Direct and Indirect Classifiers

Dissertation

zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
der Universität Dortmund

Dem Fachbereich Statistik
der Universität Dortmund
vorgelegt von

Andrea Peters
aus Viersen

Dortmund 2003

Prüfungskommission:	Prof. Dr. W. Urfer (Vorsitzender)
	Prof. Dr. C. Weihs (Gutachter)
	Prof. Dr. W. Krämer (Gutachter)
	PD Dr. B. Lausen (Gutachter)
	Dr. D. Enache (Beisitzer)
Tag der mündlichen Prüfung:	16. Dezember 2003

Danksagung

Allen, die zum Entstehen meiner Arbeit beigetragen haben, möchte ich an dieser Stelle meinen Dank aussprechen.

An erster Stelle danke ich PD Dr. Berthold Lausen für die Betreuung und für die Initialisierung des Themas meiner Arbeit. Neben seiner Unterstützung gilt mein Dank auch Dr. Torsten Hothorn für das aufmerksame und überaus kritische Lesen des Manuskriptes, ebenso wie für die zahlreichen Diskussionen über den Sinn und Unsinn verschiedenster Ansätze des maschinellen Lernens.
M.Sc. Janice Hegewald hat die Arbeit sicher nicht nur sprachlich korrigiert, sondern durch ihre Rückfragen auch AnstöSSe zu weiteren Überlegungen gegeben.

Für den zügigen Ablauf der Prüfung möchte ich mich bei dem Promotionsausschuss bedanken, der sehr unkompliziert einen Prüfungstermin noch im Jahre 2003 ermöglichte. Der Prüfungskommission, bestehend aus Prof. Dr. Wolfgang Urfer, Prof. Dr. Claus Weihs, Prof. Dr. Walter Krämer, PD Dr. Berthold Lausen und Dr. Daniel Enache, danke ich für die angenehme und produktive Prüfungsatmosphäre.

Wichtig für das Gelingen der Arbeit war sicher auch die gute und abgesicherte Atmosphäre an meiner Arbeitsstelle, dem Institut für Medizininformatik, Biometrie und Epidemiologie der Universität Erlangen-Nürnberg. Hierfür danke ich unserem Institutsvorstand Prof. Dr. Olaf Gefeller, der mir die erforderlichen Freiräume und die eigenverantwortliche Arbeit an meinen Projekten ermöglichte. Ebenso danke ich der Deutschen Forschungsgemeinschaft für die finanzielle Unterstützung der Arbeit im Rahmen des Projektes „Biometrische Modellbildung; Dokumentation, Planung und Auswertung" des Sonderforschungsbereiches 539.

Kollegen und Freunden möchte ich für die gegebenen emotionalen Stützen und Ablenkungen danken. Ich danke insbesondere Dipl. oec. troph. Jasmin Ecke und Dipl.-Kff. Andrea Lehnert-Batar für Diskussionen über die Feinheiten der

englischen Sprache und ebenfalls ihnen, als auch Dipl.-Inf. Werner Adler für ein zweites Korrekturlesen der „entscheidenden" Abschnitte meines Manuskriptes. Meiner Mitarbeiterin PD Dr. Annette Pfahlberg danke ich, dass sie immer geduldig meinen lästigen Fragen, unter anderem zum technischen Ablauf der Prüfung, Rede und Antwort stand. Ebenso danke ich auch allen anderen früheren und derzeitigen Kollegen, dass ich jederzeit auf ihre Unterstützung bauen konnte.
Nicht zuletzt danke ich meiner Familie für ihre Geduld und ihre Bodenständigkeit, die mich die Schwierigkeiten und zähen Phasen während der Entstehung der Dissertation nicht überbewerten lieSS. Meinem Freund Thomas Richter danke ich für sein Verständnis für meinen Zeitmangel und seine Zuhörbereitschaft in manchmal etwas hektischen Zeiten.

Erlangen, im Januar 2004

Contents

Zusammenfassung

In der medizinischen Forschung finden automatische Klassifikationsregeln ihre Anwendung häufig als Hilfsmittel zur untersucherunabhängigen Entscheidungsfindung. So kann in Screeningprogrammen mit einer solchen Klassifikationsvorschrift, welche die Probanden automatisch in „krank" und „gesund" einteilt, teures und erfahrenes Fachpersonal durch weniger teure Hilfskräfte entlastet werden. In der vorliegenden Arbeit wird diskutiert, wie sowohl das a priori Wissen eines Arztes über eine Erkrankung, als auch die medizinischen Messungen, welche nur für den Lerndatensatz, nicht jedoch für spätere Testdatensätze erhoben werden, zur Verbesserung einer solchen automatischen Klassifikationsvorschrift genutzt werden können.
Ausgangspunkt aller Überlegungen ist hierbei eine irreversible Erkrankung der retinalen Nervenfaserschicht, genannt der grüne Star (Glaukom). Diese Erkrankung ist derzeit Gegenstand aktueller medizinischer Forschung. Daher sind durch verschiedenste Forschungsansätze im Erlanger Glaukomregister Lerndatensätze entstanden, welche Ergebnisse vieler medizinischer Untersuchungen beinhalten, obwohl die eigentliche medizinische Diagnose dieser Erkrankung meist auf nur zwei konventionellen Untersuchungsverfahren basiert.

In der vorliegenden Arbeit wird zunächst der Ansatz der indirekten Klassifikation beschrieben. Der konventionelle indirekte Klassifikator unterscheidet zwischen so genannten „erklärenden" Variablen, welche sowohl für derzeitige, als auch für zukünftige Beobachtungen erhoben werden, und „intermediären"

Variablen, auf denen die Diagnose basiert. Dieser konventionelle Ansatz der indirekten Klassifikation kann in Situationen, in denen ein detailliertes a priori Wissen einschließlich einer medizinischen Diagnosevorschrift, basierend auf intermediären Variablen, bekannt ist, angewendet werden.
In einem weiteren Schritt wird die indirekte Klassifikation auf Situationen, in denen weniger Wissen über einen gegebenen Lerndatensatz vorhanden ist, ausgeweitet. Allgemeiner wird ein indirekter Klassifikator als ein Klassifikator definiert, der alle im Lerndatensatz vorhandenen Variablen in einer gewissen Form für die Klassifikation nutzt. Der algorithmische Vorschlag „indirect subagging" kombiniert eine beliebige Anzahl von Vorhersagemodellen für intermediäre Variablen, welche für zukünftige Probanden nicht erfaßt werden. Im Gegensatz zur konventionellen indirekten Klassifikation, kann „indirect subagging" auch in Situationen angewendet werden in denen nur wenig a priori Wissen vorhanden ist. Ein direkter Klassifikator nutzt ausschlieSSlich die erklärenden Variablen und die Klassenvariable zur Erstellung einer automatischen Entscheidungsregel.

Eine bekannte medizinische Diagnosevorschrift, wie sie im konventionellen Ansatz der indirekten Klassifikation einbezogen wird, ermöglicht die Unterscheidung zwischen einer diagnostizierten Klassenzugehörigkeit bzgl. der Diagnosevorschrift und einer wahren Klassenzugehörigkeit. Es wird im folgenden zwischen diesen beiden möglichen Erkrankungszuständen unterschieden.
Asymptotische Eigenschaften der indirekten Klassifikation werden untersucht, und es wird gezeigt, daß der konventionelle indirekte Ansatz unter bestimmten Modellannahmen Bayes konsistent bzgl. des diagnostizierten Erkrankungszustandes ist, während „indirect subagging" auch unter allgemeineren Annahmen Bayes konsistent ist.
Ein abstraktes Simulationsmodell führt zu der Erkenntnis, daSS die korrekte Formulierung der Diagnosevorschrift im konventionellen indirekten Ansatz ausschlaggebend für dessen Misklassifikationsrate ist. Insgesamt erreichten die in-

direkten Klassifikatoren niedrigere Fehlerraten, als die direkten.
Im weiteren wird ein komplexes Simulationsmodell beschrieben, welches eine ähnliche Datenstruktur, wie sie bei der Glaukomklassifikation gegeben ist, generiert. Es wird untersucht, wie indirekte und direkte Klassifikatoren auf verschiedene Varianzen der erklärenden bzw. der intermediären Variablen reagieren.
Die Anwendung der diskutierten Verfahren auf eine Fall-Kontroll-Studie von Glaukompatienten und gesunden Probanden macht den Nutzen der indirekten Klassifikation bei realen Fragestellungen deutlich. Anwendungen auf zwei weitere Datensätze weisen darauf hin, daß der Ansatz „indirect subagging" vergleichbar gute oder bessere Misklassifikationsraten wie die entsprechenden direkten Klassifikatoren zu erreichen scheint. Die Güte der indirekten Klassifikation scheint auch von dem Informationsgehalt der intermediären Variablen abzuhängen.
Abschließend wird die Durchführung der indirekten Klassifikation mit Hilfe des Zusatzpaketes **ipred** in der Programmierumgebung R demonstriert.

Abstract

Automated classification rules are often required tools in medical research. In screening programs, an automated classification rule is desired which can accurately identify the subjects as being "healthy" or "affected", allowing expensive and experienced specialised staff to be replaced by cheaper assistants. In the current thesis we examine how to make use of medical examinations and a priori knowledge, which are given for learning samples but not for later test samples, in order to improve an automated classification rule.
The starting point of all considerations was glaucoma, an affection of the retinal nerve fibre layer. Glaucoma is an irreversible disease and focus of recent research. Several medical approaches lead to learning samples in the Erlanger Eye Registry. The registry includes a magnitude of medical examinations, although the diagnosis is usually based on two conventional examination tools.

At the beginning, we describe the approach of indirect classification. The conventional indirect classifier distinguishes between explanatory variables, i.e. variables which are available for recent and future observations, and intermediate variables, i.e. variables the diagnosis is based on. This approach of indirect classification can be applied in situations where detailed a priori knowledge, including a diagnostic rule based on the intermediate variables, is given.
In the following, we extend indirect classification to situations where such a diagnostic tool is not known. We define an indirect classifier more generally, as a classification rule, which makes use of all variables given in the learning sample.

We make the algorithmic proposal "indirect subagging". Indirect subagging is a generalised indirect classification approach which combines an arbitrary number of prediction models for intermediate variables, which are not collected for future observations. In contrast to the conventional indirect approach, we can apply indirect subagging in situations where only little a priori knowledge is given. In contrast to the framework of indirect classification we define a direct classifier as a classifier which only uses the set of explanatory variables.

A given diagnostic function, incorporated into the conventional indirect classification approach, enables the distinction between an observed class membership following the diagnostic function and a true class membership. We distinguish between these two possible states of the disease in following investigations. Furthermore, we examine asymptotic properties of indirect classification and show that the conventional indirect approach is Bayes consistent with respect to the observed class membership and under certain model assumptions, while indirect subagging is Bayes consistent under more general assumptions.
An artificial simulation model leads to the conclusion, that a correct specification of the fixed diagnostic function is crucial for the performance of the conventional indirect classifier. All in all, the indirect classifiers outperform direct ones within this simulation framework.
Moreover, we develop a complex simulation setup, which generates the data structure as given by the task of glaucoma classification. We investigate the performance of direct and indirect classifiers for different variances of explanatory and intermediate variables here.
Application to a case-control study of glaucoma and healthy subjects show the gain of indirect classification. The application to two additional datasets indicate that indirect subagging performs comparably or even better than the corresponding direct classifiers. The performance of the classifier seems to depend on the diagnostic value of the intermediate variables.

Finally, we demonstrate the application of indirect classification, using the add-on package **ipred** in the programming environment R.

Chapter 1

Introduction

Medical decision making is often a complex process based on several high dimensional measurements obtained by different examinations. The experience of the physician is precondition for a reliable classification of patients to diseased or healthy. Especially in screening programs, the observers are often unexperienced assistants, hence, the construction of automated classification rules is a desirable aim. Biostatistic and machine learning propose a magnitude of such automated classification techniques, usually constructed on study populations.

On the one hand, there are mechanisms which search for similar clusters within a study population where the class membership of the observations is not known, e.g. the state of disease for each patient is unknown. This is called "unsupervised learning". On the other hand, one can be interested in the development of a rule which assigns future patients. Such rules are constructed on a study population (learning sample) with known class membership of the observations. We call that "supervised learning". This manuscript focuses on the latter classification technique.

Supervised classification rules are usually assessed with respect to their error rates, i.e. by the proportion of misclassified patients. Error rates can either be estimated by applying a rule on an independent sample, called "test sample" or

by the calculation of an error rate estimator (Efron and Tibshirani, 1997; Schiavo and Hand, 2000).

However, the starting point of all following consideration was the task of medical decision making in glaucoma diagnosis. Glaucoma is a neuro-degenerative disease which affects the retinal nerve fibre layer. Nowadays it is the second most frequent cause for blindness worldwide, see Coleman (1999). This disease is irreversible, however, state-of-the-art therapy can slow down its progression. Therefore, early detection of glaucoma is helpful for the conservation of visual faculty. A case-control study of glaucomatous and healthy subjects was performed.

Classification of glaucoma is embedded with a quantity of typical difficulties arising from the development of supervised classification techniques in medical applications.
A common difficulty is the aim to classify an observation based on a small number of variables, although the learning sample contains more information, i.e. additional "intermediate" variables are available. The glaucoma dataset for example includes numerous medical examinations: Three dimensional measurements of the optic nerve head indicate the proportion of degenerated retinal nerve fibres. A two dimensional photo of the eye background is an established but less informative tool to measure this decrease and visual field tests assess the visual faculty of a patient. A reduction of medical examinations is required to spare patients' time and costs for the diagnostic procedure. Hence, a classification rule should be based only on three dimensional examinations, which have the ability to detect glaucoma early.
Different solutions are proposed in literature on how to make use of additional intermediate variables, i.e. variables available in learning samples but not in later test samples. Classification approaches dealing with latent variables establish a possibility to incorporate intermediate variables but can't handle additional a

priori knowledge. The inclusion of latent variables often leads to an improvement of error rates, Vermunt and Magidson (2003) provides an overview of recent developments. Tibshirani and Hinton (1998) proposes the use of additional intermediate variables as a structuring component. In "mixture coaching" the authors subdivide a model to predict the response variable based on the explanatory variables into a model for predicting the response in different partitions of the intermediate variables and a model for predicting the partition membership of the intermediate variable from the explanatory variable only. "Response coaching" predicts intermediate and response variables from explanatory variables only and integrates over the intermediate variables. Martus (2001) applied latent variable techniques for the evaluation of diagnostic measurements of glaucoma concerning paired organs.

Furthermore, in medical diagnosis a priori information about the relationship between the disease and the outcome of some of these examinations is given. Statistical classification methods that mimic this process of medical decision making should pay attention to the distinction between such a priori knowledge and the information about measurements required to predict the parameters defining the diagnosis. In glaucoma diagnosis, a priori knowledge about the definition of the disease is given, although there exists a magnitude of morphological variations of glaucoma, e.g. primary or secondary open angle glaucoma, glaucoma with or without an increased intra-ocular pressure etc.. All variations are uniformly defined by variables describing the visual field defect and the loss of retinal nerve fibres (Lee et al., 1998). A classification rule incorporating this diagnostic information can be based on a reduced set of examinations while making use of the full information of the data. The task was to use the a priori known definition of the disease and develop an early detecting automated classification based on modern clinical examination tools. A difficulty is the correct specification of the medical decision rule, because mis-specifying can lead to poor results of the au-

tomated procedure.

An approach suggested by Hand et al. (2001) allows the incorporation of intermediate variables and connected medical a priori knowledge into a statistical classification method. As introduced above, it assumes that the outcomes of the examinations are subdivided into three groups of variables: those to be used predicting the diagnosis, those to be used defining the diagnosis and the final diagnostic variable itself. The indirect classification process is executed in two separate steps. In the first step, prediction models for the defining variables embedding all other variables of potential influence are created based on a learning sample. In the second step, these defining variables are classified according to a deterministically known classifying function to yield the final medical diagnosis. The medical a priori knowledge is used twofold in this approach: (i) it is the criterion for the subdivision of variables into the different groups and (ii) it determines the fixed classifying function used in the second step of the procedure.

Moreover, in other applications a definition of a disease is often not known. A priori knowledge can be reduced to information whether medical tests are performed on future patients or not. Even the informative matter of intermediate and explanatory variables is often not known. Consider for example classification of diabetes and healthy subjects. To our knowledge there are no standard medical measurements used to diagnose the disease and it is not clarified, whether it is e.g. age dependent or not. That means, the a priori knowledge is reduced on informations about e.g. costs and availability of medical tests. An automated rule is required, which does not depend on the correct specification of a medical decision rule and nevertheless uses all available variables of the learning sample but predicts on a reduced set of variables.

However, a given definition of a disease, as in glaucoma diagnosis, leads to a possible distinction between a true and an observed class membership. The definition reflects the experienced based diagnosis of the physician, whereas the true

diagnostic state of a patient causes an underlying data structure. For example, patients have smaller visual field defects if the true diagnostic state is healthy rather than if they are diagnosed as healthy. This type of error, called differential misclassification, is often discussed in epidemiological context (Flegal et al., 1991; Rothman and Greenland, 1998; Grimes and Schulz, 2002) but usually neglected in classification tasks. The distinction between observed and true class membership results in two possible assessments of the classification rule: a misclassification error in terms of the observed and one in terms of the true class membership.

In this thesis, we give a generalised definition of the framework of indirect classification as suggested by Hand et al. (2001). We define indirect classification as classification that incorporates information connected with the intermediate variables. In contrast, we define direct classifiers as those using the information of variables available in learning and later test samples only. We combine indirect classifiers following the framework of Hand et al. (2001) with bootstrap aggregation which leads to an improvement of error rates in application (Breiman, 1996a; Peters et al., 2003).
However, the difficulty of how to deal with situations where information about the medical decision rule is missing or moreover, where it is even not known which examinations contain more or less diagnostic value, remain unsolved. The a priori knowledge is reduced to a criterion for the subdivision of variables here. We propose a procedure which is based on the indirect approach and which makes use of a rating of variables by classification trees (Breiman et al., 1984). Combining predictive models for the intermediate variables with subagging (Bühlmann and Yu, 2002) leads to the automated classification approach "indirect subagging". A difficulty associated with the prediction of intermediate variables is deciding which model is appropriate. Indirect subagging enables us to combine an arbitrary number of regression models with subagging and Hothorn and Lausen (2003c) demonstrate in a direct classification framework that additional

variables can improve the performance of bagging (or subagging) if they contain information about the underlying data structure and they do not affect the performance if they are uninformative. Resulting, the indirect subagging approach should perform comparably to direct classification in situations where the predicted intermediate variables embed no diagnostic value and should outperform it otherwise.

Error rates are often used to assess supervised classification techniques. Throughout this thesis an analysis of error rates of indirect classifiers is performed with respect to a distinction between an observed class membership, given by the fixed classifying function and a true class membership, causing an underlying data structure (Peters and Lausen, 2003). We define differential misclassification for a classification task as well as error rates with respect to the true and with respect to the observed class membership. Error rate calculations are performed for a simple discriminant model with normally distributed variables. We investigate asymptotic properties of indirect classification in situations with and without a fixed classifying function.

The performance of indirect classification techniques in finite sample situations are analysed within different simulation studies. On the one hand we consider different structures of the decision space, on the other hand we evaluate a classification rule for glaucoma diagnosis which is based on three dimensional measurements of the optic nerve head morphology. The indirect approach classifies patients using a definition of glaucoma based on measurements obtained by visual field tests and two dimensional fundus photos but requires three dimensional information about the optic nerve head only, to classify future patients. Such a rule reduces the number of necessary examinations, which in turn decreases the amount of time demanded from patients and reduces medical costs. A simulation study which mimics the data structure of the glaucoma dataset is our tool to assess the performance of direct and indirect classifiers for the task of glaucoma

diagnosis.
To illustrate the practical application of indirect classification and to demonstrate the gain achieved by its utilisation we apply indirect classifiers to different data sets, including situations with and without a known and fixed classifying function.

More specifically, this thesis is organised as follows. In chapter 2 we discuss medical decision making for classification of glaucoma and introduce a case-control study. We describe the considered discriminant model in section 3.1 and define and contrast indirect and direct classification in sections 3.3 and 3.2. In section 4 we analyse the performance of indirect classifiers and the direct Bayes classifier under certain model assumptions and consider asymptotic properties. Classification errors with respect to the true class membership and those with respect to the observed class membership are distinguished.
Chapter 5 includes different simulation approaches. We focus on the analysis of different dependencies within training and test samples in section 5.1. In section 5.2 we discuss a simulation model which mimics data structures of a case-control study of normal and glaucoma subjects.
Applications are performed in chapter 6. The proposed techniques are implemented in a computer package, which is described in chapter 7.

Chapter 2

Classification of Glaucoma

The starting point of all considerations was the task of glaucoma classification. Typical difficulties of medical decision making occur here. We have a magnitude of variables from different medical examinations but a diagnosis which is only based on a very limited number of variables. Moreover, the given a priori knowledge enables us to formulate a simplified diagnosis of the disease.
In the following we describe the process of medical decision making which leads to the assignment of a given set of variables into explanatory, intermediate and response variables and to the simplified diagnosis of the disease. We describe a case-control study of normal and glaucomatous subjects afterwards.

2.1 Medical Decision Making of Glaucoma

Glaucoma is a slow and irreversible neuro-degenerative disease which affects the retinal nerve fibre layer and often occurs in a population of elderly people, see Coleman (1999). The diagnosis of glaucoma is based on examinations of the visual field defect and the morphology of the optic nerve head (ONH). Since the onset of the disease is usually not detected and the state-of-the-art therapy of glaucoma is to slow down its progression, early detecting classification rules are

required.

There are several possibilities to examine the visual field defects and the optic nerve head (ONH). A common examination of the visual field is performed with the octopus, this tool gives impulses of light within the visual field of a patient. The patient signals whether he or she sees the flicker or not. Photographs of the ONH (papillometry) or a three dimensional topographical analysis of it by the Heidelberg Retina Tomograph (HRT) (see Swindale et al., 2000; Mardin et al., 1999) are often used to assess the ONH morphology. The HRT is a confocal scanning laser tomograph that produces a series of 32 images, each of 256×256 pixels, which are converted to a single topography image where each pixel represents a depth value (see Heidelberg Engineering, 1997).

As mentioned above, variables which describe the visual field defect and the

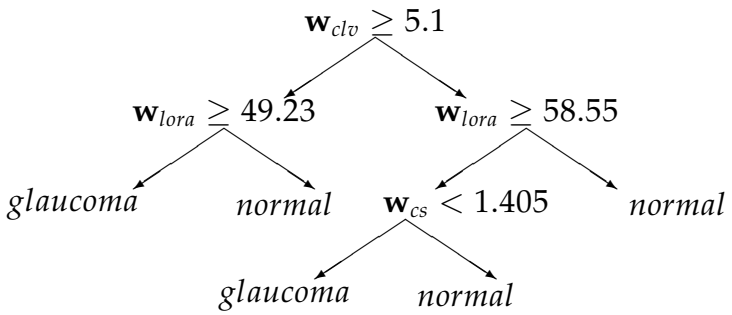

Figure 2.1: Diagnosis of glaucoma: Based on the intermediate variables $\mathbf{w}_{lora}$ (loss of rim area), $\mathbf{w}_{cs}$ (contrast sensitivity), $\mathbf{w}_{clv}$ (corrected loss variance) a patient is classified according to the graph.

proportion of retinal nerve fibres lost are used to define the disease, although a unique definition of glaucoma is controversial. For example Lee et al. (1998) observed that there are several different definitions of normal tension glaucoma, which is a special case of glaucoma. We introduce a simplified definition of glaucoma which is based on medical knowledge and depends on the following intermediate variables: Loss of rim area $\mathbf{w}_{lora}$, corrected loss variance $\mathbf{w}_{clv}$ and contrast

sensitivity $\mathbf{w}_{cs}$. Loss of rim area describes the proportion of the papilla which does not consist of the rim area, i.e. which does not consist of retinal nerve fibres, and is measured using optic nerve head photographs. Corrected loss variance describes the loss of variance of the visual field corrected by the short-time fluctuation, observed by the octopus. This examination also measures the contrast sensitivity of the eye. Based on these three intermediate variables a patient is classified as normal, if $g(\mathbf{w}) = 0$ and as glaucomatous if $g(\mathbf{w}) = 1$, where

$$\begin{aligned} g(\mathbf{w}) \quad := \quad & \chi_{\{\mathbf{w}_{clv} \geq 5.1\}}(\mathbf{w}_{clv})\chi_{\{\mathbf{w}_{lora} \geq 49.23\}}(\mathbf{w}_{lora}) + \\ & \chi_{\{\mathbf{w}_{clv} < 5.1\}}(\mathbf{w}_{clv})\chi_{\{\mathbf{w}_{lora} \geq 58.55\}}(\mathbf{w}_{lora})\chi_{\mathbf{w}_{cs} < 1.405}(\mathbf{w}_{cs}), \end{aligned} \tag{2.1}$$

with $\mathbf{w} = (\mathbf{w}_{lora}, \mathbf{w}_{cs}, \mathbf{w}_{clv})$ and $\chi(.)$ denotes the indicator function. The function $g(\mathbf{w})$ for glaucoma diagnosis is also displayed in figure 2.1.

