From the Institute for Neuro- and Bioinformatics
of the University of Lübeck
Director: Prof. Dr. Thomas Martinetz

Classification for Biomarker Search - Aspects of experimental design and a method extension

Dissertation
for Fulfillment of
Requirements
for the Doctoral Degree
of the University of Lübeck

from the Department of Computer Sciences/Engineering

Submitted by

Anna Telaar
from Bocholt

Lübeck 2012

First referee:	Prof. Dr. Thomas Martinetz
Second referee:	Prof. Dr. Joachim Selbig
Date of oral examination:	26.10.2012
Approved for printing:	Lübeck, 1.11.2012

To my Family.

Zusammenfassung

Biomarker ermöglichen die Klassifizierung von Patienten in vordefinierte Gruppen und können somit Ärzten als Unterstüzung bei der Diagnose und der Therapieauswahl dienen. Im folgenden werden Biomarker als eine Menge von Variablen verstanden, die als Prediktoren eine zuverlässige und robuste Entscheidungsregel generieren, wobei die Anzahl der Variablen möglichst klein ist. Diese Arbeit umfasst im wesentlichen drei Hauptteile: Zunächst wird ein Biomarker mit statistischen Lernverfahren bestimmt am Beispiel eines Tuberkulosedatensatzes. Der Mittelteil befasst sich mit dem Einfluss der Versuchsplanung auf die Biomarkersuche im Fall von gepoolten Proben. Abschließend wird eine Erweiterung der Methode Powered Partial Least Squares Discriminant Analysis (PPLS-DA) vorgeschlagen, um die Fehlerrate der Klassifikation noch weiter zu verringern.

Im ersten Teil wird auf der Basis von Metabolitprofilen ein Biomarker bestimmt, der zwischen Tuberkulose Patienten und infizierten, aber gesunden Personen unterscheidet. In diesem speziellen Fall werden mögliche Klassifizierungsmethoden verglichen, eine Rangliste der Metaboliten erstellt und eine Menge von Metaboliten als Biomarker-Kandidat bestimmt. Hierfür werden die zwei statistischen Lernverfahren Random Forest (RF) and PPLS-DA eingesetzt. Es wird ein Kreuzvalidierungsansatz zur Biomarkersuche vorgestellt, der auf einer validierten Rangliste von Merkmalen basiert. Mit diesem Ansatz wird ein Biomarker bestehend aus 19 Metaboliten mit der Klassifikationsmethode Random Forest (RF) bestimmt. Im Vergleich dazu hat

die Klassifikationsmethode PPLS-DA einen Biomarker aus nur 8 Metaboliten gefunden, bei gleicher Fehlerrate der Klassifikation.

Im zweiten Teil wird die Bedeutung der Wahl des experimentellen Designs für die Biomarkersuche auf Basis von Genexpressionsprofilen untersucht. Dazu werden Pooling-Designs und Einzelproben-Designs in Klassifikationsstudien verglichen. Anhand der Ergebnisse einer umfangreichen Simulationsstudie wird der Einfluss von gepoolten Proben auf die Fehlerrate der Klassifikationsmethoden und die Wahl möglicher Biomarker dargestellt. Darüber hinaus werden experimentelle Daten nachträglich "künstlich" gepoolt, um experimentelle Datensätze in die Untersuchung mit einbeziehen zu können. Es kann die in der Literatur gegebene Empfehlung untermauert werden, dass ein Einzelproben-Design für Klassifikationsstudien vorzuziehen ist. In Abhängigkeit vom Genexpressions-Muster, dem Anteil von informativen Genen für die Klassifikation, der Poolgröße, der technischen Variation und der Klassifikationsmethode wird der Einfluss eines Pooling-Designs auf die Biomarkersuche dargestellt. Es wird deutlich, dass die Klassifikationmethoden PPLS-DA, PLS-DA und RF am wenigsten anfällig gegenüber Pooling bei der Biomarkersuche sind.

Die Klassifikationsmethode PPLS-DA hat sich in dieser Arbeit als besonders vorteilhaft für die Biomarkersuche erwiesen, aufgrund dessen wird diese im dritten Teil weiter entwickelt. Die ursprüngliche PPLS-DA bestimmt den Power-Parameter über die Maximierung der Korrelation zwischen der Komponente und der Gruppenzugehörigkeit. Es wird eine Erweiterung vorgestellt, die den Power-Parameter in einem Optimierungsverfahren bestimmt, mit dem Ziel die Fehlerrate der LDA im Klassifikationsschritt zu minimieren. Hierfür werden vier mögliche Varianten vorgeschlagen. Die Fehlerrate wird untersucht für die Erweiterungen im Vergleich zur Fehlerrate der ursprünglichen PPLS-DA und der von PLS-DA. Hierbei wird die Datensatz-Struktur berücksichtigt. Die vorgeschlagene Erweiterung von PPLS-DA bildet einen Beitrag zur methodischen Entwicklung in der Biomarkersuche.

Acknowledgements

During the time I have been working on this thesis, many people supported me scientifically and personally. Without them, this thesis would not have been possible.

At first I thank Dirk Repsilber for his great support and this engagement during the last years. I am very grateful to Gerd Nürnberg for his tireless efforts, the willingness to share his experience and moral support. I am very pleased to have had the occasion to work with him.

Sincerely, I thank Thomas Martinetz, for his time and the very good cooperation.

I thank Karsten Schlettwein for his willingness and persistence in supporting my IT needs. Many thanks go to Norbert Poschadel, Ulrike Borchhardt, Daisy Zimmer, Beate Garske and the colleges of the research unit Genetic and Biometry of the Leibniz Institute for Farm Animal Biology for their helpful comments. I thank the Max Planck Institute for Infection Biology Berlin especially the director Stefan H.E. Kaufmann and January Weiner for the very good cooperation and for the financial support. Many thanks also go to the biostatistic group at the Norwegian University of Life Sciences Aas and especially to Kristian Hovde Liland for some very interesting weeks in Norway, which got me moving forward.

Without my family and Henrik, it were not possible to write this thesis, I am very glad to have them. Thank you very much.

Contents

When you have excluded the impossible,
whatever remains, however improbable,
must be the truth.

Sherlock Holmes

1 Introduction

Frequently the visit at the medical doctor includes the collection of a blood or urine sample to confirm a diagnosis. For example, is there a patient suspected of having diabetes mellitus, a blood sample might be taken to confirm the diagnosis. The concentration of glycated hemoglobin (denoted by HbA_{1c}) is used as a feature (variable), that classifies patients into two groups, a patient with a concentration measured equal or over 6.5% is classified to the group of patients with diabetes mellitus, patients with a concentration between 5.7 and 6.4% belong to the group with increased diabetes risk (Deutsche Diabetes Gesellschaft, 2010). The HbA_{1c}-value is a biomarker for diabetes mellitus. Thus, a biomarker enables the discrimination between groups using a decision rule. The HbA_{1c}-value is an example of a biomarker already established in practice. A particular advantage of this value for the diagnosis of diabetes is that only one blood sample is needed; the determination is independent of time of the day and does not require a fasting patient (Deutsche Diabetes Gesellschaft, 2010). Furthermore the HbA_{1c} has an outstanding status in practice for the blood glucose control because it is a measure of the average blood glucose concentration of the last 8 weeks. Therefore it allows a statement of the blood glucose concentration over a longer time, which is very important for the treatment of diabetics. If the HbA_{1c} value lies within a given reference range, the patient is well treated, otherwise the patient needs to be re-stabilized.

A further example for the application of a biomarker is a pregnancy test (Bracht, 2009). Based on the concentration of special hormones measured in a blood sample

or a urine sample, an assessment is made whether a woman is pregnant or not. Allergy tests (like a skin prick test) are also based on biomarkers (Boot *et al.*, 2008). The skin is incised on the forearm and brought into contact with various allergens. The size of possible welts on the skin is measured and indicates the susceptibility to an allergen.

Biomarkers such as cholesterol and blood fat level, are often non-specific and ambiguous, therefore molecular biomarkers such as gene expression and metabolite concentration are preferred regarding their validity for the classification result (Jarasch, 2011). Because of this reason, this thesis focuses on molecular biomarkers.

The search for a biomarker starts with the goal of finding a single feature which is easily measurable and allows a robust classification with respect to outer circumstances. It turns out that especially for complex disease pictures and in the case of early diagnosis, such a single feature cannot be found. For multivariate diseases like cancer, intuitively a biomarker is needed which bases on a combinations of features. This combination of features results in more valid and meaningful decisions than a single feature (Etzioni *et al.*, 2003; Pfeiffer and Bur, 2008). Therefore a biomarker can be considered as a set of features which allows decisions about group affiliation with a low number of wrongly classified samples. However the number of features should be small, so that the practical use (measurement of the features) is feasible. In this thesis, this important aspect is taken into account and special focus is put on the assembly of features to a biomarker.

In the last ten years the search for biomarkers has increased dramatically. Pharmaceutical research companies and public health authorities show an immense interest in biomarker search (Gudenus and Granzer, 2010). For clinical management decisions as in drug-development, risk assessment, diagnostic testing and treatment selection towards individualized medicine, there is an urgent need for biomarkers which are easy to measure and can replace invasive or expensive methods (Simon *et al.*, 2003).

The starting point of biomarker search are patients for which the group member-

ships are known. First (to stay in the previous example), blood samples are collected from these patients and the blood components (features) are measured. The goal is to classify persons according to a feature or a set of features. In the diabetes example, patients with diabetes mellitus have a clearly higher HbA_{1c}-value than healthy persons. Unfortunately, the search for biomarkers is not as simple like the example suggested and statistical methods are required to detect biomarker candidates. Statistical learning methods are widely and successfully applied to search for biomarkers (Moler *et al.*, 2000). With the help of importance values (measuring the importance of a feature for the discrimination between groups) it is possible to select features as possible biomarkers. Method development for biomarker search including multivariate approaches is a major challenge to develop biomarkers also for complex diseases. Nevertheless, the basis of each biomarker search is built by choosing an appropriate design (Kerr, 2003). In the design choice, however, not only statistical considerations are integrated, even outer circumstances play a role. For example, financial resources often limit the cost-intensive sample preparation for microarray studies and lead to a lower sample size. If more samples are collected than can be measured, mixing the samples (pooling) is a solution to increase the statistical power for detection of differentially expressed genes. Also pooling is applied if not enough cDNA is available for the hybridization step (e.g. for the antennae of bees). While biologists often apply pooling in microarray experiments, Kerr (2003) advised against choosing a pooling design for classification studies like biomarker searches, because the inference to the individual level suffers. This raises the question of how strong the consequences of pooling are on the results of biomarker search. Which statistical learning methods are most robust against pooling considering the performance and the selection of biomarkers? Statisticians are confronted with these fundamental questions.

In this thesis, different statistical aspects including design and choice of method are analyzed, which are important for the search of biomarker candidates.

This yields the following structure: In Chapter 2 statistical basics for the biomarker search are presented. The existing definition of a biomarker is given, and statistical aspects of a biomarker search are described. Also the statistical learning methods investigated are briefly explained.

To illustrate the biomarker identification process, an example of a biomarker search on the basis of metabolite profiles is given in Chapter 3. A biomarker which allows the discrimination between tuberculosis patients and infected but healthy persons is determined. The statistical learning method random forest is applied to detect a biomarker in this practical example. To verify how strong this biomarker depends on the method used, additionally a biomarker is identified with respect to Powered Partial Least Squares Discriminant Analysis (PPLS-DA). The origin of this Chapter based on work as a co-author for Weiner *et al.* (2011). In all following Chapters the focus is on gene expression data, because these features have often proven to be a good biomarker, and are particularly promising (Jarasch, 2011).

Chapter 4 describes the simulation of gene expression data and the publicly available gene expression data sets used in this thesis. The main part of this thesis studies the influence of the design choice in terms of pooled samples on the biomarker search presented in Chapter 5. A simulation study builds a detailed analysis of pooling with respect to linear and non-linear separable data. Biomarker candidates identified by pooling designs are compared to corresponding candidates for single sample designs to verify dependence of proposed biomarker candidates on the underlying design. Also experimental data sets are considered to underline the results. Chapter 5 bases on the publications Telaar *et al.* (2010),Telaar *et al.* (2012a).

In Chapter 3 PPLS-DA identifies a clearly lower number of features as biomarker than random forest while the number of falsely classified samples is similar. Also in Chapter 5, PPLS-DA comes out as applicable method for pooling designs, because it is most robust against pooling regarding the prediction performance and the important features for the classification. Furthermore the basic approach of this method (to combine original features to new features) seems to be advantageous.

Therefore in Chapter 6, PPLS-DA is extended to reduce the prediction error by modifying the original optimization problem. The work described in this Chapter shows an improvement of a promising method for biomarker search. A manuscript summarizing the presented findings has been submitted by Telaar *et al.* (2012b).

2 Statistical background for biomarker search

This chapter describes the background of a biomarker search. Starting with the official biomarker definition and a more theoretical definition in the context of this thesis, statistical aspects of the biomarker search are introduced. A representation of the biomarker search as a classification problem is then explained. Thereafter, selected statistical learning methods which are applied in this thesis to detect biomarkers are briefly introduced.

2.1 Basics of biomarker

2.1.1 Definition of a biomarker

The Biomarkers Definition Workgroup (2001) defines a biomarker as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention".

In the context of this thesis: A biomarker is defined as a set of features $\mathfrak{M} = \{M_1, \dots M_b\} \neq \emptyset$ (with at least one feature, i.g. $b \geq 1$) which enables a unique assignment/mapping of an unknown object to an unique predefined group. There are two main requirements:

a) the cardinality of the set $\mathfrak{M}$ ($|\mathfrak{M}| = b$) should be adequately small while

b) the assignment (classification rule) should be as accurate as possible.

Condition b) means that the number of wrongly classified objects should be as small as possible. For the first condition, a small number of selected features, Liu *et al.* (2002) state three reasons (in the mentioned publication $|\mathfrak{M}| = 20$). The first reason is, that medical doctors and biologists prefer a small number of features because otherwise the examining is not practicable. Secondly, a large number of features can lead to very time-consuming calculations. Thirdly, the inclusion of additional features does not lead to an improved prediction. Furthermore, for practical users of a biomarker the interpretability of the features building the biomarker is often important (Liggett *et al.*, 2004).

2.1.2 Biomarker search - statistical point of view

Before a biomarker is used, many developmental steps have been made (Feng *et al.*, 2004). In this thesis, the focus is on the first three steps of biomarker search. Baumgartner *et al.* (2011) describes these steps more in detail for metabolomic biomarkers. Main aspects of these steps can also be adapted to genomic or proteomic biomarker searches:

- definition of the hypothesis
- choice of study design and realization of the study
- sample preparation, analyses and biomarker identification.

A clear, pre-defined hypothesis is the beginning of every biomarker search. Next, a proper experimental design is necessary to allow derivation of valid statistical conclusions from the data (Kell, 2007; Kerr, 2003; Allison *et al.*, 2006; Yang and Speed, 2003). A good design minimizes the influence of disturbance factors like age, gender and additional diseases. The mentioned influencing quantities contribute to

the variation between the persons/samples, called biological variation. This kind of variation can be reduced by sample pooling (mixture of samples) (Allison *et al.*, 2006). Moreover pooling lowers costs, if financial constrictions play a role in the design choice.

After design selection, performing and evaluating the study yields a biomarker candidate. Whether or not it really becomes a biomarker is decided in further steps Baumgartner *et al.* (2011) (independent validation of the biomarker). In this thesis, a biomarker candidate is simply called a biomarker, having in mind that the biomarker has to be verified.

Case-control studies are frequently realized for biomarker search (Baumgartner *et al.*, 2011). This means, one group of persons which show the clinical indication (cases) and another group of persons without the clinical indication (controls) are investigated in the study. Often biomarker search is part of the analysis of clinical studies on unrelated end points, where additional molecular data have been measured. These studies can be formulated as a classification problem.

Description of a classification problem

The basis of a classification problem is a data matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_g] \in \mathbb{R}^{n \times g}$ and a response vector $\mathbf{y}$. The data matrix has n rows for the n samples/objects measured and g columns for the g features investigated, these features are also called predictors. The k^{th} column of $\mathbf{X}$ is denoted by $\mathbf{x}_k$. The vector $\mathbf{y} = [y_1, \ldots, y_n]^t \in \mathcal{G}^n$ contains the group memberships for each object. The label of objects $\{1, \ldots, n\}$ can be given as discrete variables or symbolics, here $\mathcal{G}$ is the set containing all labels $y_i, i = 1, \ldots, n$.

Each sample belongs to a unique group ν and each group has a sample size of n_ν. In this thesis, only two groups are considered and therewith $n = n_1 + n_2$. The prior probability of group 1 is denoted by π_1 and for group 2 by π_2, with $\pi_1 + \pi_2 = 1$. The

group information can also be given in form of a dummy coded matrix $\mathbf{Y} = (y_{i\nu})_{i\nu}$, $(i = 1 \ldots, n,\ \nu = 1, 2)$ as follows: The entry $y_{i\nu}$ equals 1 if sample i belongs to groups ν, otherwise the entry equals 0, $(i = 1, \ldots, n$ and $\nu = 1, 2)$.

The goal of a classification study is to determine a function $C(\mathbf{x}) : \mathbb{R}^g \longrightarrow \mathcal{G}$, $\mathbf{x} \mapsto y$ which assigns a unique group coded by the group labels with the greatest possible accuracy to each object given as a feature vector $\mathbf{x}$ (Indahl *et al.*, 2007).
Now the biomarker search can be summarized in the following way: the aim is to find a set $\mathfrak{M} = \{M_1, \ldots M_b\}$ (the biomarker) which features can be used as predictors instead of all g features with $1 \leq b < g$ for a classification function $C_{\mathfrak{M}} : \mathbb{R}^b \longrightarrow \mathcal{G}$ with good performance.

For the determination of function C, it is necessary to "learn" the properties of the data (i.e. data structure, distribution) to classify new independent samples to a predefined group. This can be done by statistical learning methods (presented in Section 2.2) which base on two steps, the training step and the test step. A classification rule/function C is built on the information of a partial data set $\mathbf{X}_{train}$ (Hastie *et al.*, 2009). The corresponding group labels are denoted by $\mathbf{y}_{train}$. The set $\{\mathbf{X}_{train}, \mathbf{y}_{train}\}$ is called the training set. In this thesis, the objects of the training set are randomly chosen with a ratio of $\phi_{\mathbf{X}}$ on the n samples of the whole data set. This procedure relates to the group size, since $\phi_{\mathbf{X}} \cdot n_\nu$ samples of group ν, $\nu = 1, 2$ contribute to the training set. The remaining samples build the test set $\{\mathbf{X}_{test}, \mathbf{y}_{test}\}$ with $\mathbf{X}_{train} \cap \mathbf{X}_{test} = \emptyset$ and a sample size per group of $(1 - \phi_{\mathbf{X}}) \cdot n_\nu$. Therewith the n objects are partitioned into a set containing objects for the training step and objects for the test step. The samples of the test sets are used to verify the learned classification function. These samples are assigned to a group label according to the learned mapping C (test step). Because the true group memberships are known, the prediction error (PE, the proportion of wrongly classified objects) can be calculated.

Measuring biomarker performance

A further aspect of biomarker search is the evaluation of the biomarker. This is already indicated in requirement b) of a biomarker in Section 2.1.1. A biomarker often is assessed with respect to its prediction accuracy (the proportion of correctly classified objects) or often by its PE. For the determination of these measures, it is important to use new independent samples not used in the training set (Lai *et al.*, 2006). The test set can be used or new independently measured samples. Therefore often so called cross-validation approaches are used to estimate the PE considering the whole data set $\mathbf{X}$ (Dudoit and Fridlyand, 2003). Several forms of cross-validation approaches exist (e.g. the classical leave-one-out cross-validation (Hastie *et al.*, 2009)). In this thesis the cross-validation approach is applied on the n objects by partitioning the data into training set and test set by the ratios ($\phi_{\mathbf{X}}$,1-$\phi_{\mathbf{X}}$). This partition can be repeated $r \in \mathbb{N}$ times. Let n_{train} denote the number of objects for the training set and n_{test} the corresponding number for the test set. With this kind of partition, there are $\binom{n}{n_{train}}$ possible combinations for the objects of the training set, in contrast to the leave-one-out cross-validation with only n repetitions. Let r denote the number of resampling steps of the training set and the test set. Calculating the PE for each test set, a vector $\mathbf{z}$ containing r PE values is obtained. This enables the estimation of the PE by the mean value and the corresponding 95% confidence interval. A confidence interval is calculated as follows: the upper bound of the confidence interval is $u(\mathbf{z}) = mean(\mathbf{z}) + 1.96 \cdot sd(\mathbf{z})/\sqrt{r}$. The lower bound is calculated likewise. If these confidence intervals overlap, no significant differences are reported, if they are disjunct, the corresponding PEs are reported as significantly different.

If parameters have to be optimized for classification methods, this is only done on the training set by a similar cross-validation procedure as described above. Random samples of the training set $\mathbf{X}_{train}$ are drawn with a ratio of $\phi_{\mathbf{X}_{train}}$ to create an inner training set, and the remaining samples build the inner test set analog to the

training and test set described above. This step is repeated r_{inner} times. This kind of cross-validation is called inner-cross-validation. The training set and the test set according to $\phi_{\mathbf{X}}$ are called outer training and outer test set.

Once a candidate biomarker is identified, it should be checked for reproducibility (Liggett *et al.*, 2004). It is important to verify the biomarker and investigate the robustness with respect to variation in the population. Therefore various statistical methods should be compared for the identification of biomarkers (Feng *et al.*, 2004).

2.2 Classification methods

In the previous Section 2.1.2, biomarker search is formulated as a classification problem. In the following, classification methods are introduced which have already shown great promise for finding biomarkers (Dettling and Buehlmann, 2003; Dudoit and Fridlyand, 2003; Fan *et al.*, 2011) and are applied throughout this thesis. The open source statistical programming language R (R Development Core Team, 2011) was used for implentations of these classification methods.

2.2.1 Linear discriminant analysis

The linear discriminant analysis (LDA) is a linear classification method which determines a linear decision boundary based on a Bayes classification. The probability for an object (determined by $\mathbf{x}$) belonging to group ν is denoted by $p_\nu(\mathbf{x})$ and is normally distributed with mean value $\mu_{(\nu)}$ and variance Σ_ν. The decision rule of the Bayes classification is now that the object with measured values $\mathbf{x}$ belongs to the group with the largest posteriori probability (Indahl *et al.*, 2007)

$$p(\nu|\mathbf{x}) = \frac{\pi_\nu p_\nu(\mathbf{x})}{p_1(\mathbf{x}) + p_2(\mathbf{x})}.$$

If no prior probabilities π_ν are known, the group proportions n_ν/n are used instead. The following descriptions are based on Huberty (1994) and Ripley (1997). If the groups have a common covariance structure $\Sigma_\nu = \Sigma$ for all ν, the linear discriminant function for group ν is defined as

$$\delta_\nu(\mathbf{x}) = \mathbf{x}^t\Sigma^{-1}\mu_{(\nu)} - \frac{1}{2}\mu_{(\nu)}^t\Sigma^{-1}\mu_{(\nu)} + log\,\pi_\nu.$$

The final classification rule is then

$$G(\mathbf{x}) = \arg\max_\nu \delta_\nu(\mathbf{x}),$$

the object with the feature values $\mathbf{x}$ is assigned to the group with the largest value of the discriminant function. In the case of two-groups, the difference between δ_1 and δ_2 is

$$\delta_1(\mathbf{x}) - \delta_2(\mathbf{x}) = \mathbf{x}^t\Sigma^{-1}(\mu_{(1)} - \mu_{(2)}) - \frac{1}{2}(\mu_{(1)}^t\Sigma^{-1}\mu_{(1)} - \mu_{(2)}^t\Sigma^{-1}\mu_{(2)}) + log(\frac{\pi_1}{\pi_2}),$$

therefore the following classification rule is achieved:

$$\delta_1(\mathbf{x}) - \delta_2(\mathbf{x}) \begin{cases} \geq 0 & \Rightarrow \mathbf{x} \mapsto \text{ group 1} \\ \text{otherwise} & \Rightarrow \mathbf{x} \mapsto \text{ group 2.} \end{cases}$$

The discriminant function δ_ν (2.2.1) can be rewritten as

$$\delta_\nu(\mathbf{x}) = \mathbf{b}_\nu^t\mathbf{x} + b_{\nu_0},$$

with $b_\nu = \mu_{(\nu)}\Sigma^{-1}$ and a constant $b_{\nu_0} = -\frac{1}{2}\mu_{(\nu)}^t\Sigma^{-1}\mu_{(\nu)} + log\pi_\nu$. The vector $\mathbf{b}_\nu$ contains the weights for each feature. Therewith the difference between the two discriminant functions δ_1 and δ_2 can also be described by

$$\delta_1(\mathbf{x}) - \delta_2(\mathbf{x}) = \mathbf{b}^t\mathbf{x} + b_0,$$

with $\mathbf{b} = (\mathbf{b}_1 - \mathbf{b}_2)$ and $b_0 = (b_{1_0} - b_{2_0})$. The classification function/rule of LDA is now

$$C_{\text{LDA}}(\mathbf{x}) := \delta_1(\mathbf{x}) - \delta_2(\mathbf{x}) : \mathbb{R}^g \to \mathcal{G}.$$

Geometrically $\delta_1(\mathbf{x}) - \delta_2(\mathbf{x}) = 0$ describes the decision boundary, a $(g-1)$-dimensional hyperplane (Bishop, 2006).
For such an LDA the R-function *lda* of the R-package `MASS` is applied and the proportions of the groups $\pi_\nu = n_\nu/n$ in the training set are used as prior probabilities.

One disadvantage of LDA is, that the required sample size for the group with the minimal sample size is $3 \cdot g$ (3 times the number of genes) (Huberty, 1994). Especially for gene expression data, this condition is not fulfilled, because often $n_\nu << 3 \cdot g$. Therefore, in this thesis the top ten features are first selected according to the p-value of the t-test (the t-test is described in Section 2.3.2). Here, the ten features with the smallest p-values are chosen; they are the predictors of the LDA. These ten features can be understand as possible biomarker candidates. How good they are as biomarker, can be evaluated by the PE of the LDA. The classification using LDA with features filtered according to the t-test is denoted by $t-$LDA.

2.2.2 Random forest

Random forest (RF) is described by its developer Breiman (2001) as follows: It grows "an ensemble of trees and lets them vote for the most popular class". Decision trees are built, each on a random sample chosen with replacement on the objects. Breiman (1996) called the remaining samples which are not considered for the creation of the tree the "out-of-bag" samples (OOB). A decision tree is built in the following steps: at each node, *mtry* of the g features are chosen. Among these features, the feature with the best split (the greatest prediction performance) at this node is selected. The trees are not pruned. In one random forest *mtry* is constant for each tree.

Breiman (2002) recommended choosing $mtry$ equal to the square root of the number of features ($\sqrt{g}$) and ≥ 1000 for the number of trees (ntree). Classifying new samples is done by dropping them down through all trees of the random forest. The group with the majority of votes depicts the predicted group membership according to the RF (Liaw and Wiener, 2002).

The classification rule of random forest is

$$C_{\mathrm{RF}}(\mathbf{x}) : \mathbb{R}^p \to \mathcal{G}.$$

with $\mathbf{x}$ belonging to group 1, if the majority of the trees classify $\mathbf{x}$ to group 1, otherwise $\mathbf{x}$ belongs to group 2.

For RF in this thesis the R package `randomForest` (Liaw and Wiener, 2002) is applied. For the experimental data, a FORTRAN code available from http://www.stat.berkeley.edu/~breiman/RandomForests/cc_software.htm is used with the same settings as for the R Code. The FORTRAN code is used because the experimental data sets have a large number of genes which lead to very time consuming calculations in R. The results produced using R and FORTRAN are similar. This was checked with simulated data.

2.2.3 Support vector machine

Support vector machines (SVM) were first proposed as a soft margin classifiers by Vapnik (1995), theory development started with works of Vapnik and Chervonenkis (1974) and Vapnik (1979). For a comprehensive introduction see Burges (1998). An SVM is based on classification according to a linear hyperplane like for LDA. For the introduction of SVM, the group memberships are assumed to be $y_i \in \mathcal{G} = \{1, -1\}$ for all samples $i = 1, \ldots, n$. The explanations are deduced from Burges (1998).

Linear SVM

The first case considered consists of two groups with data points linear separable by a hyperplane. Analogous to the approach of LDA, a hyperplane defined by $\mathbf{b}^t\mathbf{x} + b_0 = 0$ is sought. The question is how to determine the vector $\mathbf{b}$ and the constant b_0, if several solutions are possible. Now the hyperplane with the largest margin is selected, i.e. the hyperplane with the largest distance between the data points on the positive and negative side. The underlying minimization problem of the linear SVM is:

$$\begin{aligned} \min_{\mathbf{b},b_0} \quad & \tfrac{1}{2}\mathbf{b}^t\mathbf{b} \\ \text{subject to} \quad & y_i(\mathbf{b}^t\mathbf{x}_i + b_0) \geq 1. \end{aligned} \tag{2.1}$$

The data vectors which lie on the hyperplanes defining the margin $(\{\mathbf{x}_j \,|\mathbf{b}^t\mathbf{x}_j + b_0 = 1\} \cup \{\mathbf{x}_l \,|\mathbf{b}^t\mathbf{x}_l + b_0 = -1\})$ are called support vectors. If the data is non-separable, slack variables $\xi_i, i = 1, \ldots, n$, $\xi = [\xi_1, \ldots, \xi_n]$ are introduced to soften the conditions of the optimization problem (2.1). Additionally, a cost constant $\mathcal{C}$ is established to penalize the softness of the conditions. The resulting minimization problem is then

$$\begin{aligned} \min_{\mathbf{b},b_0,\xi} \quad & \tfrac{1}{2}\mathbf{b}^t\mathbf{b} + \mathcal{C}\textstyle\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & y_i(\mathbf{b}^t\mathbf{x}_i + b_0) \geq 1 - \xi_i \\ & \xi_i \geq 0. \end{aligned}$$

The sum $\sum_{i=1}^{n}\xi_i$ is used so that the accumulated violations of the conditions are minimal.

Therewith for linearly separable data, an optimization problem is given to determine a hyperplane allowing weakened conditions if necessary.

Nonlinear SVM

If the data are not linearly separable in the feature-space $\mathbb{R}^g$, the so called "kernel trick" (Aizerman *et al.*, 1964) is used to project the data in a higher dimensional space, where the data is linearly separable. The function φ maps the data in a higher dimensional Euclidean space $\mathcal{H}$, $\varphi : \mathbb{R}^g \to \mathcal{H}$. The kernel function $\mathcal{K}$ should have the following form $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \varphi(\mathbf{x}_i) \cdot \varphi(\mathbf{x}_j)$ such that only the dot product in $\mathcal{H}$ is used and φ does not need to be known. A simple example of a nonlinear pattern is shown in Figure 2.1(a). The red points illustrate the data points of group 1 and the blue points indicate samples of group 2 in $\mathbb{R}$. Figure 2.1(c) shows the data points projected to a higher dimensional space $\mathcal{H} = \mathbb{R}^2$ by the function $\varphi : \mathbb{R} \to \mathbb{R}^2$ with $\varphi(x_1) = [x_1, x_1^2]^t$. In the space $\mathbb{R}^2$ the samples are linearly separable by a hyperplane.

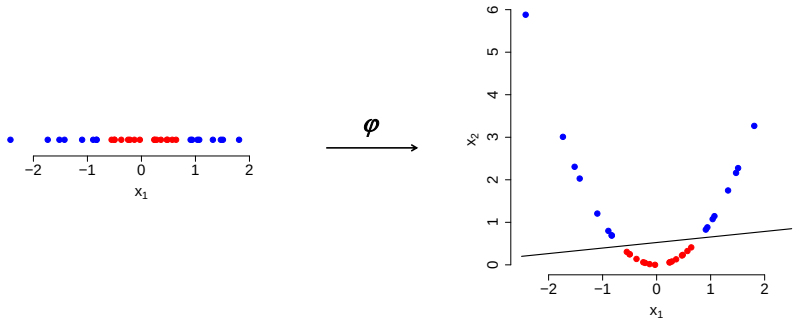


Figure 2.1: Nonlinear data points (red data points belong to group 1, blue data points to group 2) in $\mathbb{R}$ (panel left) and the same data points under the mapping φ in $\mathbb{R}^2$ (panel right) (now linear separable).

In this thesis, the following two basic kernels are included:

- linear: $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^t \mathbf{x}_j$

- radial basis function: $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = exp(-\tau ||\mathbf{x}_i - \mathbf{x}_j||^2)$, $\tau \geq 0$.

The underlying minimization problem of the nonlinear SVM is :

$$\begin{aligned} \min_{\mathbf{b},b_0,\xi} \quad & \tfrac{1}{2}\mathbf{b}^t\mathbf{b} + \mathcal{C}\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & y_i(\mathbf{b}^t\varphi(\mathbf{x}_i) + b_0) \geq 1 - \xi_i \\ & \xi_i \geq 0. \end{aligned}$$

If φ equals the identical function, this optimization problem corresponds to those of the linear SVM.
The classification rule developed by linear and nonlinear SVM can be summarized by

$$C_{\text{SVM}}(\mathbf{x}) : \mathbb{R}^g \to \mathcal{G}.$$

with $C_{\text{SVM}}(\mathbf{x}) := 1$ if $\mathbf{b}^t\varphi(\mathbf{x}) + b_0 \geq 1$ and $C_{\text{SVM}}(\mathbf{x}) := -1$ if $\mathbf{b}^t\varphi(\mathbf{x}) + b_0 < 1$.
The parameters $\mathcal{C}$ and τ (only for the radial basis function) have to be set by the user. A large value for $\mathcal{C}$ punishes the breach of the constraint more than a small one.

For solving the optimization problem in the linear and non-linear case, the optimization problem can be solved using Lagrange multipliers. Therewith the data points are given in form of dot products between vectors, and in the non-linear case, the function φ does not need to be known.

The R package e1071 (Dimitriadou *et al.*, 2009) is applied for SVM and the parameters $\mathcal{C}$ and τ are tuned within the interval $[2^{-5}, 2^4]$ and $[2^{-10}, 2^5]$, respectively, using the R-function tune.svm with a cross-validation of 10 steps. The intervals for tuning are chosen according to the suggestion of Dettling (2004). In this thesis, SVM with the linear kernel is denoted by SVML and SVM with the radial basic function is denoted by SVMR.

2.3 Dimension reduction

2.3.1 Feature Extraction

The next section mainly bases on Barker (2010), Barker and Rayens (2003), Indahl *et al.* (2007) and Nocairi *et al.* (2005). In this Section and without loss of generality, let the columns of the data matrix $\mathbf{X}$ be centered (otherwise this can be easily done by subtraction of the mean of the variable). A lot of features (genes) are often measured for only a few objects/patients especially in gene expression experiments. Classification methods like LDA cannot deliver good classification rules if the number of features g is much larger than the number of samples n. This problem is known in the literature under the "large p small n" problem, where p denotes the number of features (Hastie *et al.*, 2009). Therefore, dimension reduction techniques are often used before the final classification takes place. Feature extraction is a special form of dimension reduction, based on transformation to a lower dimensional subspace. The transformation can be linear or nonlinear. In this thesis only linear feature extraction techniques, like principal component analysis (PCA) and partial least squares (PLS), are considered.

These methods calculate components as a convex linear combination of the original features in an iterative procedure. These components are called scores. The procedure can shortly be summarized as follows: Starting with $\mathbf{X}_{\{0\}} = \mathbf{X}$, the vector $\mathbf{w}_{\{1\}} = [w_{1_{\{1\}}}, \ldots, w_{g_{\{1\}}}]^t$ containing the weight for each feature is determined and the corresponding score vector $\mathbf{t}_{\{1\}} = \mathbf{X}_{\{0\}}\mathbf{w}_{\{1\}}$ is calculated. Afterwards the already explained information in the scores is reduced of $\mathbf{X}_{\{0\}}$ resulting in the 1^{st} residuum $\mathbf{X}_{\{1\}}$ of $\mathbf{X}$. Then the next weight vector $\mathbf{w}_{\{2\}}$ is determined to calculate the next score vector $\mathbf{t}_{\{2\}}$ and so on. This is a simplified representation of the procedure, in detail there exist many variations to determine the weights and the residuum (Indahl *et al.*, 2009).

The main challenge is to determine the weights in such a way, that the important

information of the data remains in the components and the number of components (n_c, the dimension of the subspace) is lower than the original feature space ($n_c < g$). Because in this thesis the context is on classification, ideally the groups should be better separable in the subspace. In the following, two feature extraction methods are presented which base on linear transformations, partial least squares discriminant analysis (PLS-DA) and powered partial least squares discriminant analysis (PPLS-DA). In Section 2.3.1, PLS-DA is compared to other linear feature extraction methods.

Partial least squares discriminant analysis

PLS is a powerful approach often used in chemometrics (a statistical scientific discipline of model building and data analysis in chemistry) (Wold, 1995). It was developed for solving regression problems. If the response is a vector, PLS is denoted by PLS1 and if the response is a matrix, the corresponding PLS procedure is denoted by PLS2. In the PLS procedure, the weights of the linear combinations are called loading weights. The basic idea is to create the loading weights in such a way, that the covariance between $\mathbf{X}$ and $\mathbf{y}$ is maximized.

For classification, it is common to use PLS2 with the dummy matrix $\mathbf{Y}$ which indicates the group memberships (sometimes this method is also called partial least squares-discriminant analysis (PLS-DA) in the literature).
For PLS2 the first loading weights vector in the context of classification is the dominant eigenvector of the matrix

$$\mathbf{H} = \left(\sum_{\nu=1}^{2} {n_\nu}^2 \bar{x}_{(\nu)k} \bar{x}_{(\nu)l} \right)_{kl},$$

here $\bar{x}_{(\nu)k}$ denotes the mean value of feature $\mathbf{x}_k$ for group ν and $k = 1 \ldots, g$, $l = 1 \ldots, g$ (Barker and Rayens, 2003). The groups with a higher sample size get a higher weight in the calculation of the loading weights. Therefore Barker and Rayens (2003) and Nocairi *et al.* (2005) propose to use the dominant eigenvector of

the between-group covariance matrix $\mathbf{B}$:

$$\mathbf{B} = \left(\sum_{\nu=1}^{2} \frac{n_\nu}{n} \bar{x}_{(\nu)k} \bar{x}_{(\nu)l} \right)_{kl},$$

with $k = 1 \ldots, g$, $l = 1. \ldots, g$. Nocairi *et al.* (2005) shows the mathematical proof for this approach. Regarding classification problems, deflation of the response is not reasonable, because the dummy matrix $\mathbf{Y}$ contains discrete variables. Therefore the response should remain the same in each iterative step for the classification task.
In this thesis, the PLS procedure, based on the between-group covariance and with fix response matrix $\mathbf{Y}$, is denoted by PLS-DA.
In Figure 2.2 an example of coordinate transformation to a lower dimensional space using PLS-DA is shown, the dimension is reduced from three genes to only one component.

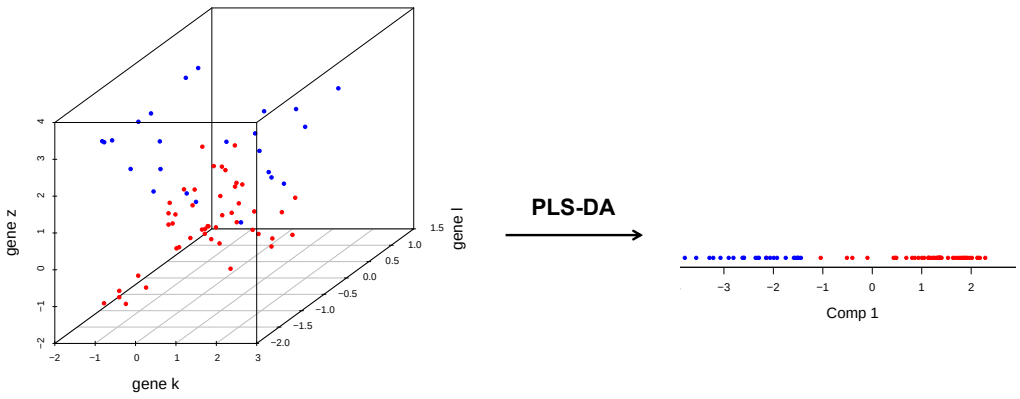


Figure 2.2: Demonstration of coordinate transformation to a lower dimensional space using PLS-DA (red data points belong to group 1, blue data points to group 2). In the left panel the original feature space is shown. The right panel shows the data in the new coordinate system, here the first component already leads to a separation of the groups.

PLS-DA can be described as the following maximization problem:

Let $[\mathbf{w}_{\{1\}}, \ldots, \mathbf{w}_{\{a-1\}}]$ *be the matrix containing the already determined loading*

weights, then the a^{th} loading weight is the vector $\mathbf{w}$ *which maximizes the between-group covariance of* $\mathbf{X}_{\{a-1\}}\mathbf{w}$ *with respect to* $\mathbf{Y}$*. Therewith according to Nocairi* et al. *(2005),* $\mathbf{w}_{\{\mathbf{a}\}}$ *is the dominant eigenvector of the between-group covariance matrix* $\mathbf{B}_{\{a-1\}}$ *and the corresponding score vector is* $\mathbf{t}_{\{\mathbf{a}\}} = \mathbf{X}_{\{a-1\}}\mathbf{w}_{\{\mathbf{a}\}}$.

The main steps of a PLS-DA algorithm are:

- determination of the loading weights vector $\mathbf{w}_{\{a\}}$ as dominant eigenvector of the between-group covariance matrix $\mathbf{B}_{\{a-1\}}$
- normalization of $\mathbf{w}_{\{a\}}$: $\mathbf{w}_{\{a\}} \mapsto \frac{\mathbf{w}_{\{a\}}}{||\mathbf{w}_{\{a\}}||}$
- calculation of the scores: $\mathbf{t}_{\{a\}} = \mathbf{X}_{\{a-1\}}\mathbf{w}_{\{a\}}$
- calculation of the loadings: $\mathbf{q}_{\{a\}} = \mathbf{X}^t_{\{a-1\}}\mathbf{t}_{\{a\}}$
- deflation of $\mathbf{X}_{\{a-1\}}$: $\mathbf{X}_{\{a\}} = \mathbf{X}_{\{a-1\}} - \mathbf{t}_{\{a\}}\mathbf{q}^t_{\{a\}}$.

Because all components are determined analogously, it is suffices to look at the first loading weight vector.
PLS-DA is related to Fishers canonical discriminant analysis (FCDA) where the loadings are calculated as the dominant eigenvector of $\mathbf{T}^{-1}\mathbf{B}$, here $\mathbf{T}$ denotes the total sum of squares and cross product matrix.

In this thesis, PLS-DA is applied as preprocessing step for dimension reduction. For the final classification the components are used as predictors of a LDA, as suggested for the final classification (Indahl *et al.* (2007)).

An enhanced version of PLS-DA with inclusion of prior probabilities (π_ν) in the estimation of $\mathbf{B}$ is recommended in Indahl *et al.* (2007). In this version, the importance of each group does not longer depend on the empirical prior probabilities. Therewith a direct opportunity is given to put more weight on special groups for the calculation of the loading weights. Moreover $\mathbf{X}$ is transformed to a $n \times 2$ matrix $\mathbf{Z} = \mathbf{X}\mathbf{W}_0$, here $\mathbf{W}_0 = \mathbf{X}^t\mathbf{Y}\sqrt{\mathbf{\Pi}}(\mathbf{Y}^t\mathbf{Y})^{-1}$ is the transformation matrix and $\mathbf{\Pi} = diag(\{\pi_\nu | \nu = 1, 2\})$

is a diagonal matrix with the square root of the prior probabilities as diagonal entries. Therewith a $g \times g$ eigenvalue problem ($\mathbf{B} \in \mathbb{R}^{g \times g}$) is reduced to a 2×2 eigenvalue problem (2=number of groups), because the between-group covariance matrix according to $\mathbf{Z}$ is considered. Afterwards the eigenvalue is transformed by $\mathbf{W}_0$. This approach is implemented in the software R in own PLS-DA code.

Indahl *et al.* (2007) gives a comprehensive overview of PLS-DA, for deeper insight Barker and Rayens (2003) and Nocairi *et al.* (2005) are recommended.

The advancement of PLS-DA presented in the following allows the concentration on a smaller number of features. This reduces the influence of features with little classification information.

