol{\beta}\|_1 + \alpha\|\boldsymbol{\beta}\|_2^2$ is called the elastic net penalty and it is illustrated together with the Ridge and Lasso penalties in Figure 6.8.

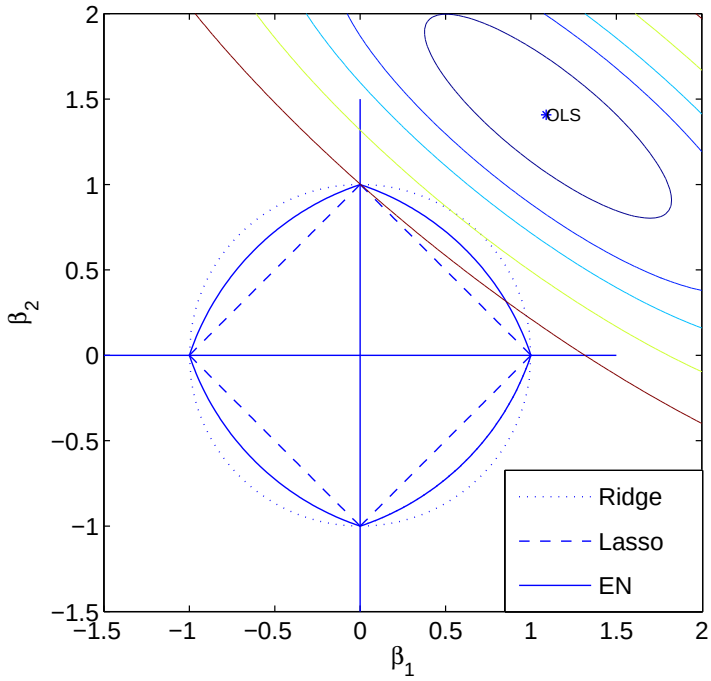


Figure 6.8: Illustration of the estimation with Ridge ($\|\boldsymbol{\beta}\|_2^2 \leq 1$), Lasso ($\|\boldsymbol{\beta}\|_1 \leq 1$) and LARS-EN ($(1-\alpha)\|\boldsymbol{\beta}\|_1 + \alpha\|\boldsymbol{\beta}\|_2^2 \leq 1$ with $\alpha = 0.5$).

[5][Zou & Hastie 2005]
[6][Zou & Hastie 2005]

Algorithm

Zou and Hastie[7] says that we can transform the naive elastic net problem into an equivalent Lasso problem on the augmented data

$$\boldsymbol{X}^* = (1+\lambda_2)^{-1/2} \begin{bmatrix} \boldsymbol{X} \\ \sqrt{\lambda_2}\boldsymbol{I}_p \end{bmatrix} \quad , \quad \boldsymbol{y}^* = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0}_p \end{bmatrix} \quad . \tag{6.30}$$

The normal equations, yielding the OLS solution, to this augmented problem are

$$\begin{aligned} \left(\frac{1}{\sqrt{1+\lambda_2}}\right)^2 \begin{bmatrix} \boldsymbol{X} \\ \sqrt{\lambda_2}\boldsymbol{I}_p \end{bmatrix}^T \begin{bmatrix} \boldsymbol{X} \\ \sqrt{\lambda_2}\boldsymbol{I}_p \end{bmatrix} \hat{\boldsymbol{\beta}}^* &= \frac{1}{\sqrt{1+\lambda_2}} \begin{bmatrix} \boldsymbol{X} \\ \sqrt{\lambda_2}\boldsymbol{I}_p \end{bmatrix}^T \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0}_p \end{bmatrix} \Leftrightarrow \\ \frac{1}{\sqrt{1+\lambda_2}} \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda_2 \boldsymbol{I}_p^T \boldsymbol{I}_p\right) \hat{\boldsymbol{\beta}}^* &= \boldsymbol{X}^T\boldsymbol{y} \quad . \end{aligned} \tag{6.31}$$

We see that $\frac{1}{\sqrt{1+\lambda_2}}\hat{\boldsymbol{\beta}}^*$ is the Ridge regression estimate with parameter λ_2. Hence, performing Lasso on this augmented problem yields an elastic net solution. Consequently, we can with advantage use the LARS algorithm with the Lasso modification to find the Lasso solution to this augmented problem. Summing up, the naive elastic net estimator is a two-stage procedure: For a fixed λ_2 the Ridge regression estimator is found and then a Lasso-type shrinkage is performed through LARS.

However, the naive elastic net does not perform satisfactory unless it is very close to either Lasso or Ridge regression. This is because a double amount of shrinkage will occur. Therefore, the naive elastic net solution is scaled, and using a scaling with $1+\lambda_2$ the variable selection property is preserved and the double amount of shrinkage is avoided. Furthermore, minimax optimality is obtained. When the Ridge regression is combined with Lasso, the direct shrinkage $\frac{1}{1+\lambda_2}$ is not needed and is removed by rescaling. It is unnecessary because the Lasso shrinkage controls the variance. We will come back to this in Section 6.4.5.

Finally, since the problem is augmented, then, in particular when $p \gg n$ the computations are slowed down. Therefore, the inversion of the matrix $\boldsymbol{X}_A^T\boldsymbol{X}_A$ (the correlation of the active independent variables, cf. the LARS algorithm) can with advantage be done by an up or down dating of the Cholesky factorization of $\boldsymbol{X}_A^T\boldsymbol{X}_A$ from the previous step.

[7] [Zou & Hastie 2005, Lemma 1]

Choice of parameters

The algorithm uses the LARS implementation with the Lasso modification, and hence we have the parameter λ_2 to adjust, but also the number of iterations for the LARS algorithm can be used. The larger λ_2, the more weight is put on the Ridge constraint and the number of active variables increases. The Lasso constraint is weighted by the number of iterations. Few iterations correspond to a high value of λ_1, and vice versa. The number of iterations can also be used to ensure a low number of active variables comparable to the procedure in Forward Selection. Cross-validation on RSS(λ_2, ite), the RSS as a function of λ_2 and the number of iterations, is used to choose good regularization parameters.

6.3.5 Sparse Principal Components

The idea behind sparse principal components is to produce principal components with sparse loadings. The sparseness reduces the number of active variables which e.g. is desirable in an inline production. The method was introduced in [Zou et al. 2004*b*] and used the elastic net (LARS-EN) to perform a regression of the principal components in order to obtain the sparseness. In [Zou et al. 2004*b*] an algorithm for producing sparse principal components is stated, but it is also mentioned how, by use of the following theorem, to perform a two-stage exploratory analysis to obtain sparse PCs.

Theorem 1[8] *$\forall i$, denote $\boldsymbol{Y}_i = \boldsymbol{U}_i\boldsymbol{D}_i$ ($\boldsymbol{D}_i$ is the i-th diagonal element in $\boldsymbol{D}$). $\boldsymbol{Y}_i$ is the i-th principal component. $\forall \lambda > 0$, suppose the $\hat{\boldsymbol{\beta}}_{Ridge}$ is the Ridge estimates given by:*

$$\hat{\boldsymbol{\beta}}_{Ridge} = \text{argmin}_{\boldsymbol{\beta}}\{\|\boldsymbol{Y}_i - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2\} \quad . \tag{6.32}$$

Let $\hat{\boldsymbol{v}} = \dfrac{\hat{\boldsymbol{\beta}}_{Ridge}}{\|\hat{\boldsymbol{\beta}}_{Ridge}\|_2}$, then $\hat{\boldsymbol{v}} = \boldsymbol{V}_i$, where $\boldsymbol{X} = \boldsymbol{U}^T\boldsymbol{D}\boldsymbol{V}$ is the SVD of $\boldsymbol{X}$.

[8] [Zou et al. 2004*b*]

Proof Using $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^T$ and $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$, we have

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{Ridge} &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{Y}_i \\
&= (\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^T + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{V}_i) \\
&= (\boldsymbol{V}(\boldsymbol{D}^2 + \boldsymbol{V}^T\lambda\boldsymbol{I}\boldsymbol{V})\boldsymbol{V}^T)^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{V}_i \\
&= (\boldsymbol{V}(\boldsymbol{D}^2 + \lambda\boldsymbol{I})\boldsymbol{V}^T)^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{V}_i \\
&= \boldsymbol{V}^{-T}(\boldsymbol{D}^2 + \lambda\boldsymbol{I})^{-1}\boldsymbol{V}^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{V}_i \\
&= \boldsymbol{V}(\boldsymbol{D}^2 + \lambda\boldsymbol{I})^{-1}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^T\boldsymbol{V}_i \\
&= \boldsymbol{V}(\boldsymbol{D}^2 + \lambda\boldsymbol{I})^{-1}\boldsymbol{D}^2\boldsymbol{V}^T\boldsymbol{V}_i \\
&= \boldsymbol{V}_i\frac{\boldsymbol{D}_i^2}{\boldsymbol{D}_i^2 + \lambda} \quad . \qquad (6.33)
\end{aligned}$$

□

Only the ith diagonal element contributes in the last derivation due to the orthonormality of $\boldsymbol{V}$. From (6.33) we see that the Ridge estimates differ from the principal directions $\boldsymbol{V}_i$ only by a constant, hence, normalizing the estimates will yield exactly the principal directions and therefore the ith approximated principal component $\hat{\boldsymbol{U}}_i = \boldsymbol{X}\hat{\boldsymbol{V}}_i$ is given as a result of $\hat{\boldsymbol{V}}_i = \frac{\hat{\beta}}{\|\hat{\beta}\|_2}$.

The Ridge estimates do not give sparse solutions and therefore LARS-EN is used. The Ridge penalty, though, should be kept to ensure reconstruction of the principal components.

In this project only an exploratory analysis is considered, hence, first the PCA is performed and then the LARS-EN is used to find sparse approximations.

The adjusted total variance of the Sparse PCs which take into account the correlations among the Sparse PCs $\hat{\boldsymbol{U}}_i$ can be calculated by use of a QR decomposition $\hat{\boldsymbol{U}} = \boldsymbol{Q}\boldsymbol{R}$, where $\boldsymbol{Q}$ is orthonormal and $\boldsymbol{R}$ is upper triangular. The adjusted variance of $\hat{\boldsymbol{U}}_i$ is $\boldsymbol{R}_{j,j}^2$ the jth diagonal element squared, cf. [Zou et al. 2004*b*, Sec. 3.4].

6.4 Additions

This section examines additional features of the utilized methods than those described so far.

The first five sections examine the shrinkage effect of Lasso, Ridge regression and LARS-EN on the residuals. During the experiments residuals with trends were ob-

served as if a constant term is missing in the model. However, this cannot be the case because the response variable is centered. The trends were in particular observed when early stopping was used in the LARS-EN algorithm. This is due to the coefficient shrinkage and can likewise be observed for Ridge and Lasso.

In the sixth section, regression with dummy variables as dependent variables is examined and the relation to Discriminant Analysis investigated.

6.4.1 Shrinkage in Lasso

A simple regression example with one independent variable is constructed

$$y_i = 0.5x_i + \epsilon_i \quad , \quad i = 1, ..., n \quad , \tag{6.34}$$

where $\epsilon_i \in N(0, \sigma^2)$ with $\sigma = 10^{-3}$. For a simulation with $n = 1000$, the residuals and the true observation versus the estimated values obtained from the Lasso regression are seen in Figure 6.9. Notice, how the residuals show a more and more pronounced linear departure from the usual *random noise* pattern, hereafter referred to as trends, as the shrinkage increases, i.e. λ increases. Consequently, the effect of the coefficient shrinkage is underestimation. Recall, the issue of prediction accuracy which can be solved by coefficient shrinkage by sacrificing some bias. The underestimation is exactly a trade off bias for smaller variance with respect to prediction accuracy.

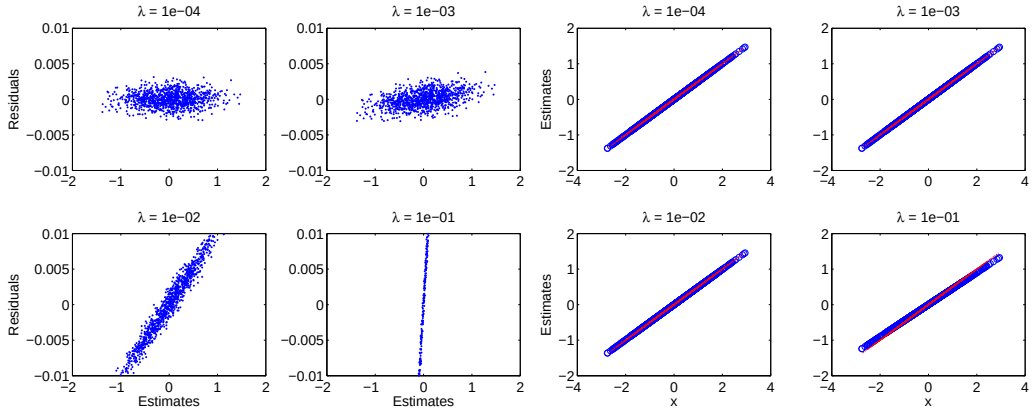
6.4.2 Shrinkage in Ridge

As seen for Lasso, shrinkage with Ridge results in underestimation. A two-dimensional data set with $n = 1000$ is constructed, where

$$yi = 0.5x_{1i} - 0.6x_{2i} + \epsilon_i \quad , \tag{6.35}$$

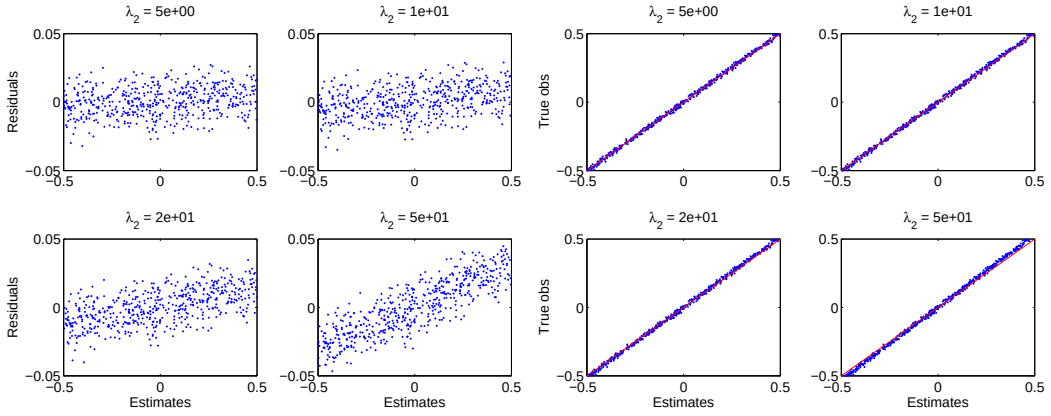
where $\epsilon_i \in N(0, \sigma^2)$ with $\sigma = 10^{-2}$ and $\Sigma_{xx} = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}$. The measured observations and the residuals as functions of the estimates are illustrated in Figure 6.10.

The residuals begin to show trends as λ is increased. Again, the underestimation is a trade off bias for smaller variance with respect to prediction accuracy.



(a) Residuals (b) Estimates

Figure 6.9: (a): Residuals versus estimated values, obtained from Lasso shrinkage for four values of λ. (b): Estimated values as function of x, the function $y = 0.5x$ is marked with red and the estimated values obtained by Lasso are marked with blue circles, for four values of λ.



(a) Residuals (b) True observations

Figure 6.10: Residuals and true observations versus the estimated values. The red line marks $y = x$.

6.4.3 Early stopping in LARS-EN

A small example with a synthetic data set is given here. The data set has 9 variables and 50 observations, $\boldsymbol{X}_{50\times 9}$. Note, that the following experiments show the same trends when the number of observations is larger, for example 5000, and therefore a general tendency is described. It is created from normally distributed data transformed to be correlated gradually with the response variable, $\boldsymbol{y}$. The correlation matrix of the response and regression variables is

$$Corr([\boldsymbol{y}, \boldsymbol{X}]) = \begin{pmatrix} 1.00 & -0.81 & -0.69 & -0.47 & -0.39 & -0.15 & 0.20 & 0.02 & 0.14 & 0.10 \\ -0.81 & 1.00 & 0.54 & 0.16 & 0.28 & -0.21 & -0.18 & -0.13 & 0.19 & 0.12 \\ -0.69 & 0.54 & 1.00 & -0.04 & 0.16 & -0.03 & -0.07 & 0.04 & -0.18 & -0.01 \\ -0.47 & 0.16 & -0.04 & 1.00 & -0.06 & 0.17 & 0.15 & -0.06 & -0.23 & -0.15 \\ -0.39 & 0.28 & 0.16 & -0.06 & 1.00 & 0.10 & -0.88 & 0.02 & -0.05 & 0.13 \\ -0.15 & -0.21 & -0.03 & 0.17 & 0.10 & 1.00 & -0.09 & -0.02 & -0.44 & -0.42 \\ 0.20 & -0.18 & -0.07 & 0.15 & -0.88 & -0.09 & 1.00 & -0.03 & -0.08 & -0.23 \\ 0.02 & -0.13 & 0.04 & -0.06 & 0.02 & -0.02 & -0.03 & 1.00 & -0.48 & -0.57 \\ 0.14 & 0.19 & -0.18 & -0.23 & -0.05 & -0.44 & -0.08 & -0.48 & 1.00 & 0.27 \\ 0.10 & 0.12 & -0.01 & -0.15 & 0.13 & -0.42 & -0.23 & -0.57 & 0.27 & 1.00 \end{pmatrix}.$$

Other correlation matrices were also examined, e.g. for a correlation matrix of $\begin{pmatrix} 1 & & \rho \\ & \ddots & \\ \rho & & 1 \end{pmatrix}$, where $\rho \in [0.60, 0.99]$, the results are similar. Again, this indicates that the tendencies illustrated in the following are general.

The simulated observations; the true observations, are estimated using LARS-EN. The LARS-EN algorithm is run with $\lambda = 10^{-6}$ and the results at each iteration are illustrated in Figure 6.12, and 6.13. In Figure 6.11 the singular values of the data matrix, $\boldsymbol{X}$, are plotted.

It is seen that early stopping in this case produces trends in the residuals. The trends are caused by underestimation. Underestimation was also seen with Lasso, and since the number of iterations correspond to the weight of the Lasso constraint, this was expected. Furthermore, the singular values indicate that the first 6 variables included are of greater importance as their singular values are larger, this is verified by the residual plots.

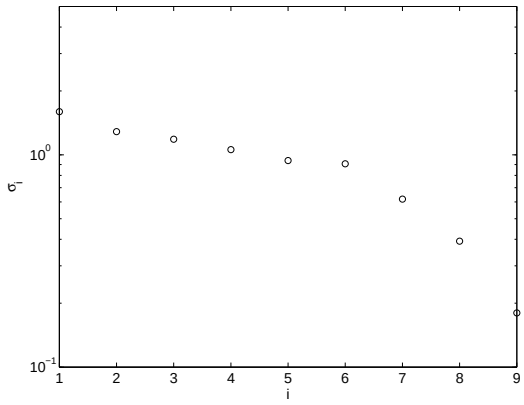


Figure 6.11: Singular values of the test example with 9 variables.

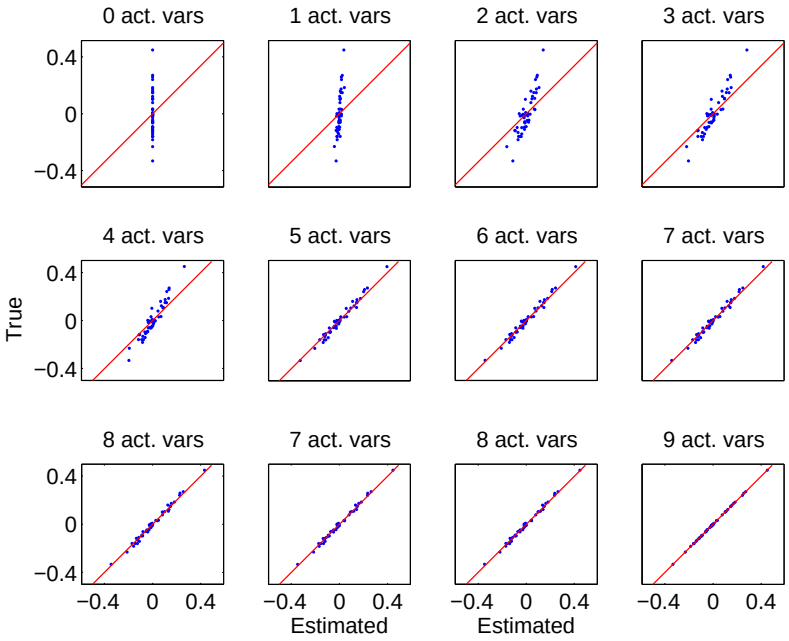


Figure 6.12: True observations versus estimated values of $\boldsymbol{y}$ at each LARS-EN iteration. The red line marks $y = x$.

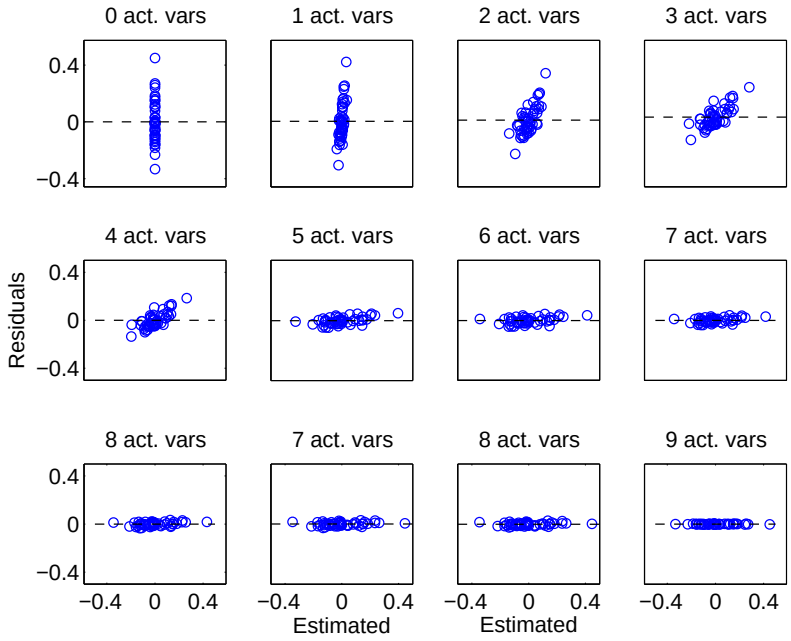


Figure 6.13: Residuals for LARS-EN on the test example with 9 variables at each LARS-EN iteration.

6.4.4 Regularizing with λ in LARS-EN

Regularizing with λ in LARS-EN produces trends in the residuals as when early stopping is used for regularization. The same synthetic data set is used as in Section 6.4.3.

LARS-EN performs in accordance with the theory it selects the variable with the greatest absolute correlation with the response variable at each iteration. The residuals and true observations versus the estimates are illustrated for different values of λ in Figure 6.14 and 6.15. The algorithm is iterated till all possible variables are entered for the given λ. Note, that the trends in the residuals are opposite of those observed with early stopping. Hence, overestimation is the result of regularizing with λ in LARS-EN. On the face of it this was not expected since λ is the weight put on the Ridge constraint, and Ridge regression tends to underestimate. However the Ridge estimates of LARS-EN require $\frac{1}{1+\lambda_2}$ shrinkage to control the variance, but it is not performed because the Lasso shrinkage controls the variance, as we shall see in the next section.

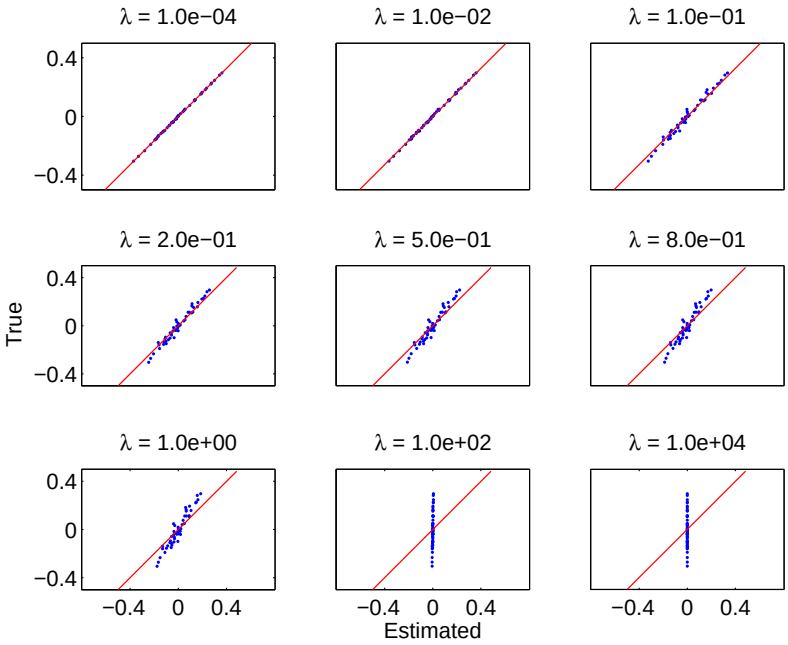


Figure 6.14: True versus estimated values of $\boldsymbol{y}$. There are 9 active variables in all cases. The red line marks $y = x$.

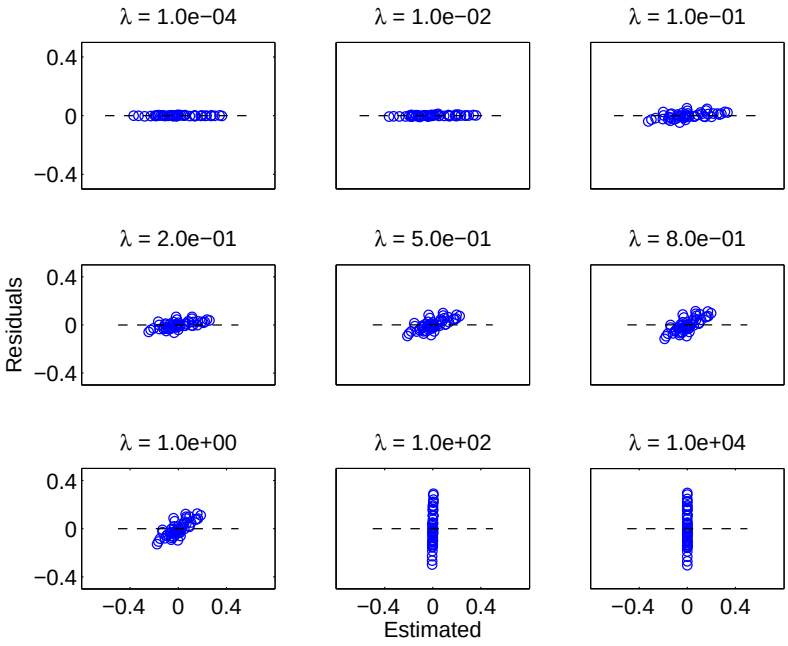


Figure 6.15: Residuals for LARS-EN on the te6st example with 9 variables. There are 9 active variables in all cases.

6.4.5 Early stopping and λ regularization

The data from Section 6.4.3 is used. Here, both early stopping and regularization with λ is considered. For each value of λ, the number of iterations is chosen based on the *Mean Squared Error* MSE. The residuals and true observations versus the estimates for different values of λ are illustrated in Figure 6.16 and 6.17.

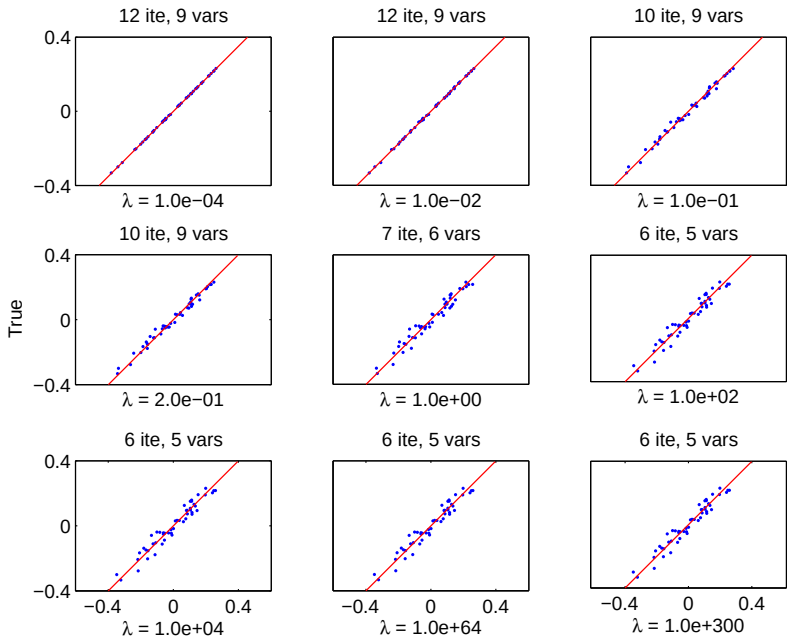


Figure 6.16: True versus estimated values of $\boldsymbol{y}$. The red line marks $y = x$.

Using both early stopping and λ regularization gives in this case no trends in the residuals. The two trends cancel out. Recall, that the trends were opposite for the early stopping and the λ regularization (cf. Figure 6.13 and 6.15).

However, it may not always be possible to choose the number of iterations and a λ such that the training data is not over fitted. Hence, trends in the residuals caused by over- or underestimation may be observed for the LARS-EN algorithm.

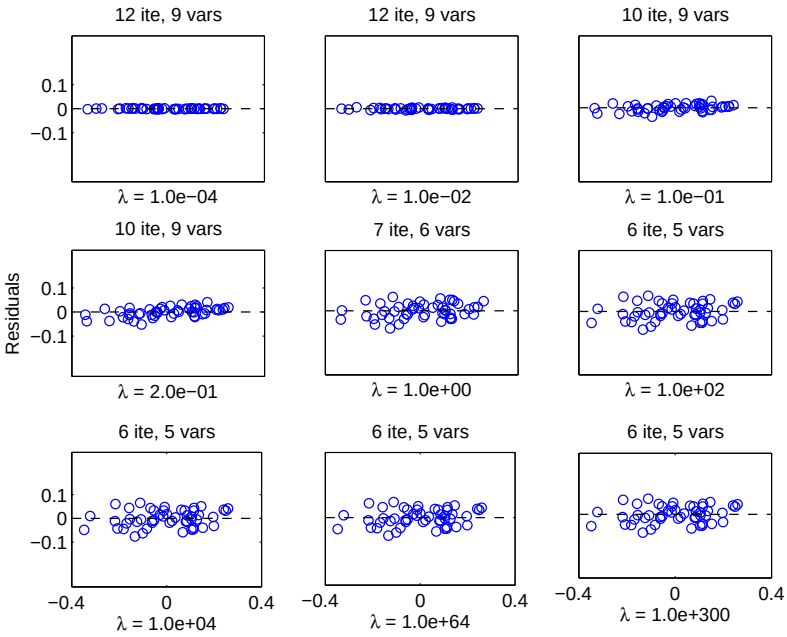


Figure 6.17: Residuals for LARS-EN on the test example with 9 variables.

6.4.6 Classification via regression

The results obtained by Discriminant Analysis should be comparable to those obtained by multivariate linear regression, cf. [StatSoft 2005]. This is examined in the following.

Two classes

In the Discriminant Analysis, each class is held against the others separately. Using a dummy variable as dependent variable is straight forward in the two class situation, as we will see in the following. The dummy variable takes on the values 1 for class a and -1 for class b. The classification is done by classifying all observations with a predicted value greater than zero, as belonging to class a, and values smaller than zero to class b. This is similar to what is done with the score functions in the Discriminant Analysis. Recall, that for two scoring functions s_a and s_b we have the classification rule: *if* $s_a - s_b > 0$ *then the observation belongs to class* a, cf. Section 6.2.2.

Comparison of Discriminant Analysis and regression with dummy variables
In the following, the resemblance between regression with a dummy variable and Bayesian Discriminant Analysis is illustrated.

Consider, the two class situation with p independent variables $\boldsymbol{x}_1, ..., \boldsymbol{x}_p$. Let $\boldsymbol{X} = [\boldsymbol{x}_1 ... \boldsymbol{x}_p]$, $\boldsymbol{\mu}$ be the mean of all observations and $\boldsymbol{\mu}_a$ and $\boldsymbol{\mu}_b$ the means for the classes a and b, respectively. The dependent variable is a dummy variable of n_b minus ones and n_a ones centered to have zero mean and is denoted $\boldsymbol{y}$.

We have

$$\begin{aligned}
\boldsymbol{X}^T\boldsymbol{y} &= \begin{bmatrix} x_{11} & \dots & x_{1n} \\ \vdots & & \vdots \\ x_{p1} & \dots & x_{pn} \end{bmatrix} \begin{bmatrix} \frac{1}{n_a} \\ \vdots \\ \frac{1}{n_a} \\ \frac{-1}{n_b} \\ \vdots \\ \frac{-1}{n_b} \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{n_a}\sum_{i\in a} x_{1i} - \frac{1}{n_b}\sum_{i\in b} x_{1i} \\ \vdots \\ \frac{1}{n_a}\sum_{i\in a} x_{pi} - \frac{1}{n_b}\sum_{i\in b} x_{pi} \end{bmatrix} \\
&= \hat{\boldsymbol{\mu}}_a - \hat{\boldsymbol{\mu}}_b \quad , \qquad (6.36)
\end{aligned}$$

and the OLS parameter estimates are then

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\hat{\boldsymbol{\mu}}_a - \hat{\boldsymbol{\mu}}_b) \end{aligned} \tag{6.37}$$

Recall, that the discriminating function for the GLM is given by

$$\begin{aligned} y = \boldsymbol{x}^T\hat{\boldsymbol{\theta}} &= 0 \Leftrightarrow \\ \boldsymbol{x}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\hat{\boldsymbol{\mu}}_a - \hat{\boldsymbol{\mu}}_b) &= 0 \quad . \end{aligned} \tag{6.38}$$

In the Bayesian Discriminant Analysis the discriminating function is given by

$$s_a - s_b = \frac{1}{2}(\hat{\boldsymbol{\mu}}_a + \hat{\boldsymbol{\mu}}_b)^T\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_b - \hat{\boldsymbol{\mu}}_a) + \boldsymbol{x}^T\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_a - \hat{\boldsymbol{\mu}}_b) = 0 \quad . \tag{6.39}$$

There is a resemblance between (6.38) and (6.39). The relation between $\hat{\boldsymbol{\Sigma}}$ and $\boldsymbol{X}^T\boldsymbol{X}$ is not one to one. However, if $\boldsymbol{X}$ is standardized both are estimates of the dispersion matrix. In the Discriminant Analysis the within sums of squares deviation matrix is utilized. Furthermore, the Bayesian discriminating function includes a constant, but this is not necessary if $\boldsymbol{y}$ is centered.

Example
Figure 6.18 illustrates a small example of a data set with two classes and two observations. Note that variable x_1 can discriminate the two classes entirely while variable x_2 cannot. The discriminant function from running the discriminant procedure in SAS is

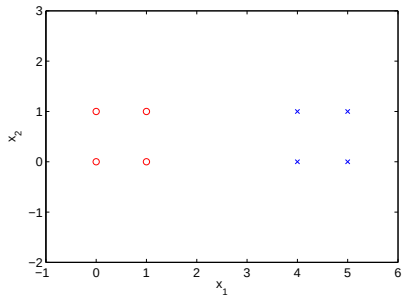


Figure 6.18: Data set with two classes and two variables.

$$\begin{aligned} s_a &= -0.75 + 1.5x_1 + 1.5x_2 \\ s_b &= -30.75 + 13.5x_1 + 1.5x_2 \\ s_a - s_b &= 30 - 12x_1 \end{aligned} \tag{6.40}$$

Hence, we have the classification rule: *if* $s_a - s_b > 0$ *then class* a. Choosing a dummy variable y equal to 1 for class a and -1 for class b and running `proc glm` in SAS should then yield a regression function equivalent to that in (6.40). The regression function becomes

$$y = 1.1765 - 0.47059x_1 \tag{6.41}$$

Multiplying by a factor 25.5 this gives numerically the same function as in (6.40), and hence the classification rules: *if* $s_a - s_b > 0$ *then class* a and *if* $y > 0$ *then class* a are equivalent[9].