2.2 Case-Control Study

Data from a cross-sectional study including 85 glaucomatous and 85 normal eyes from the Erlangen Glaucoma Registry are given (cf. Mardin et al., 1999, 2003). Only the measurements of the first examination of one eye of each patient are taken. The variables are obtained by HRT, papillometric and visual field examinations and include anamnestic information. Normal and glaucomatous subjects are matched by age and sex, to adjust for possible confounding.

We assume that for future examinations only HRT data will be available. Hence, only these sets of variables are used as explanatory variables, there are 62 HRT explanatory variables. The HRT variables include several measurements of the volume and areas of certain portions of the papilla. Some major HRT variables are displayed in table 2.1 and their detailed description is given in section 5.2.

The main characteristic of glaucoma is a reduced number of retinal nerve fibres. Consequently, the morphology of the papilla becomes less prominent, volumes

describing the upper part of the papilla decrease and those describing the proportion of volume not covered by retinal nerve fibres increase. Figure 2.2 shows a two dimensional profile of a model of a healthy (solid line) and a glaucomatous papilla (dashed line), respectively.

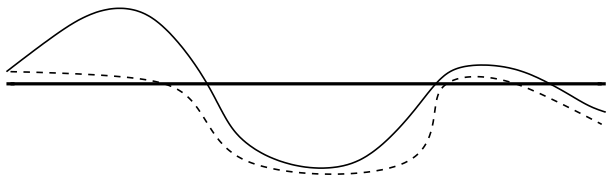

Figure 2.2: Two dimensional profile of a model of a healthy (solid line) and a glaucomatous (dashed line) papilla.

The HRT variables *varg* and *abrg* reflect the change in the upper part of the papilla, *eag*, *mhcg* and *vbrg* describe the increase of volume not covered by retinal nerve fibres. *tmg* is a measure for steepness and *ag* reflects the size of the papilla. We assign variables obtained by papillometry and visual field examinations (see section 2.1) to intermediate variables. To describe the structure of the study population, we also report the anamnestic variables *age* and intra-ocular pressure (*iop*) at the examination day.

Table 2.1 includes summary statistics of some explanatory and intermediate variables as well as of some clinical characteristics of the population. The matching by age and sex guarantees a similar distribution of these variables in both groups. The wilcoxon p-values of intermediate and explanatory variables indicate the differences between normal and glaucoma subjects.

		normal	glaucoma	p-value
clinical	*number*	85	85	-
characteristics	*sex* (f/m)	48/37	48/37	-
	iop	16.00 [14.00, 19.00]	16.00 [14.00, 19.00]	0.60
	age	56.00 [51.00, 62.00]	56.00 [51.00, 62.00]	0.93
intermediate	$\mathbf{w}_{clv}$	1.40 [0.50, 2.30]	32.60 [0.50, 2.30]	0.00
variables	$\mathbf{w}_{cs}$	1.47 [1.29, 1.58]	1.20 [1.29, 1.58]	0.00
	$\mathbf{w}_{lora}$	45.31 [36.55, 55.86]	69.84 [36.55, 55.86]	0.00
explanatory	*ag*	2.44 [2.07, 2.86]	2.57 [2.07, 2.86]	0.09
variables	*abrg*	0.86 [0.38, 1.37]	1.58 [0.38, 1.37]	0.00
	eag	1.54 [1.03, 2.12]	2.04 [1.03, 2.12]	0.00
	varg	0.35 [0.29, 0.50]	0.15 [0.29, 0.50]	0.00
	vbrg	0.19 [0.05, 0.37]	0.50 [0.05, 0.37]	0.00
	mhcg	0.07 [0.03, 0.11]	0.12 [0.03, 0.11]	0.00
	tmg	-0.15 [-0.22, -0.07]	-0.03[-0.22, -0.07]	0.00

Table 2.1: Median, lower and upper quantile ($[\cdot, \cdot]$) and p-value of wilcoxon rank sum test of intermediate variables and explanatory variables obtained by HRT examinations and some clinical characteristics. *iop* is intra-ocular pressure, a detailed description of explanatory variables is given in chapter 5.2.1.

Chapter 3

Indirect and Direct Classification

We assume a situation with three types of variables: the response variable (class membership), a set of explanatory variables (available for the prediction of future observations) and a set of intermediate variables (variables available in learning samples but not collected for future observations). In the medical context a reduction of examinations especially avoiding invasive medical tests of future patients is appropriate to spare patients' time and costs. Our aim is to construct classifiers which use the full information of the learning sample but classify a future observation based on only the reduced set of explanatory variables.

Furthermore, we distinguish between two types of class memberships: (i) the diagnosis given by the observer, i.e. for example the fixed classifying function of glaucoma diagnosis described in figure 2.1 and (ii) the true (but not observable) state of a patient. In medical application a discrepancy between observed and true diagnostic state occurs e.g. in situations where the misclassification depends on exposure. Older subjects are, for example, more often misclassified as having glaucoma, since this is an age dependent disease. In epidemiology this phenomena is called differential misclassification.

In this chapter we describe the considered discriminant model more formally and define direct and indirect classification.

3.1 The Discriminant Model

Let $\mathcal{L} := \{(\mathrm{y}_i, \mathbf{w}_i, \mathbf{x}_i), i = 1, \ldots, n\}$ denote a learning sample of n independent observations and three groups of variables: the responses $\mathrm{y}_i \in \{1, \ldots, J\}$ are the class labels, the intermediate variables are q-dimensional vectors $\mathbf{w}_i = (\mathrm{w}_{i1}, \ldots, \mathrm{w}_{iq})^\top \in \mathbb{R}^q$ and the p explanatory variables are denoted by $\mathbf{x}_i = (\mathrm{x}_{i1}, \ldots, \mathrm{x}_{ip})^\top \in \mathbb{R}^p$.
We assume the explanatory variables are related to the intermediate variables by a function

$$\begin{array}{rccc} k: & \mathbb{R}^p & \longrightarrow & \mathbb{R}^q \\ & \mathbf{x}_i & \longmapsto & \mathbf{w}_i, \end{array}$$

i.e. $\mathbf{w}_i = k(\mathbf{x}_i) + \varepsilon$, where ε is a random error. The class labels y_i are related to the intermediate variables by

$$\begin{array}{rccc} g: & \mathbb{R}^q & \longrightarrow & \{1, \ldots, J\} \\ & \mathbf{w}_i & \longmapsto & \mathrm{y}_i, \end{array}$$

which classifies an observation based on intermediate variables only: $\mathrm{y}_i = g(\mathbf{w}_i)$. Furthermore, we assume an underlying data structure, determined by a true class membership variable y_i^0. In a medical context, the assigned intermediate variable $\mathrm{y}_i = g(\mathbf{w}_i)$ represents the diagnosis of an observer, whereas y_i^0 is the real state of a patient. Therefore, explanatory, intermediate and response variables given the true state of the patient are distributed following

$$(\mathrm{Y}, \mathbf{W}, \mathbf{X}) \sim \mathcal{F}, \tag{3.1}$$

where $(\mathrm{Y}, \mathbf{W}, \mathbf{X})$ are random variables and random vectors with realizations $(\mathrm{y}_i, \mathbf{w}_i, \mathbf{x}_i), i = 1, \ldots, n$. Restricted learning samples $\mathcal{L}_{\mathrm{y},\mathbf{x}} := \{(\mathrm{y}_i, \mathbf{x}_i), i = 1, \ldots, n\}$ and $\mathcal{L}_{\mathbf{w},\mathbf{x}} := \{(\mathbf{w}_i, \mathbf{x}_i), i = 1, \ldots, n\}$ are random samples from the conditional marginal distributions $\mathcal{F}_{\mathrm{Y},\mathbf{X}}$ and $\mathcal{F}_{\mathbf{W},\mathbf{X}}$, respectively.

We define a classifier as a function

$$\begin{aligned} C: \ \mathbb{R}^p &\longrightarrow \{1, \ldots, J\} \\ \mathbf{x}_i &\longmapsto \mathrm{y}_i. \end{aligned}$$

The classifier, denoted by $C(\mathbf{x}_{\text{new}}; \mathcal{L})$, assigns a future explanatory variable $\mathbf{x}_{\text{new}}$ to a class membership and is trained on a given learning sample $\mathcal{L}$.
The two frameworks of direct and indirect classification differ in their usage of the underlying data structure and restricted learning samples.
Furthermore, the discriminant model results in three possible errors: The first reflects the difference between true and observed diagnosis, the second is the misclassification error of a classifier with respect to the true and the third is the misclassification error with respect to the observed class membership. A detailed description of these possible misclassification results is given in chapter 4.

3.2 Direct Classifiers

Direct classification methods try to estimate the relation between explanatory $\mathbf{x}_i$ and observed response variable y_i. i.e. the composition $g \circ k$. More formally, a direct classifier

$$C^{\mathrm{d}}(\mathbf{x}_{\text{new}}) := C^{\mathrm{d}}\left(\mathbf{x}_{\text{new}}; \mathcal{L}_{\mathrm{y},\mathbf{x}}\right)$$

predicts y_{new}-values for a new observation $\mathbf{x}_{\text{new}}$ based on a learning sample of responses and explanatory variables, see figure 3.1. Examples of direct classifiers are linear discriminant analysis (LDA^{d}) and classification trees ($CTREE^{\mathrm{d}}$). Bootstrap aggregation of classification trees ($bagging - CTREE^{\mathrm{d}}$) leads to a reduction of error rates in many applications, Breiman (1996a, 1998).
We repeat these direct classifiers in this section, since we use them as comparison to indirect classification techniques.

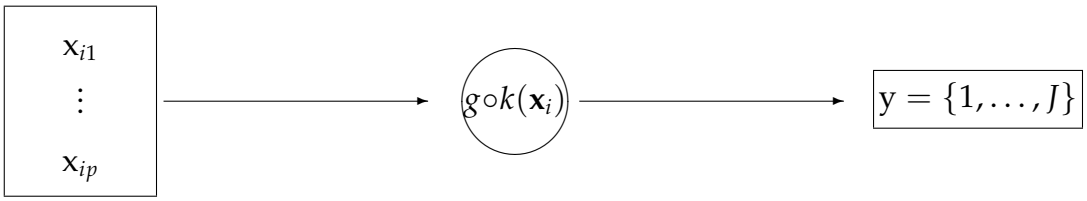

Figure 3.1: Direct classification rules are constructed based on the explanatory variables $\mathbf{x}_i := (x_{i1}, \ldots, x_{ip})^\top$ and estimate the composition $g \circ k(\cdot)$.

Linear Discriminant Analysis We apply the classical methods of linear discriminant analysis (LDA^{d}) as described by Fisher (1936) and Rao (1948).
The LDA^{d} seeks a linear combination $\mathbf{xa}$ of the explanatory variables which maximises the ratio of its between-group variance to its within-group variance, where $\mathbf{x} \in \mathbb{R}^{n \times p}$ is the matrix of explanatory variables of a given learning sample, $\mathbf{a} = (\mathbf{a}_1, \ldots, \mathbf{a}_{J'})$ and $\mathbf{a}_i \in \mathbb{R}^p$ for $i = \{1, \ldots, J'\}$ with $J' = \min\{p, J-1\}$.
Let $S_j = (n_j - 1)^{-1}(\mathbf{x}_j - \bar{\mathbf{x}}_j 1_{n_j}^\top)(\mathbf{x}_j - \bar{\mathbf{x}}_j 1_{n_j}^\top)^\top$, the within group j variance matrix. We have $\mathbf{x}_j \in \mathbb{R}^{p \times n_j}$ is the matrix for the j-th group with n_j observations, $\bar{\mathbf{x}}_j \in \mathbb{R}^p$ is the vector of means over the observations and $1_{n_j} = (1, \ldots, 1)^\top \in \mathbb{R}^{n_j}$, where $j = \{1, \ldots, J\}$. The (pooled) within group variance is defined by
$W = (n - J)^{-1} \sum_{j=1}^{J} (n_j - 1) S_j$ and the between group variance is
$B = (J-1)^{-1} \sum_{j=1}^{J} n_j (\bar{\mathbf{x}}_j - \bar{\mathbf{x}})(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})^\top$, where $\bar{\mathbf{x}} \in \mathbb{R}^p$ is the vector of means over all observations.
Differentiating the ratio of between and within group of variance $\lambda = \dfrac{\mathbf{a}^\top B \mathbf{a}}{\mathbf{a}^\top W \mathbf{a}}$ with respect to $\mathbf{a}$ and equating to zero yields $B\mathbf{a} - \lambda W \mathbf{a} = 0$. This eigenvalue equation has solutions for values of λ satisfying $|B - \lambda W| = 0$, i.e. it has solutions for eigenvalues of $W^- B$. The eigenvector $\mathbf{a}_1$ corresponding to the largest eigenvalue is the direction which leads to a maximum separation. We call this direction the first discriminant function and we have $J' = \min\{p, J-1\}$ distinct eigenvalues. We calculate subsequent eigenvectors corresponding to maxima of λ under the

constraints $\mathbf{a}_k^\top W \mathbf{a}_h = 0$ for $h = 1, \ldots, k-1$, i.e. $\mathbf{a}_k^\top \mathbf{X}$ and $\mathbf{a}_h^\top \mathbf{X}$ are uncorrelated for $h \neq k$ and $\mathbf{X} \in \mathbb{R}^p$ the explanatory random vector.
In a situation with $J = 2$ groups the eigenvector $\mathbf{a}_1$ is the only discriminant function, resulting in the allocation of a new observation to group j' if the discriminant score $h(\mathbf{x}_{\text{new}}) := \mathbf{a}_1^\top \mathbf{x}_{\text{new}}$ is $min_{j=1}^{J} |h(\mathbf{x}_{\text{new}}) - \mathbf{a}_1^\top \bar{\mathbf{x}}_j| = h(\mathbf{x}_{\text{new}}) - \mathbf{a}_1^\top \bar{\mathbf{x}}_{j'}$. $h(\mathbf{x}_{\text{new}})$ is called Fisher's linear discriminant function, more details are given in Hand (1997).

Classification Tress Given a learning sample $\mathcal{L}_{y,\mathbf{x}}$ where the response variable $y \in \{1, \ldots, J\}$ defines the class membership we construct a classification tree by recursive partitioning the learning sample, following Breiman et al. (1984). At each node $\mathcal{L}_{y,\mathbf{x}}$ is divided into two daughter nodes according to a splitting criterion which allows us to choose the best splitting covariate and the corresponding cut-point. We need an impurity measure to access the decrease of impurity from one node to its daughter nodes. Let $i(t)$ be the impurity measure of node t and $i(t_R)$, $i(t_L)$ the impurity measures of the right and left daughter nodes.
Hence we want to maximise

$$max_s \Delta i(t, s) = max_s \{i(t) - (i(t_L)p_L + i(t_R)p_R)\},$$

where p_L and p_R are the proportions of observations falling in the left and right daughter nodes and s is the split, corresponding to a certain variable and a cut-point.
Let s^* be the best split at node t, the tree decrease in impurity is given by

$$\Delta I(t, s^*) = \Delta i(t, s^*) p(t),$$

where $p(t) = \frac{\text{“number of observations in node } t''}{\text{“number of total observations”}}$ is the weight of the node. Breiman et al. (1984) define the total impurity of a tree as T as

$$I(T) = \sum_{t \in \tilde{T}} I(t) = \sum_{t \in \tilde{T}} i(t) p(t),$$

where $\tilde{T}$ is the set of terminal nodes and T is the set of splits used together with the order in which they were used.
We use Gini's index as impurity measure. The index is defined by

$$i(t) = \sum_{j' \neq j} p(j|t) * p(j'|t),$$

where $p(j|t), p(j'|t)$ are the proportions of observations in node t belonging to classes j and j'. Hence the Gini index is smaller the greater the differences in proportions $p(j|t)$ and $p(j'|t)$ are, i.e. the better node t separates the groups.

Bootstrap Aggregation An aggregated direct classifier C_A^d is defined by

$$C_A^{\mathrm{d}}(\mathbf{x}_{\mathrm{new}}; \mathcal{L}_{\mathrm{y},\mathbf{x}}) := E_{\mathcal{F}_{\mathrm{Y},\mathbf{X}}} C^{\mathrm{d}}(\mathbf{x}_{\mathrm{new}}; \mathcal{L}_{\mathrm{y},\mathbf{x}}), \tag{3.2}$$

i.e. the expectation of $C^{\mathrm{d}}(\mathbf{x}_{\mathrm{new}}; \mathcal{L}_{\mathrm{y},\mathbf{x}})$ with respect to the distribution of the learning sample. It is estimated by the bootstrap:

$$\hat{C}_A^{\mathrm{d}}(\mathbf{x}_{\mathrm{new}}; \mathcal{L}_{\mathrm{y},\mathbf{x}}) = E_{\hat{\mathcal{F}}_{\mathrm{Y},\mathbf{X}}} C^{\mathrm{d}}(\mathbf{x}_{\mathrm{new}}; \mathcal{L}^*_{\mathrm{y},\mathbf{x}}), \tag{3.3}$$

where the expected value is over $\mathcal{L}^*_{\mathrm{y},\mathbf{x}}$, a random sample with replacement from the empirical distribution function $\hat{\mathcal{F}}_{\mathrm{Y},\mathbf{X}}$. $\hat{C}_A^{\mathrm{d}}(\mathbf{x}_{\mathrm{new}}; \mathcal{L}_{\mathrm{y},\mathbf{x}})$ is approximated by drawing a finite number of bootstrap samples, for details see Breiman (1996a). More detailed a bootstrap aggregated direct classifier is calculated in three steps:

1. Draw B samples $\mathcal{L}^{*(1)}_{\mathrm{y},\mathbf{x}}, \ldots, \mathcal{L}^{*(B)}_{\mathrm{y},\mathbf{x}}$ of size n with replacement.

2. Calculate a classifier for each bootstrap sample $\mathcal{L}^{*(b)}_{\mathrm{y},\mathbf{x}}$, $b = 1, \ldots, B$.

3. Classify a new observation $\mathbf{x}_{\mathrm{new}}$ by majority voting over all predicted class memberships $C^{\mathrm{d}}(\mathbf{x}_{\mathrm{new}}; \mathcal{L}^{*(1)}_{\mathrm{y},\mathbf{x}}), \ldots, C^{\mathrm{d}}(\mathbf{x}_{\mathrm{new}}; \mathcal{L}^{*(B)}_{\mathrm{y},\mathbf{x}})$.

In the following we consider especially bootstrap aggregated classification trees (*bagging*$^{\mathrm{d}}$) as bootstrap aggregated classifiers. Classification trees are quite unstable in the sense that small changes in the learning sample can lead to large

differences in the resulting trees. Breiman (1996a) shows that bagging can give substantial gains in accuracy. We stabilise classification trees by 50 bootstrap replications.

3.3 Indirect Classifiers

Indirect classification methods make use of all variables available in the learning sample. In the following we consider two situations. On the one hand, the classifying function g is known and fixed and on the other hand it is unknown and therefore has to be estimated. In clinical context a known function g is given if e.g. a disease is defined based on certain cut-points of clinical parameters. Situations with an unknown function g occur whenever the learning sample contains more "informative" variables than the later test sample. We are interested in replacing the missing measurements with estimations based on the explanatory variables only and use this additional information to improve the indirect classifier.
Let $\hat{k}$ be any appropriate predictive model for the intermediate variables and denote $\hat{\mathbf{w}}_{\text{new}} := \hat{k}(\mathbf{x}_{\text{new}}; \mathcal{L}^{(k)})$, where $\mathcal{L}^{(k)}$ contains a subset of observations and $p+q$ variables and $\mathcal{L}^{(k)} \subseteq \mathcal{L}_{\mathbf{w},\mathbf{x}}$. Hence, $\hat{k}$ predict future intermediate variables $\mathbf{w}_{\text{new}}$ for a new observation $\mathbf{x}_{\text{new}}$ based on a restricted learning sample of explanatory and intermediate variables only.
We generally define an indirect classifier by

$$\begin{aligned} C^{\text{ind}}(\mathbf{x}_{\text{new}}) &:= C^{\text{ind}}(\mathbf{x}_{\text{new}}; \mathcal{L}) \\ &= C^{\text{ind}}(\mathbf{z}_{\text{new}}; \mathcal{L}^{(C)}), \end{aligned} \tag{3.4}$$

where $\mathbf{z}_{\text{new}} = (\mathbf{x}_{\text{new}}^\top, \hat{\mathbf{w}}_{\text{new}}^\top)^\top$. $\mathcal{L}^{(C)}$ has $m \leq n$ observations and
$\mathcal{L}^{(C)} \subseteq \mathcal{L}_{\text{y},\hat{\mathbf{w}},\mathbf{x}} = \{(\text{y}_i, \hat{\mathbf{w}}_i, \mathbf{x}_i), i = 1, \ldots, n\}$, i.e. a learning sample including response, predicted intermediate and explanatory variables.

Known classifying function Indirect classification methods in situations with known classifying function use g and estimate k instead of the composition $g \circ k$, see section 3.1. It predicts future y_{new}-values for a new observation $\mathbf{x}_{new}$ based on the complete learning sample $\mathcal{L}$ by applying g to the predicted intermediate variables. The indirect classifier of equation (3.4) becomes:

$$C^{ind}(\mathbf{x}_{new}; \mathcal{L}) = g(\hat{k}(\mathbf{x}_{new}; \mathcal{L}_{\mathbf{w},\mathbf{x}})), \tag{3.5}$$

see figure 3.2.

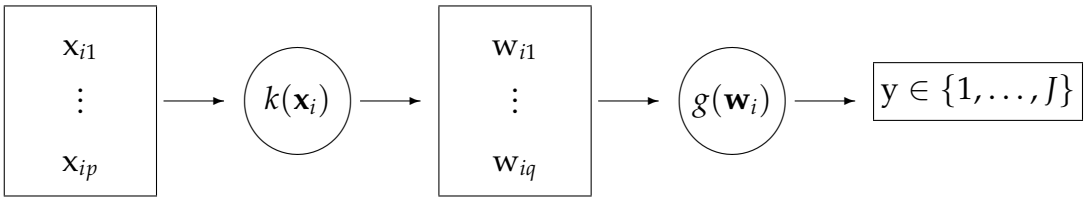

Figure 3.2: Indirect classification with known classifying function g: Models are constructed based on the explanatory variables $\mathbf{x}_i := (x_{i1}, \ldots, x_{ip})^\top$ to predict the intermediate variables $\mathbf{w}_i := (w_{i1}, \ldots, w_{iq})^\top$. Suspects are classified according g.

We can use any appropriate prediction model to estimate the function k. We focus on linear models (LM^{ind}) and regression trees ($RTREE^{ind}$) as predictors for the intermediate variables in the following.
Furthermore, we try to improve the misclassification error by bagging indirect classification and define the aggregated indirect classifier as

$$C_A^{ind}(\mathbf{x}_{new}) \;=\; E_{\mathcal{F}} C^{ind}(\mathbf{x}_{new}; \mathcal{L}) = E_{\mathcal{F}} g(\hat{k}(\mathbf{x}_{new}; \mathcal{L}_{\mathbf{w},\mathbf{x}})), \tag{3.6}$$

where the expectation is with respect to learning samples $\mathcal{L}$.
$C_A^{ind}(\mathbf{x}_{new}; \mathcal{L})$ is estimated by the bootstrap in the usual way

$$\hat{C}_A^{ind}(\mathbf{x}_{new}; \mathcal{L}) = E_{\hat{\mathcal{F}}} C^{ind}(\mathbf{x}_{new}; \mathcal{L}^*_{\mathbf{w},\mathbf{x}}) = E_{\hat{\mathcal{F}}} g(\hat{k}(\mathbf{x}_{new}; \mathcal{L}^*_{\mathbf{w},\mathbf{x}})), \tag{3.7}$$

and approximated by the following procedure:

1. Draw B bootstrap samples $\mathcal{L}^{*(1)}_{\mathbf{w},\mathbf{x}}, \ldots, \mathcal{L}^{*(B)}_{\mathbf{w},\mathbf{x}}$ from $\mathcal{L}_{\mathbf{w},\mathbf{x}}$.

2. Construct a predictor $\hat{k}$ for each bootstrap sample $\mathcal{L}^{*(b)}_{\mathbf{w},\mathbf{x}}, b = 1, \ldots, B$.

3. A new observation $\mathbf{x}_{\text{new}}$ is classified by majority voting over all predictions of the response $g(\hat{k}(\mathbf{x}_{\text{new}}; \mathcal{L}^{*(b)}_{\mathbf{w},\mathbf{x}}))$ for $b = 1, \ldots, B$.