Powered partial least squares discriminant analysis

PPLS-DA is a specialized version of PLS-DA and was introduced by Liland and Indahl (2009). A power parameter is established to improve the calculation of the loading weights for a better separation of the groups. Moreover, the maximization of the covariance is replaced by the maximization of the correlation.
In the usual PLS-DA approach the loading weight of a feature is determined, upon other terms, by the product of the correlation between the feature to the response and the standard deviation of the feature. Hence, a balanced influence is given by the correlation and the variance. This often makes models based on PLS stable (Indahl, 2005). Now for prediction, dominance $\mathbf{X}$-variance which is irrelevant does not lead to optimal models, therefore Indahl (2005) propose Powered PLS (PPLS) which allows the user to control the importance of the correlation part relative to the standard deviation part by a power parameter γ in the regression context. The power parameter γ enables a flexible different weighting of the correlation part and the standard deviation part of loading weights, which will be explained in detail now. The columns of the dummy matrix $\mathbf{Y}$ are denoted by $\mathbf{y}_1$ and $\mathbf{y}_2$. The matrix

$\mathbf{W}_0$, containing possible loading weights vectors as columns, can be factorized into three matrices: $\mathbf{W}_0 = \mathbf{CSP}$.

Here the matrix $\mathbf{C}$ contains the correlation between $\mathbf{X}$ and $\mathbf{Y}$, the matrix $\mathbf{S}$ contains the standard deviation of $\mathbf{X}$ and the matrix $\mathbf{P}$ contains the standard deviation of $\mathbf{Y}$ and the prior probabilities π_ν and group sizes n_ν:

$$\mathbf{C} = (\zeta_k \cdot |corr(\mathbf{x_k}, \mathbf{y}_\nu)|)_{k\nu},$$

with ζ_k the sign of $corr(\mathbf{x_k}, \mathbf{y}_\nu)$, $k = 1, \ldots, g$, $\nu = 1, 2$,

$$\mathbf{S} = diag(\{sd(\mathbf{x}_k) | \, k = 1, \ldots, g\})$$

and

$$\mathbf{P} = diag(\{n \cdot sd(\mathbf{y}_\nu)\frac{\sqrt{\pi_\nu}}{n_\nu} | \, \nu = 1, 2\}).$$

Now, to the entries of the matrices $\mathbf{C}$ and $\mathbf{S}$ an exponent is added depending on the so called power parameter $\gamma \in U$, resulting in $\mathbf{C}(\gamma) = (\zeta_k \cdot |corr(\mathbf{x_k}, \mathbf{y}_\nu)|^{\frac{\gamma}{1-\gamma}})_{k\nu}$, $k = 1, \ldots, g$, $\nu = 1, 2$ and $\mathbf{S}(\gamma) = diag(\{sd(\mathbf{x}_k)^{\frac{1-\gamma}{\gamma}} | \, k = 1, \ldots, g\})$. The power parameter γ can be chosen from the open interval $U = (0, 1)$. The matrix $\mathbf{W}_0(\gamma) = \mathbf{C}(\gamma)\mathbf{S}(\gamma)\mathbf{P}$ contains the candidate loading weights as columns (depending on γ)

$$\mathbf{w}_\nu(\gamma) = K_{\nu,\gamma} \cdot \omega_{\nu,\gamma},$$

with $\omega_{\nu,\gamma} = \left[\zeta_1 \cdot |corr(\mathbf{x_1}, \mathbf{y}_\nu)|^{\frac{\gamma}{1-\gamma}} \cdot sd(\mathbf{x_1})^{\frac{1-\gamma}{\gamma}} \ldots, \zeta_g \cdot |corr(\mathbf{x_g}, \mathbf{y}_\nu)|^{\frac{\gamma}{1-\gamma}} \cdot sd(\mathbf{x_g})^{\frac{1-\gamma}{\gamma}}\right]^t$ and $K_\nu = n \cdot sd(\mathbf{y}) \cdot \sqrt{\pi_g}/n_\nu$, $\nu = 1, 2$.

If the power parameter γ tends towards 0 ($\gamma \searrow 0 \Rightarrow \frac{1-\gamma}{\gamma} \to \infty$, $\frac{\gamma}{1-\gamma} \to 0$), the standard deviation part is mainly used for the calculation of the loading weights and features with a large standard deviation get larger absolute weights than the remaining ones. If the power parameter tends towards 1 ($\gamma \nearrow 1 \Rightarrow \frac{1-\gamma}{\gamma} \to 0$, $\frac{\gamma}{1-\gamma} \to \infty$),

the correlation part is mainly used and features with a large correlation towards $\mathbf{y}_\nu$ get larger absolute weights. The end points of the interval U are checked in separate steps in the PPLS-DA algorithm. If γ equals 1, only the feature(s) with the largest correlation get a non-zero loading weight and if γ equals 0 only the feature(s) with the largest standard deviation get a non-zero loading weight. The power parameter γ enables different weighting of features with large correlation or standard deviation.

The determination of the loading weight of PPLS-DA could be summarized as follows (here reported for the first component for reasons of simplicity): First again the transformation $\mathbf{Z}(\gamma) = \mathbf{X}\mathbf{W}_{\{0\}}(\gamma)$ is performed to reduce the dimension. For the transformed matrix $\mathbf{Z}(\gamma)$ the between group sum of squares and cross-product matrix including prior probabilities can be calculated as follows
$\mathbf{B}_\Pi(\gamma) = n\mathbf{Z}(\gamma)^t\mathbf{Y}(\mathbf{Y}^t\mathbf{Y})^{-1}\mathbf{\Pi}(\mathbf{Y^t}\mathbf{Y})^{-1}\mathbf{Z}(\gamma)$ and the total variance matrix is obtained as $\mathbf{T}_\Pi(\gamma) = n\mathbf{Z}(\gamma)^t\mathbf{V}_\Pi\mathbf{Z}(\gamma)$ with $\Pi = diag(\{\pi_1, \pi_2\})$ and $\mathbf{V}_\Pi = diag(\{v_i | v_i = n\pi_g/n_g, i = 1, \ldots, n\})$.

The maximization problem of PPLS-DA is then

$$\arg\max_{\gamma\in[0,1]} \frac{\mathbf{a}^t\mathbf{B}_\mathbf{\Pi}(\gamma)\mathbf{a}}{\mathbf{a}^t\mathbf{T}_\mathbf{\Pi}(\gamma)\mathbf{a}}. \tag{2.2}$$

The solution of the optimization problem (2.2) is denoted by γ_{max}. To avoid singular matrices $\mathbf{T}_\Pi(\gamma)$ and to get a numerically more stable solution, Liland and Indahl (2009) substitute the maximization problem (2.2) by the maximization problem $\arg\max_{\gamma\in[0,1]} \text{cca}(\mathbf{Z}(\gamma), \mathbf{Y})$, where cca denotes the canonical correlation. Indahl *et al.* (2009) showed that these procedures are equivalent.

The power parameter γ_{max} is determined by maximization of the canonical correlation:

$$\gamma_{max} = \arg\max_{\gamma\in[0,1]} \text{cca}(\mathbf{Z}(\gamma), \mathbf{Y}).$$

In the algorithm of PPLS-DA in the R package pls, the R function *optimize* is

used to search for the maximum. The final loading weight vector is then $\mathbf{w} = \mathbf{W_0}(\gamma_{max})\mathbf{a}_{\gamma_{max}}$.

For each component, a power parameter is separately calculated as explained above. For PPLS-DA used in this thesis, the R function *cppls* is applied with the parameters lower=0 and upper=1, defining the range of the interval $U = [lower, upper]$.

Analogous to PLS-DA, the components of PPLS-DA could be used as predictor for a classification method. In this thesis, PPLS-DA components build the predictors for an LDA, as suggested by (Liland and Indahl, 2009). For convenience PLS-DA combined with LDA is abbreviated by PLS-DA and PPLS-DA combined with LDA by PPLS-DA as in Liland and Indahl (2009).

Determination of the number of components

For PLS-DA and for PPLS-DA it is necessary to choose the number of components. This is done in an inner cross-validation approach (on the training set, see Section 2.1.2) as recommended by e.g. Liland and Indahl (2009) and Boulesteix (2004).

Successively for component number $a, a = 1 \ldots, n_c$ the corresponding models on the inner training set are built and the PE of LDA for the inner test set is stored. This step is repeated r_{inner} times. Afterwards, the mean PE for each fix number of components is calculated. Finally the number of components used, $n_{c(opt)}$, leads on average to the smallest PE over all r_{inner} repetitions.
In this work the PLS-DA and the PPLS-DA models are based on ten inner-cross-validation steps ($r_{inner} = 10$) to determine this optimal number of components.

Comparison to other feature extraction methods

In the following, four linear feature extraction methods are compared, PLS-DA, PPLS-DA using $\gamma = 0.5$, PCA and CCA, corresponding to Borga (2001). All four

methods are based on the following eigenvector problem for the determination of the weights for the linear combination:

Let $[\mathbf{w}_{\{\mathbf{1}\}}, \ldots, \mathbf{w}_{\{\mathbf{a-1}\}}]$ *be the matrix containing the already determined (loading) weights, then* $\mathbf{w}_{\{a\}}$ *is the dominant eigenvector of* $\mathbf{E}^{-1}_{\{a-1\}}\mathbf{D}_{\{a-1\}}$.

The dominant eigenvector of $\mathbf{E}^{-1}_{\{a-1\}}\mathbf{D}_{\{a-1\}}$ maximizes the so called Rayleigh quotient $\frac{\mathbf{w}^t\mathbf{D}_{\{a-1\}}\mathbf{w}}{\mathbf{w}^t\mathbf{E}_{\{a-1\}}\mathbf{w}}$. The different methods investigate different forms of the matrices $\mathbf{E}_{\{a-1\}}$ and $\mathbf{D}_{\{a-1\}}$.

To simplify the comparison between the method PPLS-DA using $\gamma = 0.5$ to the other methods, the transformation of the matrix $\mathbf{X}$ to $\mathbf{Z}$ is omitted, because it only reduces the complexity of the eigenvalue problem (see Section 2.3.1).

Table 2.1 summarizes the matrices for the determination of the first (loading) weights vector.

Table 2.1: Form of the matrices $\mathbf{E}$ and $\mathbf{D}$

name	**E**	**D**
PLS-DA	**I**	**B**
PPLS-DA using $\gamma = 0.5$	**T**	**B**
CCA	**T**	**B**
PCA	**I**	**T**

The identity matrix is denoted by $\mathbf{I}$ and the total covariance matrix of $\mathbf{X}$ by $\mathbf{T}$. PCA investigates only the variance of the data matrix $\mathbf{X}$. Therefore PCA works only for dimension reduction of classification problems if the "among-group variability soundly dominates the within-group variability" (Barker and Rayens, 2003), because then the maximization of the variance leads to a discrimination of the groups. If the data structure shows a larger within-groups variability than that among groups, PLS-DA is preferable to PCA. For all methods except PCA, the matrix $\mathbf{D}$ equals the between-groups covariance matrix $\mathbf{B}$. Therefore all other methods use the information of the group memberships (supervised methods). Moreover PPLS-DA using

$\gamma = 0.5$ corresponds to CCA.

2.3.2 Techniques to detect biomarkers

The identification of a biomarker can be understood as a dimension reduction procedure, because a subset of the g features is chosen as biomarker. Therefore a criterion is needed to decide which features are important for discrimination between the groups and which are not. This means the features need to be evaluated towards their discrimination ability, resulting in an ordered list of the features according their evaluation (ranking list). The evaluation can be done by univariate techniques or multivariate techniques (Lai *et al.*, 2006). The univariate method assesses each feature separately, to determine an importance value for each feature. An example of a univariate method is the t-test, where a p-value is calculated for each feature. In contrast, multivariate methods focus on a combination of features. An example is two features which separately are considered to have a low or no discrimination ability between groups, but in combination the groups are linear separable (Guyon and Elisseeff, 2003; Haynes and Rees, 2006).

Multivariate feature selection methods can especially be useful for complex classification problems (e.g. Breast cancer patients with poor or good diagnostics) with large collinearity among the features and where a single gene cannot build a biomarker. Especially for biomarker studies dealing with early detection of disease, combination of features can be advantageous (Etzioni *et al.*, 2003; Han *et al.*, 2009). Intuitively a biomarker containing more than one feature is preferable in such cases, because a combination of features is more robust than a single feature with respect to variation in the data.

Often correlation is a measure of the multivariate methods to create a importance value (Lai *et al.*, 2006).

Summarizing, users should consider both univariate and multivariate methods, because the performance depends on the considered data set.

In the following, a univariate method which is applied in this thesis is introduced to determine a ranking list.

Univariate techniques

A first step to find a candidate biomarker could be a statistical test which detects discriminating features between two or more groups. Liu *et al.* (2002) propose to choose the features according to a ranking list created e.g. by the t-statistic or the χ^2-statistic. Also Hedenfalk *et al.* (2001) apply the t-statistic to select genes. Dettling and Buehlmann (2003) suggest using Wilcoxon's rank sum statistic. In this thesis the t-test is chosen, because this simple univariate method is known to outperform more complex feature selection methods in some studies according to prediction accuracy (Haury *et al.*, 2011).

t-test The t-test goes back on the work of William Sealy Gosset (Student, 1908). The assumptions for using a t-test are a normally distributed population and equal variances for the groups $\nu = 1, 2$. Considering a feature k as random variable (X_k), differences in the mean values of the groups are tested for a realization $\mathbf{x}_k$ of X_k. For these case of a two-sided test, the null hypothesis is that the group's mean values are equal $H_0 : \mu_{(1)k} = \mu_{(2)k}$ and the alternative hypothesis (in which we are interested) is that the group's mean values are different $H_1 : \mu_{(1)k} \neq \mu_{(2)k}$, where $\mu_{(\nu)k}$ denotes the unknown mean of group ν for feature k. The t-statistic for a two-sided test is (Sumpf and Moll, 2004)

$$T(\mathbf{x_k}) = \frac{\bar{x}_{(1)k} - \bar{x}_{(2)k}}{s} \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}},$$

here $\bar{x}_{(\nu)k}$ denotes the mean value of $\mathbf{x_k}$ for group ν, $\mathbf{x}_{(\nu)k}$ denotes the values of feature $\mathbf{x}_k$ for group ν and $s = \big((n_1 - 1)var(\mathbf{x}_{(1)k}) + (n_2 - 1)var(\mathbf{x}_{(2)k})\big) / (n_1 + n_2 - 2)$. Let d denotes the test statistic of $\mathbf{x_k}$, $d = T(\mathbf{x_k})$. Then the p-value is the probability, that for any other realization $\mathbf{x}'_k$ of X_k the test statistic (denoted by d') is equal

or greater than the observed one d assuming the null hypothesis, ($p = P(d' \geq d\,|H_0)) \cup P(d' \geq -d\,|H_0)$). A feature ranking list based on the p-value, is constructed as a top list containing the features with the smallest p-value first. That means, that the probability of randomly finding a higher discrepancy than $|d|$ is small for these features. A small probability speaks against the null hypothesis (Bender and Lange, 2007).

Multiple testing The following descriptions are mainly conform with Bland and Altman (1995). In the context of metabolite or gene expression studies not only the mean value difference with respect to a single gene is interested, all genes measured are considered. For each gene a test is applied (multiple testing). For g genes, therewith g null hypotheses ($H_{0_1}, \ldots, H_{0_g}$) are given with the same significance level α. Considering each test separately, if H_{0_j} is correct, the probability to reject H_{0_j} is α (called comparison-wise error) for $j = 1 \ldots, g$. In the case that several null hypotheses are correct at the level of α and assuming that the test statistics are independent, the probability that more than one null hypothesis is rejected, is larger than α namely $(1 - (1 - \alpha)^g)$ (called family-wise error). Taking now the number of multiple tests into account, the multiple tests are corrected towards the described problem with respect to the significance level. One approach to control the significance level α is the Bonferroni method: If the aim is that all of the correct g null hypotheses are separately rejected with a probability of 0.05, than choose for each single test $\alpha = 0.05/g$. This easy way to control α has the disadvantage that the statistical power (rejection of the null hypothesis when the null hypothesis is false) is reduced. Therefore this Bonferroni method is an example of a too conservative method. Benjamini and Hochberg (1995) propose the false discovery rate (FDR) as expected value of the number of false positives (type I error) divided by the number of false and true positives. This value should be smaller than α. In this context "positive" denotes a significant result. If the divisor is zero the FDR is defined as zero. Storey (2003) described the so called q-value as "the smallest FDR value for

which the test is significant". With the q-value choice, it is possible to control the expected number of false positives.
In this thesis the R-package `stats` is applied for a t-test and an FDR correction based on Storey (2003) (R-package `qvalue`, Dabney *et al.* (2009)) is used.

Multivariate techniques

Multivariate techniques to build a ranking list are given, for example, by the importance values available for the multivariate methods. In this thesis, the mean decrease accuracy, the importance value of RF is applied and the absolute loading weights of the first component of PLS-DA and PPLS-DA.

Mean decrease accuracy The RF mean decrease accuracy (MDA) is calculated for a feature k as the proportion of correctly predicted samples in the OOB samples subtracted by the proportion of correctly predicted samples of the OOB samples if the values of feature k are permuted. This difference is averaged over each tree in the forest and divided by the standard error. The MDA is an importance value for each feature, which also depends on all other features. Therefore this importance value is determined in a multivariate way.

Importance value of PLS-DA and PPLS-DA It is often stated, that feature extraction methods like PLS-DA are helpful for the final classification, but that these components are difficult to interpret especially for biologists and medicals. Statements about the importance of the original features are not so easy to make, because for the final classification the components are used as predictors for classification methods. Which original features are important for the classification can be decided by considering the corresponding loading weights. Loading weights are weights for the linear combination of the original features and the loading weights are suitable as an importance value of the original features for the classification. In this thesis, the absolute values of the loading weight for the first component is used as importance

value of the methods PPLS-DA and PLS-DA.

Creating a biomarker

Once creating a ranking list, the next question is how to select the features which form a biomarker?

In Liu *et al.* (2002) a fix number of the top 20 features of a ranking list is considered (e.g. χ^2-statistics scores, t-statistics scores). Although they mentioned that the optimal number of selected features depends on the data set, the classification method and on the technique selecting the features used. The restriction to a fixed number of features building a biomarker candidate can impair the establishment of a classification rule. Including additionally one or two further features could improve the prediction significantly. Moreover there exists not only a single set with the best prediction performance. For sets with different cardinalities a similar prediction performance could be achieved (Lai *et al.*, 2006). Dudoit and Fridlyand (2003) suggested selecting the first k features of the ranking list for $k = 10, 50, 100, 500, 1000, g$ (for data sets with $g > 1000$) to see how the reduction of the number of predictors influences the performance of the classification. For a biomarker search with the aim of a small number of features, more precise increments are preferential. Therefore in this thesis the top $k = 2, \ldots, g$ features of the ranking list are successively chosen (see Chapter 3). According to the second requirement of a biomarker, the set corresponding to the lowest cardinality is chosen to build the biomarker.
In the literature many feature selection techniques are discussed see for example Langley (1994); Inza *et al.* (2004), like filtering in a previous step features (which correspond here to the t-test as filter method) and so-called wrapper approaches which use a classifyer to evaluate the feature importance to select the feature subset (Kohavi and John, 1997). On the one hand filter methods select feature subsets independently of a classifier, which can be a drawback (Darzi and Asghari, 2011), but on the other hand they are not so costly with respect to computational time.

Wrapper approaches which are much more time consuming use as measure for the feature importance the prediction accuracy which seems the best way to evaluate the feature according to their importance for the classification (Das, 2001). Therefor wrapper approaches should result in a feature subset with a higher prediction accuracy and are so advantageous for biomarker search.

3 Example of biomarker discovery with statistical learning methods in the context of tuberculosis

3.1 Introduction

A practical example of biomarker search with statistical learning methods is introduced to show the different steps of biomarker search and the statistical aspects which should be considered. This example of biomarker search is in the context of tuberculosis. Tuberculosis is an infection disease caused by Mycobacterium tuberculosis. Infection does, however, not necessary cause the active disease. Only a few people come down with active tuberculosis (TB) (World Health Organization, 2011). Although the tuberculosis incidences, the tuberculosis cases and the number of deaths from tuberculosis have been falling, tuberculosis is still a major health problem with 8.8 million incident cases and 1.45 million deaths in 2010 (World Health Organization, 2011). Therefore the development of new therapeutic options is one point on the Stop TB Strategy established by the WHO in 2006 in the 2015 global targets for reductions in this disease. The Max-Planck Institute for Infection Biology in Berlin is one of several institutes which study the disease tuberculosis. The researchers are especially concerned with the development of biomarkers for active tuberculosis (Jacobsen *et al.*, 2008; Maertzdorf *et al.*, 2011; Jacobsen *et al.*,

2007). The focus is on the differences between the people remaining healthy after an infection in contrast to the people developing the disease. A biomarker discriminating between these persons is important for further developments of drugs and vaccines (Maertzdorf *et al.*, 2011).

The example of biomarker identification considered in this chapter bases on metabolite profiles separating between patients with active tuberculosis and persons which are infected but healthy.

3.2 Design and data

The design was given because all data were collected before the dissertation started in the years 2006 and 2007. In general three groups of persons are distinguished. The non-infected persons (TST^-), the infected but healthy persons (TST^+) and the persons with active tuberculosis (TB), see Table 3.1. In this study 136 blood samples are measured, each sample belongs to one of the three groups. All persons are HIV negative and all TB patients have not received a therapy at the time of sample collection. The TST^- and the TST^+ group are divided into 17 men and 29 women with a mean age of 27.7 years in the TST^- group and of 27.3 years in the TST^+ group. The TB group contains 19 men and 25 women and have a mean age of 26.3 years. In total 389 metabolites are measured by the mass spectrometry analysis.

Table 3.1: Overview of the metabolites study

group	**infection status**	**sample size**
TB	tuberculosis patients	44
TST^+	healthy, infected persons	46
TST^-	healthy, non-infected persons	46

Normalized data are used, provided by the Max Planck Institute for Infection Biology, Berlin. Block normalization (the median value of each run-day block is cal-

culated as reference size) has been applied to minimize the inter-day instrument variation. Missing values are only replaced if 50% of the values for the metabolite considered are reported, then the missing value is substituted by the minimum of the values. The experimental procedure including description of the specimens, sample preparation, mass spectrometry analysis and data curation, are in detail described in Weiner *et al.* (2011).

3.3 Comparison of classification methods

The aim is to discriminate between the TB and the TST^+ samples, only these two groups are concerned for all further steps of the identification of a biomarker, resulting in a 90×389 data matrix $\mathbf{X}$. All shown results in this chapter are based on $r = 500$ resampling steps of the outer training and outer test set. As a kind of a pre-study, the six classification methods introduced in Chapter 2 (SVML, SVMR, RF, PLS-DA, PPLS-DA and t-LDA) are compared, to find out which method fits best to the metabolite data, like it is suggested by Feng *et al.* (2004). To enable comparable results, all classification methods use the same training and test sets. A ratio of $\phi_{\mathbf{X}} = 0.7$ is chosen as ratio of the training set on the whole data set $\mathbf{X}$ and also for PPLS-DA and PLS-DA the ratio of the inner training set on the outer training set ($\phi_{\mathbf{X_{train}}}$) is set to 0.7. As classification error the mean PE results and the corresponding 95% confidence intervals are investigated, Figure 3.1 shows the results for $r = 500$ resampling steps of the training and test set. RF shows the lowest PE around 0.050, SVML, SVMR and PPLS-DA have a similar PE around 0.070 and t-LDA and PLS-DA show the largest PE around 0.082. The method RF is chosen for further applications for the search of a biomarker, because this method seems to be the most adequate for this metabolite data set and the discrimination between the TB and the TST^+ persons.

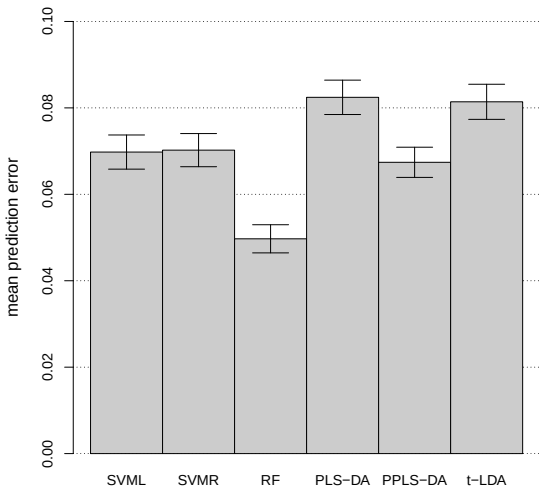


Figure 3.1: Mean PEs (height of the bars) and 95% confidence intervals for the metabolite data for six different classification methods.

3.4 Biomarker search using random forest

Before the actual selection of a biomarker started, the influence of the parameters $\phi_{\mathbf{X}}$ and $mtry$ on the PE of RF and the reproducibility of the ranking list are studied for different resampling steps of the training set (Section 2.1.2).

3.4.1 Influence of the parameters $\phi_{\mathbf{X}}$ and $mtry$ on the prediction

For $\phi_{\mathbf{X}}$ (the ratio of the objects chosen for the training set out off all 90 objects) the following different parameter settings are investigated $\phi_{\mathbf{X}} = 0.5, 0.6, 0.7, 0.8, 0.9, 0.95$. Table 3.2 gives an overview of the parameter $\phi_{\mathbf{X}}$ and the corresponding sample sizes of the training and the test set. For $\phi_{\mathbf{X}} = 0.9$, the test sets contain only 9 objects for the calculation of the PE. Therewith if l subjects are classified wrongly, the PE for this run is $l/9$. For $\phi_{\mathbf{X}} = 0.95$ the proportion of wrongly classified objects is even larger, because the step size of the possible error rates is $1/4$. Hence, the PE depends on the sample size of the test set. Also for $mtry$ (the number

of features chosen at each node to find the best split) different values are chosen, $mtry = 10, 15, 20, 25, 30, 50, 70, 100, 200$. The rule of thumb ($mtry = \sqrt{g} \approx 20$ with $g = 389$) equals therewith 20. For all possible combination of the parameters $\phi_{\mathbf{X}}$ and $mtry$, the results are illustrated in Figure 3.2. All 95% confidence intervals are overlapping except for $mtry = 100, 200$. These intervals are not shown to preserve clarity in Figure 3.2. No great differences in the mean PE are found for different choices of $\phi_{\mathbf{X}}$ in combination with different choices of $mtry$ except for $mtry = 100, 200$. A very high number of features (100 or 200) to choose at each node of the trees grown in RF, results in a higher error than for smaller choices of $mtry$ for all different $\phi_{\mathbf{X}}$. The mean PE is the lowest for $\phi_{\mathbf{X}} = 0.8$ for $mtry$ between 20 and 30. Because $mtry = 20$ equals the rule of thumb, this value is chosen for all further considerations. For the ratio $\phi_{\mathbf{X}}$ of the training set on the whole data set 0.8 is chosen, because this value shows the lowest mean PE.

Table 3.2: Overview of parameter $\phi_{\mathbf{X}}$ and resulting sample size of the training set

$\phi_{\mathbf{X}}$	0.5	0.6	0.7	0.8	0.9	0.95
n_{train}	45	54	63	72	81	86
n_{test}	45	36	27	18	9	4

3.4.2 Determining the biomarker

The aim is to find a biomarker discriminating between TST^+ and TB persons. This set as predictor should result in a minimal PE and its cardinality should be minimal. For the determination of the biomarker, the MDA (mean decrease accuracy, the importance value of RF) is applied (see Section 2.3.2).

In a first step, according to MDA a ranking list is created. Figure 3.3 shows this metabolite ranking (y-axis) in dependency on the mean MDA (x-axis). The red points depict the mean value of the MDA for 500 resampling steps of the objects for the training and the test set. The red lines represent the corresponding 95% confidence intervals. The black points illustrate exemplarily the value of MDA for

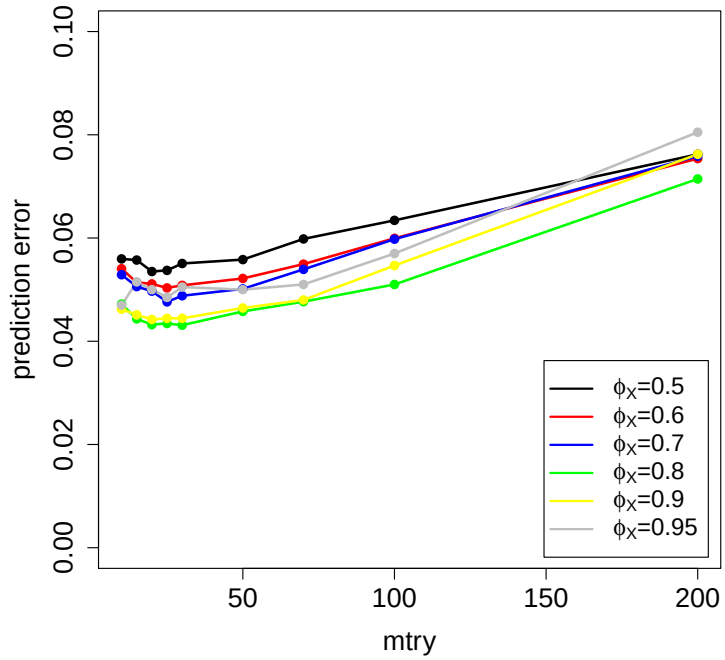


Figure 3.2: Mean PE of RF plotted against *mtry* for different choices of $\phi_{\mathbf{X}} = 0.5, 0.6, 0.7, 0.8, 0.9, 0.95$.

one single run of RF (one partition of the data set in a training and a test set). The metabolite with number 250 is the most important one for the classification. For the next 12 metabolites of the ranking list, large overlapping of the 95% confidence intervals are seen. Therewith these features have similar importance for the classification. Overall, the more the importance value decreases, the more the range of the 95% confidence intervals decreases. Regarding the black points, it can be seen how strong the ranking list can vary between different partitions of the objects into training and test set. Therefore it is preferable to repeat the calculation of the importance value several times to get a more robust (against varying individuals) and a more reproducible ranking list. In the following, a procedure is proposed for the determination of a feature set as candidate biomarker according to a ranking list. Figure 3.4 shows a simplified scheme illustrating the steps for determination of a feature list for a candidate biomarker.

The calculation of the metabolite ranking list according to the MDA is based on

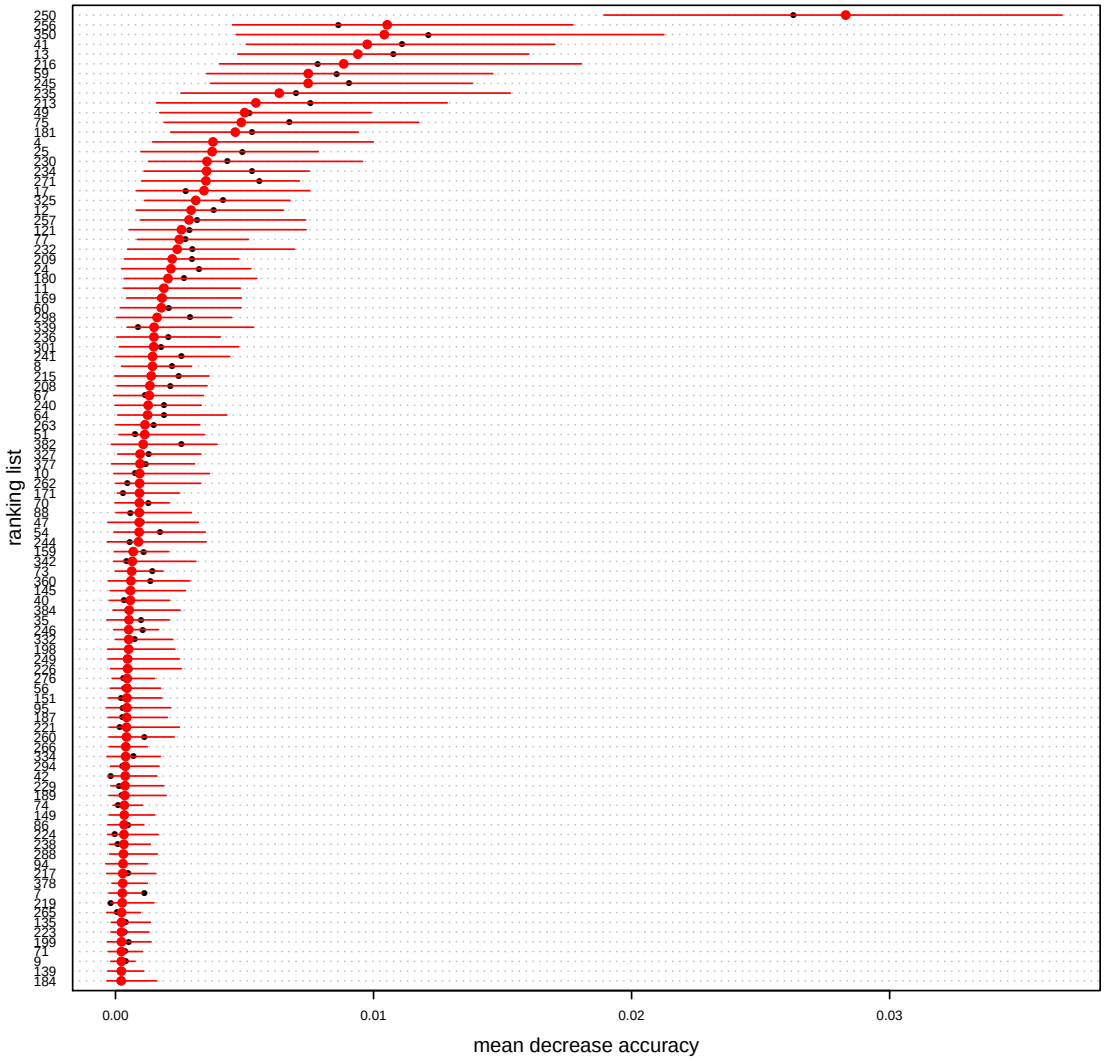


Figure 3.3: Plot of the mean MDA (red points) with 95% confidence intervals for the metabolite ranking (only the first 200 features are shown) averaged over 500 repetitions. The black points depict the MDA of a single random forest.

inner cross-validation steps (see Section 2.1.2) to determine a ranking list for the underlying outer training set. Thus for an outer training set a ranking list (M_1,..., M_{389}) is created based on r_{inner} repetitions of the sampling for the inner training and inner test set. Each set $\{M_1,\ldots, M_k\}$ of the top $k = 2\ldots, 389$ metabolites of the ranking list, build the predictors for RF. The corresponding PEs of the outer test set are calculated, resulting in 388 PE values. The resampling of the outer training and outer test set is repeated $r = 500$ times. Resulting in a matrix with 388 rows and 500 columns, each entry determines the PE for a certain number of metabolites used and a special choice of the outer training and the outer test set.

Then the set with the lowest cardinality, leading to a PE which could not be improved by adding further metabolites, is identified as biomarker.

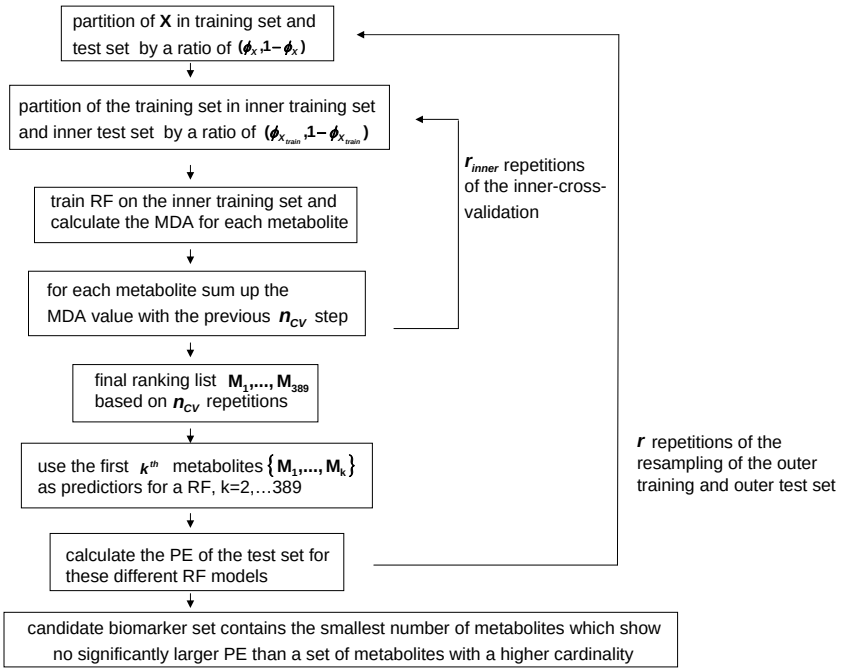


Figure 3.4: Simplified scheme for the determination of a biomarker with RF.

As an example the mean PE for each number of metabolites used and the corresponding 95% confidence intervals are shown in Figure 3.5 using $r_{inner} = 100$. The mean PE clearly decreased for the first 6 sets of top ranked metabolite ($\{M_1, M_2\}, \ldots,$ $\{M_1, \ldots, M_7\}$). For a certain cardinality of the metabolite set, no improvement is

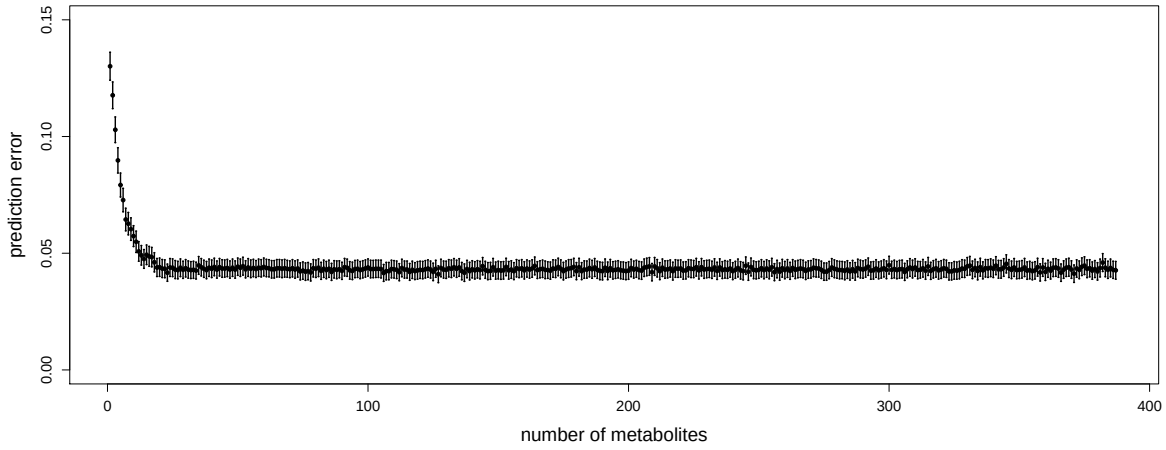


Figure 3.5: Mean PEs and corresponding 95% confidence intervals of RF plotted against numbers of metabolites used.

shown, all 95% confidence intervals are overlapping. The mean PE, belonging to a predictor set containing the first 19 top ranked metabolites, lies in all 95% confidence intervals of the PEs for predictor set with a higher cardinality (see Figure 3.6). Therefore a predictor set is proposed containing these first 19 metabolites. The corresponding mean PE to this biomarker is 0.046 with a 95% confidence interval of $[0.042, 0.050]$.

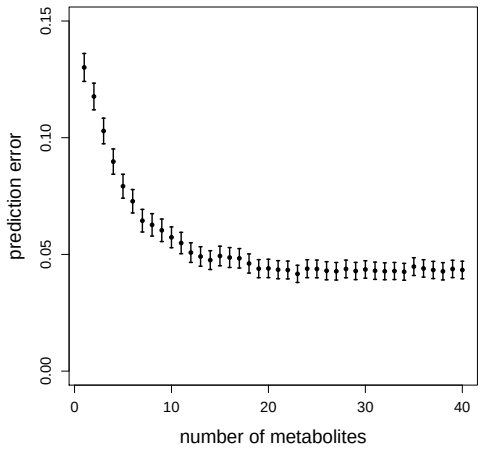


Figure 3.6: Mean PEs and corresponding 95% confidence intervals of RF for up to 40 metabolites used.

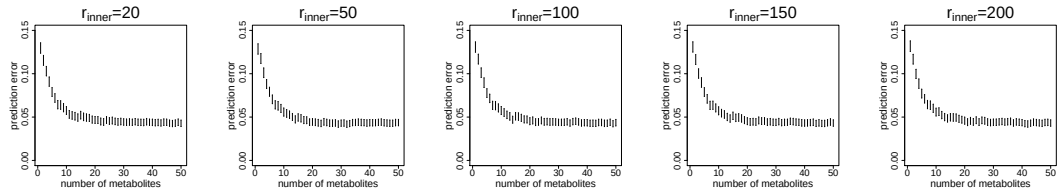
Summarizing the introduced approach to determine a biomarker, the prediction ability is compared for stepwise rising cardinalities of metabolite sets. Successively, additional features are added which is a kind of forward selection (Langley, 1994). Therewith a gapless observation of the PE in dependency on the number of predictors used is possible. In the literature similar procedures are proposed. For example, Dudoit and Fridlyand (2003) proposed to consider sets of features also according to a ranking list, but with a fix cardinality of $k = 10, 50, 100, 500, 1000, g$ (in their study $g > 1000$). These ranking lists are calculated according to the t-statistic and the Wilcoxon statistic without inner cross-validation and the final classification methods used are e.g. RF, SVMR and SVML. Borgia *et al.* (2009) use also the MDA value to select a biomarker, but first they filtered the features by the p-value of the Mann-Whitney test. Therefore first a univariate importance value is applied to select features which have even alone a good discrimination ability. Hence, features which lead only in combination with other features to group separation are potentially excluded from the study, and a loss of valuable features for biomarker candidates is possible.

Influence of r_{inner}

Additionally the influence of the number of inner cross-validation steps r_{inner} to create the ranking list is studied. For r_{inner} five different values are considered: $20, 50, 100, 150, 200$. Only small differences in the PE of the outer test set are found for the different choices of r_{inner} and for $k, k = 1, \ldots, 50$ metabolites used (see Figure 3.7), and there are no significant differences. Comparing the ranking lists (data not shown), for more than 50 repetitions ($r_{inner} \geq 50$) the first 20 metabolites of the ranking list are equal. The metabolites of lower positions differ more than the top 20 metabolites. This is not surprising, when Figure 3.3 is in mind. Especially for the metabolites more at the bottom of the ranking list, the MDA values are very similar. A choice of $r_{inner} = 100$ is suggested, which is a compromise between prediction accuracy of the position of the ranking list (estimation of the importance

value) and running time (number of repetitions of the inner cross-validation).

Similar results are found by using the mean decrease Gini index instead of the MDA value, which is also an importance value of RF for each feature (data not shown).



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

Figure 3.7: Mean PEs and corresponding 95% confidence intervals of RF depending on up to 50 metabolites for different choices of r_{inner}.

Summarizing the results, the parameter choice of $\phi_{\mathbf{X}}$ and $mtry$ shows no great influence on the PE except for extreme values of $mtry$ like 100 or 200. Breiman (2002) reports also no great influences of $mtry$ in a moderate range around the rule of thumb. Moreover, the ranking list according to the MDA can vary if only separate single runs are considered. The dependence of the ranking list on the training set is in agreement to the findings of Ein-Dor *et al.* (2004). Therefore an inner-cross-validation approach is applied to determine the ranking list. This ranking list is the basis for the biomarker specification. A biomarker is identified comprising the first top ranked 19 metabolites resulting in a mean PE of 0.046 for the discrimination between TB and TST^+ samples.

3.4.3 Comparison of the feature ranking list produced by RF and PPLS-DA

RF is selected for the determination of a biomarker, because this method shows the lowest mean PE (Figure 3.1). For PPLS-DA the PE is only slightly higher. Can this small differences in the PE, lead to larger differences in the ranking list? To answer this question, a ranking list created by PPLS-DA in an analogous way is used to check for correspondences. Thus it is possible to study how stable the

ranking list of RF is with respect to a different variable importance value of another method. The absolute loading weights of the first component are used as importance value according to PPLS-DA (see Section 2.3.2). Analogous to the procedure for RF, the parameters are $\phi_{\mathbf{X}}$=0.8 and $r_{inner} = 100$ for PPLS-DA. The findings for PPLS-DA are illustrated in Figure 3.8 which shows the mean PE in dependency on the number of metabolites used as predictors. For the first three sets (($\{M_1, M_2\}, \ldots, \{M1, \ldots, M_4\}$)) the mean PE clearly decreases. Comparing to the results of RF (Figure 3.6), for PPLS-DA already the first two metabolites as predictors achieve a lower PE.

A set with the top 8 metabolites, leads to a mean PE of PPLS-DA which is inside the 95% confidence interval corresponding to a larger number of predictors. According to PPLS-DA, a biomarker is proposed with the top 8 metabolites of the ranking list of PPLS-DA (see Figure 3.9). A mean PE of 0.059 is achieved for this set with the 95% confidence interval $[0.055, 0.064]$. In comparison, the biomarker specified by RF leads to a significantly lower PE. However for $\phi_{\mathbf{X}}$=0.8, the outer test set contains 18 objects (see Table 3.2). This means if one person is classified wrongly a PE of $1/28 \approx 0.056$ is caused. Considered over the 500 repetitions (r=500), there is no large loss in prediction accuracy for PPLS-DA in comparison to RF (with a PE of 0.046) and PPLS-DA needs less than half of the features used with RF for optimal prediction.