More than two classes

The question is how to construnct the dummy variables in the case of more than two classes. Three options seem direct:

1 For the regression with dummy variables to resemble Discriminant Analysis each class should be compared to the others separately. In the case with three classes, one might consider dummy variables of ones, zeros, and minus ones. For three classes, the dummy variables could look like this

$$\begin{bmatrix} \mathbf{1} \\ \mathbf{0} \\ \mathbf{-1} \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} \mathbf{-1} \\ \mathbf{1} \\ \mathbf{0} \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} \mathbf{0} \\ \mathbf{-1} \\ \mathbf{1} \end{bmatrix} \quad . \tag{6.42}$$

However, this includes a priori information about how the classes are related. In practice, this did not turn out to be reasonable.

2 One might consider to hold one class against all other classes as one. For three classes, the dummy variables could look like this

$$\begin{bmatrix} \mathbf{1} & \mathbf{-1} & \mathbf{-1} \\ \mathbf{-1} & \mathbf{1} & \mathbf{-1} \\ \mathbf{-1} & \mathbf{-1} & \mathbf{1} \end{bmatrix} \quad . \tag{6.43}$$

[9] $s_a - s_b > 0 \Leftrightarrow y > 0$

For the example in Figure 6.19, the class in the middle, marked with triangles cannot be distinguished from the other two if they are looked upon as one class. However, in practice these dummy variables perform well, and are therefore used in this project.

- To consider only two classes at a time. This way the number of regressions performed increases drastically with the number of classes.

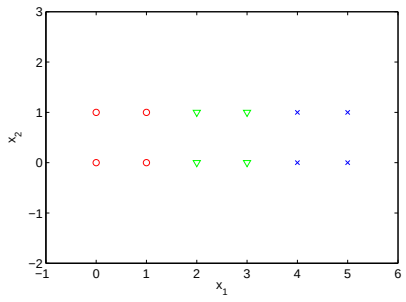


Figure 6.19: Data set with three classes and two variables.

Option 2 is the one considered in this proejct and by choosing the right classification rule the disadvantage of this option is decreased. When option 2 is used one can choose between two classification rule options:

a To classify an observation to the class it is closest to based on the estimated values for all the dummy variables. This option makes the classification of the middle class in Figure 6.19 possible.

b To classify an observation belonging to a class only if it belongs to it in the two class situation, and if it does not belong to any class, classify it to a class of unclassified observations.

In this project the observations are classified in accordance with a, but an example of option b will be given in Chapter 8.

6.5 Summing up

Two methods to segment the fungi from the images have been presented. One based on information from one spectral band that assumes the fungi have grown into cir-

cular colonies and therefore can extract spatial information, and one that exploits the information provided by all 18 spectral bands but without spatial information included.

The traditional OLS regression suffers from prediction accuracy due to over fitting in the case where $p \gg n$ (the number of variables is much larger than the number of observations). With many variables included, OLS also lacks interpretability. Discriminant Analysis suffers from the same issues as OLS when $p \gg n$. To solve these issues Forward Selection or PCA can be performed preliminary to reduce the dimensions.

The Ridge and Lasso regression reduce the variance of the prediction error by adding a constraint to the minimization of the RSS. Furthermore, the Lasso constraint reduces the dimensions. LARS reduces the dimensionality by variable selection, and can be modified to compute Lasso solutions computationally faster than the original Lasso. The LARS-EN model selection method combines the variable shrinkage of Ridge and the variable selection from LARS with Lasso modification. Hence, LARS-EN performs both regression and variable selection as well as variable shrinkage in one step.

The sparseness of these methods can be used to construct sparse principal components where not all variables are included for each principal component, i.e. both a projection of data and variable reduction is performed.

The shrinkage can cause the estimates to be under- or overestimated, but can, nonetheless, be necessary in order not to over fit.

It is shown that Discriminant Analysis and regression with dummy variables as dependent variables are closely related, and utilization of the newer model selection methods with use of dummy variables is proposed.

Chapter 7

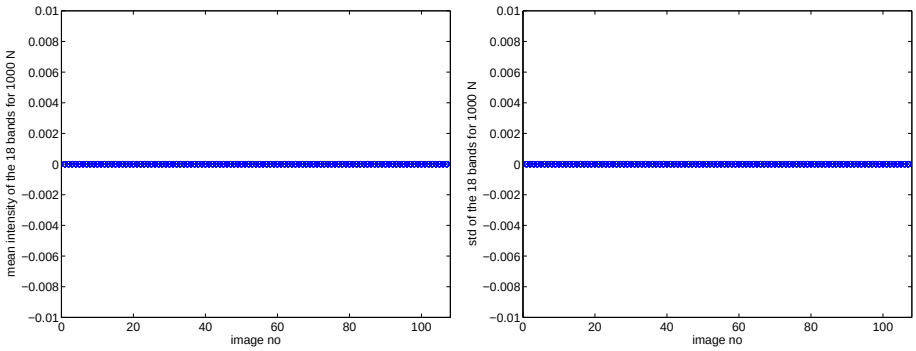
Pre-processing

This chapter describes the results of the pre-processing analyses of the images in order to extract features to use for estimation or classification. The first section illustrates the reproducibility of the images over time. The second section illustrates the results obtained from the segmentation of the fungal colonies. The following sections describe the extracted features. Five data sets are constructed from the fungi images and two data sets from the sand images.

7.1 Reproducibility

Reproducibility of the images means that if the same image is acquired at different times the results should be comparable. Previously, the equipment has been tested over a time period of seven hours, cf. Appendix A. To verify those results under the circumstances of this experiment, the reproducibility of the images over time is investigated using the 1000 NCS standard sheet chosen as background. The mean and standard deviation of sections of the background are plotted in Figure 7.1 as a function of the image number. The images were taken over approximately two hours.

There is practically no variation in the values of the background for any of the 18 spectra. Hence, the accuracy and the reproducibility of the images obtained with the equipment are satisfactory.



(a) Mean (b) Standard deviation

Figure 7.1: The mean and the standard deviation of the 1000 NCS sheet background in the images of fungi.

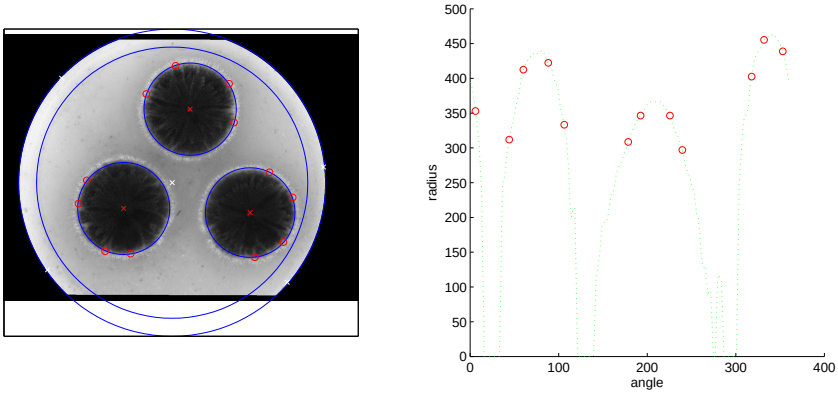
7.2 Segmentation of fungi

In this section the two segmentation methods for segmenting the colonies in the images of fungi are validated. In the first part the results of the identification of circular colonies are illustrated. In the second part the segmentation results obtained by use of Histogram Pursuit are illustrated.

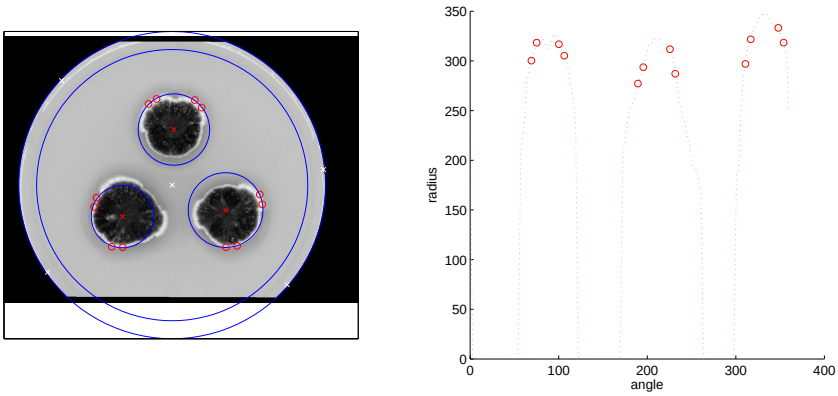
7.2.1 Identification of circular colonies

To illustrate the method, two examples from the images are given in Figure 7.2. One, where the method performs well, and one where it performs poorly. However, in the case of poor identification, the identified centers are close to the true centers of the colonies, and therefore the ROIs only include fungi, the method would in this case still work. Additional problems have arisen in the cases where the colonies have grown close to the edge of the petri dish. Such an example is illustrated in Figure 7.3. The analyzer radius if of great importance, in particular when the colonies are close to the edge of the petri dish. In Figure 7.3 two different analyzer radii have been used. When a large analyzer radius is used, two of the colonies cannot be identified because the edge of the petri dish creates light reflections that the method interprets as the light edge of the colonies. When a smaller analyzer radius is chosen, the colony that is close to the edge cannot be identified because the edge is not included in the analysis.

In the cases where the identification fails, a manual identification of the centers and

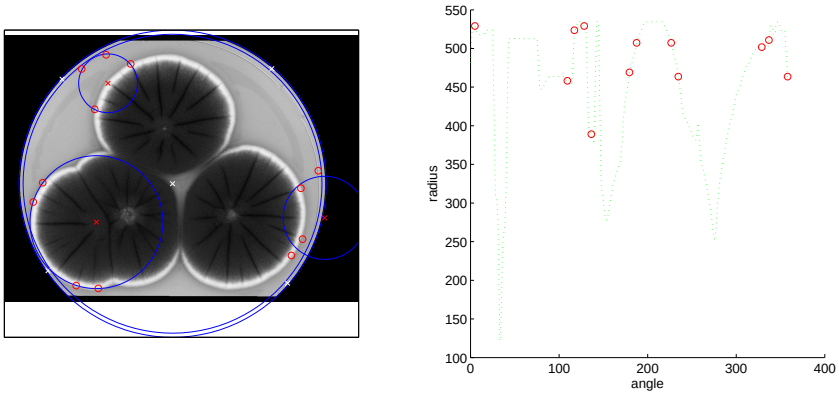


(a) Good identification

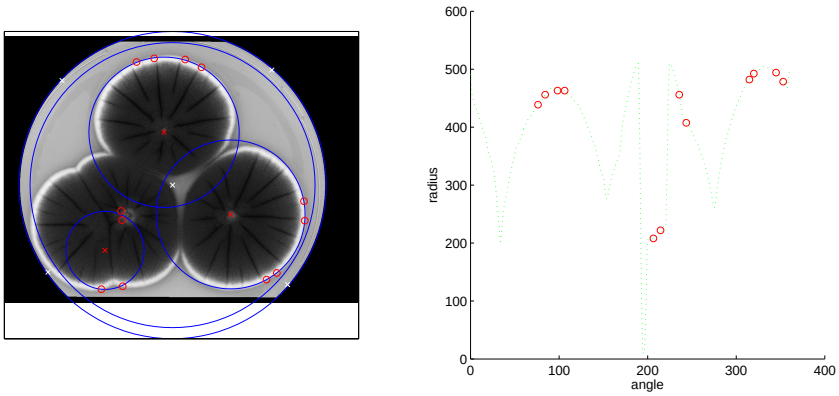


(b) Poor identification

Figure 7.2: Two examples of identification of circular colonies. Left: The 6th spectral band with the circles, the centers of the colonies, and the points on the edge of the colonies marked. Right: The distance from the detected colony to the center of the petri dish versus the angle of the scans.



(a) Large analyzer radius



(b) Small analyzer radius

Figure 7.3: One example of identification of circular colonies with two different analyzer radii. Left: The 6th spectral band with the circles, the centers of the colonies, and the points on the edge of the colonies marked. Right: The distance from the detected colony to the center of the petri dish versus the angle of the scans.

radii of the colonies were used. On the YES medium, around half of the identifications were performed manually. For the other media the method performed better.

7.2.2 Histogram Pursuit (HP)

Here, the HP algorithm[1] is used to segment the fungal colonies from the medium and the background. The first step is to segment the background, the petri dish, and the fungal colonies into three classes. The next step is to examine each of the three classes and then repetitively examine each of the subclasses obtained for further classes until a subclass no longer can be split in two or more.

The interest is to segment the colonies from the background as well as the petri dish, and if possible extract information of differences within the colonies. This is done in order to extract features to be used in a further classification of the individuals.

The subclasses obtained differ depending on the appearance of the individuals. The results are also illustrated in the article added in Appendix A. In the following subsection examples of the masks from each of the 9 groups of the 3 media and the 3 species are illustrated.

Three examples of the subclasses obtained from the HP algorithm are illustrated in Figure 7.4, for more examples see Appendix A. The colonies are well separated from both petri dish and background. Furthermore, the lighter edges and centers of the colonies can be separated. The latter might be useful since the different species differ in appearance at the edge of the colonies.

The masks for the three individuals on the YES medium in Figure A.9 are illustrated in Figure 7.5. The mask for the second repetition of isolate d of P. *melanoconidium* on YES is constructed using the projection obtained for the first repetition of same. This is done because the classes cannot be split such that the edge is detected separately. This results in the poorest mask obtained, illustrated in Figure 7.6.

Examples on the OAT medium are illustrated in Figure 7.7. The P. *polonicum* on OAT is segmented easily in all cases except one. The method does not find the edge of the third replica of isolate d. Even though an adequate projection is found automatically, the threshold is not. This is illustrated in Figure 7.8 where the histogram of the found projection is illustrated, as well as a manually chosen threshold, and the edge obtained hereby.

Examples of masks on the CYA medium are illustrated in Figure 7.9. The isolates of the three species on CYA are segmented well in all cases.

Summing up, the fungi are well separated from the media for all isolates. Furthermore, the method could separate the lighter edges from the darker centers of the fungal

[1] [Gomez 2005]

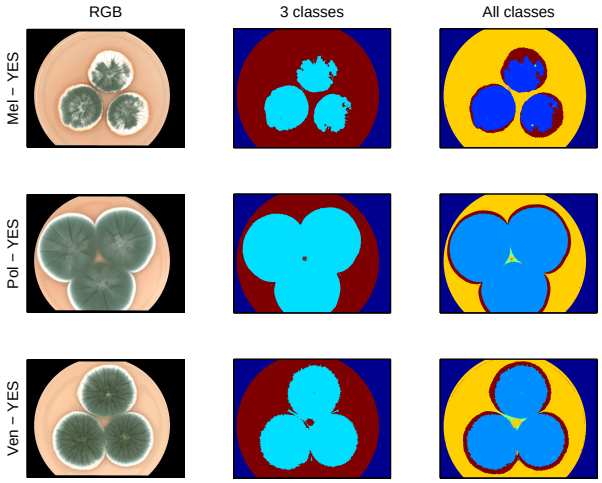
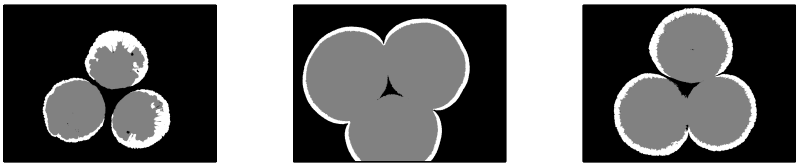


Figure 7.4: Example of segmentation of the three species on the YES medium. First column illustrates RGB representations of the multi spectral images. Second column illustrates the first segmentation into three classes. The third column illustrates the final segmentation.

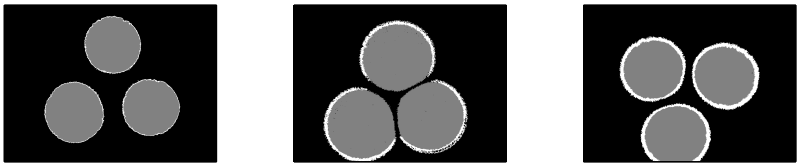


(a) Mel YES a (b) Pol YES a (c) Ven YES b

Figure 7.5: Masks for three of the individuals on YES.

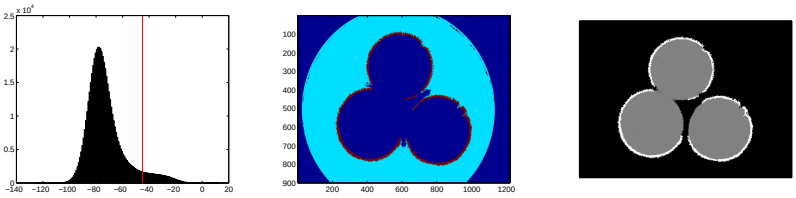
(a) Mel YES d

Figure 7.6: Mask of isolate d of P. *melanoconidium* on YES.



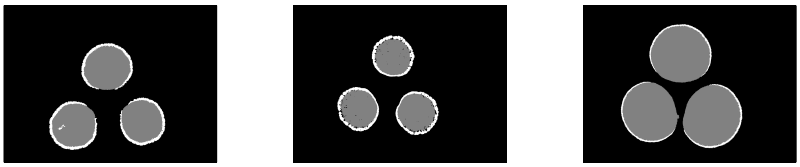
(a) Mel OAT a (b) Ven OAT a (c) Pol OAT a

Figure 7.7: Masks for three of the individuals on OAT.



(a) Histogram of projection (b) Segmented edge (c) Mask

Figure 7.8: Manually chosen threshold and the edge obtained hereby. The segmented edge is red. The medium it is segmented from is light blue and the remaining classes are dark blue. In this case the first identified fungi were separated into two classes: The middle of the medium and the colonies.



(a) Mel CYA b (b) Ven CYA c (c) Pol CYA b

Figure 7.9: Masks for three of the individuals.

colonies.

7.3 Fungi features from HP

The segmentations obtained from HP are utilized, where the ROI is the centers and the edges of the colonies as one mask. The features extracted from the ROIs are: The 1st, 5th, 10th, 30th, 50th, 70th, 90th, 95th, and 99th percentiles, the mean, the standard deviation, and the maximal intensity. The features are extracted both for the original spectra, the difference of the spectra, and the pair wise products of the spectra. Some of the features give zero for all samples and are therefore disregarded. In total there are 3754 features.

7.4 Fungi features of fungi and edge separate

In this data set the centers of the colonies and the edges of the colonies are regarded as separate masks. For the centers of the colonies the same features, as the ones described in Section 7.3, are extracted. The features extracted for the edges are: The 1st, 10th, 50th, 90th, and 99th percentiles, the mean, the standard deviation, and maximal intensity. In total there are 6219 features.

7.5 Fungi features of 10 visual bands representing RGB

This data set consists of the three linear combination of the ten visual spectra used to represent R, G and B, as illustrated in Section 5.2. The centers and the edges of the colonies are treated as one mask, and the features are the same as the ones in Section 7.3. In total there are 101 features.

7.6 Fungi features of the three bands closest to RGB

This data set consists of the three spectral bands closest to R, G and B. That is: 645nm, 505nm, and 450nm. The centers and the edges of the colonies are treated as one mask, and the features are the same as in Section 7.3. In total there are 103 features.

7.7 Spatial fungi features

This data set consits of features extracted from the ROIs identified in Section 7.2.1. The ROIs are subdivided into six sub areas, geoemtrically separated by 0.1, 0.3, 0.5, 0.7, and 0.9 times the radius of the colony. Since the fungal colonies grow from the center and outwards, the subareas reflect the age of the colony. The sub areas are illustrated in Figure 7.10. The features extracted from each sub area are: The mean, the standard deviation, the maximum intensity, and the 1st, 30th, 50th, 70th, 90th, and 99th percentiles. Each of the spectral bands are used, as well as difference images and pair wise products of the images of all spectral bands. In total there are 17496 features.

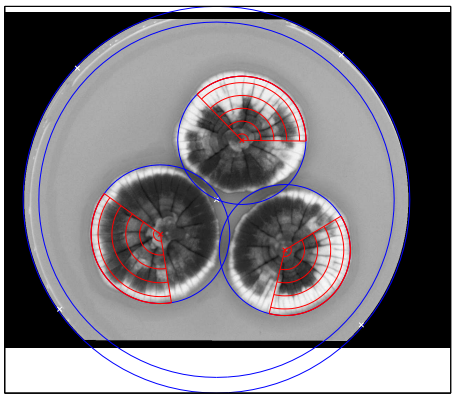


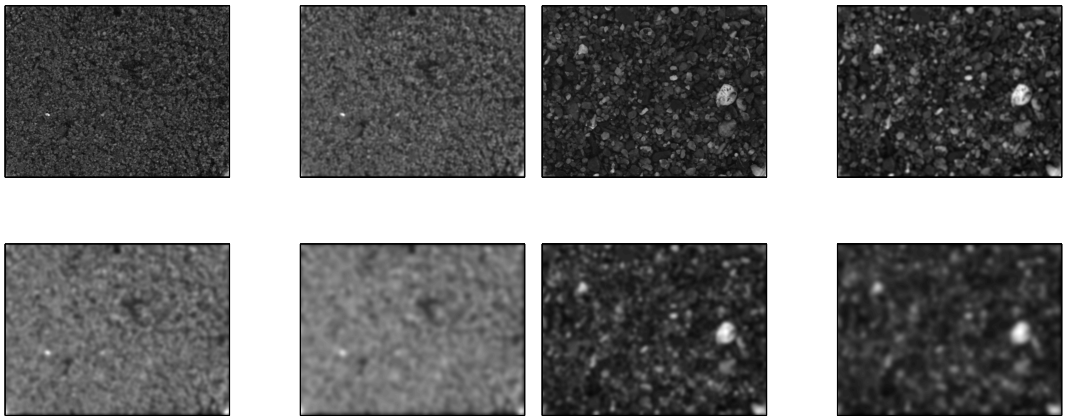
Figure 7.10: Illustration of the six sub areas of ROIs in the identification of circular colonies.

7.8 Sand features 1

For each image the features are extracted from histograms of the ROI in the 18 spectra. The features are: The mean, the standard deviation, the 1st, 5th, 10th, 50th, 90th, and 99th percentiles, and the maximum intensity in each spectrum, as well as the mean of the maximal intensities. The same features are extracted from difference and pair wise products of the 18 spectra. For some of the pair wise products of the spectra the higher percentiles are zero for all observations, and these features are disregarded, as they do not provide additional information to the classification. In total there are 667 features for each sample.

7.9 Sand features 2

The 1st, 5th, 10th, 30th, 50th, 70th, 90th, 95th, and 99th percentiles are evaluated of the original spectra, the logarithm of the spectra, the differences between the spectra, the pair wise products of the spectra, the pair wise ratios between the spectra, the opening, and the closing of the standardized image. Furthermore, scale spaces are constructed by filtering each spectral band with a Gaussian lowpass filter with standard deviations 0, 1, 2, 5, 10, 15, 20, 25, and 30. The scale spaces are illustrated in Figure 7.11. Note, the large difference between the scale spaces on the medium and large grain curve.



(a) Medium grain curve (b) Large grain curve

Figure 7.11: Illustration of scale spaces for medium and large grain curve of sand type 3. From upper left corner: standardized image of 1st spectral band, scale space image with standard deviations 5, 10, and 15.

The standard deviation, mean, kurtosis, and skewness of the scale spaces and the differences between the scale spaces are calculated. Additional features are: The mean and standard deviation of the gradient of the size fractions 1, 0.9, 0.8, 0.6, 0.4, and 0.2 of the scale space images, constructed by nearest neighbor interpolation. There are 2016 features in total.

Chapter 8
Results Fungi

This chapter describes the results obtained for the fungi data. The first section examines the ill posedness of the problems through the singular values of the data matrices. The second section illustrates the results obtained with traditional Discriminant Analysis. The third section lists the results obtained using LARS-EN with dummy variables. The fourth section describes an analysis of variance on the experiment; testing which of the effects are significantly different from zero. Finally, the fifth section examines the significance of the additional information provided by including information from an extra medium.

If nothing else is mentioned each medium is considered separately, leaving 36 observations in three equally sized classes.

In Discriminant Analysis and analysis of variance, the observations are assumed to be normally distributed. For most of the groups and the examined variables, tests of normality[1] are accepted at a 10% level of significance. Furthermore, the analyses are considered robust to small non-compliances.

8.1 Singular values

The singular values can be used as an indication of whether a problem is ill or well posed. The singular values of the four[2] data sets of features for the fungi samples on

[1] Tests of nonnormality conducted were: Shapiro-Wilk and Kolmogorov-Smirnov, cf. [NIST/SEMATECH 2006], both calculated in SAS.

[2] The data sets of spatial features is not included because p is too large.

YES are illustrated in Figure 8.1. It is seen that there is a gap in the singular values between number 36 and 37. This reveals a numerical rank of 36, corresponding to the number of observations. Furthermore, the first singular value is large compared to the second, leaving a small gap between the first and second singular values. It is therefore expected that one dimension can explain a large part of the variance in the data, and that at least 36 variables should be enough to include in the analyses. The same tendencies are illustrated for the data on OAT and CYA, cf. Appendix E, Figure E.1 and E.2.

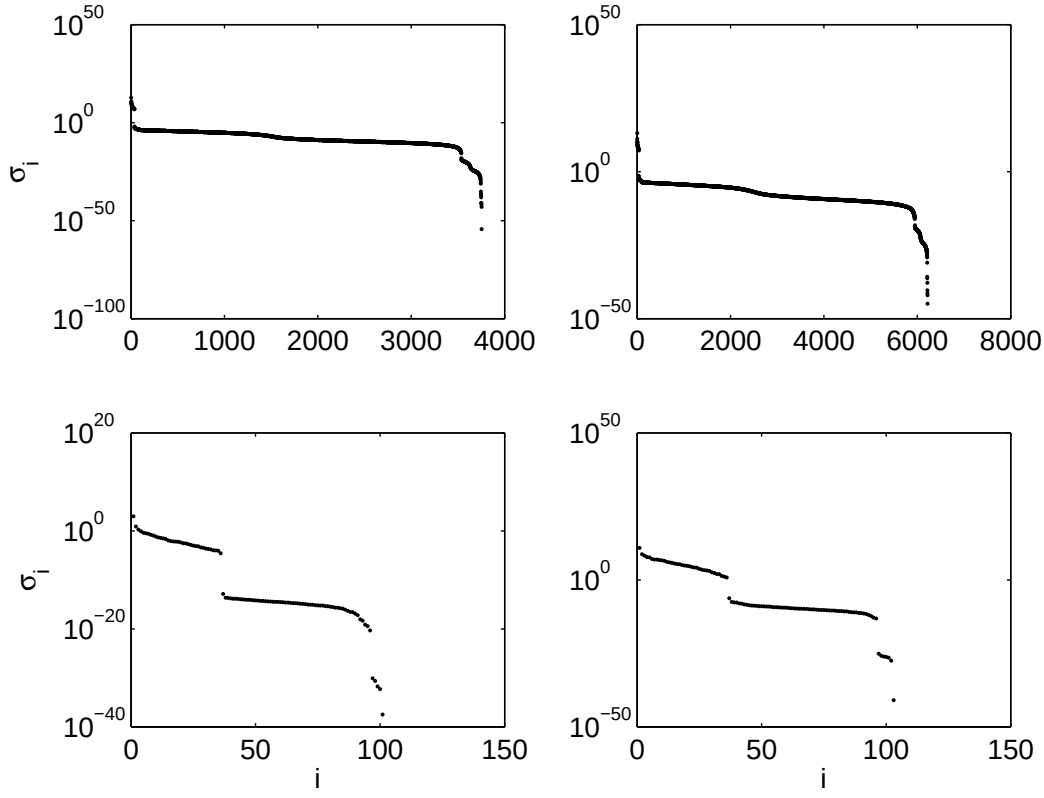


Figure 8.1: Plot of singular values for the fungi data sets on YES. From upper left corner: Features from edges and centers of the colonies together, edges and centers separated, linear combinations of the visual bands to represent RGB and the three bands closest to RGB.

In the following, if nothing else is mentioned, the data set of all spectra with the edges and centers of the colonies together is used. The reason for this is illustrated in Section 8.3.

8.2 Discriminant Analysis

Performing Discriminant Analysis requires a subset of variables or principal components in order not to over fit training data. Linear discriminant functions are used for the classification. Recall, that the linear discriminant functions assume homogeneity of variance, i.e. that the dispersion of the classes are equal. This assumption is tested with Levene's test of homogeneity[3]. Only the data set with fungi and edge as one mask is examined here. If nothing else is mentioned the results are from the data on YES.

With Forward Selection based on Wilk's Λ-tests of the original variables only two variables are needed to classify all observations correctly with leave-one-out cross-validation. These variables are the first two variables in Table 8.1. With 2-fold cross-validation, i.e. one training set of eighteen observations, and one test set of eighteen observations, DA2 is chosen for both sets, but DA1 is substituted by DA3 for one of the sets, cf. Table 8.1. Levene's test of equal variance is at a 5% level of significance accepted for the two combinations of variables.

Var	Image	Parameter	Bands (nm)
DA1	Difference	99th percentile	cyan & amber (505&590)
DA2	Difference	30th percentile	ultra blue & red (430&645)
DA3	Difference	5th percentile	ultra blue & NIR (430&870)

Table 8.1: The three variables selected according to Wilk's Λ in the Discriminant Analysis on the YES medium.

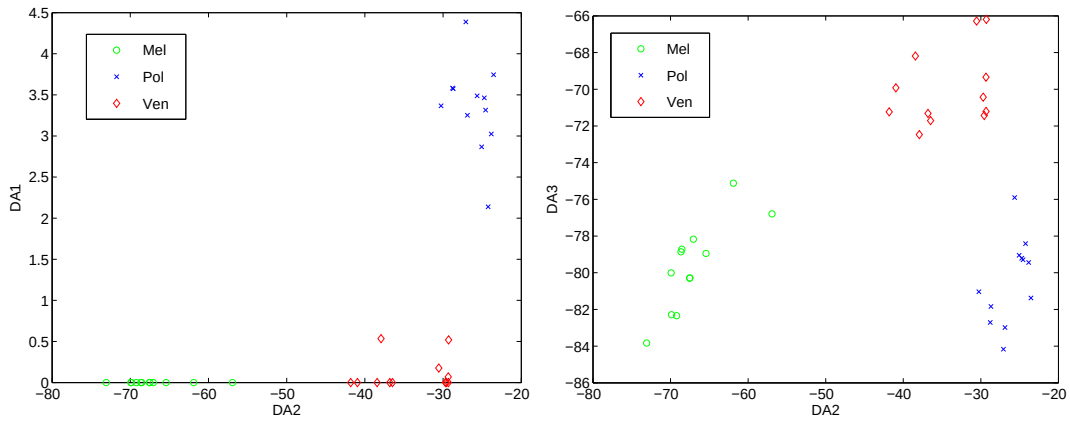
Figure 8.2 illustrates scatter plots of the three selected variables. P. *polonicum* has larger differences between cyan and amber than P. *venetum* and P. *melanoconidium*. P. *melanoconidium* has larger absolute differences between ultra blue and red than P. *polonicum* and P. *venetum*. P. *venetum* has smaller absolute differences between ultra blue and NIR(870nm) than P. *polonicum* and P. *melanoconidium*.

When only one variable is selected for each validation, six of the P. *venetum* observations are misclassified as P. *melanoconidium* (17% of all observations).

Discriminant Analysis combined with PCA requires ten PCs in order to obtain no misclassifications.

Performing Discriminant Analysis on the data on CYA and OAT the results are not as good as on YES. On CYA there are two misclassifications when ten variables are se-

[3]Levene's test is used instead of Bartlett's test of equality in variance since it is less sensitive to departures from normality, cf. [NIST/SEMATECH 2006].



(a) DA1 vs. DA2

(b) DA3 vs. DA2

Figure 8.2: Scatter plots of DA1 and DA3 versus DA2. Green: P. *melanoconidium*, blue: P. *polonicum*, and red: P. *venetum*.

lected and leave-one-out cross-validation used. On OAT there is one misclassification when ten variables are selected and leave-one-out cross-validation used.

8.3 LARS-EN with dummy variables

The LARS-EN method with dummy variables is used to identify the three species in this section. Both leave-one-out, 6-fold, and 2-fold CV as well as different λs and numbers of iterations have been tested for each of the three media. Furthermore, the different data sets are compared on the YES medium.

The test and train results are illustrated in Figure 8.4 and 8.5 for the YES medium. The results are satisfactory. The YES medium is best in the sense that when this medium is used, fewer features are needed in order to obtain no misclassifications. Only two features for each dummy variable (in total six variables) are necessary in that situation for both leave-one-out and 6-fold CV. If only one variable is used to regress each dummy variable, i.e. three variables in total, the error rate is 3%, or only one misclassification. The three selected variables are listed in Table 8.2. Note, that only five of the spectral bands are included: Ultra blue, cyan, amber, red, and NIR(870nm). Note, that EN1=DA2, EN2=DA3, and EN3=DA1. The values of the variables for the three species are illustrated in Figure 8.3. It is seen that EN1 discriminates P. *melanoconidium* with respect to the two other species, EN2 discriminates P. *venetum*,

and EN3 discriminates P. *polonicum*.

Var	Specie	Image	Parameter	Bands (nm)
EN1	Mel	Difference	30th percentile	ultra blue & red (430&645)
EN2	Ven	Difference	10th percentile	ultra blue & NIR (430&870)
EN3	Pol	Difference	99th percentile	cyan & amber (505&590)

Table 8.2: The variables selected by LARS-EN for the YES medium with leave-one-out CV and maximal one feature for each validation. Specie describes which of the three species is discriminated from the remaining. EN1=DA2, EN3=DA1 and EN3=DA3.

