

# Raman-spektroskopische Diagnostik von primären Hirntumoren mit Hilfe weicher chemometrischer Klassifikationsmethoden



---

seit 1558

## **Dissertation**

zur Erlangung des akademischen Grades  
doctor rerum naturalium (Dr. rer. nat.)

vorgelegt dem Rat der  
Chemisch-Geowissenschaftlichen Fakultät  
der Friedrich-Schiller-Universität Jena

von Dipl.-Chem. Claudia Beleites  
geboren am 6. Mai 1978 in Bad Nauheim

User beware, one must take so much care  
in matters statistical, be right and aware  
lest LODs lose their meaning,  
false positives be teeming  
though nothing worth noting is there!

— *Unpublishable Humor* by Amrita Ray & Daniel E. Weeks

Gutachter:

1. Gutachter: Prof. Dr. Jürgen Popp, Institut für physikalische Chemie,  
Friedrich-Schiller-Universität Jena

2. Gutachter: Prof. Dr. Reiner Salzer, Bioanalytische Chemie,  
Technische Universität Dresden

Tag der Verteidigung: 11. September 2014

# Inhaltsverzeichnis

<b>Wichtige Abkürzungen und Formelzeichen</b>	<b>vii</b>
<b>I Einführung</b>	<b>1</b>
<b>1 Motivation und Ziel der Arbeit</b>	<b>2</b>
<b>2 Die untersuchten Hirntumore</b>	<b>5</b>
2.1 Astrozytome und Glioblastome . . . . .	6
2.2 Lymphome . . . . .	7
2.3 Hirntumordiagnostik . . . . .	8
<b>3 Schwingungsspektroskopie</b>	<b>12</b>
3.1 Praktische Aspekte . . . . .	14
3.2 Raman- und IR-Spektroskopie von primären Hirntumoren . . . . .	18
<b>4 Chemometrische Datenanalyse</b>	<b>21</b>
4.1 Spektren als multivariate Daten . . . . .	21
4.2 Datenvorbehandlung . . . . .	22
4.3 Bilineare Modelle . . . . .	24
4.3.1 Partial Least Squares (PLS) Regression und Principal Component Analysis (PCA) . . . . .	25
4.4 Klassifikation . . . . .	26
4.4.1 Lineare Diskriminanzanalyse (LDA) . . . . .	27
4.4.2 Logistische Regression . . . . .	29
4.4.3 Weitere Klassifikationsmethoden . . . . .	31
4.4.4 Modellqualität: zufällige und systematische Fehler . . . . .	32
4.4.5 Erforderliche Trainingsprobenzahl . . . . .	32
4.5 Erweiterte Klassifikationskonzepte: Einklassen- und weiche Klassifikation . . . . .	34
4.6 Robustheit, Modellstabilität und Aggregation . . . . .	38
4.7 Resampling . . . . .	40
4.8 Validierung . . . . .	42
4.8.1 Validierungsschemata . . . . .	42
4.8.2 Resampling-basierte Validierung . . . . .	44
4.8.3 Validierung von aggregierten Modellen . . . . .	45
4.8.4 Statistische Unabhängigkeit von Trainings- und Testdaten . . . . .	46
4.8.5 Kenngrößen für die Qualität der Vorhersagen . . . . .	48
4.9 Validierung mit weicher Referenz . . . . .	59

<b>II</b>	<b>Entwicklungen und Untersuchungen zur Validierung von chemometrischen Modellen im Rahmen der Dissertation</b>	<b>63</b>
<b>5</b>	<b>Planung der erforderlichen Patientenzahlen</b>	<b>64</b>
5.1	Abschätzung der notwendigen Trainingsprobenzahl durch Messen der Lernkurve . . . . .	64
5.2	Abschätzen der notwendigen Testprobenzahl . . . . .	67
5.2.1	Modellvergleiche und Optimierung, Festlegen von Hyperparametern	68
<b>6</b>	<b>Vergleich der Validierungsschemata für biospektroskopische Daten</b>	<b>72</b>
<b>7</b>	<b>Messung der Modellstabilität mit iterierter Kreuzvalidierung</b>	<b>74</b>
7.1	Wiederholungsmessungen, Modellaggregation und <i>bolstered error estimation</i>	75
<b>8</b>	<b>Weiche Kenngrößen für die Qualität von Klassifikationsmodellen</b>	<b>77</b>
8.1	Die Zuordnungsmatrix bei weicher Klassifikation . . . . .	77
8.2	Berechnung der Kenngrößen bei weicher Klassifikation . . . . .	78
8.2.1	Abweichung von der idealen Zuordnungsmatrix . . . . .	79
8.2.2	Absoluter und quadrierter Fehler . . . . .	81
8.3	Vergleich der weichen und harten Kenngrößen . . . . .	82
8.4	Implementierung . . . . .	84
<b>III</b>	<b>Experimente, Materialien und Methoden</b>	<b>87</b>
<b>9</b>	<b>Proben</b>	<b>88</b>
9.1	Präparation . . . . .	88
<b>10</b>	<b>Raman-Messungen</b>	<b>91</b>
<b>11</b>	<b>Referenzdiagnose</b>	<b>94</b>
<b>12</b>	<b>Datenanalyse</b>	<b>97</b>
12.1	Vorbehandlung der Raman-Spektren . . . . .	97
12.2	Einteilung und Charakterisierung der Datensätze . . . . .	101
12.2.1	Grading von Astrozytomen . . . . .	101
12.2.2	Differentialdiagnostik von Lymphomen und Astrozytomen . . . . .	105
12.3	Klassifikationsmodelle . . . . .	105
12.3.1	Aggregierte Modelle . . . . .	109
12.4	Modellvalidierung . . . . .	110
<b>IV</b>	<b>Ergebnisse und Diskussion</b>	<b>111</b>
<b>13</b>	<b>Einfluss des Einbettmediums</b>	<b>112</b>
<b>14</b>	<b>Detaillierte Referenzdiagnosen</b>	<b>114</b>