Considering the underlying ranking list more in detail, Figure 3.10 shows the importance values corresponding to the ranking lists with respect to RF and PPLS-DA (the green lines belong to the top 19 metabolites for RF and top 8 for PPLS-DA). For each method, the metabolite 250 shows an importance value which is nearly twice as large as for the remaining metabolites. For RF the top 19 metabolites correspond to a cut-off value of 0.0020 for the MDA, all metabolites with a lower importance value do not further improve the PE of RF. The corresponding value for PPLS-DA is 0.0748.

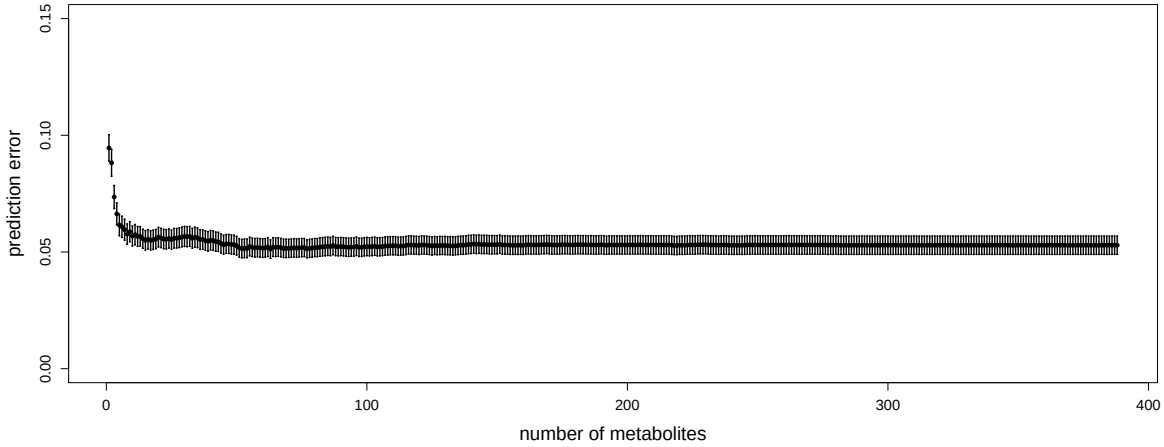


Figure 3.8: Mean PEs and corresponding 95% confidence intervals of PPLS-DA plotted against numbers of metabolites used.

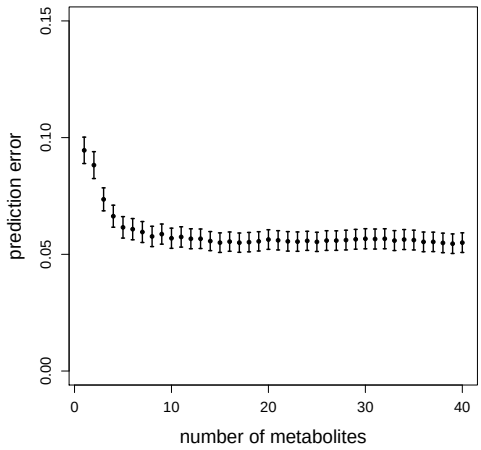
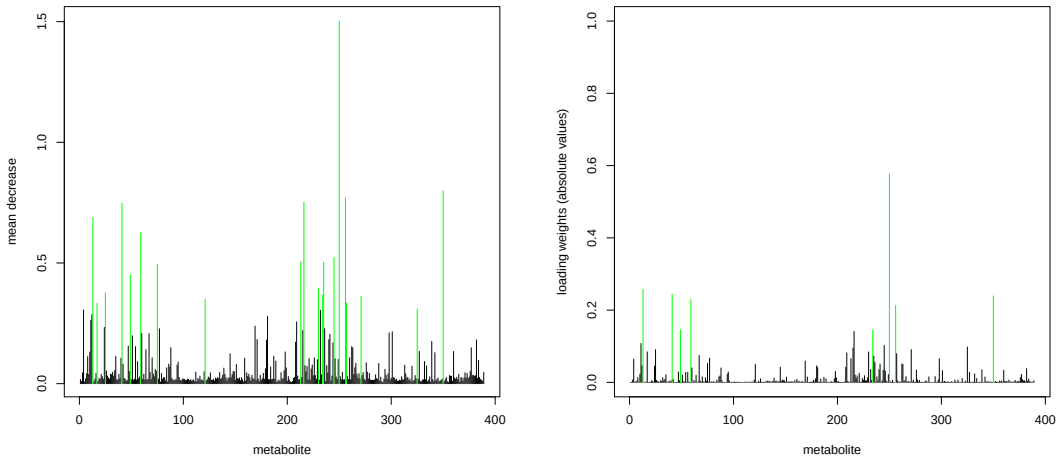


Figure 3.9: Mean PEs and corresponding 95% confidence intervals of PPLS-DA for up to 40 metabolites used.



(a) (b)

Figure 3.10: Averaged importance value for the metabolites according to RF (a) and according PPLS-DA (b), the green lines belong to the proposed biomarkers.

Comparing the ranking lists more in detail, Table 3.3 opposes the ranking lists based on RF for the first 19 metabolites and for PPLS-DA for the first 8, based on the importance values shown in Figure 3.10.

Table 3.3: Biomarker determined with respect to RF and PPLS-DA

	position											
method	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
RF	250	350	256	216	41	13	59	245	213	235	75	49
PPLS-DA	250	13	41	350	59	256	49	234				
	position											
method	**13**	**14**	**15**	**16**	**17**	**18**	**19**					
RF	230	25	234	271	121	257	17					

Regarding the first 3 metabolites of the ranking lists, metabolite 250 is in both lists on the top, the next 3 positions differ. All 8 metabolites of the biomarker detected by PPLS-DA can also be found in the biomarker detected by RF. Although both importance values base on different approaches, the features on the top positions of the ranking lists almost coincide, only the positions differ.

Summarizing the results, a candidate biomarker set is determined according to RF with 19 metabolites for the discrimination between TB and TST+ persons. For PPLS-DA a candidate biomarker set is created containing only 8 metabolites with a similar PE than for RF. The ranking lists do not differ much on the top positions for PPLS-DA and RF for this metabolite data set. This leads to the assumption, that the top 8 metabolites of PPLS-DA are robust with respect to a different importance measure and build a promising biomarker, which has to be validated further.

In this chapter a biomarker is identified, first different classification methods are compared to choose the one which fits best to the data. Then the variable importance value of this classification method is applied to create a ranking list building the basis for the determination of a biomarker. These main steps are further considered in dependency on the choice of the experimental design in Chapter 5.

4 Description of microarray data

This chapter describes the simulated data and the publicly available experimental data sets which are used in the following chapters.
First covariance measures are explained then the simulation study and the experimental data sets are introduced and the covariance structure is studied.

Covariance structure

For detailed description of the covariance structure of the data, two measures are applied analogous to Saebø *et al.* (2008). These are the condition index, first used in Belsley *et al.* (1980), and the absolute value of the covariances between the principal components and the response vector as used in Helland and Almøy (1994). The condition index $\kappa_k = \sqrt{\lambda_1/\lambda_k}$, $k = 1, \ldots, g$ is a measure for the linear dependence between the features, with λ_k being the k^{th} eigenvalue of $cov(\mathbf{X})$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_g$. The increase of the first five condition indexes $(\kappa_1, \kappa_2, \kappa_3, \kappa_4, \kappa_5)$ reflects the collinearity of the features (Saebø *et al.*, 2008). A rapid increase of the first five condition indexes indicates, that the features have a strong linear dependence, a weak increase implies a weak dependence. Considering now the principal components of $\mathbf{X}$, like in Saebø *et al.* (2008), the relevance of a component is measured by means of the absolute value of the covariances $(|cov(\mathbf{z_k}, \mathbf{y})|)$ between the

This chapter describes the simulated data and the experimental data sets used in Telaar *et al.* (2010), Telaar *et al.* (2012a), Telaar *et al.* (2012b). Therefore this chapter bases on these publications.

principal component $\mathbf{z_k} = \mathbf{X}\,\mathbf{e_k}$ and the class vector $\mathbf{y}$. Here y_i equals 1 if sample i belongs to group $\nu = 1$, otherwise y_i equals -1, $i = 1, \ldots, n$. The eigenvector belonging to the k^{th} largest eigenvalue is denoted by $\mathbf{e}_k$. Helland and Almøy (1994) infer, that data sets with relevant components which have small eigenvalues are difficult to predict.

4.1 Gene expression data simulation

Gene expression data are often measured by means of microarrays (Malone and Oliver, 2011). For microarray data two data scales are considered, the original scale (measured intensities) and the log-scale (after normalization). The values for a gene k, $k = 1, \ldots, g$ on the scale of measured intensities can be understood as realizations u_k of a random variable. For the data analysis during the normalization process the data are transformed to the log scale (Irizarry *et al.*, 2003), here normally distributed random variables $log(u_k)$ are assumed (Quackenbush, 2002). Then the gene expression can be modeled on the log scale as $log(u_{ik}) = \mu_k + \varepsilon_{b_i} + \varepsilon_{t_i}, i = 1, \ldots, n$ (n number of single samples) with mean gene expression level μ_k, biological variation ε_{b_i} and technical variation ε_{t_i} which are independently, identically and normally distributed with mean zero and biological variance σ_b^2 and technical variance σ_t^2 respectively (Zhang *et al.*, 2007). On the log scale the gene expression values are denoted by $X_{ik} = log(u_{ik})$. Therefore on the original scale the random variables are lognormally distributed with mean value $e^{\mu_k + \frac{\sigma_b^2 + \sigma_t^2}{2}}$ and variance $e^{2\mu_k + \sigma_b^2 + \sigma_t^2} \cdot (e^{\sigma_b^2 + \sigma_t^2 - 1})$.

Considering a two group classification problem, 60 single samples are simulated per class ($n_1 = n_2 = n/2 = 60$) and 1000 genes partitioned in an informative part and a non-informative part for the classification, because some genes show group information in their expression data and others do not.

The non-informative part of the data matrix consists of normally distributed random variables with mean $\mu_k = 8$, a biological variance $\sigma_b^2 = 0.04$ and a technical variance σ_t^2. As mean value $\mu_k = 8$ is chosen, because it is usually the modal value of experimental gene expression data and σ_b^2 is set equal to 0.04, which occurred in analyses of microarray data (Rudolf, 2011). For the technical noise, three levels are investigated $\sigma_t^2 \in \{0, 1/4\sigma_b^2, \sigma_b^2\}$.

Four scenarios for the informative part of the data matrix are simulated, which are illustrated in Figure 4.1:

(a) Scenario 1: Differentially expressed features with a mean class difference of Δ, $\Delta = \{[0.1, 0.5], 0.2, 0.5\}$ (Figure 4.1(a)).

(b) Scenario 2: A pattern by threshold, one group with values inside a defined interval and the other group with values lying outside the interval. However, both classes have the same mean value (Figure 4.1(b)).

(c) Scenario 3: Each two-dimensional linear pattern with of two linear dependent features (Figure 4.1(c)).

(d) Scenario 4: Each two-dimensional "circular" pattern with the values of one group inside a circle and the values of the other group outside the circle (Figure 4.1(d)).

The scenarios are examples for possible biomarkers which could occur in real data. Scenario 1 refers to the well known type of differentially expressed genes, where for example the class of disease subjects has a higher mean value of gene expression for a gene than the group of healthy subjects.

A feature of the form of scenario 2 can also be a biomarker: the gene expression for the non-infected subjects lies in a special reference range and the gene expression for the infected subjects lies outside the range. An example (not for gene expression biomarker but in general) is also the HbA_{1c} (see Chapter 1): Patients with a

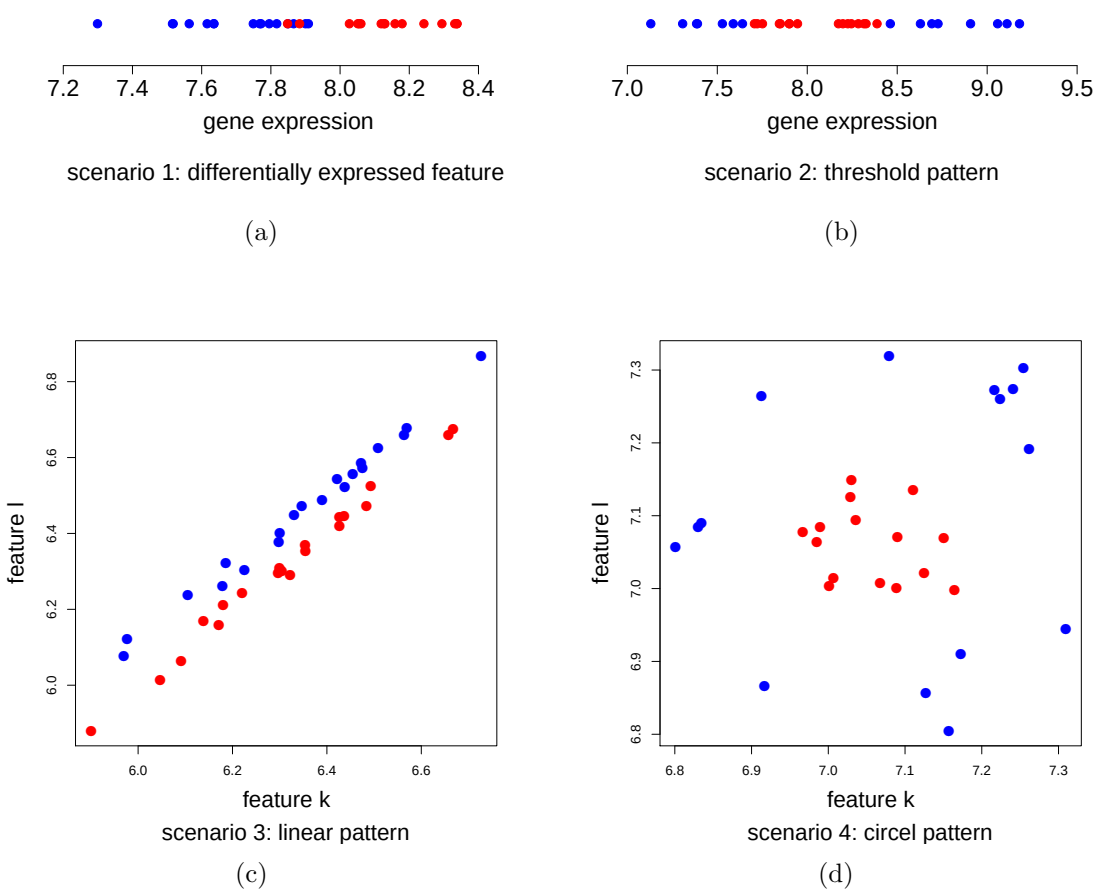


Figure 4.1: Simulated pattern scenarios for the gene expression data with $\sigma_t^2 = 0$. Upper part the two one-dimensional patterns: (a) scenario 1 and (b) scenario 2, lower part the two 2−dimensional patterns: (c) scenario 3 and (d) scenario 4.

HbA_{1c} in a reference interval are well threated, if the HbA_{1c} lies outside the interval, the patient needs to be re-stabilized. Scenario 4 is a specialized two-dimensional form of scenario 2, one group with a clearly lower variance than the other group. Two linearly dependent features (scenario 3) represent a set of correlated features. Scenario 3 is an example, where each gene of the pattern individually considered, cannot clearly separate the groups, but the combination of these two genes leads to a separation of the groups.

On the one hand, the two classes are linearly separable if the underlying data is of the form of scenario 1 or 3, and therefore easier to classify with linear classification methods like t-LDA, PLS-DA, PPLS-DA and SVML. On the other hand, scenario 2 and 4 are more complex patterns which can not be separated in a linear way.

The separation of the two classes of a circular pattern, as in scenario 4, is already mentioned in Russel and Norvig (2009).

The features belonging to the informative part are denoted as informative simulated features (ISF). The informative part is simulated in different ratios to the non-informative part by choosing proportions of $1\%, 10\%$ and 20% ISF. The different proportions represent a small, middle and high information content for the separation of the groups.

Simulation of scenario 1 The differentially expressed features of scenario 1 are simulated with a mean class difference $\delta \in \Delta$, $\Delta = \{[0.1, 0.5], 0.2, 0.5\}$.
For a gene k, the mean $\mu_{(1)k}$ of group 1 is randomly chosen according to the uniform distribution from the interval $[6, 10]$. The gene expression values which belong to the subjects of group 1 are chosen from $N(\mu_{(1)k}, \sigma_b^2 + \sigma_t^2)$. The gene expression values of the group 2 individuals are drawn from $N(\mu_{(1)k} + \delta, \sigma_b^2 + \sigma_t^2)$.

Five cases are taken into account for the value combination of δ and σ_t^2. For case 1, 2 and 3, δ is chosen according to the uniform distribution from the interval $[0.1, 0.5]$ and only the technical variance differs. Case 1 has a technical variance of zero, case 2 of one-quarter of the biological variance ($\sigma_t^2 = \frac{1}{4}\sigma_b^2$), and case 3 is simulated with a technical variance of the same size as the biological variance ($\sigma_t^2 = \sigma_b^2$). The differentially expressed features of case 4 have a mean class difference $\delta = 0.2$ and $\sigma_t^2 = \sigma_b^2$. The simulated data of case 5 also have the large technical variation ($\sigma_t^2 = \sigma_b^2$), and a higher mean class difference $\delta = 0.5$.
Considering the data structure of case 3 as example, the condition indexes are shown in Figure 4.2. The first 5 condition indexes are 1.00 1.01 1.04 1.05 1.06. As expected, this increase is the weakest for all data sets considered, because of the described simulation procedure. Figure 4.3(a) illustrates also the weak linear dependency among the features, but there are irrelevant components (low absolute covariance) with large eigenvalues. This can impair the prediction using (P)PLS-DA. For all

other cases also a weak dependence is found (data not shown).

Simulation of scenario 2 For a feature k of scenario 2 (the threshold pattern) $n = n_1 + n_2$ single sample values are drawn from normally distributed random variables with variance $\sigma_b^2 + \sigma_t^2$ and a mean μ_k randomly chosen (according to the uniform distribution) from the interval $[6, 10]$ for each feature. After ordering these random numbers, the first $\frac{1}{2}n_1$ values are assigned to group 1, the next n_2 to group 2 and the last $\frac{1}{2}n_1$ belong again to group 1.
Accounting for the condition index, as example for 1% and $\sigma_t^2 = 0$, the sequence is 1.00, 1.13, 1.15, 1.15, 1.16 and again a weak increase is found. Also the eigenvalue plot (data not shown) supports this.

Simulation of scenario 3 Without loss of generality, let k and l be the features which build the pattern of scenario 3. Then $X_{ik(1)}$ for sample i in group 1, and $X_{jk(2)}$ for sample j in group 2, $i \neq j$ are simulated as normally distributed random variables with mean $\mu_k = c$ and variance $\sigma_b^2 + \sigma_t^2$. The value c is chosen randomly (according to the uniform distribution) from the interval [6,10]. The second feature, l, is calculated as follows $X_{il(1)} = X_{ik(1)} + \varepsilon_i$ for group 1 and $X_{jl(2)} = X_{jk(2)} + \delta + \varepsilon_j$ for group 2, with $\delta = 0.1$ and $N(0, 0.01^2)$-distributed ε_i and ε_j. These two features (l, k) are linearly dependent and therefore strongly correlated. For every pair of features (l, k) a new mean value and new random variables are chosen.
Again the dependence among the features is considered for $\sigma_t^2 = 0$ and with 1% ISF, the first five condition indexes are : 1.00, 1.08, 1.10, 1.11, 1.12 (Figure 4.2). Analog to scenario 2, the features show a weak linear dependence which is also illustrated by Figure 4.3(b), but many components are found with large eigenvalues in combination with low covariances. Therefor the prediction using (P)PLS-DA is even more difficult than for scenario 1 where this combination occurres rarely (see Figure 4.2).

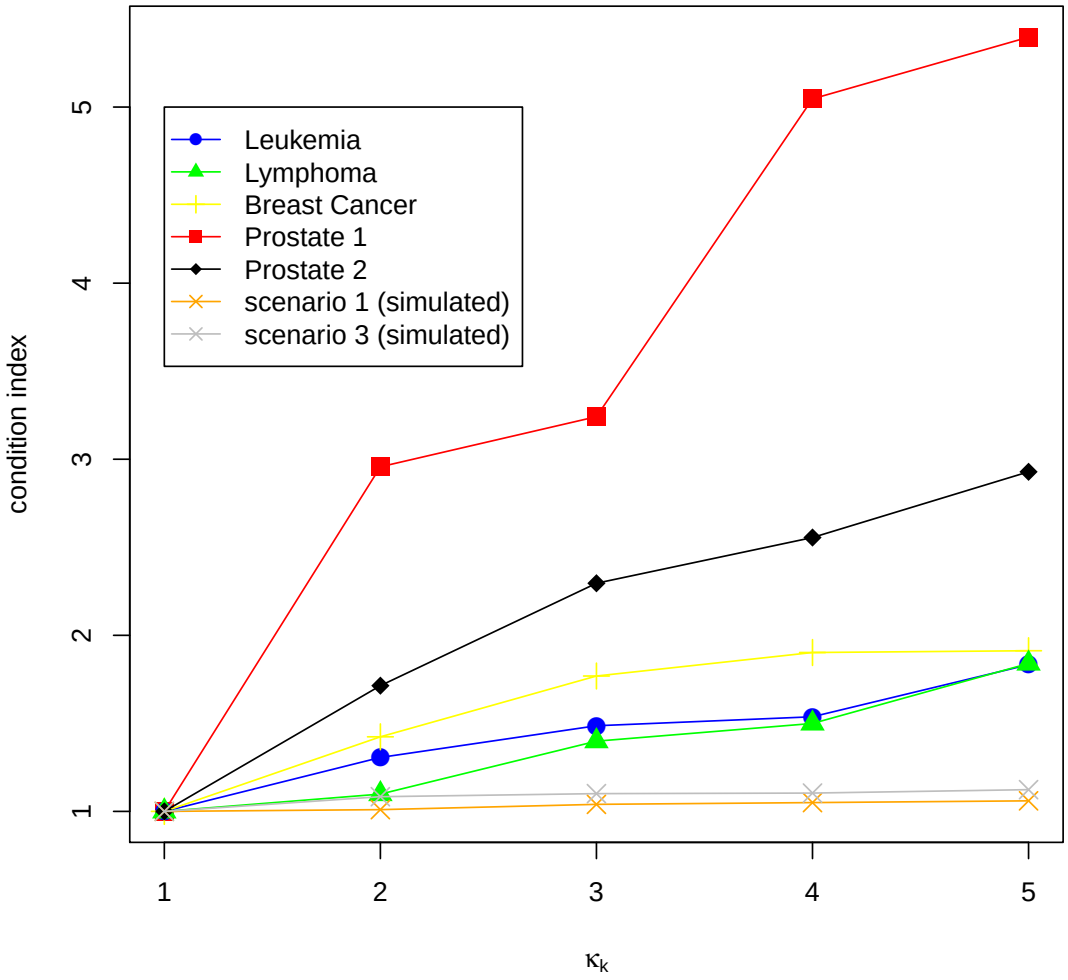
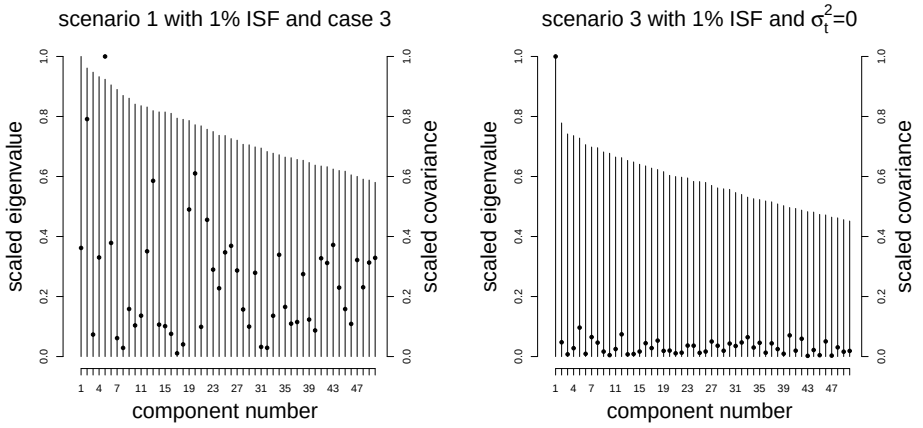


Figure 4.2: Condition index κ_k for the first five eigenvalues of the 5 experimental data sets and for the simulated data for two scenarios as example (scenario 1 with 1% ISF for case 3 and scenario 3 with 1% ISF and $\sigma_t^2 = 0$).

Simulation of scenario 4 The circle pattern, scenario 4, is simulated for features k and l as follows for group 1: The value d_i is uniformly drawn from the interval $[0, 360]$ for each sample i of group 1 and $X_{ik(1)}$ and $X_{il(1)}$ are calculated as $X_{ik(1)} = r_1 cos(d_i) + \varepsilon_{i_k}$ and $X_{il(1)} = r_1 sin(d_i) + \varepsilon_{i_l}$ with $r_1 = 0.1$ and $\varepsilon_{i_t} \sim N(0, 0.0004)$, $t = k, l$. Analogously for class 2 the gene expression is simulated for the genes k and l with $r_2 = 0.25$. With a randomly chosen value m according to the uniform distribution from the interval $[6, 10]$, the values $X_{.k}$ and $X_{.l}$ are shifted by adding m. Then the mean values of the features k and l are m. The variances of group 1



(a) (b)

Figure 4.3: Plot of the first 50 largest eigenvalues $\mathbf{e}_k$ of cov($\mathbf{X}$) (bars) and of the absolute covariance between $\mathbf{z}_k$ and $\mathbf{y}$ (dots) for the simulated data.

and group 2 differ. For gene k and l the variance is $\frac{r_1^2}{2} + \sigma_\varepsilon^2 = 0.0054$ for group 1 and $\frac{r_2^2}{2} + \sigma_\varepsilon^2 = 0.0316$ for group 2 (see Appendix B).
As well for this scenario with $\sigma_t^2 = 0$ and 1% ISF the first five condition indexes show a weak increase (1.00, 1.06, 1.07, 1.08, 1.09).

Summarizing, the patterns of the simulation study have attributes concerning the convex hull (set of all convex linear combinations). All four simulated patterns can be divided into two types, for scenario 1 and 3, the convex hull of the one class is not a cover of the other class (pattern type I) and for scenario 2 and 4, the convex hull of the one class is a cover of the other class (pattern type II). Moreover the simulated data sets show for each scenario a weak linear dependence among the features.

For the training and test set of the simulated data, 30 single samples per group are randomly chosen for the training set and the remaining 30 single samples per group build the test set. Therefore for the simulated data $\mathbf{X}_{train}$ and $\mathbf{X}_{test}$ are 60×1000 matrices with 60 samples and 1000 genes.

4.2 Experimental data

Additionally, five publicly available experimental microarray data sets are used in this thesis. The experimental data sets are summarized in Table 4.1 containing informations about the group size, number of genes, proportion of differentially expressed genes and original publication. For the determination of the number of differentially expressed genes (N_{DEGs}) a t-test and an FDR correction is used (see Section 2.3.2). Only genes are accepted as differentially expressed with a q-value below 0.05 (Storey, 2003) and differentially expressed genes are abbreviated with DEGs.

Analogous to the description of the covariance structure of the simulated data, the condition indexes are illustrated for the first five largest eigenvalues (scaled to the first eigenvalue) in Figure 4.2 and the plot of the first 50 largest scaled eigenvalues and the corresponding scaled covariances between $\mathbf{z_k}$ and $\mathbf{y}$ (Figure 4.4).

Table 4.1: Overview of the experimental data sets

name	n_1/n_2	g	N_{DEGs}	N_{DEGs} in %	original publication
Leukemia	47/25	3571	1445	40.46	Golub *et al.* (1999)
Lymphoma	58/19	7129	1739	24.39	Shipp *et al.* (2002)
Breast Cancer	44/34	4997	54	1.08	van't Veer (2002)
Prostate 1	50/52	6033	2393	35.26	Singh *et al.* (2002)
Prostate 2	41/62	42129	595	1.40	Lapointe *et al.* (2004)

Leukemia

The Leukemia data were downloaded from the Whitehead Institute website. The training set and the test set are merged to get a higher sample size and sample from these are drawn to get new proportions 0.7 and 0.3 for the training and test set. The R code for data preprocessing from http://svitsrv25.epfl.ch/R-doc/library/multtest/doc/golub.R is used, which was developed according to Dudoit *et al.* (2002). The data set consists of two groups, 25 patients with acute myeloid leukemia and 47 patients with acute lymphoblastic leukemia and the gene expression

values for 3571 genes.

The condition indexes show a weak increase for this data set (1.00, 1.31, 1.49, 1.537, 1.83). This and the plot of the eigenvalues (Figure 4.4(a)) lead to the assumption of a weak linear dependency between the genes, but not so weak like for scenario 1 of the simulated data where the genes are simulated independently from each other. For the Leukemia data set, the more relevant components have the largest eigenvalues (Figure 4.4(a)) which implies potential for good prediction performance of (P)PLS-DA . Moreover, this data set has the highest proportion of DEGs (40.46%, Table 4.1).

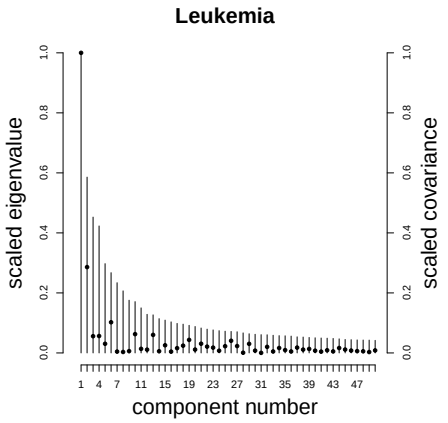
Lymphoma

This data set was downloaded from the website http://www.broadinstitute.org/mpr/lymphoma/. The data are GC-RMA (robust multi-array average) normalized. Two groups are considered, 58 patients with diffuse large B-cell lymphomas and 19 patients with B-cell lymphoma, follicular lymphoma. Only genes with a non-zero variance are used in our analysis, which leads to 7129 genes.

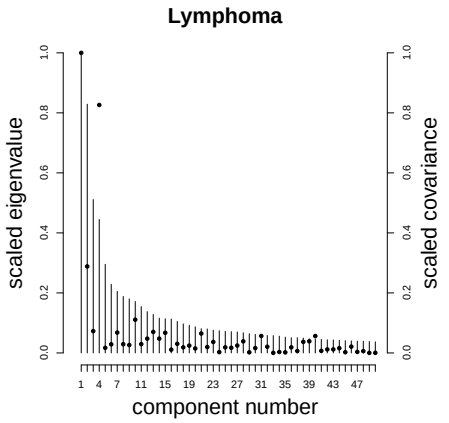
The between-variable dependencies are comparable to the Leukemia data set (condition indexes: 1.00, 1.10, 1.40, 1.50, 1.84). The covariance structure (Figure 4.4(b)) is also comparable to those of the Leukemia data set and the total number of DEGs is marginally higher than for the Leukemia data set, but the proportion on the total number of genes is clearly lower (24.3%, Table 4.1).

Breast Cancer

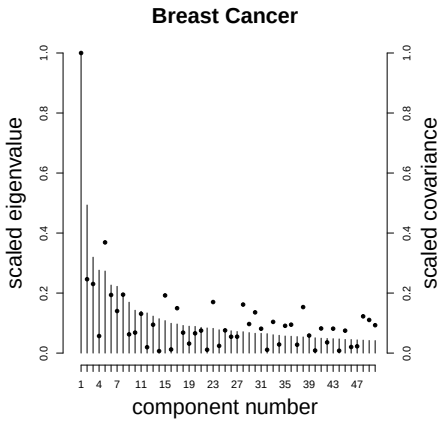
This normalized and filtered data set was downloaded from http://homes.dsi.unimi.it/~valenti/DATA/MICROARRAY-DATA/R-code/Do-Veer-data.R. The normalization was performed according to van't Veer (2002). In this data set, only the two groups with the highest sample size are included: 34 patients with distant metastases within 5 years and 44 patients without, after at least 5 years. The total number



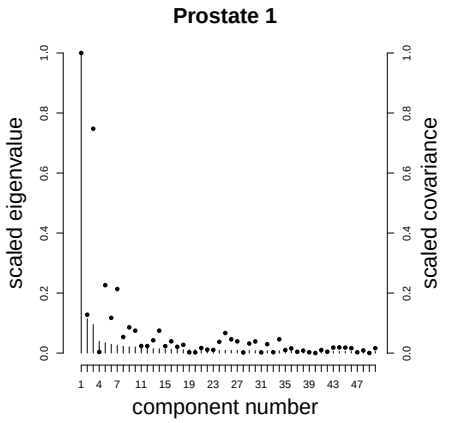
(a)



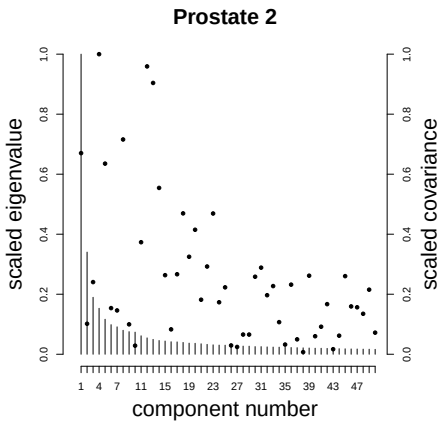
(b)



(c)



(d)



(e)

Figure 4.4: Plot of the first 50 largest eigenvalues $\mathbf{e}_k$ of $\mathrm{cov}(\mathbf{X})$ (bars) and of the absolute covariance between $\mathbf{z}_k$ and $\mathbf{y}$ (dots) for the experimental data sets.

of genes is 4997.

The condition indexes increases weakly for the first five eigenvalues (1.00, 1.42, 1.77, 1.90 and 1.91), but slightly faster than for the Leukemia and Lymphoma data set (Figure 4.2). The eigenvalue plot (Figure 4.4(c)) illustrates also a weak linear dependence between the features and many components which low eigenvalues and large covariance which indicates a difficult prediction using P(PLS-DA). The proportion of DEGs is the lowest for all experimental data sets (1.08%, see Table 4.1).

Prostate 1

This data set contains 52 tumor and 50 non-tumor cases and was downloaded from http://stat.ethz.ch/~dettling/bagboost.html. The preprocessing is described in Dettling and Buehlmann (2003) and the final data set contains 6033 genes.

This data set shows a rapid increase of the condition index from κ_1 to κ_5 (1.00, 2.96, 3.24, 5.046, 5.397), describing a strong linear dependency of the genes (Figure 4.2). This property is also indicated by the plot of the eigenvectors (Figure 4.4(d)) and most of the components with low eigenvalues have also a low covariance. This data set has a high proportion of DEGs (32.26%, Table 4.1).

Prostate 2

The normalized data set was downloaded, from http://bioinformatics.mdanderson.org/TailRank/. A description of the normalization can be found at http://bioinformatics.mdanderson.org/TailRank/tolstoy-new.pdf. In this data set, only the two groups with 41 patients with normal prostate tissue and the 62 patients with primary tumors are included.

The condition index shows a rapid increase (1.00, 1.71, 2.10, 2.56, 2.93) for the first five eigenvalues (Figure 4.2), but more moderate than for the Prostate 1 data set. Figure 4.4(e) illustrates that also relevant components have small eigenvalues which makes the prediction using P(PLS-DA) more difficult. The proportion of DEGs is

very low (1.4%, Table 4.1) and similar to those of the Breast Cancer data set, but the total number of genes is the largest (42129) for all experimental data sets.

5 Pooling design for biomarker search

This Chapter concerns the influence of a pooling design used for biomarker search in comparison to a single sample design and is based on work published in Telaar *et al.* (2010) and Telaar *et al.* (2012a).

5.1 Motivation

Pooling, the combination of single samples, is a common procedure in microarray technology, because the biological variation is reduced and less arrays are needed in comparison to a single sample design. If fewer arrays are needed, the financial cost are reduced with respect to the acquisition and to the sample preparation. Furthermore if not enough cDNA for the hybridization step for the microarray experiment exists, pooling is a solution to accomplish the experiment anyway (Simon *et al.*, 2002).

In the literature, pooling is studied in detail only for searching for differentially expressed genes. In a simulation study, Peng *et al.* (2003) opined that "pooling biological samples appropriately is statistically valid" and cost-effective for microarray experiments if group level differences are the objective. Kendziorski *et al.* (2005) concluded that pooling is advantageous for designs with only one or two arrays per group (with the aim to identify differentially expressed genes), but not for designs

with a large number of arrays. The drawbacks of pooling are discussed e.g. in the technical note of Affymetrix (2004). For example individual sample information like age and gender get lost if individuals in the pools are not perfectly matched. Moreover outliers or misclassified samples cannot be identified after pooling. It is known that sample pooling leads to biases. Mary-Huard *et al.* (2007) divide this bias into log bias, caused by the log transformation, and pooling bias which reflects the different proportions of the samples on a pool.

In the literature only a few statements are found concerning pooling and biomarker search. Kerr (2003) stated that the design of a classification study, like for biomarker search, should not consist of pooled samples, because data is required at the "individual level". Allison *et al.* (2006) point out more precisely why pooling should be avoided for biomarker search: "pooling interferes with the ability to accurately assess inter-individual variation and covariation". This is based on the fact that variance and covariance structure is changed by pooling. Moreover, Sadiq and Agranoff (2008) mentioned that the use of pooled samples may lead to a loss of features which ca build a biomarker. From the statistical point of view, pooling should be avoided for biomarker search. Nevertheless, this aspect of design choice is not widely known, and sometimes the circumstances do not allow to follow the advise. It is common practice to use pooled samples to search for biomarkers in animal experiments, like Searfoss *et al.* (2003), who pooled samples of rat ileum tissue for a microarray analysis. Jacobsen *et al.* (2007) also apply pooled samples for finding possible biomarkers discriminating patients infected with tuberculosis but healthy and those who suffer under tuberculosis.

In the literature no clear statements have been found concerning the influence on the validity of the experiment if a pooling design is used for biomarker search. Therefore a study was started to allow statements about the consequences of a pooling design in comparison to a single sample design for classification.

Now the total number of (available) single samples is denoted by n_{total}, the number of samples used for the single sample arrays by n_S and the number of single samples used for the pools by n_{S_P} (clearly $n_{total} \geq n_S, n_{S_P}$). The number of arrays for the single sample design is denoted by a_S and the number of arrays for the pooling design by a_P. The total number of arrays which can be financed is called a_{total}.

Comparing a single sample design and a pooling designs, two possible concepts are distinguished. Concept I is the comparison of a single sample design with a pooling design in which the number of single sample arrays is higher than the number of pooling arrays ($a_S > a_P$), but the number of single samples used is equal ($n_S = n_{S_P}$). The main focus is on concept I because this case frequently occurs in practice. The pooling design is often chosen because microarray experiments are cost-intensive (especially the labeling and amplification steps). Therefore the number of single samples is often higher than the number of arrays which can be financed ($n_{total} > a_{total}$), but the aim is to use as many single samples as possible.

Concept II is the comparison of a single sample design with a pooling design with the same number of arrays ($a_S = a_P$). That means $n_S < n_{S_P}$, not all single samples used for the pools are measured in the single sample design. Concept II is analyzed to study the influence of reduced sample size.

In this Chapter, at first some theoretical basics of pooling are described. For analyzing the effect of pooled samples on biomarker search, a simulation study is carried out to study the influence on the PE of different patterns of informative features, the proportion of informative features and different technical variation levels for concept I and concept II. Moreover, for concept I sets of important features (possible biomarkers) are compared detected by the pooling design and the single sample design. Furthermore, four experimental data sets are pooled artificially and analyzed regarding concept I and concept II.

5.2 Pooling design

5.2.1 Statistical modeling of pooling

Pooling, the combination of several RNA samples, takes place before measuring gene expression of these RNA samples. Therefore, first the gene expression of the single samples are considered without technical variation, $u_{ik(b)} = \mu_k + \varepsilon_{b_i}$ denotes the gene expression of sample $i, i = 1, \ldots, n_{S_P}$ for gene k without technical variation with μ_k as mean value of gene k and biological variation ε_{b_i}, analogous to the notation in Section 4.1. The gene expression $u_{p_j k}$ of a pool $p_j, 1 \leq j \leq n_P$, is a convex linear combination of the single samples which form the pool, weighted with an individual proportion of these single samples:

$$u_{p_j k} = \sum_{i=m_p(j-1)+1}^{m_p \cdot j} \omega_i u_{ik(b)} + \varepsilon_{t_{p_j}},$$

where $1 \leq m_p \leq n_{S_P}$ is the number of mixture components within a pool and $n_P = \frac{n_{S_P}}{m_p}$ is the total number of pools. The objects $1, \ldots, n_{S_P}$ are assumed to be ordered in a sequence how they are used to from the pools. Here ω_i denotes the proportion of the i-th sample in a pool with $\sum_{i=m_p(j-1)+1}^{m_p \cdot j} \omega_i = 1$. The technical variation of pool p_j is denoted by $\varepsilon_{t_{p_j}}$, which is independently, identically and normally distributed with mean zero and technical variance σ_t^2. The deviation of weight ω_i from $\frac{1}{m_p}$ is the pooling bias, caused by different proportions of the single samples on a pool (Mary-Huard *et al.*, 2007). The size of a pool is determined by m_p. The pools are built on the original scale, afterwards the gene expression values of the pools are transformed to the log-scale. The Jensen's inequality for convex linear combinations of convex functions (Jensen, 1906) adapted to the pooling situation describes the log bias. The gene expression of a pool p_j on the log scale is then $log(u_{p_j k}) = X_{p_j k}$. The resulting bias is called log bias (Mary-Huard *et al.*, 2007). Therewith, assuming single samples drawn from a normal-distributed population with mean μ_k and variance $\sigma_b^2 + \sigma_t^2$, the corresponding pools underlie a normal-distributed population with a mean $\mu_k + \varepsilon_{log}$

(added to a term describing the log bias) and variance $\frac{\sigma_b^2}{m_p} + \sigma_t^2$. Pooling reduces the biological variance from σ_b^2 to $\frac{\sigma_b^2}{m_p}$. Here two main conditions are assumed for pooling designs:

(i) each individual sample contributes only to one pool and

(ii) the pool size m_p is the same for all pools, as recommended in the literature for pooling designs (Zhang *et al.*, 2007).

In this thesis only ideal pools are considered where each sample contributes with the same weight $\omega_i = \omega$, for $i = 1, \ldots, n_{S_P}$. Therewith the pooling bias is disregarded.

5.2.2 Comparison of concept I and concept II for a single sample design and a pooling design

To study the influence of a pooling design for biomarker search, it is necessary to compare the properties of pooling designs and single sample designs. Considering a classification task with two groups, the pools are randomly built inside the groups.

Figure 5.1 illustrates a single sample design (upper part) in comparison to a pooling design (bottom part) for one group in general. The single samples are separately analyzed, each on a single array. Resulting from the expression measurement, the data are considered on the original scale as mentioned above. After the log-transformation, the data are approximately normally distributed on the log-scale (Geller *et al.*, 2003). For the pooling design, first the samples are mixed to pools with a pool size m_p and then these pools are analyzed on an array.

Constant number of samples - concept I

In concept I, all single samples used for the single sample design contribute to the pools ($n_S = n_{S_P}$).