Although Hothorn (2003) proved that bootstrap aggregating over estimations of the conditional class probabilities is Bayes consistent in a direct classification framework, we choose majority voting over all predictions of the response, since indirect classification does not offer reliable estimates of the class probabilities in any circumstances. Moreover Breiman (1996a) indicates that majority voting leads roughly to the same results as averaging over estimations of the conditional class probabilities. Furthermore, note that we aggregate with respect to the predicted class membership variable since we are interested in reducing the misclassification error rather than to improve the prediction model for the intermediate variables. Friedman (1997) gives a discussion on the connection of mean squared error and misclassification error in the context of estimated class probabilities. Using regression trees we denote the procedure outlined above "$bagging - RTREE^{\text{ind}}$" (bootstrap aggregated indirect classification using $RTREE$) and do 50 bootstrap replications, see Peters et al. (2002a).

Unknown Classifying Function In the application of an indirect classifier with an unknown classifying function $g(\cdot)$ several difficulties occur. One has to decide which model is appropriate to predict the intermediate variables and how to construct a classification rule. The relationship between explanatory, intermediate and response variables is not known and has to be estimated. Therefore, the classification rule should incorporate explanatory variables as well as predicted intermediate variables. To prevent the chance of over-fitting, prediction models for the intermediate variables should be trained on an independent learning sam-

ple $\mathcal{L}^{(k)} \subseteq \mathcal{L}_{\mathbf{w},\mathbf{x}}$ from the learning sample $\mathcal{L}^{(C)}$, where the classifier C^{ind} is trained on. More formally we require $\mathcal{L}^{(k)} \cap \mathcal{L}^{(C)} = \emptyset$, with respect to the n observations. Moreover, it is not known whether a "good" prediction of the intermediate variables, assessed by e.g. the sum of squares, results into a "good" indirect classification rule, assessed by its misclassification error. In the following we propose an algorithm which combines an arbitrary number of regression models for the intermediate variables with a subsample aggregated classifier (subagging). Bühlmann and Yu (2002) derived theoretical results of the effect of subagging decision trees in a direct classification context. They show that subagging is Bayes consistent if the size of the subsamples is chosen to be 50% of the size of the full learning sample and performs comparable to the bagging approach of Breiman (1996a).

In situations where model assumptions concerning the discriminant model of section 3.1 are not fulfilled and explanatory variables are related to response variables directly we ensure that the proposed algorithm achieves at least comparable results to direct classifiers by training the classifier on an independent learning sample which includes explanatory, predicted intermediate and response variables.

We denote this procedure by "indirect subagging" (ISB) and it works as follows:

1. Random sampling $m = [a * n]$ observations B times $\mathcal{L}^{*(1)}, \ldots, \mathcal{L}^{*(B)}$ without replacement from $\mathcal{L}$, where $0 < a < 1$ and let $\mathbf{x}^{*^{(b)}}$ denote the matrix of explanatory variables from $\mathcal{L}^{*^{(b)}}$, $b = 1, \ldots, B$ and $[\ldots]$ is the floor operation.

2. Compute r predictive models $\hat{k}_1(\cdots), \ldots, \hat{k}_r(\cdots)$ for the intermediate variables based on the explanatory variables using the so called "out-of-subag" sample $\mathcal{L}^{*^{(b)(k)}} := \mathcal{L} \backslash \mathcal{L}^{*^{(b)}}$.

3. Construct a classification tree based on the original explanatory variables as well as the predicted intermediate variables of the subag sample

$$\mathcal{L}^{*^{(b)(C)}} := (\mathbf{x}^{*^{(b)}}, \hat{k}_1(\mathbf{x}^{*^{(b)}}), \ldots, \hat{k}_r(\mathbf{x}^{*^{(b)}})).$$

4. Apply step II) and III) to all subag samples and classify a new observation by majority voting.

We use subagging as an criterion to split the learning sample into $\mathcal{L}^{(k)}$ and $\mathcal{L}^{(C)}$, respectively and vary the parameter $a \in \{0.25, 0.5, 0.75\}$ in applications of indirect subagging on real data sets.

We choose regression models incorporated in the indirect subagging approach covering different types of possible relationships between explanatory and intermediate variables: Linear models are appropriate to estimate linear dependencies, regression trees estimate tree-based dependencies and projection pursuit regression is appropriate to model nearly every continuous relationship (Friedman and Stuetzle, 1981; Meyer et al., 2003). Consequently, we calculate different indirect subagging classifiers using linear models (LM^{isb}) or regression trees ($RTREE^{\text{isb}}$) to predict intermediate variables for future datasets. The classifier $LM + RTREE^{\text{isb}}$ uses both techniques and $LM + RTREE + PPR^{\text{isb}}$ combines linear models, regression trees and projection pursuit method, where the plus sign indicates the combination of the different regression methods.

In contrast to "indirect subagging", Hothorn (2003) and Hothorn and Lausen (2003b) propose the combination of different classification models via a bootstrap aggregated classifier in a direct classification framework. Their approach "bundling" combines different direct classifiers, whereas "indirect subagging" deals with an indirect classification approach. Our proposal improves the discriminant value of the predictors where the final classifier is based on, rather than combining different types of direct classifiers.

More detailed, indirect subagging applies the described regression models to each "out-of-subag" sample $\mathcal{L}^{*^{(b)(k)}}, b = 1, \ldots, B$ in a first step. In a second step, a classification tree is fitted based on the extended set of predictors of the subag samples. These extended sets of predictors consist of the original explanatory

variables and one set of predicted intermediate variables for each fitted regression model. Consequently, the classification trees constructed within the indirect subagging algorithm $LM + RTREE + PPR^{\text{isb}}$ are based on three times as many predicted intermediate variables as these trees fitted within the algorithm LM^{isb}. We draw $B = 50$ subag samples to improve indirect subagging.

Chapter 4

Error Rates

As described in section 3.1 we differentiate between two types of class memberships: a true class membership y_i^0, which causes an underlying data structure, and an observed class membership y_i, which mimics the diagnosis assigned by a physician, see Peters and Lausen (2003). This results in different types of errors caused by a wrong class labelling and by a wrong decision of the applied classification rule.

In the following we use the notation of Efron and Tibshirani (1997). Q_1 denotes the misclassification loss function, where

$$Q_1(z_1, z_2) = \begin{cases} 1, & \text{if} \quad z_1 \neq z_2 \\ 0, & \text{if} \quad z_1 = z_2. \end{cases} \tag{4.1}$$

Given a new observation $(y_{new}^0, y_{new}, \mathbf{w}_{new}, \mathbf{x}_{new})$, where $y_{new} = g(\mathbf{w}_{new})$ is the observed diagnosis and y_{new}^0 is the true disease state, we have three types of losses:

1. $Q_1(y_{new}^0, y_{new})$, loss caused by the observed diagnosis,

2. $Q_1(y_{new}, C(\mathbf{x}_{new}; \mathcal{L}))$, loss caused by classifier $C(\mathbf{x}_{new}; \mathcal{L})$,

3. $Q_1(y_{new}^0, C(\mathbf{x}_{new}; \mathcal{L}))$, overall loss caused by the whole classification process.

Resulting true error rates, i.e. the probabilities that the classifier will misclassify a new observation, corresponding to the three loss functions are denoted by

$$\begin{aligned} \mu_1 &:= E[Q_1(\mathrm{Y}^0_{\mathrm{new}}, \mathrm{Y}_{\mathrm{new}})], \\ \mu_2(C(\mathbf{X}_{\mathrm{new}}; \mathcal{L})) &:= E[Q_1(\mathrm{Y}_{\mathrm{new}}, C(\mathbf{X}_{\mathrm{new}}; \mathcal{L})] \text{ and} \\ \mu_3(C(\mathbf{X}_{\mathrm{new}}; \mathcal{L})) &:= E[Q_1(\mathrm{Y}^0_{\mathrm{new}}, C(\mathbf{X}_{\mathrm{new}}; \mathcal{L})], \end{aligned}$$

where the expectation is with respect to the unconditional distribution of the learning sample. The μ_1 error describes the discrepancy between the true state of the patient and the observed diagnosis and is therefore a measure for the degree of label correctness in test samples. This type of error, called differential misclassification, is often discussed in epidemiological context (Flegal et al., 1991; Rothman and Greenland, 1998; Grimes and Schulz, 2002). In epidemiology it describes the difference between the true and the observed state of a patient, where the proportion of misclassification depends on exposure. In the medical field this situation may occur, for example, lunge cancer is less likely to be detected within a group of non-smokers or glaucoma is more often false diagnosed within a population of elderly people.

The misclassification rate with respect to the observed diagnosis (μ_2) corresponds to misclassification results in real data situations, whereas the μ_3 error rates of the classifier is the true misclassification error, i.e. the error with respect to the true state of the patient. Although a good classification rule should minimise this kind of error, there is often no opportunity to assess it in reality.

Consequently, the expected true error rates μ_2^{d} and μ_3^{d} for direct classification, i.e. error rates over a design set of a given size, are denoted by

$$\begin{aligned} \mu_2^{\mathrm{d}}(C^{\mathrm{d}}(\mathbf{X}_{\mathrm{new}}; \mathcal{L}_{\mathrm{y},\mathbf{x}})) &= E_{\mathcal{F}'_{\mathrm{Y},\mathbf{X}}} E_{\mathcal{F}'_{\mathrm{Y}_{\mathrm{new}},\mathbf{x}_{\mathrm{new}}}} Q_1(\mathrm{Y}_{\mathrm{new}}, C^{\mathrm{d}}(\mathbf{X}_{\mathrm{new}}; \mathcal{L}_{\mathrm{y},\mathbf{x}})) \text{ and} \\ \mu_3^{\mathrm{d}}(C^{\mathrm{d}}(\mathbf{X}_{\mathrm{new}}; \mathcal{L}_{\mathrm{y},\mathbf{x}})) &= E_{\mathcal{F}'_{\mathrm{Y},\mathbf{X}}} E_{\mathcal{F}'_{\mathrm{Y}^0_{\mathrm{new}},\mathbf{x}_{\mathrm{new}}}} Q_1(\mathrm{Y}^0_{\mathrm{new}}, C^{\mathrm{d}}(\mathbf{X}_{\mathrm{new}}; \mathcal{L}_{\mathrm{y},\mathbf{x}})), \end{aligned}$$

where $E_{\mathcal{F}'_{\mathrm{Y}_{\mathrm{new}},\mathbf{x}_{\mathrm{new}}}}$ and $E_{\mathcal{F}'_{\mathrm{Y}^0_{\mathrm{new}},\mathbf{x}_{\mathrm{new}}}}$ refers to unconditional expectation (with respect to the true class membership) over future observations $(\mathrm{Y}_{\mathrm{new}}, \mathbf{X}_{\mathrm{new}})$ and

$(\mathrm{Y}^0_{\text{new}}, \mathbf{X}_{\text{new}})$. The expectation $E_{\mathcal{F}'_{\mathrm{Y},\boldsymbol{x}}}$ is over the unconditional distribution of the restricted learning samples $\mathcal{L}_{\mathrm{y},\boldsymbol{x}}$.

Furthermore, for indirect classification, the expected true error rates are given by

$$\mu_2^{\text{ind}}(C^{\text{ind}}(\mathbf{X}_{\text{new}}; \mathcal{L})) = E_{\mathcal{F}'} E_{\mathcal{F}'_{\mathrm{Y}_{\text{new}}, \mathbf{X}_{\text{new}}}} Q_1(\mathrm{Y}_{\text{new}}; C^{\text{ind}}(\mathbf{X}_{\text{new}}; \mathcal{L})) \text{ and} \tag{4.2}$$

$$\mu_3^{\text{ind}}(C^{\text{ind}}(\mathbf{X}_{\text{new}}; \mathcal{L})) = E_{\mathcal{F}'} E_{\mathcal{F}'_{\mathrm{Y}^0_{\text{new}}, \mathbf{X}_{\text{new}}}} Q_1(\mathrm{Y}^0_{\text{new}}; C^{\text{ind}}(\mathbf{X}_{\text{new}}; \mathcal{L})) \tag{4.3}$$

$E_{\mathcal{F}'}$ refers to the unconditional distribution of the complete learning sample $\mathcal{L}$.

In contrast to misclassification loss, let Q_2 denote squared error loss

$$Q_2(\hat{k}(\mathbf{X}_{\text{new}}; \mathcal{L}_{\mathbf{w},\boldsymbol{x}}), \mathbf{w}_{\text{new}}) = \sum_{j=1}^{q} (\hat{k}(\mathbf{X}_{\text{new}}; \mathcal{L}_{\mathbf{w},\boldsymbol{x}})_j - (\mathrm{w}_{\text{new}})_j)^2. \tag{4.4}$$

The mean squared error is defined by

$$\mu_{\text{MSE}} = E_{\mathcal{F}'} E_{\mathcal{F}'_{\mathbf{W}_{\text{new}}, \mathbf{X}_{\text{new}}}} Q_2(\mathbf{W}_{\text{new}}, \hat{k}(\mathbf{X}_{\text{new}}; \mathcal{L}_{\mathbf{w},\boldsymbol{x}})), \tag{4.5}$$

where again $E_{\mathcal{F}'}$ is with respect to learning samples $\mathcal{L}_{\mathbf{w},\boldsymbol{x}}$ and $E_{\mathcal{F}'_{\mathbf{W}_{\text{new}}, \mathbf{X}_{\text{new}}}}$ refers to unconditional expectation over future explanatory and intermediate variables $(\mathbf{W}_{\text{new}}, \mathbf{X}_{\text{new}})$.

However, there is no indication that improving the prediction model $\hat{k}$ with respect to mean squared error μ_{MSE} does improve the misclassification errors μ_2 or μ_3 of the indirect classifier C^{ind} simultaneously.

4.1 Bayes and Indirect Classifier using a known Classifying Function - Asymptotic Properties

The Bayes error is the minimum possible error rate in a given learning sample and therefore provides a lower bound on any error rate which may be achieved by a real classification rule. In the following we analyse the performance of μ_1, μ_2 and μ_3 errors of the Bayes classifier and an indirect classifier with known classifying

function under a certain discriminant model. Note that we calculate the Bayes error following the framework of direct classification, hence it neglects knowledge connected with the group of intermediate variables and therefore stands for the minimal error rate achieved on a reduced learning set.
The discriminant model of chapter 3.1 is a framework which allows a magnitude of different distributions of the learning sample. The calculation of error rates under these general model assumptions is highly complicated. In the following section we restrict the model to normally distributed explanatory and intermediate variables to analyse the performance of error rates of direct and indirect classifiers.

Simple discriminant model In order to analyse asymptotic behaviours of the discussed classifiers we have to restrict ourselves to a "simple" discriminant model, i.e. we impose model assumptions of normally distributed explanatory and intermediate variables. We assume a binary true state $\mathrm{y}_i^0 \in \{0,1\}$ and a binary class membership variable $\mathrm{y}_i \in \{0,1\}$ describing the observed diagnosis. $\mathbf{X} \in \mathbb{R}^p$ denotes the multivariate normal random vector of explanatory variables,
$\mathbf{X} \sim N_p(\mathbf{c}, \sigma_{\mathbf{x}}^2 \mathcal{I}_p)$, where $\mathcal{I}_p$ represents the identity matrix of dimension p and $\sigma_{\mathbf{x}}^2 > 0$. Given the true state of the patient Y^0, we have $\mathbf{c} = \begin{cases} \mathbf{c}^{(1)}, & \text{if} \quad \mathrm{y}^0 = 1 \\ \mathbf{c}^{(0)}, & \text{if} \quad \mathrm{y}^0 = 0 \end{cases}$,
where $\mathbf{c}, \mathbf{c}^{(0)}, \mathbf{c}^{(1)} \in \mathbb{R}^p$.
We decompose $\mathbf{X} = \mathbf{c} + \varepsilon_{\mathbf{X}}$ and define $\varepsilon_{\mathbf{X}} \sim N_p(0, \sigma_{\mathbf{x}}^2 \mathcal{I}_p)$ as the measurement error of explanatory variables.
We calculate error rates for a simplified situation with one intermediate variable W, i.e. $q = 1$. The random variable $\mathrm{W} \in \mathbb{R}$ is a linear transformation of the explanatory variables: $\mathrm{W} = \beta^\top \mathbf{X} + \varepsilon_W$, where $\varepsilon_{\mathrm{W}} \sim N(0, \sigma_{\mathrm{w}}^2)$ is the measurement

error of the intermediate variables. The observed class membership is defined by

$$\mathrm{Y} \;=\; \begin{cases} 1, & \text{if} \qquad \mathrm{W} > \delta \\ 0, & \text{else,} \end{cases}$$

where δ is a given cut-point.

As an indirect classification rule, we incorporate the ordinary least squares estimator $\hat{\beta}$ to predict the intermediate variables. The intermediate variable of a new observation $(\mathrm{y}^0_{\text{new}}, \mathrm{y}_{\text{new}}, \mathrm{w}_{\text{new}}, \mathbf{x}_{\text{new}})$ is predicted following

$$\hat{\mathbf{w}}_{\text{new}} \;=\; \hat{\beta}^\top \mathbf{x}_{\text{new}},$$

where $\hat{\beta} = (\mathbf{x}^\top \mathbf{x})^{-1}\mathbf{x}^\top \mathbf{w}$, $\mathbf{x} = (\mathbf{x}_1 \cdots \mathbf{x}_n)^\top \in \mathbb{R}^{n\times p}$, $\mathbf{w} = (\mathrm{w}_1 \cdots \mathrm{w}_n)^\top \in \mathbb{R}^{n\times q}$ and $q = 1$, i.e. the outcomes of the learning sample.

Differential misclassification The μ_1 error rate represents the difference between true state of the patient and observed diagnosis, which depends on the intermediate variables, in given test samples. In our setup it is

$$\begin{aligned} \mu_1 &= \pi_0 \cdot P(\mathrm{Y} = 1 | \mathrm{Y}^0 = 0) + \pi_1 \cdot P(\mathrm{Y} = 0 | \mathrm{Y}^0 = 1) \\ &= \pi_0 \cdot \left\{ 1 - \Phi\left(\frac{\delta - \mathbf{c}^{(0)^\top}\beta}{\sigma} \right) \right\} + \pi_1 \cdot \Phi\left(\frac{\delta - \mathbf{c}^{(1)^\top}\beta}{\sigma} \right), \end{aligned}$$

where $\pi_0 = P(\mathrm{Y}^0 = 0), \pi_1 = P(\mathrm{Y}^0 = 1)$, $\sigma^2 := \mathsf{var}(\mathrm{W}) = \sigma^2_{\mathbf{x}} \beta^\top \beta + \sigma^2_{\mathrm{w}}$ and $\Phi(\cdot)$ is the standard normal distribution function. The given cut-point δ is chosen with respect to $0 < \delta - \beta^\top c^{(0)}$ and $0 > \delta - \beta^\top c^{(1)}$, hence $\Phi\left(\frac{\delta - \mathbf{c}^{(0)^\top}\beta}{\sigma}\right) \in [0.5, 1]$ and decreases with increasing σ and $\Phi\left(\frac{\delta - \mathbf{c}^{(1)^\top}\beta}{\sigma}\right) \in [0, 0.5]$ and increases with increasing σ. Resulting, differential misclassification is high for large variances of the measurement error of explanatory or intermediate variables. The change of the μ_1 error rate is displayed in figure 4.1. For this graphical representation we choose $\beta = (0.1, \ldots, 0.1)^\top \in \mathbb{R}^{10}$, $\delta = 2$, $c^{(0)} = (0, \ldots, 0)^\top \in \mathbb{R}^{10}$,
$c^{(1)} = (4, \ldots, 4)^\top \in \mathbb{R}^{10}$ and display $\sigma_{\mathbf{x}}$ and $\sigma_{\mathbf{w}} \in \{0, \ldots, 5\}$ on the x and y axis. The z axis describes the value of the resulting μ_1 error.

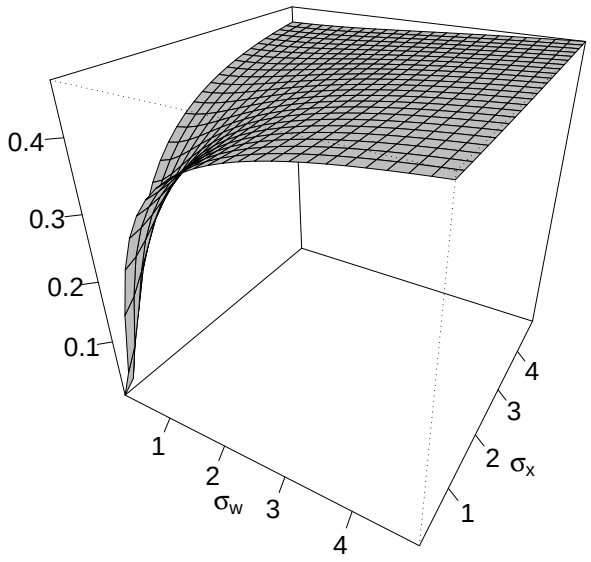

Figure 4.1: μ_1 error plotted against increasing variation of explanatory and intermediate variables.

Observed misclassification The discrepancy between predicted class membership following a classification rule and the observed class membership is measured in terms of the μ_2 error rate. In a real data example, calculated error rates reflect the μ_2 error rate, because one does usually not distinguish between a real state and an observed class membership variable. We investigate the performance of an indirect classifier with respect to this observable error rate and calculate the μ_2 errors of the Bayes and the indirect classifier. The Bayes classifier is given by

$$C^{\text{Bayes}}(\mathbf{x}_{\text{new}}) = \begin{cases} 1, & \text{if } P(\mathrm{Y}=1|\mathbf{X}=\mathbf{x}_{\text{new}}) = \max_{k\in\{0,1\}} P(\mathrm{Y}=k|\mathbf{X}=\mathbf{x}_{\text{new}}) \\ 0, & \text{else.} \end{cases}$$

Hence, it allocates a new observation $\mathbf{x}_{\text{new}}$ to class "1" whenever $P(\mathrm{Y}=1|\mathbf{X}=\mathbf{x}_{\text{new}}) > P(\mathrm{Y}=0|\mathbf{X}=\mathbf{x}_{\text{new}})$. We rewrite

$$\begin{aligned} & P(\mathrm{Y}=1,\mathbf{X}=\mathbf{x}_{\text{new}}|\mathrm{Y}^0) &>& P(\mathrm{Y}=0,\mathbf{X}=\mathbf{x}_{\text{new}}|\mathrm{Y}^0) \\ \Leftrightarrow\ & P(\beta^\top\mathbf{x}_{\text{new}}+\varepsilon_{\mathrm{W}}>\delta|\mathrm{Y}^0) &>& P(\beta^\top\mathbf{x}_{\text{new}}+\varepsilon_{\mathrm{W}}\leq\delta|\mathrm{Y}^0) \\ \Leftrightarrow\ & 0 &>& \delta-\beta^\top\mathbf{x}_{\text{new}}. \end{aligned}$$

The $\mu_2(C^{\text{Bayes}}(\mathbf{x}_{\text{new}}))$ error of the Bayes classifier is calculated:

$$\begin{aligned} & \mu_2(C^{\text{Bayes}}(\mathbf{X}_{\text{new}})) \qquad (4.6) \\ =\ & \pi_0\cdot P(\mathrm{Y}\neq C^{\text{Bayes}}(\mathbf{X})|\mathrm{Y}^0=0)+\pi_1\cdot P(Y\neq C^{\text{Bayes}}(\mathbf{X})|\mathrm{Y}^0=1) \\ =\ & \pi_0\cdot\left\{P(\mathrm{W}>\delta,\beta^\top\mathbf{X}<\delta|\mathrm{Y}^0=0)+P(\mathrm{W}<\delta,\beta^\top\mathbf{X}>\delta|\mathrm{Y}^0=0)\right\}+ \\ & \pi_1\cdot\left\{P(\mathrm{W}>\delta,\beta^\top\mathbf{X}<\delta|\mathrm{Y}^0=1)+P(\mathrm{W}<\delta,\beta^\top\mathbf{X}>\delta|\mathrm{Y}^0=1)\right\}, \end{aligned}$$

where $(W,\beta^\top\mathbf{X}|\mathrm{Y}^0=j)$, $j=\{0,1\}$ is bivariate normal with mean $(\beta^\top\mathbf{c}^{(j)},\beta^\top\mathbf{c}^{(j)})^\top$ and restricted covariance matrix $\mathsf{cov}(W,\beta^\top\mathbf{X}|\mathrm{Y}^0=j)=\begin{pmatrix}\sigma^2 & \sigma^2-\sigma_{\mathrm{w}}^2\\ \sigma^2-\sigma_{\mathrm{w}}^2 & \sigma^2-\sigma_{\mathrm{w}}^2\end{pmatrix}$.
The μ_2 error rate for the described indirect classifier is calculated analogously:

$$\begin{aligned} & \mu_2(C^{\text{ind}}(\mathbf{X}_{\text{new}};\mathcal{L}_{\mathbf{w},\mathbf{x}})) \qquad (4.7) \\ =\ & \pi_0\cdot P(g(\mathrm{W})\neq g(\hat{\mathrm{W}})|\mathrm{Y}^0=0)+\pi_1\cdot P(g(\mathrm{W})\neq g(\hat{\mathrm{W}})|\mathrm{Y}^0=1) \\ =\ & \pi_0\cdot\left\{P(\mathrm{W}>\delta,\hat{\mathrm{W}}<\delta|\mathrm{Y}^0=0)+P(\mathrm{W}<\delta,\hat{\mathrm{W}}>\delta|\mathrm{Y}^0=0)\right\}+ \\ & \pi_1\cdot\left\{P(\mathrm{W}>\delta,\hat{\mathrm{W}}<\delta|\mathrm{Y}^0=1)+P(\mathrm{W}<\delta,\hat{\mathrm{W}}>\delta|\mathrm{Y}^0=1)\right\}, \end{aligned}$$

where $(W,\hat{W}|\mathrm{Y}^0=j)$, $j=\{0,1\}$ is bivariate normal with mean $(\beta^\top\mathbf{c}^{(j)},\hat{\beta}^\top\mathbf{c}^{(j)})^\top$ and restricted covariance matrix $\mathsf{cov}(W,\hat{W}|\mathrm{Y}^0=j)=\begin{pmatrix}\sigma^2 & \sigma_{\mathbf{x}}^2\hat{\beta}^\top\beta\\ \sigma_{\mathbf{x}}^2\hat{\beta}^\top\beta & \sigma_{\mathbf{x}}^2\hat{\beta}^\top\hat{\beta}\end{pmatrix}$.
We rewrite the summands of the Bayes classifier (4.6):

$$\begin{aligned} & P(\mathrm{Y}\neq C^{\text{Bayes}}(\mathbf{X})|\mathrm{Y}^0) \\ =\ & P(\mathrm{Y}=0,C^{\text{Bayes}}(\mathbf{X})=1|\mathrm{Y}^0)+P(\mathrm{Y}=1,C^{\text{Bayes}}(\mathbf{X})=0|\mathrm{Y}^0) \\ =\ & P(\beta^\top\mathbf{X}+\varepsilon_{\mathrm{W}}<\delta,\beta^\top\mathbf{X}>\delta|\mathrm{Y}^0)+P(\beta^\top\mathbf{X}+\varepsilon_{\mathrm{W}}>\delta,\beta^\top\mathbf{X}<\delta|\mathrm{Y}^0). \end{aligned}$$