<b>15 Einfluss der Probenlagerung</b>	<b>117</b>
<b>16 Tumorgrading der Astrozytome</b>	<b>119</b>
16.1 Deskriptive Modelle . . . . .	120
16.2 Prädiktive Modelle . . . . .	121
16.2.1 PLS-LDA- und PLS-LR-Modelle . . . . .	122
16.2.2 Weiche Kenngrößen . . . . .	122
16.2.3 Modellstabilität und Aggregation . . . . .	125
16.2.4 Klassifikation mit den CH-Valenzschwingungsbanden . . . . .	125
16.2.5 Spektroskopische Interpretation der prädiktiven LDA-Modelle . . . . .	126
16.2.6 Diskussion . . . . .	131
<b>17 Differentialdiagnostik von Lymphomen und Astrozytomen</b>	<b>133</b>
17.1 Deskriptive Modelle . . . . .	133
17.2 Prädiktive Modelle . . . . .	134
17.2.1 Spektroskopische Interpretation . . . . .	136
17.2.2 Klassifikation mit den CH-Valenzschwingungsbanden . . . . .	138
17.2.3 PLS-LDA- und PLS-LR-Modelle . . . . .	139
17.2.4 Aggregation von 9 Modellen . . . . .	139
17.2.5 Aggregation über mehrere Spektren . . . . .	142
17.2.6 Diskussion . . . . .	143
<b>18 Zusammenfassung</b>	<b>146</b>
<b>19 Ausblick</b>	<b>149</b>
<b>V Anhang</b>	<b>159</b>
<b>A Protokolle zur Probenpräparation</b>	<b>160</b>
A.1 Präparationsprotokoll Gefrierschnitte . . . . .	160
A.2 Fixierung der Dünnschnitte . . . . .	160
A.3 Färbung mit Methyleneblau . . . . .	160
<b>B Details zu den Modellen</b>	<b>161</b>
B.1 Details zum Astrozytomgrading . . . . .	161
B.1.1 Tabellen Astrozytomgrading . . . . .	161
B.1.2 Weitere Spezifiäts-Sensitivitäts-Diagramme . . . . .	169
B.2 Details Differentialdiagnostik Astrozytom oder Lymphom . . . . .	170
B.3 „normale“ Modelle . . . . .	170
B.4 Ensemble-Modelle . . . . .	177
B.5 Ensemble-Modell: 9 Modelle und 9 Spektren aggregiert . . . . .	179
B.5.1 Weitere Spezifiäts-Sensitivitäts-Diagramme . . . . .	183
B.5.2 Ergebnisse Harte Vorhersage . . . . .	184
<b>C Software</b>	<b>187</b>
C.1 Testprotokolle und Details zu softclassval . . . . .	187
C.1.1 Testprotokoll R Pakettest . . . . .	187

## *Inhaltsverzeichnis*

C.1.2 Unit-Tests . . . . .	188
C.2 Messprogramm RamanGUI . . . . .	188
C.3 Datenbank der Tumorproben . . . . .	189
C.3.1 Daten auf dem Dateiserver . . . . .	190
C.3.2 PostgreSQL-Datenbank . . . . .	191
<b>Publikationen und Software zu dieser Arbeit</b>	<b>194</b>
<b>Referenzen</b>	<b>196</b>
<b>Glossar</b>	<b>216</b>
<b>Index</b>	<b>221</b>

# Wichtige Abkürzungen und Formelzeichen

- ALA** 5-Aminolävulinsäure.
- ANN** Künstliches neuronales Netz (engl. *artificial neural net*).
- ATR** Attenuated Total Reflection.
- BHS** Blut-Hirn-Schranke.
- BSE** Bovine Spongiforme Encephalitis.
- CARS** kohärente anti-Stokes Raman-Streuung bzw. -Spektroskopie (engl. *Coherent Anti-Stokes Raman Scattering or -Spectroscopy*).
- CBTRUS** Central Brain Tumor Registry of the United States.
- CCD** engl. *charge coupled device*.
- CT** Computertomographie bzw. -Tomogramm.
- DNS** Desoxyribonukleinsäure(n).
- FTIR** Fourier"-transform Infrarot (Spektroskopie).
- IR** Infrarot (Spektroskopie).
- LDA** Lineare Diskriminanzanalyse.
- MRT** Magnet-Resonanz-Tomographie bzw. -Tomogramm.
- MSC** engl. *multiplicative signal correction*.
- PBS** physiologische Kochsalzlösung mit Phosphatpuffer (engl. *Phosphate Buffered Saline*).
- PCA** Hauptkomponentenanalyse (engl. *principal component analysis*).
- PEG** Polyethylenglykol.
- PET** Positronen-Emissions-Tomographie.
- PLS** engl. *partial least squares*.
- PLS-DA** engl. *Partial Least Squares Discriminant Analysis*.
- PZNSL** primäres ZNS-Lymphom.
- RNS** Ribonukleinsäure(n).
- ROC** engl. *Receiver Operating Curve*.
- SERS** Oberflächenverstärkte Raman-Spektroskopie (engl. *Surface Enhanced Raman Spectroscopy*).
- SIMCA** engl. *Soft Independent Modeling of Class Analogies*.
- SPECT** engl. *Single Photon Emission Computed Tomography*.
- SVM** engl. *Support Vector Machine*.
- WHO** Weltgesundheitsorganisation (engl. *World Health Organization*).
- ZNS** Zentrales Nervensystem.