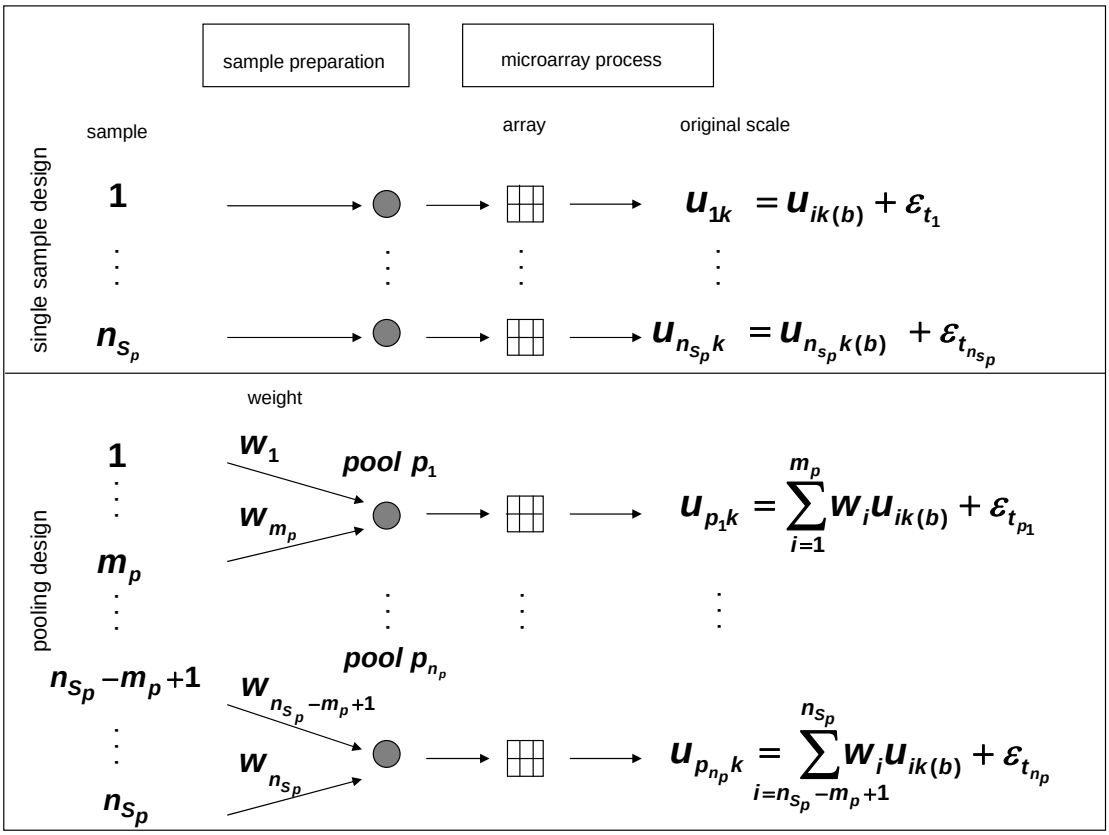


Figure 5.1: Single sample design and pooling design as example (simplified scheme).

Pooling with a pool size $m_p > 1$ therefore leads to a reduced number of arrays comparing a single sample design and a pooling design: $a_S > a_P$ because $a_S = n_S > \frac{n_{S_P}}{m_p} = a_P$. The different designs of concept I are chosen such that the statistical power of the t-test to detect differentially expressed features is nearly equal for all designs. The reduced sample size ($n_S > \frac{n_{S_P}}{m_p}$) for the pools is compensated by the diminished biological variance ($\frac{\sigma_b^2}{m_p}$) while all other conditions are kept constant (as for example the difference of the group mean and level of significance).

Constant number of arrays - concept II

The basis of concept II is to consider equal numbers of arrays for the single sample design and the pooling designs ($a_S = a_P$). It follows that $n_S < n_{S_P}$ for a pool size $m_p > 1$, because $n_S = a_S < a_P \cdot m_p = n_{S_P}$. This means more single samples contribute to the pooling design than to the single sample design for concept II.

Moreover, the statistical power of a t-test to detect differentially expressed genes for the pooling designs is higher than for the single sample design. This advantage grows with increasing pool size (because $\frac{\sigma_b^2}{m_p}$ is decreasing).

5.3 Simulation study

For the comparison of concept I and concept II with respect to the classification task, a simulation study is performed. The single sample design and the pooling designs are evaluated by the PE (prediction error, see Section 2.1.2) and possible biomarkers, using concept I or concept II.

5.3.1 Concept and implementation

At first the approach of the simulation study is introduced which describes the training and test set for the single sample design and the pooling design. Then the results are presented for concept I and concept II, first for data without technical variation and finally for data with technical variation.

Simulation of the pooling procedure

The data of the single sample design are simulated as described in Section 4.1. Because pooling takes place on the original scale, first the single sample data are simulated without technical variation. For the simulation of the pooling procedure, first the normally distributed data (the single samples) are transformed to the original scale (by taking the values as exponent to basis 2). Then, the pools are built gene-wise as the mean of the samples which form the pool. This procedure accounts for the group memberships and pool size $m_p = 2, 3, 5$. Finally, pool values are back-transformed to the log scale. Afterwards the technical variation is added to

the single sample data and the pooled data. Because pooling is most advantageous if the technical variance is smaller than the biological variance (Kendziorski *et al.*, 2003), the technical variation is chosen according to three levels $\sigma_t^2 \in \{0, 1/4\sigma_b^2, \sigma_b^2\}$ as described in Section 4.1.

Description of the training and the test set In this simulation study the total number of available samples is $n = 60 + 60 = 120$. For concept I, the training set of the single sample design consists of $n_S = n_{S_P} = 2 \cdot 30$ samples (30 per class). A design with pool size $m_p = 2(3, 5)$ leads to a sample size $n_P = 2 \cdot 15$ $(2 \cdot 10, 2 \cdot 6)$ of the training set. For concept II, the training set for the single sample design has $n_S = 2 \cdot 30, 2 \cdot 15, 2 \cdot 10, 2 \cdot 6$ samples. These samples are randomly chosen without replacement out of the corresponding $n_{S_P} = 2 \cdot 30$ single samples, which are the basis for the corresponding pooling designs. The same test set is used for all designs for both concepts, containing 30 samples per group which are single samples. This takes into account the practical application, because the biomarker should classify a new single object into a group. This is one important aspect for the comparison of a single sample design and a pooling design for biomarker search.
The simulation of the data was repeated 500 times for each parameter setting.

5.3.2 Results

To find out which method is most robust against pooling with respect to the PE, the six introduced classification methods (SVML, SVMR, RF, PPLS-DA, PLS-DA and t-LDA, see Chapter 2 for an introduction) are compared. At first the four scenarios, the two one-dimensional feature patterns (scenario 1 and scenario 2) and two two-dimensional feature patterns (scenario 3 and scenario 4) (see Section 4.1 for a detailed description), are analyzed without technical variation to study the raw pooling effect for concept I and for concept II. Furthermore, for concept I the sets

of features which are important for the classification using the methods PPLS-DA and RF are compared between the single sample and the pooling designs, because these sets can contain possible biomarkers. Additionally, the important features for PPLS-DA and PLS-DA are compared for scenario 1 without technical variation. All results presented for scenario 1 without technical variation used $\delta = [0.1, 0.5]$, which corresponds to case 1 (see Section 4.1).

Finally the results for simulated data with different choices of technical variation are shown for concept I and concept II exemplarily for scenario 1 with $\delta = [0.1, 0.5]$ and scenario 4.

Classification results using concept I for the simulated data without technical variation

Prediction error results for scenario 1 Figure 5.2 shows the mean PEs for case 1 of scenario 1 for the methods: SVML, SVMR, RF, PPLS-DA, PLS-DA and t-LDA in dependence of the designs with pool sizes $m_p = 1, 2, 3, 5$.

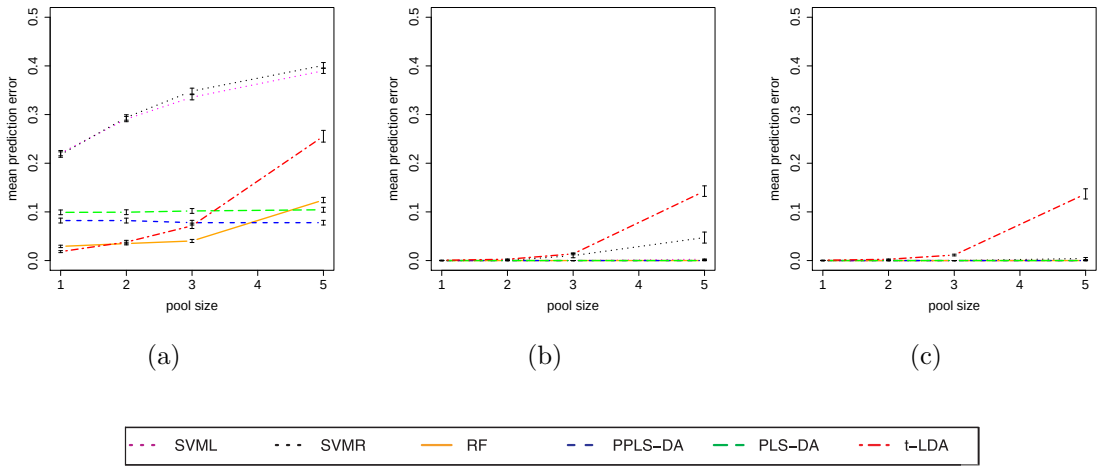


Figure 5.2: Mean PEs for scenario 1 for 1% ISF (a), 10% ISF (b) and 20% ISF (c) with 95% confidence intervals.

In Figure 5.2(a) it is presented that PPLS-DA and PLS-DA show a constant PE for all designs, for PPLS-DA the PE is around 0.08 and for PLS-DA around 0.1. All remaining methods show a clearly increasing PE with increasing pool size. RF and

$t-$LDA show a PE below 0.1 for a design with a pool size smaller or equal to 3, but a high increase of the PE from pool size 3 to 5. In comparison to the single sample design ($m_p = 1$), the PE of the design with pool size 5 is 12 times higher for t-LDA (0.25) and 8 times higher for RF (0.12). SVMR and SVML show similar and very large PEs for all designs, for the single sample design these PEs are around 0.22 and increase to 0.4 for a design with a pool size of $m_p = 5$.

The results for 10% ISF are similar to those of 20% (Figure 5.2(b) and 5.2(c)). The methods PPLS-DA, PLS-DA, RF and SVML show no pooling effect on the PE while the PE is nearly zero for all pool sizes. For 10%, a PE below 0.02 is found for t-LDA and SVMR for a pool size smaller or equal to 3, the PEs for pool size 5 increased to 0.15 for t-LDA and to 0.05 for SVMR. The method t-LDA shows nearly the same PE result for 20% ISF which was expected, because the t-LDA based only on 10 features. For SVMR and 20% ISF, now all pools sizes lead to a PE close to zero.

Prediction error results for scenario 2 For scenario 2 (the threshold pattern) without technical variation the PE results are shown in Figure 5.3 which are completely different from the results of scenario 1.

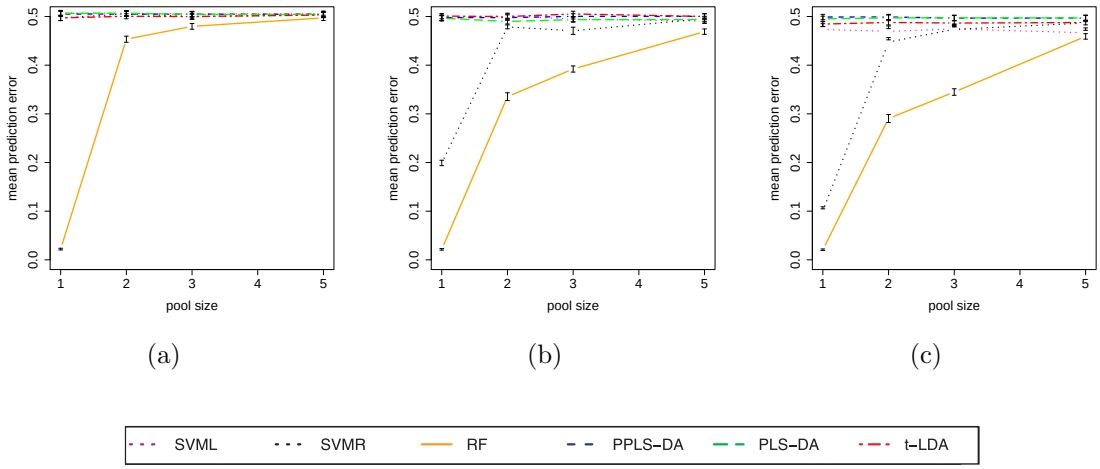


Figure 5.3: Mean PEs for scenario 2 for 1% ISF (a), 10% ISF (b) and 20% ISF (c) with 95% confidence intervals.

For 1% ISF the PEs lie around 0.5 for all methods, except for RF. For the latter,

the PE is around 0.02 for the single sample design and increases to between 0.45 and 0.49 for the pooling designs (Figure 5.3(a)).

For higher proportion of informative features with 10% and 20% ISF, only the methods RF and SVMR show a change in the PE (Figure 5.3(b) and Figure 5.3(c)). The method RF shows a decreasing PE for pool sizes 2 and 3 for an increasing proportion of informative features. The PE for SVMR decreased only for the single sample design for increasing proportion of ISF, from 0.5 for 1% ISF to 0.11 for 20% ISF.

Prediction error results for scenario 3 The PE results for scenario 3 (two linearly dependent features) are displayed in Figure 5.4. For this scenario, all methods show PEs between 0.4 and 0.5 for 1% ISF and no or only a minor pooling effect of a higher PE for an increasing pool size (Figure 5.4(a)).

For 10% ISF the PE results are clearly lower than for 1% ISF for all methods.

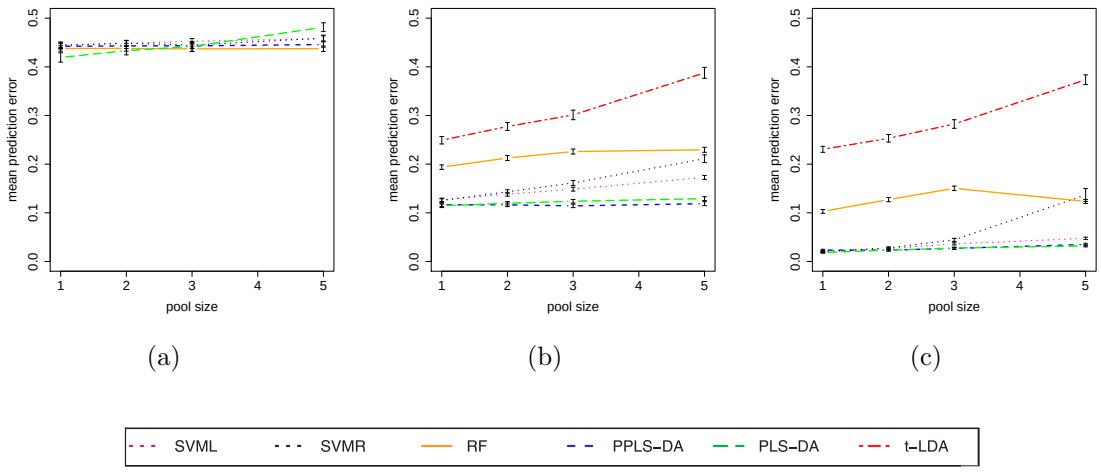


Figure 5.4: Mean PEs for scenario 3 for 1% ISF (a), 10% ISF (b) and 20% ISF (c) with 95% confidence intervals.

For PLS-DA and PPLS-DA the PEs are constant around 0.11 for all pool sizes m_p. Also SVML shows only a minor increase of the PE from 0.12 for the single sample design to 0.16 for the pooling design with pool size $m_p = 5$ (Figure 5.4(b)). This is similar for RF, but here with a generally higher PE around 0.2. SVMR and SVML

show a similar PE for the single sample design, but slightly different PEs for a pool size of $m_p = 5$. The PE of t-LDA extends from 0.12 for the single sample design to 0.38 for the design with pool size $m_p = 5$. For 20% ISF (Figure 5.4(c)), the PEs for all methods decrease and show similar shapes with increasing pool size compared to 10% ISF.

Prediction error results for scenario 4 Figure 5.5 shows the PE results of scenario 4 (the circle pattern) these results are very similar to those of scenario 2 (the threshold pattern).

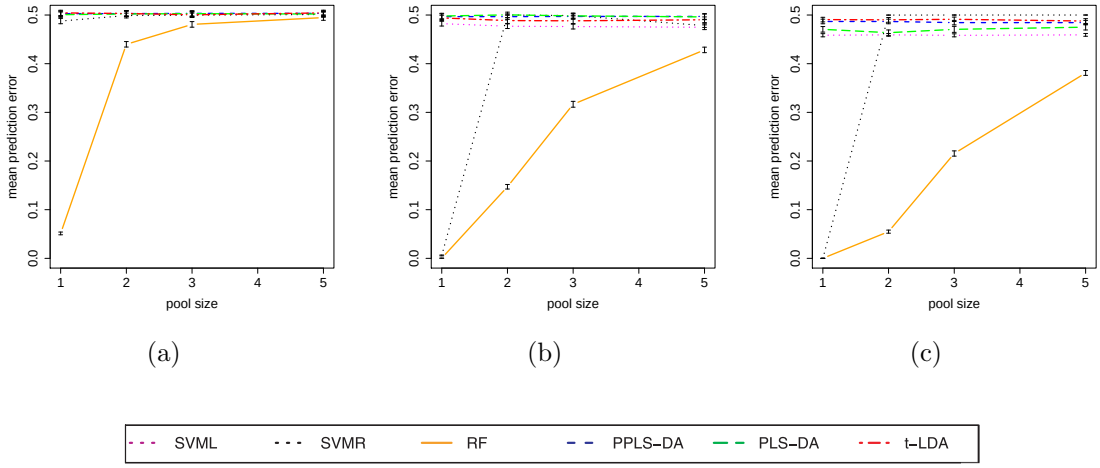


Figure 5.5: Mean PEs for scenario 4 for 1% ISF (a), 10% ISF (b) and 20% ISF (c) with 95% confidence intervals.

Considering 1% ISF, the PEs are between 0.48 and 0.5 for the methods SVMR, SVML, PPLS-DA, PLS-DA and t-LDA for all designs. The PE of RF is very low (around 0.05) for the single sample design $m_p = 1$, then the PE increases up to 8-times higher for the design with pool size $m_p = 2$. For a design with pool size $m_p = 3$, the PE of RF increases to 0.43, followed by a lower increase to 0.48 for the design with pool size $m_p = 5$.

For 10% versus 1% ISF the main differences are that the PE of SVMR is around 0.025 for the single sample design (for $m_p > 1$ the PEs are still over 0.48) and the PE of RF decreases further for designs with pool size $m_p > 1$. For 20% ISF the

PEs of SVML, SVMR, PPLS-DA, PLS-DA and t-LDA are similar to the results for 10% ISF for all designs. The PEs of RF decrease for all pool sizes $m_p = 2, 3, 5$ in comparison to 10% ISF.

Summarizing the results for the simulated data without technical variation for concept I, for pattern type I PPLS-DA and PLS-DA show a constant PE for increasing pool size. All other methods show an increasing PE for an increasing pool size especially for a low proportion of ISF. For pattern type II with 1% ISF, only RF has a low PE for the single sample design, but clearly higher PEs for the pooling designs. All other methods perform weaker with PEs around 0.5 for all designs. For increasing proportion of ISF, also SVMR shows a decreasing PE for the single sample design, but not for the pooling designs.

Comparison of important features declared of RF and PPLS-DA for concept I for the simulated data without technical variation

Now the ranking list of important features for the classification are considered which are the basis for the identification of a biomarker. On the top are features important for the discrimination between the groups. Here only the methods RF and PPLS-DA are investigated, because the PE results in the preceding Section 5.3.2 clearly shows that PPLS-DA and RF outperform the other methods. PPLS-DA shows almost no pooling effect (on the PE) for all scenarios and for patterns of type I additionally a low PE. RF indicates the lowest PE for all designs ($m_p = 1, 2, 3, 5$) for patterns of type II. The used importance value of RF is the mean decrease accuracy (see Section 2.3.2) and of PPLS-DA the absolute loading weights of the first component (see Section 2.3.2). An importance value for each feature and each scenario and proportion of ISF is calculated. Thus, a ranking list can be created to sort the features according to their importance for classification which can build the basis to compose a biomarker (see Section 2.3.2). Hence, if the term important feature

sets is used, a set of important features is indicated depending on the method used, the observed scenario and the proportion of ISF. For 1% (10%) ISF of the one-dimensional patterns, only the top 10 (100) features are investigated and for the two-dimensional patterns the top 20 (200) features. The set of important features for the classification with method M (from now on M denotes PPLS-DA or RF) is denoted by $\mathrm{D}_{m_p}^{\mathrm{M}}$ for a design with pool size $m_p = 1, 2, 3, 5$ and the set of ISF is denoted by $\mathrm{D}_{\mathrm{sim}}$. Only such features of $\mathrm{D}_{m_p}^{\mathrm{M}}$ which are simulated as informative are of interest, because only these are good possible biomarkers. Therefore, finally the intersections $\mathrm{I}_1^{\mathrm{M}} := \mathrm{D}_1^{\mathrm{M}} \cap \mathrm{D}_{\mathrm{sim}}$ are analyzed for the single sample design and the intersections $\mathrm{I}_{1:m_p}^{\mathrm{M}} := \mathrm{I}_1^{\mathrm{M}} \cap \mathrm{D}_{m_p}^{\mathrm{M}}$, $m_p = 2, 3, 5$ for the pooling designs. The cardinality of $\mathrm{I}_1^{\mathrm{M}}$ is the number of ISF which are important for the classification using the single sample design. Hence, the intersections $\mathrm{I}_{1:m_p}^{\mathrm{M}}$ consist of important ISF which are identical for a single sample design and a design with a pool size of m_p. Viewing the cardinalities of these intersections, the effects of pooling on the biomarker search are assessed in comparison to single samples. The cardinality of $\mathrm{I}_1^{\mathrm{M}}$ is taken as a reference value.

In the following, the mean cardinality values of the intersections $\mathrm{I}_{1:m_p}^{\mathrm{M}}$, $m_p = 1, 2, 3, 5$ of 500 repeated classification tasks are considered for the methods PPLS-DA and RF. Only the results for 1% and 10% ISF are shown, because in Section 5.3.2 the PE results turn out to be very similar for 10% and 20% ISF.

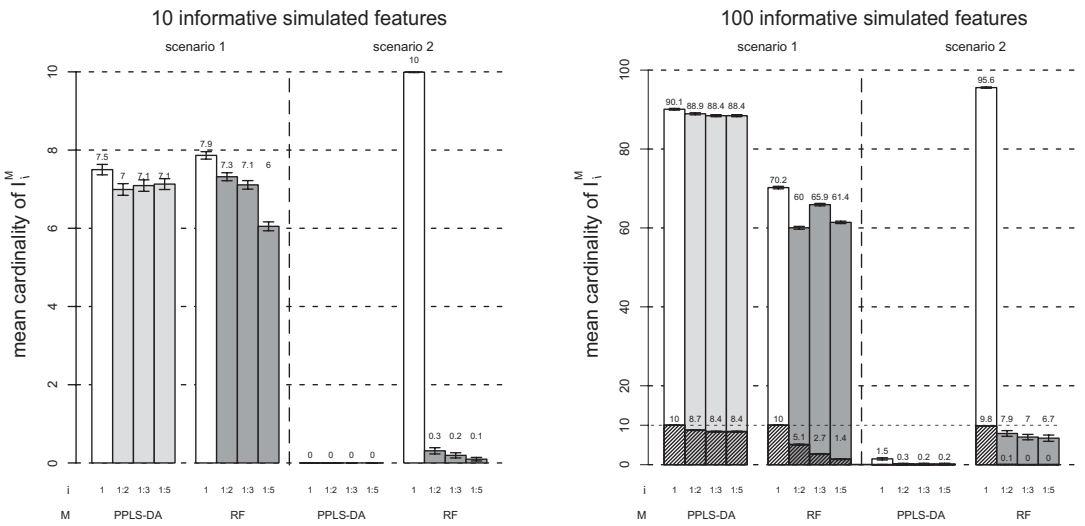
PPLS-DA and RF declared important feature sets for scenario 1 and scenario 2 with 1% ISF The results of the methods PPLS-DA and RF for scenario 1 and 2 are displayed in Figure 5.6. It contains the cardinalities of $\mathrm{I}_1^{\mathrm{M}}$, the set of important ISF for the single sample design and of $\mathrm{I}_{1:m_p}^{\mathrm{M}}$, $m_p = 2, 3, 5$, the intersections of this set with the set of important features for the pooling designs, for the methods PPLS-DA and RF.

For scenario 1, the cardinalities for PPLS-DA of $\mathrm{I}_{1:m_p}^{\text{PPLS-DA}}$ are nearly equal, with 7.1 for all pool sizes $m_p = 2, 3, 5$, while the cardinality for the single sample designs is

slightly higher with 7.5 (Figure 5.6(a)). This means, over 93% of the important ISF detected by PPLS-DA are equal for the single sample design and the pooling design. In comparison, for RF 7.9 ISF are detected as important for the single sample design and between 7.3 (for $m_p = 2$) and 6 (for $m_p = 5$) are identical with those of the important features for the pooling designs. This means 76% of the important ISF coincide between a pooling and a single sample design for pool size 5 and 92% for a pool size of 2.

For scenario 2, the threshold pattern, PPLS-DA detected no features which are informative simulated and important for the single sample design ($|\mathrm{I}_1^{\text{PPLS-DA}}| = 0$). Therewith, no important ISF could be found which coincide with the important features of the pooling designs. For RF, there are only few important ISF coinciding for the single sample and the pooling designs, even though all top ten important features for the single sample design are simulated as informative.
If now the results for scenario 1 and scenario 2 are compared, the cardinalities of the intersections $\mathrm{I}_{1:m_p}^{\mathrm{M}}$, $m_p = 2, 3, 5$, are lower for scenario 2 than for scenario 1 for both methods PPLS-DA and RF. For RF only for the single sample design the cardinality of $\mathrm{I}_1^{\mathrm{RF}}$ is larger for scenario 2 than for scenario 1.

The above results match to the PE results, for scenario 1. PPLS-DA shows nearly no influence of pooling on the PE (see Figure 5.2(a)) (nearly an equal number of features is detected as informative), but for RF the PE increases (the number of coinciding features decreases) for increasing pool size. As well for scenario 2, the results are consistent to the PE results (see Figure 5.3(a)). PPLS-DA shows a PE around 0.5 for all pooling designs ($m_p = 1, \ldots, 5$) and detects no informative simulated feature as important for the classification. RF results only for the single sample design in a low PE and also only for this design almost all ISF are identified as important.



(a) (b)

Figure 5.6: The mean cardinalities for the single sample design of I_1^M (white bars) and for the pooling designs $I_{1:m_p}^M$, $m_p = 2, 3, 5$ are shown for scenario 1 and scenario 2 for 1% ISF (a) and 10% ISF (b) for M=PPLS-DA, RF. The *first group* depicts results for the method PPLS-DA (with three light gray bars). The *second group* depicts those for RF (with three gray bars). The numbers above the bars are the mean values rounded to one decimal place. 95% confidence intervals are shown. The lower separated hatched bars in (b) represent the mean cardinality if only the top ten important features are accounted for in the feature sets.

PPLS-DA and RF declared important feature sets for scenario 1 and scenario 2 with 10% ISF Figure 5.6(b) displays the results for scenario 1 with 10% ISF (this means 100 ISF). Considering PPLS-DA for the single sample design, in average 90.1 features are declared as important and are simulated as informative. The numbers of ISF, which are important for the single sample design and the pooling designs, are only slightly lower with $|I_{1:2}^{\text{PPLS-DA}}| = 88.9$, $|I_{1:3}^{\text{PPLS-DA}}| = 88.4$ and $|I_{1:5}^{\text{PPLS-DA}}| = 88.4$. For RF, the corresponding cardinalities are lower (between 70.2 and 60) than for PPLS-DA for all pool sizes. Although both methods have similar average PEs for all different designs (see Figure 5.2(b)), for PPLS-DA more important ISF are equal between the single sample design and the pooling designs and no great dependency on the pool size is found.

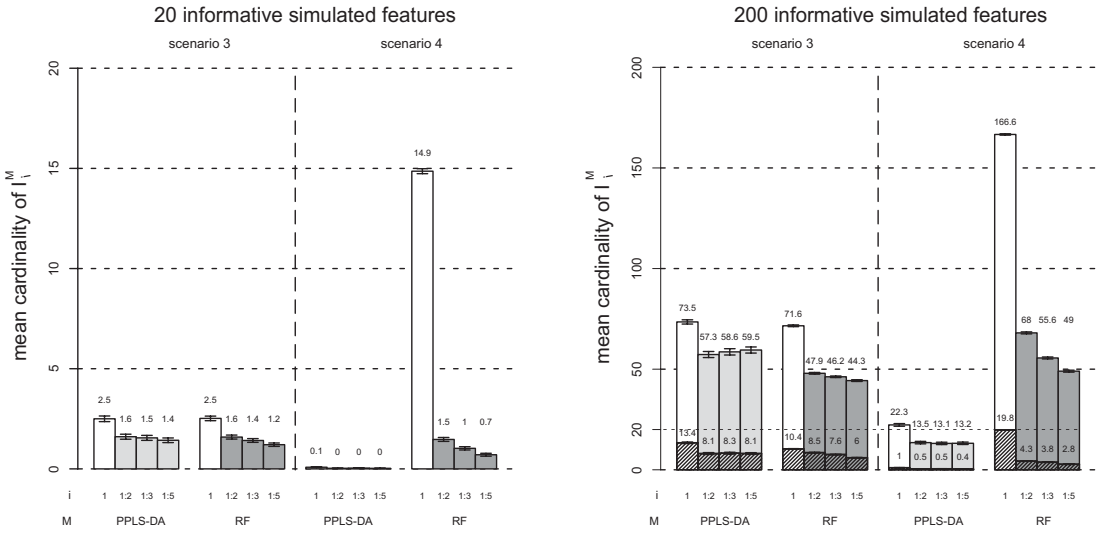
Comparing only the top ten important features for PPLS-DA for the single sample design with the ISF, ten of ten possible features are equal (Figure 5.6(b), separated hatched bars). Summarizing, the proportion of identical important ISF between pooling designs and the single sample design is between 98.6% ($m_p = 2$) and 98.1% ($m_p = 5$) for PPLS-DA and between 93.87% ($m_p = 3$) and 85.47% ($m_p = 2$) for RF. Investigating only the top ten important features, especially for RF only between 51% and 14% of the features are equal, between 87% and 84% of the features coincide for PPLS-DA.

For scenario 2 with 10% ISF less than two features are simulated as informative and detected as important for the single sample design ($I_1^{\text{PPLS-DA}} = 1.5$) concerning PPLS-DA. The intersection with the important features for the pooling designs are lower with less than 0.3 coinciding features. Different results are shown for RF. The number of identical important ISF is very large for the single sample design ($|I_1^{\text{RF}}|$=95.6). However, the comparison with the pooling designs shows low similarities (less than one tenth of $|I_1^{\text{RF}}|$).

If exclusively the top ten important features are investigate for the intersection I_1^{RF}, for RF a high cardinality (9.8) is calculated. The cardinalities of the remaining

intersections $\mathrm{I}_1^{\text{PPLS-DA}}$, $\mathrm{I}_{1:m_p}^{\text{PPLS-DA}}$, $m_p = 2, 3, 5$ and $\mathrm{I}_{1:m_p}^{\text{RF}}$, $m_p = 2, 3, 5$, are near to zero.

PPLS-DA and RF declared important feature sets for scenario 3 and scenario 4 with 1% ISF Figure 5.7 shows the mean values of $|\mathrm{I}_1^{\mathrm{M}}|$ and $|\mathrm{I}_{1:m_p}^{\mathrm{M}}|$, $m_p = 2, 3, 5$, for scenario 3 (two linearly dependent features), and scenario 4 (the circle pattern) for the methods PPLS-DA and RF. For these two-dimensional patterns 20 features are simulated as informative for the classification.



(a) (b)

Figure 5.7: The mean cardinalities for the single sample design of $\mathrm{I}_1^{\mathrm{M}}$ (white bars) and for the pooling designs $\mathrm{I}_{1:m_p}^{\mathrm{M}}$, $m_p = 2, 3, 5$ are shown for scenario 3 and scenario 4 for 1% ISF (a) and 10% ISF (b) for M=PPLS-DA, RF. The *first group* depicts results for the method PPLS-DA (with three light gray bars). The *second group* depicts those for RF (with three gray bars). The numbers above the bars are the mean values rounded to one decimal place. 95% confidence intervals are shown. The lower separated hatched bars in (b) represent the mean cardinality if only the top 20 important features are accounted for in the feature sets.

For scenario 3, the cardinalities are alike for both methods PPLS-DA and RF, with 2.5 features which are ISF and important for the classification for the single sample design, and between 1.6 and 1.2 of these features are identical to the important features for the pooling designs (Figure 5.7(a)). The proportion of identically important ISF lies between 48% and 64% for the intersections of the sets for the single sample design and the pooling designs.

For scenario 4, the number of important ISF is nearly zero for the single sample design considering PPLS-DA and hence no feature is identical to the important features of the pooling designs. In contrast to RF, this methods detect nearly 15 features as important which are informative simulated for the single sample design ($|\mathrm{I}_1^{\mathrm{RF}}| = 14.9$), but the proportion of important ISF between the single sample design and the pooling designs lies only between 10% (for a pool size $m_p = 2$) and 4.4% (for a pool size $m_p = 5$).

PPLS-DA and RF declared important feature sets for scenario 3 and scenario 4 with 10% ISF Increasing the proportion of ISF to 10% (this means 200 ISF) for scenario 3, nearly 74 of 200 ISF are identified as important using PPLS-DA for the single sample design (Figure 5.7(b)). For the pooling design between 57 (for pool size 2) and 60 (for pool size 5) important features coincide to the important ISF for a single sample design. That means between 78% and 81% of the features detected by PPLS-DA for the single sample design are also identified for a pooling design.
For RF between 44.3 (for a pool size of 5) and 47.9 (for a pool size of 2) important features are equal to the 71.6 important ISF of the single sample design. Therewith the proportion of coinciding detected features by a single sample and a pooling design lies between 61% and 67% for the method RF.

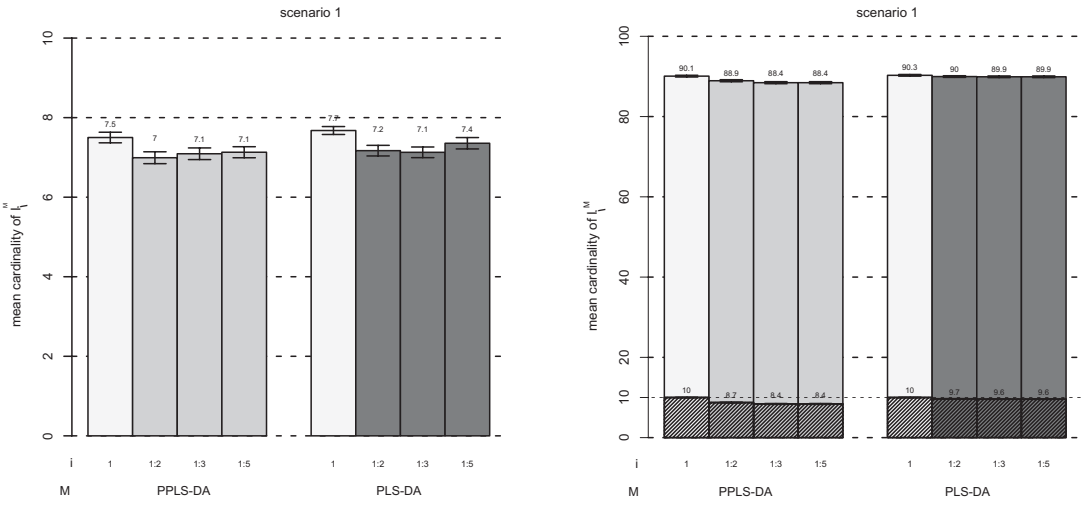
Investigating only the top twenty features, for PPLS-DA an average between 8.1 and 8.3 important features of the pooling design coincide with the 13.4 important ISF of the single sample design. For RF, between six (for a pool size of 5) and 8.6 (for a pool size of 2) important features for the pooling design coincide with the 10.9 important ISF for the single sample design.

For scenario 4 (the circle pattern) the number of important ISF for a single sample design is minor for PPLS-DA, with $|\mathrm{I}_1^{\text{PPLS-DA}}| = 22.3$ of maximal 200, the intersections with the corresponding sets of all pooling designs have similar cardinalities ($\approx$13.4).

Comparing to RF, a generally larger absolute number of coinciding features are found. However, the differences between the single sample design and the pooling designs are much larger than those for PPLS-DA, similar to scenario 2.
Considering only the top twenty important features, for RF between 4 and 2 features coincide for the single sample design and the pooling designs with a pool size of two and five. However, for the single sample design, almost all 20 possible ISF are identified as important by RF. For PPLS-DA and for the single sample design only one of the 20 important features is simulated as informative, and therefore the mean numbers of coinciding important ISF are low (between 0.5 and 0.4).

PLS-DA and PPLS-DA declared important feature sets for concept I Because PLS-DA shows no pooling effect of the PE for all scenarios and only a slightly higher PE than PPLS-DA for patterns of type I, important ISF are now compared for PPLS-DA with those of PLS-DA exemplarily for scenario 1. Moreover the PLS-DA loading weights of a single sample design are theoretically compared with those of a pooling design.

Figure 5.8 shows the mean cardinalities of $\mathrm{I}_1^{\mathrm{M}}$ and $\mathrm{I}_{1:m_p}^{\mathrm{M}}$, $m_p = 2, 3, 5$ for the methods PPLS-DA and PLS-DA. For 1% ISF the cardinalities are similar for each intersection, see Figure 5.8(a). Specifically for 10% ISF, the cardinalities for PLS-DA are larger than for PPLS-DA for all considered intersections (Figure 5.8(b)). Moreover, if only the top ten features are accounted (Figure 5.8(b), hatched bars), the coincidence of the important ISF for the single sample design and the pooling designs lies between 97% (for $m_p = 2$) and 96% (for $m_p = 5$) for PLS-DA, in comparison to PPLS-DA the corresponding coincidence lies between 87% and 84%. This large coincidence is caused by similar loading weights for the first component of PLS-DA for the single sample and the pooling design. If the log bias (see Section 5.2.1) is omitted, the equality of the loading weights can be proved which is shown in the following.



(a) (b)

Figure 5.8: The mean cardinalities for the single sample design of $\mathrm{I}_1^{\mathrm{M}}$ (white bars) and for the pooling designs $\mathrm{I}_{1:m_p}^{\mathrm{M}}$, $m_p = 2, 3, 5$ are shown for scenario 1 for 1% ISF (a) and 10% ISF (b) for M=PPLS-DA, PLS-DA. The *first group* depicts results for the method PPLS-DA (with three light gray bars). The *second group* depicts those for PLS-DA (with three gray bars). The numbers above the bars are the mean values rounded to one decimal place. 95% confidence intervals are shown. The lower separated hatched bars in (b) represent the mean cardinality if only the top ten important features are accounted for in the feature sets.

Loading weights of PLS-DA Now it is shown, that the loading weights vectors for a single sample and pooled sample design are equal for the first component of PLS-DA if the log bias is dropped. Because the R code used of PLS-DA, was implemented according to Indahl *et al.* (2007), this proof based on the corresponding PLS-DA approach. The absolute values of these loading weights are applied for the calculation of the feature ranking list. Thereby the same ranking lists are produced for both designs.

For each gene $k \in \{1, \ldots, g\}$, a corresponding column vector $\mathbf{x_k}$ (with length $n_1 + n_2 = n$) contains the gene expression of gene k for each sample. Here it can be assumed, that the elements $x_{ik}, i = 1, \ldots, n$ and $k = 1, \ldots, g$ are centered. The complete data set of the single sample design is the sample matrix $\mathbf{X} \in \mathbb{R}^{n \times g}$ and the dummy matrix $\mathbf{Y} \in \mathbb{R}^{n \times 2}$ which coded the group memberships for each sample. The $(\frac{n}{m} \times g)$ sample matrix of the pooling design with pool size m is denoted by $\mathbf{X}^p$ and by $\mathbf{Y}^p \in \mathbb{R}^{\frac{n}{m} \times 2}$ the dummy matrix of the pooling design.

First a transformation matrix $\mathbf{W}_0 = \mathbf{X}^t\mathbf{Y}\sqrt{\mathbf{\Pi}}(\mathbf{Y}^t\mathbf{Y})^{-1}$ is defined to simplify the calculation (later only the dominant eigenvector of a quadratic matrix with dimension equal to the number of groups has to be calculated). Here $\mathbf{\Pi}$ is the (2 x 2) diagonal matrix with the prior probabilities for each group as diagonal elements. Of the transformed data $\mathbf{X}^0 = \mathbf{X}\mathbf{W}_0$ the matrix of group means $\overline{\mathbf{X}}^0 = (\mathbf{Y}^t\mathbf{Y})^{-1}\mathbf{Y}^t\mathbf{X}^0$ is calculated. Then the between groups covariance matrix of $\mathbf{X}^0$ is $\mathbf{B}^0_\Pi = n\,(\overline{\mathbf{X}}^0)^t\mathbf{\Pi}\overline{\mathbf{X}}^0$. The PLS-DA loading weight vector $\mathbf{w}$ is the first dominant eigenvector $\mathbf{a}_0$ of the between groups covariance matrix $\mathbf{B}^0_\Pi$, transformed by $\mathbf{W}_0$: $\mathbf{w} = \mathbf{W}_0\mathbf{a}_0$.
In this thesis, no information on prior probabilities is assumed and so the empirical prior probabilities are used, for the single sample design $\pi_1 = n_1/n$ and $\pi_2 = n_2/n$ and for the pooling design $\pi_1 = n_{1(P)}/(n_{1(P)} + n_{2(P)})$ and $\pi_2 = n_{2(P)}/(n_{1(P)} + n_{2(P)})$ with $n_{\nu(P)}$ is the number of pools of group ν, $\nu = 1, 2$. Assuming pooling takes place on the log scale, the matrix for group means $\mathbf{X}$ and $\mathbf{X}^p$ are the same. Therefore the transformation matrices $\mathbf{W}_0 \in \mathbb{R}^{g \times 2}$ are equal for both designs, because this

matrix consists of the group means for each feature weighted with the square root of the related prior probabilities. It follows that $\mathbf{B}_{\mathbf{\Pi}}^{0,p} = \frac{1}{m}\mathbf{B}_{\mathbf{\Pi}}^{0}$ where $\mathbf{B}_{\mathbf{\Pi}}^{0,p}$ denotes the between groups covariance matrix for $\mathbf{X}^p$. Therefore the same dominant eigenvector a_0 for both designs is achieved. Note for an eigenvector $\mathbf{x}$ of $\mathbf{D}$ and $\mathbf{M} = c\mathbf{D}$, then $\mathbf{x}$ is also an eigenvector of $\mathbf{M}$: $\mathbf{Mx} = c\mathbf{Dx} = c\lambda\mathbf{x}$. Clearly if $\mathbf{x}$ is a dominant eigenvector for $\mathbf{D}$, then also for $\mathbf{M}$. This means an equivalent eigenvector problem is found for the single sample and the pooling design.
Finally, the loading weights are equal for the single sample design and the pooling design for the first component, under the assumption of no log bias.

Classification results using concept II for the simulated data without technical variation

In the following PE results for concept II (comparison of a single sample design and pooling designs with equal number of arrays, $a_P = a_S$) are exemplarily presented for pattern type I for scenario 1 with 1% ISF and scenario 3 with 1% ISF and 10% ISF and for pattern type II for scenario 2 and scenario 4 both with 10% ISF. The results of the remaining cases (data not shown) are similar to the presented results. In the figures regarding the results of concept II, for 30 arrays per group for the pooling design a pool size of 1 is used. For 15 arrays per group a pool size of two is applied to build the pools, for 10 arrays per group a pool size of three and for 6 arrays per group a pool size of five.