The summands of the μ_2 error rate of the indirect classifier (4.7) can be rewritten analogously. This representation reveals that we obtain large μ_2 errors for increased absolute values of ε_W in situations where the variance σ_x^2 is small. The expressions $\beta^\top \mathbf{X} + \varepsilon_W < \delta$ and $\beta^\top \mathbf{X} > \delta$ are determined by the measurement error of the intermediate variables ε_W. Therefore, the probability $P(Y \neq C^{\text{Bayes}}(\mathbf{X})|Y^0)$ should be larger for small variances σ_x, since either
$P(\beta^\top \mathbf{X} + \varepsilon_W < \delta, \beta^\top \mathbf{X} > \delta|Y^0)$ or $P(\beta^\top \mathbf{X} + \varepsilon_W > \delta, \beta^\top \mathbf{X} < \delta|Y^0)$ becomes large for a large negative or positive measurement error ε_W. Considering a situation, where $\text{var}(W) = \sigma^2$ causes the variance of $\text{var}(\beta^\top \mathbf{X}) = \sigma^2 - \sigma_w^2$. The μ_2 errors of the Bayes and indirect classifiers should decrease here, since for large σ^2 it follows $\sigma^2 \approx \sigma^2 - \sigma_w^2$ and $P(\beta^\top \mathbf{X} - \mathbf{W} < \delta, \beta^\top \mathbf{X} > \delta|Y^0)$ decreases with increasing σ_x^2. We display the performance of the $\mu_2(C^{\text{Bayes}}(\cdot))$ error in figure 4.2, using the same parameters as in figure 4.1.
Moreover, considering asymptotic properties of the indirect classifier, from the consistency of $\hat{\beta}$ follows the consistency of the indirect classifier.
Lemma:
Given the data structure described in chapter 4.1, the indirect classifier using the ordinary least squares estimate to predict intermediate variables of future observations is Bayes consistent with respect to the μ_2 error rate, i.e.

$$P(|\mu_2(C^{\text{Bayes}}(\mathbf{X}_{\text{new}})) - \mu_2(C^{\text{ind}}(\mathbf{X}_{\text{new}}; \mathcal{L}_n))| > \varepsilon) \overset{n\to\infty}{\longrightarrow} 0 \; \forall \varepsilon > 0,$$

where $\mathcal{L}_n$ is a sequence of learning samples $\mathcal{L}_n := \{(w_i, \mathbf{x}_i), i = 1, \ldots, n\}$.

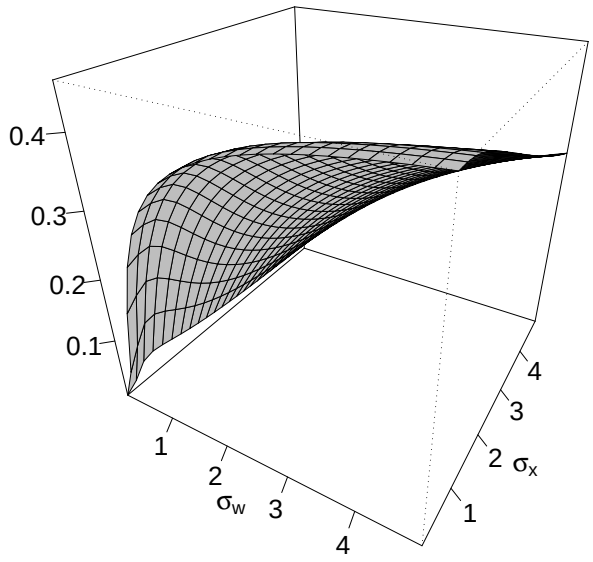

Figure 4.2: $\mu_2(C^{\text{Bayes}}(\cdot))$ error plotted against increasing variation of intermediate and explanatory variables.

Proof:

We rewrite

$$|\mu_2(C^{\text{Bayes}}(\mathbf{X}_{\text{new}})) - \mu_2(C^{\text{ind}}(\mathbf{X}_{\text{new}}; \mathcal{L}_n))| \leq \tag{4.8}$$

$$\left|\pi_0 \left(\int_{-\infty}^{\delta}\int_{\delta}^{\infty}(f_{(\mathrm{W},\hat{\mathrm{W}}_n|\mathrm{Y}^0=0)}(\mathrm{x},\mathrm{y}) - f_{(\mathrm{W},\mathbf{X}^\top\beta|\mathrm{Y}^0=0)}(\mathrm{x},\mathrm{y}))dx\,dy\right)\right| + \tag{4.9}$$

$$\left|\pi_0 \left(\int_{\delta}^{\infty}\int_{-\infty}^{\delta}(f_{(\mathrm{W},\hat{\mathrm{W}}_n|\mathrm{Y}^0=0)}(\mathrm{x},\mathrm{y}) - f_{(\mathrm{W},\mathbf{X}^\top\beta|\mathrm{Y}^0=0)}(\mathrm{x},\mathrm{y}))dx\,dy\right)\right| +$$

$$\left|\pi_1 \left(\int_{-\infty}^{\delta}\int_{\delta}^{\infty}(f_{(\mathrm{W},\hat{\mathrm{W}}_n|\mathrm{Y}^0=1)}(\mathrm{x},\mathrm{y}) - f_{(\mathrm{W},\mathbf{X}^\top\beta|\mathrm{Y}^0=1)}(\mathrm{x},\mathrm{y}))dx\,dy\right)\right| +$$

$$\left|\pi_1 \left(\int_{\delta}^{\infty}\int_{-\infty}^{\delta}(f_{(\mathrm{W},\hat{\mathrm{W}}_n|\mathrm{Y}^0=1)}(\mathrm{x},\mathrm{y}) - f_{(\mathrm{W},\mathbf{X}^\top\beta|\mathrm{Y}^0=1)}(\mathrm{x},\mathrm{y}))dx\,dy\right)\right|,$$

where $\hat{\mathrm{W}}_n = \hat{\beta}_n^\top \mathbf{X}$, $\hat{\beta}_n := (\mathbf{x}^\top\mathbf{x})^{-1}\mathbf{x}^\top\mathbf{w}$, $\mathbf{x} = (\mathbf{x}_1 \cdots \mathbf{x}_n)^\top$ and $\mathbf{w} = (\mathrm{w}_1 \cdots \mathrm{w}_n)^\top$. $f_{(\mathrm{W},\hat{\mathrm{W}}_n|\mathrm{Y}^0=j)}(\mathrm{x},\mathrm{y})$ is the conditional bivariate normal density function with mean

$(\beta^\top \mathbf{c}^{(j)}, \hat{\beta}^\top \mathbf{c}^{(j)})^\top$ and covariance $\text{cov}(W, \hat{W}_n | \text{Y}^0 = j) = \begin{pmatrix} \sigma^2 & \sigma^2_{\mathbf{x}} \hat{\beta}_n^\top \beta \\ \sigma^2_{\mathbf{x}} \hat{\beta}_n^\top \beta & \sigma^2_{\mathbf{x}} \hat{\beta}_n^\top \hat{\beta} \end{pmatrix}$ and $f_{(\text{W}, \mathbf{X}^\top \beta | \text{Y}^0 = j)}(\text{x}, \text{y})$ is the bivariate normal density function with mean $(\beta^\top \mathbf{c}^{(j)}, \beta^\top \mathbf{c}^{(j)})^\top$ and covariance $\text{cov}(W, \beta^\top \mathbf{X} | \text{Y}^0 = j) = \begin{pmatrix} \sigma^2 & \sigma^2_{\mathbf{x}} - \sigma^2_{\mathbf{w}} \\ \sigma^2 - \sigma^2_{\mathbf{w}} & \sigma^2 - \sigma^2_{\mathbf{w}} \end{pmatrix}$, with $j \in \{0, 1\}$. We know that under the assumed normal distribution, $\hat{\beta}_n$ is consistent for β and therefore $\mathbf{x}^\top \hat{\beta}_n$ is consistent for $\mathbf{x}^\top \beta$.

From the consistency of the estimator $\mathbf{x}^\top \hat{\beta}_n$ follows the convergence in law (Theorem 2.3.5, Lehmann, 1999), i.e.

$$\mathcal{F}_{\text{W}, \mathbf{X}^\top \hat{\beta}_n} \overset{n \to \infty}{\longrightarrow} \mathcal{F}_{\text{W}, \mathbf{X}^\top \beta},$$

where $\mathcal{F}_{\cdot, \cdot}$ notate conditional distributions. This is equivalent to

$$\int \int_S f_{(\text{W}, \mathbf{X}^\top \hat{\beta}_n | \text{Y}^0)}(\text{x}, \text{y}) d\text{x}\, d\text{y} \overset{n \to \infty}{\longrightarrow} \int \int_S f_{(\text{W}, \mathbf{X}^\top \beta | \text{Y}^0)}(\text{x}, \text{y}) d\text{x}\, d\text{y},$$

$\forall$ sets S for which the probabilities in question are defined and for which the boundary of S has probability zero under the distribution of $(\text{W}, \mathbf{X}^\top \beta | \text{Y}^0)$ (Theorem 1.7, p. 343, Lehmann, 1991). Hence, each addend of equation (4.9) converges to zero and

$$P(|\mu_2(C^{\text{Bayes}}(\mathbf{x}_{\text{new}})) - \mu_2(C^{\text{ind}}(\mathbf{x}_{\text{new}}; \mathcal{L}_n))| > \varepsilon) \leq P(\underbrace{\text{expression}(4.9)}_{\longrightarrow 0} > \varepsilon) \overset{n \to \infty}{\longrightarrow} 0.$$

□

True misclassification Although in real data examples we can only assess the error with respect to the observed class membership, the μ_3 error quantifies the discrepancy between the true state of the patient and the predicted class memberships. The Bayes rule for the μ_3 error is

$$C^{\text{Bayes}}(\mathbf{x}_{\text{new}}) = \begin{cases} 1, & \text{if } P(\text{Y}^0 = 1 | \mathbf{X} = \mathbf{x}_{\text{new}}) = \max_{k \in \{0,1\}} P(\text{Y}^0 = k | \mathbf{X} = \mathbf{x}_{\text{new}}) \\ 0, & \text{else.} \end{cases}$$

Hence, it allocates to class "1" whenever
$P(\mathrm{Y}^0 = 1|\mathbf{X} = \mathbf{x}_{\text{new}}) > P(\mathrm{Y}^0 = 0|\mathbf{X} = \mathbf{x}_{\text{new}})$. We rewrite

$$A = (\mathbf{c}^{(0)} - \mathbf{c}^{(1)})^\top (\sigma_{\mathbf{X}}^2 \mathcal{I}_{q\times q})^{-1} (\mathbf{x}_{\text{new}} - \frac{1}{2}(\mathbf{c}^{(0)} + \mathbf{c}^{(1)})) > \log\left(\frac{\pi_0}{\pi_1}\right),$$

for details see Ripley (1996), pp. 21-22.
If $\mathbf{X}$ comes from class "0" then $A \sim N(\frac{1}{2}\zeta^2, \zeta^2)$ and $A \sim N(-\frac{1}{2}\zeta^2, \zeta^2)$ if $\mathbf{X}$ comes from class "1", where $\zeta = \sqrt{(\mathbf{c}^{(0)} - \mathbf{c}^{(1)})^\top (\sigma_{\mathbf{x}}^2 \mathcal{I}_{q\times q})^{-1} (\mathbf{c}^{(0)} - \mathbf{c}^{(1)})}$. Hence,

$$\begin{aligned} &\mu_3(C^{\text{Bayes}}(\mathbf{X}_{\text{new}})) \\ &= \pi_0 \cdot \Phi\left(-\frac{1}{2}\zeta + \frac{1}{\zeta}\log\left(\frac{\pi_1}{\pi_0}\right)\right) + \pi_1 \cdot \Phi\left(-\frac{1}{2}\zeta - \frac{1}{\zeta}\log\left(\frac{\pi_1}{\pi_0}\right)\right). \end{aligned} \tag{4.10}$$

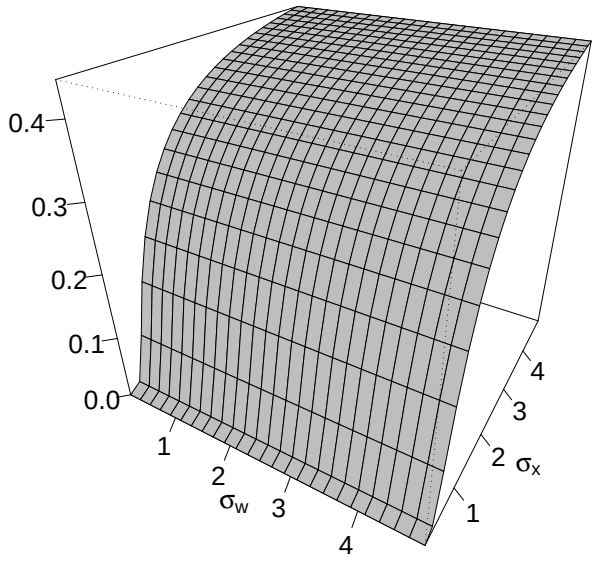

Figure 4.3: $\mu_3(C^{\text{Bayes}}(\cdot))$ error plotted against increasing variation of intermediate and explanatory variables.

The performance of the μ_3 Bayes error is displayed in figure 4.3, choosing the parameters of figure 4.1. We calculate the $\mu_3(C^{\text{ind}}(\mathbf{x}_{\text{new}}; \mathcal{L}_{\mathbf{w},\mathbf{x}}))$ error analogously:

$$\begin{aligned} &\mu_3(C^{\text{ind}}(\mathbf{X}_{\text{new}}; \mathcal{L}_{\mathbf{w},\mathbf{x}})) \\ &= \pi_0 \cdot \left\{ 1 - \Phi\left(\frac{\delta - \mathbf{c}^{(0)\top}\beta}{\sqrt{\sigma_x^2 \hat{\beta}^\top \hat{\beta}}} \right) \right\} + \pi_1 \cdot \Phi\left(\frac{\delta - \mathbf{c}^{(1)\top}\beta}{\sqrt{\sigma_x^2 \hat{\beta}^\top \hat{\beta}}} \right). \end{aligned} \tag{4.11}$$

The variance of the measurement error of the intermediate variable does not influence the μ_3 Bayes error rate, see equation (4.10). In comparison, the μ_3 error rate of the indirect classifier increases with an increase of σ_w, since the estimation of $\hat{\beta}$ achieves absolute larger values for large σ_w and, consequently, the addends in expression (4.11) increase with increasing variance, compare paragraph "differential misclassification". The variance of the measurement error of the explanatory variables affects the μ_3 error of the Bayes classifier as well as the error of the indirect rule.

4.2 Indirect Subagging - Asymptotic Properties

In the previous section, we discussed asymptotic properties of an indirect classifier using a fixed classifying function by imposing very restrictive model assumptions. Results of Hothorn and Lausen (2003b) and Hothorn (2003) enable us to examine the asymptotic performance of the indirect subagging approach without these restrictions in this paragraph.

Hothorn and Lausen (2003b) proved the Bayes consistency of bootstrap aggregated classifiers in a direct classification framework. The authors do not distinguish between a true and an observed class membership, they consider a learning sample with one class membership variable reflecting the observed class membership in our framework. Furthermore they assume an aggregation that averages the conditional class probability estimators per bootstrap (or subag) sample

and chooses the class with highest average class probability, rather than the approach of majority voting. However, Breiman (1996a) indicates that these two voting algorithms lead roughly to the same results. Consequently, we conclude based on the results of Hothorn (2003) that the observed misclassification error of indirect subagging approximates the observed misclassification error of the Bayes classifier. This Bayes consistency is proofed in Theorem 1 and 2 of Hothorn (2003) for an indirect subagging approach that aggregates over conditional class probability estimators.

More specifically, Theorem 1 of Hothorn (2003) shows the Bayes consistency of a classification tree calculated on a united subag sample $\mathcal{L}^{*^{(b)(C)}}$, where
$b = 1, \ldots, B$ is the number of subag samples to be drawn and
$\mathcal{L}^{*^{(b)(C)}} \subset \mathcal{L}^{(C)} := \{(\mathrm{y}_i, \mathbf{x}_i, \hat{k}_1(\mathbf{x}_i), \ldots, \hat{k}_r(\mathbf{x}_i)), i = 1, \ldots, n\}$ is a learning sample including explanatory and response variable and a limited number of additional predictors, see section 3.3. They prove this theorem by an extension of the proof of Theorem 3, Lugosi and Nobel (1996) where the Bayes consistency of classification trees is shown.
In Theorem 2 Hothorn (2003) proves that subagging (or bagging) by averaging over conditional class probability estimators with a fixed and finite number of subag samples and classification trees corresponding to each subag sample constructed in the described way is again Bayes consistent.
Altogether we conclude that the proposed procedure of indirect subagging is Bayes consistent with respect to the observed misclassification error. However, theoretic considerations of the true misclassification error of indirect subagging are complicated, since the true class membership does not appear in learning samples but affects the distribution of intermediate, explanatory and, consequently, observed class membership variables. We avoid these calculations and refer to simulation results.

4.3 Comparison of Error Rates - Finite Sample Situation

In section 4.1 we calculated error rates of Bayes and indirect classifiers incorporating a fixed classifying function and a linear model and analysed asymptotic properties for indirect subagging and this specialised indirect classifier. In this section we indicate the performance of direct and indirect classification in finite sample situations.

We compare calculated error rates of the Bayes classifier and the indirect classifier with a known classifying function with simulated error rates of the direct classifier LDA^{d} and indirect subagging in the discriminant model described in section 4.1. We choose the linear discriminant analysis as a comparable direct classification rule since it is optimal in situations with normally distributed data and does a linear separation of the decision space of explanatory variables. The calculation of LDA^{d} error rates is widely discussed, but complicated in situations where the exact distribution of the given learning sample is not known (McLachlan, 1975; Läuter, 1992). Our focus is the comparison between direct and indirect classification rules, therefore we simulate the misclassification results for the LDA^{d} and the indirect subagging classifier. For indirect subagging we include the ordinary least squares and do 50 subag samples (LM^{isb}).

We restrict ourselves to a situation with $p = 2$, i.e. $\mathbf{X} \sim N_2(\mathbf{c}, \sigma_{\mathbf{x}}\mathcal{I}_2)$, to visualise the analysed behaviour of an indirect classifier compared to the Bayes error and the direct classification rule LDA^{d}. Additionally, for given learning samples we approximate $\mu_2(C^{\mathrm{ind}}(\mathbf{x}_{\mathrm{new}}; \mathcal{L}_{\mathbf{w},\mathbf{x}}))$ and simulate misclassification errors for the linear discriminant analysis by generating 100 observations per class and doing 1000 iterations. The error rates for the Bayes classifier and $\mu_3(C^{\mathrm{ind}}(\mathbf{x}_{\mathrm{new}}; \mathcal{L}))$ are calculated numerically.

Figures 4.4 and 4.5 display calculated error rates for $\mathbf{c}^{(0)} = (0.5, 0.5)^\top$,

$\mathbf{c}^{(1)} = (1.5, 1.5)^\top$, $\beta = (0.5, 0.5)^\top$ and $\delta = 1$. In figure 4.4 we increase $\sigma_{\mathbf{x}} = (0.1, \ldots, 5)^\top$ and set $\sigma_{\mathrm{w}} = 0.1$.

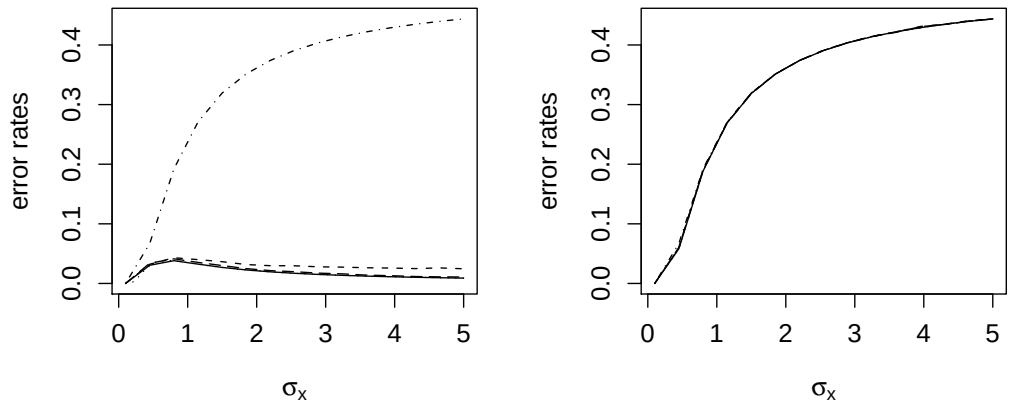

Figure 4.4: μ_2 (left side) and μ_3 (right side) misclassification errors of Bayes (dotted), direct (dashed), indirect (solid) classifiers, indirect subagging (longdash) and differential misclassification (dotdash) with increasing standard deviations $\sigma_{\mathbf{x}}$ of the measurement error of the explanatory variables and $\sigma_{\mathrm{w}} = 0.1$.

A direct classifier predicts future class memberships based on explanatory variables only. Therefore, we expect a superiority of the discussed indirect classifier compared to the LDA^{d} for large values of $\sigma_{\mathbf{x}}^2$ and small values of σ_{w}^2.
Simulated results in Figure 4.4 show, that the applied indirect classification technique outperforms the LDA^{d} with respect to the μ_2 error especially for large variances of the explanatory variables. We achieve μ_2 and μ_3 errors almost equal to the Bayes errors using indirect classification. Differential misclassification reflects the difference between the observed and the true class membership. We examined in section 4.1 that it increases with an increase of the variance of the measurement error. The divergence of these prior probabilities of true and observed

class memberships does not lead to an increase of the misclassification results of different classification techniques simultaneously (Hand et al., 1998). Resulting, the μ_1 error diverges from μ_2 and μ_3 error.
We demonstrate the performance for an increasing measurement error of the intermediate variables $\sigma_w = (0.1, \ldots, 5)^\top$ and $\sigma_x = 0.1$ in figure 4.5

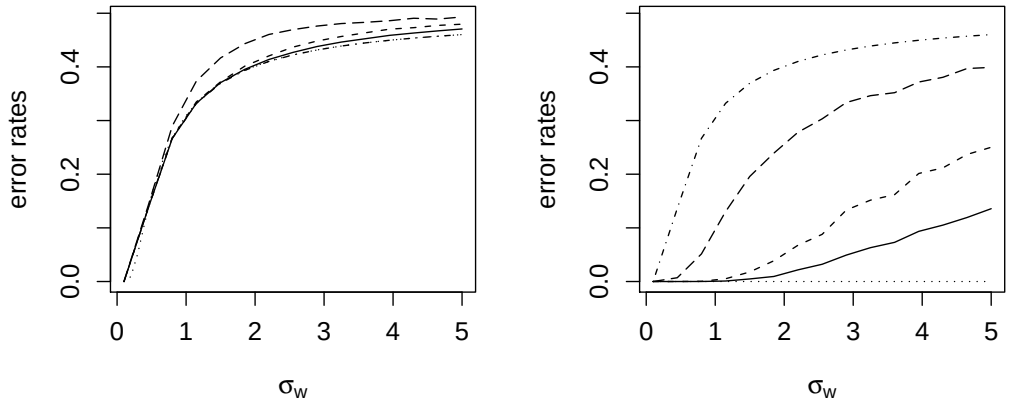

Figure 4.5: μ_2 (left side) and μ_3 (right side) misclassification errors of Bayes (dotted), direct (dashed), indirect (solid) classifiers, indirect subagging (longdash) and differential misclassification (dotdash) with increasing standard deviations σ_w and $\sigma_x = 0.1$.