## Wichtige Abkürzungen und Formelzeichen

Zeichen	Bedeutung
<i>allgemein</i>	
$s^2$	Varianz
$s$	Standardabweichung
$\bar{x}$	Mittelwert von $x$
$\mathbf{X}^T$	Transponierte von Matrix $\mathbf{X}$
<i>Spektroskopie</i>	
$I$	Intensität
$E$	Energie oder Extinktion
$\lambda$	Wellenlänge
$\nu = \frac{c}{\lambda}$	Frequenz
$\tilde{\nu} = \frac{1}{\lambda}$	Wellenzahl
$\Delta \tilde{\nu} = \tilde{\nu}_0 - \tilde{\nu} = \frac{1}{\lambda_0} - \frac{1}{\lambda}$	Wellenzahldifferenz
$\epsilon_i$	Extinktionskoeffizient der Substanz $i$
$c_i$	Konzentration von Substanz $i$
$d$	optische Pfadlänge
<i>Chemometrie allgemein</i>	
$\mathbf{X}^{(n \times p)}$	Datenmatrix (hier: Spektrenmatrix)
$n$	Anzahl der Objekte (Spektren, Proben, Messungen)
$p$	Anzahl der Variaten (Messkanäle, Datenpunkte im Spektrum)
$\epsilon$	Fehlermatrix, Fehlerterm, Residuen
<i>Klassifikation</i>	
$n_g$	Anzahl der Klassen
$G \in \{1, \dots, n_g\}$	Klasse $G$
$R \in \{1, \dots, n_g\}$	Referenzinformation (hart)
$P \in \{1, \dots, n_g\}$	Modellvorhersage (hart)
$g \in [0, 1]^{n_g}$	Klassenzugehörigkeit
$r \in [0, 1]^{n_g}$	Referenzinformation Klassenzugehörigkeit
$p \in [0, 1]^{n_g}$	Modellvorhersage Klassenzugehörigkeit
$\mathbf{G}^{(n \times n_g)}$	Klassenzugehörigkeitsmatrix
<i>Validierung</i>	
$k$	Anzahl der Sets in der Kreuzvalidierung, oder Anzahl der Ereignisse einer Binomialverteilung
$i$	Anzahl der Iterationen
$p$	Wahrscheinlichkeit bzw. Häufigkeit eines Ereignisses
$\hat{p}$	beobachtete Häufigkeit
Pr	Wahrscheinlichkeit
$\mathbf{Z}^{(n_g \times n_g)}, \mathbf{Z}$	Zuordnungsmatrix und Funktion, die die Zuordnungsmatrix berechnet
$i$	Zeilenindex der Zuordnungsmatrix (Referenzinformation)
$j$	Spaltenindex der Zuordnungsmatrix (Vorhersage durch Modell)
$\text{Sens}_G^O(\text{Spez}_G^O, \text{PPV}_G^O, \text{NPV}_G^O)$	Sensitivität (Spezifität, positiver und negativer Vorhersagewert) Klasse $G$ , berechnet mit Hilfe des Operators $O$ , siehe Kap. 4.8.5 und 8.2
$\alpha, \beta$	Fehler I. Art und Fehler II. Art

Weitere abkürzende Schreibweisen: Summen über bestimmte Teile der Zuordnungsmatrix  $\mathbf{Z}$  addieren alle Elemente, die den in den Indices angegebenen Bedingungen entsprechen. Z. B. bedeutet  $\sum \mathbf{Z}_{i,p}$  die Summe über alle Zeilen der Spalte  $P$ :  $\sum_{i=1}^{n_g} \mathbf{Z}_{i,p}$ .  $\sum \mathbf{Z}_{i \neq G,p}$  steht abkürzend für  $\sum_{i \in \{1, \dots, n_g\} | i \neq G} \mathbf{Z}_{i,p}$ , die Summe über alle Zeilen außer  $G$  der Spalte  $P$ .  $\sum_n$  steht für die Summe über alle  $n$  Fälle (Spektren, Proben).

Die im Rahmen der vorliegenden Arbeit entstandenen Artikel [CB1–CB13] die von mir geschriebenen R-Pakete `softclassval` und `arrayhelpers` (beide [CB1]) und `hyperSpec` [CB14] sowie das von Simon Fuller unter meiner Betreuung erstellte `OpenBlasThreads` [CB15] können an der Markierung „CB“ in der Referenz erkannt werden. Außerdem zitiere ich Ergebnisse aus meiner Diplomarbeit [16], die natürlich nicht Bestandteil der vorliegenden Arbeit und daher nicht gesondert markiert ist.



# **Teil I**

## **Einführung**

# 1 Motivation und Ziel der Arbeit

Das Ziel dieser Arbeit ist, neue Möglichkeiten der Hirntumordiagnostik an nativem Gewebe aufzuzeigen, um mit neuen chemometrischen Methoden den Weg zu einer neuen *in vivo* Tumordiagnostik zu eröffnen. Die experimentelle Grundlage ist Raman-Spektroskopie an Hirngewebe. Normales Hirngewebe, Astrozytome, Glioblastome und Lymphome werden auf der Grundlage ihrer Schwingungsspektren mit Hilfe einer zu diesem Zweck entwickelten chemometrischen Datenauswertung beurteilt.

Raman-Spektren können mit faseroptischen Sonden *in vivo* aufgenommen werden. Ein geeignetes chemometrisches Modell könnte aus diesen Spektren innerhalb von Millisekunden auf die Gewebeart schließen. Das führt zu zwei Szenarien für eine *in vivo* Raman-Diagnostik. Einerseits könnten Raman-Sonden als Werkzeug während der Operation benutzt werden. Der Neurochirurg könnte damit während der Operation messen, ob der Tumorrand erreicht ist. Andererseits lassen sich diese Faseroptiken in Endoskope oder Biopsienadeln integrieren. Raman-Spektroskopie könnte helfen, die Diagnostik auf der Basis von Nadelbiopsien grundlegend zu verbessern. Biopsieproben sind mitunter zu klein, um eine sichere histologische Diagnose zu stellen. Eine Raman-unterstützte Diagnostik kann unmittelbar nach der Aufnahme der Spektren zur Verfügung stehen. Ist mit einem Spektrum keine sichere Aussage möglich, so können weitere Spektren aufgenommen werden, um zu einer sicheren Diagnose zu gelangen. In einem ersten Schritt könnte so die Raman-Spektroskopie der Stereonavigation bei der Entnahme der Biopsieproben helfen und sicherstellen, dass die Biopsieprobe tatsächlich aus dem Tumor entnommen wird. Letztlich könnte ganz auf die Entnahme von Tumorgewebe verzichtet werden. Bei der Entnahme einer Biopsie können Tumorzellen entlang des Wegs der Hohlnadel verschleppt werden, so dass weitere Tumorherde entstehen können. Raman-Sonden können mit Frontlinsen ausgestattet werden, so dass sie tiefer im Inneren des Gewebes messen. Mit einer solchen Technik ließe sich das Risiko, Tumorzellen zu verschleppen, stark verringern.