Pattern type I Figure 5.9 shows the mean PE for scenario 1 (differentially expressed genes) in dependence on equal number of arrays for pooling designs and corresponding single sample designs with 1% ISF. For all methods, except for SVMR and SVML, the PE for the pooling design is clearly lower than for the single sample design with equal number of arrays. Furthermore, the differences between the PEs increase with a decreasing number of arrays for these methods. For all pooling designs, the PEs of PPLS-DA and PLS-DA are nearly constant for both methods, but

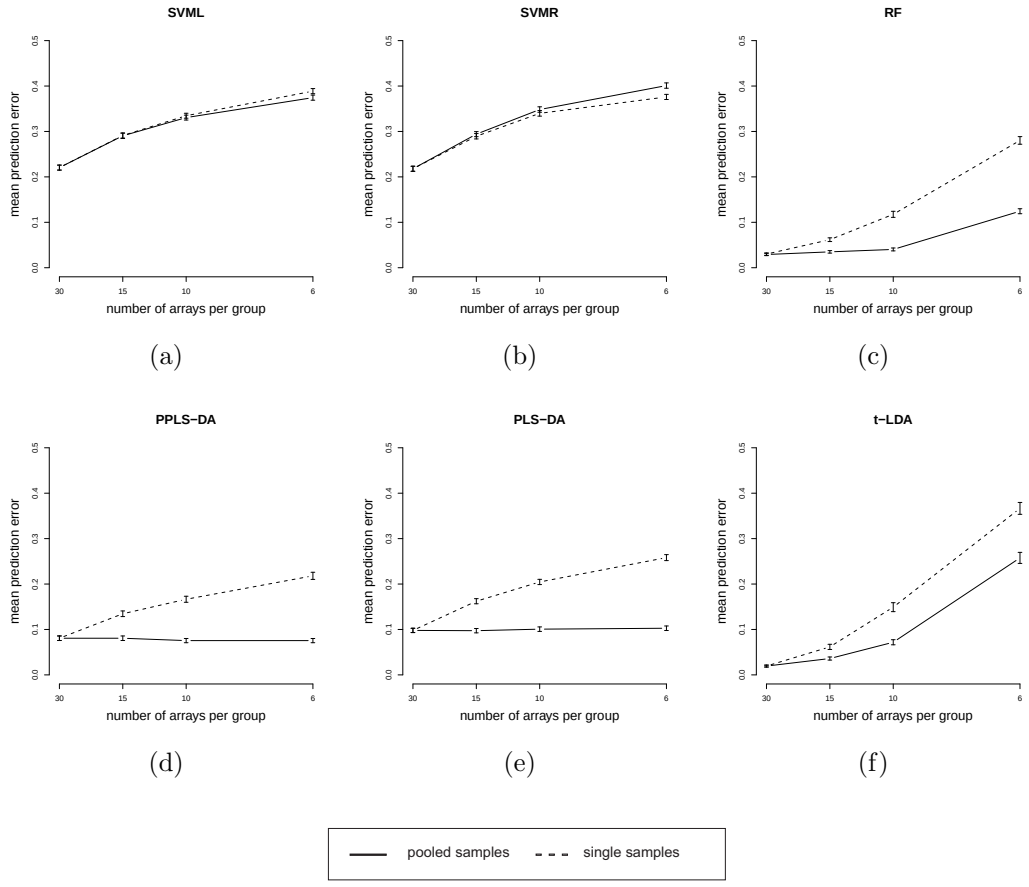


Figure 5.9: Mean PEs for scenario 1 for 1% ISF with 95% confidence intervals for concept II ($a_P = a_S$).

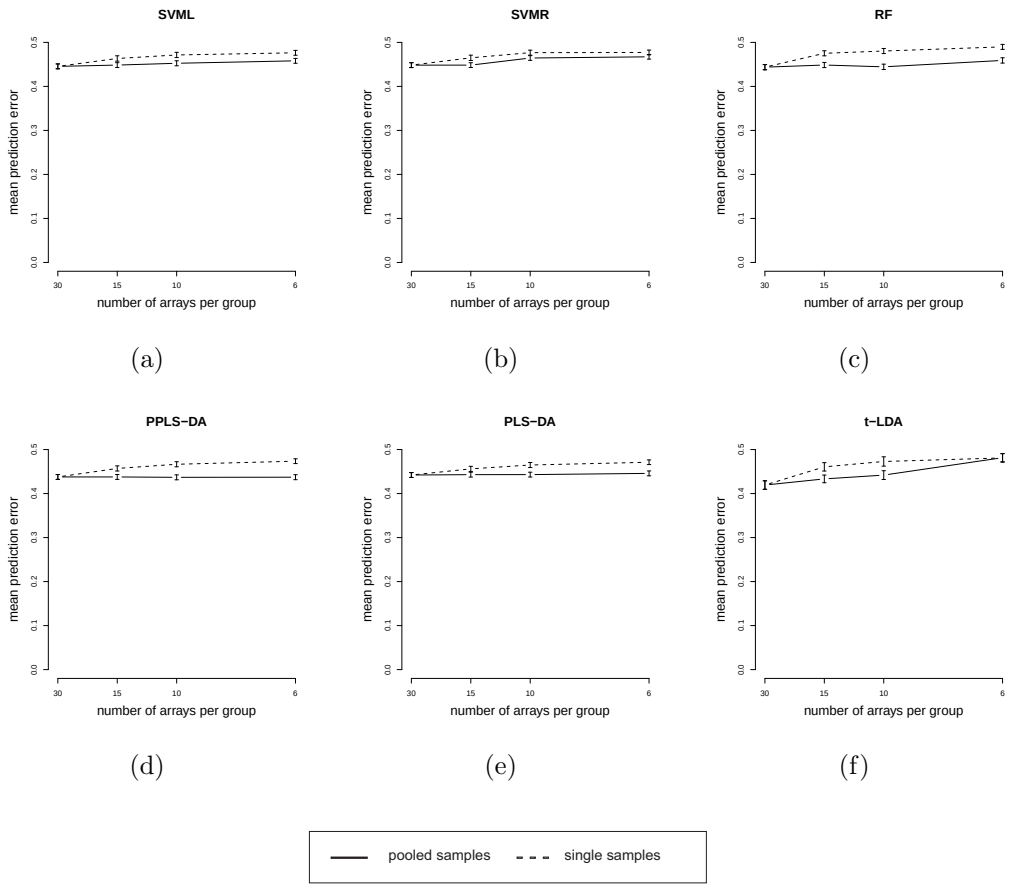


Figure 5.10: Mean PEs for scenario 3 for 1% ISF with 95% confidence intervals for concept II ($a_P = a_S$).

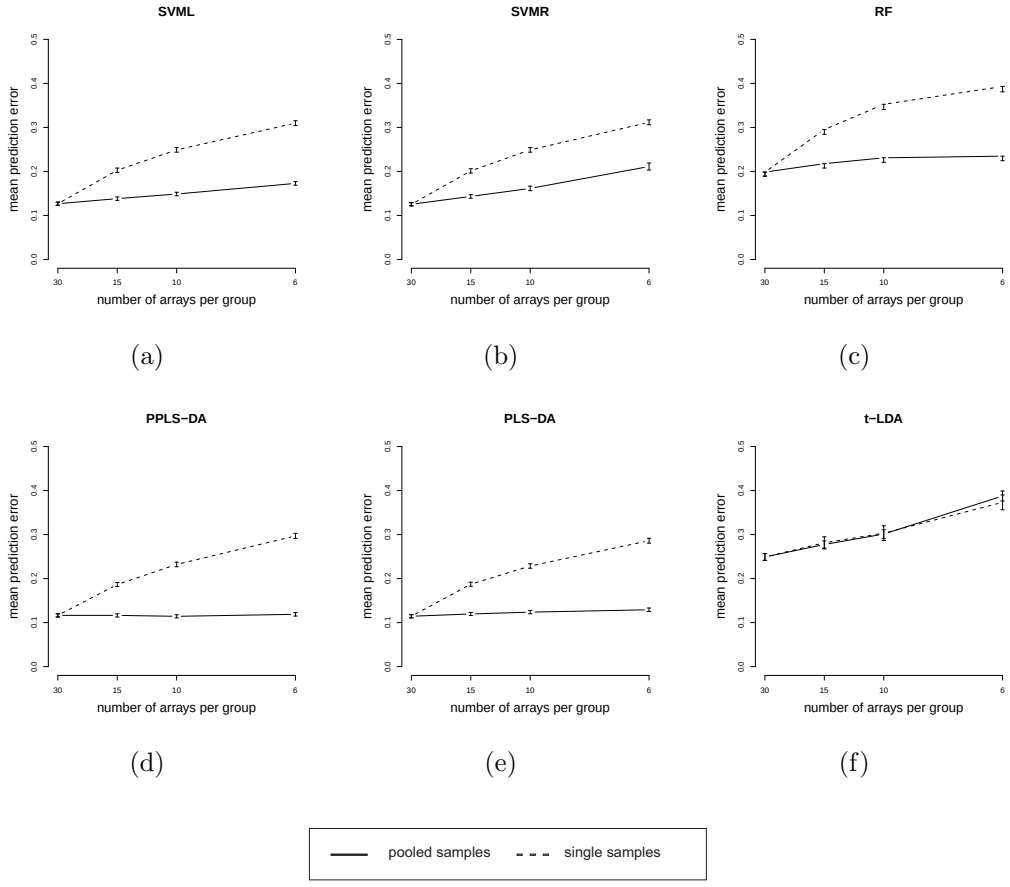


Figure 5.11: Mean PEs for scenario 3 for 10% ISF with 95% confidence intervals for concept II ($a_P = a_S$).

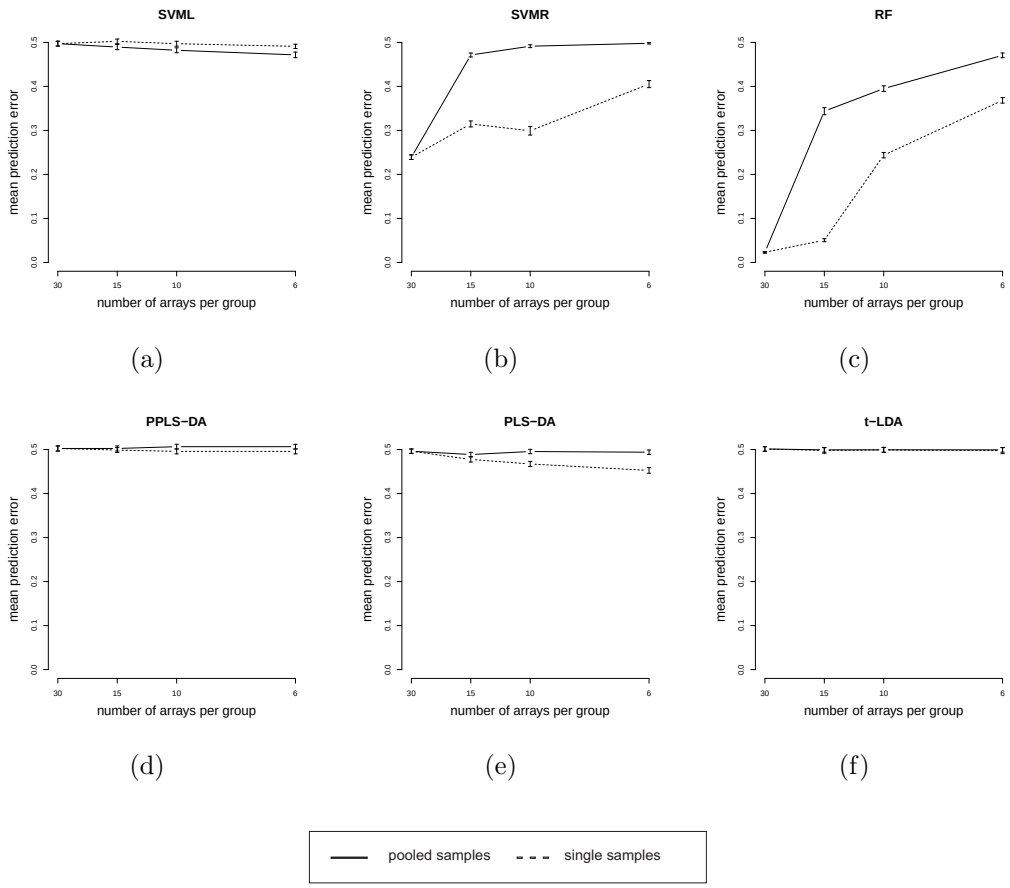


Figure 5.12: Mean PEs for scenario 2 for 10% ISF with 95% confidence intervals for concept II ($a_P = a_S$).

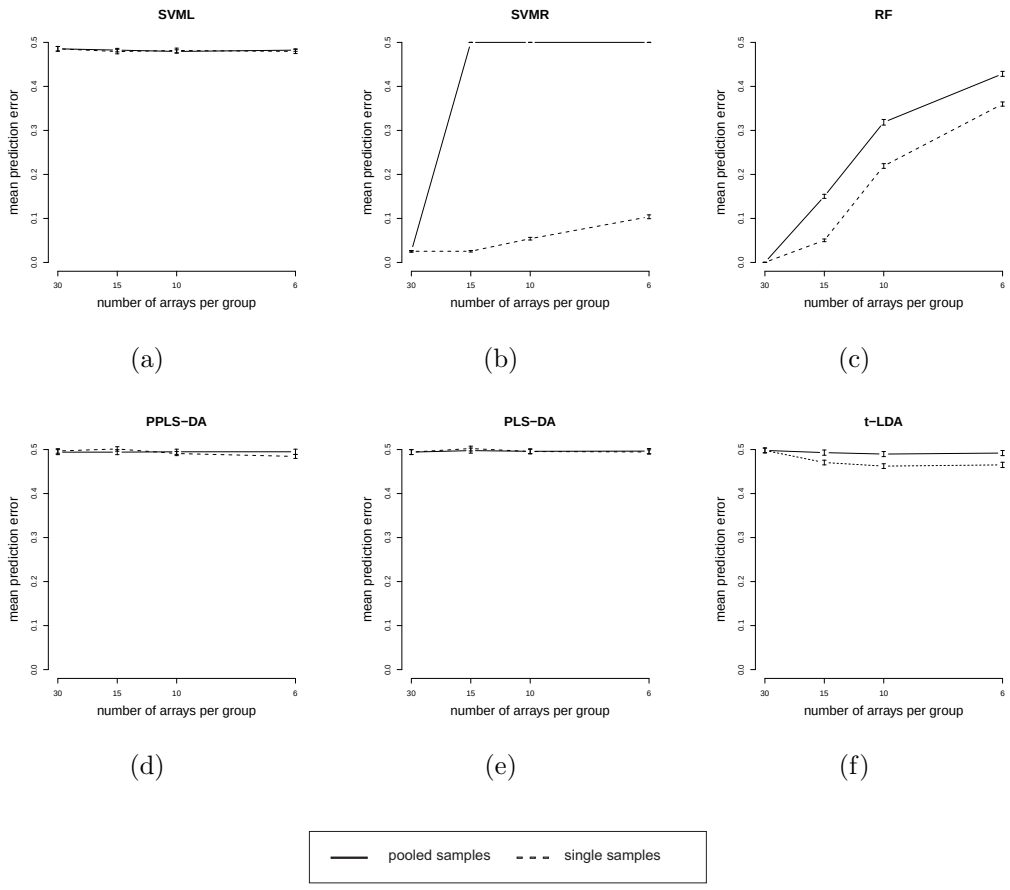


Figure 5.13: Mean PEs for scenario 4 for 10% ISF with 95% confidence intervals for concept II ($a_P = a_S$).

the PE increases with decreasing sample size for the single sample designs. For the methods SVML and SVMR the pooling designs show equal or slightly lower PEs than the single sample design for all number of arrays.

The PE results for scenario 3 with 1% ISF are illustrated in Figure 5.10. All classification methods show similar PEs (greater 0.4) for the single sample design and the pooling design with the same number of arrays.

For scenario 3 with 10% ISF the results are shown in Figure 5.11. Comparing 10% ISF to 1% ISF, all PEs are lower for the single sample and the pooling designs and number of arrays for all classification methods. Moreover for 10% ISF, the PE of the pooling design is lower than for the single sample design for all methods and each number of arrays, except for t-LDA. Only for t-LDA, the PEs are nearly equal for the single sample designs and the pooling designs. While for the pooling designs the PEs of RF, PPLS-DA and PLS-DA are nearly constant with decreasing number of arrays, the PEs for the single sample design clearly increase with decreasing number of arrays.

Pattern type II The PE results for concept II are similar for scenario 2 and scenario 4 of pattern type II. Figure 5.12 shows the results for scenario 2 (the threshold pattern) with 10% ISF. Only for the two methods, RF and SVMR, differences in the PEs are found comparing a single sample design and a pooling design with the same number of arrays (Figure 5.12). For these methods, the single sample designs have lower PEs than the pooling designs, but the differences in the PE decrease with decreasing number of arrays.

For scenario 4 (the circle pattern) with 10% ISF, only SVMR and RF have low PEs for the single sample design with 30 arrays per group. Therefore differences are shown between the single sample design and the pooling design with the same number of arrays (Figure 5.13) for these methods only. In contrary to scenario 1, the PE of RF is lower for the single sample designs than for the pooling designs. The

differences between the PE are very large for SVMR. For the single sample designs, the PE of SVMR increases from 0.02 to 0.101 with decreasing sample size per group (from 30 to 6). For all pooling designs ($m_p \geq 2$), the PE of SVMR is around 0.5. A reason for this high PE can be that this pattern type seems to be destructed after pooling.

Classification results for the simulated data with technical variation

In this Section, technical variance is simulated for the gene expression values. For concept I, the result are exemplarily illustrated for scenario 1 (pattern type I) and scenario 4 (pattern type II) with the two technical variance levels $\sigma_t^2 = 1/4\sigma_b^2$ and $\sigma_t^2 = \sigma_b^2$. For concept II, scenario 1 and scenario 4 with $\sigma_t^2 = \sigma_b^2$ is studied. The two scenarios 2 and 3 show similar results for concept I and concept II to scenario 1 and scenario 4, respectively.

First, concept I is considered for the differences between the single sample design and the pooling designs according to the PE and important ISF. Then concept II is analyzed to study the differences in the PE between the single sample design and pooling designs for equal number of arrays.

Concept I Figure 5.14 illustrates the mean PEs and corresponding 95% confidence intervals for scenario 1 for a technical variance $\sigma_t^2 = 1/4\sigma_b^2$ and $\sigma_t^2 = \sigma_b^2$. For 10% ISF and 20% ISF, the results are similar to those without technical variance (Figure 5.2(b) and Figure 5.2(c)). If the results for 1% ISF with the two technical variation levels are compared to the results without technical variation, for all classification methods the increase in the PE is larger with increasing pool size than for scenario 1 without technical variation, as expected. Especially for RF and t-LDA, the largest increase is found from $m_p = 1$ to $m_p = 5$, but for the pool sizes $m_p = 1, 2, 3$ the PEs are the lowest. PPLS-DA has the lowest increase of the PE with increasing pool size. In comparison to the results without technical variation, PLS-DA does not

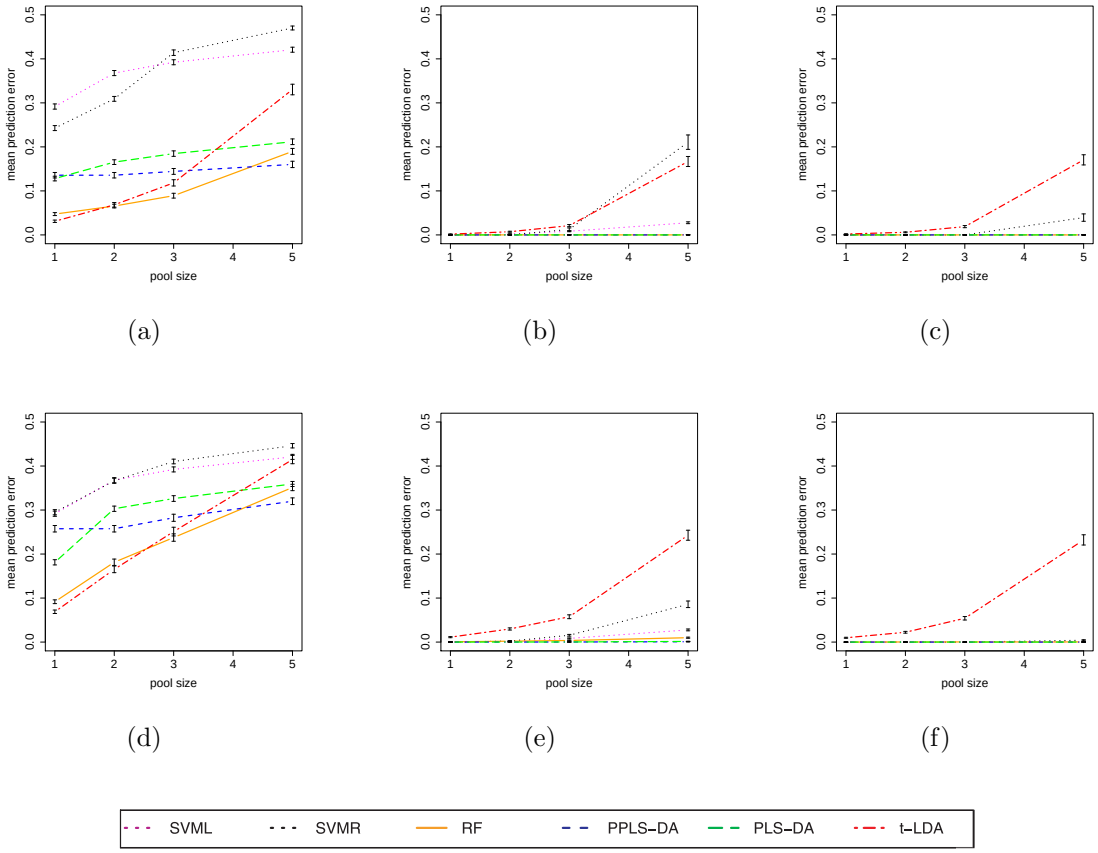


Figure 5.14: Mean PEs and 95% confidence intervals for scenario 1, in the first row for $\sigma_t^2 = 1/4\sigma_b^2$ with 1% ISF (a), 10% ISF (b) and 20% ISF (c) and in the second row for $\sigma_t^2 = \sigma_b^2$ with 1% ISF (d), 10% ISF (e) and 20% ISF (f).

longer show a constant PE. A clearly higher increase of the PLS-DA PE is found, especially between the single sample design and the pooling design with pool size 2.

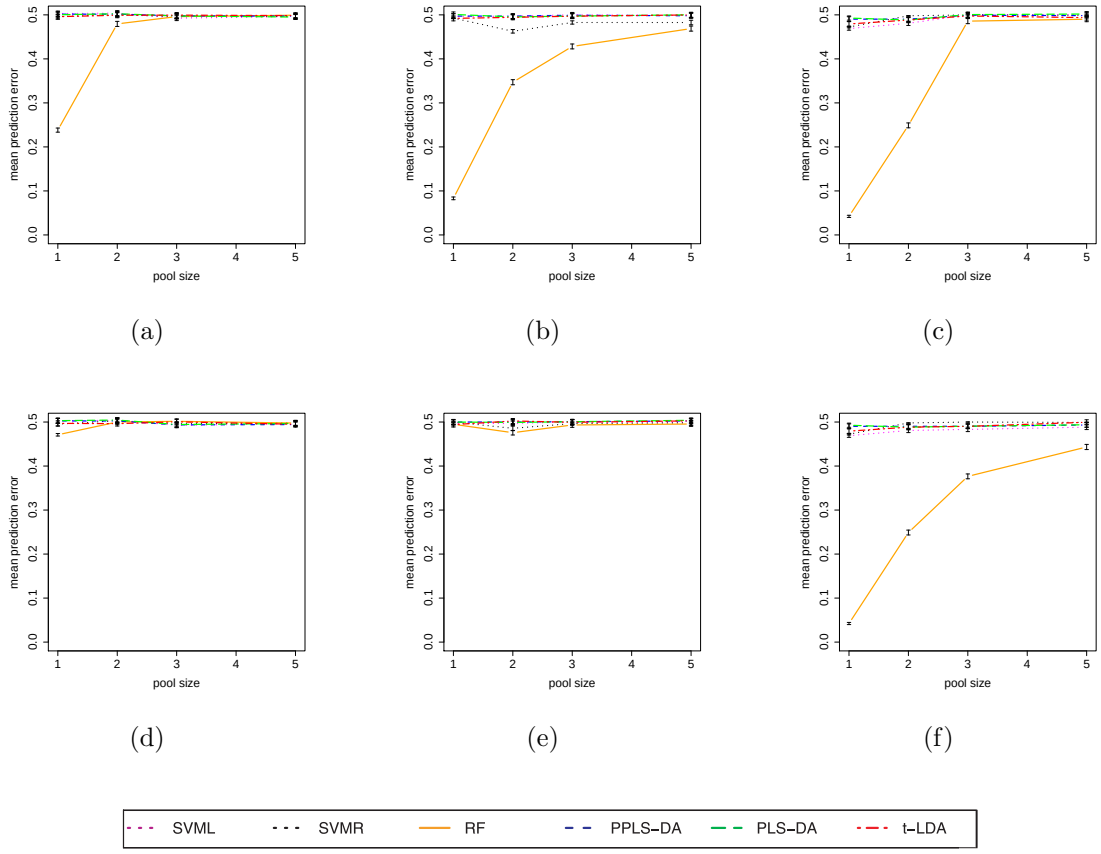


Figure 5.15: Mean PEs and 95% confidence intervals for scenario 4, in the first row for $\sigma_t^2 = 1/4\sigma_b^2$ with 1% ISF (a), 10% ISF (b) and 20% ISF (c) and in the second row for $\sigma_t^2 = \sigma_b^2$ with 1% ISF (d), 10% ISF (e) and 20% ISF (f).

Figure 5.15 summarizes the PE results for scenario 4 if technical variation is additionally simulated in two levels. In contrast to the results without technical variation (Figure 5.5), for both technical variation levels even for the single sample design, the method SVMR is not able to discriminate between the two groups. Like expected, RF shows larger PEs for increasing technical variation, especially for the pooling designs, but with a higher number of ISF this seems to be compensated to some degree.

Summarizing the PE results for concept I with technical variation, for scenario 1 also the PE of PPLS-DA is almost not influenced by an increasing pool size like in

the case without technical variation. For the gene expression data set with scenario 1 and technical variation, the classification method PLS-DA has now a clearly increasing PE.

For scenario 4 with and without technical variation, also RF shows the lowest PE for the single sample and the pooling design. In contrast to the results of scenario 4 without technical variation, for the single sample design and all pooling designs, the classification method SVMR has now larger PEs around 0.5 even for 10% ISF and 20% ISF.

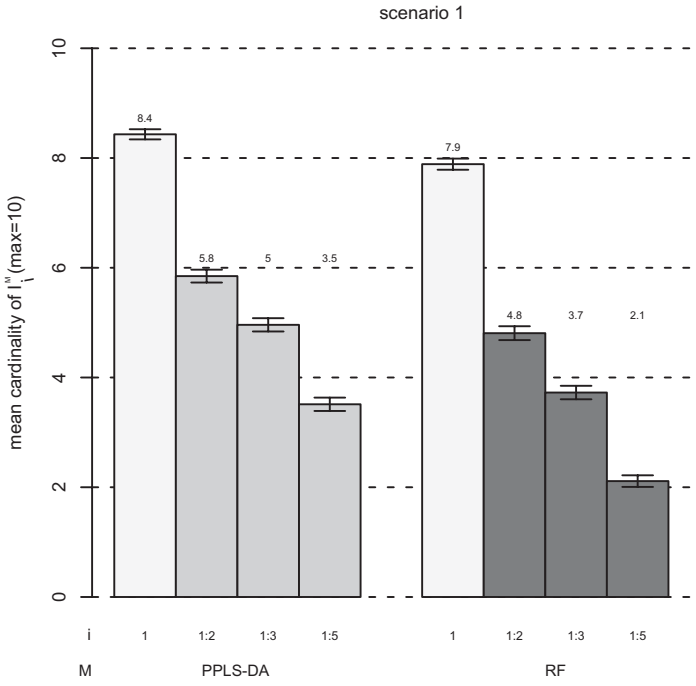


Figure 5.16: The mean cardinalities for the single sample design of I_1^M (white bars) and for the pooling designs $I_{1:m_p}^M$, $m_p = 2, 3, 5$ are shown for scenario 1 for 1% ISF for M=PPLS-DA, RF. The *first group* depicts results for the method PPLS-DA (with three light gray bars). The *second group* depicts those for RF (with three gray bars). The numbers above the bars are the mean values rounded to one decimal place. 95% confidence intervals are shown.

PPLS-DA and RF declared important feature sets for case 3 of scenario 1 with 1% ISF Now for scenario 1 with 1% ISF and a technical variance of $\sigma_t^2 = \sigma_b^2$, the important ISF are compared for the single sample design and the pooling designs

(Figure 5.16). For PPLS-DA and RF the number of important ISF which coincide between the single sample and the pooling design decreases for increasing pool size. This decrease in the number of coinciding important ISF is larger for the method RF than for PPLS-DA. For example for RF only 26 % of the important ISF are equal between the single sample design and a pooling design with pool size 5. In comparison, for PPLS-DA still up to 42 % important ISF are equal for the same comparison between the single sample design and a pooling design with pool size 5. If now the described results for scenario 1 with technical variation are compared to the corresponding results without technical variation (Figure 5.8(a)), for PPLS-DA and RF the differences between the cardinalities of the sets of important ISF for the single sample design and the pooling designs are larger for the results based on technical variation, as expected. Still PPLS-DA shows clearly a larger number of important ISF which coincide between the single sample design and the pooling design than RF.

Concept II For concept II (equal number of arrays of a single sample design and a pooling design ($a_P = a_S$)), also scenario 1 (pattern type I) and scenario 4 (pattern type II) are studied.

For scenario 1, the results are illustrated in Figure 5.17, only for 1% ISF and a technical variance of $\sigma_t^2 = \sigma_b^2$, because for 10% ISF and 20% ISF the results are similar to the corresponding results without technical variance (Figure 5.9). The methods RF, PPLS-DA, PLS-DA and t-LDA show lower PEs for the pooling design than for the single sample design. SVML has nearly equal PEs comparing the pooling designs and the single sample design with the same number of arrays. The same is found for SVMR, only for six arrays per group the PE of the pooling design is slightly larger than the PE for the single sample design. For all methods, the differences between the PE of the pooling design and the single sample design with equal number of arrays are equal or smaller than the corresponding differences for scenario 1 without technical variation. The reason is, that for the gene expression data without

technical variation, the reduction of the biological variance has a greater effect on the total variance. For the single sample design, the total variance is $\sigma_b^2 + \sigma_t^2$ and for a pooling design $\sigma_b^2/m_p + \sigma_t^2$ (see Section 5.2.1), hence for $\sigma_t^2 = 0$, a pool size m_p =5 leads to a variance reduction of 80% of the total variance. In comparison for $\sigma_t^2 = \sigma_b^2$ the variance is only reduced by 40% of the total variance, because pooling has only an effect on the biological variance.

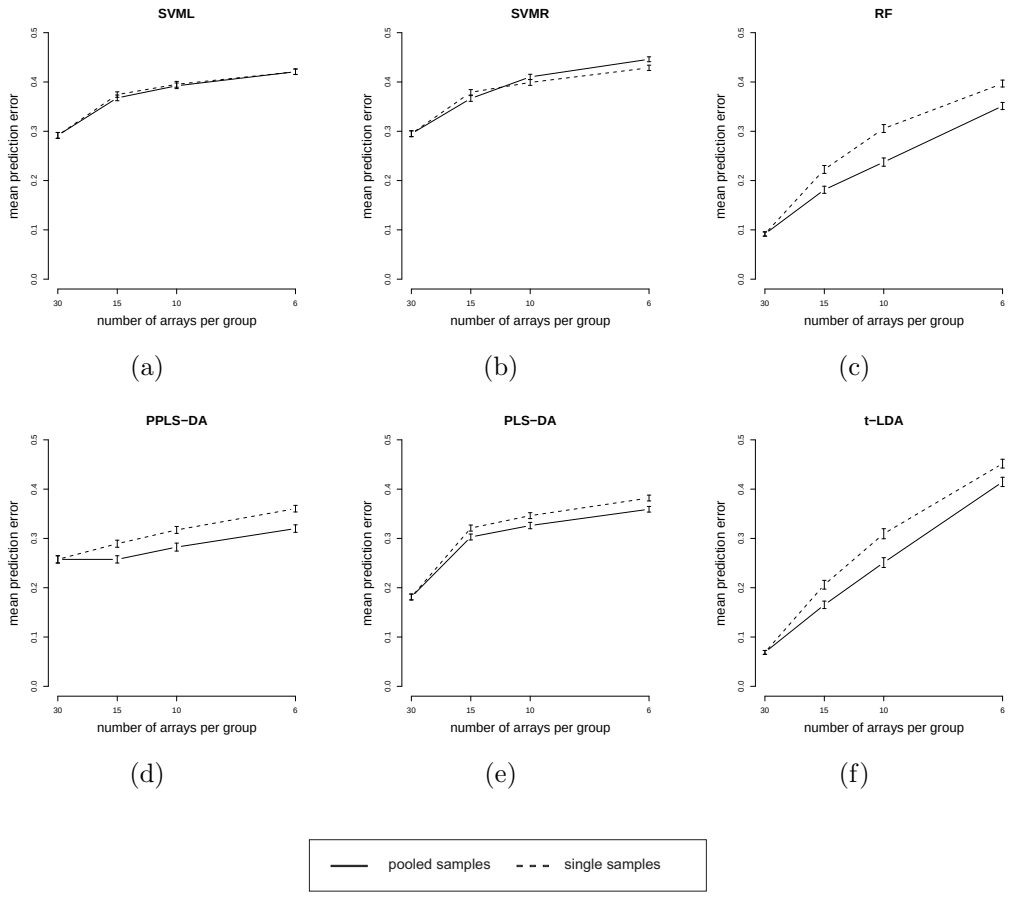


Figure 5.17: Mean PE for scenario 1 with $\sigma_t^2 = \sigma_b^2$ for 1% ISF with 95% confidence intervals for concept II ($a_P = a_S$).

Now for scenario 4 with 20% ISF and $\sigma_t^2 = \sigma_b^2$, the results of concept II are illustrated in Figure 5.18. Analogously to the results without technical variation with 1% ISF, only for RF differences are shown between the single sample design and the pooling design. However now for all number of arrays, the single sample design shows a larger PE than the pooling design. The reason can be the high technical variation, which leads to no clear separation between the groups even for single

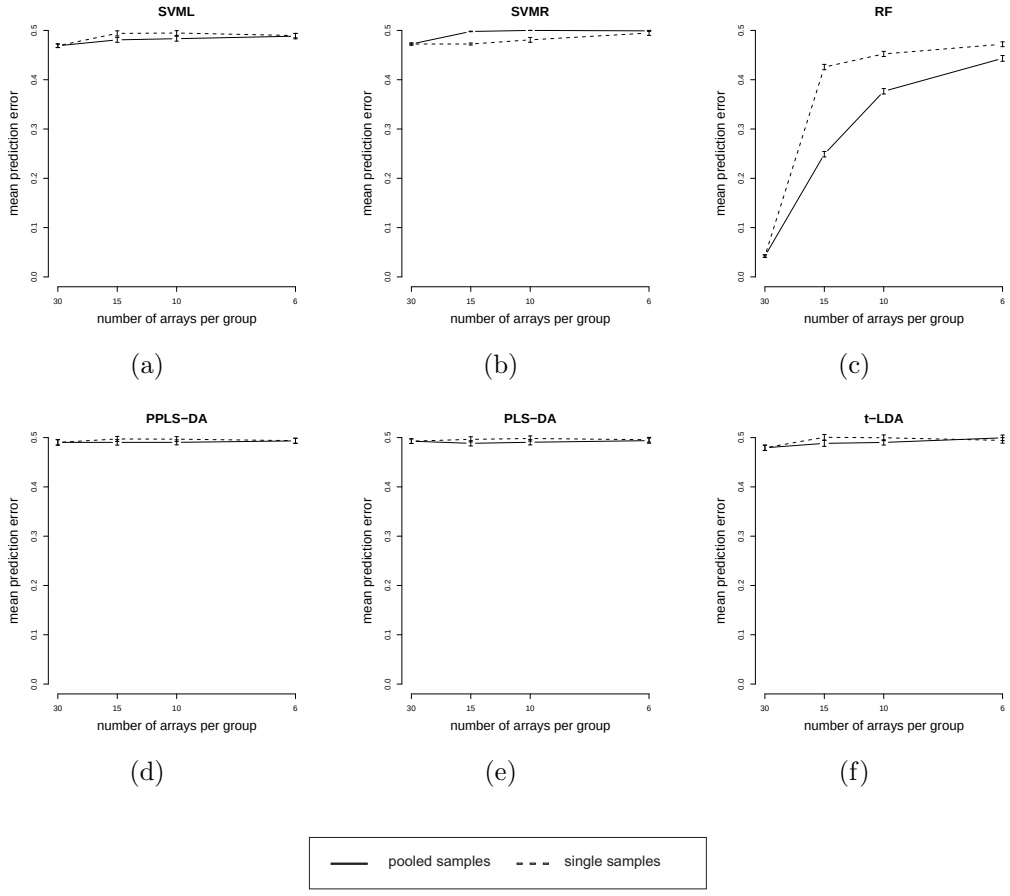


Figure 5.18: Mean PE for scenario 4 with $\sigma_t^2 = \sigma_b^2$ for 20% ISF with 95% confidence intervals for concept II ($a_P = a_S$).

samples (see Figure A.2). All other methods have similar and large PEs between the single sample design and the pooling design.

Now the results are summarized for the simulated gene expression data with technical variation. Regarding concept I for scenario 1 (pattern type I), PPLS-DA shows an almost constant PE with increasing pool size. Also PPLS-DA has a larger number of important ISF which coincide between the single sample design and the pooling design than for RF. However, these numbers of coinciding important ISF are larger for the simulated data without technical variation. For scenario 4 (patterns of type II), RF shows the lowest PE for all designs, but with a strong increase for increasing pool size like for the simulated data without technical variation. All other classification methods have large PEs around 0.5. Hence, even for the single sample design, SVMR has a large PE for scenario 4 with 10% ISF and 20% ISF, which is in contrast to scenario 4 without technical variation.

Considering concept II, the differences in the PE between the single sample design and the pooling design with the same number of arrays are lower than for the corresponding results without technical variation.

5.4 Artificial pooling of experimental data

To the best author's knowledge there is no experimental data set publically available which allows the comparison of a single sample design and pooling designs according to concept I and concept II. To be able to incorporate experimental data, publicly available experimental data sets (see Section 4.2) are theoretically pooled; this is called artificial pooling.

5.4.1 Approach and implementation

Approach of artificial pooling

The publicly available data sets are chosen according to two criteria:

(1) the number of samples per group has to be large enough to allow artificial pooling with pool size 2, 3 and 5 and

(2) already normalized data have to be available.

Therefore the Leukemia, the Prostate 1, the Prostate 2 and the Breast Cancer data set (see Section 4.2) are included. The Lymphoma data set is disregarded, because the group size of the one group is too small with 19. For the training sets of the pooling designs, the samples are randomly chosen, gene wise the mean values are calculated as pool value. For three of four data sets (Prostate 1, Prostate 2 and Breast Cancer data set) $n_{S_P} = 30$ single samples per group are randomly drawn to build the training set of the single sample design. Hence, 30 pools are calculated for $m_p = 2$ and 20 pools for a pool size 3 and 12 pools for a pool size 5 respectively. Because of the smaller number of available single samples (47 and 25 per group) in the Leukemia data set, the following numbers of pools are calculated: For $m_p = 2, (5)$ there are 10 pools (4 pools) per group, and 7 pools per group for $m_p = 3$.
For all data sets, the remaining single samples build the test set. Therewith the sample size of the test sets differs between the data sets. The classification is performed analogously to the simulated data. The artificial pooling procedure is repeated 100 times to estimate the PE for each classification method and corresponding 95% confidence intervals for the means are calculated.

5.4.2 Results

In the following the PE results are presented for the four publicly available data sets for a single sample design and pooling designs with artificial pools for concept I and for three data sets for concept II.

Classification result of concept I using experimental data

In the following, for the experimental data sets the PE results are described for concept I. The PE results can only be compared for the single sample designs to published PE results, because only the single sample design is used in the literature for these experimental data sets. However not exactly the same normalized data set, implementations and parameters for the classification methods were used. For the functions used and for parameter settings chosen by the cited authors, the mentioned papers are given as reference.

For all four experimental data sets the PE results for the single sample design and the three pooling designs are illustrated in Figure 5.19.

Leukemia data set (Figure 5.19(a)) For the single sample design, the PE of all methods is very low (below 0.06). For the methods SVML, RF, PLS-DA and PPLS-DA, the PE remains below 0.09 also for all pooling designs. For t-LDA the PE increases to 0.15 for pooling designs with a pool size larger than 2. Considering the pooling designs separately, the largest PEs are found for designs with a pool size smaller than 3 ($m_p \leq 3$) for t-LDA and for designs with a pool size $m_p = 5$, SVMR has the largest PE. Moreover for the method SVMR, the largest increase of the PE is found from 0.033 (for a single sample design) to 0.25 (for a design with pool size 5).

Comparing these PE results for the Leukemia data set to the results of the simulated data, most similarities are found to scenario 1 with 10% ISF. This can be caused by the high proportion of differentially expressed genes in the Leukemia data (40.5%, see Table 4.1). Only t-LDA shows larger PEs for $m_p = 2, 3$ than in the simulated data.

If now the PEs of the single sample design are compared to published PE results, similar PE results of RF (0.019-0.051), SVMR (0.018), SVML (0.014) and a further PLS-algorithm (0.02-0.03) are reported in Dettling (2004), Díaz-Uriarte and de Andrés (2006) and Boulesteix (2004).

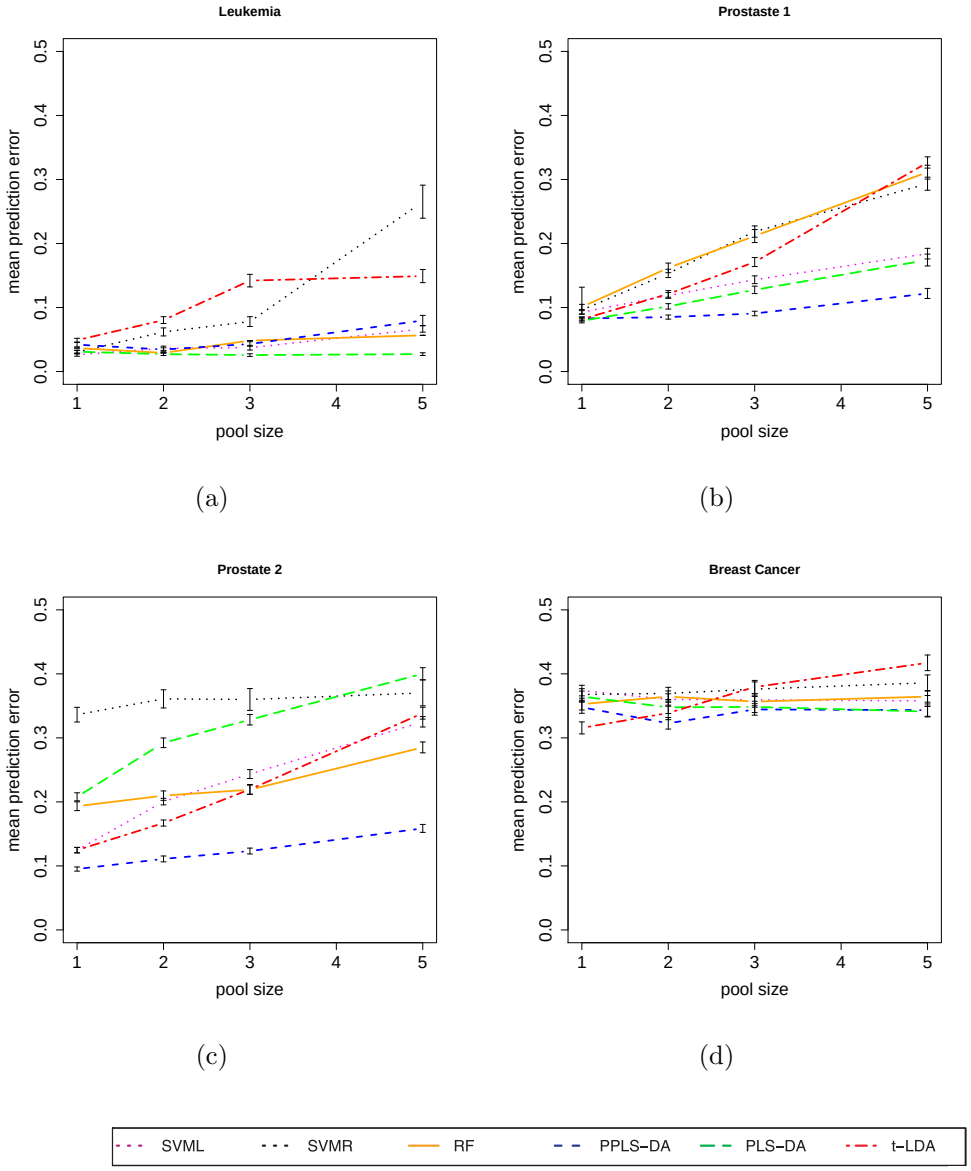


Figure 5.19: Mean PEs for the designs of concept I for the experimental data sets with 95% confidence intervals, estimated for 100 runs.

Prostate 1 data set (Figure 5.19(b)) For the single sample design, all methods show similar PEs equal or less than 0.1. The PE of PPLS-DA increases only slightly with increasing pool size from 0.08 (for a pool size $m_p = 1$) to 0.12 (for a pool size $m_p = 5$). PLS-DA and SVML show a similar increase of the PE, up to 0.17 for a pool size $m_p = 5$. For a pool size greater than 3, the PE of t-LDA increases strongly, for example for a pool size of 5 the PE is 4-times larger than for a single sample design. RF and SVMR show similar PEs for all designs, which increase linear from a single sample design to a design with pool size 5.