The indirect classifier is superior to the LDA^d with respect to the μ_2 and μ_3 errors for large variances of the measurement error of the intermediate variables. Performances of indirect classifier and Bayes classifier with respect to the observed diagnosis are calculated in equations (4.7, 4.6). These error rates are influenced by the variance of the measurement error σ_w in the same way, therefore the resulting μ_2 errors of the Bayes classifier and indirect classifier are almost the same, see Figure 4.5. In contrast to an indirect approach, a direct classification rule does

not incorporate the definition of the observed diagnosis, i.e. it does not reflect the distinction between an observed and a true class membership. A direct classifier seeks for similarities within the groups determined by the observed class membership and therefore, considering the error with respect to the true state of the patient, the direct classification rule appears to be more affected by differential misclassification than the indirect one. Indirect subagging is highly influenced by an increase of σ_w.

Chapter 5

Simulation

The usage of simulation models is an established tool to mimic data structures and to generate independent learning and test samples. Therefore, we construct simulation models with a known distribution of the generated samples to analyse general characteristics of indirect vs. direct classifiers.

First, we investigate the performance of direct classifiers versus indirect classifiers using a known classifying function and indirect subagging in situations with three different dependencies between explanatory, intermediate and response variables. For this purpose we extend the simple discriminant model of section 4.3 and mimic data structures where the intermediate variables embed all, little and none of the discriminant information. Note that the indirect classifier using the fixed classifying function $g(\cdot)$ as defined in section 3.3 works with a misleading classifying function if the intermediate variables do not embed the whole discriminant information. $g(\cdot)$ reflects a misspecified relation between intermediate and observed response variables, i.e. a misspecified decision rule of a physician, in these situations.

Second, we investigate different classifiers for the task of glaucoma classification. We examine the performance of direct and indirect classifiers for glaucoma diagnosis by an extending the simulation model of Hothorn and Lausen (2003a).

5.1 Simple Model

We generate learning and test samples using the following simulation model. We have a binary class membership variable $y_i^0 \in \{0, 1\}$ which describes the true state of the patient. Furthermore, $\mathbf{X} \in \mathbb{R}^p$ is the p dimensional multivariate normal random vector of explanatory variables $\mathbf{X} \sim N_p(\mathbf{c}, \sigma_{\mathbf{x}}^2 \mathcal{I})$, where $\mathcal{I}_p$ represents the identity matrix of dimension p. Given the true class membership of a patient we have $\mathbf{c} = \begin{cases} \mathbf{c}^{(0)}, & \text{if } y_i^0 = 0 \\ \mathbf{c}^{(1)}, & \text{if } y_i^0 = 1 \end{cases}$. The random vector of intermediate variables $\mathbf{W} \in \mathbb{R}^q$ is a linear transformation of the explanatory variables: $\mathbf{W} = \beta^\top \mathbf{X} + \varepsilon$, where $\beta \in \mathbb{R}^{q \times p}$ and $\varepsilon \in \mathbb{R}^q$ is a normal distributed measurement error with $\varepsilon \sim N_q(0, \sigma_{\mathbf{w}}^2 \mathcal{I}_q)$. We define the observed diagnosis Y of a patient as $Y = h(\mathbf{Z})$, where $\mathbf{Z} = (\mathbf{W}^\top, \mathbf{X}^\top)^\top \in \mathbb{R}^{p+q}$.

We set the parameters $\mathbf{X} \in \mathbb{R}^{10}$, $p = 10$, $\sigma_{\mathbf{x}}^2 = 2$, $\sigma_{\mathbf{w}}^2 = 1$,
$\mathbf{c} = \begin{cases} \mathbf{c}^{(0)} = (0, \ldots, 0)^\top, & \text{if } y_i^{(0)} = 0 \\ \mathbf{c}^{(1)} = (4, \ldots, 4)^\top, & \text{else.} \end{cases}$ and choose the linear dependence between explanatory and intermediate variable $\mathbf{W} \in \mathbb{R}^2$ by
$\beta_j = (1, \ldots, 1)^\top \in \mathbb{R}^{10}$, $j = \{1, 2\}$. Note that the data generating process differs from the model assumptions, introduced in section 3.1. The observed diagnosis is defined on explanatory and intermediate variables $Y = h(\mathbf{Z})$, whereas an indirect classifier incorporates a classifying function, which is based on the intermediate variables only: $C^{\text{ind}}(\mathbf{x}_{\text{new}}; \mathcal{L}) = g(\hat{\mathbf{w}}_{\text{new}})$.

5.1.1 Setups

We consider situations with a decision surface following a cut-point model on the one hand and with a tree-based decision surface on the other hand.

Setup 1 In this setup we choose a cut-point model as the relationship between $\mathbf{Z}$ and the response Y:

$$\mathrm{Y} = \begin{cases} 1, & \text{if } \gamma^\top \mathbf{Z} > 0 \\ 0, & \text{else.} \end{cases}$$

We simulate three sub-setups.

Setup 1.1: Response depends only on intermediate variables:
$\gamma = (0.5, 0.5, 0, \ldots, 0)^\top \in \mathbb{R}^{12}$,

Setup 1.2: Response depends on intermediate and explanatory variables:
$\gamma = (1/12, \ldots, 1/12)^\top \in \mathbb{R}^{12}$,

Setup 1.3: Response depends only on explanatory variables:
$\gamma = (0, 0, 0.1, \ldots, 0.1)^\top \in \mathbb{R}^{12}$

and fix the differential misclassification (difference between true and observed class membership) to about 25% for all sub-setups by the choice of the cut-point and the weightings for intermediate and explanatory variables.
The indirect classification approach incorporates a fixed classifying function, compare section 3.3. We determine this fixed classifying function in setups $1.1 - 1.3$ by

$$\begin{aligned} C^{\text{ind}}(\mathbf{x}_{\text{new}}; \mathcal{L}) &= g(\hat{k}(\mathbf{x}_{\text{new}}; \mathcal{L}_{\mathbf{w},x})) \\ &= g(\hat{\mathbf{w}}_{\text{new}}) \\ &= \begin{cases} 1, & \text{if } (0.5, 0.5) \cdot \hat{\mathbf{w}}_{\text{new}} > 0 \\ 0, & \text{else,} \end{cases} \end{aligned}$$

i.e. it is based on intermediate variables only and reflects the true data structure in setup 1.1. Note that in setups 1.2 and 1.3 this function does not reflect the functional relationship between explanatory, intermediate and response variables, it stands for a misspecified decision rule of a physician. For example, in setup 1.3

the observed class membership is defined by a linear combination of explanatory variables only, whereas the classifying function incorporated in the indirect classification approach depends only on intermediate variables and is therefore totally misspecified.

Setup 2: We consider a tree-based relationship between intermediate and response here, see figure 5.1.
More detailed, the conditions $cond_1$, $cond_2$ and $cond_3$ are in the three sub-setups:

Setup 2.1: Response depends only on intermediate variables:
$cond_1 : \mathbf{w}_1 \geq 20$; $cond_2 : \mathbf{w}_2 < 44.5$; $cond_3 : \mathbf{w}_2 < -4.5$,

Setup 2.2: Response depends on intermediate and explanatory variables:
$cond_1 : \mathbf{w}_1 \geq 20$; $cond_2 : \mathbf{x}_1 < 5.35$; $cond_3 : \mathbf{x}_1 < -1.35$,

Setup 2.3: Response depends only on explanatory variables:
$cond_1 : \mathbf{x}_1 \geq 2$; $cond_2 : \mathbf{x}_2 < 6.5$; $cond_3 : \mathbf{x}_2 < -2.5$.

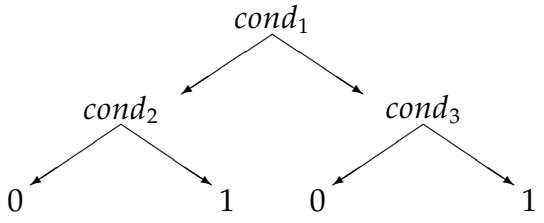

Figure 5.1: Tree-based relationship between intermediate and response.

Again the cut-points are chosen with respect to a differential misclassification of 25%. The fixed classifying function incorporated in the indirect classification framework is given by:

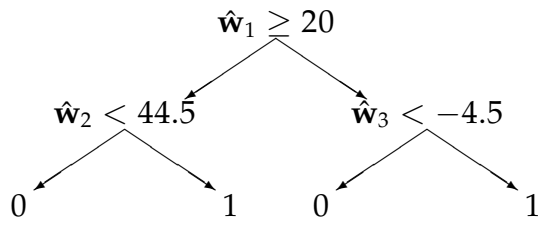

Figure 5.2: Fixed classifying function incorporated into indirect classification.

As described for setups simulating linear dependencies in the previous section, this function simultaneously reflects the true relationship between intermediate and observed class membership variables in setup 2.1 and differs from the observed diagnosis in setups 2.2 and 2.3.

The simulation model is our tool to investigate the performance of different classifiers in learning and test samples following different decision surfaces, i.e. we have six different relationships between observed class membership and explanatory and intermediate variables. Therefore we generate cases and controls of learning and a test sample following the true class membership and apply the decision surfaces to the generated data. More specifically, we generate a learning and test sample with 200 observations: 100 with true class membership "0" and 100 with true class membership "1". We do 1000 replications.

We divide the learning sample into 50% out-of-subag and subag samples for the indirect subagging classifiers. Subagging and bagging are performed with 50 subag and bootstrap samples, respectively.

5.1.2 Results

The observed misclassification μ_2, i.e. the difference between predicted class membership and observed class membership, is minimised for the indirect classification approach in most setups.
Considering the setups where the indirect classifier incorporates the true relationship between intermediate and response variable (setup 1.1 and 2.1), indirect approaches perform well as long as linear models are used to predict the intermediate variables, which is the true data structure here. In setup 1.1 the μ_2 error of LM^{ind} achieves an error rate of 1.8% and of 4.1% in setup 2.1. Indirect subagging has comparable results in these setups. In contrast, the best results of the direct classifiers are achieved by $bagging^{\text{d}}$ of about 11.8% in setup 1.1 and 24% in setup 2.1.

Direct classifiers in setups 2.2 and 2.3 are improved for these decision surfaces not following the assumptions of the discriminant model, described in chapter 3.1. $bagging^{\text{d}}$ is a tree-based classifier constructed on explanatory variables only. It performs best in setup 2.3 with an error rate of 0.8%. However it is superior to LDA^{d} even in situations where the decision space is generated by a cut-point model in setups $1.1 - 1.3$. This can be caused by subgroups generated by different distributions of the learning sample with respect to the true diagnosis.
The indirect subagging classifiers always perform at least comparable to the indirect classifiers and better in situations where indirect classifiers LM^{ind}, $RTREE^{\text{ind}}$ and $bagging - RTREE^{\text{ind}}$ incorporate misleading decision rules, i.e. fixed classifying functions which differ from the true relationship between explanatory, intermediate and response variables. They achieve good results as long as the linear model, which represents the relationship between explanatory and intermediate variables, is included. Adding several "alternative" regression models, e.g. regression trees or projection pursuit method, does not affect their performance.

		1.1	1.2	1.3	2.1	2.2	2.3
μ_1		0.250	0.250	0.250	0.242	0.251	0.248
μ_2	LDA^{d}	0.131	0.131	0.131	0.329	0.149	0.148
	$bagging^{\mathrm{d}}$	0.118	0.118	0.117	0.240	0.022	0.008
	LM^{ind}	0.018	0.013	0.250	0.041	0.304	0.369
	$RTREE^{\mathrm{ind}}$	0.205	0.206	0.250	0.333	0.319	0.295
	$bagging\text{-}RTREE^{\mathrm{ind}}$	0.137	0.137	0.249	0.275	0.279	0.264
	LM^{isb}	0.020	0.014	0.006	0.044	0.010	0.048
	$RTREE^{\mathrm{isb}}$	0.144	0.144	0.142	0.291	0.037	0.046
	$LM + RTREE^{\mathrm{isb}}$	0.020	0.014	0.006	0.044	0.010	0.047
	$LM + RTREE + PPR^{\mathrm{isb}}$	0.021	0.015	0.006	0.044	0.010	0.046
μ_3	LDA^{d}	0.161	0.158	0.161	0.107	0.264	0.202
	$bagging^{\mathrm{d}}$	0.262	0.258	0.259	0.164	0.265	0.246
	LM^{ind}	0.250	0.249	0.500	0.238	0.239	0.240
	$RTREE^{\mathrm{ind}}$	0.211	0.210	0.500	0.122	0.121	0.124
	$bagging\text{-}RTREE^{\mathrm{ind}}$	0.270	0.272	0.500	0.052	0.054	0.053
	LM^{isb}	0.246	0.248	0.250	0.246	0.255	0.240
	$RTREE^{\mathrm{isb}}$	0.287	0.285	0.285	0.139	0.283	0.241
	$LM + RTREE^{\mathrm{isb}}$	0.247	0.246	0.252	0.247	0.254	0.244
	$LM + RTREE + PPR^{\mathrm{isb}}$	0.246	0.247	0.250	0.247	0.254	0.238

Table 5.1: Simulation results of setups $1.1 - 2.3$ of the simple model. $a = 0.5$ for indirect subagging.

Considering the μ_3 error rates, i.e. the difference between predicted class membership and the true diagnostic state of a patient, direct classifiers perform satisfactorily in many situations. Especially LDA^{d} errors are about 15% in setups $1.1 - 1.3$, whereas the other classification techniques have a misclassification rate of at least 20%. The indirect subagging classifiers often achieve misclassification errors similar to the proportion of differential misclassification. In setup $2.1 - 2.3$ the indirect approach $bagging - RTREE^{\mathrm{ind}}$ achieves the best results among all considered classifiers with a misclassification rate of about 5%.

Figures 5.3 and 5.4 show the distribution of μ_2 and μ_3 error rates of setups 1.1, 1.3, 2.1 and 2.3 and for selected classifiers. Considering figure 5.3, the μ_2 error is smaller than the estimated μ_3 error for all classifiers, the differences between these error rates of indirect classifiers is larger than the differences of direct classifiers. The variance is minimised for indirect subagging in most situations, the μ_3 error rates for the indirect classifiers LM^{ind} and $bagging - RTREE^{\mathrm{ind}}$ in setup 1.3 are about 50% for each replication of the Monte-Carlo simulation.

Estimated error rates of direct, indirect and indirect subagging classifiers are less uniformly distributed for the tree-based decision surface in setups $2.1 - 2.3$, see figure 5.4.

In setup 2.1, the variance of estimated error rates of indirect subagging is again smaller than the variance of the other classification techniques but it is increased in setup 2.3. Moreover, the difference between μ_2 and μ_3 errors of the direct classifier LDA^{d} is as large as the differences corresponding to indirect classifiers and indirect subagging in setup 2.1. In setup 2.3 the direct classifier $bagging^{\mathrm{d}}$ has the larger difference. Even the relation between these two kinds of errors shifts between setups 2.1 and 2.3 for some classifiers. In setup 2.1 it is $\mu_3 < \mu_2$ for direct classifiers and $\mu_2 < \mu_3$ for LM^{ind}, whereas the relation of these three classifiers is the other way round in setup 2.3.

Altogether, indirect subagging appears to estimate the observed diagnosis

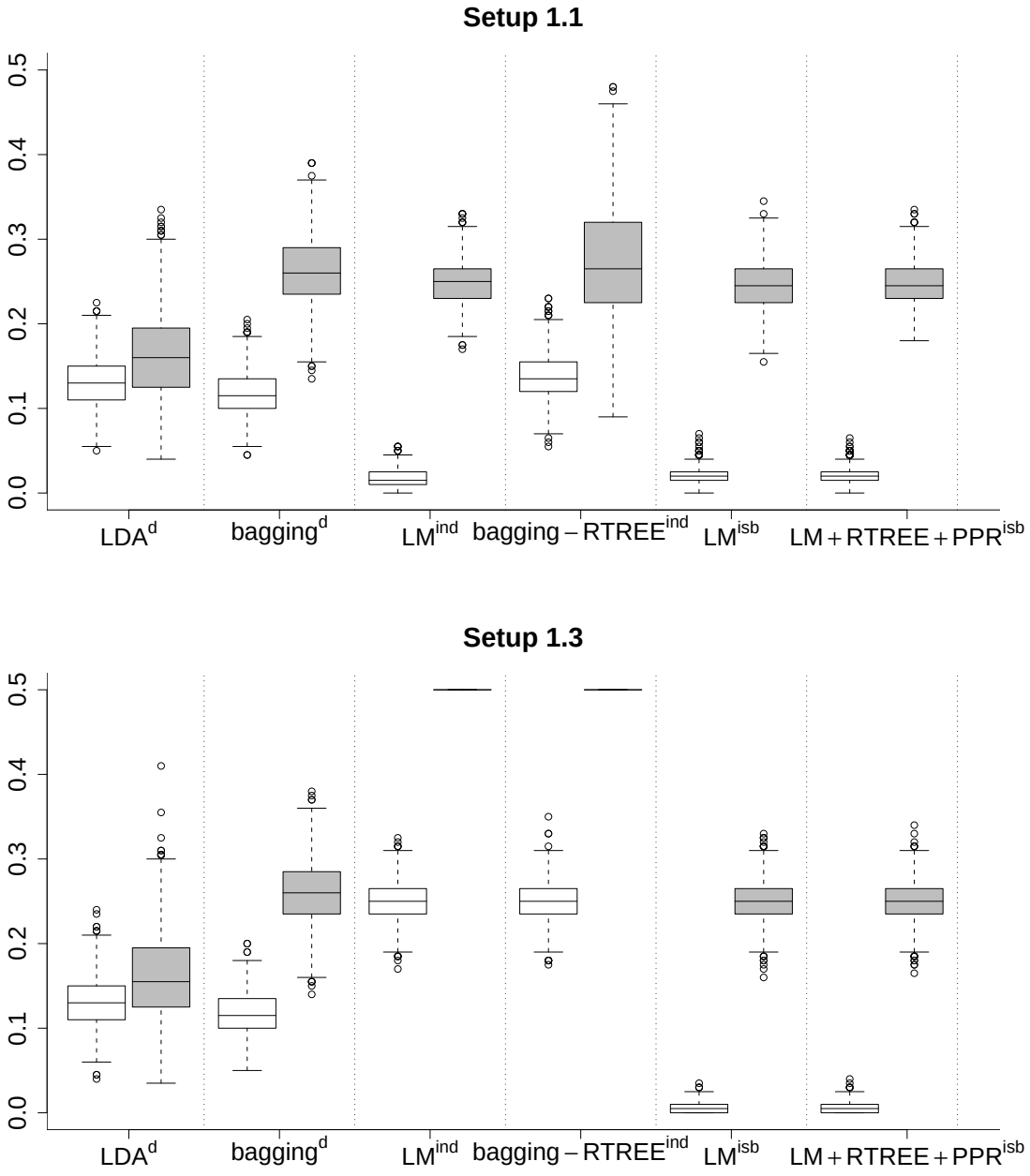

Figure 5.3: Distribution of estimated μ_2 (white boxes, left) and μ_3 (grey boxes, right) error rates of selected classifiers in setup 1.1 and 1.3.

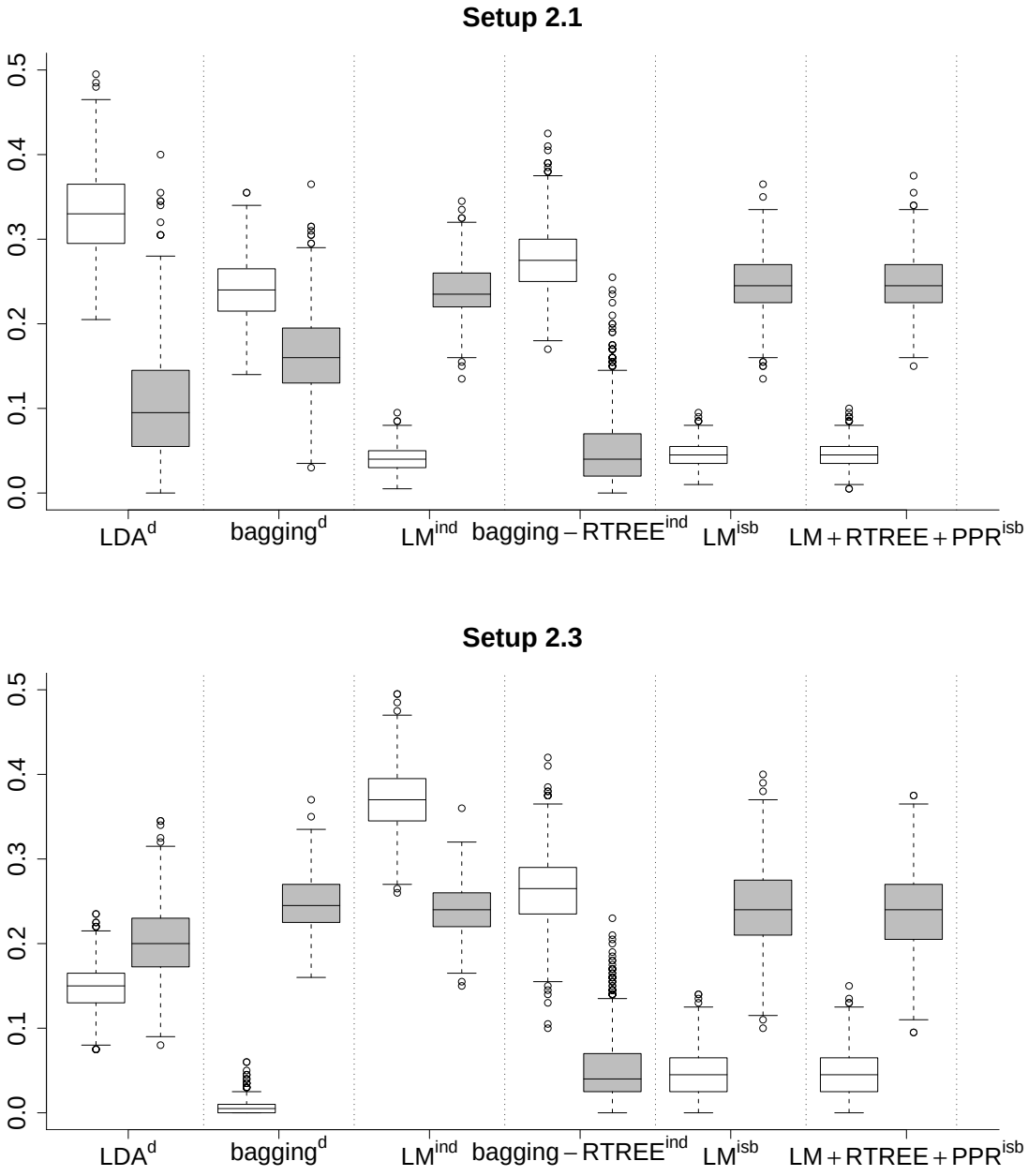

Figure 5.4: Distribution of estimated μ_2 (white boxes, left) and μ_3 (grey boxes, right) error rates of selected classifiers in setup 2.1 and 2.3.

better than the true class membership and achieves the smallest variance of all considered classification techniques. This result is plausible, since it does not include the definition of the observed class membership, which is correctly specified in setups 1.1 and 2.1 but totally misspecified in setups 1.3 and 2.3. Furthermore, it combines linear relationships and tree-based relationships and therefore, the incorporated data structures in every setup.

The direct classifiers also estimate the observed class membership better than the true class membership in most situations (setups 1.1, 1.3 and 2.3). In setups 1.3 and 2.3 the diagnosis of the observed class membership depends only on the explanatory variables and the direct classifier is appropriate to model this relationship. On the contrary, if the observed diagnosis is exclusively determined by the intermediate variables and additionally the relationship between intermediate and response variables is not the same kind of relationship between explanatory and intermediate variables, direct classifiers have smaller μ_3 errors than μ_2 errors (setup 2.1). If both relationships are linear, a direct classifier is again appropriate to model the combination of the linear relation between explanatory and intermediate variables and intermediate variables and response.

Indirect classifiers which include a fixed classifying function based on the set of intermediate variables only, mimic the process of medical decision making in setups 1.1 and 2.1. Therefore, they achieve smaller errors due to the observed diagnosis in these situations. The results of setups 1.3 and 2.3 indicate that a performance of this kind of classification technique is also affected by the underlying data structure. In these setups the observed diagnosis depends on explanatory variables only. The μ_2 error rate of the indirect classifier is smaller than its μ_3 error rate in a situation where the dependencies between intermediate and response variables are linear. This relation is the other way round in tree-based situations.