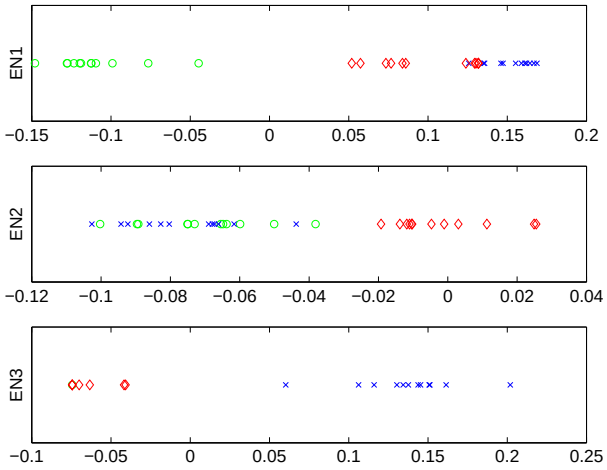
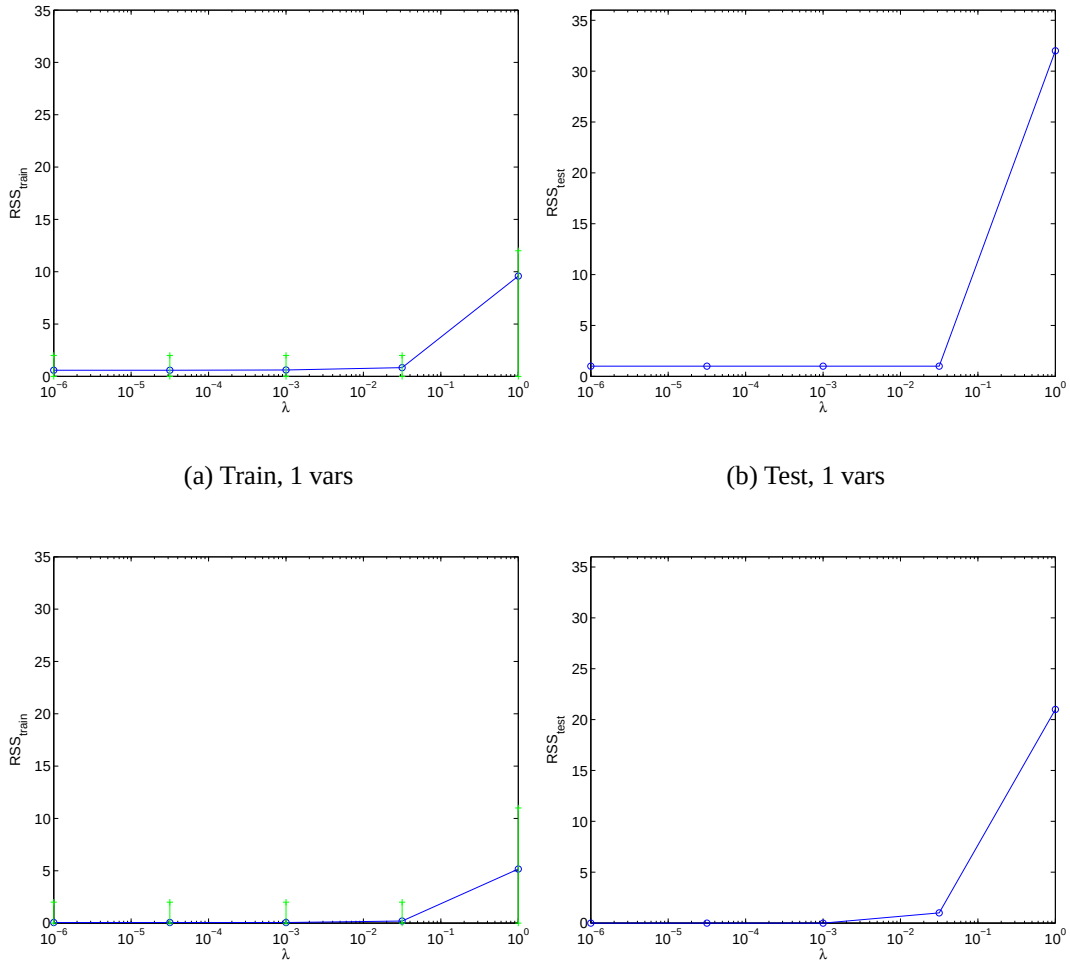


Figure 8.3: Plots of the three variables selected by LARS-EN; EN1, EN2 and EN3. Green: P. *melanoconidium*, blue: P. *polonicum*, and red: P. *venetum*.

Classification with 2-fold cross-validation was also conducted, i.e. the data was split in two sets. Including just one variable all observations can be classified correctly, but it depends on how data is split. The error rate might be higher (around nine misclassifications or 25%). The results for two different partitionings of data are illustrated in Appendix E, Figure E.3. In general, the P. *melanoconidium* specie is not a problem to classify even if an isolate is not represented in both parts of a partitioning. The P. *polonicum* and P. *venetum* species are more delicate. For the different partitionings, different sets of variables are selected, i.e. LARS-EN with dummy variables is very sensitive to the training data.

The observations have been classified belonging to the closest class. Classifying the



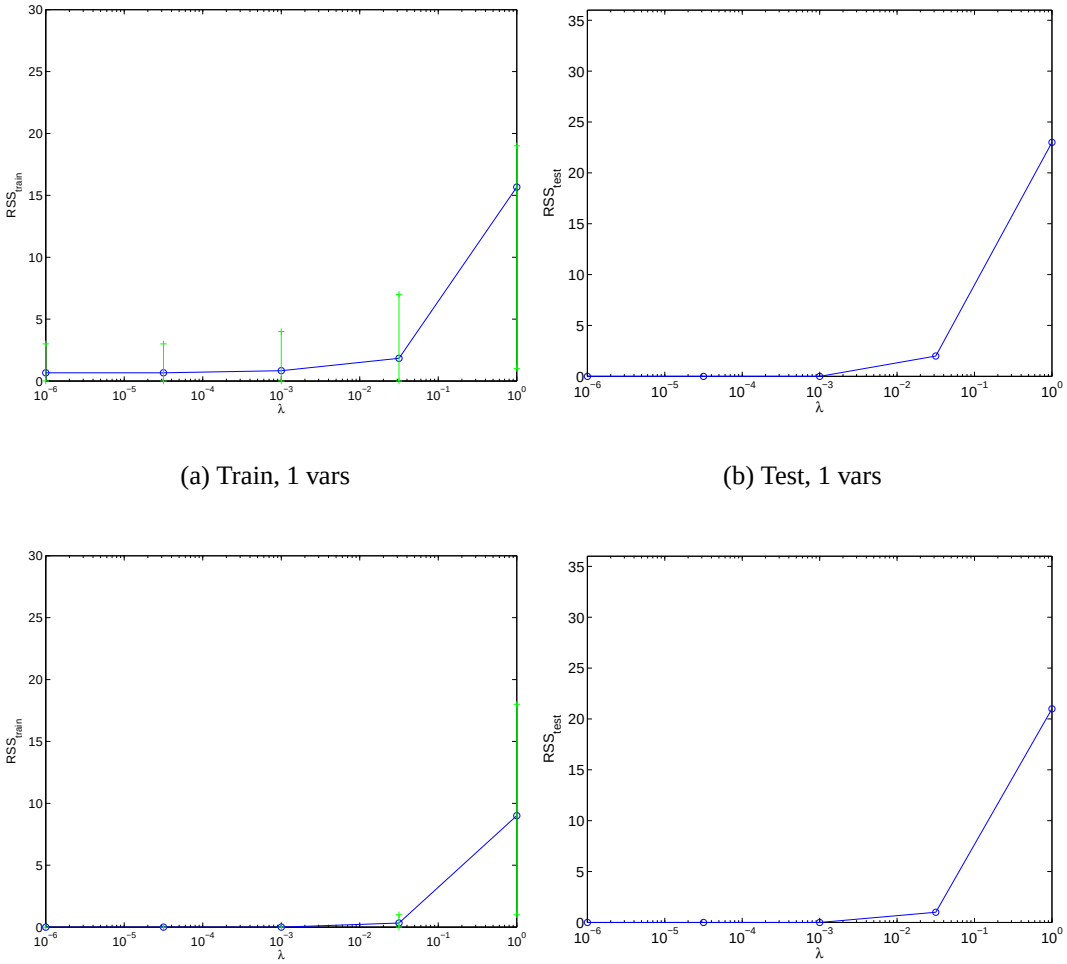
(a) Train, 1 vars

(b) Test, 1 vars

(c) Train, 2 vars

(d) Test, 2 vars

Figure 8.4: Misclassifications for leave-one-out CV on YES medium.



(a) Train, 1 vars

(b) Test, 1 vars

(c) Train, 2 vars

(d) Test, 2 vars

Figure 8.5: Misclassifications for 6-fold CV on YES medium.

observations as either belonging to a class, or being unclassified[4] gives a higher error rate. In this case, five variables are needed to classify all observations correctly.

The OAT and CYA media require more features to obtain the same error rates as the ones obtained on YES. With nine variables they yield two to four misclassifications (6-11%). On the OAT and CYA media the results of LARS-EN are illustrated in Appendix E, Figure E.4 to E.7.

Comparing the results on the YES medium with those obtained only using information from three spectral bands; the ones closest to R, G and B, illustrates that multi-spectral images are an advantage. For the three bands: 645, 505 and 450nm, four variables are required to obtain no misclassifications for both leave-one-out and 6-fold CV. The results are illustrated in Figure E.8 and E.9 in Appendix E.

Similar results are obtained if three images representing R, G and B, representations which are linear combinations of the ten visual spectral bands. Only three variables are needed to obtain low error rates (2-4 misclassifications or 6-11%), and if six variables are used there are no misclassifications. However, for the entire data set only five spectral bands have been utilized. The results are illustrated in Figure E.10 and E.11 in Appendix E.

Separating the edges and the centers of the colonies give additional features, but it does not improve the classification. Three variables are required to classify all the observations correctly. The results are illustrated in Figure E.12 and E.13 in Appendix E.

Finally, the spatial features, obtained from identification of circular colonies, result in two misclassifications (6%) when two variables are selected. The results are illustrated in Figure E.14 in Appendix E.

8.4 Three-sided analysis of variance

Considering an analysis of variance in this experiment, there are four factors: Medium (M), specie (S), isolate (I), and repetition (R). Similar analyses are described in [Conradsen 2002*a*, sec. 5.4] and [Rencher 2002, sec. 6.6.2].

Specie with respect to medium and isolate with respect to medium are cross classifications where as specie, medium, and isolate with respect to repetition, and specie with

[4] Option b in Section 6.4.6.

respect to isolate are hierarchical classifications. That is:

$$\begin{array}{ccc} S & \times & M \\ I & \times & M \end{array} \qquad \begin{array}{ccc} S & \supset & I \\ S & \supset & R \\ I & \supset & R \\ M & \supset & R \end{array} \quad .$$

The factors medium and specie are deterministic and the effects caused by these are denoted with lower case letters. The repetition factor is random and the effect caused by this factor is therefore denoted with an upper case letter. It can be argued whether the isolate factor is indeed deterministic or stochastic. As the isolates are chosen to represent a large geographic region and can be reproduced, they could be regarded deterministic. On the other hand, if another laboratory was to reproduce the experiment the isolates might not be the same and the factor could then be regarded as stochastic.

Two models are investigated: A deterministic model where isolate is deterministic, and a mixed model where isolate is stochastic.

The models are

$$X_{klj\nu} = \mu + m_k + s_l + ms_{kl} + i(s)_{j(l)} + mi(s)_{kj(l)} + R(msi)_{\nu(klj)} \tag{8.2}$$

and

$$X_{klj\nu} = \mu + m_k + s_l + ms_{kl} + I(s)_{j(l)} + mI(s)_{kj(l)} + R(msI)_{\nu(klj)} \quad , \tag{8.3}$$

where

$$\begin{aligned} k &= 1,\ldots,3 \quad (\text{medium}) \\ l &= 1,\ldots,3 \quad (\text{specie}) \\ i &= 1,\ldots,4 \quad (\text{isolate within specie}) \\ \nu &= 1,\ldots,3 \quad (\text{repetition within medium, specie and isolate}) \quad . \end{aligned}$$

Note, that the interaction between the stochastic isolate effect and the deterministic medium effect in Model (8.3) is a stochastic term.

8.4.1 Univariate analysis of variance

The Sums of Squares (SS) for the factors are calculated by use of the formulas in Table D.1 in Appendix D.2 and the results are listed in Table 8.3 for one of the variables

Variation	SS	f	SS/f
M	$8.69 \cdot 10^1$	2	$4.34 \cdot 10^1$
S	$5.48 \cdot 10^1$	2	$2.75 \cdot 10^1$
MS	$6.23 \cdot 10^1$	4	$1.56 \cdot 10^1$
I(S)	$1.36 \cdot 10^1$	9	$1.51 \cdot 10^0$
MI(S)	$2.19 \cdot 10^1$	18	$1.22 \cdot 10^0$
R(MSI)	$1.26 \cdot 10^1$	72	$1.75 \cdot 10^{-1}$
Total	$2.52 \cdot 10^2$	107	$2.36 \cdot 10^0$

Table 8.3: ANOVA for the 99th percentile of the difference between 4th and 6th spectra of the data set with fungi and edge in one, DA1.

selected in the Discriminant Analysis, DA1. For the other variables selected in the Discriminant Analysis, and the first two PCs, the results are given in Appendix E, Table E.1 to E.4.

Tests of the following null-hypotheses are conducted: That the variance of the stochastic terms are zero, and that each of the other effects are insignificant. In Appendix D.2, Table D.2 and D.3 the expected values of the SS of each effect, as well as the error effect to test against, are listed for the two models. The effect to test against is the one with the same expected SS except for the variance of the tested effect. For example if M is to be tested against R(MSI) the test size becomes $\frac{SS_M/f_M}{SS_{R(MSI)}/f_{R(MSI)}}$.

The tests corresponding to Model (8.2) are listed in Table 8.4 and in Appendix E, Table E.5 to E.8. The tests corresponding to Model (8.3) are listed in Table 8.5 and in

H_0	Test Size	F-fractile
$m_k = 0,\ k = 1,2,3$	$\frac{4.34 \cdot 10^1}{1.75 \cdot 10^{-1}} = 248$	$F(2,72)_{0.99}$ =4.91
$s_l = 0,\ l = 1,2,3$	$\frac{2.75 \cdot 10^1}{1.75 \cdot 10^{-1}} = 157$	$F(2,72)_{0.99}$ =4.91
$ms_{kl} = 0,\ k = 1,2,3,\ l = 1,2,3$	$\frac{1.56 \cdot 10^1}{1.75 \cdot 10^{-1}} = 89.0$	$F(4,72)_{0.99}$ =3.59
$i(s)_{j(l)} = 0,\ j = 1,2,3,4,\ l = 1,2,3$	$\frac{1.51 \cdot 10^0}{1.75 \cdot 10^{-1}} = 8.63$	$F(9,72)_{0.99}$ =2.66
$mi(s)_{kj(l)} = 0,\ k = 1,2,3,$ $j = 1,2,3,4,\ l = 1,2,3$	$\frac{1.22 \cdot 10^0}{1.75 \cdot 10^{-1}} = 6.97$	$F(18,72)_{0.99}$=2.20

Table 8.4: Tests based on Model (8.2) for the 99th percentile of the difference between 4th and 6th spectra of the dataset with fungi and edge in one, DA1.

Appendix E, Table E.9 to E.12.

First, Model (8.2) is examined where the isolate effect is deterministic. The results are summarized in Table 8.6. The null-hypotheses that we cannot distinguish between

H_0	Test Size	F-fractile
$m_k = 0,\ k = 1,2,3$	$\frac{4.34 \cdot 10^1}{1.22 \cdot 10^0} = 35.7$	$F(2,18)_{0.99}$ =6.01
$s_l = 0,\ l = 1,2,3$	$\frac{2.75 \cdot 10^1}{1.51 \cdot 10^0} = 18.2$	$F(2,9)_{0.99}$ =8.02
$ms_{kl} = 0,\ k = 1,2,3,\ l = 1,2,3$	$\frac{1.56 \cdot 10^1}{1.22 \cdot 10^0} = 12.8$	$F(4,18)_{0.99}$ =4.57
$\sigma^2_{I(s)} = 0$	$\frac{1.51 \cdot 10^0}{1.22 \cdot 10^0} = 1.24$	$F(9,18)_{0.67}$ =1.24
$\sigma^2_{mI(s)} = 0$	$\frac{1.22 \cdot 10^0}{1.75 \cdot 10^{-1}} = 6.97$	$F(18,72)_{0.99}$=2.20

Table 8.5: Tests based on Model (8.3) for the 99th percentile of the difference between 4th and 6th spectra of the dataset with fungi and edge in one, DA1.

media, species and isolates and their interactions are all rejected at a 3% level of significance.

H_0 /Variable	DA1(EN3)	DA2(EN1)	DA3(EN2)	PC1	PC2
$m_k = 0$	R(1%)	R(1%)	R(1%)	R(1%)	R(1%)
$s_l = 0$	R(1%)	R(1%)	R(1%)	R(1%)	R(1%)
$ms_{kl} = 0$	R(1%)	R(1%)	R(1%)	R(1%)	R(2%)
$i(s)_{j(l)} = 0$	R(1%)	R(1%)	R(1%)	R(1%)	R(1%)
$mi(s)_{kj(l)} = 0$	R(1%)	R(1%)	R(1%)	R(1%)	R(3%)

Table 8.6: Summing up the tests based on Model (8.2). A indicates the null-hypothesis is accepted and R rejected at a 5% level of significance. In parentheses is given the level of significance where the acceptance or rejection still holds. The variables DA1 and DA2 are the ones selected in Discriminant Analysis, EN1, EN2 and EN3 are the first three variables selected with LARS-EN, and PC1 and PC2 are the first two principal components.

When Model(8.3) is examined where the isolate effect is stochastic we still at a 3% level of significance reject the null-hypotheses related to the effects of media and interactions between media and isolates for all variables regarded. The results are summed up in Table 8.7. For all variables the two following null-hypothesis is accepted: That there is no significant difference between isolates.significant effect of the repetitions.

For the EN2 variable the null-hypothesis of no difference between species is accepted. Figure 8.3 thus also shows that for this variable two of the species cannot be distinguished. For all remaining variables the null-hypothesis is rejected. For the PCs and EN2 the hypothesis of no interaction effect between media and species is accepted. For DA1 and DA2, on the other hand, this null-hypothesis is strongly rejected.

Those results seem promising for the analyses made where it is desired to distinguish

H_0 /Variable	DA1(EN3)	DA2(EN1)	DA3(EN2)	PC1	PC2
$m_k = 0$	R(1%)	R(1%)	R(3%)	R(1%)	R(1%)
$s_l = 0$	R(1%)	R(1%)	A(17%)	R(1%)	R(1%)
$ms_{kl} = 0$	R(1%)	R(1%)	A(13%)	A(16%)	A(17%)
$\sigma^2_{I(s)} = 0$	A(33%)	A(69%)	A(52%)	A(25%)	A(23%)
$\sigma^2_{mI(s)} = 0$	R(1%)	R(1%)	R(1%)	R(1%)	R(3%)

Table 8.7: Summing up the tests based on Model (8.3). A indicates the null-hypothesis is accepted and R rejected at a 5% level of significance. In parentheses is given the level of significance where the acceptance or rejection still holds. The variables DA1 and DA2 are the ones selected in Discriminant Analysis, EN1, EN2 and EN3 are the first three variables selected with LARS-EN, and PC1 and PC2 are the first two principal components.

between species but not necessarily isolates. Furthermore, the results are promising if other isolates within the three species are desired classified.

8.4.2 Multivariate analysis of variance

The expansion from one to more dimensions is straight forward, the SS become variance matrices where mean values and singleton observations are replaced with mean vectors and observation vectors in the formulas. The test statistic changes to Wilk's Λ which was described in Section 6.2.2. If for example M is to be tested against R(MSI) the test statistic becomes $\frac{\det(RSS_{R(MSI)})}{\det(RSS_{R(MSI)}+RSS_M)}$ which is U-distributed instead of F-distributed. A transform from the U- to the F-distribution is utilized, as described in Appendix D.1.

The tests performed here are limited by the error degrees of freedom which in this case are two, since there are three repetitions. Hence, only two variables can be used in a multivariate analysis of variance before the examined matrices become singular and a solution thus becomes impossible.

The variables selected in the Discriminant Analysis and in LARS-EN are used as bases, but also the first two PCs are used as basis for the reasons described in the following. The results for DA1 and DA2 are listed in Table 8.8 and 8.9, and for the remaining sets of variables in Appendix E, Table E.13 to E.18.

H_0	U	q	r	F	F-fractile
$\boldsymbol{m}_k = \mathbf{0},\ k = 1,2,3$	0.0136	2	72	269	$F(4,142)_{0.99}$ =3.45
$\boldsymbol{s}_l = \mathbf{0},\ l = 1,2,3$	0.0305	2	72	168	$F(4,142)_{0.99}$ =3.45
$\boldsymbol{ms}_{kl} = \mathbf{0}$, $k = 1,2,3,\ l = 1,2,3$	0.0117	4	72	146	$F(8,142)_{0.99}$ =2.64
$\boldsymbol{i(s)}_{zj(l)} = \mathbf{0}$, $j = 1,2,3,4,\ l = 1,2,3$	0.135	9	72	13.5	$F(18,142)_{0.99}$=2.06
$\boldsymbol{mi(s)}_{zkj(l)} = \mathbf{0},\ k = 1,2,3,$ $j = 1,2,3,4,\ l = 1,2,3$	0.0496	18	72	13.8	$F(36,142)_{0.99}$=1.77

Table 8.8: Tests based on the multivariate version of Model (8.2) and the variables; 30th percentile of the difference between 1st and 8th spectra and 99th percentile of difference between 4th and 6th spectra. The correlation between the variables is $\rho = 0.92$. DA1 & DA2.

H_0	U	q	r	F	F-fractile
$\boldsymbol{m}_k = \mathbf{0},\ k = 1,2,3$	0.0509	2	18	29.2	$F(4,43)_{0.99}$ =3.93
$\boldsymbol{s}_l = \mathbf{0},\ l = 1,2,3$	0.0422	2	9	15.5	$F(4,16)_{0.99}$ =4.77
$\boldsymbol{ms}_{kl} = \mathbf{0}$, $k = 1,2,3,\ l = 1,2,3$	0.0779	4	18	11.0	$F(8,34)_{0.99}$ =3.09
$\boldsymbol{\sigma}^2_{i(s)} = \mathbf{0}$	0.440	9	18	0.96	$F(18,34)_{0.48}$ =0.96
$\boldsymbol{\sigma}^2_{mi(s)} = \mathbf{0}$	0.0496	18	72	13.8	$F(36,142)_{0.99}$=1.77

Table 8.9: Tests based on the multivariate version of Model (8.3) and the variables; 30th percentile of the diference between 1st and 8th spectra and 99th percentile of the difference between 4th and 6th spectra. The correlation between the variables is $\rho = 0.92$.

PCA is performed to obtain dimensions that describe the variance in data better. LARS-EN is used to select the two principal components that describe each of the factors best. Dummy variables that represent the factor levels are constructed for the factors and the interactions between the factors. Two principal components are not sufficient to discriminate between either of the group effects. The first and second principal components are the ones selected more frequently for all effects. Multivariate analysis of variance is therefore performed on the first two PCs. The projections that explain most of the variance in original data are hence the ones most correlated with the effect dummy variables. This seems to be a good basis for testing whether the effects are significant or not.

When the isolate effect is considered deterministic the the hypotheses that each of the other effects are insignificant are rejected at a 5% level of significance for all bases, cf. Table 8.10.

H_0 /Variable	DA1 & DA2	DA1 & DA3	DA2 & DA3	PC1 & PC2
$m_k = 0$	R(1%)	R(1%)	R(1%)	R(1%)
$s_l = 0$	R(1%)	R(1%)	R(1%)	R(1%)
$ms_{kl} = 0$	R(1%)	R(1%)	R(1%)	R(1%)
$i(s)_{j(l)} = 0$	R(1%)	R(1%)	R(5%)	R(1%)
$mi(s)_{kj(l)} = 0$	R(1%)	R(1%)	R(1%)	R(1%)

Table 8.10: Summing up the multivariate tests based on Model (8.2). A indicates the null-hypothesis is accepted and R rejected at a 5% level of significance. In parentheses is given the level of significance where the acceptance or rejection still holds. The variables DA1, DA2, and DA3 are the ones selected in Discriminant Analysis as well as LARS-EN and PC1 and PC2 are the first two principal components.

As in the univariate analysis of variance an important difference is observed when the isolate effect is considered stochastic instead of deterministic. The results are summed up in Table 8.11. It is then found that there is no significant difference between isolates, but there is still a significant difference between species and media. This leads us to assume that the experiment can be conducted for other isolates within the three species.

H_0 /Variable	DA1 & DA2	DA1 & DA3	DA2 & DA3	PC1 & PC2
$m_k = 0$	R(1%)	R(1%)	R(1%)	R(1%)
$s_l = 0$	R(1%)	R(1%)	R(1%)	R(1%)
$ms_{kl} = 0$	R(1%)	R(1%)	R(1%)	R(8%)
$\sigma^2_{I(s)} = 0$	A(52%)	A(54%)	A(93%)	A(18%)
$\sigma^2_{mI(s)} = 0$	R(1%)	R(1%)	R(1%)	R(1%)

Table 8.11: Summing up the multivariate tests based on Model (8.3). A indicates the null-hypothesis is accepted and R rejected at a 5% level of significance in general with one rejection at an 8% level. In parentheses is given the level of significance where the acceptance or rejection still holds. The variables DA1, DA2, and DA3 are the ones selected in Discriminant Analysis as well as LARS-EN and PC1 and PC2 are the first two principal components.

8.5 Tests for media

This section calculates Mahalanobi's distance between species, conducts Hotelling's T^2-test for equal means, and tests for additional information provided by each medium to the discrimination of species.

Seven PCs are included in the analyses as the covariance matrices becomes close to singular if more PCs are included. The PCs are chosen since they do not favor a particular medium like the variables chosen in LARS-EN and the Discriminant Analysis do.

Performing Discriminant Analysis with a linear discriminant function on the first seven PCs all observations are classified correctly for the YES and OAT media, but for the CYA medium two P. *venetum* observations are misclassified. No cross-validation has been performed.

Mahalanobi's distance between species is calculated for all combinations of media and listed in Table 8.12. All distances are significant, i.e. Hotelling's T^2-test[5] where the null-hypothesis that two means of the species are equal are rejected. Hotelling's T^2-test assumes that the classes have equal dispersion. Levene 's test of equality in variance[6] rejects that the covariance matrices are equal at a 5% level of significance. However, Hotelling's T^2-tests are considered anyway.

Medium/Distance	Mel-Pol	Mel-Ven	Pol-Ven
YES	216 (<1%)	53 (<1%)	140 (<1%)
OAT	73 (<1%)	54 (<1%)	19 (<1%)
CYA	41 (<1%)	35 (<1%)	21 (<1%)
YES & OAT	1582 (<1%)	142 (<1%)	480 (<1%)
YES & CYA	763 (<1%)	356 (<1%)	1410 (<1%)
OAT & CYA	217 (<1%)	642 (<1%)	195 (<1%)
YES & OAT & CYA	3710 (2%)	4440 (1%)	3669 (2%)

Table 8.12: Mahalanobi's distances between the species for each of the media. The calculations are based on the first seven PCs. For the features of the edge and fungi in one. In parentheses are given the p-values of Hotelling's T^2-test of the null-hypothesis that the means of the two species are equal.

The distances between species on the different media are larger on the YES medium

[5]Hotelling's T^2-test is reviewed in Appendix D.3.

[6]Levene's test is used instead of Bartlett's test of equality in variance since it is less sensitive to departures from normality, cf. [NIST/SEMATECH 2006].

than the other two media. However, the distance between P. *melanoconidium* and P. *venetum* is largest on the OAT medium.

Each medium is now regarded as additional information to the same observation. One observation then has p variables belonging to the YES medium, p variables belonging to the OAT medium, and p variables belonging to the CYA medium[7]. Tests of the null-hypothesis that a medium does not contribute to the discrimination compared to one or two media are conducted. This test corresponds to a test of the last p variables belonging to the same medium do not contribute to the discrimination[8]. The test statistics are compared to fractiles in the F(12,20) and F(12,13) distributions for the base with three and two media, respectively. The p-values of these tests are listed in Table 8.13.

Base Media	Distance	Test Medium		
		YES	OAT	CYA
YES & OAT & CYA	Mel-Pol	20%	58%	89%
YES & OAT	Mel-Pol	<1%	<1%	-
YES & CYA	Mel-Pol	<1%	-	5%
OAT & CYA	Mel-Pol	-	2%	10%
YES & OAT & CYA	Mel-Ven	44%	26%	12%
YES & OAT	Mel-Ven	17%	16%	-
YES & CYA	Mel-Ven	<1%	-	<1%
OAT & CYA	Mel-Ven	-	<1%	<1%
YES & OAT & CYA	Pol-Ven	18%	84%	40%
YES & OAT	Pol-Ven	<1%	6%	-
YES & CYA	Pol-Ven	<1%	-	<1%
OAT & CYA	Pol-Ven	-	<1%	<1%

Table 8.13: P-values for tests of the null-hypothesis that the variables belonging to the test medium do not contribute to the discrimination. For the features of the edges and the centeres of the fungal colonies in one.

Discriminating between any two of the species, all hypotheses that one of the media do not contribute compared to the other two are accepted 10% level of significance. Hence, leaving two media that contribute. Discriminating between P. *melanoconidium* and P. *venetum*, it is accepted (at a 16% level of significanca) that OAT and YES do not contribute to the discrimination with respect to each other.

[7]Recall, that $p = 7$.

[8]The test is reviewed in App. D.4

8.6 Summing up and discussion

The discussion is divided into three parts: Results obtained for the classification of the three species, comparison of data sets, and comparison between Discriminant Analysis and LARS-EN with dummy variables.

Identification of *Penicillium* fungi

The three species can be classified correctly by use of only two variables in Discriminant Analysis. That is, differences between cyan and amber, and ultra blue and red are enough to distinguish between the species. The amount of cross-validation can be discussed, however, as only two variables are utilized and the discriminant functions are linear in the analysis, the amount of over fitting ought to be minimal.

Three-sided analysis of variance shows that the effects: Media, species, and their interactions are significant at a 5% level. The isolate effect is statistically significant if it is considered deterministic, but insignificant if it is considered stochastic. Hence, it can be assumed that the three species can be distinguished if the experiment is repeated with the same as well as other isolates.

Mahalanobi's distances between species are significantly different from zero, which underlines the fact that they can be discriminated. The distances illustrate that the visual appearance, and not the genetic relation, is strongest as the smallest distances are observed between P. *polonicum* and P. *venetum*.

Furthermore, it can be assumed, statistically, that one of the three media does not contribute further to the discrimination compared to using the two other media. Hence, using two media should be sufficient. Since the distances have been largest on YES, and YES and OAT, statistically, can be assumed not to contribute to the discrimination with respect to each other, the best choice of media must be YES and CYA. However, as it was seen for both Discriminant Analysis and LARS-EN, the species can be discriminated using just one medium. The YES medium is the one that gives the best results for the classification.

Choice of data set

The masks where the edges and the centers of the colonies are treated as one have provides better features than if they are treated separately. Furthermore, the multispectral images are an advantage to RGB as fewer variables are required when all spectra are included. The features from the RGB representation obtained by linear combinations of the ten visual spectra performs slightly worse, though comparable, to using all spectra separately. However, the features selected contain information from

ten spectra whereas the features selected from all spectra separately only uses five of the spectral bands.

Comparison of methods

LARS-EN with dummy variables is more sensitive to which observations are in the test respective training sets for few-fold cross-validation, e.g. 2-fold cross-validation, compared to Discriminant Analysis. Furthermore, the Discriminant Analysis discriminates between all species at the same time, and not as LARS-EN between one class and remaining classes. That is each variable is used to discriminate between all species in Discriminant Analysis. Hence, the Discriminant Analysis only requires two variables to classify all observations correctly, compared to at least three with LARS-EN (corresponding to one misclassification with leave-one-out CV).

Chapter 9
Results Sand

This chapter describes the results obtained of the estimations of the moisture content in the sand samples.

As the knowledge of the sand type is a priori, it seemss reasonable to make a model for each sand type. Models for each sand type are selected and compared using: Forward Selection combined with OLS, PCA combined with Forward Selection and OLS, Ridge regression, Lasso regression, LARS-EN, and sparse PCs.

The number of observations is not proportional for the three grain curves within each sand type. Hence, when making one model for one sand type more weight will be put to the medium grain curve, as there are more observations on this. Therefore, models for each grain curve are also examined.

The first section describes the reason for transforming the dependent variable. The second section illustrates how the images can be used to identify sand type and grain curve. The third section illustrates that some of the problems are ill posed. The fourth section examines models for the five sand types for different model selection techniques. The fifth section examines models for the grain curves. The sixth section briefly describes the features that are included in the models. Finally, the seventh section sums up and discusses the results obtained.

9.1 Logarithmic transformation

In general, the variance of the moisture content tends to increase with the moisture content. The measured moisture content observations have therefore been logarithmi-

cally transformed, and this have provided better results. In Figure 9.1 the residuals are illustrated of OLS on sand type 1, fine grain curve with and without a logarithmic transform of the measured moisture content.

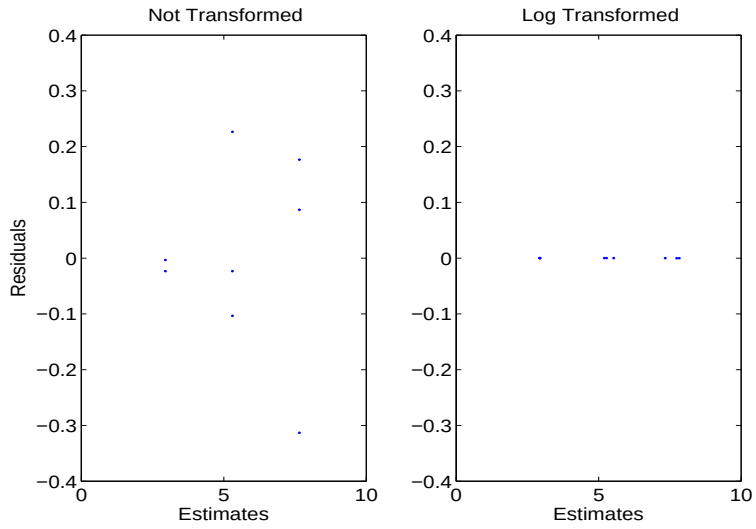


Figure 9.1: Residuals from OLS on 10 variables selected with Forward Selection.

The variance of the residuals increases with the size of the estimates, it is known that this trebd can be reduced by transforming the dependent variables, cf. [Conradsen 2002*a*, Chapt. 4]. Data shows less trends in the residuals and yields lower standard deviations when the moisture content observations are logarithmically transformed. The tendencies are caused by a larger variance for the higher moisture content measures. However, the trends are not as obvious as in Figure 9.1 for all groups of the data. In the following only the logarithmic transformed moisture content measures will be regarded.

9.2 Sand types and grain curves

The five sand types can be discriminated entirely based on the first two canonical variables, as illustrated in Figure 9.2. As the sand types can be discriminated visually, and as they physically are gathered from five distinct geographic places, it is reasonable to choose models for the five sand types separately.

Figure 9.3 illustrates the first two canonical variables discriminating between grain curves in the three sand types 1, 3, and 5. It is possible to discriminate between grain

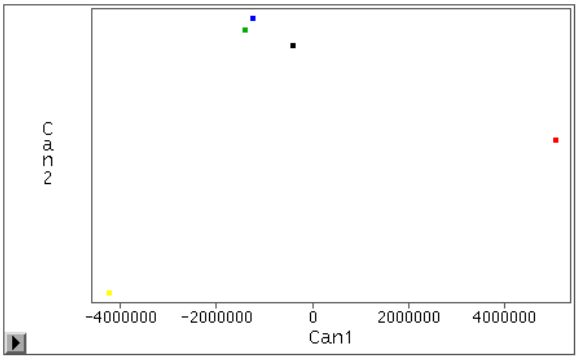
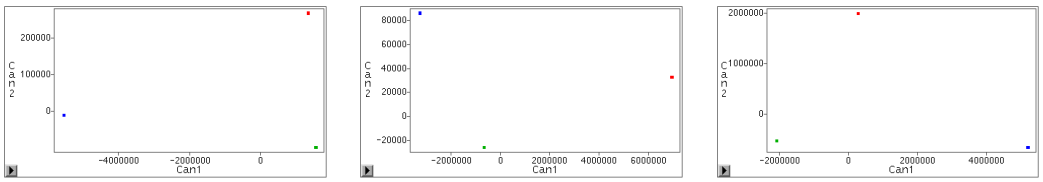


Figure 9.2: Plot of the first two canonical variables. The five sand types are marked with: 1: red, 2: yellow, 3: green, 4: blue, and 5: black.

curves, however, models will be constructed both including all grain curves and separately for each grain curve within the sand types.