Diese Arbeit untersucht für beide Szenarien jeweils eine Fragestellung. Für das Szenario der Diagnostik während einer Operation steht die Erkennung der verschiedenen Tumorgrade von Astrozytomen im Zentrum dieser Arbeit. Astrozytome und Glioblastome gehören zu den Tumoren der Glia, den Gliomen<sup>(a)</sup>. Die vollständige chirurgische Entfernung des Tumors ist einer der wichtigsten unabhängigen Faktoren für die Vorhersage der medianen Lebenserwartung des Patienten [17, 18]. Die Gliome werden daher in der Regel so weit wie möglich chirurgisch entfernt. Bei vielen Tumoroperationen außerhalb des Gehirns wird ein Sicherheitsabstand zum Tumor eingehalten, der sicherstellt, dass der Tumor komplett entfernt wird. Das ist im Gehirn nicht möglich, da das normale Hirngewebe auch in unmittelbarer Nähe des Tumors geschont werden muss [19, 20]. Allerdings unterscheiden sich die Gliome visuell oft nur wenig vom umliegenden Gewebe und die Übergänge zwischen den verschiedenen Tumorstadien verlaufen fließend.

---

<sup>(a)</sup> Im folgenden steht *Astrozytom* als Oberbegriff für Astrozytome und Glioblastome.

Zudem wachsen Astrozytome infiltrativ, Tumorzellen wandern in normales Gewebe ein. Der maligne (bösartige) Teil des Tumors soll bei einer Gliomoperation auf jeden Fall entfernt werden. Demgegenüber werden niedriggradige Anteile des Tumors gegebenenfalls nicht entfernt, um nicht umliegendes normales und funktionales Hirngewebe zu gefährden. Deshalb werden bei einer solchen Operation genaue Informationen über das Gewebe am Resektionsrand benötigt. Aus Sicht der Neurochirurgen besteht damit ein großer Bedarf an Werkzeugen, die bei der Einstufung der verschiedenen Gewebe *in vivo* während der Operation helfen. Ein Ziel dieser Arbeit ist daher die Differenzierung zwischen normalem Hirngewebe sowie niedriggradigem und hochgradigem Astrozytomgewebe.

Eine Raman-Biopsie könnte bei der Unterscheidung zwischen Astrozytomen und Lymphomen helfen. Lymphome werden anders als Astrozytome oft ohne weiteren chirurgischen Eingriff mit einer Chemotherapie behandelt. Die Therapieplanung erfolgt meist anhand von Biopsieproben. Histologisch ist dabei die Unterscheidung von Astrozytomen und Lymphomen mitunter schwierig. Die Biopsieproben sind oft zu klein, um eine eindeutige Diagnose zu stellen. Daher wird als zweite Fragestellung die Unterscheidung von Astrozytomen (inklusive der Glioblastome) und Lymphomen untersucht.

Raman-Spektren geben Summeninformationen über die biochemische Zusammensetzung der gemessenen Probe wieder und können in eine Diagnostik umgerechnet werden. Dazu sind chemometrische Auswertemethoden erforderlich. Eine Differentialdiagnostik entspricht aus chemometrischer Sicht einer Klassifikation: die Probe wird einer der möglichen Differentialdiagnosen zugeordnet. In dieser Arbeit werden Klassifikationsmodelle entwickelt, die anhand der gemessenen Raman-Spektren normales Gewebe, niedriggradiges und hochgradiges Tumorgewebe beziehungsweise normales, Astrozytom- oder Lymphomgewebe unterscheiden.

Gliome wachsen sehr heterogen. Daher enthält eine Tumorprobe häufig verschiedene morphologisch unterschiedliche Gewebe und Zellpopulationen sowie Zellpopulationen, die nicht eindeutig einem Tumorgrad zugeordnet werden können, zum Beispiel, weil gerade eine weitere Entdifferenzierung stattfindet. Bisher wurden solche Proben sowohl aus der Bildung der Klassifikationsmodelle als auch aus der Validierung ausgeschlossen. Der Ausschluss dieser Grenzfälle ist allerdings in mehrfacher Hinsicht problematisch. In der vorliegenden Arbeit hätte knapp  $\frac{1}{3}$  aller Proben und fast die Hälfte aller Spektren ausgeschlossen werden müssen. Zwar untersucht diese Arbeit verglichen mit den üblichen Studiengrößen in der Biospektroskopie mit 86 Patienten eine große Patientenzahl. Diese liegt jedoch immer noch Größenordnungen unter den einschlägigen Empfehlungen für statistische Modellierung [21–25]. Darüberhinaus entstünde durch das Ausschließen der Grenzfälle ein scheinbar einfacheres Problem und es besteht die Gefahr, dass Modelle, die nur eindeutige Proben bei der Modellbildung berücksichtigen, nicht auf Grenzfälle übertragbar sind. Die Einstufung von Grenzfällen ist aber das Ziel der hier entwickelten Diagnostik. Diese Arbeit verwendet daher erstmals in der Biospektroskopie Klassifikationsmodelle, die auch Grenzfälle bei der Modellbildung berücksichtigen.

Allerdings reicht es nicht aus, ein gutes Klassifikationsmodell zu erstellen: die Leistungsfähigkeit des Modells muss auch nachgewiesen werden. Das ist Aufgabe der Modellvalidierung. Im Theorieteil dieser Arbeit werden daher verschiedene Möglichkeiten zur Validierung von Klassifikationsmodellen für spektroskopische Datensätze zusammengetragen, untersucht und auch neu entwickelt.