5.2 Glaucoma Model

We evaluate the discussed direct and indirect classifiers based on the example of glaucoma classification. We consider glaucoma classification as an example of medical decision making. Hence a true classifying function $g(\cdot)$ is known and we incorporate it into the indirect classification approach. We discussed some details about the disease in section 2.1 and introduced the classifying function. In this section we propose a simulation model which mimics data structures of the given case-control study described in section 2.2.

5.2.1 Model

In the following we present a simulation model which mimics HRT measurements as well as the outcome of visual field and papillometric data. Parameters included in the simulation model are extracted from the case-control study introduced in section 2.2, the simulated data structure is similar to the real-life data structure in the case-control study. Therefore, this model is our tool to investigate the performance of direct and indirect classifiers in the task of glaucoma classification.

We simulate HRT and visual field measurements as well as parameters extracted from photographs of the optic nerve head. More specifically, we derive 26 variables for the HRT simulation characterising the optic nerve head morphology, two variables $\mathbf{w}_{cs}$ (contrast sensitivity) and $\mathbf{w}_{clv}$ (corrected loss variance) characterising the visual field defect and one variable $\mathbf{w}_{lora}$ (loss of rim area) approximating measurement results from fundus photography.

For HRT measurements we use the simulation model proposed by Hothorn and Lausen (2003a). This model calculates the morphology of the optic nerve head (ONH) as introduced in Swindale et al. (2000), where the surface (s_η) is

defined by:

$$s_\eta(u,v) = \\ -\left(\frac{z}{1+e^{(r-r_0)/s}} + a(u-u_0) + b(v-v_0) + c(u-u_0)^2 + d(v-v_0)^2 + z_0\right),$$

with $r = \|(u-u_0, v-v_0)\|_2$, the 2-norm, and parameter vector

$$\eta = (z, z_0, r_0, s, u_0, v_0, a, b, c, d).$$

The optic nerve head has the shape of a cup, a two dimensional profile of a model image is displayed in figure 5.5.

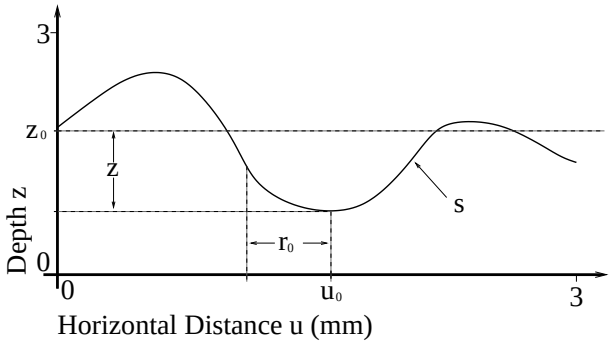

Figure 5.5: Two dimensional profile of a model image along the horizontal axis at $v_0 = v$ illustrating the meaning of some of the model parameters.

Swindale et al. (2000) discuss the interpretation of the different parameters, and a detailed description is given in table 5.2.

The arguments u and v vary between 0 and 3 mm, which is the area scanned by the HRT. The parameter vector $\eta^{(N)}$ is estimated from the normal subjects and $\eta^{(G)}$ from the glaucoma subjects of the case-control study introduced in the previous chapter. These vectors help to achieve a more realistic model of an artificial normal and glaucomatous ONH morphology and are summarised in table 5.3. Given a surface s_η we simulate predictors derived by the HRT as described in Hothorn

(u_0, v_0)	centre of the cup
z	measure of cup depth
z_0	baseline height of all images
s	steepness of slope of the cup walls
r_0	distance of the cup wall (at half-height) from the canter of the cup
a	overall component of tilt along the nasotemporal axis
b	overall component of tilt along the vertical axis
c	overall curvature along the nasotemporal axis
d	overall curvature along the vertical axis

Table 5.2: Description of the elements of the parameter vector η.

	u_0	v_0	z_0	r_0	z
$\eta^{(N)}$	1.489	1.404	0.755	0.525	0.726
$\eta^{(G)}$	1.489	1.404	0.755	0.640	0.697
	s	a	b	c	d
$\eta^{(N)}$	0.118	0.009	0.027	0.046	-0.045
$\eta^{(G)}$	0.094	0.013	0.012	-0.019	-0.083

Table 5.3: Values for HRT simulation.

and Lausen (2003a) and Hothorn and Lausen (2002). The surface is computed on a grid over the scanning area $[0, 3]^2$ segmented into 40×40 points. To model the measurement error by the HRT we add independent identically distributed normal errors with variance σ^2_{HRT} to the surface at each point of the grid. Increasing variances of the measurement error σ^2_{HRT} of the explanatory variables leads to an increasing overlap of distributions of the variables corresponding to normal and glaucomatous eyes respectively.

Several morphological variables are extracted from the simulated surfaces, which

mimic parameters of the commercial HRT software (Heidelberg Engineering, 1997). Figure 5.6 displays Swindale et al. (2000) model of the ONH morphology of a normal and a glaucomatous eye respectively for the parameters of table 5.3.

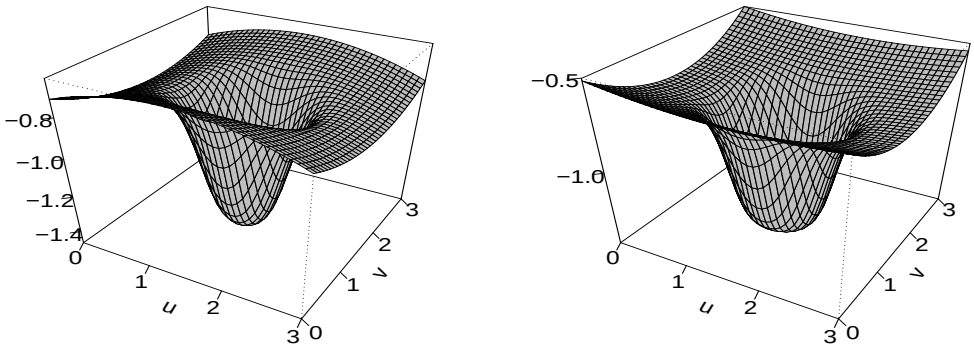

Figure 5.6: Morphology of a normal (left) and glaucoma (right) ONH.

The size of the cup, describing the ONH, is determined by a circle called contour line. Volumes and areas are extracted with respect to a reference plane. We define the reference plane as the mean height of the contour line minus 30 μm. Figure 5.7 visualises the position of contour line and reference plane in a two dimensional profile. More specifically, the HRT simulation calculates 31 values describing the ONH morphology: area global (*ag*) is the overall area of the papilla; area below reference global (*abrg*) is the area surrounded by the cut of the surface and the reference plane; effective area global (*eag*) is the area above the reference plane within the contour line; volume above reference global (*varg*) is the volume above

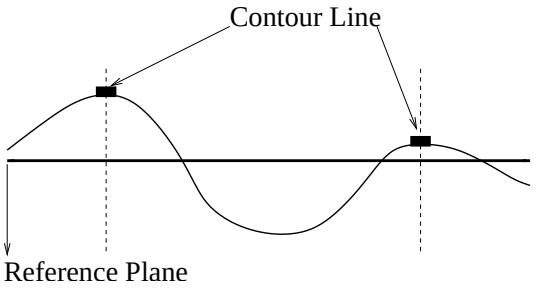

Figure 5.7: Two dimensional profile through the model image with contour line and reference plane.

the reference plane within the contour line; volume below reference global (*vbrg*) is the volume of the cup below the reference plane; mean height in contour global (*mhcg*) describes the depth of the cup and third moment global (*tmg*) the kurtosis of the cup. These global measurements are subdivided into segments of 90^o: temporal, nasal, inferior and superior. That means there are four additional results corresponding to each described morphological value. For example, the kurtosis is subdivided into: *tmt*, *tmn*, *tmi* and *tms*, which is third moment temporal, third moment nasal, etc..

We obtain variables describing the visual field defect ($\mathbf{w}_{cs}, \mathbf{w}_{clv}$) and the proportion of intact nerve fibres ($\mathbf{w}_{lora}$) by transformations of simulated HRT values $\mathbf{x}^{(HRT)}$ plus a normal measurement error with expectation zero and variance $\sigma^2_{\mathbf{w}_{cs}}, \sigma^2_{\mathbf{w}_{clv}}$ or $\sigma^2_{\mathbf{w}_{lora}}$, respectively. We simulate $\mathbf{w}_{cs}, \mathbf{w}_{clv}$ and $\mathbf{w}_{lora}$ by

$$\begin{aligned} \mathbf{w}_{lora} &= \alpha^{(lora)} + \beta^{(lora)^\top}\mathbf{x}^{(HRT)} + \varepsilon_{\mathbf{w}_{lora}}, \\ \mathbf{w}_{clv} &= \alpha^{(clv)} + \beta^{(clv)^\top}\mathbf{x}^{(HRT)} + \varepsilon_{\mathbf{w}_{clv}} \text{ and} \\ \mathbf{w}_{cs} &= \alpha^{(cs)} + \beta^{(cs)^\top}\mathbf{x}^{(HRT)} + \varepsilon_{\mathbf{w}_{cs}}. \end{aligned} \tag{5.1}$$

The parameter vectors $\beta^{(lora)}, \beta^{(clv)}$ and $\beta^{(cs)}$ and the intercepts $\alpha^{(lora)}, \alpha^{(clv)}$ and $\alpha^{(cs)}$ are extracted from the case-control study. Each element of the parameter vectors is set to zero except those corresponding to the HRT variables summarised

in table 5.4.

	$\alpha^{(lora)}$	$\alpha^{(clv)}$	$\alpha^{(cs)}$
Intercept	45.490	25.710	0.9440
	$\beta^{(lora)}$	$\beta^{(clv)}$	$\beta^{(cs)}$
abri	00.000	88.680	0.0000
abrn	00.000	00.000	-0.2840
abrs	37.980	170.63	0.0000
abrt	00.000	-132.83	0.0000
eag	00.000	-28.350	0.0000
eat	00.000	00.000	0.7962
vbrg	-40.54	00.000	0.0000
vbri	130.21	00.000	0.0000
mhcg	51.160	82.380	-1.4601
mhct	00.000	00.000	0.9726
tmi	00.000	00.000	-0.6972
tms	60.050	00.000	0.0000

Table 5.4: Parameters $\beta^{(lora)}$, $\beta^{(clv)}$ and $\beta^{(cs)}$ and intercepts $\alpha^{(lora)}$, $\alpha^{(clv)}$ and $\alpha^{(cs)}$ for the transformation of HRT variables to $\mathbf{w}_{cs}$, $\mathbf{w}_{clv}$ and $\mathbf{w}_{lora}$. The variables in the first column describe quantities derived from the HRT simulation.

Since the three intermediate variables have different measurement scales, we determine the variances $\sigma^2_{\mathbf{w}_{cs}}, \sigma^2_{\mathbf{w}_{clv}}$ and $\sigma^2_{\mathbf{w}_{lora}}$ of the measurement errors by the coefficient of variation τ, where $\tau = \frac{\sigma_{\mathbf{w}_{clv}}}{\eta_{\mathbf{w}_{clv}}} = \frac{\sigma_{\mathbf{w}_{lora}}}{\eta_{\mathbf{w}_{lora}}} = \frac{\sigma_{\mathbf{w}_{cs}}}{\eta_{\mathbf{w}_{cs}}}$. We extract $\eta_{\mathbf{w}_{cs}}, \eta_{\mathbf{w}_{clv}}$ and $\eta_{\mathbf{w}_{lora}}$ from the study population and simulate an increasing overlap of the distributions of the variables of the two classes by increasing the coefficient of variation τ.

The classifying function $g(\mathbf{w})$, where $\mathbf{w} = (\mathbf{w}_{clv}, \mathbf{w}_{cs}, \mathbf{w}_{lora})^\top$ in this simulation model is given by:

$$g(\mathbf{w}) = \begin{cases} glaucoma, & \text{if } \chi_{\mathbf{w}_{clv} \geq 163.15}(\mathbf{w}_{clv}) + \\ & \chi_{\mathbf{w}_{cs} < 0.2559}(\mathbf{w}_{cs}) + \chi_{\mathbf{w}_{lora} \geq 147.42}(\mathbf{w}_{lora}) \geq 1 \\ normal, & \text{else} \end{cases} \tag{5.2}$$

and $\chi(\cdot)$ is the indicator function. The function $g(\mathbf{w})$ is chosen with respect to an μ_1 error, i.e. a difference between true and observed diagnosis of about 10% for $\tau = \sigma^2_{HRT} = 0.05$. This function is incorporated into indirect classification approaches. $g(\mathbf{w})$ is based on visual field and papillometric values only and stands for the diagnosis of a physician. Hence, using this simulation model, the true state of the patient and the diagnosis are known. The true state is given by the simulation setup and the observed diagnosis is given by the classifying function $g(\mathbf{w})$.

5.2.2 Setups

We choose two simulation setups to compare error rates of direct and indirect classification methods. For each setup samples of 200 patients are generated with 100 observations classified to "glaucoma" and 100 observations classified to "normal", following the classifying function (5.2). This configuration of the samples mimics a realistic situation of a case-control study, where observations are collected following their observed class membership.

In the first setup we increase the variance of the simulated explanatory HRT variables. In other words, we examine the performance of direct and indirect classifiers with highly varying predictive values in the simulation model imitating the task of glaucoma classification. 4 levels of variation for the HRT simulation are used: $\sigma^2_{HRT} = 0.05, 0.20, 0.35$ and 0.50. The coefficient of variation for the intermediate variables is fixed at $\tau = 0.05$.

The second setup models the discrepancy between true state and observed diagnosis by altering the coefficient of variation τ. Hence, direct classifiers are trained on observed diagnosis diverging from the true underlying data structure and indirect classifiers model intermediate variables based on a highly varying learning sample. The coefficient of variation τ varies for $0, 0.02, 0.05, 0.08$ and 0.10, the variance of the measurement error in the HRT simulation is fixed at $\sigma^2_{HRT} = 0.20$. Fixed and varying parameters of both setups are summarised in table 5.5. We construct direct and indirect classifiers based on the learning sample

	fixed parameters	varying parameters
Setup 1	$\tau = 0.05$	$\sigma^2_{HRT} \in \{0.05, 0.20, 0.35, 0.50\}$
Setup 2	$\sigma^2_{HRT} = 0.20$	$\tau \in \{0, 0.02, 0.05, 0.08, 0.10\}$

Table 5.5: Parameter settings for setup 1 and 2.

and test them on the test sample. The size of the Monte-Carlo Study is 1000.

5.2.3 Results

Simulated misclassification errors for the diagnosis $g(\mathbf{w})$ with fixed variance of the intermediate variables and increasing variance of the explanatory HRT variables are summarised in table 5.6.

The discrepancy between observed diagnosis and true state of the patient increases with σ^2_{HRT}. Consequently, misclassification results of all applied classifiers are affected as well. In real-data examples one assesses the error between observed diagnosis and predicted class membership (μ_2). Direct classifiers perform comparably to indirect ones for small variances of the explanatory variables. At $\sigma^2_{HRT} = 0.2$ this relation shifts and indirect methods outperform direct ones as long as regression models are included which reflect, in this application, the true data structures. The indirect classifier LM^{ind} achieves especially good results for

	σ^2_{HRT}	0.05	0.2	0.35	0.5
μ_1		0.099	0.231	0.338	0.400
μ_2	LDA^{d}	0.099	0.216	0.236	0.230
	$CTREE^{\mathrm{d}}$	0.113	0.269	0.328	0.336
	$bagging^{\mathrm{d}}$	0.119	0.233	0.274	0.273
	LM^{ind}	0.109	0.199	0.203	0.181
	$RTREE^{\mathrm{ind}}$	0.152	0.273	0.317	0.315
	$bagging - RTREE^{\mathrm{ind}}$	0.098	0.223	0.280	0.289
	LM^{isb}	0.099	0.187	0.208	0.200
	$RTREE^{\mathrm{isb}}$	0.099	0.223	0.269	0.272
	$LM + RTREE^{\mathrm{isb}}$	0.099	0.188	0.208	0.199
	$LM + RTREE + PPR^{\mathrm{isb}}$	0.099	0.187	0.207	0.201
μ_3	LDA^{d}	0.001	0.131	0.299	0.379
	$CTREE^{\mathrm{d}}$	0.019	0.177	0.309	0.382
	$bagging^{\mathrm{d}}$	0.028	0.117	0.249	0.344
	LM^{ind}	0.014	0.191	0.326	0.393
	$RTREE^{\mathrm{ind}}$	0.068	0.198	0.311	0.368
	$bagging - RTREE^{\mathrm{ind}}$	0.003	0.126	0.268	0.350
	LM^{isb}	0.002	0.153	0.284	0.358
	$RTREE^{\mathrm{isb}}$	0.003	0.093	0.236	0.335
	$LM + RTREE^{\mathrm{isb}}$	0.002	0.151	0.284	0.357
	$LM + RTREE + PPR^{\mathrm{isb}}$	0.002	0.153	0.283	0.356

Table 5.6: Simulated error rates for several classifiers, $\tau = 0.05$ and increasing σ^2_{HRT}. We set $a = 0.5$ for indirect subagging classifiers.

large values of σ^2_{HRT}, indirect subagging performs comparably.
The μ_3 errors of the direct classifiers are smaller than the μ_3 errors of the indirect classifiers. Altogether, the μ_3 error rate of all classifiers seem to be influenced more by the classification technique (tree-based or linear) rather than by the classification framework (direct or indirect).
Simulated misclassification rates of the second scenario with increasing variance of the intermediate variables, are summarised in table 5.7.

Most applied indirect classifiers achieve smaller misclassification rates for small values of τ, considering the μ_2 error rate. Indirect classification trees outperform direct classification trees, and bagging reduces misclassification rates further. For larger values of τ the situation changes and direct methods are comparable to indirect classification incorporating the fixed classifying function. Indirect subagging achieves an estimated misclassification error of about $3-5\%$ less than the errors of all other classifiers in a situation with $\tau = 0.08$ and $\tau = 0.1$, respectively. The regression model incorporated in the indirect classification framework determines the performance of the classifier. LM^{ind} achieves the smallest simulated misclassification results among all applied classifiers and values of $\tau \in \{0, 0.02, 0.05\}$, indirect subagging performs well as long as linear models are included in this setup. For $\tau \geq 0.05$ the indirect subagging classifiers, which incorporate linear models gain the best results among all considered classification techniques. For $\tau = 0.1$ these misclassification errors are about 24.6%, i.e. about 4% smaller than the best direct classifier LDA^{d}. If the linear regression model is not included in the indirect subagging approach, these classifiers achieve misclassification errors which are about $3-5\%$ larger than the results of the indirect subagging classifiers which include linear regression models. This result is plausible, since the simulation model includes linear transformations to model the intermediate variables, see section (5.1). Considering the error between the true state of the patient and the classification results, direct methods are superior

	τ	0	0.02	0.05	0.08	0.1
μ_1		0.158	0.177	0.231	0.268	0.285
μ_2	LDA^{d}	0.110	0.140	0.216	0.269	0.291
	$CTREE^{\mathrm{d}}$	0.169	0.196	0.269	0.321	0.343
	$bagging^{\mathrm{d}}$	0.136	0.163	0.234	0.279	0.301
	LM^{ind}	0.000	0.098	0.199	0.260	0.290
	$RTREE^{\mathrm{ind}}$	0.151	0.182	0.273	0.326	0.352
	$bagging - RTREE^{\mathrm{ind}}$	0.130	0.157	0.223	0.269	0.295
	LM^{isb}	0.090	0.120	0.187	0.232	0.257
	$RTREE^{\mathrm{isb}}$	0.141	0.164	0.223	0.265	0.285
	$LM + RTREE^{\mathrm{isb}}$	0.090	0.120	0.188	0.233	0.257
	$LM + RTREE + PPR^{\mathrm{isb}}$	0.090	0.119	0.187	0.232	0.256
μ_3	LDA^{d}	0.088	0.098	0.131	0.155	0.166
	$CTREE^{\mathrm{d}}$	0.1440	0.147	0.177	0.214	0.230
	$bagging^{\mathrm{d}}$	0.099	0.098	0.116	0.135	0.146
	LM^{ind}	0.158	0.177	0.192	0.207	0.221
	$RTREE^{\mathrm{ind}}$	0.141	0.152	0.198	0.238	0.258
	$bagging - RTREE^{\mathrm{ind}}$	0.123	0.123	0.128	0.140	0.153
	LM^{isb}	0.157	0.159	0.152	0.150	0.149
	$RTREE^{\mathrm{isb}}$	0.097	0.096	0.093	0.103	0.109
	$LM + RTREE^{\mathrm{isb}}$	0.156	0.159	0.151	0.149	0.147
	$LM + RTREE + PPR^{\mathrm{isb}}$	0.156	0.159	0.153	0.150	0.148

Table 5.7: Simulated error rates for several classifiers, $\sigma^2_{HRT} = 0.2$ and increasing τ. We set $a = 0.5$ for indirect subagging classifiers.

to indirect ones in situations with a small measurement error of the intermediate variables ($\tau \in \{0, 0.02\}$). For large variance ($\tau = 0.1$) indirect subagging outperforms all other classification techniques in terms to the μ_3 error rate. However, again the tree-based methods are superior to the linear ones, perhaps because of possible subgroups generated by the distinction between a true and an observed diagnosis.

Figure 5.8 shows the distribution of simulated error rates from setup 1 for selected classifiers and variances $\sigma^2_{HRT} = \{0.05, 0.5\}$. The difference between direct and indirect classifiers is small for a small variance of the explanatory variables. The observed misclassification error is larger than the true misclassification error for all classifiers here and it varies more. Considering a situation with a large variance of the explanatory variables ($\sigma^2_{HRT} = 0.5$), the μ_2 error is smaller than the μ_3 error for all classifiers. The variance of the true misclassification error is increased compared to the situation with $\sigma^2_{HRT} = 0.05$. Indirect classifiers outperform direct ones here, as long as linear models are used to predict the intermediate variables.

Figure 5.9 shows the distribution of simulated error rates from setup 2 for selected classifiers and coefficient of variation $\tau = \{0, 0.1\}$. Most indirect classifiers outperform direct classifiers in a situation with a small variance of the measurement error of the intermediate variables $\tau = 0$. The observed misclassification error is larger than the true misclassification error for direct classifiers and smaller than the true misclassification error for most indirect classifiers. The indirect classifier LM^{ind} mimics the true data structure here and achieves optimal results.
Observed misclassification errors are larger than true misclassification errors for direct and indirect classifiers in a situation with $\tau = 0.1$. Variances of the μ_2 and μ_3 errors are in creased compared to the situation with $\tau = 0$. The difference between the performance of direct and indirect classifiers is small in this situation.
In both setups classifiers including linear modelling techniques gain better results than tree-based methods, although bagging improves misclassification errors of

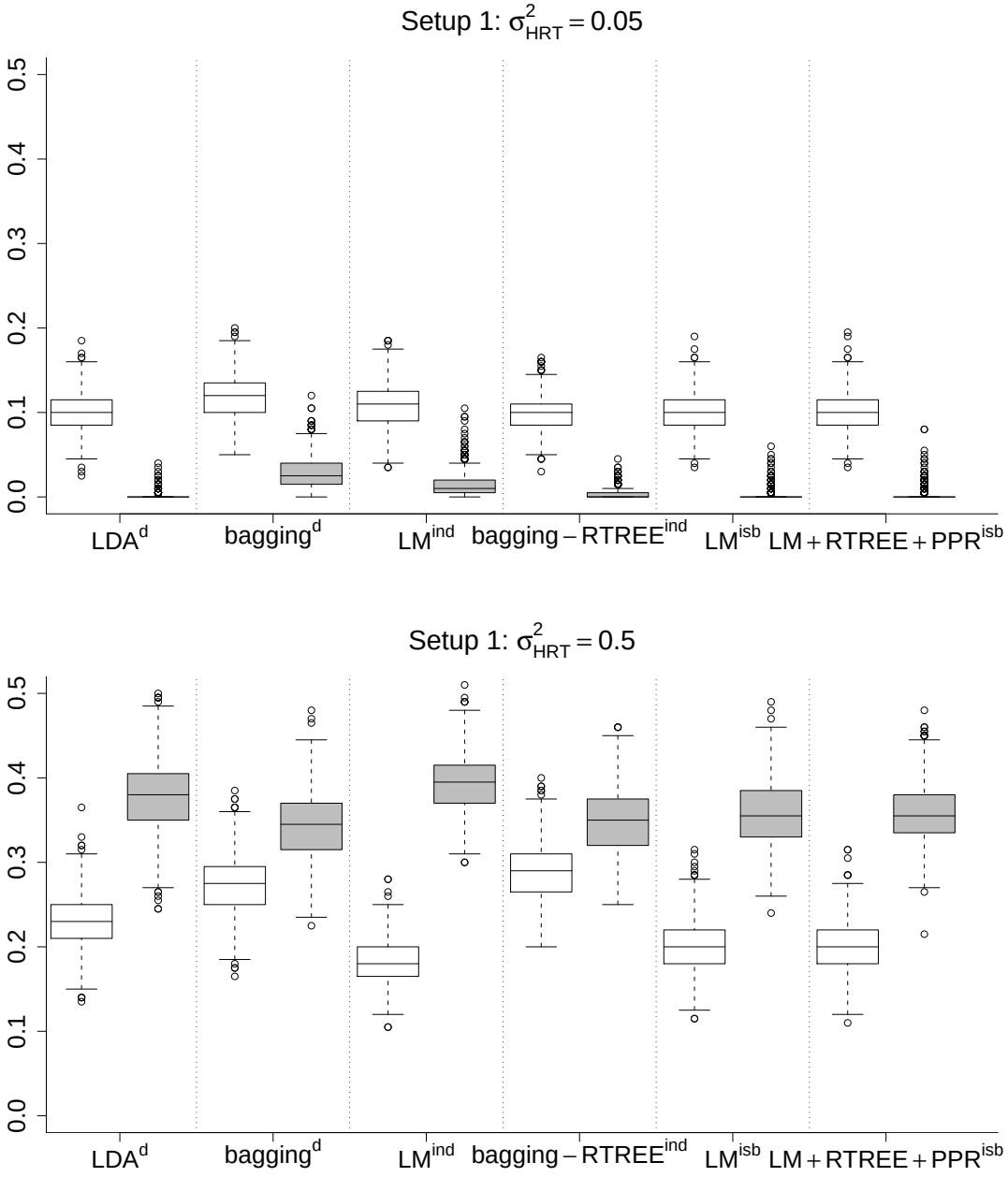

Figure 5.8: Distribution of estimated μ_2 (white boxes, left) and μ_3 (grey boxes, right) error rates of selected classifiers.