(a) Type 1 (b) Type 3 (c) Type 5

Figure 9.3: Plot of the first two canonical variables based on discrimination between grain curves. The three grain curves are marked with: Fine: red, Medium: green, and Large: blue.

9.3 Singular values

The singular values can be used as an indication of whether a problem is ill or well posed. In Figure 9.4 and 9.5 the singular values of the feature matrices for each of the five sand types are plotted. For *features 1* the singular values decay gradually and there is only a small gap around singular value number 200. These problems can be assumed to be ill posed for a feature number less than 200. It is therefore expected to be difficult to select an exact number of features less than 200 to include

in the solutions. For *features 2* the singular values reveal a numerical rank equal to the number of observations in the data set. It is therefore expected that if a small amount of variables is to be inclueded, then the number of variables should equal the number of observations. However, it might be desirable to include less variables, but a part from this gap the singular values decay gradually.

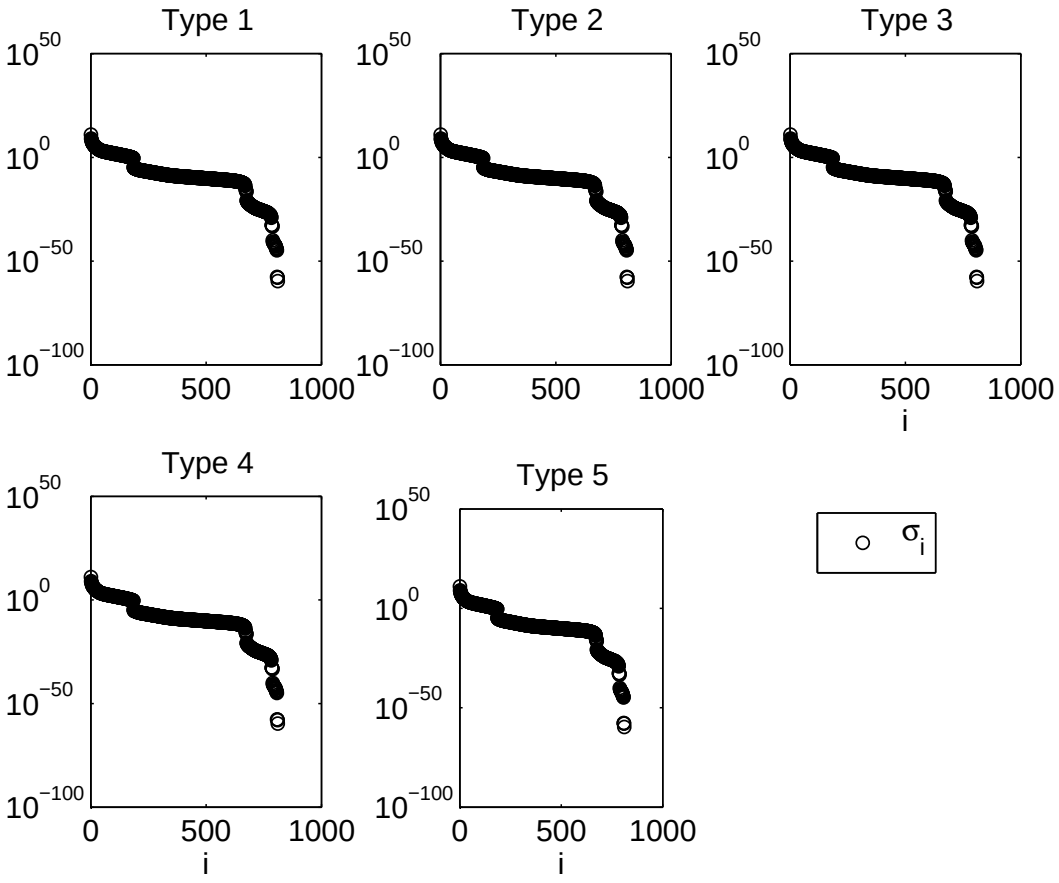


Figure 9.4: Singular values of *features 1* for the five sand types.

9.4 Models for each sand type

This section validates models selected by the methods: Forward Selection combined with OLS, PCA combined with Forward Selection and OLS, Ridge regression, Lasso regression, LARS-EN, and sparse PCs.

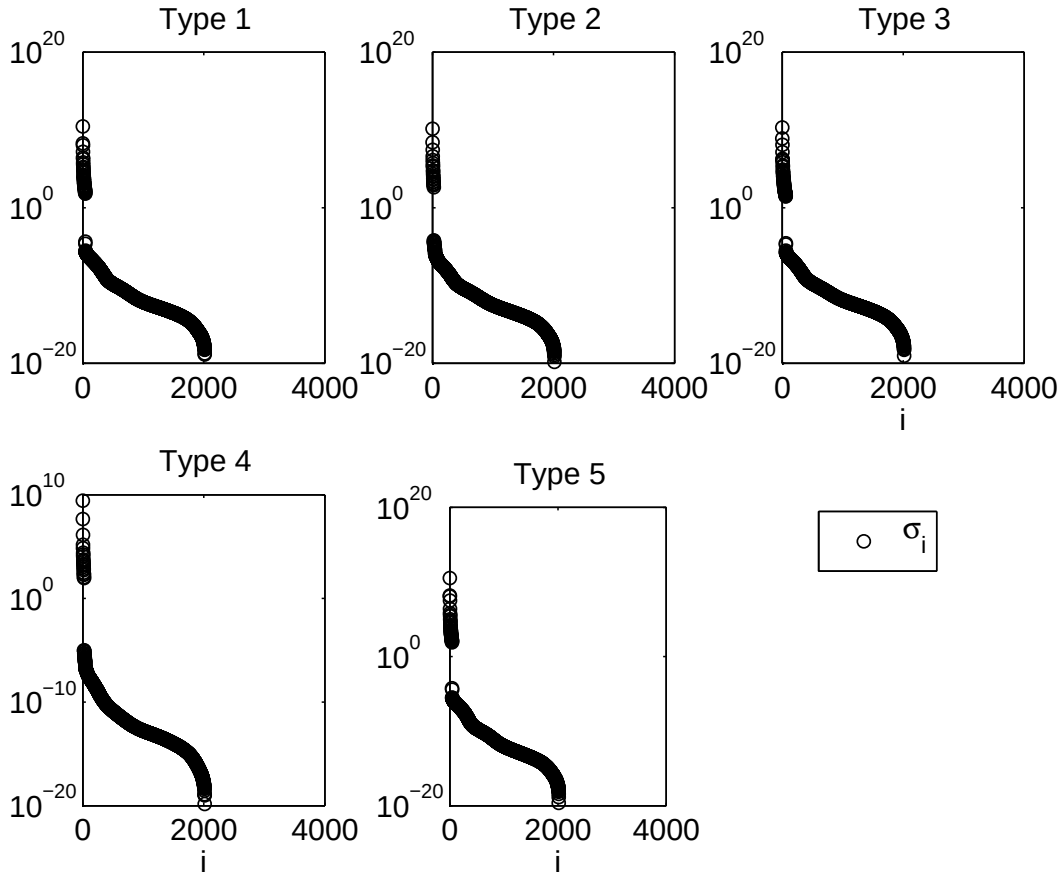


Figure 9.5: Singular values of *features 2* for the five sand types.

9.4.1 Forward Selection

Forward Selection with a significance level of 5% is performed on the original variables of both *features 1* and *features 2* and combined with OLS estimation. Standard deviations for test and training data are plotted in Figure 9.6. Using *features 2* gives much smaller errors of both training and test data, than using *features 1*.

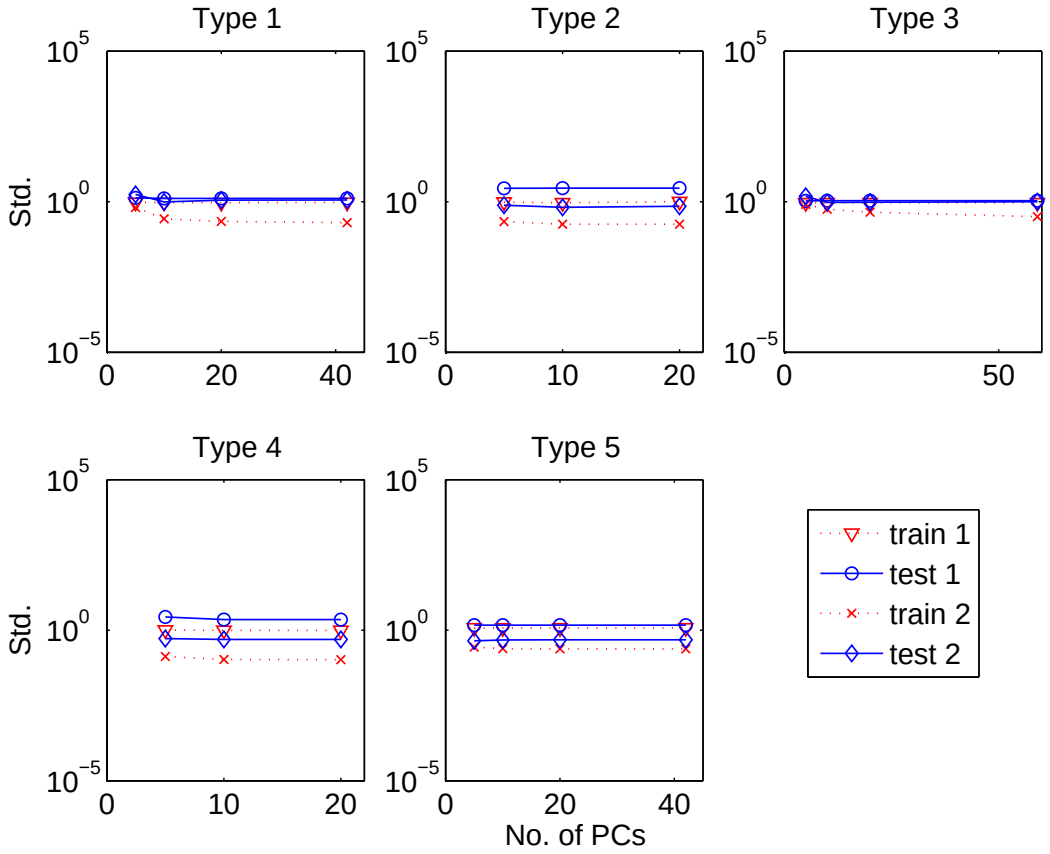


Figure 9.6: Standard deviation of OLS with Forward Selection based on leave-one-out CV for the five sand types.

The lowest standard deviations obtained on the training data of *features 2* are listed in Table 9.1. Training data is over fitted even though the number of variables included is smaller than the number of observations. For some of the sand types this method yields low standard deviations of the prediction eror despite the over fitting. However, this method is computationally very slow.

Type	Std. Train	Std. Test	No. Vars
1	0.07	1.0	10
2	0.03	0.4	10
3	0.3	0.9	10
4	0.01	0.2	20
5	0.07	0.2	5

Table 9.1: The minimum standard deviations for training and test sets of OLS with Forward Selection on features 2. The number of variables selected is also listed.

9.4.2 Principal Componenet Analysis

Forward Selection with $\alpha = 5\%$ is performed on the first 400 PCs and then the OLS estimates based on the selected variables are used for leave-one-out cross validation. Standard deviations for test and training data are plotted in Figure 9.7.

For the data set without scale space features, *features 1*, the training data fits sand type 2 and 4 better than the other sand types. These sand types only consist of samples belonging to one grain curve. For the data set with scale space features, *features 2*, some of the scale space features are selected for sand type 1, 3 and 5, but not for sand type 2 and 4. In the following only *features 2* are considered.

Combining PCA and OLS clearly over fits data, and the best results are obtained when only 5 PCs are included. The standard deviations of the test data are typically in the range from 1 to 2, except for type 4 where it is around 0.2. The results for *features 2* are summed up in Table 9.2.

Type	Std. Train	Std. Test	No. of PCs
1	0.4	1.3	5
2	0.03	1.3	5
3	0.5	1.0	10
4	10^{-16}	0.2	20
5	0.4	1.7	5

Table 9.2: The minimum standard deviations for training and test sets of PCA combined with OLS of features 2. The number of PCs included in the analysis is also listed.

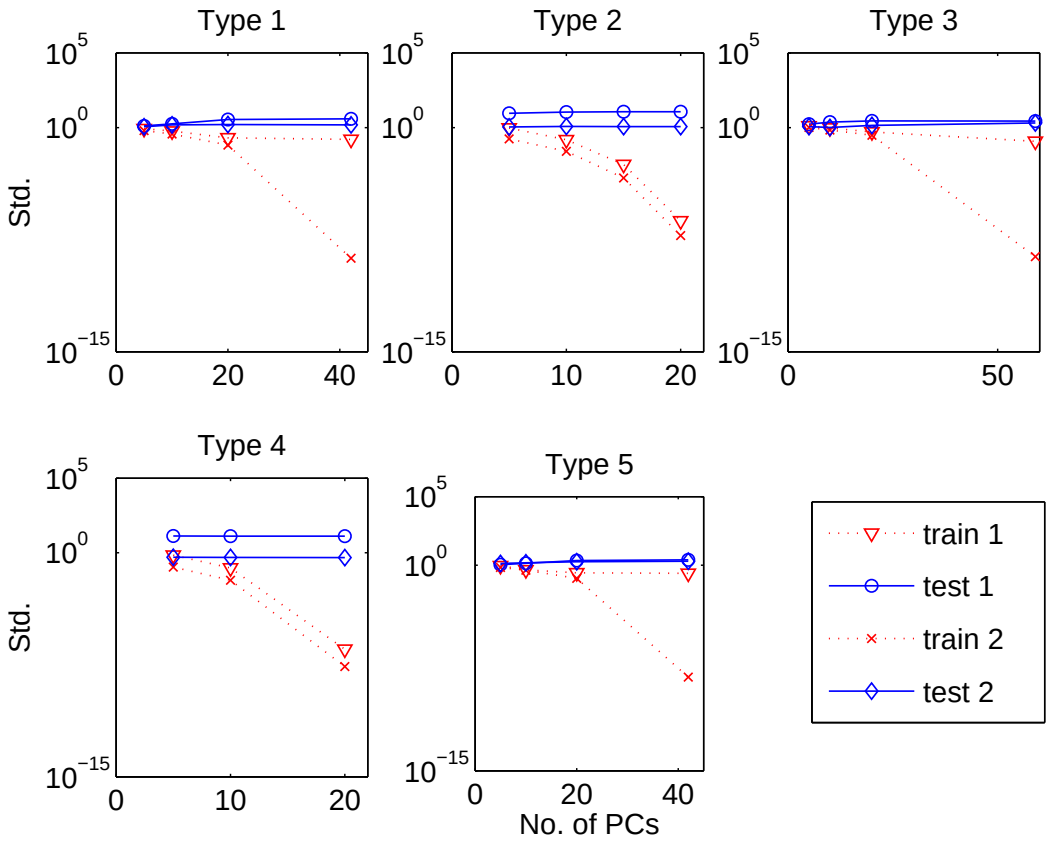


Figure 9.7: The minimum standard deviations of OLS with PCs based on leave-one-out CV for the five sand types.

9.4.3 Ridge regrssion

The Ridge regression does not reduce the dimensionality, the number of active variables is $p = 2016$ in all cases. In Figure 9.8 the MSE as function of λ is illustrated for sand type 1.

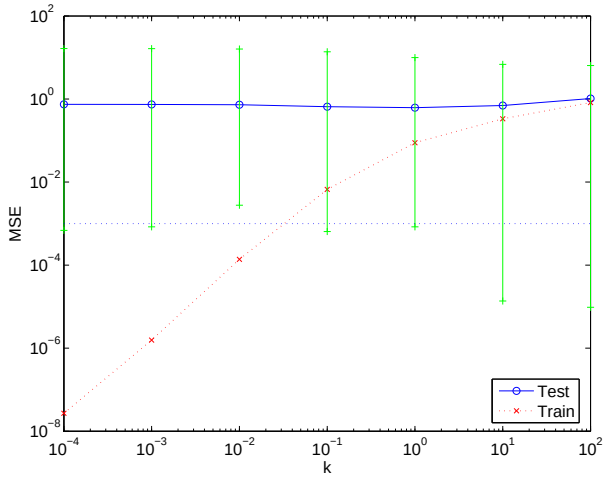


Figure 9.8: MSE as a function of λ for Ridge regression on sand type 1. The CV minimum and maximum of the MSE are illustrated with green for each value of λ.

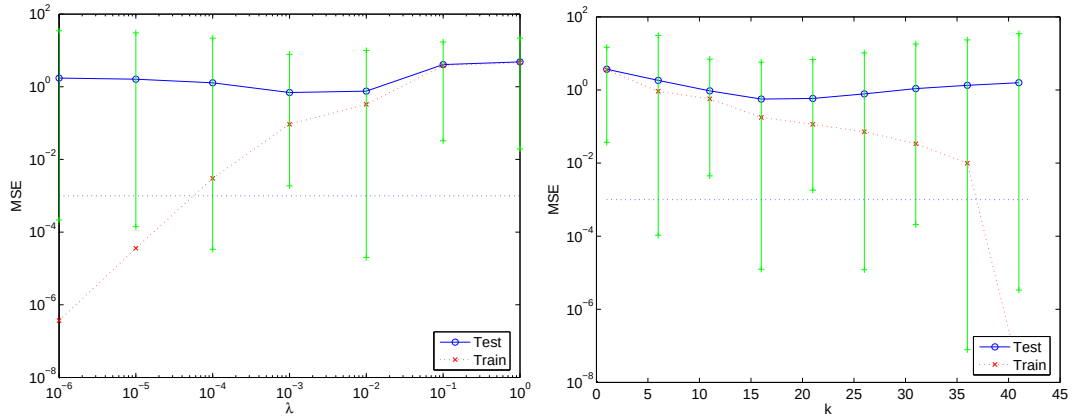
The results obtained with Ridge shrinkage are summed up in Table 9.3. The highest standard deviations are lower than for the combinations of Forwrd Selection or PCA with OLS.

Type	Std. Train	Std. Test	λ
1	0.3	0.8	10^0
2	0.2	0.4	10^1
3	0.2	0.7	10^{-1}
4	0.04	0.3	10^0
5	0.3	0.6	10^0

Table 9.3: The minimum standard deviations for training and test data in Ridge regression for the five sand types. λ is the regularization parameter chosen.

9.4.4 Lasso

The original Lasso algorithm and the Lasso modification in LARS-EN are compared for sand type 1. Figure 9.9 shows that the two algorithms give similar results but depend on the two different regularization parameters λ in Lasso and the number of iterations used in LARS-EN. The MSE of LARS-EN is illustrated as a function of the number of active parameters instead of the number of iterations. The minima are found at $\lambda \simeq 10^{-3}$ and at $k \simeq 16$ (number of active parameters in LARS-EN). When $\lambda = 10^{-3}$ there are around 20 active parameters.



(a) Lasso original

(b) Lasso in LARS-EN

Figure 9.9: MSE for Lasso on sand type 1. (a): The original Lasso with λ as regularizing parameter. (b): The Lasso modification in LARS-EN with the number of active variables, k, as regularizing parameter. The CV minimum and maximum for each value of k are illustrated with green.

Type	Std. Train	Std. Test	k
1	0.4	0.8	16
2	0.2	0.5	11
3	0.4	0.7	26
4	0.2	0.3	11
5	0.3	0.4	11

Table 9.4: The minimum standard deviations for training and test data in Lasso for the five sand types. k is the active number of variables chosen as regularization parameter.

The results obtained with Lasso for the five sand types are summed up in Table 9.4. The Lasso shrinkage improves the standard deviations compared to the traditional methods, and the standard deviations are comparable to those obtained with Ridge regression. Furthermore, Lasso reduces the dimensions and the over fitting is thereby reduced compared to Ridge regression. The dimension reduction is important in an inline production.

9.4.5 LARS-EN

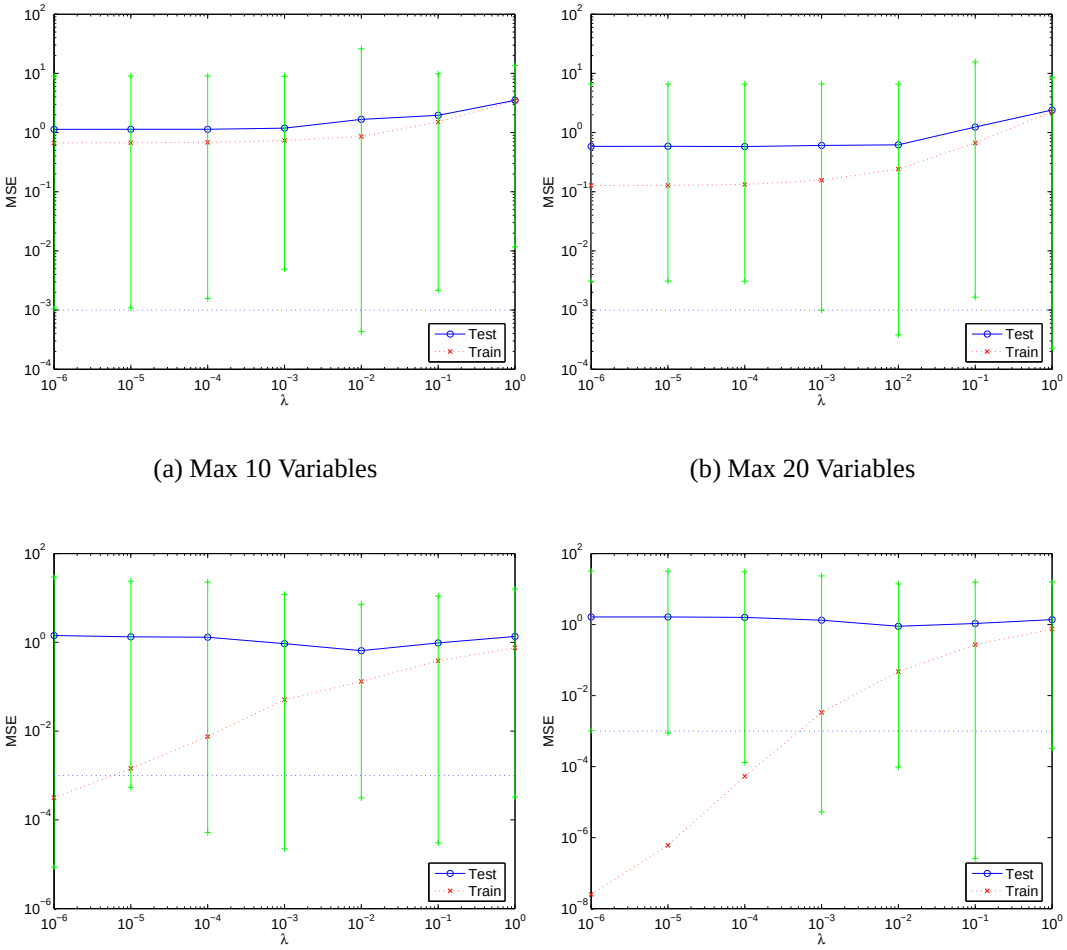
In this section LARS-EN model selection for the five sand types is examined. The two regularization parameters are chosen by means of leave-one-out CV.

In Figure 9.10 the MSE in LARS-EN as function of λ on sand type 1 for four different values of early stopping is illustrated. The early stopping, or maximal number of variables included in the model, is not necessarily the actual number included. When λ is large the number of active variables decreases. The minimum MSE of the training determines the number of iterations or equivalently the number of variables included in the model.

The standard deviation of the test set decreases as the number of variables increases consecutively with an increasing value of λ providing the necessary regularization. The minimum standard deviations of the test set are found through CV on RSS(λ, ite) and summed up in Table 9.5.

Type	Std. Train	Std. Test	λ	ite	Var
1	0.4	0.8	10^{-4}	69	20
2	0.3	0.4	10^{2}	119	118
3	0.2	0.7	10^{-2}	533	400
4	0.3	0.3	10^{2}	202	201
5	0.3	0.4	10^{-3}	19	10

Table 9.5: The minimum standard deviations for training and test in LARS-EN for the five sand types. λ and ite are the regularization parameters chosen. Var is the number of active parameters in the model selected with the given parameters for the entire data set.



(a) Max 10 Variables

(b) Max 20 Variables

(c) Max 40 Variables

(d) Max 100 Variables

Figure 9.10: MSE based on leave-one-out CV for sand type 1 using different numbers of iterations. The minimum mean standard deviations in the three cases are: 1.3, 0.8, 0.9 and 1.0, respectively, corresponding to $\lambda = 10^{-6}$, $\lambda = 10^{-2}$, $\lambda = 10^{-2}$ and $\lambda = 10^{-2}$. Note, that the range for the MSE differs in the three plots, therefore MSE of 10^{-3} is marked with a dotted line, and the minimum of the maximal MSE of the training is marked with a broken line.

9.4.6 Principal components

Performing LARS-EN model selection on the first 400 PCs instead of the original variables does not provide better results with respect to standard deviation. The results are summed up in Table 9.6. Using the PCs the training data is over fitted more than when the original variables were used.

Type	Std. Train	Std. Test	λ	ite	Var
1	$2 \cdot 10^{-2}$	0.7	10^{-4}	59	40
2	$3 \cdot 10^{-6}$	0.3	10^{-6}	97	72
3	$3 \cdot 10^{-2}$	0.9	10^{2}	11	10
4	$2 \cdot 10^{-4}$	0.3	10^{-6}	65	20
5	0.7	0.9	10^{0}	12	11

Table 9.6: The minimum standard deviations for training and test in LARS-EN on the PCs of the five sand types. λ and ite are the regularization parameters chosen. Var is the number of active parameters in the model selected with the given parameters for the entire data set.

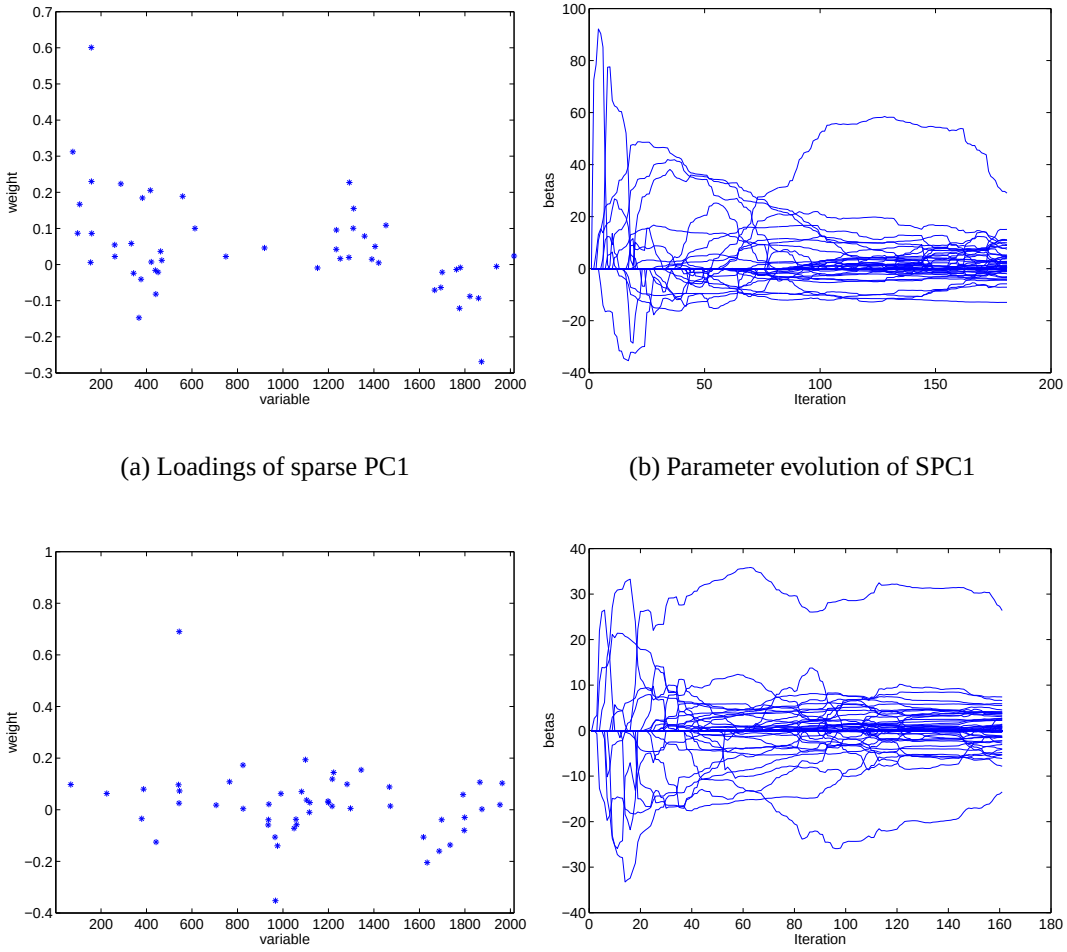
9.4.7 Sparse principal components

The Sparse PCs do not explain a greater variance than the PCs and the regression is therefore not assumed to get better, but the over fitting might. Furthermore, fewer variables are included in the anlyses which is an advantage in an inline production.

The loadings of the sparse PCs are chosen by LARS-EN with $\lambda = 10^{-6}$ and a maximum of fifty active variables. In Figure 9.11 the loadings of the two first sparse principal components are illustrated together wiht their parameter evolutions in LARS-EN. Note, how the parameters become active as the iterations progress. There are 50 active variables out of 2016, but from the loadings it is seen that only a couple of the variables are weighted more than 0.5.

Table 9.7 lists the variance and cumulated variance of the first ten PCs, as well as the cumulated variance, and cumulated adjusted variance of the corresponding sparse PCs. The amount of variance explained by the sparse PCs is very low. However, they are still adequate in model building, as we will see in the following.

Performing model selection with LARS-EN on the first twenty sparse PCs yield approximately the same standard deviations of the training data as on the PCs, but the



(a) Loadings of sparse PC1

(b) Parameter evolution of SPC1

(c) Loadings of sparse PC2

(d) Parameter evolution of SPC2

Figure 9.11: The loadings and parameter evolutions of the two first sparse PCs. There are 50 active variables in each sparse PC. The two features weighted more than 0.5 are for SPC1: the 157th feature, the 30th percentile in the 9th spectral band, and for SPC2: the 543th feature, the 1st percentile of the multiplication between the 1st and 6th spectral bands.

Variance(%)/SPC	1	2	3	4	5	6	7	8	9	10
PC	53.7	20.2	11.2	7.5	1.9	1.2	1.1	0.6	0.6	0.4
Cum PC	53.7	73.9	85.1	92.6	94.5	95.7	96.7	97.4	97.9	98.2
Cum SPC	0.4	0.8	1.0	1.2	1.3	1.3	1.4	1.4	1.4	1.5
Cum Adj SPC	0.4	0.8	1.0	1.2	1.3	1.3	1.4	1.4	1.4	1.5

Table 9.7: The variance and cumulated variance of the PCs, the cumulated variance of the sparse PCs, and the cumulated adjusted variance of the sparse PCs for the first ten PCs of sand type 1.

standard deviation of the test data is now comparable to that of the training, cf. Table 9.8. Compared to OLS regression on a similar low number of PCs the standard deviation is decreased. Hence, the over fitting is decreased. Furthermore, the number of variables included in the calculations is decreased as each sparse PC only contains loadings for fifty of the original variables. The decreased number of variables is an advantage in an inline production.

Type	Std. Train	Std. Test	λ	Var	Sparseness
1	0.6	0.7	10^{-2}	10	50
2	0.06	0.4	10^{-4}	15	50
3	0.6	0.8	10^{-2}	15	50
4	0.4	0.5	10^{-4}	5	50
5	0.6	0.8	10^{-5}	8	50

Table 9.8: The minimum standard deviations for training and test in LARS-EN on the sparse PCs of the five sand types. λ is the regularization parameters chosen, ite is slightly larger than the number of active variables, Var. Sparseness is the number of active variables in the sparse PCs.

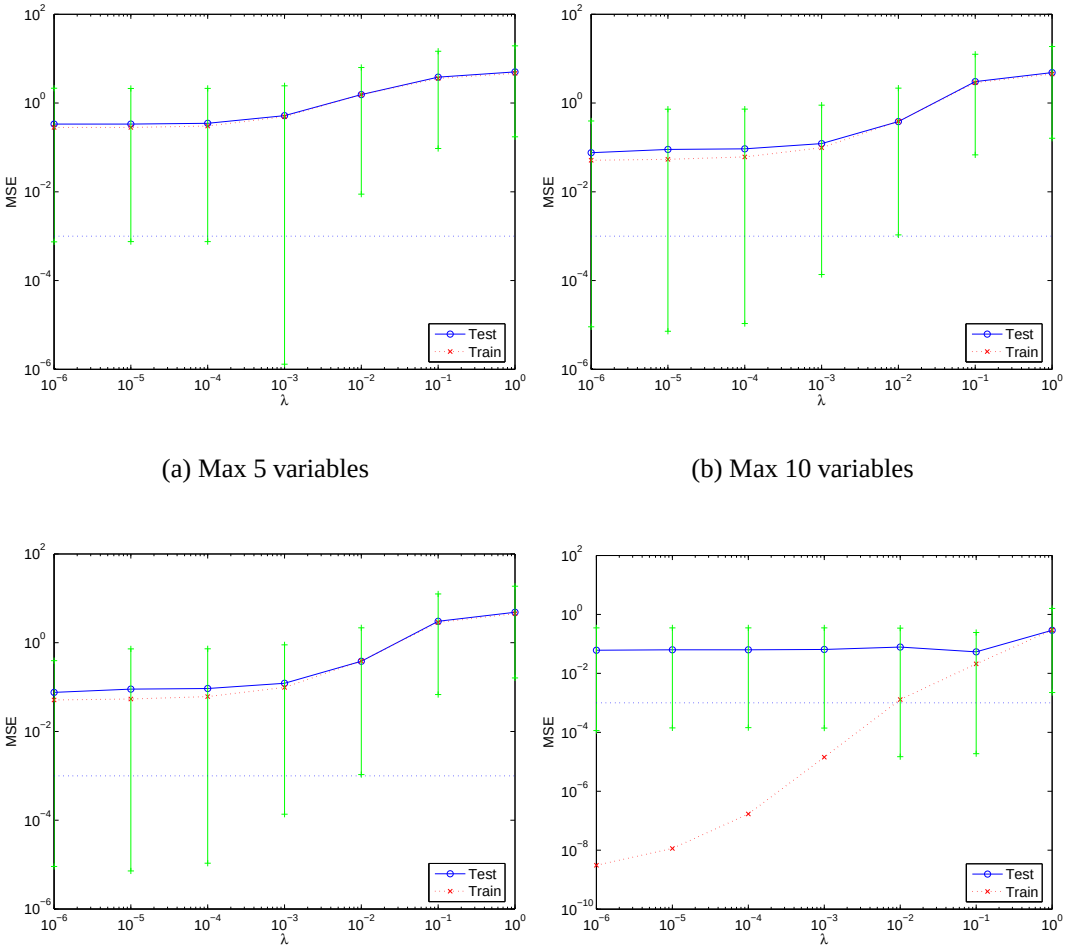
9.5 Models for each sand type and grain curve

In this section only LARS-EN model selection is considered as it is computationally much faster and in the previous analyses it has provided at least as good results as the other methods.

9.5.1 LARS-EN

LARS-EN is used to select models for sand type 1, 2 and 3 divided into the three grain curves fine, medium and large. The selection of λ and the number of iterations/active variables is illustrated in Figure 9.12 for sand type 1 and the fine grain curve. From maximally ten to forty active variables the difference in MSE is small, and the lowest number of variables might be to prefer.