Die Ränder der Astrozytome zu erkennen ist das Ziel der neuen Diagnostik. Noch wich-

tiger als Grenzfälle zum Erstellen der Diagnostik zu nutzen, ist deshalb, die Qualität der vorgenommenen Einstufung auch für die Grenzfälle nachzuweisen. Im Theorieteil wird also eine einheitliche Behandlung von Kenngrößen für die Qualität der Klassifikationsergebnisse entwickelt, die einerseits existierenden Kenngrößen für Klassifikationsmodelle wie Sensitivität und Spezifität entsprechen, andererseits aber auf Grenzfälle anwendbar sind. Diese neue Validierungsmethodik ermöglicht es, die modellierten Klassengrenzen gezielt zu testen.

Ein solcher Nachweis der Vorhersagequalität für Grenzfälle ist nicht nur für das hier vorgestellte Astrozytomgrading wichtig, sondern für alle Anwendungen von Klassifikationsmodellen, bei denen der Übergang zwischen den Klassen physisch vorkommen kann. In der biospektroskopischen Tumordiagnostik ist das häufig, zum Beispiel, wenn Gewebeararten unterschieden werden sollen, die eine Reihe steigender Tumorgrade bilden oder der Übergang von Dysplasie zu Tumor erkannt werden soll. Grenzfälle treten auch auf, wenn Mischungen aus den einzelnen Klassen mit einer niedrigen räumlichen Auflösung untersucht werden, so dass im Messvolumen eine Mischung vorliegt. Weiterhin betrifft diese Problematik jegliche Diagnostik, die die Referenzmethode besonders im Hinblick auf Grenzfälle und Proben unterstützen soll, die nicht oder nicht sicher eingestuft werden können.

## 2 Die untersuchten Hirntumore

Primäre Tumore des zentralen Nervensystems (ZNS) und des Gehirns treten nach Angaben des Central Brain Tumor Registry of the United States (CBTRUS) [26] mit einer Inzidenz von 20,6 Neuerkrankungen pro Jahr und 100 000 Einwohnern auf (alterskorrigiert). Davon sind knapp  $\frac{1}{3}$  Gliome. Die in dieser Arbeit untersuchten diffusen und anaplastischen Astrozytome und die Glioblastome machen etwa  $\frac{3}{4}$  der Gliome aus. Tabelle 2.1 gibt eine Übersicht über epidemiologische Daten der betrachteten Hirntumore.

Primäre intrakranielle Tumore werden nach einer Empfehlung der Weltgesundheitsorganisation (engl. *World Health Organization*, WHO) in vier Malignitätsgrade unterteilt, die Histologie und voraussichtliches Verhalten widerspiegeln [29, 30], eine Übersicht gibt Tabelle 2.2.

**Tabelle 2.1** Klinisches Verhalten und Epidemiologie der untersuchten Hirntumore [17, 19, 20, 26–28].

Tumor	Inzidenz <sup>a</sup>	Anteil <sup>b</sup>	WHO-Grad	mediane Überlebenszeit / a
Diffuse Astrozytome	0,1 – 0,6	0,7 – 2,8	II	4 – 5
Anaplastische Astrozytome	0,4 – 0,5	1,7 – 3,2	III	1 – 2
Glioblastome	3,2	15 – 20	IV	0,5
<i>alle Tumore des Neuroepithels</i>	6,6	32		
Lymphome PZNSL	0,1 – 0,8	3 – 5	IV	knapp 1 Studien 2,5 - 8

<sup>a</sup> in 100 000 Personenjahren

<sup>b</sup> an allen primären Tumoren des ZNS in Prozent

**Tabelle 2.2** Grading der primären intrakraniellen Tumore nach histologischen Kriterien [26, 31, 32].

WHO-Grad	Malignität	Histologische Kriterien	mediane Überlebenszeit / a
°I	benigne	Gut differenziertes Gewebe	5 – 50
°II	semibenigne	Einzelne atypische Zellen; Kernatypien; noch gut differenziertes Gewebe	4 – 5
°III	maligne	Viele atypische Zellen; Mitosen; Ursprungsgewebe entdifferenziert, aber noch erkennbar	1 – 3
°IV	hochmaligne	Entdifferenziertes Gewebe, viele Mitosen; Nekrosen; Endothelproliferation	0,5 – 2

## 2.1 Astrozytome und Glioblastome

Astrozytome sind Tumore, die aus Astrozyten (Sternzellen) entstehen. Die Astrozyten gehören zur Glia, dem Stützgewebe des Gehirns. Die Astrozyten regeln insbesondere den Flüssigkeits- und Ionenhaushalt sowie den pH-Wert des umgebenden Hirngewebes. Astrozyten umgeben die Nervenzellen und trennen sie sowohl von der Hirnhaut als auch von den Blutgefäßen im Hirn. Dabei bilden sie zusammen mit dem Gefäßendothel die Blut-Hirn-Schranke (BHS). Entsprechend versorgen sie die Nervenzellen mit Nährstoffen und sorgen für den Abtransport von Stoffwechselprodukten. Sie stabilisieren das Hirngewebe mechanisch. Beim Erwachsenen teilen sich die Astrozyten normalerweise nur selten. Bei Bedarf können sie sich jedoch teilen und zum Beispiel bei Verletzungen Glianarben und Gliosen bilden [33].

Astrozyten können zu Astrozytomen entarten. Die Astrozytome gehören zu den Gliomen und somit zu den neuroepithelialen Tumoren. Knapp die Hälfte aller primären Hirntumore sind Gliome. Gliome sind überproportional häufig bösartige Neubildungen und stellen gut  $\frac{3}{4}$  der malignen primären Hirntumore. Prognose und Therapie hängen stark vom Grad des Glioms ab (Tabelle 2.1).

Über Ursachen und Risikofaktoren zur Entstehung von Astrozytomen ist nur wenig bekannt. Allergien sind negativ mit Gliomen korreliert, während Rauchen keine Rolle spielt [34]. Bestimmte erbliche Syndrome gehen mit einem erhöhten Risiko einher, an einem Gliom zu erkranken [32]. In den letzten Jahren wurden weitere mit Gliomen assoziierte genetische Veränderungen gefunden [35, 36]. Der beobachtete Zusammenhang zwischen Tumor und Genotyp ist allerdings bei Oligodendrogliomen oder Astrozytomen mit oligodendroglialen Anteilen ausgeprägter als bei Astrozytomen [37]. Ionisierende Strahlung ist der einzige gesicherte Umwelteinfluss, der mit einem erhöhten Risiko einhergeht [38]. Weiterhin werden elektromagnetische Wellen im Radiofrequenzbereich als Ursache kontrovers diskutiert [39–43], und die WHO hat eine Einstufung als „möglicherweise krebserregend“ bei „eingeschränkter Datenlage“ vorgenommen [44].