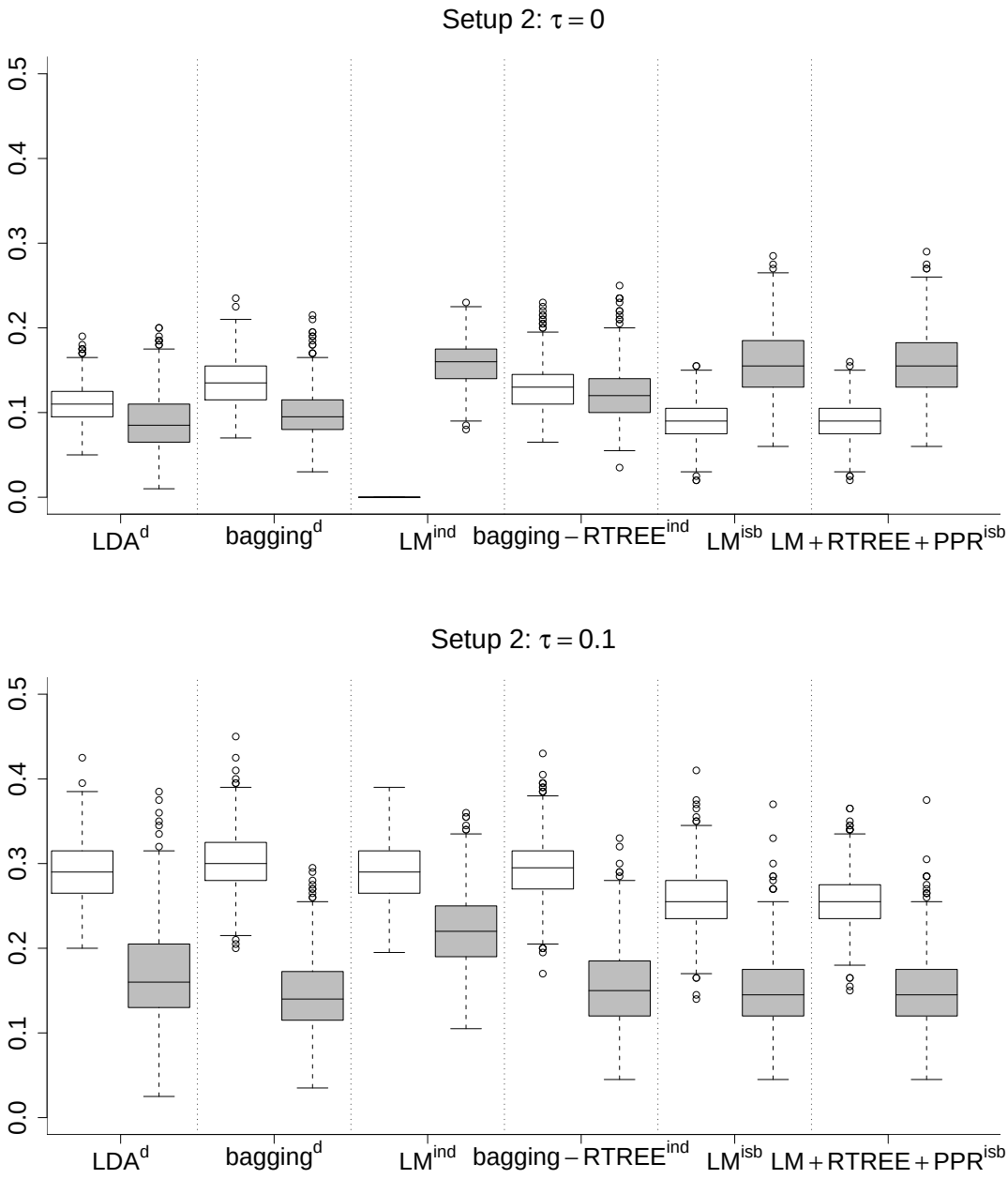

Figure 5.9: Distribution of estimated μ_2 (white boxes, left) and μ_3 (grey boxes, right) error rates of selected classifiers.

regression and classification trees. An indirect classifier includes the diagnostic function that is the discrepancy between true state and observed diagnosis. Therefore, the performance of an indirect classification rule depends on the performance of the diagnostic function as well as the measurement errors incorporated in the given learning and test samples. Nevertheless, the indirect method outperforms the direct method for situations with small variances of the measurement error of the intermediate variables and large variances of the measurement error of the explanatory variables.

Chapter 6

Applications

The evaluation of classification techniques in terms of their application to real data sets is an established tool. We apply direct and indirect classifiers to different data sets to avoid the criticism that the artificial data favours a particular methods. On the one hand we consider a situation where the true classifying function is known. We use the diagnosis of glaucoma described in chapter 2.1 to formulate an indirect classification rule using this a priori knowledge. On the other hand we calculate direct and indirect subagging classifiers for two data sets where no fixed classifying function is given. For indirect subagging classifiers we vary the proportion of the subag sample between 25, 50 and 75% of the total datasets.

6.1 Glaucoma Classification

We apply direct and indirect classification techniques to the cross sectional study described in chapter 2.1. We assume that for future examinations only HRT data are available. Hence, only the 62 HRT variables are used as explanatory variables $\mathbf{x}$. The HRT variables include several measurements of the volume and areas of certain sections of the papilla.

We calculate direct and indirect classifiers using the fixed classifying function as well as indirect subagging introduced in chapter 3.3. Bagging and subagging are performed using $B = 50$ random samples. 50 10-fold-cross-validations (CV) are calculated on the given set of observations to compute an estimate of misclassification error. The results of the indirect and direct classification approaches are listed in Table 6.1.

Classifier	Error Rate Estimation		
LDA^{d}	0.219		
$CTREE^{\mathrm{d}}$	0.273		
$bagging^{\mathrm{d}}$	0.191		
LM^{ind}	0.243		
$RTREE^{\mathrm{ind}}$	0.213		
$bagging - RTREE^{\mathrm{ind}}$	0.179		
$a =$	0.25	0.5	0.75
LM^{isb}	0.174	0.188	0.209
$RTREE^{\mathrm{isb}}$	0.180	0.189	0.207
$LM + RTREE^{\mathrm{isb}}$	0.178	0.188	0.211
$LM + RTREE + PPR^{\mathrm{isb}}$	<u>0.165</u>	0.179	0.202

Table 6.1: Error rate estimation via 10 fold-cross validation for different direct and indirect classifiers applied onto the Glaucoma Data.

The CV error is 24.25% for the indirect approach using LM^{ind}. The error is smaller if a regression tree is used to predict the intermediates, namely 21.3%. $bagging - RTREE^{\mathrm{ind}}$ results in a misclassification error estimate of 17.9%. Direct classification by LDA^{d} achieves an error estimation of 21.9%. Training a $CTREE^{\mathrm{d}}$ on the given data set of explanatory variables, the misclassification error estimate is 27.30%. The error can be reduced onto 19.1% by bagging the tree. Compar-

ing bagged direct and indirect classifiers, the reduction in estimated misclassification error from 19.1% to 17.9% indicates the gain achieved by using a priori knowledge. Furthermore indirect subagging performs comparably to the indirect approach using a fixed classifying function. The proportion of out-of-subag and subag samples influences the performance of these classifiers. In situations where predictive models for the intermediate variables are calculated on 75% of the data, whereas the classification rule is calculated on 25%, indirect subagging performs best. It achieves an estimated misclassification rate of 16.5% if we include all considered regression techniques ($LM + RTREE + PPR^{\text{isb}}$).

6.2 Datasets with Unknown Classifying Function

We consider two datasets where the a priori knowledge defines explanatory and intermediate variables only. We apply direct classifiers and indirect subagging to these datasets.

Dystrophy Data Duchenne Muscular Dystrophy (DMD) is a genetically transmitted disease, passed from a mother to her children. Affected female offspring usually suffer no apparent symptoms, male offspring with the disease die at young age. Although female carriers have no physical symptoms they tend to exhibit elevated levels of certain serum enzymes or proteins.

The dystrophy dataset contains 209 observations of 75 female DMD carriers and 134 female DMD non-carrier. The data is given by Andrews and Herzberg (1985). It includes 6 variables describing age of the female and the serum parameters serum marker creatine kinase (CK), serum marker hemopexin (H), serum marker pyruvate kinase (PK) and serum marker lactate dehydroginase (LD).
The serum markers CK and H may be measured rather inexpensively from frozen serum, while PK and LD require fresh serum. Therefore, we assign age, CK and

H to explanatory and LD and PK to intermediate variables.

	0.25	0.5	0.75
LDA^{d}		0.155	
$CTREE^{\mathrm{d}}$		0.157	
$bagging^{\mathrm{d}}$		<u>0.118</u>	
LM^{isb}	0.124	0.123	<u>0.117</u>
$RTREE^{\mathrm{isb}}$	0.130	0.139	0.143
$LM + RTREE^{\mathrm{isb}}$	0.122	0.122	<u>0.119</u>
$LM + RTREE + PPR^{\mathrm{isb}}$	<u>0.120</u>	<u>0.120</u>	<u>0.116</u>

Table 6.2: Results of dystrophy dataset.

The classifying function is not known for this dataset, hence only the new algorithm can be applied as indirect classifiers. Direct classifiers are based on the explanatory variables CK and H only. Tibshirani and Hinton (1998) analysed that the usage of the additional intermediate variables LD and PK does not improve misclassification results very much. Consequently, the new approach is comparable to the direct classifiers. Error rates are approximately the same for all considered proportions of the subag sample.

Pima Indian Diabetes The diabetes dataset was collected by the National Institute of Diabetes and Digestive and Kidney Diseases and contains 768 observations from females of Pima Indian heritage, including 500 healthy subjects and 268 diabetics. It contains 8 variables describing the number of times pregnant, plasma glucose concentration (glucose tolerance test), diastolic blood pressure (mm Hg), triceps skin fold thickness (mm), 2-hour serum insulin (mu U/ml), body mass index (weight in kg/(height in m)^2), diabetes pedigree function and age (years), see Smith et al. (1988).

We assign body mass index, diabetes pedigree function and age to explanatory

variables since they are easy to collect and do not need serum probes, the remaining variables are assigned to intermediate variables.

	0.25	0.5	0.75
LDA^{d}		0.310	
$CTREE^{\mathrm{d}}$		0.325	
$bagging^{\mathrm{d}}$		0.317	
LM^{isb}	0.295	0.298	0.298
$RTREE^{\mathrm{isb}}$	0.300	0.305	0.315
$LM + RTREE^{\mathrm{isb}}$	0.295	0.297	0.300
$LM + RTREE + PPR^{\mathrm{isb}}$	0.296	0.297	0.299

Table 6.3: Results of diabetes dataset with explanatory variables body mass index, pedigree function and age.

Again no fixed classifying function is available, hence we apply direct classifiers to the three explanatory variables and indirect subagging to the full information of the learning sample. The results indicate the gain of indirect classification. The inclusion of intermediate variables lowers the misclassification rate of direct vs. indirect approach by about 2%. We achieve smallest estimated misclassification errors in situations with a large out-of-subag sample.

Chapter 7

Implementation

The software to calculate indirect classifiers and error rate estimations, used in the current thesis were implemented in the R package **ipred**. In the following we will introduce the programming environment R and the package **ipred** and demonstrate its functionality with the example of direct and indirect glaucoma classification.

7.1 R - A Statistical Programming Environment

The R environment is a coherent system which provides a programming language, high level graphics, interfaces to other languages and debugging facilities. R is available as free software under http://www.r-project.org.
The programming language is a dialect of the S language which was developed at Bell Laboratories. The principle designer of the S language was John Chambers. Since 1997, an R core group has existed which focuses on the maintenance and the development of efficient and scientifically valuable software, especially in areas of statistics. The object oriented programming and the opportunity to define objects associated with attributes enables us to provide user friendly interfaces. We avoid the re-specification of arguments within different working steps

by detaching these objects to corresponding methods via so called generic functions. Many modern statistical techniques have been implemented in R, mainly supplied as packages. R packages are available under http://cran.r-project.org.

7.2 ipred: Improved Predictors and Error Rate Estimators

The **ipred** package is an attempt to create a unified interface for improved predictors and error rate estimators, Peters et al. (2002b).
In the current thesis we discussed improved predictions of class memberships by bootstrap aggregation of classification trees and by indirect classification approaches. The function `bagging` implements e.g. bootstrap aggregation of classification trees in a direct classification framework. It also provides an interface for the new classification technique "bundling", which combines an arbitrary number of classifiers with bagging (Hothorn and Lausen, 2003b). The approaches of indirect classifiers with known and unknown classifying function, discussed in this thesis are implemented in the functions `inclass` and `inbagg`. Furthermore, different re-sampling based error rate estimators can be calculated with `errorest`, particularly it includes cross-validation, a bootstrap estimator and its bias corrected version .632+ (Efron and Tibshirani, 1997).
The package also contains different datasets, e.g. the glaucoma and dystrophy datasets, discussed in chapter 2.1 and section 6.2, respectively. In the following we demonstrate the performance of the functions by their application to the glaucoma data set.

The functions inbagg and inclass Approaches of indirect classification are implemented in the functions `inclass(...)` and `inbagg(...)`.
`inclass(...)` performs indirect classification as described in Hand et al. (2001),

and `inbagg(...)` implements indirect classification combined with bagging as well as the new approach of indirect subagging described in chapter 3.3. To provide a user friendly interface with as few necessary specifications as possible, we define generic functions which determine e.g. print methods.
In the following we demonstrate how to fit the indirect classifier for glaucoma classification. We want to use linear models to predict the intermediate variables (LM^{ind}).

```
R> # load glaucoma dataset
R> data(GlaucomaMVF)
R> fit.ind <- inclass(formula = Class~clv+lora+cs~., data = GlaucomaMVF,
+       pFUN = list(list(model = lm)), cFUN = classify)
```

We have four non-optional arguments to fit this indirect classifier. The `formula` consists of three parts:
`formula = response~intermediate~explanatory`. The left hand side assigns the response variable, the middle part defines the intermediate variables and the right hand side the explanatory variables. With `data = GlaucomaMVF` we assign the dataset in use. The argument `cFUN` is either a fixed classifying function which assigns the intermediate (and explanatory) variables to a certain class membership or a list with elements `formula, model, predict` and `training.set`, specifying how to estimate the relationship between intermediate and response variables. We give a more specific description of the list `cFUN` below, in the context of the functionality of the function `inbagg(...)`. For our application a fixed classifying function is given and displayed in figure 2.1. We combine this fixed function with indirect classification by setting `cFUN = classify`.
The argument `pFUN` is a list of lists with elements specifying how to model the functional relationship between intermediate and explanatory variables. We applied indirect classifiers which predict each intermediate variable with the same regression model, e.g. the linear regression model for the classifier LM^{ind}. `inclass(...)` provides the opportunity to use different models to predict intermediate variables. For example, if we want to model the variables $\mathbf{w}_{lora}$ and $\mathbf{w}_{cs}$ by regression trees based on all explanatory variables and $\mathbf{w}_{clv}$ by a linear model based on some explanatory variables only, we choose

```
R> pFUN <- list(list(formula = cs+lora~., model = rpart),
+                    list(clv~ag+abrg+varg+vbrg, model = lm))
```

and assign `pFUN` to the function `inclass(...)`. The generic determined print method of the fitted object `fit.ind` of class `inclass` gives information about the intermediate variables and the predictive model:

```
R> fit.ind


 Indirect classification, with 3 intermediate variables:
 clv lora cs


 Predictive model per intermediate is lm
```

Furthermore, the fitted object contains among other information the fitted regression models corresponding to each intermediate variable and the classifying function. These values can be extracted easily by calling

```
R> fit.ind$model.intermediate
$clv


Call:
formula.list[[i]]$model(formula = formula, data = data)


Coefficients:
(Intercept)           ag           at           as           an           ai
   123.6510   -1197.0967    1549.9025    1220.8971    1140.7330    1131.0894
        eag          eat          eas          ean          eai         abrg
  4296.2945   -4494.1603   -4353.4151   -4281.0329   -4169.7150     130.4761

                                    ⋮
```

A detailed documentation is given in the appendix.

The usage of the function `inbagg(...)` is straightforward. Beneath necessary arguments to specify predictive models for intermediate and response variables, one has to provide information about how the subsampling procedure shall be performed. Hence, we fit an indirect subagging classifier with 25 subsamples and an out-of-subag of 50% using linear regression models (LM^{ind}) by

```
R> fit.isb <- inbagg(formula = Class~clv+lora+cs~., data = GlaucomaMVF,
+     pFUN = list(list(model = lm)),
+     cFUN = list(formula = Class~.),
+     nbagg = 25, ns = 0.5, replace = FALSE)
```

The arguments `formula, data, pFUN` and `cFUN` are used similarly as described for `inclass(...)` above. `pFUN` is again a list of lists where each element specifies a regression model for the intermediate variables. Note that indirect subagging enables us to combine an arbitrary number of regression models, whereas the number of models is limited by the number of intermediate variables in the traditional indirect classification approach using a fixed classifying function. A list assigned to `cFUN` specifies how to calculate the classification rule. This list has the arguments `formula`, specifying which variables are to be used to predict the response variables, `model` specifying the classifier, `training.set` indicating whether the classifier shall be trained based on the subag sample and the argument `predict` specifying a predictive function of the regression model if necessary. The arguments `nbagg, ns` and `replace` determine how many subag samples are to be drawn, the size of each subag sample and whether to draw with or without replacement.
The default of the argument `cFUN` is set to a classification tree, trained on the subag sample. Hence, we only add the information that the tree shall be trained on explanatory and intermediate variables by `cFUN = list(formula= Class~.)`. The print output looks as follows

```
R> fit.isb
 Indirect bagging, with 25 bootstrap samples and intermediate variables:
 clv lora cs
```

We get information about the number of subag samples and the selected intermediate variables.

The function errorest We use $10-$fold cross-validation to estimate error rates in different applications. However, **ipred** provides an interface for different resampling based estimators. Our aim was an interface which requires as few user specifications as possible and which covers a wide range of error rate estimators

for any classifier. Therefore, we defined generic functions, which determine the appropriate method.
The user can specify different error rate estimations of any classifier implemented in R using `errorest` with the basic arguments `formula, data, model, predict` and `estimator`. We perform a $10-$fold cross validation of the linear discriminant analysis by calling:

```
R> # exclude intermediate variables
R> study.group <- GlaucomaMVF[ ,
+                   !(names(GlaucomaMVF) %in% c(``clv'', ``cs'', ``lora''))]
R> # define predict function
R> mypredict.lda <- function(object, newdata) predict(object, newdata)$class
R> error.lda <- errorest(Class~., data = study.group,
+                   model = lda, predict = mypredict.lda)
```

Note that the glaucoma dataset includes explanatory, intermediate and response variables. The formula based interface enables us to define response and explanatory variables easily. We reduce the input dataset `data = study.group` to explanatory and response variables only, since we use the short-cut "`formula = Class~.`". The argument `model = lda` determines the classification techniques, and the method for predicting a fixed classification rule is specified by `mypredict.lda`.
Due to a unified interface the definition of the wrapper function `mypredict.lda` is necessary, since `predict.lda` does not return the predicted class membership by default (as required). A description of all optional arguments of `errorest(...)` is given in the appendix. For improved classifiers implemented in **ipred**, the standard output is adapted. For example the error rate for an bootstrap aggregated classifier is estimated by:

```
R> error.bagging <- errorest(Class~., data = study.group, model = bagging)
```

For indirect classifiers `inclass(...)` and `inbagg(...)` the input arguments are more complex, since we have to distinguish between explanatory, intermediate and response variables. We specify a crossvalidation of the classifier $bagging - RTREE^{\text{ind}}$ defined in chapter 3.3 by

```
R> error.ind <- errorest(Class~clv+lora+cs~., data = GlaucomaMVF,
+     model = inbagg, cFUN = classify, pFUN = list(model = rpart,
+     training.set = ``bag''), replace = TRUE, ns = 1)
```

where `cFUN = classify` is the fixed classifying function of figure 3.2 and `pFUN` specifies how to calculate predictive models for the intermediate variables. The argument `replace = TRUE` is logical and indicates whether we want to perform bagging or subagging, `ns = 1` specifies the proportion of observations to be drawn. The print output of the calculated error rate of linear discriminant analysis is:

```
R> error.lda


Call:
errorest.data.frame(formula = Class ~ ., data = study.group,
    model = lda, predict = mypredict.lda)

         10-fold cross-validation estimator of misclassification error


Misclassification error:  0.2229
```

Hence it returns the call, information about the chosen error rate estimator and the estimation itself.

Chapter 8

Discussion and Outlook

Automated classification techniques are often useful in medical applications. An automated rule can be applied by any even unexperienced observer to pre-select subjects into healthy or diseased, e.g. during medical screening programs.
We concentrate on situations where these rules are based on a reduced set of variables, although more variables are available in the study population. This corresponds to situations in the medical field, where the number of examinations is reduced, to avoid invasive procedures, spare patients' time, or reduce medical costs.
Furthermore, a priori knowledge about a disease is often available. In the example of glaucoma diagnosis the a priori knowledge is the definition of glaucoma, in many other applications it includes the separation between medical tests performed only for the current study population (learning sample) and those performed for future patients (test sample).
An exact definition of a disease given by a physician leads to the distinction between two types of diagnosis: (i) the diagnosis given by the observer and (ii) the diagnosis reflecting the true state of a patient. In epidemiology the difference between the observed and true state of the patient is known as differential misclassification (Song and Ahn, 2002; Bratcher and Stamey, 2002).

The indirect classification approach (Hand et al., 2001) is a framework that combines medical and statistical knowledge. A learning sample is subdivided into variables available for future patients (explanatory variables) and variables defining a disease and not observed in the future (intermediate variables). Afterwards a classification rule is constructed with respect to medical a priori information. The advantage of this procedure is that the classification rule is based on a reduced set of necessary diagnostic tests, while incorporating the medical a priori information from the full set of measurements.
We define the framework of indirect classification more generally as suggested by Hand et al. (2001) and distinguish between indirect and direct classification techniques. The basic difference between these techniques is that direct classifiers require the same variables available in learning and test samples, whereas indirect classifiers replace missing future examinations incorporating a priori knowledge.

However, a statistical difficulty is the choice of an analysis technique, which is able to model the given data adequately. The incorporation of classification trees in a medical context is widely discussed (Zhang and Singer, 1999; Lausen et al., 1994; Ciampi, 1991). Combining classifiers with bagging reduces misclassification errors in both the direct and indirect approaches. We define bagging for indirect classification in situations where the a priori knowledge includes the definition of a disease.

As described above, the a priori knowledge often includes the distinction between recent and future examinations only. We propose an algorithm which makes use of all variables available in the learning sample and classifies future observations based only on a reduced set of variables. The procedure "indirect subagging" enables us to calculate an arbitrary number of regression models to predict variables missing for future observations and incorporates them into the classification technique subagging. In the literature different approaches have been proposed, which combine different classification techniques or im-

prove them with additional variables. Mojirsheibani (2001) discuss an iterated classification rule, based on additional pseudo-predictors. Other approaches for the combination of different classification techniques are suggested by LeBlanc and Tibshirani (1996) and Mojirsheibani (1999). Our algorithmic proposal "indirect subagging" has several advantages:

- it avoids the model selection problem i.e. the choice of an adequate regression model to predict the intermediate variables;
- it does an automated rating of variables for the classification task since it is a tree based method; and
- it prevents over-fitting since it calculates predictive models for additional variables on the so called "out-of-subag" sample that is independent from the subag sample that the classification rule is based on.