In comparison to the simulated gene expression data, the PE results of the Prostate 1 data set show most resemblance to the PE results of pattern type I. Considering published PE results for the single sample design, the PE results of Díaz-Uriarte and de Andrés (2006) for SVML are equal to our result (0.064), for RF the PE is slightly lower with 0.077. Dettling (2004) reported a PE of RF (0.09) which lies in the 95% confidence intervals of the results presented in this thesis and for SVMR a PE (0.0788) which is outside the 95% confidence interval of the mean PE shown here. Boulesteix (2004) calculated also a PE of 0.078 of a PLS-algorithm for the single sample design which corresponds to our results of PLS-DA.

Prostate 2 data set (Figure 5.19(c)) For the Prostate 2 data set, the PE results differ more between the methods than for the other data sets. SVML and t-LDA show a PE below 0.13 for the single sample design. For the pooling designs, these PEs increases up to 0.34 for a pool size $m_p = 5$. The PE of RF is nearly constant (around 0.21) for pool sizes smaller or equal to 3, but for a pool size of 5 the PE increases to 0.28. For all designs, PLS-DA shows a slightly higher PE than SVML. PPLS-DA clearly shows the lowest PE for all designs. These PE increases from 0.09 for the single sample design to 0.16 for a pool size of 5. For SVMR the PE is between 0.33 and 0.37 for all designs. That PE is estimated out of 50 runs, because the calculations are very time-consuming.

Now in contrast to the simulated gene expression data, the PE results of the Prostate 2 data set differ most to the results of the simulation study. The reason can be the much higher number of genes (42129) in the experimental data instead of 1000 which were considered in simulated data. For this data set, the results of the single sample data cannot be compared to other published results, because different classification methods are applied.

Breast Cancer data set (Figure 5.19(d)) For all pool sizes and all methods, the PEs are larger than 0.3. For all designs, the PE is between 0.32 and 0.36 for the methods PLS-DA, PPLS-DA, RF, SVMR and SVML. For the single sample design, the PE of t-LDA is the lowest (0.31) and increases to 0.42 for a pool size $m_p = 5$. Comparing these results to the results of the simulated data, scenario 3 with 1% ISF shows also for the single sample design, that t-LDA has the lowest PE, but increases more than RF, SVML, SVMR, PLS-DA and PPLS-DA. Now for RF and SVML, the PE of the single sample design is compared to published PE results for these classification methods. For RF, a PE of 0.342 and for SVML a PE of 0.325 are reported in Díaz-Uriarte and de Andrés (2006) for the single sample design. For RF the PE is covered by the estimated 95% confidence intervals for the single sample design shown here, the PE of SVML is only slightly higher.

Classification result of concept II using experimental data

The mean PE results for the experimental data sets are illustrated for the Leukemia data set, the Prostate 1 data set and the Breast Cancer data set for concept II (comparison of a single sample design and pooling designs with equal number of arrays, $a_P = a_S$). The Prostate 2 data set show similar results (data not shown).
For all designs for the Leukemia data set, Figure 5.20 shows the PE results for concept II. Comparing the pooling designs and the single sample design with equal

number of arrays, the pooling design has a lower PE than the single sample design for every classification method except for SVMR, which has nearly identical PE results for both designs. For a decreasing number of arrays the differences between the PEs of the single sample and the pooling designs increase. The differences are particularly small, for the methods PLS-DA and SVML for 10 arrays per group. However, for 4 arrays per group, the PEs of these two methods are twice as large for the single sample design than for the pooling design. The maximum difference in the PE between the single sample and the pooling design is found for RF for 6 arrays per group, with a PE nearly three times larger than for the pooling design. For t-LDA and PPLS-DA the differences are comparable between a single sample design and a pooling design.

The results for the Prostate 1 data set are shown in Figure 5.21. The differences between the single sample design and the pooling design with the same number of arrays are similar to the corresponding differences for the Leukemia data set. SVMR and RF have similar results for both designs. In general, the PE for the single sample design is larger than for the pooling design and this difference increases with decreasing number of arrays. Especially for PPLS-DA, the PE for the single sample design is larger than that of the pooling design for 6 arrays per group (over two times larger).

For the Prostate 2 data set, the results of concept II (data not shown) are very similar to the results of the simulated data for scenario 3 with 10% ISF, for all classification methods except for t-LDA. The PE of t-LDA is also clearly larger for the single sample designs than for the corresponding pooling designs.

Analyzing the Breast cancer data set with respect to concept II (Figure 5.22), if differences in the PE are found between the single sample and the pooling design they are comparable small, even for only 6 arrays per group. The findings for the Breast cancer data set are analog to those of scenario 3 with 10% ISF (Figure 5.10).

The PEs for the single sample designs start around 0.35 instead of 0.45 for scenario 3.

Summarizing PE results for concept II and all data sets, for the single sample designs, the PE generally increases with a decreasing number of arrays more than for the pooling designs with decreasing number of arrays.

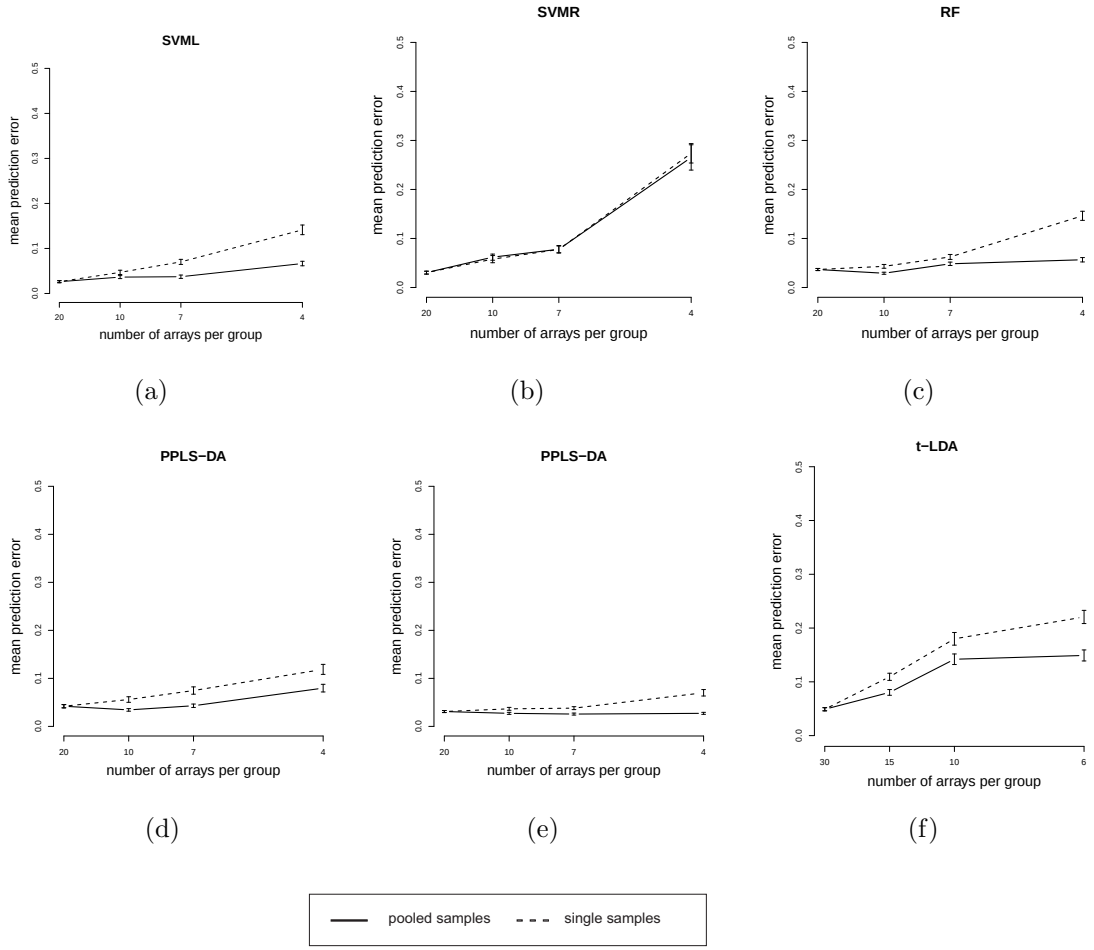


Figure 5.20: Mean PE for the Leukemia data set with 95% confidence intervals for concept II ($a_P = a_S$).

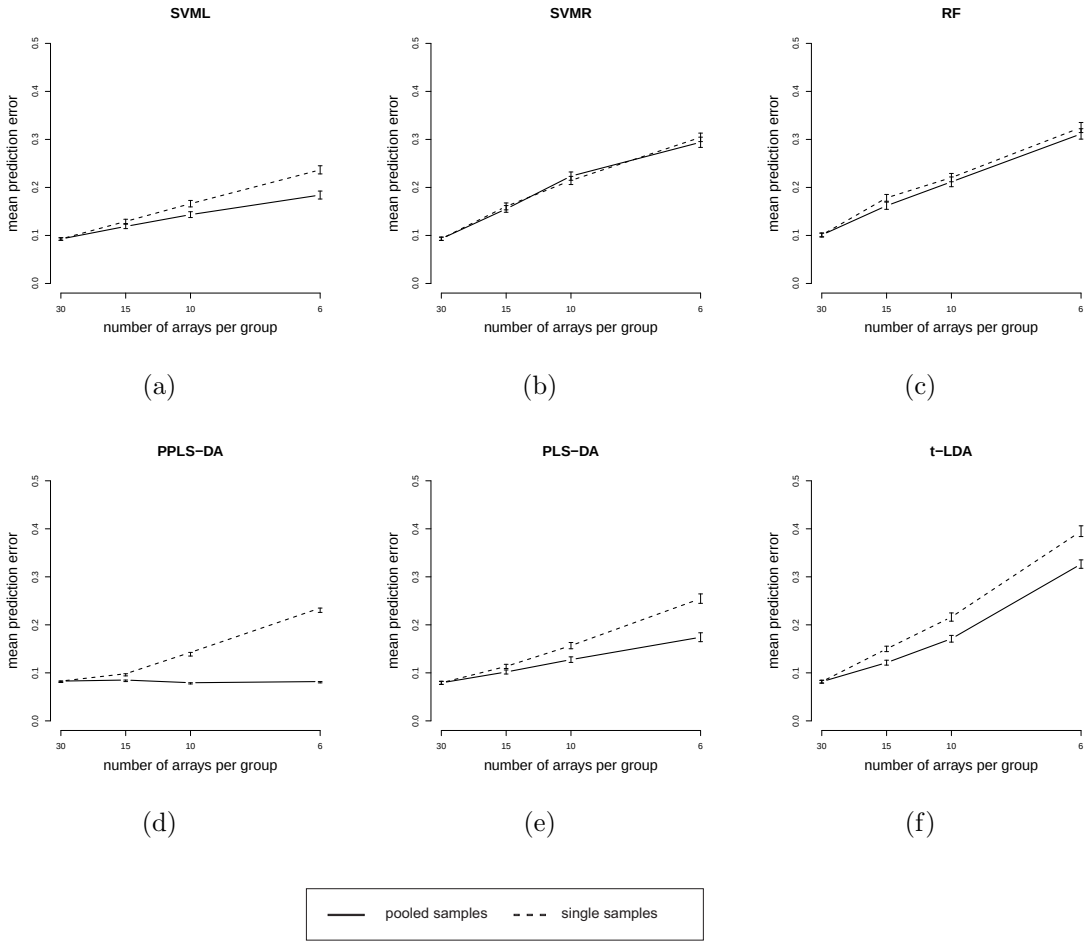


Figure 5.21: Mean PE for the Prostate 1 data set with 95% confidence intervals for concept II ($a_P = a_S$).

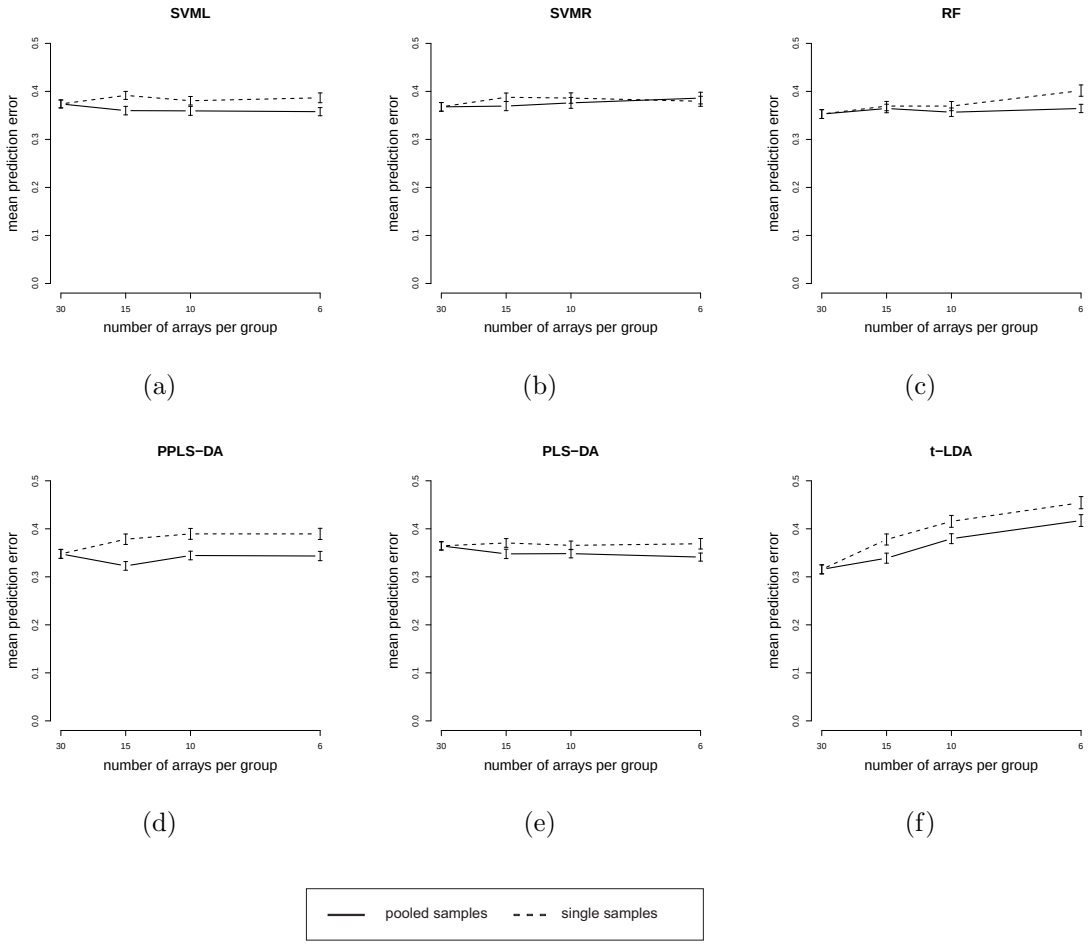


Figure 5.22: Mean PE for the Breast Cancer data set with 95% confidence intervals for concept II ($a_P = a_S$).

5.5 Discussion

Summary of main results

Pooling is known to affect performance of statistical learning, however, it is not always avoidable. In the presented study the pooling effect is considered by regarding the PE and important ISF for biomarker search in microarray gene expression experiments with the help of a simulation study. Furthermore the results are extended to publicly available data sets which are pooled artificially. The main focus is on concept I, with a lower number of arrays for the pooling designs than for a single sample design ($a_P < a_S$), but based on the same number of single samples ($n_{S_P} = n_S$).

The simulation approach employs six statistical learning methods, tested for four patterns differentiating between two groups, and with three different settings for numbers of ISF (1%,10% and 20%) and three technical variation levels ($\sigma_t^2 \in \{0, 1/4\sigma_b^2, \sigma_b^2\}$). Each combination is analyzed for the single sample design as well as for three pooling designs with increasing pool sizes.

Overall, looking at concept I there is no method which outperforms all other methods with respect to pooling designs regarding PE and important ISF independent of the other parameters for the simulated data. However, remarkable differences are found between the methods. Generally, if the technical variation increases, the differences between the single sample design and the pooling design increases considering the PE.

A pooling effect on the prediction error can only be assessed for cases in which statistical learning results in low PEs. For PPLS-DA and PLS-DA this is mostly the case for patterns of type I, where these methods show no pooling effect on the PE if no technical variation is simulated. If technical variation is simulated (Figure 5.14), PPLS-DA is the only methods which shows a nearly constant PE for increasing pool size. For patterns of type II, without and with technical variation, only RF mainly

shows the lowest PE, but with a strong pooling effect. The methods PPLS-DA and RF are further examined to analyze pooling effects on sets of important ISF. For patterns of type I and II, PPLS-DA shows a large proportion of important ISF which coincide between the single sample design and the pooling design. However, for RF and patterns of type II, much larger numbers of important ISF which coincide in absolute values have been found.

Regarding concept I for two of the four experimental data sets (Prostate 1 and Prostate 2), only PPLS-DA shows a minor pooling effect with respect to PE and also the lowest PE for all designs. PLS-DA shows the lowest and a nearly constant PE for all designs for the Leukemia data set. The Breast cancer data set seems more complicated for classification, because all classification methods and all designs lead to PEs over 0.3. Moreover, this data set has the lowest proportion of differentially expressed genes compared to the other experimental data sets.

Considering concept II (equal number of arrays for the single sample design and the pooling design, $a_S = a_P$), the influence of small sample sizes on the classification methods is studied. The PEs increase with decreasing number of arrays for the single sample design for all methods which are able to separate the classes with small PEs (Figure 5.9 - Figure 5.13, Figure 5.20 and Figure 5.21). For the pooling design and pattern type I, the PE increases slower for decreasing number of arrays per group than for the single sample design. In contrast for pattern type II, the PE of the pooling design increases strongly for decreasing numbers of arrays for classification method which show a low PE for the single sample design.

For the experimental data sets, the findings of concept II are analog to those of the simulated data for pattern type I.

Statistical learning methods and pooling

Pooling has different effects on the results of the statistical learning methods used. In general, pooling leads to a reduced biological variance. Therewith for pooling designs different variance structures are found in the training and the test set, because the training set consists of pooled samples and the test set always of single samples. Using concept I for a pooling design, pooling is accompanied by a reduced sample size. Carefully the design is chosen in such a way that the statistical power of the t-test to detect differentially expressed genes remains similar for all designs in concept I.

Simulated data Regarding the performance of classification methods for pooling designs for the simulated data, it should be distinguished between pattern type I and pattern type II.

Considering concept I and patterns of type I, the pooling effect measured by means of the PE of a classification task, depends on the chosen learning method, but is in general lower than for pattern type II. The reason is, that patterns of type I remain stable after pooling. For scenario 1 which is based on the group mean values, the group mean values of the pooled samples are only slightly biased by the log-transformation compared to those of the single samples if no technical variation is simulated. Moreover the biological variance is reduced, this affects particularly the simulated data without technical variation. For patterns of type II, the patterns are destructed after pooling (see Figure A.2). This causes the clearly higher PEs of the pooling designs compared to those of the single sample design in classification problems. For the case of low proportion of ISF (1%) and no technical variation, PPLS-DA and PLS-DA show no pooling effects. Their PEs are mainly low for patterns of type I, but larger for pattern type II (see Figure 5.3 and Figure 5.5). The large number of important ISF which coincide between the single sample design and the pooling design, especially for PLS-DA, is caused by nearly identical loading

weights for both designs (see Section 5.3.2).

Regarding concept II, it is possible to compare the pooling effect to the effect of decreasing sample size on the PE. For patterns of type I without technical variation, the PE results of PPLS-DA and PLS-DA for the pooling designs are constant although the sample size decreases while for the single sample design the PE increases (see Figure 5.9, Figure 5.10 and Figure 5.11). For RF and t-LDA, the pooling effect on the PE curve is similar to the PE curve of single sample designs for decreasing number of arrays). Nevertheless, for pattern type I the PEs of the classification methods are mainly lower for the pooling designs. This can be explained by the reduced biological variance using pooling designs (see Section 5.2.2). For patterns of type I with technical variation, all methods show an increasing PE for the single samples and the pooling design. Like already mentioned, simulating technical variation reduces the influence of the reduced biological variance (σ_b^2/m_p), especially if technical variance is larger than biological variance ($\sigma_t^2 < \sigma_b^2$).

Considering pattern type II with 10% and 20% ISF without technical variation (Figure 5.12 and Figure 5.13), only RF and SVMR have PEs below 0.5, because these two methods deal best with the more complex pattern structures. For RF a higher number of ISF seems to compensate the technical variation influence on the PE. In contrast to SVMR, where also for the larger proportion of ISF large PE values are shown. Moreover, the PE results of RF and SVMR for the pooling design clearly show higher PEs because of the mentioned destruction of the pattern after pooling. It can be speculated in general, that for pattern type II, the reduced variance does not have the same effect as for patterns of type I.

Experimental data Overall, for the experimental data sets, PPLS-DA and PLS-DA are weakly influenced by pooling with respect to the PE. Therefore especially these methods are considered. For concept I, and for two experimental data sets, Prostate 1 and Prostate 2, the classification method PPLS-DA clearly shows lower

PEs than PLS-DA, especially for the pooling designs. Hence, the optimization of the power parameter of PPLS-DA leads to a great difference in the PEs between PPLS-DA and PLS-DA for these two data sets (for more details see subsection 5.5). PLS-DA shows nearly no pooling effect for the Leukemia data set and a low PE, a reason could be the large number of differentially expressed genes, which can also explain the similar results to the simulated cases with patterns of type I.
Regarding concept II, the Leukemia data set and the Prostate 1 data set show that the reduced biological variance leads to smaller PEs of PPLS-DA, PLS-DA, t-LDA and RF for the pooling designs than for the single sample design with the same number of arrays. The same results are found for the Breast Cancer data set, but the differences in the PEs are smaller. For SVMR no differences between the single sample design and the pooling design are found for all data sets, the reduced biological variance has no effect on the PE.

Comparison of the results for experimental data and simulated data Most probably experimental data do not contain only one pattern type. Experiences with simulated data on this thesis have shown that statistical learning methods mostly use differentially expressed genes even when more complex patterns are included in the data. Therefore, the results of the experimental data show most similarities to patterns of type I. So the same methods are identified as most robust against pooling like for the simulated data for patterns of type I, namely PLS-DA and PPLS-DA. Therewith the pooling effect on the PE is relatively small, because the mean class values are maintained in the data after pooling (Figure A.1). Comparing with the results of the simulation study, the method SVML shows a low to moderate increase and SVMR a more considerable increase for increasing pool size. For none of the four experimental data sets, RF shows such a high difference between the PE of single sample designs and pooling designs as for patterns of type II in the simulated data. Overall, PPLS-DA and PLS-DA show the lowest pooling effect on the PE. The results of the Leukemia data set and the Prostate 1 data set for concept II are

very similar to the corresponding results for scenario 1. The corresponding results of the Breast cancer data set and the Prostate 2 data set are similar to the findings of scenario 3 with 1% ISF and 10% ISF.

Overall, for all four experimental data sets, the artificial pooling results mostly correspond to the results of the simulated data concerning scenario 1, especially for the Leukemia data set and for the Prostate 1 data set. For these two data sets, the reason for the similarity is the high number of differentially expressed genes in the Leukemia data set (32%) and in the Prostate 1 data set (39.6%) (see Table 4.1).

PPLS-DA and PLS-DA in the context of pooling

A closer look is taken at the methods PPLS-DA and PLS-DA because of their good performances for differentially expressed features (scenario 1). In general, for the pooling design, the optimal numbers of components for PPLS-DA and PLS-DA are lower than for the single sample design (results not shown). For scenario 1 with 1% ISF without technical variation (Figure 5.2), the PEs are slightly lower for PPLS-DA than for PLS-DA. The corresponding cardinalities of important ISF set for the single sample design ($\mathrm{I}_1^{\mathrm{M}}$) and of the intersections of these set with the important features of the pooling designs ($\mathrm{I}_{1:m_p}^{\mathrm{M}}$, $m_p = 2, 3, 5$) displays no great differences for M=PPLS-DA and M=PLS-DA (Figure 5.8).

Moreover, in Section 5.3.2 it is shown that the PLS-DA loading weights for the first component are equal for all designs under the assumption of negligible log deviation and no technical variation. As mentioned in Section 5.2.1, pooling takes place on the original scale, so the mean value of a class for the single sample data is not exactly the mean value of the class for the pooling data on the log scale (Jensen's inequality, Jensen (1906)). Therefore, this assumption does not hold in our simulation study, however the comparison of important ISF for the single sample design ($\mathrm{I}_1^{\text{PLS-DA}}$) and the important features for the pooling designs ($\mathrm{D}_{m_p}^{\text{PLS-DA}}$) (Figure 5.8) does not show large differences between the different designs with pool size $m_p = 2, 3, 5$ (for example Figure 5.8). This leads to the supposition that this log bias has no great

impact (on PLS-DA) and hence is very small. For higher technical variation levels the difference between the loading weights of the single sample design and the pooling design increases. Therefore also the PEs differ more for e.g. scenario 1 with 1% and $\sigma_t^2 \in \{1/4\sigma_b^2, \sigma_b^2\}$ (Figure 5.14(d) and Figure 5.14(a)).

Summarizing, the influence of pooling designs on biomarker search with respect to the PE and important feature sets for the methods PLS-DA, depends also highly on the technical variation. For data sets with very low technical variation, PLS-DA is robust against pooling. However, increasing technical variation, leads to larger PEs for the pooling designs than for the single sample design. Regarding PPLS-DA and especially the power parameter γ, it can be observed that with increasing pool size the average γ value increases and also with increasing technical variation level the power parameter increases (see Table 5.1).

Table 5.1: Mean γ-value for scenario 1with 1% ISF.

	pool size m_p			
σ_t^2	1	2	3	5
0	0.51	0.53	0.55	0.59
$1/4\sigma_b^2$	0.51	0.55	0.57	0.63
σ_b^2	0.51	0.58	0.59	0.64

For the experimental data, the PE results of PPLS-DA and PLS-DA differ more than those of the simulated data. The reason seems to be the influence of γ on the loading weights calculation (see Section 2.3.1). The parameter γ is determined such that the canonical correlation of $\mathbf{XW_0}(\gamma)$ and $\mathbf{Y}$ is maximal, with a $g \times 2$ transformation matrix $\mathbf{W_0}(\gamma)$, which contains possible loading weight vectors (see Section 2.3.1). The optimal γ value is between 0.51 (for $m_p = 1$) and 0.59 (for $m_p = 5$) for the simulated data without technical variation (Table 5.1). This range increases with increasing technical variation to [0.51,0.64] for $\sigma_t^2 = \sigma_b^2$. This means that especially the calculated loading weights with γ close to 0.5 are similar to the loading weights

of PLS-DA. However, the optimal γ-value for the experimental data deviates further from $\gamma = 0.5$ than for the simulated data. (For γ-values near 1, the correlation part has a greater impact to the calculation of the loading weights. For γ-values near 0, the standard deviation part has a greater impact to the calculation of the loading weights.) Therefore, for the experimental data, the loading weights differ more between PLS-DA and PPLS-DA. It follows, that the components from PLS-DA and PPLS-DA which are applied as predictors for a LDA, also differ. Moreover for γ-values near 1 or near 0, a lower number of features contribute to the loading weights of PPLS-DA. This is especially an advantage for data sets with a large number of features, like for the Prostate 2 data set and can explain the large differences between the PEs of PLS-DA and PPLS-DA for this data set. Because for data sets with such a high number of features, a lot of features may not contain information for the discrimination between the groups and can be considered as noise (Saebø *et al.*, 2008). Therefore it is advantageous to set their loading weights to zero.

Constraints and benefits of methods

The reported results are based on the chosen pattern scenarios, proportions of ISF, the statistical learning methods used and pool sizes, therefore they cannot easily be generalized. Only small pool sizes are investigated ($m_p = 2, 3, 5$), and so one can expect that for $m_p > 5$ the results between single sample design and pooling design deviate much more. Simulating gene expression of a pool on the original scale, the mean values of the samples forming the pool is calculated. This exact pool value will not occur in experimental data, but pooling weights can be interpreted as an additional technical variation (Zhang *et al.*, 2007). Furthermore, the test set of a statistical learning method consists of new single sample data, because biomarker classification is applied on single samples with a much more sensitive technique (PCR vs. microarray technique).

Two types of patterns, type I (scenario 1 and scenario 3) and type II (scenario 2 and scenario 4) are investigated. For type I, the convex hulls of the one class is not a

cover of the other class. (For scenario 3 the convex hulls are always disjunct, for scenario 1 only for large mean class differences.) The similarities in the results between these two patterns of type I occur also because one of the two linearly dependent features of scenario 3 is based on the construction of a differentially expressed feature with a mean class difference of 0.1. For a higher mean class difference (> 0.1), the classification information is dominated by the differentially expressed gene in the pattern. The patterns of type I are linearly separable which is advantageous for methods like LDA. For type II, the convex hulls of the one class cover the other. This means the pattern structure after pooling depends on the individual samples chosen, which form the pool and the parameters for the simulation of the pattern. For example, considering the circle pattern (scenario 4, Figure 4.1d), if the radius of the inner circle is much smaller than the outer one, the intersection of the convex hulls of the two classes is smaller and the probability of disjunct classes is also larger after pooling. Moreover the scenarios of pattern type II are based on the same mean values and different variances for the two classes, this may cause the similar results for this pattern type.

For the experimental data, pools are artificially built by choosing $m_p = 2, 3, 5$ samples randomly and calculating the mean value. For choosing samples for the pools, the group memberships are taken into account. Therefore possible subgroups could be mixed.

In general the experimental data have a higher number of genes as our simulated data. Hence, the differences between the PE of the Prostate 2 data and the simulated data can possibly be explained by this discrepancy.

Summarizing the differences between the experimental data and simulated data, the experimental data sets have a clearly higher number of genes g, a higher number of differentially expressed genes, more than one pattern scenario and maybe subgroups in a group.

Significance of results

The presented results illustrate which consequences can be expected if a pooling design is used instead of a single sample design in the context of biomarker search and point out how important it is to choose an adequate design with respect to the underlying questioning.

If only differentially expressed features are presented in the data and only little technical variation, PPLS-DA and especially PLS-DA show the highest (relative and absolute) coincidence of important features with those also simulated as informative. For high technical variation, PPLS-DA outperforms PLS-DA. Investigating more complex patterns (scenario 2 and scenario 4) simulated without technical variance, it can be seen that, the PE increases even for small sizes of pools ($m_p = 2$) to nearly 0.5 if only few ISF (1%) are available (Figure 5.3 (a), Figure 5.5 (a)).

For practical reasons, a biomarker signature should not consist of too many features. For example, for quantitative PCR, more components of a biomarker signature lead to increased cost and work. Moreover, less features lead to better possibilities for biological interpretation of the biomarker in terms of understanding regulatory mechanism (see Section 2.1.1).

This aspect is taken into account by assessing only the top ten (top 20) important features of the 10% ISF cases. For the influence of pooling with respect to a possible biomarker created by a ranking list, variable importance values of RF, PPLS-DA and PLS-DA are taken into account, because these methods show the best PE results. The intersection of these reduced sets of considered important features is lower than for the higher number of important feature sets, except for PPLS-DA and PLS-DA. Even if pooling does not lead to an increase in PEs for RF (scenario 1 for 10% ISF, Figure 5.2(b), less than half of the important ISF, of the single sample design coincide with the sets of RF-declared important features for the designs with pool size $m_p = 2, 3, 5$ (see Figure 5.6 (a), lower separated bars). Hence, searching for biomarkers with a pooling design leads to a different choice of biomarkers as with

a single sample design. The size of the discrepancies depends on the pool size, the scenario, the proportion of ISF, the technical variation in the data, and the learning method used. Altogether, the assumption of Sadiq and Agranoff (2008) can be verified, that "pooling serum samples may lead to loss of potential biomarkers". If pooling can not be avoided and differentially expressed genes can be expected in the data, and considering the results of the simulation study and the experimental data, PPLS-DA and PLS-DA should be preferred for biomarker search. Mainly low PEs are found, and the sets of the important features for the classification using PLS-DA for the single sample design and the pooling design nearly coincide for the simulated data. Moreover, if the structure of the informative data is unknown, additionally RF should be investigated, because RF shows the lowest PEs for all designs and pattern type II. Furthermore the highest cardinality of the intersection of important ISF is found by RF.

Regarding the results of concept II: On the one hand, if the number of arrays used is fix and the number of single samples available is higher ($n_{total} > a_{total}$), a pooling design seems to be preferable, if lots of differentially expressed genes are present in the data (Figure 5.9, Figure 5.11, Figure 5.20 and Figure 5.21). However, in this case only patterns of type I, could be detected. On the other hand, if the data contain most probably low numbers of differentially expressed genes, a single sample design is preferable.

Feng *et al.* (2004) state that different methods should be compared to find the best approach for the underlying data set according to the lowest prediction error for the identification of biomarkers. Concluding, it is suggested to start with the methods PPLS-DA, PLS-DA and RF if the samples are pooled.

What are the consequences of pooling regarding decrease of prediction accuracy or increasing problems to find biomarker candidates?

Pooling leads to larger or equal PE depending on the classification method used, the pool size and technical variation. Especially for RF the loss of possible biomarkers

is high already for the simple case of differentially expressed genes. For PPLS-DA and PLS-DA this loss is much smaller.

6 Extension of the method PPLS-DA for improved classification

In this Chapter, an extension of the method PPLS-DA is proposed to improve the prediction results. It is based on a manuscript submitted by Telaar *et al.* (2012b).

6.1 Motivation

In Chapter 3, the method PPLS-DA turns out to be especially suitable for biomarker search. There, the biomarker obtained by PPLS-DA contains only half of the number of features compared to the biomarker identified by RF, while a similar PE is reported for both methods. Therewith particularly requirement b), a small cardinality, of a biomarker, is fulfilled, (Section 2.1.1). Further in Chapter 5, PPLS-DA is most robust against pooling with respect to the PE and important features. In addition, it could be seen that the power parameter γ has a substantial influence on the PE (Section 5.5).

This Chapter examines the possibilities of improving the PE of PPLS-DA with regard to the choice of the parameter γ. The theoretical backgrounds of PPLS-DA are described in Section 2.3.1.

In the ordinary PPLS-DA, the power parameter γ is optimized over the interval $U = [0, 1]$ such that the correlation is maximized between $\mathbf{Z} = \mathbf{XW}$ and $\mathbf{Y}$. The corresponding γ is denoted by γ_{max}. To analyze the influence of the power parame-

ter on the PE, first the usual interval for the determination of the power parameter [0,1] is decomposed in 20 sub-intervals $U_t = [0.05 \cdot (t-1), 0.05 \cdot t]$ with $t = 1, \ldots, 20$. Figure 6.1 shows the mean PE results for scenario 1 with 1% ISF and $\sigma_t^2 = 0$ (case 1) plotted against the intervals U_t used for the optimization of γ. The displayed PE curve is of wavy shape. Starting at a PE of 0.07 for the interval $U_1 = [0, 0.05]$, the PE first decreases to a minimum PE of 0.05 for the interval $U_2 = [0.05, 0.1]$, increases then to a maximum of 0.17 for the interval $U_6 = [0.25, 0.3]$, followed by a decrease to a global minimum for the PE of 0.014 for the interval $U_{17} = [0.8, 0.85]$ and finally increases to 0.05 for the interval $U_{20} = [0.95, 0.1]$. The smallest PE is achieved with a power parameter optimized with respect to the correlation inside the interval $[0.8, 0.85]$. The usual optimization of the power parameter γ inside the whole interval $U = [0, 1]$ with respect to correlation leads to an average γ of 0.51 (see Table 5.1) with a mean PE of 0.1. It follows that for scenario 1 with 1% ISF and $\sigma_t^2 = 0$ a clearly lower PE can be achieved by optimizing the power parameter only inside the interval $[0.8, 0.85]$.

Therefore in the following, four alternative versions to optimize the power parameter of PPLS-DA are introduced which optimize the power parameter towards prediction accuracy by taking a cross-validation approach to avoid over-fitting. In three versions (A, A1 and A2), the power parameter and the number of components are determined according to the lowest PE of a LDA using the PPLS-DA components as predictors. In a fourth version (B) the mean squared deviation of the posterior probabilities of the LDA from the dummy response matrix is applied as objective function. Furthermore, all results of the four extensions of PPLS-DA are compared to PLS-DA with respect to the PE for simulated data sets and five publicly available experimental data sets, already described in Chapter 4.

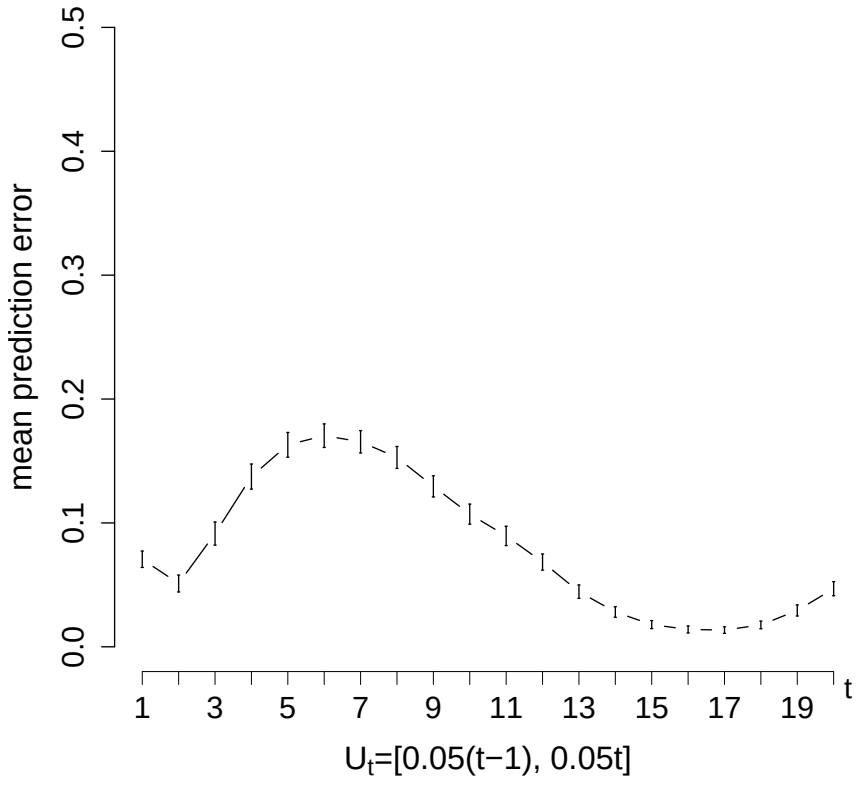


Figure 6.1: Mean PEs and corresponding 95% intervals for scenario 1 with 1% ISF and $\sigma_t^2 = 0$ plotted against different intervals U_t for the optimization of γ.

6.2 Proposal of a new optimization

For improvement of the classification using PPLS-DA, the power parameter γ is optimized with respect to the prediction accuracy of LDA in an inner cross-validation approach. For such an inner cross-validation, the data set $\mathbf{X}$ is first splitted according to a proportion of $(\phi_{\mathbf{X}}, 1 - \phi_{\mathbf{X}})$ in an outer training set and outer test set (see Section 2.1.2). Using only the outer training set to optimize the γ-value, this set is additionally partitioned randomly into an inner training set and an inner test set by the proportion of $\phi_{\mathbf{X}_{train}}$ and 1-$\phi_{\mathbf{X}_{train}}$ of the outer training set. This partition is randomly repeated r_{inner} times. The outer test sets serve to validate the learned classification function by LDA based on the components calculated according to the

optimized power parameter.

Moreover as possible γ-values, equidistant fixed values in [0,1] are taken into account with a step size s_γ. Therewith a sequence of γ-values is obtained $(\gamma_1, \ldots, \gamma_{\frac{1}{s_\gamma}+1})$.

Four versions are examined for optimizing γ and the number of PPLS-DA components to be used as input for the LDA. For all versions, the optimization of the power parameter γ depends on the choice of the parameters r_{inner} and s_γ.

Short overview of the optimization schemes: In version A all components are used to find the optimal γ which minimizes the PE, while version A1 only uses the first component to determine the optimal γ. Version A2 successively optimizes each component separately. Finally, version B optimizes like version A2, but with respect to the posterior probabilities of the LDA.
The final number of components used is determined in an additional cross-validation step for versions A1, A2 and B. Only in version A, the number of components and the power parameter are optimized in the same step.

In the following a one up to five component model of PPLS-DA is taken into account, $n_c = \{1, \ldots, 5\}$. Therewith one up to five components of PPLS-DA are used as predictors for the final classification with a LDA.

The following notations are used:

- PE_{tl} denotes a PE of the inner test set of the t^{th} partition of the outer training set, calculated for PPLS-DA based on a parameter with level l
- $\overline{\mathrm{PE}}_{.l}$ denotes a mean PE over all inner test sets calculated for PPLS-DA based on a parameter with level l

6.2.1 Version A

For each fixed γ-value $(\gamma_i, i = 1, \ldots, \frac{1}{s_\gamma} + 1)$ the optimal number of components of PPLS-DA is determined as follows (see Figure 6.2): For the r_{inner} different inner test

sets, the PE for a one up to five component model is calculated ($n_c = 1, \ldots, 5$). For a fix γ_i all components are determined according to this γ_i. Therewith over all r_{inner} repetitions a matrix $(\mathrm{PE}_{t\,n_c})_{tn_c}$, $t = 1, \ldots, r_{inner}, n_c = 1, \ldots, 5$ is obtained. Each entry corresponds to the PE of a inner test set and a certain number of components. Then for each γ_i the average PE of the inner test sets is separately calculated for the five different numbers of components $(\overline{\mathrm{PE}}_{\cdot n_c}(\gamma_i))$ $(n_c = 1, \ldots, 5)$. Hence for each γ_i, the optimal number of components $(n_{c_{i(opt)}})$ is selected according to the smallest mean PE over the r_{inner} test sets: $n_{c_{i(opt)}} = \arg \min_{\{n_c|1,\ldots,5\}} (\overline{\mathrm{PE}}_{\cdot n_c}(\gamma_i))$. Now the final optimal γ of version A denoted by γ^A_{opt}, corresponds to the minimal mean PE of the inner test sets $\gamma^A_{opt} := \arg \min_{\{\gamma_i|1,\ldots,\frac{1}{s_\gamma}+1\}} (\overline{\mathrm{PE}}_{\cdot n_{c_{i(opt)}}}(\gamma_i))$ and the corresponding optimal number of components of version A is than denoted by $n^A_{c_{(opt)}}$. So $n^A_{c_{(opt)}}$ PPLS-DA components are calculated of the outer training set all with the power parameter γ^A_{opt}. These components are then the predictors of a LDA, which is used to predict the outer test set.

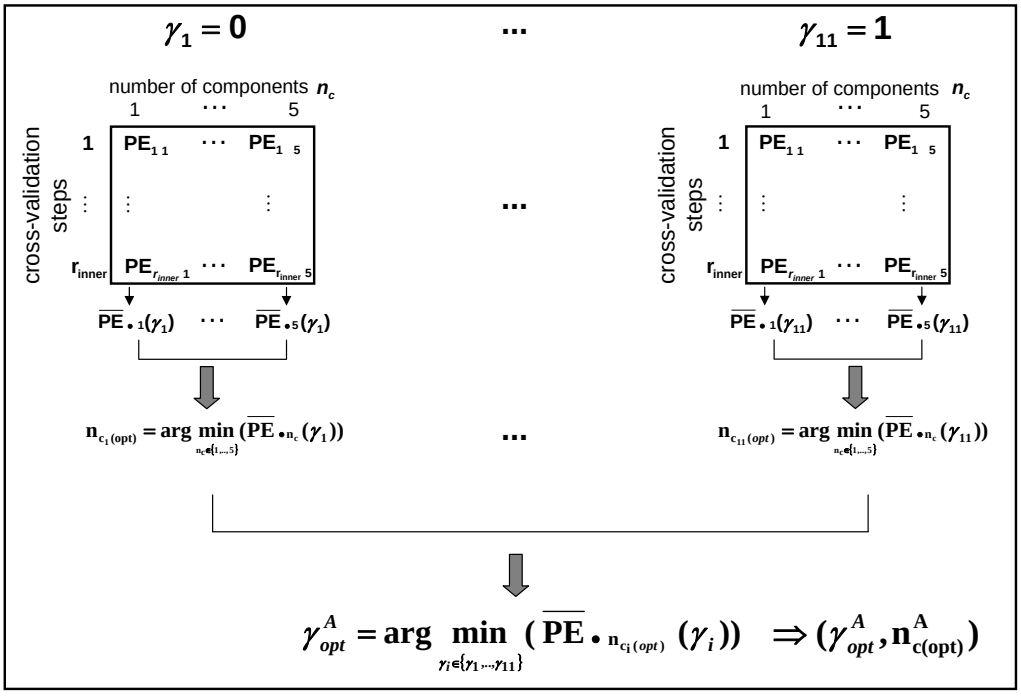


Figure 6.2: Illustration of the algorithm to find γ^A_{opt} and $n^A_{c_{(opt)}}$ of the extension of PPLS-DA according to version A with $s_\gamma = 0.1$.