*Diffuse, fibrilläre, protoplasmatische* und *gemistozytäre* Astrozytome werden nach der WHO-Klassifikation als „noch gutartig“ (°II) eingestuft und gehören zu den niedriggradigen Gliomen, wobei gemistozytäre Astrozytome manchmal auch als hochgradig eingestuft werden [45]. Astrozytome °II neigen jedoch dazu, in höhergradige (maligne) Tumore überzugehen. Die Wahrscheinlichkeit, dass ein niedriggradiges Gliom entartet, wird auf 20 – 90 % geschätzt [17, 45]. Dabei geschehen Änderungen auf molekularer Ebene nicht unbedingt synchron mit morphologischen Veränderungen [30, 46–49]. Die mediane Überlebenszeit bei Astrozytomen °II liegt bei 4 – 5 Jahren. Die Patienten können in weitere Untergruppen unterteilt werden, mit medianen Überlebenszeiten zwischen etwa 3 und über 7 Jahren [17]. Gutartige Astrozytome werden oft nur beobachtet. Erst bei stärkeren Symptomen oder Tumorwachstum folgt eine Operation, bei Rezidiven auch Strahlentherapie sowie möglicherweise eine Chemotherapie.

*Anaplastische* Astrozytome (WHO °III) sind bereits eindeutig maligne (bösartig) und gehören daher zu den hochgradigen Gliomen. Bei diesen Tumoren beträgt die mediane Überlebenszeit zwischen 1 und 3 Jahren. Auch bei diesen Patienten gibt es Untergruppen, deren mediane Überlebenszeiten zwischen  $1\frac{1}{2}$  und 5 Jahren liegen [18].

Schreitet die Entdifferenzierung des Gewebes weiter fort, geht der Tumor schließlich in ein *Glioblastom* (WHO °IV) über. Auch die höchste Stufe der Entdifferenzierung anderer

Zellarten der Glia, zum Beispiel der Oligodendrozyten, heißt Glioblastom. Die Mehrzahl der Glioblastome entsteht direkt als Tumor °IV [32]. Die Diagnose „Glioblastom“ wird nach histologischen Gesichtspunkten gestellt, aus molekularbiologischer Sicht handelt es sich um eine heterogene Gruppe von Tumoren [50]. Liegt bei der Erstdiagnose bereits ein Tumor dieser höchsten Malignität vor, so bedeutet das eine mediane Lebenserwartung von etwa  $\frac{1}{2}$  Jahr [51].

Anaplastische Astrozytome und Glioblastome werden nach denselben Therapierichtlinien behandelt [19, 20]. Die Standardtherapie ist Resektion (chirurgische Entfernung), gefolgt von Strahlentherapie. Auch hier ist eine Chemotherapie möglich. Generell sind Astrozytome jedoch relativ resistent gegen Bestrahlung und Chemotherapeutika [52, 53].

*Pilozytische* Astrozytome (WHO °I) sind klinisch von den hier betrachteten Astrozytomen verschieden. Sie treten hauptsächlich bei Kindern und Jugendlichen auf und sind nicht Gegenstand der vorliegenden Arbeit.

**Infiltratives Wachstum und Polymorphie:** Die Gen- und Proteinexpression von morphologisch ähnlichen Gliomgeweben variiert und die biochemischen Veränderungen bei der fortschreitenden Entdifferenzierung sind recht kontinuierlich [30, 46–49]. Weiterhin erfolgen diese Übergänge nicht im gesamten Tumor gleichzeitig, so dass die Tumore räumlich heterogen sind. Eine Therapie kann weitere Veränderungen wie eine Selektion von Zellen bewirken.

Zellen der Astrozytome wandern in das umliegende normale Gewebe ein [45]. Bis zu 2 cm um den „soliden“ Tumor herum wird ein Tumorzellanteil von ca. 10 % gefunden. Das bedeutet, dass etwa 6 % aller Tumorzellen außerhalb des soliden Tumorgewebes sind. Im Abstand von 2 bis 4 cm sind noch etwa 1 % der Zellen Tumorzellen. In dieser Zone befinden sich damit knapp 2 % aller Tumorzellen [27]. Daher ist eine chirurgische Entfernung aller Tumorzellen nicht möglich. Als Kriterium für die „vollständige“ Tumorresektion wird meist die Kontrastmittelaufnahme im Magnetresonanz-Tomogramm (MRT) nach der Operation herangezogen [45]. Eine solche vollständige Tumorresektion entfernt etwa 90 bis 95 % der Tumorzellen, so dass größenordnungsmäßig  $10^{10}$  Zellen im Patienten verbleiben [27]. Nähert man die Astrozyten als Kugeln von 10 µm Durchmesser, so ergibt sich ein Restvolumen in der Größenordnung von 5 cm<sup>3</sup>.

Astrozytome und Glioblastome sind also sehr heterogene Tumore. Ein Tumor besteht oft aus verschiedenen morphologisch unterschiedlichen Geweben und Geweben mit gemischten Zellpopulationen (Infiltration).

## 2.2 Lymphome

Lymphome kommen im Gehirn als primäre Tumore oder als Metastasen vor. Die Angaben zur Inzidenz der primären Lymphome des zentralen Nervensystems (PZNSL) in der Literatur schwanken stark<sup>(a)</sup> und liegen zwischen 0,1 und 0,8 pro 100 000 Personenjahre

<sup>(a)</sup> Epidemiologische Daten aus Krebsregistern sind mit deutlichen Unsicherheiten behaftet. Oft ist unklar, welcher Anteil der Kranken tatsächlich von den entsprechenden Registern erfasst wird. Coté *et al.* verglichen Daten von Krebsregistern mit denen von AIDS-Registern für AIDS-Patienten mit Lymphom. Nur  $\frac{1}{3}$  der Patienten wurde korrekt in beiden Registern geführt. Etwa die Hälfte der restlichen Patienten war nur im Krebs-Register, die andere Hälfte nur im AIDS-Register erfasst.