The minimum standard deviation is found using CV on RSS(λ, ite). The results are summed up in Table 9.9.



(a) Max 5 variables

(b) Max 10 variables

(c) Max 40 variables

(d) Max 100 variables

Figure 9.12: MSE based on leave-one-out CV for sand type 1 and medium grain curve. There are 24 observations in the data set.

The selected models are evaluated and the residuals as well as the measured observations are plotted versus the estimated values. Figure 9.13 and 9.14 illustrate examples for two of the data sets. In the first figure a slight underestimation is observed, caused by the early stopping where only 40 variables are activated, but a part from this the data behave neatly and the standard deviation is low. In the second figure a slight overestimation is observed, caused by the coefficient shrinkage of the 74 active variables with a large value of λ. However, the trend is not big since early stopping is used to control the variance and undo some of the overestimation. In general, the residual plots have small trends of either under- or overestimation, as illustrated in the examples. However, the trends are acceptable.

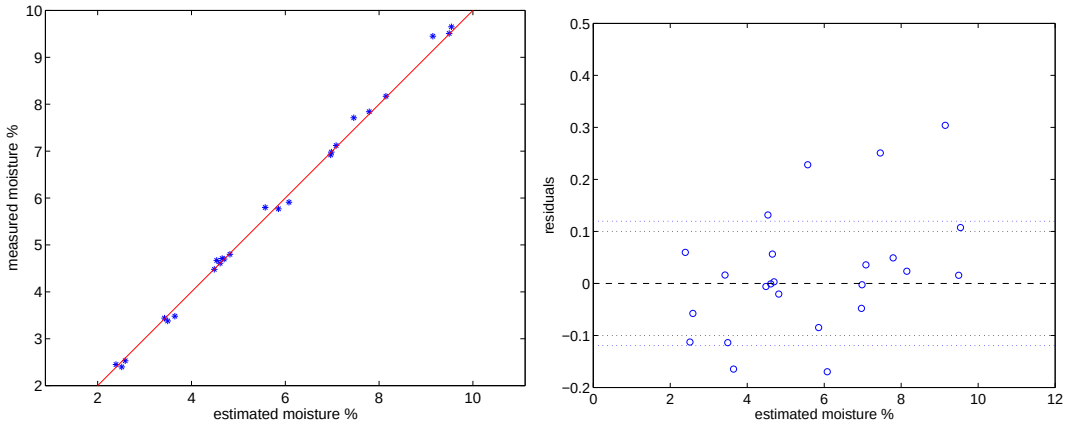
There also seems to be an outlier in the dataset. Three of the observations are samples from one bucket with the intended moisture level of 7.5%, but are all measured to approximately 6%. The one with the highest measure, however, has the smallest estimate. Furthermore, the variation is larger for the samples with higher moisture content than for those with lower, even after the logarithmic transformation of the moisture content measures.

Finally, the sample variation might be reduced through collecting more samples on the fine and large grain curves. There should be enough samples on the medium grain curve, it is thus also for the medium grain curve the lowest standard deviations are observed.

Type	Grain Curve	Std. Train	Std. Test	λ	ite	Var
1	F	0.2	0.5	10^{0}	101	100
3	F	0.3	0.8	10^{1}	75	74
5	F	0.2	0.2	10^{-2}	53	40
1	M	0.1	0.2	10^{-2}	57	40
3	M	0.3	0.4	10^{-3}	59	20
5	M	0.3	0.4	10^{0}	116	115
1	L	0.2	0.4	10^{3}	268	267
3	L	0.1	0.4	10^{0}	42	41
5	L	0.2	0.4	10^{-3}	11	10

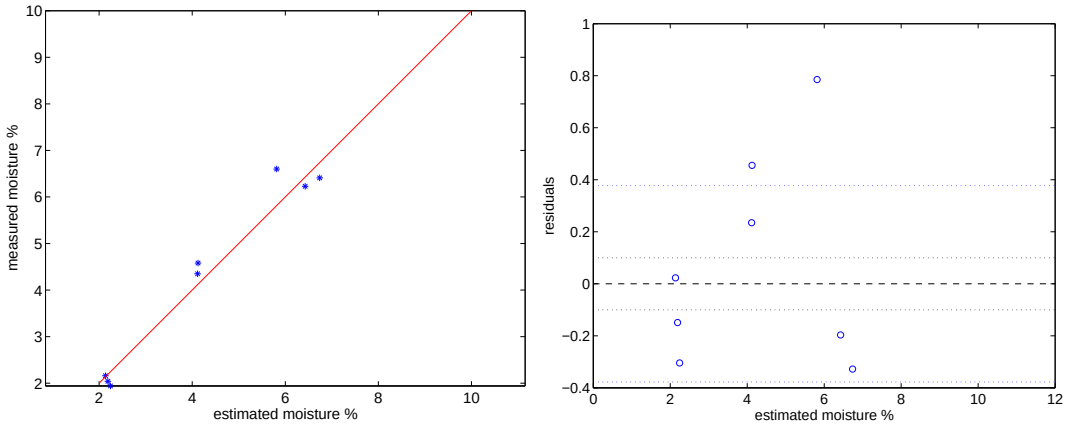
Table 9.9: The minimum standard deviations for training and test sets in LARS-EN for each grain curve on the three sand types 1, 3, and 5. λ and ite are the regularization parameters chosen. Var is the number of active variables in the model selected with the given parameters for the entire data set.

Recall, that the prediction error of leave-one-out CV often has a large variance even though it is unbiased. Therefore, 6- and 7-fold CV is tried on the medium grain curves as they have sufficient observations. The results are summed up in Table 9.9



(a) Features 1 (b) Features 2

Figure 9.13: Measured moisture content and residuals as functions of the estimated moisture content on sand type 1, medium grain curve, and the parameters listed in Table 9.9.



(a) Features 1 (b) Features 2

Figure 9.14: Measured moisture content and residuals as functions of the estimated moisture content on sand type 3, fine grain curve, and the parameters listed in Table 9.9.

Type	Grain Curve	Std. Train	Std. Test	λ	ite	Var
1	M	0.1	0.2	10^{-2}	57	40
2	M	0.3	0.4	10^{0}	132	131
3	M	0.3	0.5	10^{-3}	59	20
4	M	0.3	0.3	10^{0}	203	202
5	M	0.3	0.4	10^{0}	116	115

Table 9.10: The minimum standard deviations for training and test in LARS-EN for each grain curve in the three sand types 1, 3, and 5. λ and ite are the regularization parameters chosen with 6- or 7-fold CV. Var is the number of active variables in the model selected with the given parameters for the entire data set.

9.6 Selected features

The scale space features were particularly useful when more than one grain curve was included in the model. The features most often selected were features from differences between spectra and pair wise relations between spectra. The spectra included in the model varies from sand type to sand type. All spectra are included, but more often features with information from the two NIR bands are selected.

9.7 Summing up and discussion

The scale space features were particularly useful when more than one grain curve was included in the model. Both the singular values and the results obtained with OLS shows that *features 2* are better than *features 1*. Hence, the additional features in this data set provide additional information to the other features. Furthermore, information from the NIR spectra of 875 and 940nm is always included in the selected models. It is known that subtracting the two NIR bands of 870 and 970nm can reflect information of water content in materials[1]. The spectral bands are not quite the same, but the results indicate that the NIR spectra are important in the estimation of the moisture content.

Ridge regression, Lasso and LARS-EN yield lower standard deviations than Forward Selection and PCA combined with OLS. Hence, the coefficient shrinkage is an advantage. Furthermore, Lasso and LARS-EN select a subset of variables to include in the model. If the estimation is to be implemented in the construction line, the time is an issue, and evaluating less variables is therefore a plus. Finally, LARS-EN gives more options and additionally provides the Lasso solutions, and it is computationally much

[1] [Carstensen 2006]

faster than both Lasso and Ridge. Therefore the LARS-EN model selection is to prefer. Finally, sparse principal components have been a good alternative to principal components, in particular if the sparseness is of importance. The sparse principal components use fewer variables and therefore tend to over fit less than principal components.

The results have been best when models have been selected for each sand type and grain curve separately. Leave-one-out and 6- or 7-fold CV gives comparable results for all sand types and medium grain curve. Though, the over fitting is slightly smaller with 6- or 7-fold CV, recall, that the prediction error of leave-one-out CV often has large variance even though it is unbiased.

The standard deviations of the prediction error is around 0.4 for most of the models selected with LARS-EN, corresponding to a standard deviation of 0.1-0.3 for the training data.

Recall, that the samples collected from the same buckets of sand do not have the same moisture content measures. The standard deviations within these repetitions are 0.01-0.35. The means of these standard deviations are 0.1, 0.06, 0.2, 0.03, and 0.1 for the five sand types, respectively. The variations are larger for sand type 1, 3, and 5 which are also the sand types yielding the largest variations in the prediction error.

Comparing the variations of the sampling repetitions with the prediction errors, around one third of the prediction error is likely to be a consequence of the repetition sample variation.

Chapter 10
Conclusion

Conclusions from various aspects of the project are made. Therefore this chapter has been divided into three parts. Conclusions for each set of data: The identification of *Penicillium* fungi, and estimation of moisture content in sand samples. Additionally, conclusions from comparisons of the traditional multivariate, statistical methods with the newer model selection methods are likewise treated separately.

Identification of *Penicillium* fungi

With a 0% error rate for both leave-one-out and 2-fold cross-validation, the results have been very promising. These results have been obtained using only the YES medium. Furthermore, only two to three variables are needed to separate the species. The three variables that have discriminated best between the species include information from five of the spectral bands: Ultra blue, cyan, amber, red, and NIR(870nm).
Summing up, the three species P. *melanoconidium*, P. *polonicum*, and P. *venetum* can be identified objectively from just one medium.

The good classification results are in accordance with the results of Hotelling's T^2-tests. The tests have shown that there, statistically, is a significant difference between the means of the three species on the three media.

Statistically, it can be assumed that three media do not include additional information to the discrimination compared to using just two media. Furthermore, it can be assumed that the YES and OAT media do not provide additional information to one another in the discrimination. In addition to that Mahalanobi's distances have been largest on the YES medium.
Summing up, the best choice of media is YES and CYA.
However, in practice, the YES medium has shown sufficient to discriminate the species

completely.
It is a big advantage that one medium is sufficient to identify the species since it is both expensive and time consuming to inoculate the isolates on various media.

Mahalanobi's distances between P. *polonicum* and P. *venetum* have been smaller than between P. *polonicum* and P. *melanoconidium* using the features considered. This observation indicates that the considered features reflect the visual appearance and not the genetic relation.
Consequently, the best discrimination has been based on the appearance of the fungal colonies.

Finally, using images of all eighteen spectral bands has provided the best classification results. However, using linear combinations of the ten visual bands as representations of R, G, and B only has performed slightly worse in the sense that more variables have been included in the classification model. If species that are more difficult to identify are considered, it is therefore recommendable to gather all spectra.

Estimation of moisture content in sand

The standard deviations of the prediction errors obtained with both leave-one-out and 6- or 7-fold cross-validation have been around 0.4, corresponding to standard deviations of 0.1-0.3 for the training data. LARS-EN has shown useful to computationally fast select only a subset of variables to include in the model. These qualities are of importance if the estimation is to be implemented in a construction line.

Due to the fact that the images only capture the surface of a sand sample, and that the moisture content is particularly delicate exactly at the surface, due to vaporization, a certain sample variation must be expected. Furthermore, the measured moisture content is a measure of the moisture content in the entire sample. Hence, the relation between the small amount of sand captured by the image and the entire sand sample measured could cause some variation. Finally, the sand samples collected in the petri dishes from the same buckets of sand do not have the same moisture content measures. The sample variations of the repetitions correspond to approximately one third of the prediction errors. Because of the many sources giving rise to sample variations it is unlikely to obtain much lower standard deviations of the prediction error.

The scale space features have been useful in models with more than one grain curve and features with information from the NIR spectra have been included in the best models for all the sand types.

Comparison of methods

The Histogram Pursuit algorithm has only failed twice in segmenting the fungal colonies,

where as the identification of circular colonies has failed in half of the cases on the YES medium. Furthermore, fewer variables have been required to classify the species correctly with the features from the HP segmentation. The HP algorithm is therefore preferable to segment the fungal colonies.

LARS-EN with dummy variables has shown more sensitive to which observations are in the test respective training sets for few-fold cross-validation, e.g. 2-fold cross-validation, compared to Discriminant Analysis. Furthermore, the Discriminant Analysis discriminates between all species at the same time, and not as LARS-EN between one class and the remaining classes. Hence, the Discriminant Analysis often requires fewer variables as each variable is used to discriminate between all species. The disadvantage of the Discriminant Analysis is that it is computationally much slower than LARS-EN.

The shrinkage methods Ridge regression and Lasso have provided good results compared to Forward Selection and PCA combined with OLS for the sand data. Lasso is preferable to Ridge as the number of variables is reduced considerably. LARS-EN have provided slightly better results than Ridge and Lasso, and as both the Ridge and Lasso solutions can be obtained computationally faster via the LARS-EN, LARS-EN is to prefer.

Using LARS-EN on the PCs has shown to give larger standard deviations as training data has been over fitted. However, sparse principal components have turned out to be a good alternative to principal components, especially if the sparseness is of importance. The sparse principal components use fewer variables and therefore tend to over fit less than the principal components.

Chapter 11
Future Work

This chapter gives some ideas on future work related to this project.

Identification of fungal species

The experiment could be conducted with:

- Other isolates.
- Other genera.

This would confirm the results obtained and produce an objective reference classification model for future use.

Furthermore, if necessary, information from the images of the back side of the fungal colonies could be included in the models.

Estimation of moisture content in sand

- Use of a multi-spectral camera that takes consecutive images of a larger surface covered with a thin layer of a sand sample. This might reduce some of the sample variation. Furthermore, the relation between the amount of sand imaged and the amount of sand used for the reference measure of the moisture content would be better.
- Study of vaporization from the sand samples through imaging of the same sand sample over time. To examine the influence of vaporization from the surface of

the sand sample since the surface is the part of the sand sample that is captured in the image.

Methods

- Modify LARS-EN with dummy variables to regress more than one dependent variable at a time. Hence, each selected variable is used for all of the dummy variables. However, it is not straight forward how the variables should be selected; if it is the one correlated most with all of the dependent variables or the one correlated most with one of the dependent variables. Furthermore, this also influences the equiangular direction.
- Use of maximum likelihood estimation of the effects in the fungi experiment instead of sums of squares of deviations. Meyer[1] derived the "restricted maximum likelihood" (REML) for a multivariate mixed model with two effects. The REML overcomes the bias of the ML caused from ignoring the loss in degrees of freedom due to fitting of fixed effects. Furthermore, the method transforms to canonical variables which has the advantage of giving weight to features explaining a maximum of variance of an effect and weighting out the features that add little extra information given the other features.

[1] [Meyer 1985]

Bibliography

Carstensen, J. M. (2006). Internal communication.

Christensen, M., Miller, S. L. & Tuthill, D. (1994), 'Color standards - a review and evaluation in relation to penicillium taxonomy', *Mycol. Res.* **98**, 635–644.

Chung, D. H. & Sapiro, G. (2000), 'Segmenting skin lesions with partial-differential-equations-based image processing algorithms', *IEEE Transactions on Medical Imaging* **19**(7), 763–767.

Conradsen, K. (2002*a*), *En introduktion til statistik, Bind 2*, IMM/Informatik og Matematisk Modellering.

Conradsen, L. (2002*b*), Statistiske analyser af to-dimensionale elektroforese-geler, Master's thesis, Technical University of Denmark.

Dorge, T., Carstensen, J. M. & Frisvad, J. C. (2000), 'Direct identification of pure penicillium species using image analysis', *Journal of Microbiological Methods* **41**, 121–133.

Du, C. J. & Sun, D. W. (2005), 'Comparison of three methods for classification of pizza topping using different colour space transformations', *Journal of Food Engineering* **68**(3), 277–287.

Duda, R. O., Hart, P. E. & Stork, D. G. (2001), *Pattern Classification*, John Wiley & Sons.

Efron, B., Hastie, T., Johnstore, I. & Tibshirani, R. (2003), Least angle regression, Technical report, Statistics Department, Stanford University.

Engstrom, N., Hansson, F., Hellgren, L., Tomas, J., Nordin, B., Vincent, F. & Wahlberg, A. (1990), 'Computerized wound image analysis', *Pathogenesis of Wound and Biomaterial-Associated Infections* pp. 189–193.

Folm-Hansen, J. (1999), On Chromatic and Geometrical Calibration, Phd thesis, Technical University of Denmark.

Friedman, J. H. (1987), 'Exploratory projection pursuit', *Journal of the American Statistical Association* **82**(397), 294–266.

Friedman, J. & Tukey, J. (1974), 'A projection pursuit algorithm for exploratory data analysis', *IEEE Trans. on Computers* **23**(9), 881–889.

Frisvad, J. C. (2006). Internal communication.

Frisvad, J. C. & Samson, R. A. (2004), 'Polyphasic taxonomy of penicillium subgenus penicillium: A guide to identification of food and air-borne terverticilliate penicillia and their mycotoxins', *Stud. Mycol.* **49**, 1–173.

Frisvad, J. C., Smedsgaard, J., Larsen, T. O. & Samson, R. A. (2004), 'Mycotoxins, drugs and other extrolites produced by species in penicillium subgenus penicillium', *Stud. Mycol.* **49**, 201–242.

Fu, W. J. (1998), 'Penalized regressions: The bridge versus the lasso', *J. Computational and Graphical Statistics* **7**(3), 397–316.

Ganster, H., Pinz, A., Rohrer, R., Wildling, E., Binder, M. & Kittler, H. (2001), 'Automated melanoma recognition', *IEEE Trans. Med.Imaging* **20**(3), 233–239.

Garcia, C. & Tziritas, G. (1999), 'Face detection using quantized skincolor regions merging and wavelet packet analysis', *IEEE Transactions on Multimedia* **1**(3), 264–277.

Gill, P. E., Murray, W. & Wright, M. H. (1981), *Practical Optimization*, Academic Press.

Goldberg, D. E. (1989), *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison Wesley.

Gomez, D. D. (2005), Development of an image based system to objectively score the severity of psoriasis, Phd thesis, Technical University of Denmark.

Gutenev, A., A., Skladnev, V. N. & Varvel, D. (2001), 'Acquisition-time image quality control in digital dermatoscopy of skin lesions', *Computerized Medical Imaging and graphics* **25**, 495–499.

Hance, G., Umbaugh, S., Moss, R. & Stoecker, W. (1996), 'Unsupervised color image segmentation, with application to skin tumor border', *IEEE engineering in medicine and biology* **15**(1), 104–111.

Hansen, M. E. (2003), Indexing and Analysis of Fungal Phenotypes Using Morphology and Spectrometry, Phd thesis, Technical University of Denmark.

Hansen, P. C. (1998), *Rank-Deficient and Discrete Ill-Posed Problems*, SIAM.

Hastie, T. & Tibshirani, R. (1990), *Generalized Additive Models*, Chapmann and Hall.

Hastie, T., Tibshirani, R. & Friedman, J. (2001), *The Elements of Statistical Learning*, Springer.

Hilger, K. B. (2001), Exploratory Analysis of Multivariate Data, Phd thesis, Technical University of Denmark.

Hoerl, A. E. & Kennard, R. W. (1970), 'Ridge regression: Biased estimation for nonorthogonal problems', *Technometrics* **12**, 55–67.

Ihlow, A. & Seiffert, U. (2004), 'Automating microscope colour image analysis using the expectation maximisation algorithm', *proceedings of the 26th DAGM Symposium in Pattern Recognition, Springer-Verlag* pp. 536–54.

Jollife, I. T. (2002), *Principal Component analysis*, Springer Series in Statistics.

Leng, C., Lin, Y. & Wahba, G. (2004), A note on the lasso and related procedures in model selection, Technical Report 1091r, National University of Singapore and University of Wisconsin-Madison.

Luo, G., Chutatape, O., Li, H. & Krishnan, S. (2001), 'Abnormality detection in automated mass screening system of diabetic retinopathy', *Proceedings 14th IEEE Symposium on Computer-Based Medical Systems* pp. 132–137.

Maglogiannis, I. (2004), 'Design and implementation of a calibrated store and forward imaging system for teledermatology', *Journal of Medical Systems* **28**(5), 455–467.

Meyer, K. (1985), 'Maximum likelihood estimation of variance components for a multivariate mixed model with equal desgin matrices', **41**, 153–165.

NIST/SEMATECH (2006), *e-Handbook of Statistical Methods*. http:www.itl.gov/div898/handbook/.

Nunez, F., Diaz, M. C., Rodriguez, M., Aranda, E., Martin, A. & Asensio, M. A. (2000), 'Effects of substrate, water activity, and temperature on growth and verrucosidin production by penicillium polonicum isolated from dry-cured ham', *Journal of Food Protection* **63**(2), 231–236.

Phung, S. L., Bouzerdoum, A. & Chai, D. (2005), 'Skin segmentation using color pixel classification: analysis and comparison', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(1), 148–154.

Pitt, J. I. (1979), *The genus* Penicillium *and its telemorphic states* Eupenicillium *and* Talaromyces, Academic Press.

Raper, K. B. & Thom, C. (1949), *Manual of the Penicillia*, Williams & Wilkins.

Rencher, A. C. (2002), *Methods of Multivariate Analysis*, John Wiley & Sons.

Samson, R. A. & Frisvad, J. C. (1993), 'New taxonomic approaches for identification of food-borne fungi', *Int. Biodegr. Biodet.* **32**, 99–116.

Samson, R. A. & Frisvad, J. C. (2005*a*). http://www.studiesinmycology.org/en/content/37/polyphasic/species/ibn1_copy17/toon.

Samson, R. A. & Frisvad, J. C. (2005*b*). http://www.studiesinmycology.org/en/content/47/polyphasic/species/ibn1_copy17/toon.

Samson, R. A. & Frisvad, J. C. (2005*c*). http://www.studiesinmycology.org/en/content/49/polyphasic/species/ibn1_copy17/toon.

Samson, R. A., Seifert, K. A., Kuijper, A. F. A., Houbraken, J. A. M. P. & Frisvad, J. C. (2004), 'Phylogenetic analysis of penicillium subgenus using partial β-tubulin sequences', *Stud. Mycol.* **49**, 175–200.

Skettrup, M. (2003), Multivariat dataanalyse af 2d-elektroforesegeler, Master's thesis, Technical University of Denmark.

StatSoft, I. (2005). http://www.statsoft.com/textbook/stdiscan.html#discriminant.

Taxt, T., Hjort, N. L. & Eikvil, L. (1991), 'Statistical classification using a linear mixtureof two multinormal probability densities', *Pattern recognition letters* **12**, 731–737.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *J. R. Statist. Soc. B* **58**(No. 1), 267–288.

Turk, M. A. & Pentland, A. P. (1991), 'Face recognition using eigenfaces', *Proceedings CVPR 1991* pp. 586–591.

Vander, Y., Haeghen, Y., Naeyaert, J. M. & Lemahieu, I. (2000), 'An imaging system with calibrated color image acquisition for use in dermatology', *IEEE transactions on medical imaging* **19**(7), 722–730.

Vhrel, M. & Trussell, H. (1999), 'Color device calibration: A mathematical formulation', *IEEE Trans. Image Process* **8**, 1796–1806.

Windfeld, K. (1992), Application of Computer Intensive Data Analysis Methods to The Analysis of Digital Images and Spatial Data, Phd thesis, Technical University of Denmark.

Wyszecki, G. & Stiles, W. (1982), *Color Science: Concepts and Methods, Quantitaive Data and Formulae*, John Wiley & Sons.

Zhang, X. & Chutatape, O. (2004), 'Detection and classification of bright lesions in color fundus images', *2004 International Conference on Image Processing* **1**, 139–142.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *J. R. Statist. Soc. B* **67**(Part 2), 301–320.

Zou, H., Hastie, T. & Tibshirani, R. (2004*a*), On the 'degrees of freedom' of the lasso, Technical report, Stanford University.

Zou, H., Hastie, T. & Tibshirani, R. (2004*b*), Sparse principal component analysis, Technical report, Statistics Department, Stanford University.

Appendix A

Precise Acquisition and Unsupervised Segmentation of Multi-Spectral Images.

The article in this appendix has been submitted to the special issue of Elsevier Computer Vision and Image Understanding on 'Advances in Vision Algorithms and Sytems Beyond the Visible Spectrum'.

Precise acquisition and unsupervised segmentation of multi-spectral images.

David Delgado Gomez Line Harder Clemmensen
Bjarne K. Ersbøll Jens Michael Carstensen[1]

Informatics and Mathematical Modelling, Building 321
Technical University of Denmark, DK-2800 Lyngby, Denmark

Abstract
In this work, an integrated imaging system to obtain accurate and reproducible multi-spectral images and a novel multi-spectral image segmentation algorithm are proposed. The system collects up to 20 different spectral bands within a range that vary from 395nm to 970nm. The system is designed to acquire geometrically and chromatically corrected images in homogeneous and diffuse illumination, so images can be compared over time. The proposed segmentation algorithm combines the information provided by all the spectral bands to segment the different regions of interest. Three experiments are conducted to show the ability of the system to acquire highly precise, reproducible and standardized multi-spectral images and to show its applicabilities in different situations.

Keywords: Image acquisition, multi-spectral image analysis, illumination, exploratory data analysis, image segmentation, pattern recognition.

A.1 Introduction

According to Wyszecky [Wyszecki & Stiles 1982], color is defined as the aspect of visual perception by which an observer may distinguish differences between two structure-free fields of view of the same size and shape. Since the beginning of image analysis, several color models have been developed with the goal of enhancing the contrast of the different structures embedded. These color spaces have made the segmentation of the interesting structures easier in several problems. For instance, two

[1]*Email addresses:* ddg@imm.dtu.dk(David Delgado Gomez), s001376@serv1.imm.dtu.dk(Line Harder Clemmensen), be@imm.dtu.dk(Bjarne K. Ersbøll), (Jens Michael Carstensen)

of these color spaces, the CIE-XYZ and the CIE-Lab [Wyszecki & Stiles 1982] have been successfully applied to the segmentation of dermatological lesions [Ganster, Pinz, Rohrer, Wildling, Binder & Kittler 2001][Hance, Umbaugh, Moss & Stoecker 1996]. These two color spaces are frequently used in Dermatology because of the uniformity of the CIE-Lab color space. This uniformity that helps to understand how different two colors will look to a human observer is directly connected with dermatologist's visual lesion evaluation. These two color spaces are a linear and a non-linear transformation of the RGB color space. The CIE-XYZ is defined by

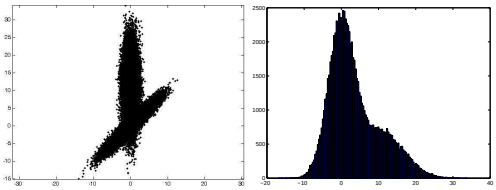
$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{bmatrix} 0.41 & 0.36 & 0.18 \\ 0.21 & 0.71 & 0.07 \\ 0.02 & 0.12 & 0.95 \end{bmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \quad ,$$

and the CIE-Lab by

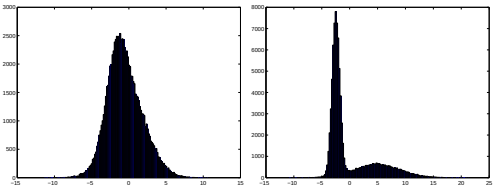
$$\begin{aligned} L &= 116(\frac{Y}{Y_n})^{\frac{1}{3}} - 16 \\ a &= 500[(\frac{X}{X_n})^{\frac{1}{3}} - (\frac{Y}{Y_n})^{\frac{1}{3}}] \\ b &= 200[(\frac{Y}{Y_n})^{\frac{1}{3}} - (\frac{Z}{Z_n})^{\frac{1}{3}}] \quad , \end{aligned}$$

where X_n, Y_n and Z_n are the X, Y, Z coordinates of a reference white patch.
Other color spaces have also been developed aiming at enhancing the interesting structures in other image analysis areas. For example, the YCbCr color space has been widely applied in facial and skin detection [Garcia & Tziritas 1999][Phung, Bouzerdoum & Chai 2005], the HSV in food assessment and fungi detection [Du & Sun 2005][Ihlow & Seiffert 2004], and the CIE-Luv in diabetes and retinopathy detection [Luo, Chutatape, Li & Krishnan 2001][Zhang & Chutatape 2004]. However, the appearance of new multi-spectral equipments that capture more than just the trichromatic bands, have emerged the need of finding new transformations that include the information provided by the new bands.

An approach that has been considered to overcome this problem is principal component analysis (PCA) [Jollife 2002]. This multivariate statistical technique consists



(a) Two dimensional dataset (b) 1st PC



(c) 2nd PC (d) Bimodal projection

Figure A.1: A two dimensional dataset, its two principal components (PC) and a bimodal projection of the dataset.

in an eigenvalue analysis of the covariance matrix for a multidimensional stochastic variable. Given a random n-dimensional variable, the i^{th} principal component is the linear combination, with normed coefficients, of the original variables which is uncorrelated with the $i-1$ first principal components and it has the largest variance. This i^{th} principal component correspond to the eigenvector associated with the i^{th} largest eigenvalues of the covariance matrix. PCA has the property that, frequently, some of the components reveal the wanted structures.

However, although this technique has successfully been applied in some data reduction and classification problems [Turk & Pentland 1991], it is not able to provide a suitable solution in other classification problems. An example of this is illustrated applying PCA to the dataset displayed in Figure A.1 (a). This synthetic dataset was generated according to a mixture of two Gaussian populations with 20000 and 10000 data points, means [0,0] and [0,10] and covariance matrices $\begin{pmatrix} 9 & 9 \\ 9 & 10 \end{pmatrix}$ and $\begin{pmatrix} 1 & 0 \\ 0 & 25 \end{pmatrix}$, respectively. The two principal components obtained are shown in Figure A.1 (b) and (c). Note, that none of the two principal components are able to separate the Gaussian populations. Moreover, it is shown in Figure A.1 (d) that it is possible to find a bimodal one-dimensional projection that separates both populations. Therefore, there exits a need to find an optimal projection from a classification point of view that enhances the different structures in the image.

This need is added to the already existing challenge of collecting precise and reproducible images so images collected at different times can precisely be compared. Different research projects in color calibration [Vhrel & Trussell 1999] and illumination control [Vander, Haeghen, Naeyaert & Lemahieu 2000] have been developed with the goal of achieving these two goals. The consequence of these studies is the appearance of new equipments which aims at obtaining precise images within last years. For instance, in dermatology, Magliogiannis [Maglogiannis 2004], developed a system that aimed at reducing the shadows produced by the human body curvature. However, as it was shown by Gutenev [Gutenev, A., Skladnev & Varvel 2001], there are at least two current problems in the acquisition of the images: specular reflection and misalignments. Lack of precision in the image acquisition has been prevented using suitable methods to objectively evaluate the images.

In this work, two solutions are proposed to deal with the two situations: an imaging system to collect precise and reproducible images and an algorithm to find suitable projections which easily segment interesting areas in the images. In section two, an integrated imaging system to obtain accurate and reproducible multi-spectral images is proposed. The well defined and diffuse illumination of the optically closed scene aims to avoid shadows and specular reflections. Furthermore, the system has been developed to guarantee the reproducibility of the collected images. This allows for comparative studies of time series of images. In order to segment the interesting structure of the images, a novel segmentation algorithm, the histogram pursuit, is presented in section three. This algorithm combines the information provided by all the different spectral bands to enhance the main structures of the image. The performance of both the equipment and the histogram pursuit algorithm to achieve the above commented goal is tested and shown in section four. The obtained results and extensions of the developed work are discussed in section five.

A.2 Collecting multi-spectral images

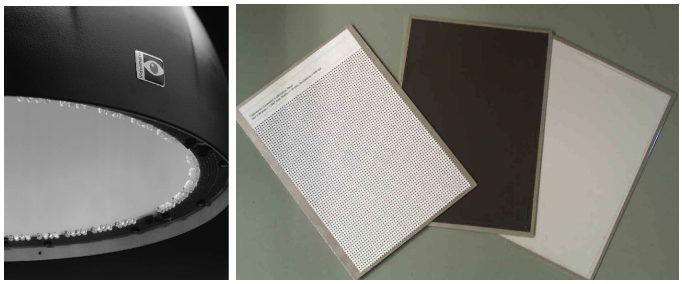
The acquisition of the multi-spectral images was conducted in collaboration with Videometer[2]. The proposed equipment, Videometer Lab, is composed of a camera, light emitting diodes and a integrating sphere. The equipment has been designed to produce completely diffuse light that avoid shadows and specular reflections. The system acquires the multi-spectral images by fast strobe illumination from light emitting diodes (LEDs) at up to 20 different wavelengths.

Figure A.2 left shows the equipment. Figure A.2 right displays a sketch of the set-up. It displays the position of the camera and the diodes inside of the sphere and the

[2]www.videometer.com

(a) Imaging equipment (b) Light set-up

Figure A.2: The camera system.



(a) Diodes (b) Sheets

Figure A.3: Positioning of the diodes in the camera set-up and calibration sheets.

place where the object is located. Figure A.3 displays the position of the diodes inside the equipment. The camera resolution is a 1380×1035. In order to increase the accuracy and reproducibility of the images a radiometric and a geometric calibration are conducted [Folm-Hansen 1999]. The radiometric calibration aims at eliminating problems with uneven intensities and vignetting, and to standardize the measurement scale. With this goal in mind two sheets of the natural color system (NCS) from the Scandinavian Color Institute were selected as calibration targets (NCS 1500 and NCS 8000). The equipment collects an image of each sheet. Then a non-linear calibration function is estimated and applied to each image pixel during the further image acquisition. The geometric calibration is conducted to make sure that aberrations, such as distortion, decentering and thin prism aberrations, do not affect the accuracy of the images. An image of a white sheet with black spots is grabbed with the camera for each wavelength. This calibration target is shown in figure A.3 right, together with the radiometric sheets. The collected multi-spectral images are thresholded and the center of gravity of each spot is calculated. A third order polynomial is applied to warp the centers of gravity to a given target. This is done for each band in the multi-spectral image in order to assure co-site registration.

A.3 Segmenting the lesion: Histogram pursuit

The core of the proposed segmentation algorithm is found in Friedman's projection pursuit algorithm [Friedman & Tukey 1974]. Projection pursuit(PP) is a statistical technique developed to find interesting structures in the data. Interesting structures are found via linear projections in which the distribution of the projected data differs as much as possible from the Gaussian distribution. Friedman justifies the non-interest of the normal distribution based on a series of properties as all the projections of a multivariate normal distribution are normal or that, for a fixed variance, the normal distribution has the least information (Fisher, negative entropy). The deviation from a Gaussian is measured through an index that measures the non-normality of the projected data.

In 1D, Friedman looks for a projection of the sphered data Z, $X = \alpha^T Z$, such that the index

$$I(\alpha) = \frac{1}{2}\sum_{j=1}^{J}(2j+1)\left[\frac{1}{N}\sum_{i=1}^{n}P_j(2\Phi(\alpha^T z_i)-1)\right]^2$$

is maximized. P_j is the Legendre polynomium of order j and $\Phi(X)$ is the standard normal density function.