The out-of-subag sample includes the observations not included in the subag sample, for bagging these subsamples are called out-of-bag and bag, respectively. The idea of using these independent data sets for the improvement of classifiers is discussed in literature several times. Hothorn (2003) proposes "bundling" which combines several classifiers with bagging, Rao and Tibshirani (1997) use the out-of-bootstrap to calculate weights for model averaging and Breiman (1996b) use the out-of-bag to form estimates, for example, for node probabilities or node errors in decision trees.
In contrast to these improvements of direct classification techniques, indirect subagging offers the opportunity to make use of the out-of-subag sample within an indirect classification framework. This method increases the discriminant value of the set of predictors, which are available for the final classifier, rather than making use of the discriminant value of the out-of-subag.

We investigate the performance of direct and indirect classifiers in situations with known distribution of learning and test samples. Furthermore, we distin-

guish between different types of classification errors: the difference between true and observed state of a patient (differential misclassification), misclassification with respect to the observed state (observed misclassification) and a misclassification with respect to the true state (true misclassification). We derive explicit formulae for these errors of the indirect classifier using a fixed classifying function and the Bayes classifier in an artificial example with normal distributed variables.

Results indicate that indirect classifiers outperform direct ones for small measurement errors of intermediate variables and large measurement errors of the explanatory variables with respect to the observed misclassification. The true misclassification errors of all considered classifiers increase simultaneously with increasing variance of the intermediate variables. They perform differently if we increase the variance of the intermediate variables only.

Error rates of true and observed misclassification of all considered classifiers diverge for an increasing differential misclassification error. This result confirms theoretical results of Hand et al. (1998). They analysed that error rates of optimal classification rules dealing with the true and with the observed class membership, respectively, are very different in situations where the prior probabilities corresponding to true and observed class membership are markedly different. In our framework differential misclassification reflects the difference between these prior probabilities. Hand (2001) gives examples showing different interpretations of diagnostic performance.

Considering asymptotic behaviours of indirect classifiers with respect to the observed diagnosis, we demonstrate Bayes consistency of a specified indirect classifier using a fixed classifying function under constrictive model assumptions concerning the distribution of learning and test samples. Indirect subagging is Bayes consistent with respect to the observed diagnosis under more general model assumptions.

Furthermore, we investigate the performance of direct and indirect classification techniques in different simulation models. We generate different structures of the decision space using normal distributed explanatory and intermediate variables with fixed variances.
Considering the observed misclassification error, the indirect classifier using the fixed classifying function performs best, as long as all parameters within this approach are correctly specified. Indirect subagging outperforms direct classifiers using a fixed classifying function and does not require as detailed specifications. It achieves results comparable to direct methods in situations where additional variables do not influence the response variable, i.e. the diagnosis of a patient, and it outperforms direct classifiers in situations where the intermediate variables contribute some diagnostic value.
The true misclassification error is small for the direct classifier linear discriminant analysis in a decision space following an easy cut-point model. Indirect classification achieves the best results for a tree based decision space.
A modification of the simulation model of the optic nerve head by Hothorn and Lausen (2003a) and the subsequent Monte-Carlo study indicate that an indirect classifier is an appropriate tool for glaucoma classification in situations with large variance of explanatory variables and small measurement errors of the intermediate variables.

The framework of indirect classification has been applied to the problem of glaucoma classification. This is an example where the given a priori knowledge covers a simple classifying function, which can be incorporated into a classification task. The set of variables from different examination tools has been structured into explanatory, intermediate and response variables. The division of variables has been performed, considering the important aspect of glaucoma that patients do usually not detect the onset of the disease. However, early detection is of main importance, since adequate therapy can slow down the progression of

the disease. It is known that damage in the optic nerve head precede visual field defects of the patients (Eid et al., 1997). A good classification rule should be based on measurements of the optic nerve head, which are able to detect early damage in the retinal nerve fibre layer. The HRT is an appropriate tool for detecting early damage Kamal et al. (1999); Iester et al. (1997), hence, the ideal explanatory variables are HRT variables. Moreover, the definition of the disease is based on the optic nerve head morphology and the visual field defects of the patient. The three intermediate variables employed in the procedure also belong to these two areas. Estimated error rates indicate that indirect approaches outperform direct ones in this application. Indirect subagging performs comparably well to indirect classification rules incorporating a fixed classifying function, but requires less specifications. Furthermore, we analysed datasets which embed less a priori information, in particular, we do not know a fixed classifying function. The first dataset include information of a group of women that are carriers and non-carriers of Duchenne Muscular Dystrophy (Dystrophy Data) and the second dataset is a study population of female Pima Indians with healthy and diabetes affected subjects (Diabetes data). The subdivision of the available variables in the study population into explanatory and intermediate variables was with respect to the cost of the medical examinations in the case of the Dystrophy data and with respect to the effort of the collection of the data in the case of the Diabetes data. Since no classifying function is known, which determines the class membership, we compare indirect subagging with direct approaches. Misclassification errors are reduced by indirect subagging in the Diabetes dataset and comparable for the Dystrophy Data. Application to data sets indicates that the percentage of observations used to calculate the regression models and the classification rule influences the performance of the indirect subagging classifier.

All in all, the application of the indirect approaches demonstrate the fruitful synergy of medical knowledge acquisition and statistical classification methods.

Bagging the classification process leads to a further reduction of the estimated error rates in the considered applications. The advantages of the indirect approach are of interest for various areas of medical decision making challenged by the fact that patients may be investigated with several tools. A natural first aim is to reveal the relationship between different examinations in order to decide whether some medical examinations can be disregarded. The division of variables used in the indirect classification is an approach similar to the framework of graphical modelling (Cox and Wermuth, 1996), which can be seen as an example for an unknown relationship between intermediate and response. Extracted a priori information from this procedure of structuring offers the opportunity to build an indirect classification rule based on a reduced set of examinations. The indirect classifier can be used to provide statistical insight in knowledge based decision support (Wetter, 2002). Another approach would be using clustering techniques in the predictive step. Torgo and da Costa (2000) proposed a method which integrates a clustering technique with regression trees. Tibshirani and Hinton (1998) make use of the structuring information of the group of intermediate variables. In contrast, an indirect classifier provides a flexible framework to incorporate medical knowledge. The extension to indirect subagging is an opportunity to do indirect classification in situations where no fixed classifying function is given.
A distinction between error rates corresponding to the diagnosis given by an observer and those describing a misclassification of the true diagnostic state of the patient is often not considered in classification tasks. Theoretical and simulation results indicate that this distinction can lead to different results about the performance of classification techniques.

Altogether, results of theoretical considerations, simulation and application suggest that a fixed classifying function should be included in an indirect classification method, whenever it is defined correctly, see sections 5.2 and 6.1. The gain is an understandable decision rule, which combines medical a priori knowledge

and statistical classification techniques and uses the full information of a given learning sample.
In contrast to that, in situations where additional and informative intermediate variables are given, but a fixed classifying function is not known or doubtful, one should use indirect subagging. On the one hand this technique again uses the full information of the learning sample and should therefore outperform direct classifiers, which neglect these additional informations (section 5.1). On the other hand, the interpretability of indirect subagging is as difficult as it is for any other direct bootstrap aggregated classification technique. Furthermore, the discriminant value of the set of intermediate variables determines the performance of indirect classification. In situations, where they are not informative, one should use direct classifiers, compare section 6.2.

However, several difficulties remain unsolved. It is not known, how to optimise indirect subagging with respect to the percentage of out-of-subag and subag samples. Moreover, since we know that good regression models do not necessarily result in a good classification rule, boosting approaches could be developed which would optimise the new procedure in terms of the misclassification error. Even the linkage of indirect classification using a fixed classifying function with recent developments in boosting algorithms may lead to further improvements. Chawla et al. (2002) proposes a technique to extend bagging and boosting for massive datasets. We assess the performance of classifiers with respect to their error rates. Investigating sensitivity and specificity rather than the total error rate for direct and indirect classification could simultaneously lead to different interpretations of their performance.

Appendix A

Implementation

The documentation of the functions `inclass(...)` and `inbagg(...)` is given in the following.

`inbagg` *Indirect Bagging*

Description

Function to perform the indirect bagging and subagging.

Usage

```
inbagg.data.frame(formula, data, pFUN=NULL,
  cFUN=list(model = NULL, predict = NULL, training.set = NULL),
  nbagg = 25, ns = 0.5, replace = FALSE, ...)
```

Arguments

`formula`	formula. A `formula` specified as `y~w1+w2+w3~x1+x2+x3` describes how to model the intermediate variables `w1, w2, w3` and the response variable `y`, if no other formula is specified by the elements of `pFUN` or in `cFUN`

`data`	data frame of explanatory, intermediate and response variables.
`pFUN`	list of lists, which describe models for the intermediate variables, details are given below.
`cFUN`	either a fixed function with argument `newdata` and returning the class membership by default, or a list specifying a classifying model, similar to one element of `pFUN`. Details are given below.
`nbagg`	number of bootstrap samples.
`ns`	proportion of sample to be drawn from the learning sample. By default, subagging with 50% is performed, i.e. draw 0.5*n out of n without replacement.
`replace`	logical. Draw with or without replacement.
`...`	additional arguments (e.g. `subset`).

Details

A given data set is subdivided into three types of variables: explanatory, intermediate and response variables.

Here, each specified intermediate variable is modelled separately following `pFUN`, a list of lists with elements specifying an arbitrary number of models for the intermediate variables and an optional element `training.set = c("oob", "bag", "all")`. The element `training.set` determines whether, predictive models for the intermediate are calculated based on the out-of-bag sample (`"oob"`), the default, on the bag sample (`"bag"`) or on all available observations (`"all"`). The elements of `pFUN`, specifying the models for the intermediate variables are lists as described in `inclass`. Note that, if no formula is

given in these elements, the functional relationship of `formula` is used.

The response variable is modelled following `cFUN`. This can either be a fixed classifying function as described in Peters et al. (2003) or a list, which specifies the modelling technique to be applied. The list contains the arguments `model` (which model to be fitted), `predict` (optional, how to predict), `formula` (optional, of type `y~w1+w2+w3+x1+x2` determines the variables the classifying function is based on) and the optional argument `training.set = c("fitted.bag", "original", "fitted.subset")` specifying whether the classifying function is trained on the predicted observations of the bag sample (`"fitted.bag"`), on the original observations (`"original"`) or on the predicted observations not included in a defined subset (`"fitted.subset"`). Per default the formula specified in `formula` determines the variables, the classifying function is based on.

Note that the default of `cFUN = list(model = NULL, training.set = "fitted.bag")` uses the function `rpart` and the predict function `predict(object, newdata, type = "class")`.

Value

An object of class `"inbagg"`, that is a list with elements

`mtrees` a list of length `nbagg`, describing the prediction models corresponding to each bootstrap sample. Each element of `mtrees` is a list with elements `bindx` (observations of bag sample), `btree` (classifying function of bag sample) and `bfct` (predictive models for intermediates of bag sample).

`y`	vector of response values.
`W`	data frame of intermediate variables.
`X`	data frame of explanatory variables.

Author(s)

Andrea Peters <Peters.Andrea@imbe.imed.uni-erlangen.de>

References

David J. Hand, Hua Gui Li, Niall M. Adams (2001), Supervised classification with structured class definitions. *Computational Statistics & Data Analysis* 36, 209–225.

Andrea Peters, Berthold Lausen, Georg Michelson and Olaf Gefeller (2003), Diagnosis of glaucoma by indirect classifiers. *Methods of Information in Medicine* 1, 99-103.

See Also

`rpart`, `bagging`, `lm`

Examples

```
library(mvtnorm)
y <- as.factor(sample(1:2, 100, replace = TRUE))
W <- rmvnorm(200, mean = rep(0, 3))

X <- rmvnorm(200, mean = rep(2, 3))
colnames(W) <- c("w1", "w2", "w3")
colnames(X) <- c("x1", "x2", "x3")
DATA <- data.frame(y, W, X)
```

```
pFUN <- list(list(formula = w1~x1+x2, model = lm, predict = mypredict.lm),
list(model = rpart))

inbagg(y~w1+w2+w3~x1+x2+x3, data = DATA, pFUN = pFUN)
```

`inclass` *Indirect Classification*

Description

A framework for the indirect classification approach.

Usage

```
inclass(formula, data, pFUN = NULL, cFUN = NULL, ...)
```

Arguments

`formula`	formula. A `formula` specified as `y~w1+w2+w3~x1+x2+x3` models each intermediate variable `w1, w2, w3` by `wi~x1+x2+x3` and the response by `y~w1+w2+w3` if no other formulas are given in `pFUN` or `cFUN`.
`data`	data frame of explanatory, intermediate and response variables.
`pFUN`	list of lists, which describe models for the intermediate variables, see below for details.
`cFUN`	either a function or a list which describes the model for the response variable. The function has the argument `newdata` only.
`...`	additional arguments, passed to model fitting of the response variable.

Details

A given data set is subdivided into three types of variables: those to be used predicting the class (explanatory variables) those to be used defining the class (intermediate variables) and the class membership variable itself (response variable). Intermediate variables are modelled based on the explanatory variables, the class membership variable is defined on the intermediate variables.

Each specified intermediate variable is modelled separately following `pFUN` and a formula specified by `formula`. `pFUN` is a list of lists, the maximum length of `pFUN` is the number of intermediate variables. Each element of `pFUN` is a list with elements:
`model` - a function with arguments `formula` and `data`;
`predict` - an optional function with arguments `object, newdata` only, if `predict` is not specified, the predict method of `model` is used;
`formula` - specifies the formula for the corresponding `model` (optional), the formula described in `y~w1+w2+w3~x1+x2+x3` is used if no other is specified.

The response is classified following `cFUN`, which is either a fixed function or a list as described below. The determined function `cFUN` assigns the intermediate (and explanatory) variables to a certain class membership, the list `cFUN` has the elements `formula, model, predict` and `training.set`. The elements `formula, model, predict` are structured as described by `pFUN`, the described model is trained on the original (intermediate variables) if `training.set="original"` or if `training.set = NULL`, on the fitted values if `training.set = "fitted"` or on observations not included in a specified subset if `training.set = "subset"`.

A list of prediction models corresponding to each intermediate variable, a predictive function for the response, a list of specifications for the intermediate and for the response are returned.
For a detailed description on indirect classification see Hand et al. (2001).

Value

An object of class `inclass`, consisting of a list of

`model.intermediate`
: list of fitted models for each intermediate variable.

`model.response`
: predictive model for the response variable.

`para.intermediate`
: list, where each element is again a list and specifies the model for each intermediate variable.

`para.response`
: a list which specifies the model for response variable.

Author(s)

Andrea Peters <Peters.Andrea@imbe.imed.uni-erlangen.de>

References

David J. Hand, Hua Gui Li, Niall M. Adams (2001), Supervised classification with structured class definitions. *Computational Statistics & Data Analysis* 36, 209–225.

Andrea Peters, Berthold Lausen, Georg Michelson and Olaf Gefeller (2003), Diagnosis of glaucoma by indirect classifiers. *Methods of Information in Medicine* 1, 99-103.

See Also

bagging, inclass

Examples

```
data(Smoking)
# Set three groups of variables:
# 1) explanatory variables are: TarY, NicY, COY, Sex, Age
# 2) intermediate variables are: TVPS, BPNL, COHB
# 3) response (resp) is defined by:

classify <- function(data){
  data <- data[,c("TVPS", "BPNL", "COHB")]
  res <- t(t(data) > c(4438, 232.5, 58))
  res <- as.factor(ifelse(apply(res, 1, sum) > 2, 1, 0))
  res
}

response <- classify(Smoking[ ,c("TVPS", "BPNL", "COHB")])
smoking <- data.frame(Smoking, response)

formula <- response~TVPS+BPNL+COHB~TarY+NicY+COY+Sex+Age

inclass(formula, data = smoking, pFUN = list(list(model = lm, predict =
mypredict.lm)), cFUN = classify)
```

Bibliography

D. Andrews and A. Herzberg. *Data*. Springer-Verlag, Berlin, 1985.

T. L. Bratcher and J. D. Stamey. Estimation of poisson rates with misclassified counts. *Biometrical Journal*, 44(8):946–956, 2002.

L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996a.

L. Breiman. Out-of-bag estimation. Technical report, Statistics Department, University of California Berkeley, Berkeley CA 94708, 1996b.

L. Breiman. Arcing classifiers. *The Annals of Statistics*, 26(3):801–49, 1998.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. Wadsworth, California, 1984.

P. Bühlmann and B. Yu. Analyzing bagging. *The Annals of Statistics*, 30(4):927–961, 2002.

N. Chawla, L. Hall, K. Bowyer, T. Moore, and W. Kegelmeyer. Distributed pasting of small votes. *Lecture Notes in Computer Science*, 2364:52–61, 2002.

A. Ciampi. Generalized regression trees. *Computational Statistics & Data Analysis*, 12(1):57–78, 1991.

A. L. Coleman. Glaucoma. *The Lancet*, 354:1803–10, 1999.

D. Cox and N. Wermuth. *Multivariate Dependencies*. Chapman & Hall, 1996.

B. Efron and R. Tibshirani. Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560, 1997.

T. M. Eid, G. L. Spaeth, L. J. Katz, A. Azuara-Blanco, J. Agusburger, and J. Nicholl. Quantitative estimation of retinal nerve fiber layer height in glaucoma and the relationship with optic nerve head topography and visual field. *Journal of Glaucoma*, 6(4):221–30, 1997.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

K. Flegal, P. Keyl, and F. Nieto. Differential misclassification arising from nondifferential errors in exposure measurements. *American Journal of Epidemiology*, 134:1233–1244, 1991.

J. H. Friedman. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1:55–77, 1997.

J. H. Friedman and W. Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76(376):817–823, 1981.

D. A. Grimes and K. F. Schulz. Bias and causal associations in observational research. *Lancet*, 359:248–252, 2002.

D. J. Hand. *Construction and assessment of classification rules*. Jon Wiley & Sons, 1997.

D. J. Hand. Measuring diagnostic accuracy of statistical prediction rules. *Statistica Neerlandica*, 55(1):3–16, 2001.

D. J. Hand, H. G. Li, and N. M. Adams. Supervised classification with structured class definitions. *Computational Statistics & Data Analysis*, 36:209–225, 2001.

D. J. Hand, J. J. Oliver, and A. D. Lunn. Discriminant analysis when the classes arise from a continuum. *Pattern Recognition*, 31(5):641–650, 1998.

Heidelberg Engineering. *Heidelberg Retina Tomograph*. Heidelberg Engineering GmbH, Tiergartenstraße 17, D-69121 Heidelberg, 1997.

T. Hothorn. Bundling classifiers with an application to glaucoma diagnosis. Dissertation, University Dortmund, 2003.

T. Hothorn and B. Lausen. Bagging combined classifiers. *Classification, Clustering and Data Analysis, Studies in Classification, Data Analysis and Knowledge Organization (K. Jajuga, A. Sokołowski and H.-H. Bock eds), Springer, Heidelberg*, pages 177–184, 2002.

T. Hothorn and B. Lausen. Bagging tree classifiers for laser scanning images: a data and simulation based strategy. *Artificial Intelligence in Medicine*, 27(1):65–79, 2003a.

T. Hothorn and B. Lausen. Bundling classifiers by bagging trees. *Preprint, Friedrich-Alexander-University Erlangen-Nuremberg*, 2003b. URL `http://www.mathpreprints.com`.

T. Hothorn and B. Lausen. Double-bagging: Combining classifiers by bootstrap aggregation. *Pattern Recognition*, 36(6):1303–1309, 2003c.

M. Iester, F. S. Mikelberg, P. Courtright, and S. M. Drance. Correlation between the visual field indices and heidelberg retina tomograph parameters. *Journal of Glaucoma*, 6(2):78–82, 1997.

D. S. Kamal, A. C. Viswanathan, D. F. Garway-Heath, R. A. Hitchings, D. Poinoosawmy, and C. Bunce. Detection of optic disc change with the heidelberg retina tomograph before confirmed visual field change in ocular hypertensives converting to early glaucoma. *British Journal of Ophthalmology*, 83:290–294, 1999.

B. Lausen, W. Sauerbrei, and M. Schumacher. Classification and regression trees (CART) used for the exploration of prognostic factors measured on different scales. In P. Dirschedl and R. Ostermann, editors, *Computational Statistics*, pages 483–496, Heidelberg, 1994. Physica-Verlag.

M. LeBlanc and R. Tibshirani. Combing estimates in regression and classification. *Journal of the American Statistical Association*, 91(436):1641–1650, 1996.

B. L. Lee, R. Bathija, and R. N. Weinreb. The definition of normal-tension glaucoma. *Journal of Glaucoma*, 7(6):366–71, 1998.

E. Lehmann. *Theory of Point Estimation*. Chapman & Hall, New York, 1991.

E. Lehmann. *Elements of Large-Sample Theory*. Springer, New York, 1999.

G. Lugosi and A. Nobel. Consistency of data-driven histogram methods for density estimation and classification. *The Annals of Statistics*, 24(2):687–706, 1996.

J. Läuter. *Stabile multivariate Verfahren: Diskriminanzanalyse - Regressionsanalyse - Faktoranalyse*. Akademie Verlag, Berlin, 1992.

C. Mardin, T. Hothorn, A. Peters, A. Jünemann, G. Michelson, and B. Lausen. New glaucoma classification method based on standard HRT parameters by bagging classification trees. *Journal of Glaucoma*, 12(4):340–346, 2003.

C. Y. Mardin, F. K. Horn, J. B. Jonas, and W. M. Budde. Preperimetric glaucoma diagnosis by confocal scanning laser tomography of the optic disc. *British Journal of Ophthalmology*, 83(3):299–304, 1999.

P. Martus. A measurement model of disease damage in paired organs. *Biometrical Journal*, 43(8):927–940, 2001.

G. McLachlan. Confidence intervals for the conditional probability of misclassification in discriminant analysis. *Biometrics*, 31:161–167, 1975.

D. Meyer, F. Leisch, and K. Hornik. The support vector machine under test. *Neurocomputing*, 55(1-2):169–186, 2003.

M. Mojirsheibani. Combining classifiers via discretization. *Journal of the American Statistical Association*, 94(446):600–609, 1999.

M. Mojirsheibani. An iterated classification rule based on auxiliary pseudo-predictors. *Computational Statistics & Data Analysis*, 38:125–138, 2001.

A. Peters, T. Hothorn, and B. Lausen. Glaucoma diagnosis by indirect classifiers. *Classification, Clustering and Data Analysis, Studies in Classification, Data Analysis and Knowledge Organization (K. Jajuga, A. Sokołowski and H.-H. Bock eds), Springer, Heidelberg*, pages 465–470, 2002a.

A. Peters and B. Lausen. Direct and indirect classification in clinical research. *Biometrical Journal*, 45(8):1023–1041, 2003.

A. Peters, B. Lausen, G. Michelson, and O. Gefeller. Diagnosis of glaucoma by indirect classifiers. *Methods of Information in Medicine*, 42(1):99–103, 2003.

A. Peters, T.Hothorn, and B. Lausen. ipred: Improved predictors. *R News*, 2(2): 33–36, 2002b.

C. R. Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society, Series B*, 10:159–203, 1948.

J. S. Rao and R. Tibshirani. The out-of-bootstrap method for model averaging and selection. Technical report, University of Toronto, 1997.

B. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

K. Rothman and S. Greenland. *Modern Epidemiology*. Lippincott-Raven, 2nd edition, 1998.

R. A. Schiavo and D. J. Hand. Ten more years of error rate research. *International Statistical Review*, 68(3):295–310, 2000.

J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes. Using the adap learning algorithm to forecast the onset of diabetes mellitus. *Proceedings of the Symposium on Computer Applications and Medical Care*, IEEE Computer Society Press:261–265, 1988.

J. Song and C. Ahn. Cluster randomization trials: a simulation study. *Biometrical Journal*, 44(3):375–390, 2002.

N. V. Swindale, G. Stjepanovic, A. Chin, and F. S. Mikelberg. Automated analysis of normal and glaucomatous optic nerve head topography images. *Investigative Ophthalmology & Visual Science*, 41(7):1730–42, 2000.

R. Tibshirani and G. Hinton. Coaching variables for regression and classification. *Statistics and Computing*, 8:25–33, 1998.

L. Torgo and J. P. da Costa. Clustered multiple regression. *Data Analysis, Classification and Related Methods*, pages 217–222, 2000.

J. K. Vermunt and J. Magidson. Latent class models for classification. *Computational Statistics & Data Analysis*, 41:531–537, 2003.

T. Wetter. Editorial: Lessons learnt from bringing knowledge-based decision support into routine use. *Artificial Intelligence in Medicine*, 24:195–203, 2002.

H. Zhang and B. Singer. *Recursive Partitioning in the Health Sciences*. Springer, 1999.

List of Tables

List of Figures