[20, 26, 27], beziehungsweise etwa 0,5 – 7 % der hirneigenen Tumore [19, 20, 26, 31, 55–58]. Immunsuppression und Infektion mit HIV erhöhen das Erkrankungsrisiko stark. Die Inzidenz bei HIV-positiven Patienten ist um 3 Größenordnungen erhöht. Für Patienten mit Acquired Immune Deficiency Syndrome (AIDS) liegt sie bei 4 – 10 pro 1000 Personenjahre [19, 54, 57, 59]. Mit der hochaktiven antiretroviralen Therapie ist die Inzidenz zurückgegangen [19, 59, 60]. Diamond *et al.* [59] sprechen von einem Rückgang von 8,4 auf 1,1 je 1000 Personenjahre.

PZNSL sind hochmaligne Tumore. Die mediane Lebenserwartung ab Diagnose wird unbehandelt mit etwa 2 Monaten angegeben [19, 31, 55, 56]. Coté *et al.* [54] fanden mediane Überlebenszeiten von 2 Monaten für Lymphom-Patienten mit AIDS und 5–7 Monaten für Patienten ohne AIDS [54]. Resektion verbessert die Lebenserwartung nicht. Trotzdem erfordert die histologische Sicherung der Diagnose eine Biopsie [19, 55, 57, 58]. Strahlentherapie verbessert die mediane Überlebenszeit auf 1–1½ Jahre. In Therapiestudien mit Chemo- bzw. kombinierter Chemo- und Strahlentherapie wurden mediane Überlebenszeiten von 2½–8 Jahre beobachtet [19, 55, 61].

### 2.3 Hirntumordiagnostik

Magnetresonanztomographie (MRT) ist das wichtigste bildgebende Verfahren zur Hirntumordiagnose, gegebenenfalls ergänzt durch Computertomographie (CT) und Angiographie. CTs haben meist eine höhere Ortsauflösung (Größenordnung 100 µm) als MRTs (Größenordnung 1–3 mm). Andererseits haben MRTs für weiche Gewebe einen wesentlich besseren Kontrast. Vor der Operation aufgenommene MRTs (oder CTs) dienen während der Operation zur Stereonavigation. Allerdings verschieben sich die Gewebe im Gehirn durch das Öffnen des Schädels, Bewegungen aufgrund des Pulses, den Einschnitt und das folgende Anschwellen der Gewebe sowie das Beiseiteschieben von normalem Gewebe (engl. *brain shift*). Nimsky *et al.* [62] beobachteten Verschiebungen von bis zu 24 mm an der Hirnrinde. Am Tumorrand traten bei  $\frac{2}{3}$  der Operationen Verschiebungen von mehr als 3 mm auf. Moderne Stereonavigationssysteme bieten mit vor der Operation aufgenommenen MRTs und CTs eine Navigationsgenauigkeit von etwa 2 bis 3 mm [63], sodass die höhere Ortsauflösung des CT keinen Vorteil bringt. Nach Angaben von Dr. Kirsch [64] arbeiten Neurochirurgen gegenwärtig auf bis zu 1 mm genau. Bei Astrozytomen ist der Tumorrand oft nicht unter dem Mikroskop erkennbar. Dann ist die Genauigkeit der Stereonavigation und damit auch der Operation durch den *brain shift* begrenzt.

Standarddiagnostik ist daher MRT mit und ohne Kontrastmittel, oder CT, wenn MRT nicht möglich ist [19]. Die Schnittbilder liefern wertvolle Hinweise nicht nur zu Tumorgöße und -lage, sondern auch zu Art und Verhalten [18]. Einen wichtigen Anhaltspunkt zur Malignität des Tumors liefert die Kontrastmittelaufnahme. Allerdings zeigen ca. 15–40 % der niedriggradigen Gliome Kontrastmittelaufnahme und etwa ein Drittel der Gliome, die kein Kontrastmittel aufnehmen, sind hochgradige Tumore [17, 31]. Außerhalb der klinischen Routinediagnostik können auch *Single Photon Emission Computed Tomography* (SPECT), Positronen-Emissions-Tomographie (PET), Magnet-Resonanz-Spektroskopie und funktionelle MRT zur Diagnostik von Hirntumoren genutzt werden. PET mit  $^{18}\text{F}$ -Desoxyglucose oder  $^{11}\text{C}$ -Methionin wird zur Rezidivdiagnostik empfohlen [65].

Die genaue Diagnose des Tumorgades und der Art einschließlich der Abgrenzung der



Gliome von Lymphomen, anderen primären Hirntumoren und Metastasen benötigt einen histologischen Befund. Entsprechende Proben werden bei einer stereotaktischen Biopsie oder während der Tumoresektion gewonnen. Diese Proben sind jedoch oft zu klein, um den gegebenenfalls sehr viel größeren Tumor zu repräsentieren [17, 18, 66]. Die genaue Tumorart kann histologisch nicht aus nekrotischem Gewebe bestimmt werden. Deshalb wird die Probe möglichst vom Rand des Tumors genommen [18]. Diese Proben spiegeln nicht immer die Malignität des gesamten Tumors wider. Diener *et al.* [19] geben jedoch an, dass mit stereotaktischer Navigation für über 90 % der Patienten aus den Biopsieproben eine sichere Diagnose gestellt werden kann.

Lymphome des ZNS treten oft um die Blutgefäße herum auf (*perivaskulär*). Die Differentialdiagnostik muss sie letztlich außer von Astrozytomen noch von einer Reihe weiterer Erkrankungen wie zum Beispiel multipler Sklerose oder Toxoplasmose unterscheiden [20, 55, 56]. Um Lymphome herum finden sich oft astrozytische Reaktionen, so dass dort auch die Glia verändert ist [45]. Auf die Biopsie zur Differentialdiagnostik von Lymphomen und Astrozytomen kann verzichtet werden, wenn im Liquor oder Glaskörper Lymphomzellen gefunden werden [20].