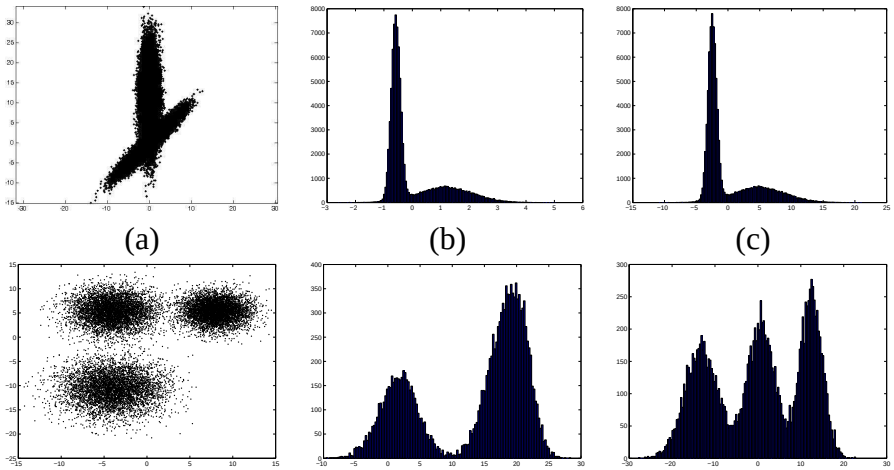


Figure A.4: Top row: a) A two dimensional dataset composed of two Gaussian populations. b) Histogram of the projected data obtained using the combination found by PP. c) Histogram of the projected data obtained using the combination found by HP. Bottom row: d) A three dimensional dataset composed of three Gaussian populations. e) Histogram of the projected data obtained using the combination found by PP. f) Histogram of the projected data obtained using the combination found by HP.

Once an interesting projection has been found, the information obtained by this projection is removed and the algorithm looks for the next informative view. This process consists in transforming the data so that the density of the transformed data Z^{k+1} is as close as possible to the old data Z^k under the constraint that its marginal density is normal. This produces the closest distribution in the sense of the relative entropy distance measure

$$\int log(Z^k/Z^{k+1})Z^k dZ \quad .$$

As it can be observed in Figure A.4 (b), Friedman's algorithm finds a projection that separates the two populations embedded in the synthetic dataset analyzed previously with PCA. This indicates that, from a classification point of view, maximizing the non-gaussianity of the projected data is a more appropriate criterion than to maximize the variance. However, maximizing the non-gaussianity of the projected data is too general. This may in datasets with more than two classes, or datasets that have some non-gaussian variables, e.g. uniform variables, result in the projection found by PP to be not optimal and thereby require more than just one projection. This would cause the computational inconvenience of having to analyze each projection found in order

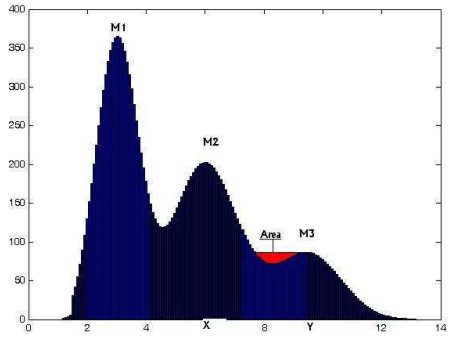


Figure A.5: Region where HP calculates the index.

to discover the combination that enhances the desired structure. This fact is illustrated in Figure A.4. Figure A.4 (e) shows the histogram of the data projected on the first projection obtained by PP of the dataset illustrated in Figure A.4 (d). This dataset is composed of three Gaussian populations with 5000 data points each, means $[10-1], [1015], [2215]$ and covariance matrices $\begin{pmatrix} 9 & 0 \\ 0 & 9 \end{pmatrix}$, $\begin{pmatrix} 6 & 0 \\ 0 & 6 \end{pmatrix}$ and $\begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}$. Note, that the first projection found by PP discriminates one of the populations with respect to the others. If the desired structure is not the discriminated, then a second projection must be obtained in order to discriminate the wanted structure. However, there exists a one-dimensional projection that separates all of the three populations.

In order to find this combination, the proposed algorithm modifies Friedman's index in order to incorporate information about the number of structures included in the image. If the image to be analyzed is assumed to have n classes, the index associated to a specific projection is defined as the $n-1$ largest area between two consecutive modes in the histogram of the projected data. The region where the HP algorithm calculates its index is labeled with and $Area$ and displayed in Figure A.5. If M_{min} represents the minimum histogram value calculated in the two maximums that define the area (x and y), n_{bins} is the number of bins between these two maximums, and $H(i)$ is the value of the i^{th} bin, then the index is calculated by:

$$I(\lambda) = \left(\sum_{i=x}^{i=y} min(H(i), M_{min}) \right) - M_{min} \times n_{bins}.$$

Notice that this index is scale invariant. If the found combination is $\sum_i^{n_{bands}} \alpha_i B_i$, then the combination $\beta(\sum_i^{n_{bands}} \alpha_i B_i)$,$\beta \in \Re$, has the same index. In order to force the algorithm to provide only projections with n modes, the algorithm gives an index of zero to all projections with a number of modes different to n. A pseudo-code to

calculate the index is given by:

```
Let H be an obtained histogram, and n the
 number of classes in the image

1- Smooth H to remove insignificant maxima.

2- Detect all the local maxima of the
    smoothed histogram. Set n_max to the
    desired number of maximums in H.

3- If n_max is equal to n then
    3.a- FOR i equal 1 to n-1
           find the area between maximum i and
           maximum i+1
    3.b- Index equal to the n-1 largest area.

4- Else
    Index=0.

5- Return Index
```

The optimization in this work is conducted using genetic optimization [Goldberg 1989].

A.4 Experimental results

In this section, three experiments are conducted to test the accuracy and applicability of the proposed equipment and segmentation techniques. The first experiment aims to show the accuracy and reproducibility of the obtained images. The last two experiments show the results obtained by the segmentation technique in two different databases: a dermatological and a mycology database.

Experiment 1: Testing the performance of the Videometer-Lab to collect reproducible and accurate images

The first experiment aims at demonstrating the accuracy of the system and the reproducibility of the acquired images. Reproducibility means that if the same image is collected at different times, the results should be comparable. This fact is really important when the objective is to detect and evaluate changes in bitemporal images. It

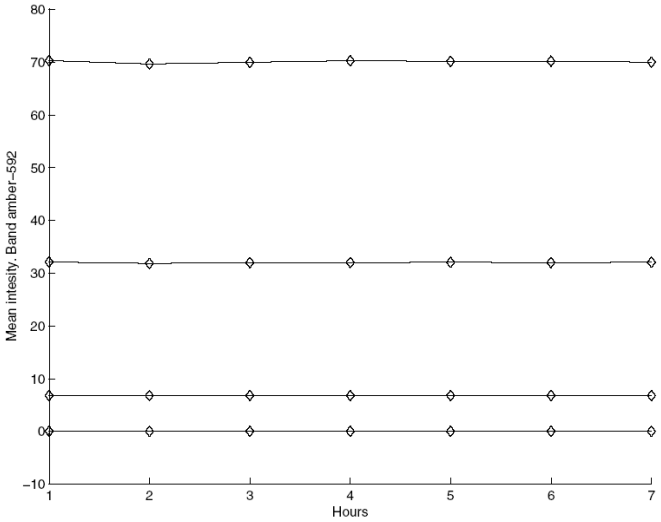


Figure A.6: Variation in the measurements of the NCS respect to the time that the equipment was turned on in the amber band, 592 nm.

guarantees that the differences in two images taken some time apart do not depend on the conditions under which they have been taken. For instance, this quality is of prime importance in applications such as evaluation of dermatological lesions where it is important to ensure that differences in the obtained measures depend only of changes in the lesion.

In order to assess the reproducibility of the images, the equipment was kept turned on during 7 hours. The set-up was calibrated every hour and images of four Natural Color System sheets ($1500N, 2500N, 5000N, 8500N$) from Scandinavian Color Institute were collected. The NCS sheets are all painted and have very small variation. The mean of each spectral band of the collected images was calculated. If the system performs accurately, the mean should not vary significant with respect to time. Marks were placed in the NCS sheets to calculate the mean in approximately the same area.

Figure A.6 shows the evolution of the measures with respect to time of the four NCS sheets in the amber band ($592nm$). Results obtained in the other bands are similar to that obtained in this band. From the figure, it is noticed that the variation is minimal. After the first hour, where the equipment reached thermal equilibrium, the differences are inappreciable. Moreover, for fixed NCS sheet, the variance of the obtained measurements for each band is minimal.

In table A.1, the variance of the measurements obtained for each band of the different NCS sheets is displayed. This small variance guarantees that measures obtained in the

Band/NCS number	1500 N	2500 N	5000 N	8500 N
Blue 472	0.0007	0.0105	0.0236	0.0348
Green 515	0.0002	0.0012	0.0028	0.0074
Amber 592	0.0013	0.0371	0.1295	0.1563
Red 630	0.0012	0.0078	0.0199	0.0222
Near IR 875	0.0010	0.0062	0.0434	0.0366
Ultra Blue 428	0.0058	0.0057	0.0141	0.0320
Cyan 503	0.0003	0.0011	0.0023	0.0086
Orange 612	0.0004	0.0066	0.0234	0.0319
Near IR 940	0.0001	0.0076	0.0501	0.0726

Table A.1: Variance of the seven means obtained for each NCS sheet in each spectral band.

image depend only on the structure being analyzed and it shows the robustness of the equipment.

Experiment 2:Segmenting 9 multi-spectral band psoriasis images

The goal of the third experiment is to assess the use of multi-spectral images when analyzing dermatological lesions. Nowadays, the medical tracking of dermatological diseases is imprecise. The main reason is the lack of suitable objective methods to evaluate the lesions. Presently, the severity of the disease is scored by doctors just through their visual examination. Doctors visually assess the lesion and make scorings and journal notes of the current condition. These notes and perhaps some photographs are usually the only memory of what the lesion looked like at the corresponding visit. Image analysts have tried to provide different solutions to these problems during the last decades [Engstrom, Hansson, Hellgren, Tomas, Nordin, Vincent & Wahlberg 1990]. However, difficulties in correctly acquiring the images [Gutenev et al. 2001], the limited information provided by the trichromatic images and the presence of artifacts such as hair [Chung & Sapiro 2000] cause that precise and objective scores of the severity of the lesions cannot be obtained. In order to evaluate the benefits of using multi-spectral images, a collection of eight multi-spectral psoriasis images were collected in collaboration with the dermatological department of Gentofte Hospital in Denmark. These multi-spectral images were composed of nine spectral bands ranging from 472 nm to 940 nm.

The nine bands of one of the collected images together with their associated wavelengths are displayed in Figure A.7. It is seen that one of the bands mainly shows the

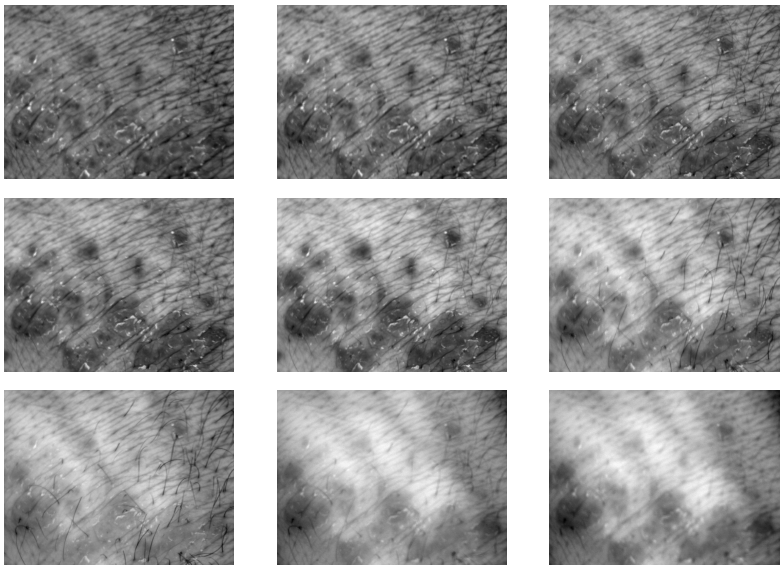


Figure A.7: The nine multi-spectral bands of one of the images. Top Left: ultra-blue, 428. Top Center: blue, 472. Top Right: Cyan, 503. Middle Left: green, 515. Middle Center: amber, 592. Middle Right: orange, 612. Bottom Left: red, 630. Bottom Center: near infrared 875. Bottom Right: near infrared 940.

hair and the veins ($630nm$). This situation was also observed in the other psoriasis images which presented these two structures (Figure A.8 (A) and (B)). This fact indicates that the multi-spectral images provided a more informative representation of the lesion than the traditional RGB images. This extra information can be used to obtain a more precise evaluation of the lesion where hair and veins are removed.

In order to statistically assess the information provided by the extra bands, the images were segmented using the HP algorithm. The HP algorithm found a projection where the lesion exhibited a considerable contrast with respect to the other structures involved in the image (Figure A.8 (C)).The data in these projections are distributed approximately according to a mixture of two Gaussians. The parameters of this model can be estimated [Taxt, Hjort & Eikvil 1991] and the lesion extracted via discriminant analysis. Results of the segmentation are shown in Figure A.8 (D). It is observed that the nine multi-spectral bands provide enough information to precisely separate the lesion from the other parts of the images. The segmented images were used to assess the information provided by the extra bands in terms of Mahalanobis distances between classes. Given two classes X and Y with observations $X_1, ..., X_{n_1}$ belonging to X and observations $Y_1, ..., Y_{n_2}$ belonging to Y, Mahalanobis distance between X and Y is defined by

$$(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2),$$

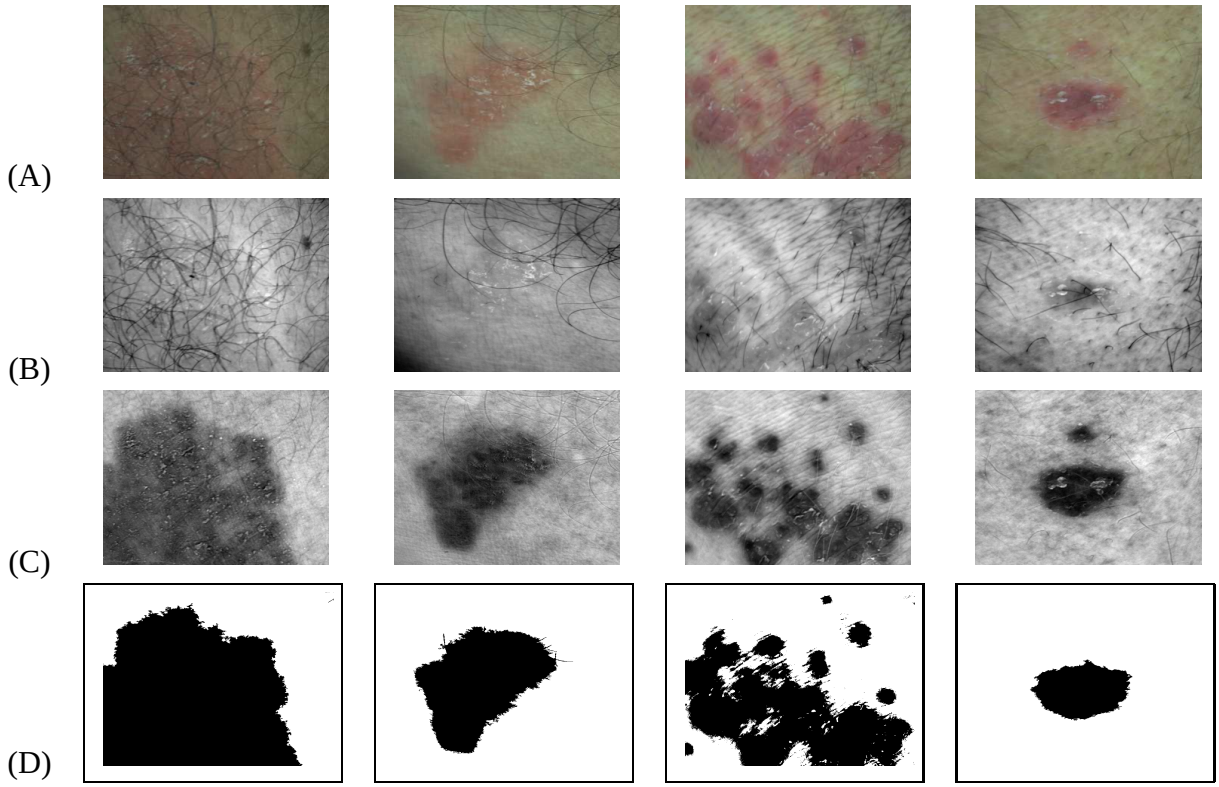


Figure A.8: (A) Four psoriasis images. (B) Spectral band $630nm$. (C) Projection image founf by the HP algorithm. (D) Lesion Segmentation.

Image	Mahalanobis distance using the RGB bands	Mahalanobis distance using the nine bands
1	10.0793	12.8460
2	2.9048	10.3904
3	7.3857	12.2284
4	14.8222	17.4322
5	1.8920	23.4698
6	23.4068	38.4291
7	7.1864	9.9264
8	18.0009	31.8217

Table A.2: Mahalanobis distances between the lesion and the other structures involved in the image.

where μ_1 and μ_2 are the mean of classes X and Y respectively. and $\sum$ is defined by

$$\Sigma = \frac{1}{n_1 + n_2 - 2}(\sum_i (X_i - \bar{X})(X_i - \bar{X})^T + \sum_i (Y_i - \bar{Y})(Y_i - \bar{Y})^T).$$

The mahalanobis distances, for each of the eight images, between the lesion and the class composed of the other structures in the image (healthy skin, hair,...) using the nine bands and using only a RGB approximation are shown in Table A.2. It can be observed that the distance increases considerably when the nine bands are used. However, a more meaningful measure based on these measures is to statistically test the null hypothesis that the six extra bands does not contribute to a better discrimination. Specifically, if the extra six variables do not contribute to a better discrimination, then

$$Z = \frac{n_1 + n_2 - p - 1}{q} \frac{n_1 n_2 (D_p - D_q)}{(n_1 + n_2)(n_1 + n_2 - 2) + n_1 n_2 D_q}$$

follows a $F(q, n1 + n2 - p - 1)$ distribution, where n_1 and n_2 are the number of observations on each class, p is the total number of variables, q is the number of variables that are to be tested if they do or do not contribute to a better discrimination and D_p and D_q are the mahalanobis distances between classes using all the variables and all the variables except the last q. Results showed that statistically the null hypothesis was rejected with a significance level of 1%. This means that the last six variables strongly contribute to a better discrimination.

Experiment 3:Segmenting 18 multi-spectral band fungi images

Classification of fungi is of importance for several reasons; for a further phylogenetic study or to reveal new species or isolates to use in e.g. food or medical industries.

Traditionally, the classification has been performed by means of chemical and visual studies of the fungi. In the last decade digital image analysis has also been utilised for the classification, but till now it has been based on RGB images as in [Hansen 2003].

The species can be differentiated by macroscopic features, microscopic features and behaviours like e.g. thermophilicity (whether or not they can grow at high temperatures). The macroscopic features are the ones captured by the image acquisition.

The Penicillium genus was chosen due to the large knowledge of and well identified isolates. Penicillium is a filamentous fungi also known as mold. Most of the species are found in the soil and in the air. They are known to produce mycotoxins. The mycotoxins can cause infections when in contact with humans, though, depending of the type of mycotoxin. The fungi can also be used to produce antibiotics, antitoxins and other drugs.

Multi-spectral images with 18 wavelengths are examined. Three species are examined: polonicum, venetum and melanoconodium of the Penicillium genus. It is assumed that the many spectra additionally can reveal some chemical information about the fungi compared to the ordinary RGB images. Within each specie four different isolates were chosen, all obtained from the IBT Culture Collection held at BioCentrum-DTU. They were chosen with geographical origin in different countries to get a greater variance within each specie. Each isolate was grown on three different media: OAT (Oatmeal Agar), YES (Yeast Extract Sucrose Agar) and CYA (Czapek Yeast Extract Agar), with three replicas on each medium to obtain the variance wihin each isolate. The isolates are grown on three media to get acces to more information. This is the usual practice when isolates are to be identified. In total there are 108 samples.

The first step is to segment the background, the petri dish and the fungi into three classes. The next step is to examine each of the three classes and then repetively examine each of the subclasses obtained for furthere classes until a subclass no longer can be split in two or more. The interest is to segment the fungi from the background as well as the petri dish, and if possible extract information of differences within the fungi. This is done in order to extract features to be used in a further classification of the species. The first step is straight forward in all cases where as the following examinations differ depending on the appearance of the individuals.

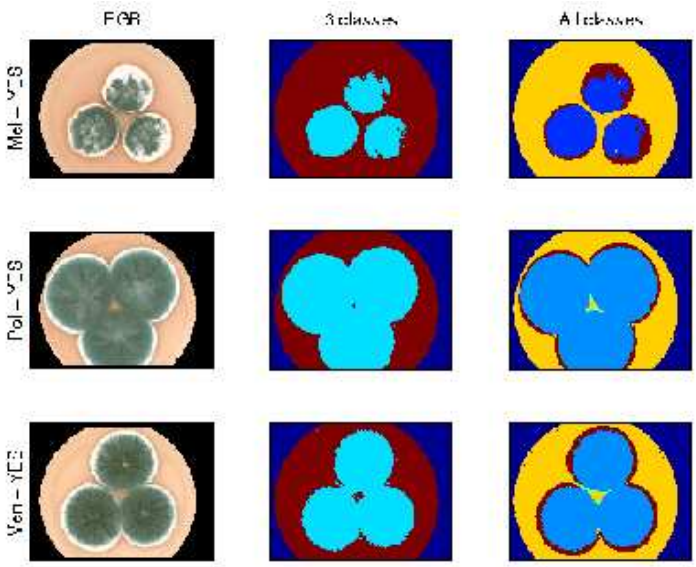


Figure A.9: Segmentation of a a melanoconidium, polonicum, and a venetum species all on the YES medium with IBT numbers: 3445, 22439 and 21549, respectively. First coloumn illustrates RGB representations of the multi spectral images. Second coloumn illustrates the first segmentation in to three classes. The third coloumn illustrates the final segmentation.

Results of the segmentation

Figure A.9 shows examples of segmentations within the images of the three species grown on the YES medium. The fungi are well seperated from both petri dish and background, and furthermore, the lighter edge of the fungi can be separated from the darker center of the fungi. The latter can be usefull since the different species differ in appearance at this point. The images are foremost split into 3 classes; the background, the medium and the fungi. As this is not sufficient the medium and the fungi classes are further examined for subclasses. Subdividing further, the lighter edge is separated from the medium class and small segments of the medium is separated from the fungi class.

Figure A.10 illustrates two examples on the OAT medium where the lighter edge of the fungi are segmented from the medium classes. Another example of a melanoconidium on YES medium is shown. In this case the lighter areas of the fungi are classified as fungi first time, but partitioning further gives a subdivision of the fungi area.

In Figure A.11 the division of the segmented medium was performed using three classes. For the venetum isolate in the middle row the segmented fungi was divided

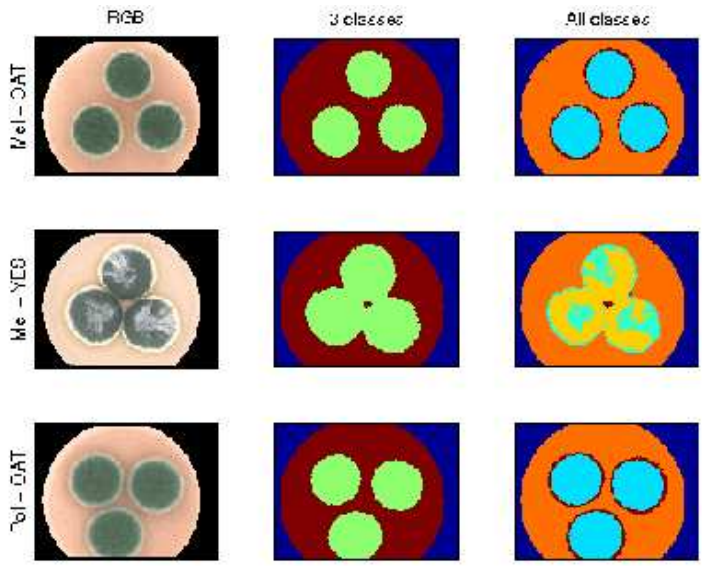


Figure A.10: Segmentation of a polonicum, and two melanoconidium species on the OAT and YES media with IBT numbers: 22439, 3445 and 10031, respectively. First coloumn illustrates RGB representations of the multi spectral images. Second coloumn illustrates the first segmentation in to three classes. The third coloumn illustrates the final segmentation.

further as it contained some of the medium. The edge of the fungi was not identified when first dividing the segmented medium, but at the following segmentation. The divisions of the media may be usefull for examinations of the chemicals the fungi produce during the growth.

Figure A.12 illustrates isolates where the fungi can be divided into more subgroups than two; the edge and the center of the fungi. Two melanoconidium isolates and one venetum isolate are shown on the CYA and YES media.

Segmentations of multi-spectral images of the three Penicillium species on the three different media have been conducted. Examples from each group have been illustrated. There are three examples where the appearance of the fungi have some variance within the 9 groups and these are also illustrated. The results shown illustrate that the fungi are well separated from the media for different isolates. Furthermore, the method can be used to find subclasses within the fungi.

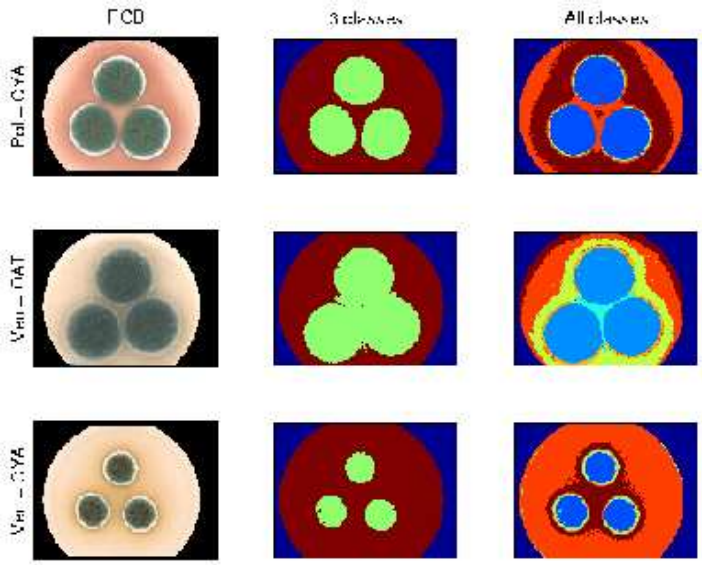


Figure A.11: Segmentation of a polonicum and two venetum species all on the CYA and OAT media with IBT number: 15982, 23039 and 16215, respectively. First coloumn illustrates RGB representations of the multi spectral images. Second coloumn illustrates the first segmentation in to three classes. The third coloumn illustrates the final segmentation.

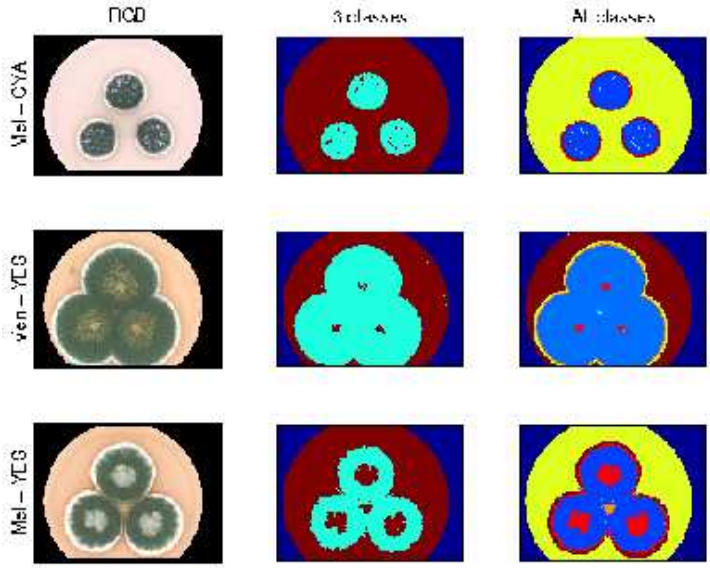


Figure A.12: Segmentation of a melanoconidium and two venetum species all on the CYA and YES media with the IBT numbers: 21534, 23039 and 21534, respectively. First coloumn illustrates RGB representations of the multi spectral images. Second coloumn illustrates the first segmentation in to three classes. The third coloumn illustrates the final segmentation where each of the three classes first found are examined for further divisions.

A.5 Conclusions

In this work, a system to collect precise and reproducible multi-spectral dermatological images has been proposed. The system can collect up to twenty different spectral bands. These bands are composed by the RGB tri-chromatic spectral bands plus seventeen extra bands that can be choosen in the range going from ultra-blue to near infrared (from 395 nm to 970 nm). The reproducibility of the equipment has been tested. A novel algorithm that combines the information of all the spectral bands in order to segment the interesting areas have also been provided. Results indicate that the equipment and the segmentation algorithm are suitable tools to measure changes in the evolution of dermatological disease. Furthermore, it has been observed that the six extra bands provide more information than the classical RGB images. This information can be used to remove noise such as hair or occlusions and to obtain more precise measures to characterize the lesion. Furthermore, the applicability of the equipment and the segmentation algorithm was tested on a second data base of fungi images. It was shown that fungi as well as some structures in the fungi can be segmented to obtain features for further classification. Results point out the proposed imaging system as a suitable tool for obtaining measures that characterize the objects under study.

A.6 Acknowledgment

The authors would like to thank to the SITE Project funded by a grant from the Danish Technical Research Foundation (Project Number STVF 56-00-0123) for supporting the present work. The author would also like to thanks to the dermatologists Lone Skov and Bo Bang of Gentofte Hospital of Denmark and to the anonymous patients, for their collaboration during the image acquisition. The authors also thank Jens Christian Frisvad from Biocentrum department at the Technical University of Denmark for providing the fungi samples.

Appendix B

Mycotoxins produced by P. mel, P. pol and P. ven

Natutal products produced by the three species [Frisvad et al. 2004]:

P. *melanoconidium*:

1 Penicillic acid, dehydropenicillic acid, orsellinic acid

2 Verrucosidin, normethylverrucosidin

3 Xanthomegnin, viomellein, vioxanthin

4 Penitrem A, B, C, D, E, F

5 Roquefortine C & D, melagrin, oxaline

6 Sclerotigenin

P. *polonicum*:

1 Penicillic acid, dehydropenicillic acid, orsellinic acid

2 Verrucosidin, normethylverrucosidin

3 Verrucosidin, verrucosinol, puberuline A, fructigenine A, dehydroverrucosine, demethylverrucosine, rugulosuvine, lecytryptophanyldiketopiperazine

4 Cyclopeptine, dehydrocyclopeptine, cyclopepin, cyclopenol, 3-methoxyviridicatin, viridicatol

5 Anacine

6 Asperterric acid

7 Methyl-4-(2-(2R)-hydroxyl-3-butynyloxy) bezonate

8 Nephrotoxic glycopeptides

P. *venetum*:

1 Cyclopeptine, dehydrocyclopeptine, cyclopepin, cyclopenol, 3-methoxyviridicatin, viridicatol

2 Terrestric acid

3 Roquefortine C

4 Atrovenetins

5 Corymbiferan lactone C & D, corymiferone

Appendix C
RGB representations of fungi

Some of the images seem more yellow than the visual appearance - in particular venetum on CYA, since agar as well as the red liquid drops on top of these samples are see-through, and hence the wavelengths for green light are still reflected by the agar or the green fungi underneath the drops, respectively.

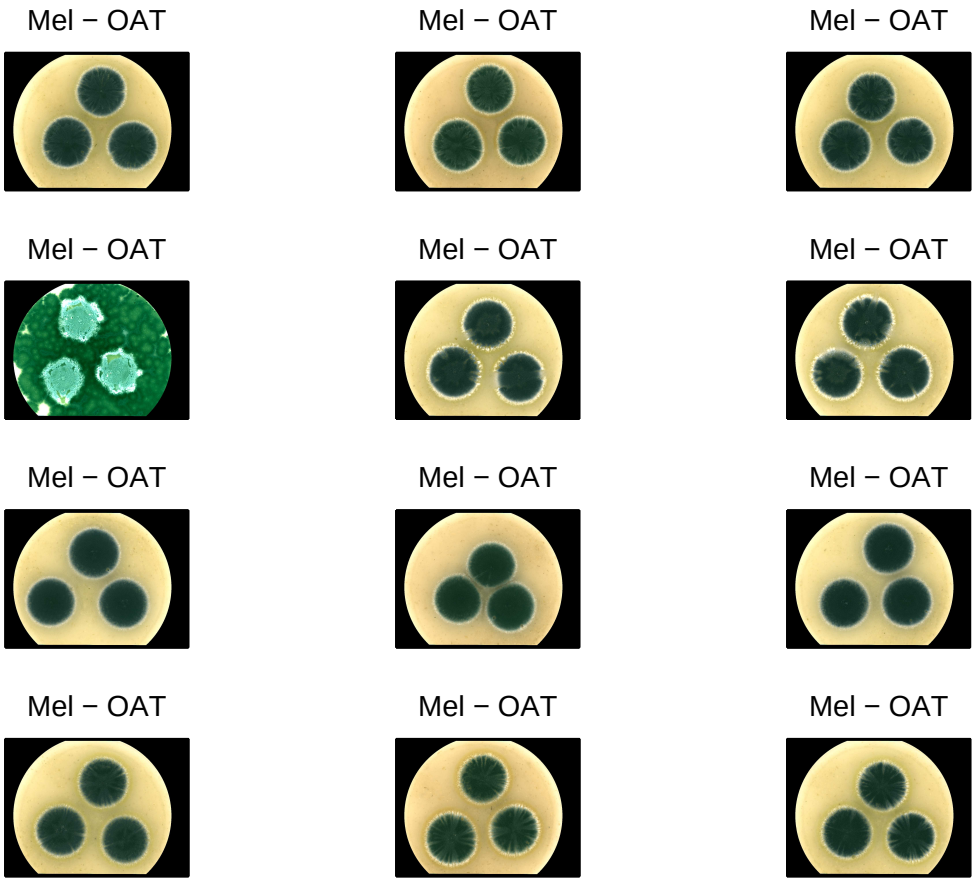


Figure C.1: RGB representation of melanoconidium on OAT.