6.2.2 Version A1

Opposite to version A, now only the first component is optimized (see Figure 6.3). In a first step, an $r_{inner} \times (\frac{1}{s_\gamma} + 1)$ matrix $(\mathrm{PE}_{t\,\gamma_i})_{t\,\gamma_i}$, with $t = 1, \ldots, r_{inner}$, $\gamma_i = \gamma_1, \ldots, \gamma_{\frac{1}{s_\gamma}+1}$ is evaluated regarding the PE of an inner test set according to a one component model ($n_c = 1$). Analogous to Version A, γ_{opt}^{A1} is the γ-value which leads to the smallest average PE over all inner test sets ($\gamma_{opt}^{A1} = \arg \min_{\{\gamma_i | 1, \ldots, \frac{1}{s_\gamma}+1\}} (\overline{\mathrm{PE}}_{\cdot\,\gamma_i})$). In a second step, the optimal number of components ($n_{c_{(opt)}}^{A1}$) is determined in an inner cross-validation with 10 repeats of the resampling of the inner training set and inner test set. In this step all components are calculated according to γ_{opt}^{A1}. Finally for each of the $n_{c_{(opt)}}^{A1}$ components based on the outer training set the loading weights vector is calculated according to γ_{opt}^{A1}. Then, the PE for the outer test set is calculated, like in version A.

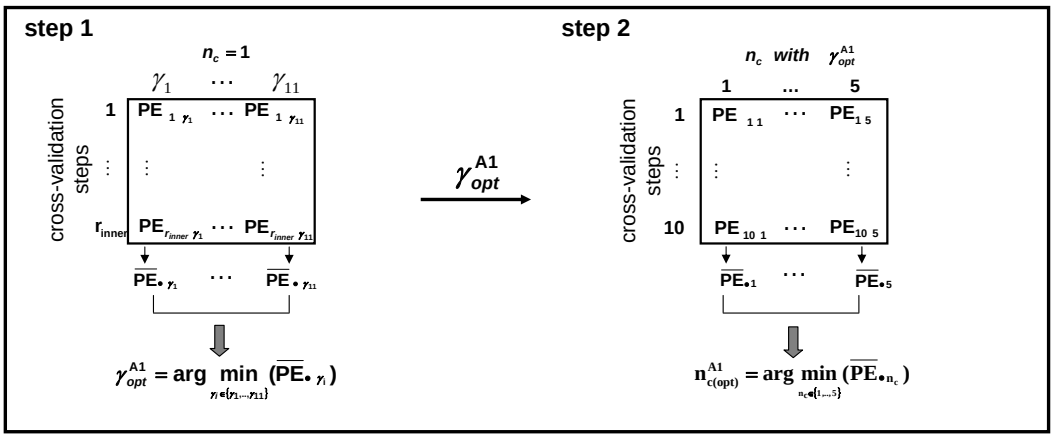


Figure 6.3: Illustration of the algorithm to find γ_{opt}^{A1} and $n_{c_{(opt)}}^{A1}$ of the extension of PPLS-DA according to version A1 with $s_\gamma = 0.1$.

6.2.3 Version A2

In version A and version A1, every loading weights vector is calculated with the same optimal γ. Now, the power parameter γ is optimized for each single component resulting in five optimal γ-values $\gamma_{opt}^{A2} = (\gamma_{opt,1}^{A2}, \ldots, \gamma_{opt,5}^{A2}) \in [0, 1]^5$. This is

illustrated in Figure 6.4. In a first step, a $r_{inner} \times (\frac{1}{s_\gamma} + 1)$ matrix $(\mathrm{PE}_{t\,\gamma_i})_{ti}$, with

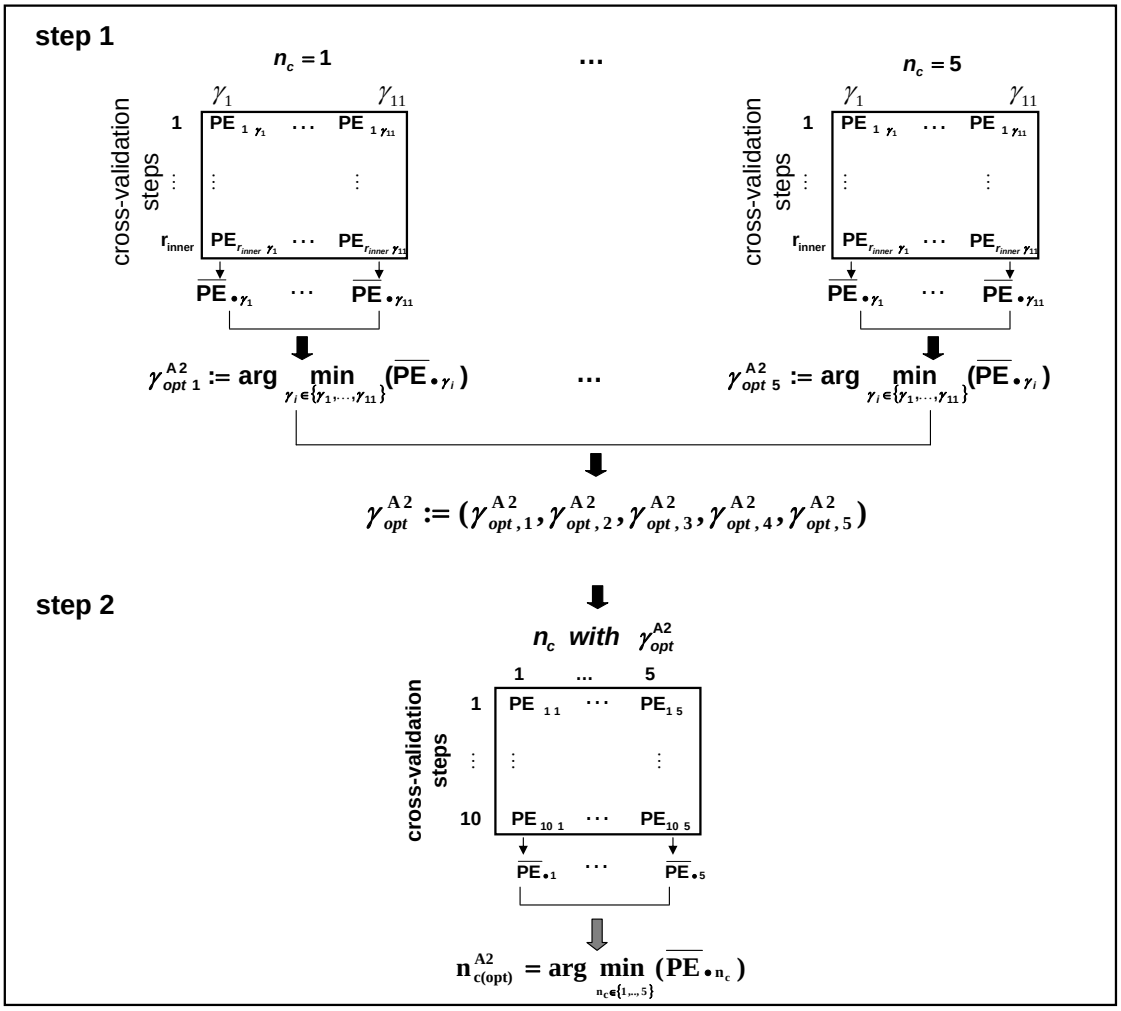


Figure 6.4: Illustration of the algorithm to find γ^{A2}_{opt} and $n^{A2}_{c_{(opt)}}$ of the extension of PPLS-DA according to version A2 with $s_\gamma = 0.1$.

$t = 1, \ldots, r_{inner}$, $\gamma_i = 1, \ldots, \frac{1}{s_\gamma} + 1$ is achieved with PE entries for each component separately. For the first loading weights vector, the optimal γ-value is calculated according to the lowest average PE of all inner test sets based on a one component model ($n_c = 1$). Next, the optimal γ is determined for the second component in a two component model with a first component calculated using $\gamma^{A2}_{opt,1}$. This means, $\gamma^{A2}_{opt,2}$ is the γ value leading to the minimal mean PE of the inner test sets based on a two component model using $\gamma^{A2}_{opt,1}$ and $\gamma^{A2}_{opt,2}$.

The following three loading weights are calculated analogously. Therewith, five op-

timal γ-values ($\gamma_{opt}^{A2} = (\gamma_{opt,1}^{A2}, \ldots, \gamma_{opt,5}^{A2}) \in [0,1]^5$) are selected, one for each of the five components. In a second step, the optimal number of components ($n_{c_{(opt)}}^{A2}$) in an inner cross-validation of the outer training set with 10 repeats is determined. The $n_{c_{(opt)}}^{A2}$ components are calculated for the outer training set using the corresponding optimal power parameter values of γ_{opt}^{A2}. Finally the PE of the outer test set is calculated.

6.2.4 Version B

The R-function predict.lda, which is used to predict group memberships for new samples according to their feature values, yields among other values the posterior probabilities $p_{i,1}$, $p_{i,2}$ for the groups $\nu = 1, 2$ for each sample i of the corresponding test set. A mean squared deviation of these posterior probabilities from the corresponding dummy vector ($y_{i\nu}$ is one if sample i belongs to group ν, otherwise $y_{i\nu}$ is zero) is used as a new measure for optimization: $z = \frac{1}{2}\sum_{i=1}^{n_{test}} \left((p_{i,1} - y_{i1})^2 + (p_{i,2} - y_{i2})^2\right) / n_{test}$, where n_{test} is the number of samples of the considered test set. All further steps are analogous to Version A2, but the optimization problem is to minimize z instead of the PE to find the optimal power parameter γ_{opt}^{B}.

6.3 Results

The results for the simulated data for scenario 1 are based on 100 repeated simulations, and for the experimental data $r = 100$ different outer training and outer test sets are sampled. In the following the mean PE values and the corresponding 95% confidence intervals are calculated for the outer test sets. The optimal number of components is also calculated as average over the 100 repetitions.

In this Section, the results of PPLS-DA using γ_{max} and the extensions (A, A1, A2 and B) of PPLS-DA are described and compared among each other, followed by a comparison with the ordinary PLS-DA.

6.3.1 Results for the simulated data

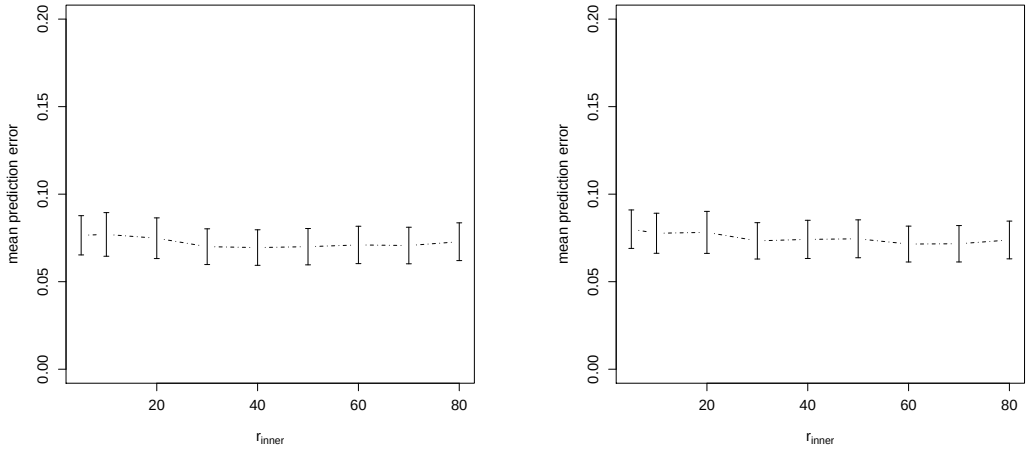
Choice of optimization settings r_{inner} and s_γ

At first more in detail scenario 1 with 1% ISF, a mean group difference of $\delta = [0.1, 0.5]$ and a technical noise of $\sigma_t^2 = \sigma_b^2$ (case 3) is considered. This scenario of differentially expressed features is chosen with a mean class difference between $[0.1, 0.5]$ and a large technical noise level, because this is most realistic for experimental data. The dependency of γ on the optimization of the step size s_γ, and on the number of internal cross-validation steps r_{inner} is shown exemplarily for version A.

Figure 6.5 illustrates the mean PE results for case 3 of scenario 1. Figure 6.5(a) shows the results for a step size (s_γ) of 0.05 plotted against the number of inner cross-validation steps (r_{inner}) and Figure 6.5(b) shows the corresponding findings for $s_\gamma = 0.1$. The straight line at 0.217 depicts the PE of PPLS-DA using γ_{max} and the corresponding 95% confidence interval. For this case, the influence of r_{inner} and s_γ can be neglected, because all confidence intervals for PPLS-DA using γ_{opt}^A overlap. Therefore, all further results of the simulated data are shown with following parameters of the optimization: $s_\gamma = 0.1$ and $r_{inner} = 50$.

PE results for simulated data

Table 6.1 summarizes the PE results for the simulated data and Table A.1 in Appendix A shows the corresponding 95% confidence intervals. The corresponding optimal numbers of components are shown in Table 6.2. Only scenario 1 is considered with 1% ISF, different mean class differences δ and different technical variance $\sigma_t^2 \in \{0, \frac{1}{4}\sigma_b^2, \sigma_t^2\}$. PPLS-DA using γ_{max} shows significantly larger PEs than PPLS-DA for all four extensions for the optimization of γ, except for case 4 of scenario 1



(a) (b)

Figure 6.5: Mean PE of PPLS-DA for simulated data using γ_{opt}^{A} plotted against r_{inner} for $s_\gamma = 0.1$ (a) and $s_\gamma = 0.05$ (b) .

(with 1% ISF, $\delta = 0.2$ and $\sigma_t^2 = \sigma_b^2$). Especially for DEGs with $\delta \in [0.1, 0.5]$, the PE of PPLS-DA with γ_{opt}^{A} is only one-tenth of the PE for PPLS-DA using γ_{max} in the case without noise, one-fifth for a minor noise level ($\sigma_t^2 = \frac{1}{4}\sigma_b^2$) and still one-third for a large noise level with $\sigma_t^2 = \sigma_b^2$.

Comparing all extensions among each other, the PEs are equal. Only for case 4 of scenario 1 with 1% ISF, with the low mean class difference ($\delta = 0.2$), equal PEs are found for PPLS-DA using γ_{opt}^{A}, γ_{opt}^{A1} and γ_{opt}^{A2}. And they are significantly lower than the PE of PPLS-DA using γ_{opt}^{B}.

For all simulated data, for all extensions of PPLS-DA it is optimal to use between 1.5 and 3.1 components. While PPLS-DA using γ_{max} leads to optimal numbers of components between 2.4 and 2.9.

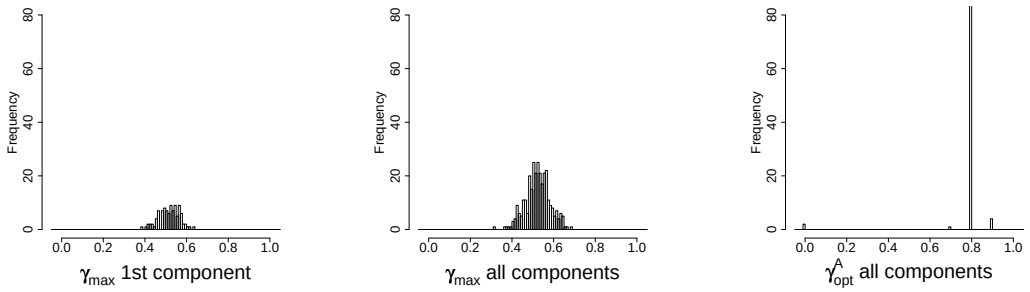
Table 6.1: Mean PE for scenario 1 with 1% ISF for $r_{inner} = 50$ and s_γ=0.1

scenario 1			PE of PPLS-DA with						PE of PLS-DA
case	σ_t^2	δ	γ_{max}	γ_{opt}^{A}	γ_{opt}^{A1}	γ_{opt}^{A2}	γ_{opt}^{B}	$\gamma = 0.5$	
1	0	[0.1,0.5]	0.097	0.018	0.016	0.017	0.017	0.102	0.102
2	$\frac{1}{4}\sigma_b^2$	[0.1,0.5]	0.121	0.027	0.028	0.027	0.028	0.134	0.136
3	σ_b^2	[0.1,0.5]	0.217	0.076	0.075	0.073	0.073	0.234	0.231
4	σ_b^2	0.2	0.377	0.326	0.338	0.317	0.403	0.388	0.385
5	σ_b^2	0.5	0.059	0.006	0.006	0.006	0.007	0.060	0.059

Table 6.2: The mean number of components used for simulated data for $r_{inner} = 50$ and n_γ=0.1

simulated data			PPLS-DA with						PLS-DA
case	σ_t^2	δ	γ_{max}	γ_{opt}^A	γ_{opt}^{A1}	γ_{opt}^{A2}	γ_{opt}^B	$\gamma = 0.5$	
1	0	[0.1,0.5]	2.6	1.85	2	2.1	2.9	2.7	2.7
2	$\frac{1}{4}\sigma_b^2$	[0.1,0.5]	2.8	2.0	2	2.1	3	2.7	2.9
3	σ_b^2	[0.1,0.5]	2.9	2.0	2.4	2.6	2.5	2.8	2.9
4	σ_b^2	0.2	2.7	2.5	2.6	2.7	2.3	2.7	2.7
5	σ_b^2	0.5	2.4	1.7	1.6	1.5	3.1	2.5	2.5

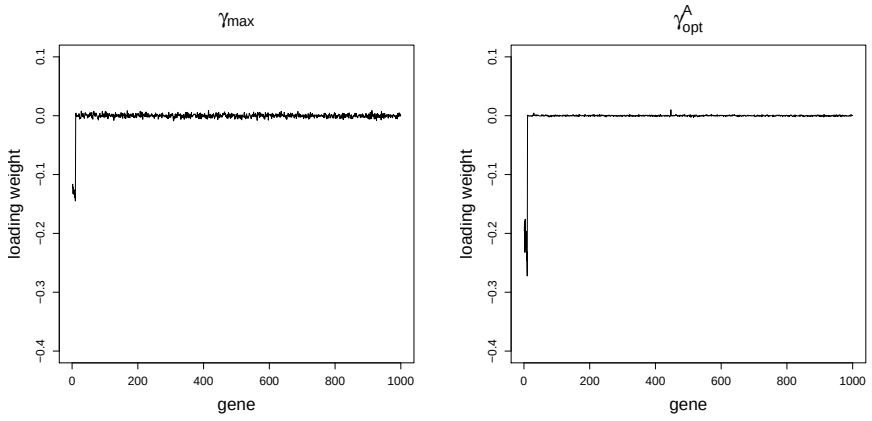
Considering the frequency distributions of the γ-values for γ_{opt}^A, Figure 6.6(c) shows the corresponding histograms for case 3 of scenario 1 with 1% ISF. For version A a modal value of 0.8 is shown. This is in contrast to PPLS-DA using γ_{max} which delivers values between 0.38 and 0.65 with similar frequencies (Figure 6.6(a)).



(a) (b) (c)

Figure 6.6: Histograms of γ_{max} and γ_{opt}^A for case 1 of scenario 1.

Figure 6.7 illustrates the dependency of the choice of the optimal power parameter on the loading weights. For r_{inner}=50 and s_γ=0.1 for scenario 1 with 1% ISF and case 3, the loading weights of the first component with γ_{max} and γ_{opt}^A are shown. The first 10 genes, which are simulated as differentially expressed, receive the highest absolute loading weight values for PPLS-DA using γ_{max} (Figure 6.7(a)) and γ_{opt}^A (Figure 6.7(b)). For γ_{opt}^A, these loading weights of the informative genes are increasing in absolute values, and especially the non-informative genes receive loading weight values closer to zero in comparison to the loading weights induced by γ_{max}.



(a) (b)

Figure 6.7: Average loading weights of the first component for scenario 1 with 1% ISF and $\sigma_t^2 = \sigma_b^2$ and $\delta = [0.1, 0.5]$ (case 3). Loading weights for the first component as calculated by PPLS-DA are shown with the power parameter γ_{max} (a) and γ_{opt}^A (b) using 50 inner cross-validation steps and a step size of $s_\gamma = 0.1$. The basis are the results of 100 choices of the outer training and outer test set.

Comparing PLS-DA and PPLS-DA using $\gamma = 0.5$ or γ_{max} (Table 6.1), equal PEs are found for all cases of the simulated data. Therewith the extensions show significantly lower PEs than PLS-DA, except for case 4 and PPLS-DA using γ_{opt}^B. The number of components used for PLS-DA is equal to the corresponding number of components used for PPLS-DA using γ_{max} or $\gamma = 0.5$. Considering the number of components, mostly it is a little larger for PLS-DA than for the extentions, except for extension B (Table 6.2).

6.3.2 Results for the experimental data sets

Now the results for the experimental data sets are shown which base on a cross-validation approach (see Section 2.1.2) with $r = 100$ repetitions of the sampling into the outer training set and outer test set.

Influence of r_{inner}

As for the simulated data, first the influence of the step size (s_γ) and the number of inner cross-validation steps (r_{inner}) for the determination of the optimal power parameter are tested prior to analyzing the experimental datasets.

Figure 6.8 shows the mean PE plotted against r_{inner} and $s_\gamma = 0.1$ for the Leukemia data set. Analogous to the simulated data, the 95% confidence intervals for the PEs of the outer test sets overlap for all considered numbers of r_{inner}, so that again $r_{inner} = 50$ is chosen. Because the 95% confidence intervals for the PE for $s_\gamma = 0.1$ and $s_\gamma = 0.05$ overlap (data not shown), the step size is set to $s_\gamma = 0.1$ as for the simulated data.

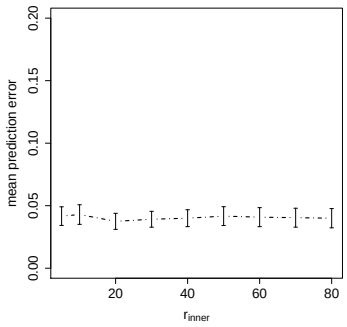


Figure 6.8: Mean PE of PPLS-DA for the Leukemia data using γ^A_{opt} plotted against r_{inner} with step size $s_\gamma = 0.1$.

Influence of the proportion of the inner training set and inner test set

For the proportion of the outer training set on the whole experimental data set, $\phi_{\mathbf{X}} = 0.7$ is applied. Therewith the mean PE results of the outer test set base on the same sample size for all different choices of $\phi_{\mathbf{X}_{train}}$. As a consequence, any bias due to different sample sizes of the outer test set is avoided.

For the experimental data sets, the effect of the proportion of the inner training set and inner test set is studied for the values $\phi_{\mathbf{X}_{train}} \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$. For

Version A, the effect of the choice of $\phi_{\mathbf{X}_{train}}$ on the outer test set PE is shown only exemplarily for the Leukemia data set (Figure 6.9), because this method is most time-consuming compared to the other extensions. For this data set, the 95% confidence intervals overlap for $\phi_{\mathbf{X}_{train}} = 0.5, 0.6, 0.7, 0.8$, only for $\phi_{\mathbf{X}_{train}} = 0.9$ a significant larger outer test set PE is shown.

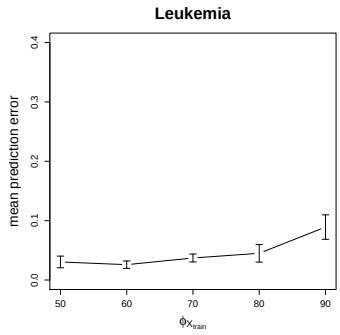


Figure 6.9: Mean PE of PPLS-DA using γ_{opt}^{A} with 95% confidence intervals for different $\phi_{\mathbf{X}_{train}}$ values for the Leukemia data set.

The findings for version A1 are shown in Figure 6.10, for different choices of $\phi_{\mathbf{X}_{train}}$ all experimental data sets show no significant differences in the outer test set PEs. For the Lymphoma data set the outer test set PE varies most for the different proportions of the inner training set on the outer training set ($\phi_{\mathbf{X}_{train}}$). Especially for the Prostate 2 data set, nearly constant PEs are found for different $\phi_{\mathbf{X}_{train}}$ choice.

For version A2 and the Leukemia, Lymphoma, Prostate 1 and Prostate 2 data set, the influence of $\phi_{\mathbf{X}_{train}}$ is illustrated in Figure 6.11. Also for this version no significant differences in the outer test set PEs are shown between different choices of $\phi_{\mathbf{X}_{train}}$.

For version B, Figure 6.12 presented the dependency of the PE for the outer test set on the choice of $\phi_{\mathbf{X}_{train}}$ for the Prostate 1 data set and the Prostate 2 data set. Only very few differences are found for the Prostate 1 data set, but they are not significant. For the Prostate 2 data set, the PE decreases slightly for $\phi_{\mathbf{X}_{train}} = 0.7$ to $\phi_{\mathbf{X}_{train}} = 0.9$, but no significant differences are shown.

Summarizing, for the versions A1, A2 and B no significant differences are shown

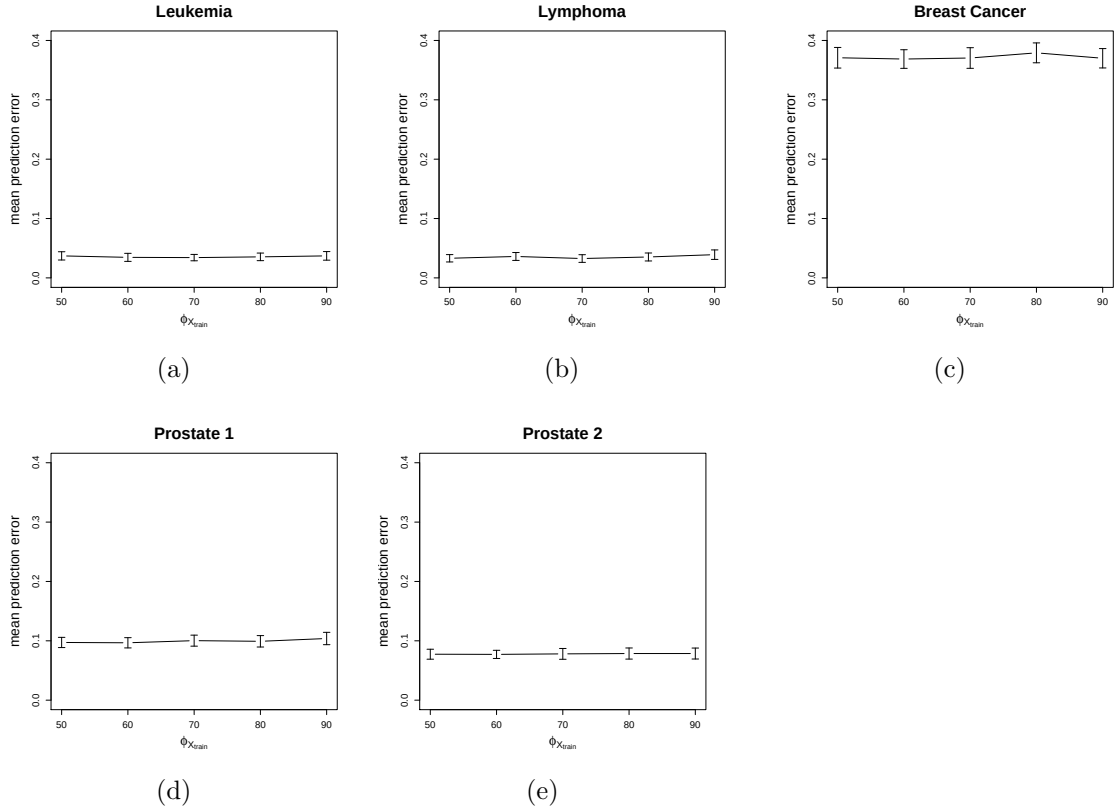
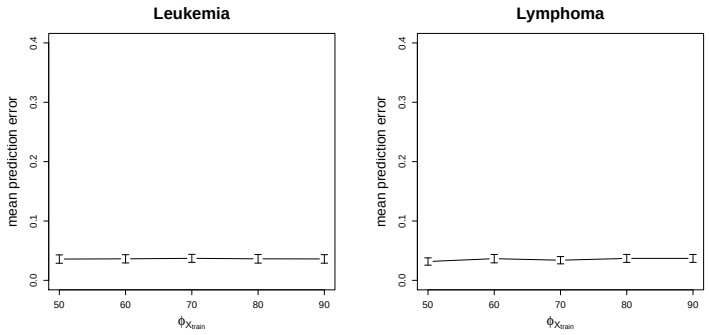
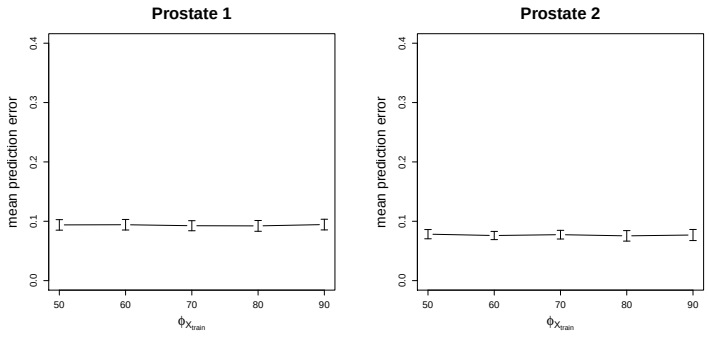


Figure 6.10: Mean PE of PPLS-DA using γ_{opt}^{A1} with 95% confidence intervals for different $\phi_{\mathbf{X}_{train}}$ values for the five experimental data sets.

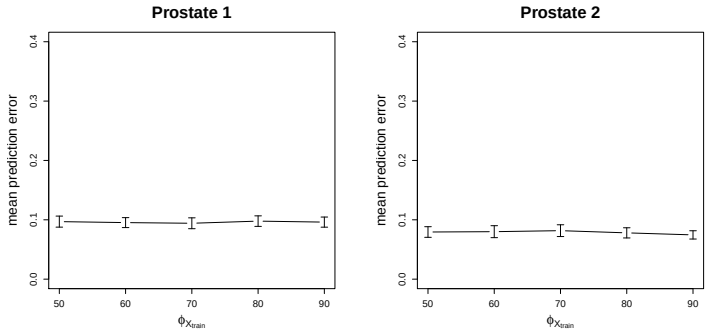


(a) (b)



(c) (d)

Figure 6.11: Mean PE of PPLS-DA using γ_{opt}^{A2} with 95% confidence intervals for different $\phi_{\mathbf{X}_{train}}$ values for four experimental data sets.



(a) (b)

Figure 6.12: Mean PE of PPLS-DA using γ_{opt}^{B} with 95% confidence intervals for different $\phi_{\mathbf{X}_{train}}$ values for two experimental data sets.

regarding the PE of the outer test set for different choices of $\phi_{\mathbf{X}_{train}}$. Considering the Leukemia data set, for version A and for values of $\phi_{\mathbf{X}_{train}} = 0.5, 0.6, 0.7, 0.8$ also no significant differences are found, but in comparison for $\phi_{\mathbf{X}_{train}} = 0.9$ a significant higher PE is shown. Therefore all further results are calculated with $\phi_{\mathbf{X}_{train}} = 0.7$.

PE results using the parameter $r_{inner} = 50$, $s_\gamma = 0.1$ and $\phi_{\mathbf{X}_{train}} = 0.7$

Table 6.3 contains the mean PE results for all five experimental data sets for all considered methods, and corresponding 95% confidence intervals are shown in Table A.2 in Appendix A. For these results, the average numbers of components used are shown in Table 6.4.

Table 6.3: Mean PE for experimental data with $r_{inner} = 50$ and n_γ=0.1

	PE of PPLS-DA using						PE of PLS-DA
data set	γ_{max}	γ_{opt}^{A}	γ_{opt}^{A1}	γ_{opt}^{A2}	γ_{opt}^{B}	$\gamma = 0.5$	
Leukemia	0.056	0.037	0.034	0.037	0.038	0.028	0.032
Lymphoma	0.044	0.029	0.033	0.034	0.033	0.033	0.029
Breast Cancer	0.346	0.382	0.370	0.372	0.365	0.380	0.374
Prostate 1	0.089	0.084	0.100	0.093	0.106	0.081	0.080
Prostate 2	0.076	0.079	0.078	0.077	0.082	0.213	0.216

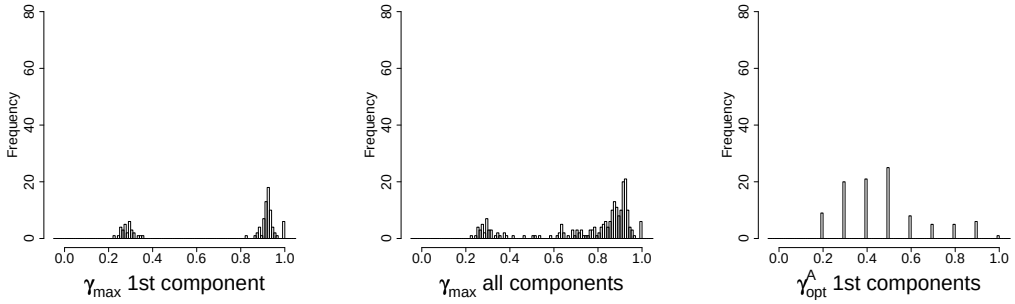
Table 6.4: The mean number of components used for the experimental data sets

	PPLS-DA using						PLS-DA
data set	γ_{max}	γ_{opt}^{A}	γ_{opt}^{A1}	γ_{opt}^{A2}	γ_{opt}^{B}	$\gamma = 0.5$	
Leukemia	2.2	2.2	2.3	2.3	3.1	1.8	2.2
Lymphoma	2.5	3.3	3.0	2.9	3.2	2.4	2.8
Breast Cancer	2.4	2.2	1.9	2.6	2.5	2.3	2.4
Prostate 1	2.5	3.4	1.8	3.0	3.1	4.0	4.1
Prostate 2	2.9	3.7	2.9	3.5	3.5	4.2	4.3

Leukemia data set

For this data set, the PE of PPLS-DA using γ_{max} (0.056) is significantly larger than for the extensions A, A1 and A2 of PPLS-DA with PEs between 0.034 and 0.037. The PE of PPLS-DA using γ_{opt}^{B} is not significantly different to the PE of PPLS-DA

using γ_{max}. The extensions show no significant differences among each other with respect to the PE. PPLS-DA using γ_{max} and all extensions use a similar number of components (in average between 2.2 and 2.3), except for PPLS-DA using γ_{opt}^{B}, which uses more components (in average 3.1). The histogram of γ_{opt}^{A}-values is shown in



(a) (b) (c)

Figure 6.13: Histograms of γ_{max} and γ_{opt}^{A} for the Leukemia data set. Values of γ_{max} detected by PPLS-DA for the first component (a) and for all components (b). In panel (c) the γ-values are shown, detected for γ_{opt}^{A} of version A for the extension of PPLS-DA.

Figure 6.13(c) with a modal value of 0.5. In comparison, in the histogram of γ_{max}-values (Figure 6.13 (a) and (b)), there are two accumulations points of γ_{max}-values for the first component, one around 0.3 and the other around 0.9. The same is true for most γ_{max}-values of all components.

Comparing the PE results of the extensions to those of PLS-DA, no significant differences are found. The PE of PPLS-DA using γ_{max} is significantly larger than the PEs of PPLS-DA using $\gamma = 0.5$ and PLS-DA. PPLS-DA with $\gamma = 0.5$ uses also in average the lowest number of components (1.8).

Lymphoma data set

The PE of PPLS-DA using γ_{opt}^{A} is significantly lower than the PE of PPLS-DA using γ_{max}. For all other extensions (A1, A2, B) the PEs are equal to the PE of PPLS-DA using γ_{max}. The averaged numbers of components used for all extensions are between 2.9 and 3.3. PPLS-DA with γ_{max} uses in average 2.5 components.

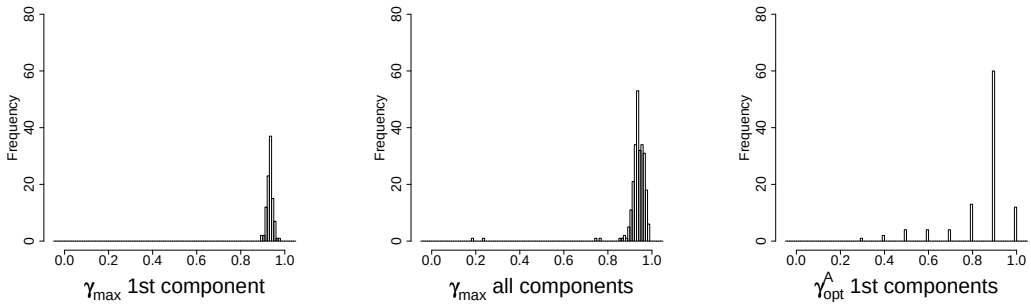
Considering PLS-DA, the PE is equal to the PEs of all extensions and of PPLS-DA using $\gamma = 0.5$, and also significantly lower than for PPLS-DA using γ_{max}. PPLS-DA with $\gamma = 0.5$ used the smallest average number of components (2.4), which is similar to the number of components used by PPLS-DA with γ_{max}.

Breast Cancer data set

For this data set, PPLS-DA using γ_{max} and the extensions A1, A2 and B show equal PEs, in average between 0.346 (PPLS-DA using γ_{max}) and 0.372 (PPLS-DA using γ_{opt}^{A2}). For extension A the PE is even larger than for PPLS-DA using γ_{max}. The number of components used is between 1.9 (PPLS-DA using γ_{opt}^{A1}) and 2.6 (PPLS-DA using γ_{opt}^{A2}).
Also PLS-DA shows a similar PE compared to PPLS-DA using γ_{max} or $\gamma = 0.5$ and to all extensions of PPLS-DA. The same is true for the number of components.

Prostate 1 data set

PPLS-DA using γ_{max} shows nearly an equal PE to PPLS-DA using γ_{opt}^{A}, γ_{opt}^{A1} and γ_{opt}^{A2}. For the extension B the PE is even larger than for PPLS-DA using γ_{max}. For all extensions, the optimal number of components is between 1.8 and 3.4 components while PPLS-DA using γ_{max} uses 2.5 components. Comparing the histogram of γ_{opt}^{A} and γ_{max} (Figure 6.14), nearly equal modal values are found which fits to the similar PE results.
Investigating PLS-DA, the PE is equal to the PE of PPLS-DA using γ_{max} or $\gamma = 0.5$. Compared to the PEs of the extensions, PLS-DA has an equal or lower PE. PPLS-DA using $\gamma = 0.5$ and PLS-DA show significantly lower PEs than PPLS-DA using γ_{opt}^{A1} and γ_{opt}^{B}. However PLS-DA and PPLS-DA using $\gamma = 0.5$ use the largest average number of components (4.1 and 4).



(a) (b) (c)

Figure 6.14: Histograms of γ_{max} and γ_{opt}^{A} for the Prostate 1 data set.

Prostate 2 data set

The PEs of the extensions of PPLS-DA are all equal to the PE of PPLS-DA using γ_{max}. Moreover, between 2.9 and 3.7 components are used.
PPLS-DA using $\gamma = 0.5$ and PLS-DA show equal PEs, but they are significantly higher than for PPLS-DA using γ_{max} and for all extensions. Considering the number of components used, PPLS-DA with $\gamma = 0.5$ and PLS-DA use more components (in average 4.2 and 4.3) in comparison to all extensions, and to PPLS-DA using γ_{max}.

Number of features with non-zero loading weights

Now the number of features which are really used for the calculation of the components is considered. Table 6.5 summarizes the mean number of features with a non-zero loading weight for PPLS-DA using γ_{max}, all proposed extensions of PPLS-DA, PPLS-DA using $\gamma = 0.5$ and PLS-DA. The second column of this table contains the total number of features (g). Over all data sets, most features have non-zero loading weights for PLS-DA. For the Leukemia and the Lymphoma data set no great differences are found in the number of features with non-zero loading weights between the extensions, PPLS-DA using $\gamma = 0.5$ and PLS-DA. For these data sets, PPLS-DA using γ_{max} leads to between 200 and 300 more features which have no

influence on the calculation of the components. Considering the Breast Cancer data set, only PPLS-DA using $\gamma = 0.5$ include all 4997 genes for the calculation of the components and extension A leads to the lowest number of features used (3881). For the Prostate 1 data set, great differences are shown in the number of used features for the calculation of the components between all methods. Version B used in average only around 842 features of 6033, in contrast PLS-DA and PPLS-DA with $\gamma = 0.5$ use all 6033 features. Also for the Prostate 2 data set with the large total number of genes (g=42129), in relation to that total number the power parameter does not lead to a clearly lower number of features with non-zero loading weights.

Table 6.5: The mean number of features with non-zero loading weight

data set	g	PPLS-DA using						PLS-DA
		γ_{max}	γ_{opt}^{A}	γ_{opt}^{A1}	γ_{opt}^{A2}	γ_{opt}^{B}	$\gamma = 0.5$	
Leukemia	3571	3266	3535	3533	3496	3533	3571	3571
Lymphoma	7129	6368	6685	6677	6681	6674	6685	6685
Breast Cancer	4997	4483	3881	4738	4879	4095	4997	4497
Prostate 1	6033	5556	5307	1022	3738	842	6033	6033
Prostate 2	42129	41015	41643	40462	42052	41034	42129	41015

6.4 Discussion

Comparison between PPLS-DA using γ_{max} and the extensions

The optimization criterion for the power parameter γ_{max} in the ordinary PPLS-DA is towards canonical correlation, and does not need to be best for prediction. The four extensions of PPLS-DA introduced optimize γ with respect to prediction using an inner cross-validation.

The PEs of the outer test sets for PPLS-DA using γ_{max} are improved or showed at least equal values by optimizing the power parameter with respect to prediction using LDA for at least three of the four versions of the extensions, for all simulated data and the experimental data sets (Table A.1 and Table A.2).

Simulated data Comparing the histograms of the optimal γ-values found by the extensions and PPLS-DA using γ_{max}, the reason for the lower PE for the extensions can be traced back to the down-weighting of the non-informative features for the simulated data (see Figure 6.6 and Figure 6.7). For the extensions, the loading weights for these features are near or equal to zero. The influence of the features, which are not informative for the discrimination (and can be interpreted as noise), is reduced, because the impact on the calculation of the components is lower for the extensions than for PPLS-DA using γ_{max}. Values of γ near to one lead to a preference of features which show a high correlation to the dummy response (in the simulation study these features are the differentially expressed genes). For the simulated data, the PEs for all PPLS-DA extensions are very similar, except for the simulated data with the low mean class difference (case 4: $\delta = 0.2$ and $\sigma_b^2 = \sigma_t^2$). Changing the optimization criterion from correlation towards prediction also leads to a lower average number of components for the simulated data, except for case 1 and case 2 for version B, which uses posterior probabilities of the LDA as objective function.

Experimental data For the experimental data, the γ-values determined by maximization of the canonical correlation (γ_{max}) are larger than the $\gamma-$values detected by the extensions of PPLS-DA. However, even if the analyses of simulated data show lower PE values for larger choices of γ, for the Leukemia data set and the Lymphoma data set the extension A shows a significantly lower PE than PPLS-DA using γ_{max}. For the Leukemia data set also the extensions A1 and A2 lead to significantly lower PEs.

For the Prostate 1 data set, PPLS-DA using γ_{max} shows a significantly lower PE than extension B. Extensions A, A1, A2 show equal PEs to those observed for PPLS-DA using γ_{max}. The PEs of PPLS-DA using γ_{max} and of three out of the four extensions are equal for the Breast Cancer set (except for extension A). For the Prostate 2

data set, all PEs of PPLS-DA using γ_{max} and the extensions are equal. Comparing the optimal power parameters γ_{opt}^{A1}, γ_{opt}^{A2} and γ_{opt}^{B} among each other for the Prostate 1 and the Prostate 2 data set, the γ-values for the first components are similar, as expected, but for all further components the optimal $\gamma-$values differ much more than for the Leukemia data set and the Lymphoma data set (data not shown).

Summarizing, the findings for the simulated data are equal for all extensions and show significantly lower PEs than PPLS-DA using γ_{max}, except for case 4 ($\Delta = 0.2$ and $\sigma_b^2 = \sigma_t^2$). For the experimental data, the results are also significantly lower or equal for the extensions considering the PE, except for the Prostate 1 data set for extension B and the Breast Cancer data set for extension A of PPLS-DA.

Comparing only PPLS-DA using γ_{max} and PPLS-DA with the extensions, for the experimental data, it can be concluded that the extensions only lead to smaller PEs than for PPLS-DA using γ_{max}, if the increase of the condition index (see Chapter 4) $\kappa_k, k = 1 \ldots, 5$ is weak and the proportion of DEGs is large (Leukemia and Lymphoma data sets). This is in contrast to the simulated data. The only explanations considered are the noise and the dependencies of the features in the experimental data which could not be adequately simulated.