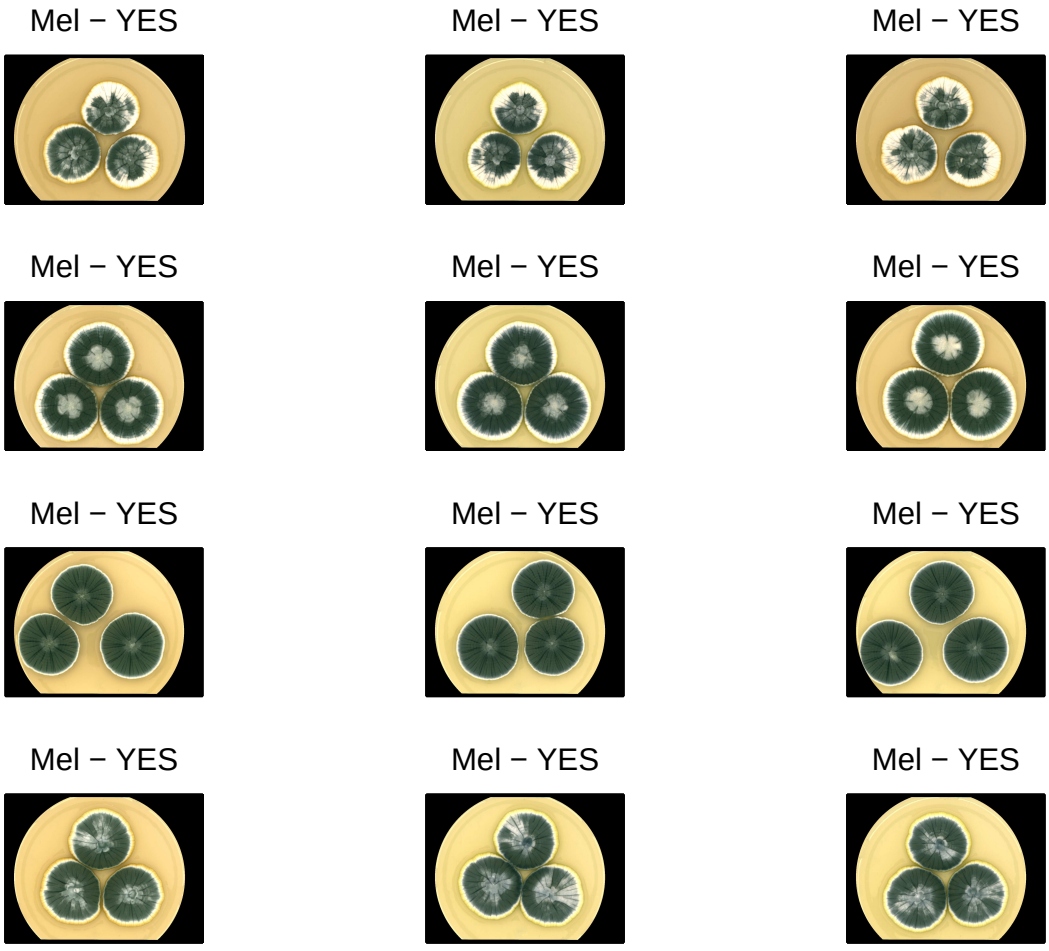


Figure C.2: RGB representation of melanoconidium on YES.

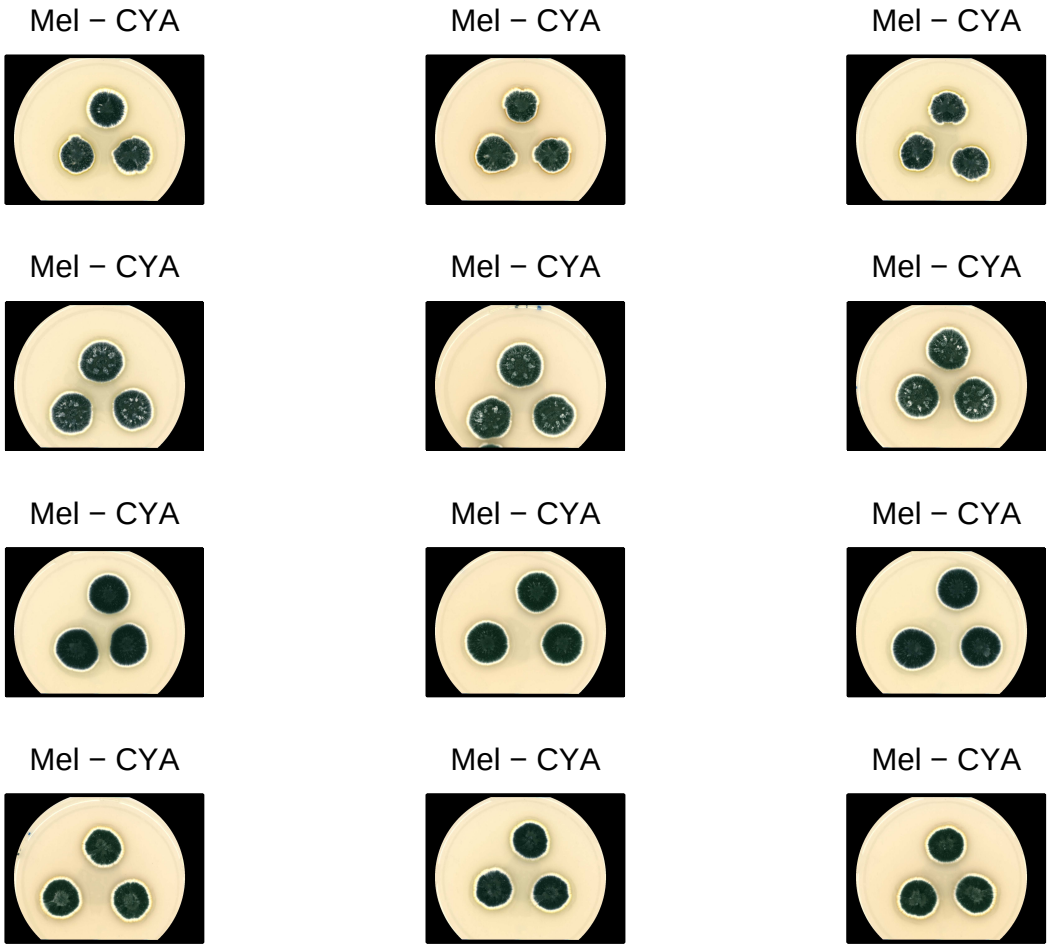


Figure C.3: RGB representation of melanoconidium on CYA.

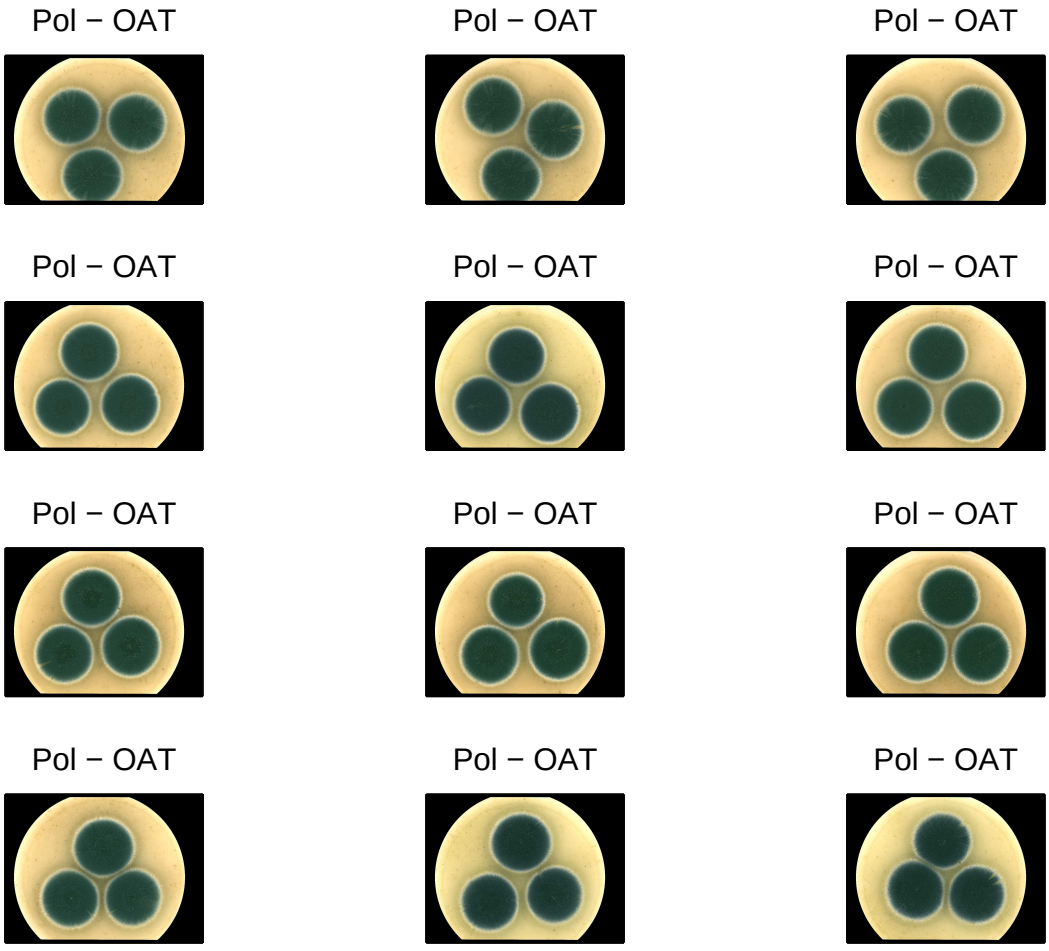


Figure C.4: RGB representation of polonicum on OAT.

Figure C.5: RGB representation of polonicum on YES.

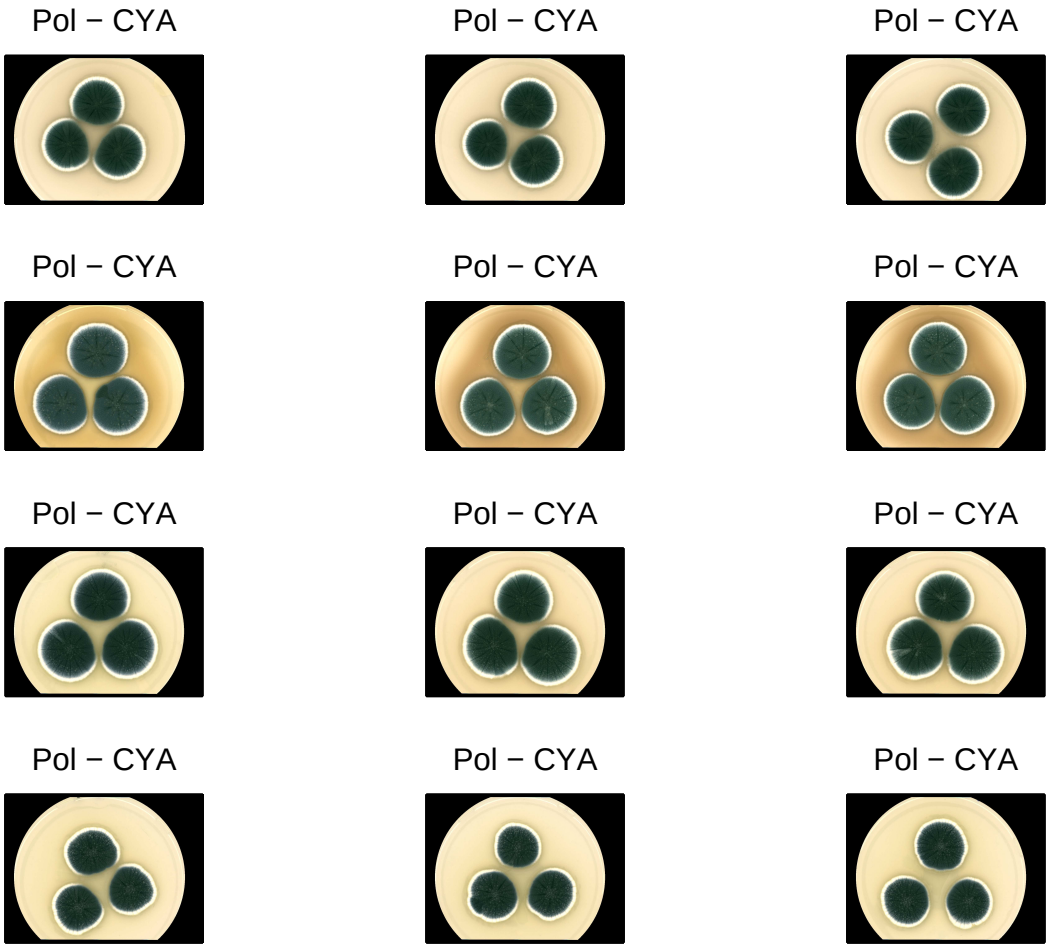


Figure C.6: RGB representation of polonicum on CYA.

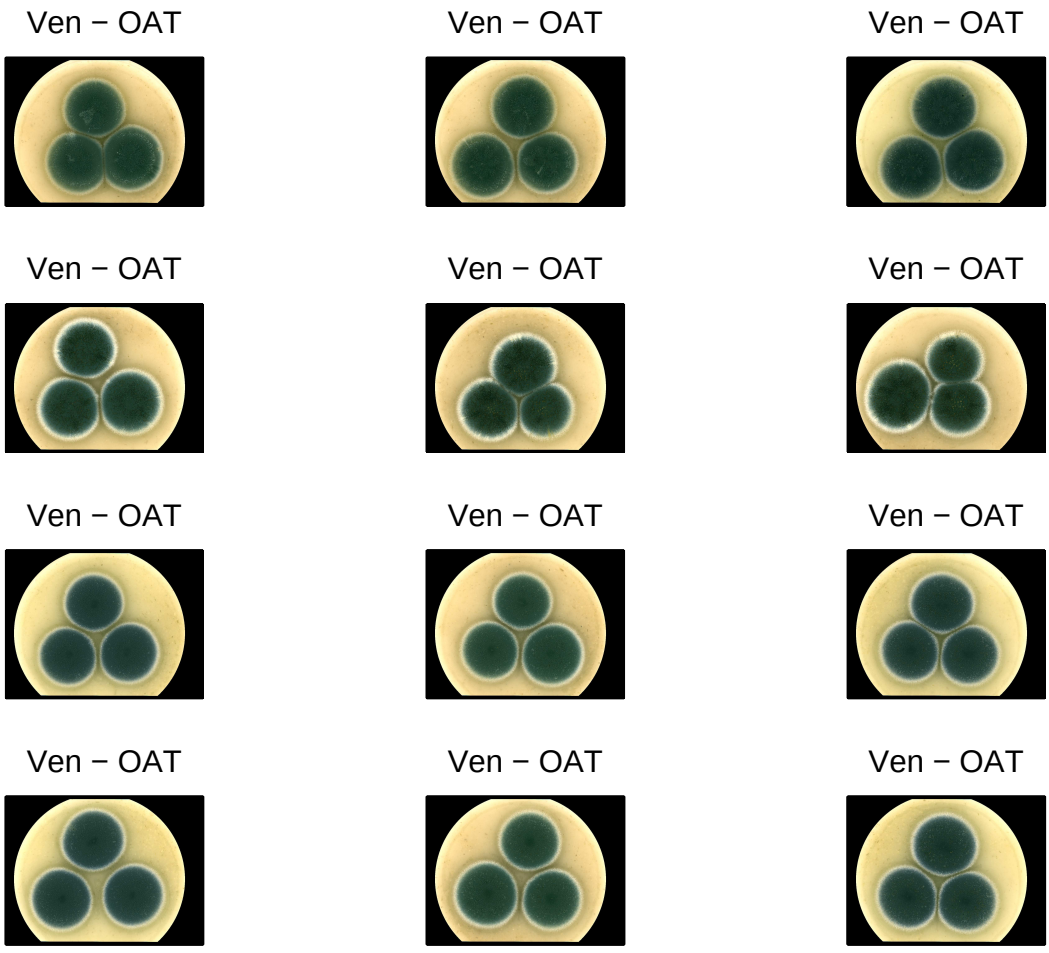


Figure C.7: RGB representation of venetum on OAT.

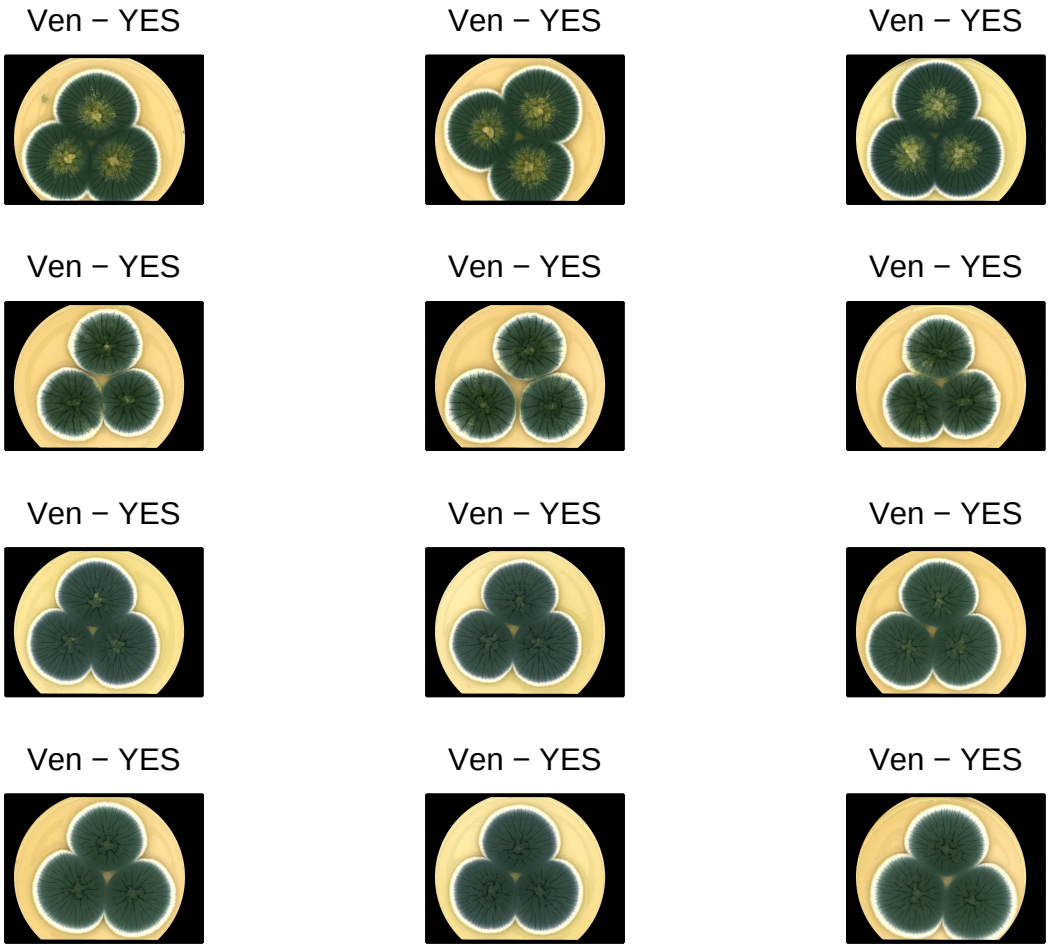


Figure C.8: RGB representation of venetum on YES.

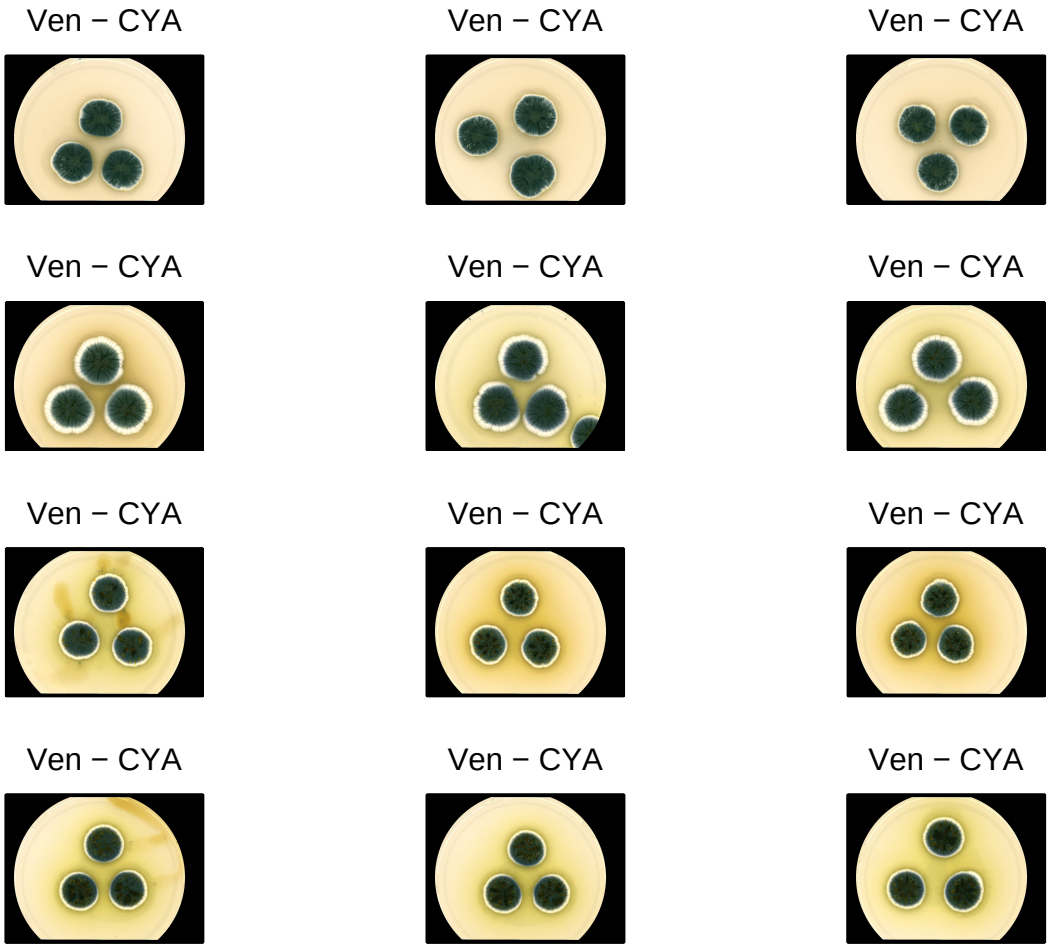


Figure C.9: RGB representation of venetum on CYA.

Appendix D
Mathematics and Statistics

D.1 Approximation of U-distribution by F-distribution

Theorem A[Conradsen 2002*a*] Let U be $U(p,q,r)$-distributed and

$$\begin{aligned} t &= \begin{cases} 1 & p^2+q^2=5 \\ \sqrt{\frac{p^2q^2-4}{p^2+q^2-5}} & p^2+q^2\neq 5 \end{cases} \\ v &= \frac{1}{2}(2r+q-p-1) \quad . \end{aligned}$$

Then

$$F = \frac{1-U^{1/t}}{U^{1/t}} \cdot \frac{vt+1-\frac{1}{2}pq}{pq}$$

is approximately distributed as

$$F(pq, vt+1-\frac{1}{2}pq) \quad .$$

The approximation is exact if either p or q equals 1 or 2.
Proof Omitted.

□

D.2 Three-sided Analysis of Variance

Variation	Formula for calculating SS
M	$3 \cdot 4 \cdot 3 \cdot \sum_{k=1}^{3} (\bar{X}_{k\ldots} - \bar{X})^2$
S	$3 \cdot 4 \cdot 3 \cdot \sum_{l=1}^{3} (\bar{X}_{.l.} - \bar{X})^2$
MS	$3 \cdot 4 \cdot \sum_{k=1}^{3} \sum_{l=1}^{3} (\bar{X}_{kl.} - \bar{X}_{k\ldots} - \bar{X}_{.l.} + \bar{X})^2$
I(S)	$3 \cdot 3 \cdot \sum_{l=1}^{3} \sum_{j=1}^{4} (\bar{X}_{.lj} - \bar{X}_{.l.})^2$
MI(S)	$3 \cdot 3 \cdot \sum_{k=1}^{3} \sum_{j=1}^{4} (\bar{X}_{k.j} - \bar{X}_{k\ldots} - \bar{X}_{..j} + \bar{X})^2 +$ $3 \cdot \sum_{k=1}^{3} \sum_{l=1}^{3} \sum_{j=1}^{4} (\bar{X}_{klj} - \bar{X}_{kl.} - \bar{X}_{k.j} - \bar{X}_{.lj} + \bar{X}_{k\ldots} + \bar{X}_{.l.} + \bar{X}_{..j} - \bar{X})$
R(MSI)	$\sum_{k=1}^{3} \sum_{l=1}^{3} \sum_{j=1}^{4} \sum_{\nu=1}^{3} (\bar{X}_{klj\nu} - \bar{X}_{klj.})^2$
Total	$\sum_{k=1}^{3} \sum_{l=1}^{3} \sum_{j=1}^{4} \sum_{\nu=1}^{3} (\bar{X}_{klj\nu} - \bar{X})^2$

Table D.1: Sums of Squares for models [Conradsen 2002*a*].

Variation	E(SS/f)	Test against
M	$\sigma^2 + 3 \cdot 3 \cdot 4 \cdot \sigma^2_M$	R(MSI)
S	$\sigma^2 + 3 \cdot 3 \cdot 4 \cdot \sigma^2_S$	R(MSI)
MS	$\sigma^2 + 3 \cdot 4 \cdot \sigma^2_{MS}$	R(MSI)
I(S)	$\sigma^2 + 3 \cdot 3 \cdot \sigma^2_{I(S)}$	R(MSI)
MI(S)	$\sigma^2 + 3 \cdot \sigma^2_{MI(S)}$	R(MSI)
R(MSI)	σ^2	Total

Table D.2: Expected Sums of Squares for Model 8.2 [Conradsen 2002*a*].

Variation	E(SS/f)	Test against
M	$\sigma^2 + 3 \cdot \sigma^2_{MI(S)} + 3 \cdot 3 \cdot 4 \cdot \sigma^2_M$	MI(S)
S	$\sigma^2 + 3 \cdot \sigma^2_{MI(S)} + 3 \cdot 3 \cdot \sigma^2_{I(S)} + 3 \cdot 3 \cdot 4 \cdot \sigma^2_S$	I(S)
MS	$\sigma^2 + 3 \cdot \sigma^2_{MI(S)} + 3 \cdot 4 \cdot \sigma^2_{MS}$	MI(S)
I(S)	$\sigma^2 + 3 \cdot \sigma^2_{MI(S)} + 3 \cdot 3 \cdot \sigma^2_{I(S)}$	MI(S)
MI(S)	$\sigma^2 + 3 \cdot \sigma^2_{MI(S)}$	R(MSI)
R(MSI)	σ^2	Total

Table D.3: Expected Sums of Squares for Model 8.3 [Conradsen 2002*a*].

D.3 Hotelling's T^2-test

For two normallly distributed classes $\pi_1 \leftrightarrow N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\pi_2 \leftrightarrow N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ of n_1 and n_2 observations and p variables, respectively, Mahalanobi's distance is given by [Conradsen 2002*a*]:

$$D^2 = (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)^T \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2) \quad , \tag{D.4}$$

where $\hat{\boldsymbol{\mu}}_i$ is the sample mean of class i and $\hat{\boldsymbol{\Sigma}}$ is the within group variance matrix weighted by $n_1 + n_2 - 2$. Hotelling's T^2-test [Conradsen 2002*a*]:

$$H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_1 : \mu_1 \neq \mu_2 \quad , \tag{D.5}$$

is given by the test size:

$$Z = \frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)} \cdot \frac{n_1 n_2}{n_1 + n_2} D^2 \tag{D.6}$$

which under the null-hypothesis is $F(p, n_1 + n_2 - p - 1)$-distributed.

D.4 Test of contribution to discrimination

Consider two normallly distributed classes $\pi_1 \leftrightarrow N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\pi_2 \leftrightarrow N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ of n_1 and n_2 observations and p variables, respectively. A test of the null-hypotesis that the last q variables do not contribute to the discrimination is given by the test size [Conradsen 2002*a*]:

$$Z = \frac{n_1 + n_2 - p - 1}{q} \frac{n_1 n_2 (D_p^2 - D_{p-q}^2)}{(n_1 + n_2)(n_1 + n_2 - 2) + n_1 n_2 D_{p-q}^2} \quad , \tag{D.7}$$

where D_p^2 is Mahalanobi's distance based on the first p variables. Under the null-hypothesis $Z \in F(q, n_1 + n_2 - p - 1)$.

Appendix E
Results Fungi

E.1 Singular values

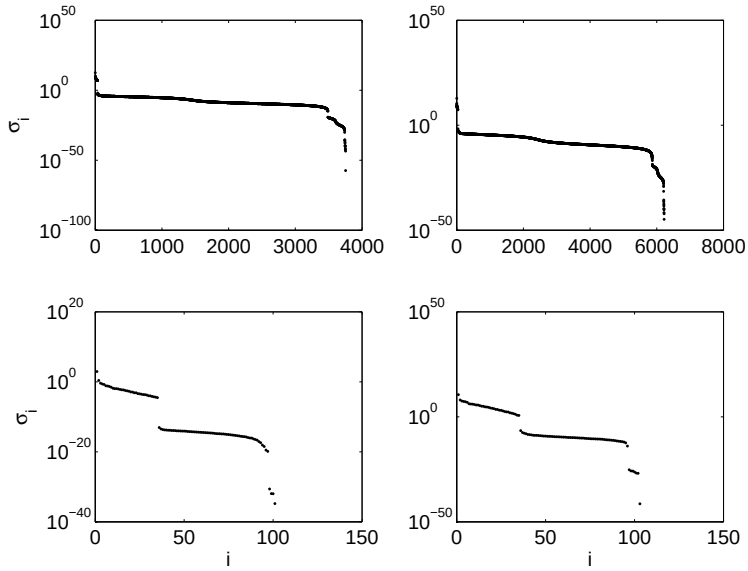


Figure E.1: Plot of singular values for the fungi datasets on OAT. From upper left corner: features from edge and fungi together, edge and fungi separate, linear combinations of the visual bands to represent RGB and the three badnds closest to RGB.

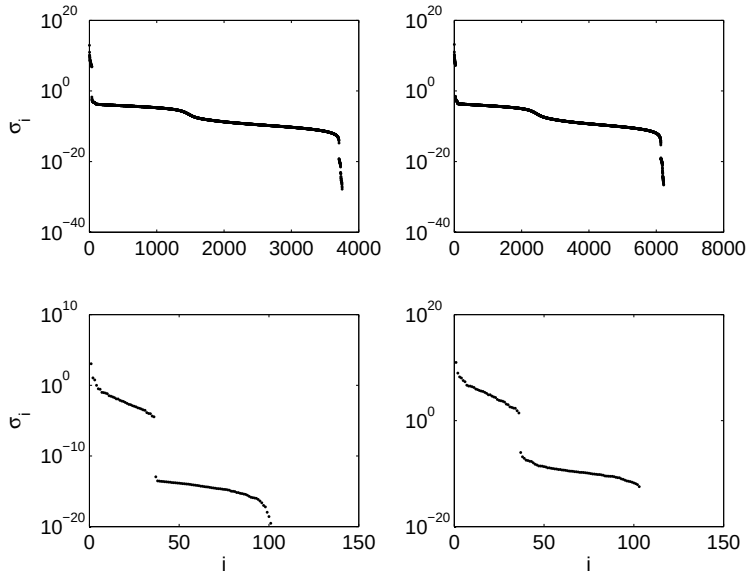


Figure E.2: Plot of singular values for the fungi datasets on CYA. From upper left corner: features from edge and fungi together, edge and fungi separate, linear combinations of the visual bands to represent RGB and the three badnds closest to RGB.

E.2 Analysis of Variance

E.2.1 RSS for ANOVA Tables

Variation	SS	f	SS/f
M	$3.45 \cdot 10^3$	2	$1.72 \cdot 10^3$
S	$2.51 \cdot 10^3$	2	$1.25 \cdot 10^3$
MS	$1.28 \cdot 10^4$	4	$3.20 \cdot 10^3$
I(S)	$1.39 \cdot 10^3$	9	$1.54 \cdot 10^2$
MI(S)	$3.93 \cdot 10^3$	18	$2.18 \cdot 10^2$
R(MSI)	$3.74 \cdot 10^2$	72	$5.20 \cdot 10^0$
Total	$2,45 \cdot 10^4$	107	$2.29 \cdot 10^2$

Table E.1: ANOVA for the 95th percentile of difference between 1st and 11th spectra of the dataset with fungi and edge in one.

Variation	SS	f	SS/f
M	$2.92 \cdot 10^3$	2	$1.46 \cdot 10^3$
S	$1.12 \cdot 10^3$	2	$5.59 \cdot 10^2$
MS	$7.33 \cdot 10^3$	4	$1.83 \cdot 10^3$
I(S)	$2.33 \cdot 10^3$	9	$2.58 \cdot 10^2$
MI(S)	$5.63 \cdot 10^3$	18	$3.13 \cdot 10^2$
R(MSI)	$3.47 \cdot 10^2$	72	$4.82 \cdot 10^0$
Total	$1.97 \cdot 10^4$	107	$1.84 \cdot 10^2$

Table E.2: ANOVA for the 10th percentile of multiplication between 1st and 12th spectra of the dataset with fungi and edge in one.

Variation	SS	f	SS/f
M	$1.10 \cdot 10^5$	2	$5.48 \cdot 10^4$
S	$2.34 \cdot 10^4$	2	$1.17 \cdot 10^4$
MS	$5.18 \cdot 10^3$	4	$1.30 \cdot 10^3$
I(S)	$8.04 \cdot 10^3$	9	$8.94 \cdot 10^2$
MI(S)	$1.05 \cdot 10^4$	18	$5.84 \cdot 10^2$
R(MSI)	$1.37 \cdot 10^3$	72	$1.90 \cdot 10^1$
Total	$1.58 \cdot 10^5$	107	$1.48 \cdot 10^3$

Table E.3: ANOVA for the 1st PC of the dataset with fungi and edge in one.

Variation	SS	f	SS/f
M	$4.08 \cdot 10^4$	2	$2.04 \cdot 10^4$
S	$6.31 \cdot 10^3$	2	$3.16 \cdot 10^3$
MS	$2.09 \cdot 10^3$	4	$5.21 \cdot 10^2$
I(S)	$3.50 \cdot 10^3$	9	$3.89 \cdot 10^2$
MI(S)	$4.77 \cdot 10^3$	18	$2.65 \cdot 10^2$
R(MSI)	$9.45 \cdot 10^3$	72	$1.31 \cdot 10^2$
Total	$6.69 \cdot 10^4$	107	$6.26 \cdot 10^2$

Table E.4: ANOVA for the 2nd PC of the dataset with fungi and edge in one.

E.2.2 Tests for univariate ANOVA

H_0	Test Size	F-fractile
$\sigma_R^2 = 0$	$\frac{5.20 \cdot 10^0}{2.29 \cdot 10^2} = 0.02$	$F(72, 107)_{0.01}$=0.60
$m_k = 0,\ k = 1, 2, 3$	$\frac{1.72 \cdot 10^3}{5.20 \cdot 10^0} = 332$	$F(2, 72)_{0.99}$ =4.91
$s_l = 0,\ l = 1, 2, 3$	$\frac{1.25 \cdot 10^3}{5.20 \cdot 10^0} = 242$	$F(2, 72)_{0.99}$ =4.91
$ms_{kl} = 0,\ k = 1, 2, 3,\ l = 1, 2, 3$	$\frac{3.20 \cdot 10^3}{5.20 \cdot 10^0} = 617$	$F(4, 72)_{0.99}$ =3.59
$i(s)_{j(l)} = 0,\ j = 1, 2, 3, 4,\ l = 1, 2, 3$	$\frac{1.54 \cdot 10^2}{5.20 \cdot 10^0} = 29.7$	$F(9, 72)_{0.99}$ =2.66
$mi(s)_{kj(l)} = 0,\ k = 1, 2, 3,$ $j = 1, 2, 3, 4,\ l = 1, 2, 3$	$\frac{2.18 \cdot 10^2}{5.20 \cdot 10^0} = 42.0$	$F(18, 72)_{0.99}$ =2.20

Table E.5: Tests based on Model (8.2) for the 95th percentile of difference between 1st and 11th spectra of the dataset with fungi and edge in one.

H_0	Test Size	F-fractile
$\sigma_R^2 = 0$	$\frac{4.82 \cdot 10^0}{1.84 \cdot 10^2} = 0.03$	$F(72, 107)_{0.01}$=0.60
$m_k = 0,\ k = 1, 2, 3$	$\frac{1.46 \cdot 10^3}{4.82 \cdot 10^0} = 303$	$F(2, 72)_{0.99}$ =4.91
$s_l = 0,\ l = 1, 2, 3$	$\frac{5.59 \cdot 10^2}{4.82 \cdot 10^0} = 116$	$F(2, 72)_{0.99}$ =4.91
$ms_{kl} = 0,\ k = 1, 2, 3,\ l = 1, 2, 3$	$\frac{1.83 \cdot 10^3}{4.82 \cdot 10^0} = 380$	$F(4, 72)_{0.99}$ =3.59
$i(s)_{j(l)} = 0,\ j = 1, 2, 3, 4,\ l = 1, 2, 3$	$\frac{2.58 \cdot 10^2}{4.82 \cdot 10^0} = 53.6$	$F(9, 72)_{0.99}$ =2.66
$mi(s)_{kj(l)} = 0,\ k = 1, 2, 3,$ $j = 1, 2, 3, 4,\ l = 1, 2, 3$	$\frac{3.13 \cdot 10^2}{4.82 \cdot 10^0} = 64.8$	$F(18, 72)_{0.99}$ =2.20

Table E.6: Tests based on Model (8.2) for the 5th percentile of multiplication between 1st and 7th spectra of the dataset with fungi and edge in one.

H_0	Test Size	F-fractile
$\sigma_R^2 = 0$	$\frac{1.90 \cdot 10^1}{1.48 \cdot 10^3} = 0.01$	$F(72,107)_{0.01}$=0.60
$m_k = 0,\ k = 1,2,3$	$\frac{5.48 \cdot 10^4}{1.90 \cdot 10^1} = 2890$	$F(2,72)_{0.99}$ =4.91
$s_l = 0,\ l = 1,2,3$	$\frac{1.17 \cdot 10^4}{1.90 \cdot 10^1} = 616$	$F(2,72)_{0.99}$ =4.91
$ms_{kl} = 0,\ k = 1,2,3,\ l = 1,2,3$	$\frac{1.30 \cdot 10^3}{1.90 \cdot 10^1} = 68.2$	$F(4,72)_{0.99}$ =3.59
$i(s)_{j(l)} = 0,\ j = 1,2,3,4,\ l = 1,2,3$	$\frac{8.94 \cdot 10^2}{1.90 \cdot 10^1} = 47.1$	$F(9,72)_{0.99}$ =2.66
$mi(s)_{kj(l)} = 0,\ k = 1,2,3,$ $j = 1,2,3,4,\ l = 1,2,3$	$\frac{5.84 \cdot 10^2}{1.90 \cdot 10^1} = 30.8$	$F(18,72)_{0.99}$ =2.20

Table E.7: Tests based on Model (8.2) for the 1st PC of the dataset with fungi and edge in one.

H_0	Test Size	F-fractile
$\sigma_R^2 = 0$	$\frac{1.31 \cdot 10^2}{6.26 \cdot 10^2} = 0.21$	$F(72,107)_{0.01}$=0.60
$m_k = 0,\ k = 1,2,3$	$\frac{2.04 \cdot 10^4}{1.31 \cdot 10^2} = 156$	$F(2,72)_{0.99}$ =4.91
$s_l = 0,\ l = 1,2,3$	$\frac{3.16 \cdot 10^3}{1.31 \cdot 10^2} = 24.1$	$F(2,72)_{0.99}$ =4.91
$ms_{kl} = 0,\ k = 1,2,3,\ l = 1,2,3$	$\frac{5.21 \cdot 10^2}{1.31 \cdot 10^2} = 3.97$	$F(4,72)_{0.99}$ =3.59
$i(s)_{j(l)} = 0,\ j = 1,2,3,4,\ l = 1,2,3$	$\frac{3.89 \cdot 10^2}{1.31 \cdot 10^2} = 2.96$	$F(9,72)_{0.99}$ =2.66
$mi(s)_{kj(l)} = 0,\ k = 1,2,3,$ $j = 1,2,3,4,\ l = 1,2,3$	$\frac{2.65 \cdot 10^2}{1.31 \cdot 10^2} = 2.02$	$F(18,72)_{0.98}$ =2.01

Table E.8: Tests based on Model (8.2) for the 2nd PC of the dataset with fungi and edge in one.

H_0	Test Size	F-fractile
$\sigma_R^2 = 0$	$\frac{5.20 \cdot 10^0}{2.29 \cdot 10^2} = 0.02$	$F(72,107)_{0.01}$=0.60
$m_k = 0,\ k = 1,2,3$	$\frac{1.72 \cdot 10^3}{2.18 \cdot 10^2} = 7.90$	$F(2,18)_{0.99}$ =6.01
$s_l = 0,\ l = 1,2,3$	$\frac{1.25 \cdot 10^3}{1.54 \cdot 10^2} = 8.15$	$F(2,9)_{0.99}$ =8.02
$ms_{kl} = 0,\ k = 1,2,3,\ l = 1,2,3$	$\frac{3.20 \cdot 10^3}{2.18 \cdot 10^2} = 14.7$	$F(4,18)_{0.99}$ =4.58
$\sigma_{I(s)}^2 = 0$	$\frac{1.54 \cdot 10^2}{2.18 \cdot 10^2} = 0.71$	$F(9,18)_{0.31}$ =0.71
$\sigma_{mI(s)}^2 = 0$	$\frac{2.18 \cdot 10^2}{5.20 \cdot 10^0} = 42.0$	$F(18,72)_{0.99}$ =2.20

Table E.9: Tests based on Model (8.3) for the 95th percentile of difference between 1st and 11th spectra of the dataset with fungi and edge in one.

H_0	Test Size	F-fractile
$\sigma_R^2 = 0$	$\frac{4.82 \cdot 10^0}{1.84 \cdot 10^2} = 0.03$	$F(72,107)_{0.01}$=0.60
$m_k = 0,\ k = 1,2,3$	$\frac{1.46 \cdot 10^3}{3.13 \cdot 10^2} = 4.68$	$F(2,18)_{0.97}$ =4.29
$s_l = 0,\ l = 1,2,3$	$\frac{5.59 \cdot 10^2}{2.58 \cdot 10^2} = 2.16$	$F(2,9)_{0.83}$ =2.17
$ms_{kl} = 0,\ k = 1,2,3,\ l = 1,2,3$	$\frac{1.83 \cdot 10^3}{3.13 \cdot 10^2} = 5.86$	$F(4,18)_{0.99}$ =4.58
$\sigma_{I(s)}^2 = 0$	$\frac{2.58 \cdot 10^2}{3.13 \cdot 10^2} = 0.83$	$F(9,18)_{0.48}$ =0.83
$\sigma_{mI(s)}^2 = 0$	$\frac{3.13 \cdot 10^2}{4.82 \cdot 10^0} = 64.8$	$F(18,72)_{0.99}$ =2.20

Table E.10: Tests based on Model (8.3) for the 10th percentile of multiplication between 1st and 12th spectra of the dataset with fungi and edge in one.

H_0	Test Size	F-fractile
$\sigma_R^2 = 0$	$\frac{1.90 \cdot 10^1}{1.48 \cdot 10^3} = 0.01$	$F(72,107)_{0.01}$=0.60
$m_k = 0,\ k = 1,2,3$	$\frac{5.48 \cdot 10^4}{5.84 \cdot 10^2} = 93.9$	$F(2,18)_{0.99}$ =6.01
$s_l = 0,\ l = 1,2,3$	$\frac{1.17 \cdot 10^4}{8.94 \cdot 10^2} = 13.1$	$F(2,9)_{0.99}$ =8.02
$ms_{kl} = 0,\ k = 1,2,3,\ l = 1,2,3$	$\frac{1.30 \cdot 10^3}{5.84 \cdot 10^2} = 2.22$	$F(4,18)_{0.90}$ =2.29
$\sigma_{I(s)}^2 = 0$	$\frac{8.94 \cdot 10^2}{5.84 \cdot 10^2} = 1.53$	$F(9,18)_{0.79}$ =1.53
$\sigma_{mI(s)}^2 = 0$	$\frac{5.84 \cdot 10^2}{1.90 \cdot 10^1} = 30.8$	$F(18,72)_{0.99}$ =2.20

Table E.11: Tests based on Model (8.3) for the 1st PC of the dataset with fungi and edge in one.

H_0	Test Size	F-fractile
$\sigma_R^2 = 0$	$\frac{1.31 \cdot 10^2}{6.26 \cdot 10^2} = 0.01$	$F(72,107)_{0.01}$=0.60
$m_k = 0,\ k = 1,2,3$	$\frac{2.04 \cdot 10^4}{2.65 \cdot 10^2} = 77.0$	$F(2,18)_{0.99}$ =6.01
$s_l = 0,\ l = 1,2,3$	$\frac{3.16 \cdot 10^3}{3.89 \cdot 10^2} = 8.13$	$F(2,9)_{0.99}$ =8.02
$ms_{kl} = 0,\ k = 1,2,3,\ l = 1,2,3$	$\frac{5.21 \cdot 10^2}{2.65 \cdot 10^2} = 1.97$	$F(4,18)_{0.86}$ =1.99
$\sigma_{I(s)}^2 = 0$	$\frac{3.89 \cdot 10^2}{2.65 \cdot 10^2} = 1.47$	$F(9,18)_{0.77}$ =1.48
$\sigma_{mI(s)}^2 = 0$	$\frac{2.65 \cdot 10^2}{1.31 \cdot 10^2} = 2.02$	$F(18,72)_{0.98}$ =2.01

Table E.12: Tests based on Model (8.3) for the 2nd PC of the dataset with fungi and edge in one.

E.2.3 Tests for Multivariate ANOVA

H_0	U	q	r	F	F-fractile
$\boldsymbol{\sigma}_R^2 = \mathbf{0}$,	0.919	72	107	0.06	$F(144, 212)_{0.01}$=0.70
$\boldsymbol{m}_k = \mathbf{0},\ k = 1, 2, 3$	0.0141	2	72	264	$F(4, 142)_{0.99}$ =3.45
$\boldsymbol{s}_l = \mathbf{0},\ l = 1, 2, 3$	0.0425	2	72	137	$F(4, 142)_{0.99}$ =3.45
$\boldsymbol{ms}_{kl} = \mathbf{0}$, $k = 1, 2, 3,\ l = 1, 2, 3$	0.00852	4	72	175	$F(8, 142)_{0.99}$ =2.64
$\boldsymbol{i(s)}_{zj(l)} = \mathbf{0}$, $j = 1, 2, 3, 4,\ l = 1, 2, 3$	0.0824	9	72	19.6	$F(18, 142)_{0.99}$ =2.06
$\boldsymbol{mi(s)}_{zkj(l)} = \mathbf{0},\ k = 1, 2, 3,$ $j = 1, 2, 3, 4,\ l = 1, 2, 3$	0.0312	18	72	18.4	$F(36, 142)_{0.99}$ =1.77

Table E.13: Tests based on the multivariate version of Model (8.2) and the variables; 99th percentile of difference between 4th and 6th spectra and 10th percentile of difference between 1st and 12th spectra. The correlation between the variables is $\rho = 0.40$. DA1 & EN2.

H_0	U	q	r	F	F-fractile
$\boldsymbol{\sigma}_R^2 = \mathbf{0}$,	0.967	72	107	0.02	$F(144, 212)_{0.01}$=0.70
$\boldsymbol{m}_k = \mathbf{0},\ k = 1, 2, 3$	0.0141	2	72	264	$F(4, 142)_{0.99}$ =3.45
$\boldsymbol{s}_l = \mathbf{0},\ l = 1, 2, 3$	0.0407	2	72	141	$F(4, 142)_{0.99}$ =3.45
$\boldsymbol{ms}_{kl} = \mathbf{0}$, $k = 1, 2, 3,\ l = 1, 2, 3$	0.00245	4	72	341	$F(8, 142)_{0.99}$ =2.64
$\boldsymbol{i}(\boldsymbol{s})_{zj(l)} = \mathbf{0}$, $j = 1, 2, 3, 4,\ l = 1, 2, 3$	0.101	9	72	1.69	$F(18, 142)_{0.95}$ =1.68
$\boldsymbol{mi}(\boldsymbol{s})_{zkj(l)} = \mathbf{0},\ k = 1, 2, 3,$ $j = 1, 2, 3, 4,\ l = 1, 2, 3$	0.0224	18	72	22.4	$F(36, 142)_{0.99}$ =1.77

Table E.14: Tests based on the multivariate version of Model (8.2) and the variables; 30th percentile of difference between 1st and 8th spectra and 10th percentile of difference between 1st and 12th spectra. The correlation between the variables is $\rho = 0.46$. DA2 & EN2.

H_0	U	q	r	F	F-fractile
$\boldsymbol{\sigma}_R^2 = \mathbf{0}$,	0.869	72	107	0.11	$F(144, 212)_{0.01}$=0.70
$\boldsymbol{m}_k = \mathbf{0},\ k = 1, 2, 3$	0.00199	2	72	761	$F(4, 142)_{0.99}$ =3.45
$\boldsymbol{s}_l = \mathbf{0},\ l = 1, 2, 3$	0.0469	2	72	128	$F(4, 142)_{0.99}$ =3.45
$\boldsymbol{ms}_{kl} = \mathbf{0}$, $k = 1, 2, 3,\ l = 1, 2, 3$	0.150	4	72	28.1	$F(8, 142)_{0.99}$ =2.64
$\boldsymbol{i}(\boldsymbol{s})_{zj(l)} = \mathbf{0}$, $j = 1, 2, 3, 4,\ l = 1, 2, 3$	0.0909	9	72	18.3	$F(18, 142)_{0.99}$ =2.06
$\boldsymbol{mi}(\boldsymbol{s})_{zkj(l)} = \mathbf{0},\ k = 1, 2, 3,$ $j = 1, 2, 3, 4,\ l = 1, 2, 3$	0.0645	18	72	11.6	$F(36, 142)_{0.99}$ =1.77

Table E.15: Tests based on the multivariate version of Model (8.2) and the first two PCs. For the features of the edge and fungi in one. The correlation bewtween the variables is $\rho = 0.00$. PC1 & PC2.

H_0	U	q	r	F	F-fractile
$\boldsymbol{\sigma}_R^2 = \mathbf{0}$,	0.919	72	107	0.06	$F(144, 212)_{0.01}$=0.70
$\boldsymbol{m}_k = \mathbf{0},\ k = 1, 2, 3$	0.0610	2	18	25.9	$F(4, 43)_{0.99}$ =3.93
$\boldsymbol{s}_l = \mathbf{0},\ l = 1, 2, 3$	0.0384	2	9	16.4	$F(4, 16)_{0.99}$ =4.77
$\boldsymbol{ms}_{kl} = \mathbf{0}$, $k = 1, 2, 3,\ l = 1, 2, 3$	0.0583	4	18	13.3	$F(8, 34)_{0.99}$ =3.09
$\boldsymbol{\sigma}_{i(s)}^2 = \mathbf{0}$	0.469	9	18	0.87	$F(18, 34)_{0.39}$ =0.87
$\boldsymbol{\sigma}_{mi(s)}^2 = \mathbf{0}$	0.0312	18	72	18.4	$F(36, 142)_{0.99}$ =1.77

Table E.16: Tests based on the multivariate version of Model (8.3) and the variables; 99th percentile of difference between 4th and 6th spectra and 10th percentile of difference between 1st and 12th spectra. The correlation between the variables is $\rho = 0.40$.

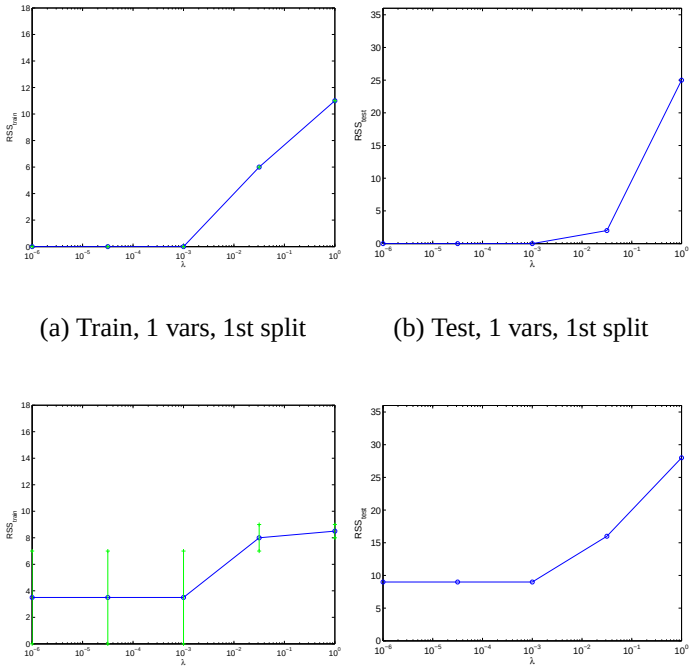
H_0	U	q	r	F	F-fractile
$\boldsymbol{\sigma}_R^2 = \mathbf{0}$,	0.967	72	107	0.02	$F(144, 212)_{0.01}$=0.70
$\boldsymbol{m}_k = \mathbf{0},\ k = 1, 2, 3$	0.0427	2	18	32.6	$F(4, 43)_{0.99}$ =3.93
$\boldsymbol{s}_l = \mathbf{0},\ l = 1, 2, 3$	0.0100	2	9	35.9	$F(4, 16)_{0.99}$ =4.77
$\boldsymbol{ms}_{kl} = \mathbf{0}$, $k = 1, 2, 3,\ l = 1, 2, 3$	0.0335	4	18	19.0	$F(8, 34)_{0.99}$ =3.09
$\boldsymbol{\sigma}_{i(s)}^2 = \mathbf{0}$	0.625	9	18	0.50	$F(18, 34)_{0.07}$ =0.52
$\boldsymbol{\sigma}_{mi(s)}^2 = \mathbf{0}$	0.0224	18	72	22.4	$F(36, 142)_{0.99}$ =1.77

Table E.17: Tests based on the multivariate version of Model (8.3) and the variables; 30th percentile of difference between 1st and 8th spectra and 10th percentile of difference between 1st and 12th spectra. The correlation between the variables is $\rho = 0.46$.

H_0	U	q	r	F	F-fractile
$\boldsymbol{\sigma}_R^2 = \mathbf{0}$,	0.869	72	107	0.11	$F(144, 212)_{0.01}$=0.70
$\boldsymbol{m}_k = \mathbf{0},\ k = 1, 2, 3$	0.00928	2	18	79.7	$F(4, 34)_{0.99}$ =3.93
$\boldsymbol{s}_l = \mathbf{0},\ l = 1, 2, 3$	0.154	2	9	6.19	$F(4, 16)_{0.99}$ =4.77
$\boldsymbol{ms}_{kl} = \mathbf{0}$, $k = 1, 2, 3,\ l = 1, 2, 3$	0.460	4	18	2.02	$F(8, 34)_{0.92}$ =1.98
$\boldsymbol{\sigma}_{i(s)}^2 = \mathbf{0}$	0.323	9	18	1.43	$F(18, 34)_{0.82}$ =1.43
$\boldsymbol{\sigma}_{mi(s)}^2 = \mathbf{0}$	0.0645	18	72	11.6	$F(36, 142)_{0.99}$ =1.77

Table E.18: Tests based on the multivariate version of Model (8.3) and the first two PCs. For the features of the edge and fungi in one. The correlation between the variables is $\rho = 0.00$.

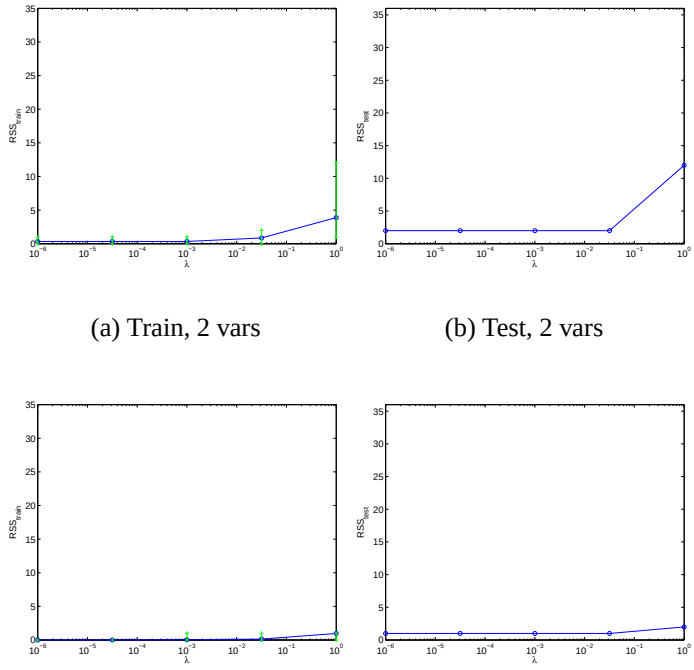
E.3 LARS-EN with dummy variables



(a) Train, 1 vars, 1st split (b) Test, 1 vars, 1st split

(c) Train, 1 vars, 2nd split (d) Test, 1 vars, 2nd split

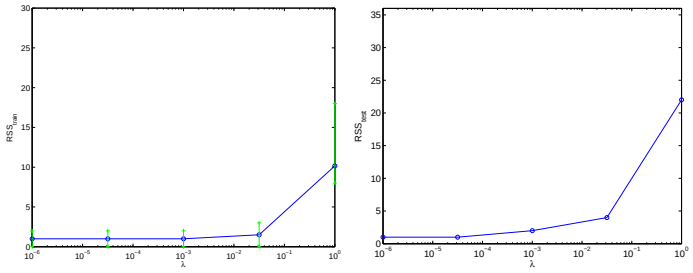
Figure E.3: Misclassifications 2-fold CV on YES medium. Two partitionings of data has been used.



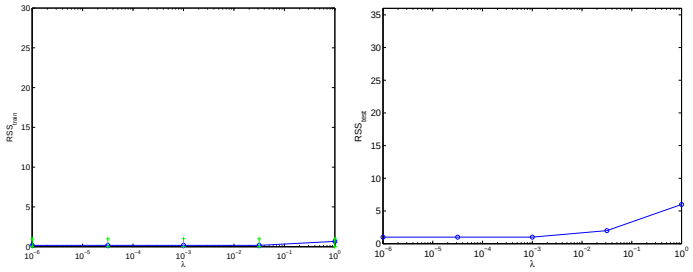
(a) Train, 2 vars (b) Test, 2 vars

(c) Train, 9 vars (d) Test, 9 vars

Figure E.4: Misclassifications for leave-one-out CV on OAT medium.

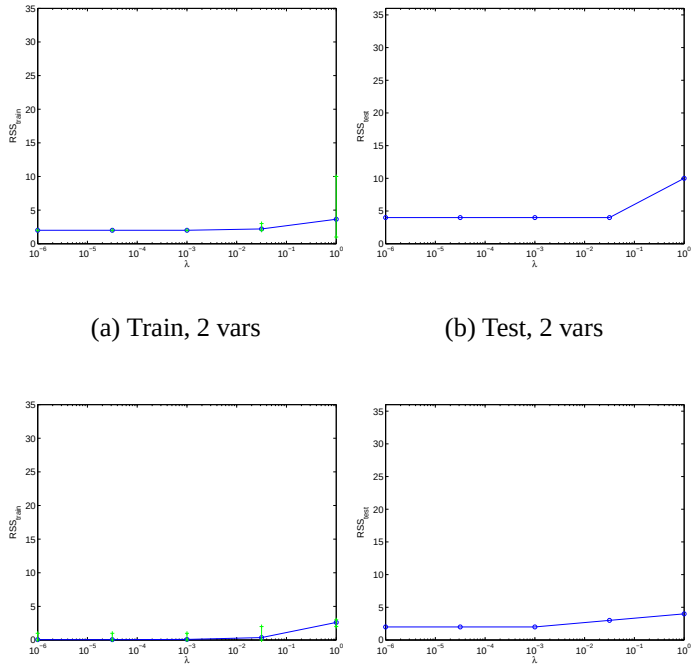


(a) Train, 2 vars (b) Test, 2 vars



(c) Train, 9 vars (d) Test, 9 vars

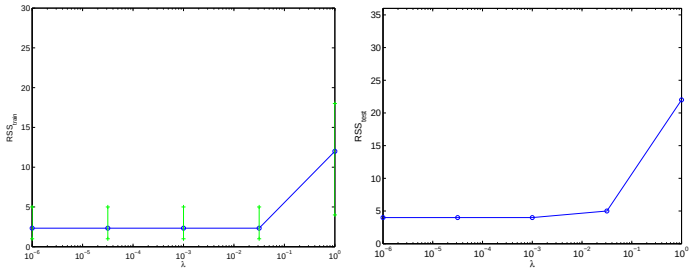
Figure E.5: Misclassifications for 6-fold CV on OAT medium.



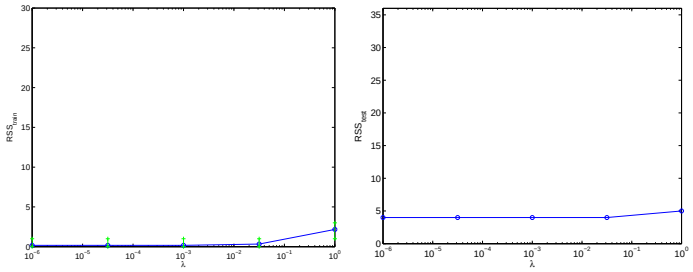
(a) Train, 2 vars (b) Test, 2 vars

(c) Train, 9 vars (d) Test, 9 vars

Figure E.6: Misclassifications for leave-one-out CV on CYA medium.

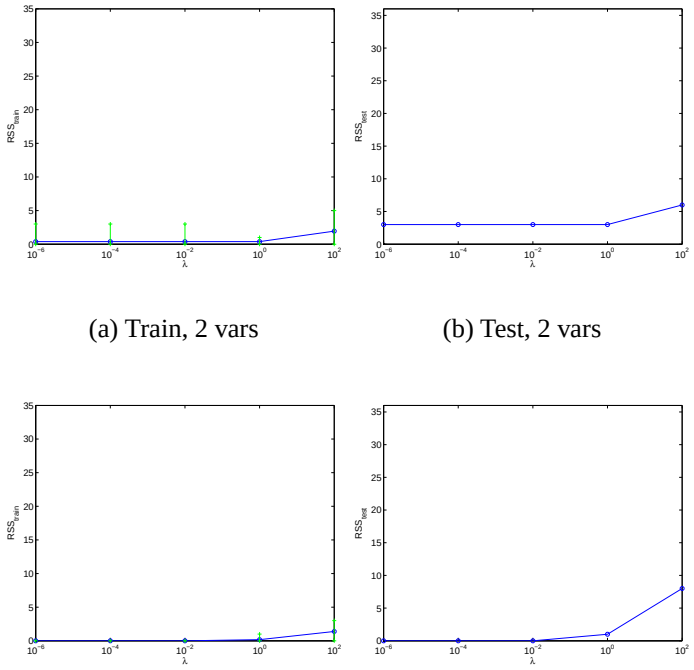


(a) Train, 2 vars (b) Test, 2 vars



(c) Train, 9 vars (d) Test, 9 vars

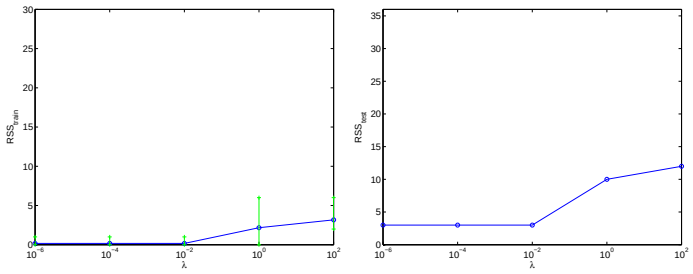
Figure E.7: Misclassifications for 6-fold CV on CYA medium.



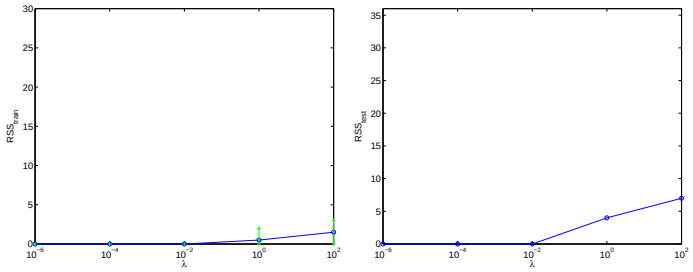
(a) Train, 2 vars (b) Test, 2 vars

(c) Train, 3 vars (d) Test, 3 vars

Figure E.8: Misclassifications for leave-one-out CV on YES medium. Dataset of three spectral bands closest to RGB.

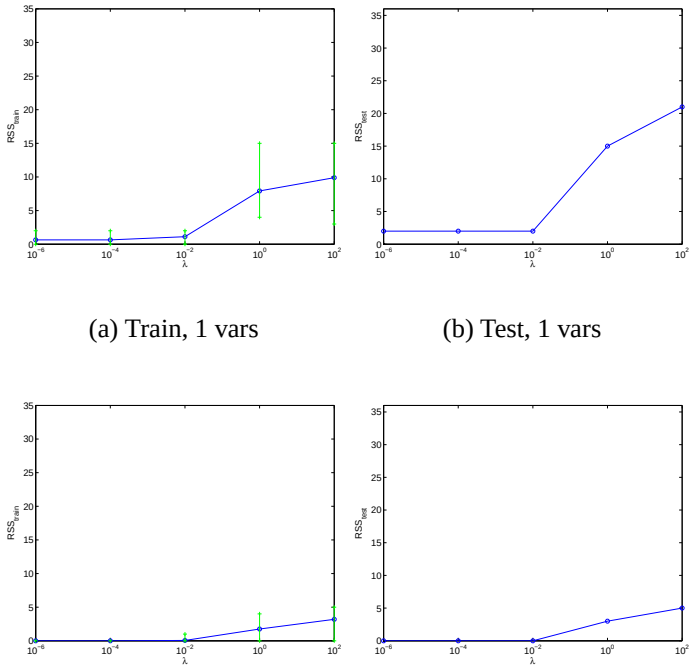


(a) Train, 3 vars (b) Test, 3 vars



(c) Train, 4 vars (d) Test, 4 vars

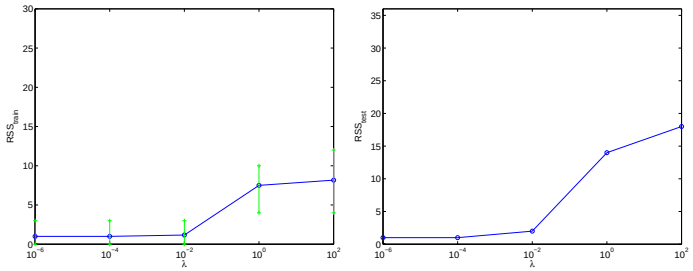
Figure E.9: Misclassifications for 6-fold CV on YES medium. Dataset of three spectral bands closest to RGB.



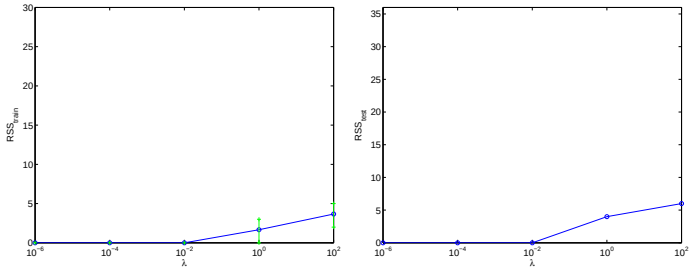
(a) Train, 1 vars (b) Test, 1 vars

(c) Train, 2 vars (d) Test, 2 vars

Figure E.10: Misclassifications for leave-one-out CV on YES medium. Dataset of the linear combinations of the 10 visual spectra to represent RGB.

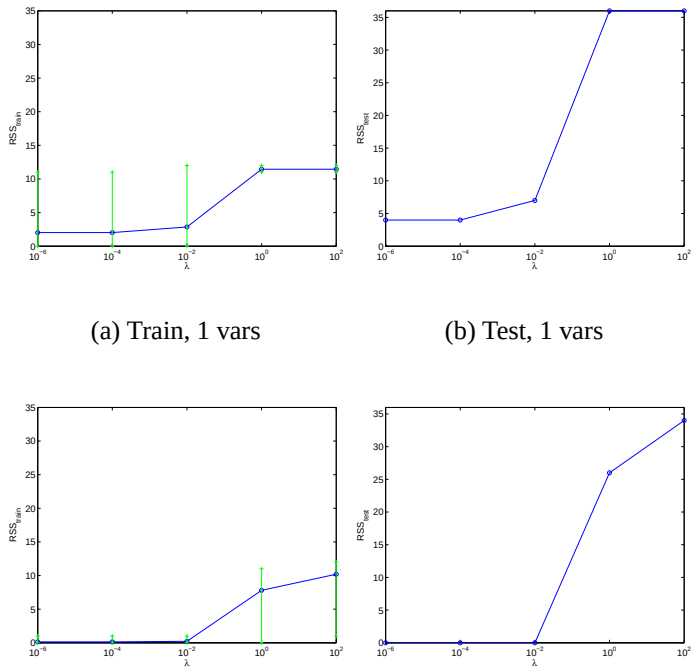


(a) Train, 1 vars (b) Test, 1 vars



(c) Train, 2 vars (d) Test, 2 vars

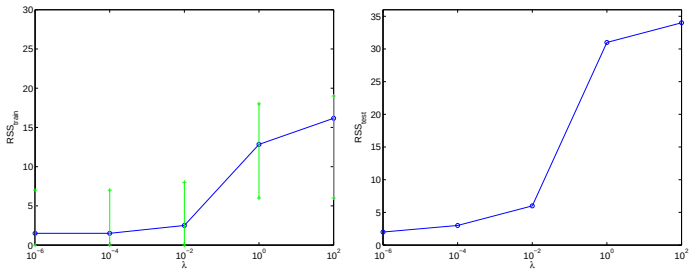
Figure E.11: Misclassifications for 6-fold CV on YES medium. Dataset of the linear combinations of the 10 visual spectra to represent RGB.



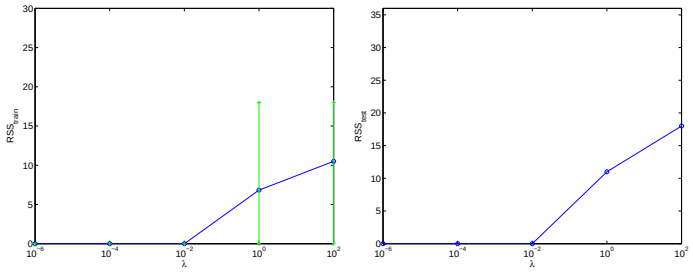
(a) Train, 1 vars (b) Test, 1 vars

(c) Train, 2 vars (d) Test, 2 vars

Figure E.12: Misclassifications for leave-one-out CV on YES medium. Dataset of the fungi and edge seprarate.

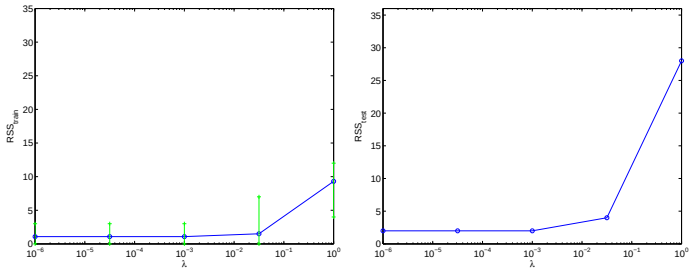


(a) Train, 2 vars (b) Test, 2 vars



(c) Train, 3 vars (d) Test, 3 vars

Figure E.13: Misclassifications for 6-fold CV on YES medium. Dataset of the fungi and edge seprarate.



(a) Train, 2 vars (b) Test, 2 vars

Figure E.14: Misclassifications for leave-one-out CV on YES medium. Dataset of the geometrical features.