Note, that the true informative genes for the experimental data are not known, and the proportion of differentially expressed genes most likely is much larger than for the simulated data, therefore the comparison of the results for simulated and experimental data is not straightforward. Moreover simply simulating a higher proportion of DEGs, leads not to similar results (data not shown). The part of reality which could not be modeled in the simulated data might be a sort of noise or a data structure that cannot be improved, regardless our choice of γ. If this part of the noise could be removed from the real data, the relative improvements might be just as good as with the simulated data.

Comparison between PPLS-DA using γ_{max}, $\gamma = 0.5$ and PLS-DA

The development of PPLS-DA followed the development of PPLS as a natural extension of the power methodology to handle discrete responses. Several factors motivated this advancement to PLS-DA. First, the application of powers enables focusing on fewer explanatory variables in the loading weights, smoothing over some of the noise in the remaining variables. Second, focus can be shifted between the correlation and standard deviation parts of the loading weights, which is even more important for discrete responses. Finally, the maximization criterion is moved from the between-group variation matrix (B) to the product of the between group variation matrix and the inverse of the within-group variation matrix ($\mathbf{W}^{-1}\mathbf{B} = \mathbf{T}^{-1}\mathbf{B}$). This has the effect of moving from covariance maximization to a correlation maximization.

In our study, PPLS-DA with $\gamma = 0.5$ (applying no power parameter) and PLS-DA show always equal PEs for the simulated and the experimental data sets. Hence, for this case the different optimization tasks show no great differences with respect to the PEs of the outer test sets, which can be understood because all data sets are standardized. Including a power parameter, the PE of PPLS-DA using γ_{max} is equal to the PE of PLS-DA for all simulated data. However, the number of components used is in average lower or equal for PPLS-DA using γ_{max} than for PLS-DA.

For two of the five experimental data sets (Leukemia and Lymphoma), the PEs of PPLS-DA using γ_{max} are significantly higher than the PEs of PLS-DA. For these data sets, the proportions of differentially expressed genes are large (24.4% and 40.5%) and the genes are only weakly linear dependent (considering the condition indexes $\kappa_k, k = 1, \ldots, 5$). Otherwise, for the Prostate 2 data set the PE of PPLS-DA using γ_{max} is significantly lower than for PLS-DA, and the number of components used is also in average lower. This data set contains only a low proportion of differentially expressed genes (1.4%), and the total number of genes is very high (42129). Moreover, for this data set, the genes show a stronger linear dependency (rapid in-

crease of the condition index) than for the Leukemia or the Lymphoma data set. Summarizing, data sets with a weak increase of $\kappa_k, k = 1 \ldots, 5$ indicate no improvement for the PE when using PPLS-DA using γ_{max} instead of PLS-DA. Concerning percentage of DEGs, the PE of PPLS-DA using γ_{max} is equal to the PE of PLS-DA only for a small percentage of DEGs (Breast Cancer data and the simulated data with scenario 1 with 1% ISF and case 3). For a weak increase of $\kappa_k, k = 1 \ldots, 5$ and a high percentage of DEGs, the PE for PPLS-DA using γ_{max} is even larger than for PLS-DA (Leukemia and Lymphoma data sets).

A rapid increase of the condition index $\kappa_k, k = 1 \ldots, 5$ and a large proportion of DEGs (Prostate 1), using PPLS-DA with γ_{max} instead of PLS-DA does not improve the PE. On the contrary, for a rapid increase of the condition index κ_k and the case of a small percentage of DEGs (Prostate 2), the PE can be improved by employing PPLS-DA using γ_{max} instead of PLS-DA.

Comparison between PLS-DA and the extensions of PPLS-DA

Considering the simulated data, the PEs of the extensions A, A1 and A2 are significantly lower than the PE of PLS-DA, except for case 4. For the experimental data, for three of the five data sets (Leukemia, Lymphoma and Breast Cancer), the PEs of PLS-DA and the extensions of PPLS-DA are equal. The PEs are also equal for the Prostate 1 data set for PLS-DA and PPLS-DA using γ_{opt}^{A} and γ_{opt}^{A2}. The PEs of PPLS-DA using γ_{opt}^{A1} and γ_{opt}^{B} are significantly larger than for PLS-DA, for the Prostate 1 data set. For the Prostate 2 data set, the PEs of all extensions are clearly lower than for PLS-DA.

First it is concluded, that equal PEs between PLS-DA and the extensions of PPLS-DA are caused by a weak between-feature dependence (weak increase of κ_k), independent of the proportion of DEGs. Second, a data set with a strong collinearity between the features and a low number of DEGs, in contrary shows a clearly lower PE for the extensions than for PLS-DA.

Comparison between PPLS-DA using $\gamma = 0.5$ and PLS-DA

The PEs for PLS-DA and PPLS-DA using $\gamma = 0.5$ are equal for all simulated and all experimental data sets investigated. Maximization of the covariance or maximization of the correlation without the power parameter, results in equal PEs of the outer test set.

Conclusions and Outlook

Data sets with a high proportion of differentially expressed gens and a weak linear dependency (like the Leukemia data set and the Lymphoma data set) most probably show good prediction results for PLS-DA. There is no gain using PPLS-DA with powers (γ_{max} or extensions) here. On the contrary, for a rapid increase of the condition index, a low proportion of differentially expressed genes and a large total number of genes, using PPLS-DA with γ_{max} clearly improves the prediction error compared to PLS-DA. In cases where PPLS-DA using γ_{max} gives no advantages over PLS-DA, using the extensions of PPLS-DA (optimizing the power parameter) for prediction can be advantageous. One aspect of future work is to validate these conclusions by additional experimental data sets as well as further simulations implementing a more complex covariance structure.

7 Summary and Outlook

The aim of this thesis was to study aspects of classification studies with regard to biomarker search. Biomarker search has become very important in the last years. Especially in cancer research the hope is to identify promising molecular biomarkers based on gene expression and metabolite concentration. It is an important task to choose the right design, here a distinction was made between single sample design and pooling design. Also the application of statistical learning methods towards biomarker search is a major challenge. These areas are reflected in the three parts of this thesis.

In the first part, an example of biomarker search is given in the context of tuberculosis to demonstrate the course of identification of a biomarker (Chapter 3). The goal is to identify a biomarker which enables the classification between metabolites profiles of tuberculosis patients and persons infected but still healthy. Comparing different classification methods by their PEs on the metabolite data set, the method RF shows the lowest PE and is therefore chosen for analyzing the importance of the metabolites for the classification. In detail, the mean decrease accuracy, a variable importance value of RF, is used to order all metabolites in a ranking list. Taking a cross-validation approach, a biomarker is determined containing 19 metabolites. To check the resulting biomarker with an other method, the classification method with the second lowest PE, PPLS-DA, is also used to determine a biomarker. The result is a biomarker containing only 8 metabolites. The PEs achieved with both

biomarkers are similar, even so PPLS-DA needs less than half of the features. Moreover the biomarker detected by PPLS-DA is a subset of the biomarker detected by RF. The specialness of the proposed approach is, that the influence of the mixture of the training set is taken into account. The variable importance values are calculated for different samples of the inner training set, and then the mean value of each importance value is used to create the final ranking list. In the case of this data set, RF and especially PPLS-DA turn out to be very promising for successful biomarker search with respect to a low PE and a small cardinality of the biomarker. In a next step the biomarker needs to be verified with independently measured new samples.

One main aspect of biomarker search is the design choice. This essential decision for the experiment is discussed in detail in the second part of Chapter 5. Here the consequences of pooling for classification studies are investigated. In practice biologists or medical doctors often apply pooled samples, because of financial restrictions or insufficient amount of cDNA for a single sample hybridization. The consequences of pooling designs are always presented in comparison to a single sample design for simulated data and experimental data which are artificially pooled. In this work it is distinguished between feature patterns, for which the convex hull of one group is not a cover of the other group (pattern type I) and feature patterns for which it is a cover (pattern type II). The influence of a pooling design depends on the underlying data structure, the pool size and the classification method used. For simulated data with pattern type I, such as differentially expressed genes, PLS-DA and especially PPLS-DA are most robust against pooling with respect to the PE and the set of important features. In contrast for patterns of type II, presented in form of a threshold pattern, pooling leads to clearly higher PEs especially for a larger pool size. Moreover the intersection of important features sets for a single sample design and a pooling design contains only a few features. That means in pooling designs features forming a threshold are difficult to identify as a biomarker. The method RF shows the best results for pattern type II, but overall a strong dependence from

the pool size is visible.

Additionally experimental data are artificially pooled and the PEs of classification methods are used to decide which classification method is most robust against pooling. For the four experimental data sets investigated the lowest pooling effect on the PE is found using PPLS-DA and PLS-DA.

As recommendation for praxis, a single sample design should be preferred for biomarker search like suggested in the literature. If pooling cannot be avoided, pool sizes as small as possible should be chosen and the statistical learning methods PPLS-DA or RF should be considered first for the detection of biomarkers.

A main point of future work is to analyze experiments where (both) single sample and pooled sample data are available from the same objects. Such an analysis was unfortunately not possible because no suitable experimental data for this comparison have been found.

In Chapter 3 it turns out, that PPLS-DA delivers a biomarker for the discrimination of tuberculosis patients and infected but healthy persons with a few features. Further in Chapter 5 this method shows the lowest pooling effect, considering simulated and experimental data sets. Therefore PPLS-DA is refined to improve the prediction accuracy in a third part of this thesis (Chapter 6). A wrapper approach is proposed in four different versions to include the classification method LDA into the optimization of the power parameter γ. For simulated data a great improvement of the PE is achieved with the extensions. For the experimental data sets investigated, the reduction of the PE also succeeded, but not for all data sets and the differences are smaller. A R-package including the extensions of PPLS-DA is in work.

In this thesis only the method LDA is applied as final classification method. Of course other classification methods can be used, like SVM. The extensions can also be adapted for PPLS for regression, only the objective function has to be changed for example towards minimizing the root mean square error. Moreover the presented

four extensions of PPLS-DA can be used as starting point of further modifications of this method. For example, the optimization problem can be changed to further reduce the influence of features without information for the discrimination between the groups. One possibility is to add a cut-off value, that means that absolute loading weights smaller than this value are set to zero. Hence features with a small absolute loading weight which can be considered as noise, have no longer impact on the components. Another direction for further improvement is the modification of the measure of variable importance. In this thesis the absolute loading weight value is used exclusively as variable importance measure for PPLS-DA using γ_{max}. If instead as variable importance measure the absolute loading weight value averaged over the components and weighted with the coefficients $\mathbf{b}$ of the LDA is used, no improvement can be achieved (data not shown). However for PPLS-DA with the optimization of the power parameter with respect to prediction, this approach might improve the results.

PPLS-DA is a linear feature extraction technique, therefore pattern in form of a threshold pattern cannot be identified for the discrimination of the groups. To enable also the detection of this kind of pattern, applying first a kernel trick can be a solution. In the whole thesis only two-group classification problems are discussed. However the investigations can be extended to more than two groups. Of course the pattern types considered for the choice of the experimental design can also be formulated for three groups and similar conclusions are expected. The extensions of PPLS-DA can also be used in this case, because PPLS-DA and LDA work for more than two groups. Only the number of components considered in the extended versions of PPLS-DA should be increased, for an increasing number of groups.

Further metabolite data and mainly microarray data are investigated. However in the last years, sequencing of DNA molecules (the sequence of the nucleotide bases are determined), becomes more and more popular. This kind of method has several advantages, but until now it is very expensive and microarray analysis is still an alternative method with similar performance. It remains for further work to check

if for sequencing data also the use of RF and PPLS-DA is advantageous.
Moreover for specialized questions of prediction, it can be advantageous to split the prediction error into two measures. Considering a classification between infected and non-infected persons, then a measure accounts for both the infected patients classified as non-infected (false negative rate) and non-infected persons declared as infected (false positive rate). Also other more detailed measures like sensitivity (true positive rate) and specificity (true negative rate) can be applied. Then it can be checked if among the new prediction measures the methods, RF and PPLS-DA, are still preferable.

In this thesis three aspects of biomarker search are considered to point out the importance of statistical approaches in this research area. The presented results suggest the two classification methods, PPLS-DA and RF as promising for biomarker search.

A Appendix

Additional Figures

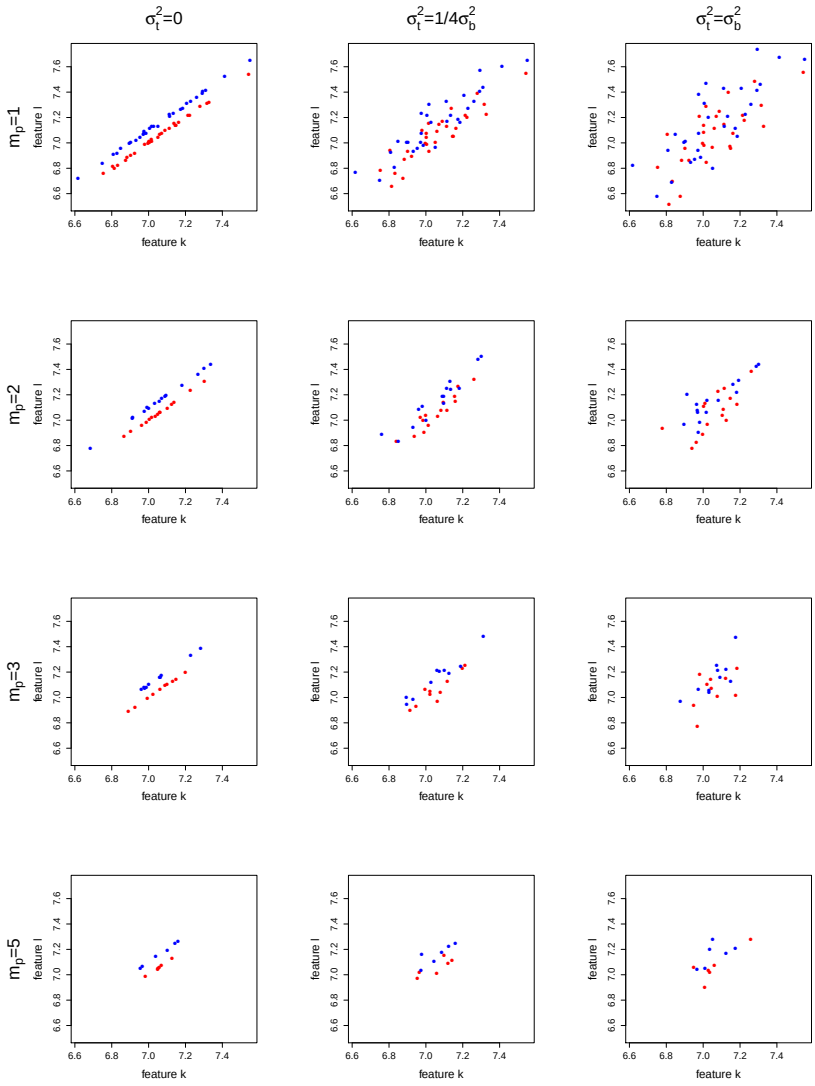


Figure A.1: Illustration of scenario 3 (two linear dependent features) before pooling and after pooling.

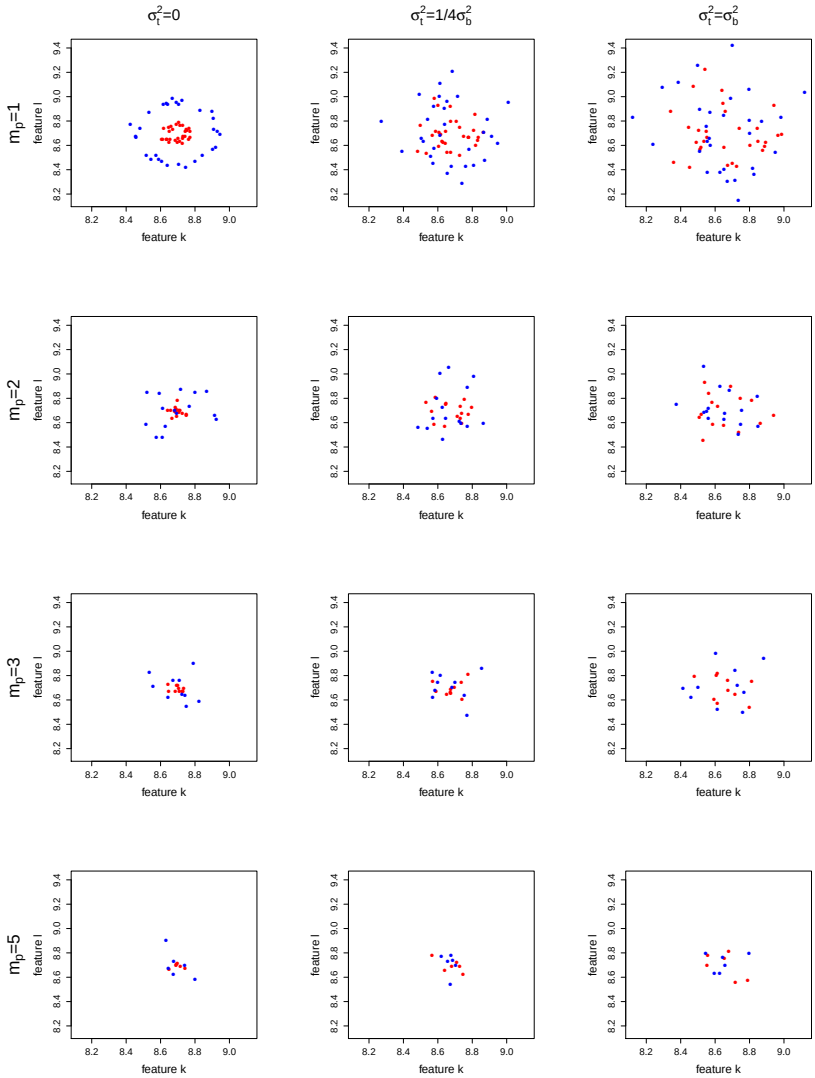


Figure A.2: Illustration of scenario 4 (circle pattern) before pooling and after pooling.

Additional Tables

Table A.1: 95% confidence intervals for scenario 1 with 1% ISF for $r_{inner} = 50$ and s_γ=0.1.

case	γ_{max}	PE of PPLS-DA using γ_{opt}^{A}	γ_{opt}^{A1}	γ_{opt}^{A2}	γ_{opt}^{B}	$\gamma = 0.5$	PE of PLS-DA
1	[0.0845,0.1089]	[0.0130,0.0237]	[0.0122,0.0205]	[0.0123,0.0213]	[0.0121,0.0212]	[0.0899,0.1131]	[0.0906,0.1138]
2	[0.1082,0.1341]	[0.0215,0.0322]	[0.0222,0.0341]	[0.0213,0.0320]	[0.0223,0.0343]	[0.1231,0.1455]	[0.1249,0.1477]
3	[0.2023,0.2323]	[0.0639,0.0888]	[0.0647,0.0859]	[0.0632,0.0835]	[0.0615,0.0851]	[0.2199,0.2475]	[0.2171,0.2449]
4	[0.3637,0.3892]	[0.3082,0.3428]	[0.3210,0.3553]	[0.3009,0.3338]	[0.3806,0.4257]	[0.3719,0.3981]	[0.3752,0.4001]
5	[0.0456,0.0584]	[0.0042,0.0078]	[0.0043,0.0077]	[0.0041,0.0079]	[0.0045,0.0085]	[0.0528,0.0642]	[0.0042,0.0078]

Table A.2: 95% confidence intervals for experimental data with $r_{inner} = 50$ and s_γ=0.1 .

data set	PPLS-DA using						PLS-DA
	γ_{max}	γ_{opt}^{A}	γ_{opt}^{A1}	γ_{opt}^{A2}	γ_{opt}^{B}	$\gamma = 0.5$	
Leukemia	[0.0469,0.0656]	[0.0304,0.0437]	[0.0288,0.0395]	[0.0303,0.0438]	[0.0320,0.0448]	[0.0227,0.0323]	[0.0269,0.0373]
Lymphoma	[0.0369,0.0509]	[0.0233,0.0349]	[0.0262,0.0391]	[0.0278,0.0400]	[0.00273,0.0396]	[0.0252,0.0400]	[0.0229,0.0344]
Breast Cancer	[0.3282,0.3631]	[0.3652,0.3983]	[0.3531,0.3878]	[0.3543,0.3892]	[0.3480,0.3815]	[0.3641,0.3958]	[0.3573,0.3905]
Prostate 1	[0.0822,0.0950]	[0.0759,0.0919]	[0.0910,0.1096]	[0.0840,0.1010]	[0.0971,0.1157]	[0.0734,0.0983]	[0.0734,0.0872]
Prostate 2	[0.0656,0.0801]	[0.0705,0.0872]	[0.0689,0.0871]	[0.0701,0.0847]	[0.0718,0.0916]	[0.2039,0.2287]	[0.2000,0.2268]

B Appendix

Comments on the circle pattern

In this section, $abs(a)$ is define as the absolute value of the real number a. The distribution function of $X_k = r \cdot cos(\varphi) + \varepsilon$ for $r > 0$ is calculated. Let φ be randomly chosen according to the uniform distribution from the interval $[-\pi, \pi]$. Therewith the density of φ is $f_\varphi(\phi) = 1/2\pi$ if $-\pi \leq \phi \leq \pi$ and zero otherwise.
First the gene k is considered. Because X_k is the sum of independent random variables, first the density of $r \cdot cos(\varphi) := Y$ is calculated.

$$\begin{aligned} P(Y \leq y_0) &= P(r \cdot cos(\varphi) \leq y_0) \\ &= P(cos(\varphi) \leq \frac{y_0}{r}). \end{aligned}$$

Three cases exist: 1) $y_0 \geq r \Rightarrow P(Y \leq y_0) = 1$. 2)$y_0 < -r \Rightarrow P(Y \leq y_0) = 0$ and 3) else. For case 3) it can be calculated with $r > 0$ and with respect to properties

of the cosine function:

$$\begin{aligned} P(Y \leq y_0) &= P(cos(\varphi) \leq \frac{y_0}{r}) \\ &= P(abs(\varphi) \geq arccos\frac{y_0}{r}) \\ &= 1 - P(abs(\varphi) \leq arccos\frac{y_0}{r}) \\ &= 1 - \frac{2arccos\frac{y_0}{r}}{2\pi} \\ &= 1 - \frac{arccos\frac{y_0}{r}}{\pi}. \end{aligned}$$

Thus as density in case 3) we get

$$\frac{\partial}{\partial y_0}(P(Y \leq y_0)) = \frac{1}{\pi}\frac{1}{\sqrt{r^2 - y_0^2}}.$$

For case 1) and 2) the density is zero. The expectation is zero because the function is symmetric. As variance it is received

$$\begin{aligned} V(Y) &= \frac{1}{\pi}\int_{-r}^{r} y_0^2 \frac{1}{\sqrt{r^2 - y_0^2}} dy_0 \\ &= \frac{1}{\pi}(-\int_{-r}^{r} y_0 \frac{y_0}{\sqrt{r^2 - y_0^2}} dy_0) \\ &= \frac{1}{\pi}[-\underbrace{(y_0\sqrt{r^2 - y_0^2})}_{=0} + \int_{-r}^{r} \sqrt{r^2 - y_0^2} dy_0)] \\ &= \frac{1}{\pi}\int_{-r}^{r} \sqrt{r^2 - y_0^2} dy_0) \\ &= \frac{r^2}{2}. \end{aligned}$$

Therefore the variance of $X_k = r \cdot cos(\varphi) + \varepsilon$ is $V(X_k) = \frac{r^2}{2} + 0.004$. Hence, for group A the variance is $0.1^2/2 + 0.004 = 0.0054$ and for group B $0.25^2/2 + 0.004 = 0.0316$. Analogously the variance and the mean value for gene l can be calculated.

C List of Abbreviations and Notations

$\mathfrak{M}$	set, containing the features building a biomarker
|I|	cardinality of a set I
$\mathbf{X}$	$n \times g$ data matrix with n objects and g features
$\mathbf{x}_k$	kth column vector of $\mathbf{X}$
$\mathcal{G}$	set containing all group labels
$\mathbf{y}$	response vector, containing the group memberships
n_ν	sample size of group ν
π_ν	prior probability of group ν
$\mathbf{Y}$	dummy matrix of group memberships (according to $\mathbf{y}$)
$\{\mathbf{X}_{train}, \mathbf{y}_{train}\}$	training set
$\{\mathbf{X}_{test}, \mathbf{y}_{test}\}$	test set
n_{train}	number of objects of the training set
n_{test}	number of objects of the test set
$\phi_{\mathbf{X}}$	ratio of n_{train} on n
$\phi_{\mathbf{X}_{train}}$	ratio of the number of objects for the inner training set on n_{train}
r	number of cross-validation steps of the whole data set
r_{inner}	number of inner cross-validation steps of the training set
sd	standard deviation
PE(s)	prediction error(s)
LDA	linear discriminant analysis

t-LDA	LDA with ten filtered features according to the t-test as predictors
$p_\nu(\mathbf{x})$	probability for an object (determined by $\mathbf{x}$) belonging to group ν
$\mu_{(\nu)}$	mean value of group ν
Σ_ν	variance of group ν
δ_ν	linear discriminant function for group ν
RF	random forest
OOB	out-of-bag
$mtry$	number of features to choose at each node to built a decision tree in RF
n_{tree}	number of tress in a RF
SVM	Support vector machine
ξ_i	ith slack variable
$\mathcal{C}$	cost constant
SVML	support vector machines with linear kernel
τ	tuning constant of the radial basis function
SVMR	support vector machines with radial kernel
PCA	principal component analysis
PLS	partial least squares
$\mathbf{X}_{\{i\}}$	ith residuum of $\mathbf{X}$
n_c	number of components
PLS1	PLS with a vector as response
PLS2	PLS with a matrix as response
PLS-DA	partial least squares discriminant analysis
$\mathbf{B}$	between-group covariance matrix
$\mathbf{T}$	total sum of squares and cross product matrix
$\mathbf{w}$	loading weight vector
$\mathbf{t}$	score vector
$\mathbf{q}$	loading vector
FCDA	Fishers canonical discriminant analysis
Π	diagonal matrix of prior probabilities of the group

$diag(\{a_1, \ldots, a_l\})$	diagonal matrix with $a_1, \ldots, a_l$ as diagonal elements
PPLS-DA	power partial least squares discriminant analysis
$corr$	correlation
γ	power parameter
$n_{c(opt)}$	optimal number of components
var	variance
$\mu_{(\nu)k}$	mean of group ν for feature k
CCA	canonical correlation analysis
MDA	mean decrease accuracy
TB	tuberculosis
TST	tuberculin skin test
TST^+	TST positive
TST^-	TST negative
κ_k	condition index
cov	covariance
u_{ik}	gene expression of gene k for sample i on the original-scale
σ_b	biological variance
σ_t	technical variance
n_{total}	total number of available single samples
X_{ik}	gene expression of gene k for sample i on the log-scale
δ	mean class difference
$X_{ik(\nu)}$	gene expression of gene k for sample i on the log-scale for group ν
ISF	informative simulated feature(s)
sin	sinus
cos	cosine
DEGs	differentially expressed genes
N_{DEGs}	number of differentially expressed genes
n_S	number of samples used for the single sample arrays
n_{S_P}	number of samples used for the pools

a_S	number of arrays for the single sample design
a_P	number of arrays for the pools
a_{total}	total number of arrays which can be financed
$u_{ik(b)}$	gene expression of gene k for sample i on the original-scale without technical variation
u_{p_jk}	gene expression of gene k for pool p_j on the original-scale
μ_k	mean gene expression level of gene k
ω_i	proportion of the $i-$th sample in a pool
$\mathrm{D}_{\mathrm{sim}}$	informative simulated features
$\mathrm{D}^{\mathrm{M}}_{m_p}$	important features for classification using method M in a design with pool size m_p
$\mathrm{I}^{\mathrm{M}}_{1}$	$= \mathrm{D}^{\mathrm{M}}_{1} \cap \mathrm{D}^{\mathrm{M}}_{\mathrm{sim}}$
$\mathrm{I}^{\mathrm{M}}_{1:m_p}$	$= \mathrm{I}^{\mathrm{M}}_{1} \cap \mathrm{D}^{\mathrm{M}}_{m_p}$ for method M important informative simulated features which coincide in the single sample design and in a design with pool size m_p
γ^A_{opt}	optimal power parameter of version A, respectively for version A1, A2,B
$n^A_{c_{opt}}$	optimal number of components for version A, respectively for version A1, A2,B
s_γ	step size for equidistant fixed values in [0,1]
$abs(a)$	absolute value of the real number a

Bibliography

Affymetrix (2004). Sample Pooling for Microarray Analysis: A Statistical Assessment of Risks and Biases. Technical Note, Part No. 701494 Rev. 2.

Aizerman, M. A., Braverman, E. A., and Rozonoer, L. (1964). Theoretical foundations of the potential function method in pattern recognition learning. In *Automation and Remote Control*, number 25, pages 821–837.

Allison, D. B., Cui, X., Page, G. P., and Sabripour, M. (2006). Microarray data analysis: From disarray to consolidation and consensus. *Nature Reviews Genetics*, **7**(1), 55–65.

Barker, M. (2010). *Partial Least Squares for Discrimination: Statistical Theory and Implementation*. LAP LAMBERT Academic Publishing.

Barker, M. and Rayens, W. (2003). Partial Least Squares for Discrimination. *Journal of Chemometrics*, **17**(3), 166–173.

Baumgartner, C., Osl, M., Netzer, M., and Baumgartner, D. (2011). Bioinformatic-driven search for metabolic biomarkers in disease. *Journal of Clinical Bioinformatics*, **1**(1).

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons.

Bender, R. and Lange, S. (2007). Was ist der p-wert? *Deutsche Medizinische Wochenschrift*, **132**, 15–23.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, **57**(1), 289–300.

Biomarkers Definition Workgroup (2001). Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics*, **69**(3), 89–95.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer Science+ Business Media, LLC.

Bland, J. M. and Altman, D. G. (1995). Multiple significance tests: The Bonferroni method. *BMJ*, **310**(6973), 170.

Boot, J. D., Chandoesing, P., de Kam, M. L., Mascelli, M. A., Das, A. M., van Wijk, R. G., de Groot, H., Verhoosel, R., Hiemstra, P. S., and Diamant, Z. (2008). Applicability and reproducibility of biomarkers for the evaluation of anti-inflammatory therapy in allergic rhinitis. *Journal Investigational Allergology Clinical Immunology*, **18**(6), 433–442.

Borga, M. (2001). Canonical correlation: a tutorial. *Available from http://people.imt.liu.se/magnus/cca.*

Borgia, J. A., Basu, S., Faber, L. P., Kim, A. W., Coon, J. S., Kaiser-Walters, K. A., Fhied, C., Thomas, S., Rouhi, O., Warren, W. H., Bonomi, P., and Liptay, M. J. (2009). Establishment of a multi-analyte serum biomarker panel to identify lymph node metastases in non-small cell lung cancer. *Journal of Thoracic Oncology*, **4**(3), 338–347.

Boulesteix, A.-L. (2004). PLS dimension reduction for classification with microarray data. *Statistical Applications in Genetics and Molecular Biology*, **3**(1).

Bracht, K. (2009). Biomarker: Indikatoren für Diagnose und Thera-

pie. *Pharmazeutische Zeitung, Available from http://www.pharmazeutische-zeitung.de/index.php?id=29346*.

Breiman, L. (1996). Out-of-bag estimation. *Technical Report, Available from ftp://ftp.stat.berkeley.edu/pub/users/breiman/*.

Breiman, L. (2001). Random Forests. *Machine Learning*, **45**, 5–32.

Breiman, L. (2002). Manual on setting up, using, and understanding random forests v3.1. *Available from http://oz.berkeley.edu/users/breiman/ Using_random_forests_V3.1.pdf*.

Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, **2(2)**, 121–167.

Dabney, A., Storey, J. D., and with assistance from Gregory R. Warnes (2009). Q-value estimation for false discovery rate control. R package version 1.18.0.

Darzi, M.; Liaei, A. H. M. and Asghari, H. (2011). Feature Selection for Breast Cancer Diagnosis: A Case-Based Wrapper Approach. *World Academy of Science, Engineering and Technology*, **77**, 1143–1145.

Das, S. (2001). Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection. *Proceedings 18th ICML*, pages 74–81.

Dettling, M. (2004). BagBoosting for Tumor Classification with Gene Expression data. *Bioinformatics*, **20**(18), 3583–3593.

Dettling, M. and Buehlmann, P. (2003). Boosting for tumor classification with gene expression data. *Bioinformatics*, **19**(9), 1061–1069.

Deutsche Diabetes Gesellschaft (2010). Stellungnahme der Deutschen Diabetes Gesellschaft, diabetesDE und des Kompetenznetzes Diabetes mellitus zur Verwendung des HbA1c-Wertes als Biomarker zur Diabetesdiagnose. *Available from http://www.deutsche-diabetes-gesellschaft.de/redaktion/news/Stellungnahme_HbA1c_final.pdf*.

Díaz-Uriarte, R. and de Andrés, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**(3), 3.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., and Weingessel, A. (2009). *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. R package version 1.5-20.

Dudoit, S. and Fridlyand, J. (2003). *Statistical analysis of gene expression microarray data*, chapter 3, pages 93–158. Chapman and Hall/CRC.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, **97**, 77–87.

Ein-Dor, L., Kela, I., Getz, G., Givol, D., and Domany, E. (2004). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, **21**(2), 171–178.

Etzioni, R., Kooperberg, C., Pepe, M., Smith, R., and Gann, P. H. (2003). Combining biomarkers to detect disease with application to prostate cancer. *Biostatistics*, **4**(4), 523–538.

Fan, Y., Murphy, T. B., Byrne, J. C., Brennan, L., Fitzpatrick, J. M., and Watson, R. W. G. (2011). Applying random forests to identify biomarker panels in serum 2d-dige data for the detection and staging of prostate cancer. *Journal of Proteome Research*, **10**(3), 1361–1373.

Feng, Z., Prentice, R., and Srivastava, S. (2004). Research issues and strategies for genomic and proteomic biomarker discovery and validation: a statistical perspective. *Pharmacogenomics*, **5**(6), 709–719.

Geller, S. C., Gregg, J. P., Hagerman, P., and Rocke, D. M. (2003). Transformation and normalization of oligonucleotide microarray data. *Bioinformatics*, **19**(14), 1817–1823.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**(5439), 531–537.

Gudenus, R. and Granzer, U. (2010). Biomarker und Surrogatendpunkte. *transkript*, **7**, 31–32.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, **3**, 1157–1182.

Han, W. K., Wagener, G., Zhu, Y., Wang, S., and Lee, H. T. (2009). Urinary biomarkers in the early detection of acute kidney injury after cardiac surgery. *Clinical Journal of the American Society of Nephrology*, **4**(5), 873–882.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer.

Haury, A.-C., Gestraud, P., and Vert, J.-P. (2011). The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS ONE*, **6**(12).

Haynes, J.-D. and Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, **7**(7), 523–534.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O. P., Wilfond, B., Borg, A., Trent, J., Raffeld, M., Yakhini, Z., Ben-Dor, A., Dougherty, E., Kononen, J., Bubendorf, L., Fehrle, W., Pittaluga, S., Gruvberger, S., Loman, N., Johannsson, O., Olsson, H., and Sauter, G. (2001). Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine*, **344**(8), 539–548.

Helland, I. S. and Almøy, T. (1994). Comparison of prediction methods when only

a few components are relevant. *Journal of the American Statistical Association*, **89**, 583 – 591.

Huberty, C. J. (1994). *Applied dicriminant analysis*. John Wiley & Sons, Inc.

Indahl, U. G. (2005). A twist to partial least squares regression. *Journal of Chemometrics*, **19**, 32–44.

Indahl, U. G., Martens, H., and Næs, T. (2007). From dummy regression to prior probabilities in PLS-DA. *Journal of Chemometrics*, **21**, 529–536.

Indahl, U. G., Liland, K. H., and Næs, T. (2009). Canonical partial least squares - a unified PLS approach to classification and regression problems. *Journal of Chemometrics*, **23**, 495–504.

Inza, I., Larrañaga, P., Blanco, R., and Cerrolaza, A. J. (2004). Filter versus wrapper gene selection approaches in DNA microarray domains. *Artificial Intelligence in Medicine*, **31**(2), 91–103.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003). Summaries of affymetrix genechip probe level data. *Nucleic Acids Research*, **31**(4).

Jacobsen, M., Repsilber, D., Gutschmidt, A., Neher, A., Feldmann, K., Mollenkopf, H. J., Ziegler, A., and Kaufmann, S. H. E. (2007). Candidate biomarkers for discrimination between infection and disease caused by mycobacterium tuberculosis. *The Journal of Molecular Medicine*, **85**(6), 613–621.

Jacobsen, M., Mattow, J., Repsilber, D., and Kaufmann, S. H. E. (2008). Novel strategies to identify biomarkers in tuberculosis. *The Journal of Biological Chemistry*, **389**(5), 487–495.

Jarasch, E.-D. (2011). Entwicklung neuer molekularer Biomarker. *BIOPRO Baden-Württemberg, Available from http://www.bio-pro.de/magazin/thema/07390/index.html?lang=de*.

Jensen, J. L. W. V. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, **30**, 175–193.

Kell, D. B. (2007). Metabolomic biomarkers: search, discovery and validation. *Expert Review of Molecular Diagnostics*, **7**(4), 329–333.

Kendziorski, C., Irizarry, R. A., Chen, K. S., Haag, J. D., and Gould, M. N. (2005). On the utility of pooling biological samples in microarray experiments. *Proceedings of the National Academy of Sciences USA*, **102**(12), 4252–7.

Kendziorski, C. M., Zhang, Y., Lan, H., and Attie, A. D. (2003). The efficiency of pooling mRNA in microarray experiments. *Biostatistics*, **4**(3), 465–477.

Kerr, M. K. (2003). Design considerations for efficient and effective microarray studies. *Biometrics*, **59**(4), 822–8.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, **97**, 273–324.

Lai, C., Reinders, M. J. T., van't Veer, L. J., and Wessels, L. F. A. (2006). A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. *BMC Bioinformatics*, **7**.

Langley, P. (1994). Selection of relevant features in machine learning. *Proceedings of AAAI Fall Symposium on Relevance*, **97**, 1–5.

Lapointe, J., Li, C., Higgins, J. P., van de Rijn, M., Bair, E., Montgomery, K., Ferrari, M., Egevad, L., Rayford, W., Bergerheim, U., Ekman, P., DeMarzo, A. M., Tibshirani, R., Botstein, D., Brown, P. O., Brooks, J. D., and Pollack, J. R. (2004). Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proceedings of the National Academy of Sciences USA*, **101**(3), 811–816.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, **2**(3), 18–22.

Liggett, W. S., Barker, P. E., Semmes, O. J., and Cazares, L. H. (2004). Measurement reproducibility in the early stages of biomarker development. *Disease Markers*, **20**(6), 295–307.

Liland, K. H. and Indahl, U. (2009). Powered partial least squares discriminant analysis. *Chemometrics*, **23**, 7–18.

Liu, H., Li, J., and Wong, L. (2002). A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Inform*, **13**, 51–60.

Maertzdorf, J., Repsilber, D., Parida, S. K., Stanley, K., Roberts, T., Black, G., Walzl, G., and Kaufmann, S. H. E. (2011). Human gene expression profiles of susceptibility and resistance in tuberculosis. *Genes Immun*, **12**(1), 15–22.

Malone, J. H. and Oliver, B. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biology*, **9**, 34.

Mary-Huard, T., Daudin, J. J., Baccini, M., Biggeri, A., and Bar-Hen, A. (2007). Biases induced by pooling samples in microarray experiments. *Bioinformatics*, **23**(13), i313–8.

Moler, E. J., Chow, M. L., and Mian, I. S. (2000). Analysis of molecular profile data using generative and discriminative methods. *Physiol Genomics*, **4**(2), 109–126.

Nocairi, H., Qannari, E. M., Vigneau, E., and Bertrand, D. (2005). Discrimination on latent components with respect to patterns. application to multicollinear data. *Computational Statistics & Data Analysis*, **48 (1)**, 139–147.

Peng, X., Wood, C. L., Blalock, E. M., Chen, K. C., Landfield, P. W., and Stromberg, A. J. (2003). Statistical implications of pooling rna samples for microarray experiments. *BMC Bioinformatics*, **4**, 26.

Pfeiffer, R. M. and Bur, E. (2008). A model free approach to combining biomarkers. *Biometrical Journal*, **50**(4), 558–570.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature Genetics*, **32**, 496–501.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ripley, B. D. (1997). *Pattern recognition and neural network*. Cambridge University Press.

Rudolf, H. (2011). Personal communication.

Russel, S. and Norvig, P., editors (2009). *Artificial Intellligence: A Modern Approach*. Prentice Hall.

Sadiq, S. T. and Agranoff, D. (2008). Pooling serum samples may lead to loss of potential biomarkers in SELDI-ToF MS proteomic profiling. *Proteome Science*, **6**, 16.

Saebø, S., Almøy, T., Aarøe, J., and Aastveit, A. H. (2008). ST-PLS: a multi-directional nearest shrunken centroid type classifier via PLS. *Journal of Chemometrics*, **20**, 54–62.

Searfoss, G. H., Jordan, W. H., Calligaro, D. O., Galbreath, E. J., Schirtzinger, L. M., Berridge, B. R., Gao, H., Higgins, M. A., May, P. C., and Ryan, T. P. (2003). Adipsin, a biomarker of gastrointestinal toxicity mediated by a functional gamma-secretase inhibitor. *Journal of Biological Chemistry*, **278**(46), 46107–46116.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C. T., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S., Ray, T. S., Koval, M. A., Last, K. W., Norton, A., Lister, T. A., Mesirov, J., Neuberg, D. S., Lander, E. S., Aster, J. C., and Golub, T. R. (2002). Diffuse large b-cell lymphoma outcome

prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, **8**(1), 68–74.

Simon, R., Radmacher, M. D., and Dobbin, K. (2002). Design of studies using DNA microarrays. *Genet Epidemiol*, **23**(1), 21–36.

Simon, R., Radmacher, M. D., Dobbin, K., and McShane, L. M. (2003). Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute*, **95**(1), 14–8.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R., and Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**(2), 203–209.

Storey, J. D. (2003). The positive false discovery rate: A bayesian interpretation and the q-value. *Annals of Statistics*, **31 (6)**, 2013–2035.

Student (1908). The probable error of a mean. *Biometrika*, **6**, 1–25.

Sumpf, D. and Moll, E. (2004). *Einführung in die Biometrie 2 - Schätzen eines Parameters und Vergleich von bis zu zwei Parametern*. BMVEL.

Telaar, A., Nürnberg, G., and Repsilber, D. (2010). Finding biomarker signatures in pooled sample designs: A simulation framework for methodological comparisons. *Advances in Bioinformatics*, **2010**, 1–11.

Telaar, A., Repsilber, D., and Nürnberg, G. (2012a). Biomarker discovery: Classification using pooled samples - a simulation study. *Computational Statistics*, pages 1–40.

Telaar, A., Liland, K. H., Repsilber, D., and Nürnberg, G. (2012b). Extensions of PPLS-DA for classification and comparison to ordinary PLS-DA. *PLoS ONE (in review)*.

van't Veer, L. e. a. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**(31), 530–536.

Vapnik and Chervonenkis (1974). Theory of pattern recognition. *(dt. Übersetzung: Wapnik und Tschervonenkis, Theorie der Mustererkennung, 1979).*

Vapnik, V. (1979). Estimation of dependences based on empirical data [in russian]. *Nauka (English translation Springer Verlag, 1982).*

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer, New York,.

Weiner, J., Padrida, S. K., Maertzdorf, J., Black, G. F., Repsilber, D., Telaar, A., Robert, P. M., Arndt-Sullivan, C., Ganoza, C. A., Faé, K. C., Walzl, G., and Kaufmann, S. H. (2011). Biomarkers of inflammation, immunosuppression and stress with active disease are revealed by metabolomic profiling of tuberculosis patients. *PLoS ONE (accepted).*

Wold, S. (1995). Chemometrics; what do we mean with it, and what do we want from it? *Chemometrics and Intelligent Laboratory Systems*, **30**, 109–115.

World Health Organization (2011). Global tuberculosis control report 2011. *Available from http://www.who.int*.

Yang, Y. H. and Speed, T. (2003). *Statistical analysis of gene expression microarray data*, chapter 2, pages 35–95. Chapman and Hall/CRC.

Zhang, W., Carriquiry, A., Nettleton, D., and Dekkers, J. C. (2007). Pooling mRNA in microarray experiments and its effect on power. *Bioinformatics*, **23**(10), 1217–24.