**5-Aminolävulinsäure:** Ein neuer Ansatz, der bei der Resektion der Tumore helfen soll, nutzt 5- oder  $\delta$ -Aminolävulinsäure ( $\text{HOOC} - \text{CH}_2 - \text{CH}_2 - \text{CO} - \text{CH}_2 - \text{NH}_2$ , ALA). ALA tritt in vielen Organismen während der Porphyrinsynthese auf [67, 68]. Anwesenheit von ALA führt zur Anreicherung von Porphyrinen in Gliomen, wie auch verschiedenen anderen Tumorgeweben und Epithelien. Die erforderliche Selektivität für eine diagnostische Nutzung der Anreicherung von Porphyrinen in Gliomen gegenüber dem umgebenden normalen Hirngewebe ist gegeben. Porphyrine können von Licht im Wellenlängenbereich von etwa 360–440 nm, also im nahen UV bis blauen sichtbaren Licht, elektronisch angeregt werden. Die Relaxation erfolgt unter Fluoreszenz oder strahlungslos über einen Triplett-Zustand. Die Fluoreszenz kann während der Operation genutzt werden, um das Tumorgewebe besser vom umliegenden normalen Gewebe zu unterscheiden (engl. *fluorescence-guided surgery*). Allerdings lassen sich auf diese Weise nur Teile des Tumors, die tatsächlich ALA aufnehmen, markieren. Dazu muss die BHS kompromittiert sein [19, 69–71].

Eine multizentrische Phase-III-Studie an 270 Patienten hat konventionelle und Fluoreszenz-geleitete Operationen bei Patienten mit malignen Gliomen verglichen. Die mediane progressionsfreie Überlebenszeit betrug 5,1 Monate mit ALA und 3,6 Monate konventionell. 65 % der mit ALA operierten Patienten waren im postoperativen MRT tumorfrei (konventionell 36 %). Die progressionsfreien Überlebensraten waren für die mit ALA operierten Patienten zwischen 4 Monaten und einem Jahr nach der Operation höher als für die konventionell operierten Patienten, später verschwand der Vorteil. Der größte relative Vorteil bestand nach 6 Monaten, als die ALA-behandelte Gruppe etwa doppelt so viele progressionsfreie Patienten hatte wie die normal operierte (41 bzw. 21 %) [70].

Der Tumorrest wurde mit Hilfe der Kontrastmittelaufnahme im MRT beurteilt. Da sowohl Kontrastmittel als auch ALA nur bei kompromittierter BHS aufgenommen werden, werden Tumorteile, die keine ALA aufnehmen, möglicherweise auch im MRT nicht erkannt [19, 71].

**Intraoperative Magnetresonanztomographie:** Eine zweite Option der Bildgebung besteht darin, während der Operation weitere MRT-Bilder aufzunehmen, die auch direkt durch das Stereonavigationssystem genutzt werden. Dabei wird zwischen den sogenannten Niedrigfeldgeräten mit Magnetfeldstärken in der Größenordnung von 0,15 T und Hochfeldgeräten mit 1,5 oder 3 T unterschieden. Die Magnete von Hochfeldgeräten sind sehr viel größer und schwerer (Größenordnung: 7,5 t).

Aufgrund der geringeren Magnetfeldstärke ist die Bildqualität von intraoperativen Niedrigfeld-MRTs schlechter als die von konventionellen MRTs. Sowohl bei Hoch- als auch bei Niedrigfeldgeräten treten zudem häufig Artefakte auf, die bei konventionellen Messungen vermieden werden können [72]. Kubben *et al.* [73] (vergleiche auch [74]) schließen in ihrer Literaturstudie, dass weitere kontrollierte und vor allem prospektive Studien nötig sind, um die Leistungsfähigkeit von intraoperativen MRT bezüglich der Überlebensdauer und Lebensqualität der Patienten zu beurteilen. Sie regen insbesondere an, intraoperatives MRT direkt mit ALA zu vergleichen. Liang und Schulder [75] weisen darauf hin, dass die Vorteile für Patienten mit niedriggradigen Gliomen ausgeprägter sind als für Patienten mit hochgradigen Gliomen.

Makary *et al.* [72] erreichten mit einem Niedrigfeldgerät eine deutliche Verlängerung der Zeitspanne zwischen der ersten und der zweiten (Rezidiv-)Operation sowie eine etwas geringere Rate an Komplikationen gegenüber konventioneller Operation. Bei der Interpretation dieser Ergebnisse muss beachtet werden, dass die Studie retrospektiv erfolgte. Während der Studiendauer veränderte sich die Nutzung: das intraoperative MRT wurde zunehmend nur noch für niedriggradige Gliome verwendet. Diese haben eine wesentlich bessere Prognose als hochgradige Gliome. Es bleibt unklar, inwieweit die beobachteten Vorteile aus diesen Unterschieden in der zugrundeliegenden Erkrankung hervorgehen. Insgesamt schließen die Autoren, dass gegenwärtig (2011) die Installation von intraoperativen Niedrigfeldgeräten *nicht* gerechtfertigt sei. Senft *et al.* [76] berichten, bei Verwendung eines Niedrigfeldgerätes im Median eine vollständige Resektion (vgl. unten) zu erreichen, während konventionell im Median 0,03 ml kontrastmittelaufnehmendes Resttumorvolumen verblieben. Unklar ist, inwieweit das einen Vorteil in der Lebensqualität und in der verbleibenden Überlebensdauer bewirkt. 0,03 ml entsprechen nach der oben gegebenen Abschätzung etwa  $10^7$  bis  $10^8$  Zellen. Damit ist bereits bei der konventionellen Resektion das im Median verbleibende kontrastmittelaufnehmende Tumolvolumen um Größenordnungen kleiner als die von Moskopp und Wassmann [27] angegebenen  $10^{10}$  Zellen, die nach einer im MRT vollständigen Resektion im Gehirn verbleiben.

Senft *et al.* [76] geben die Kosten für ein intraoperativ nutzbares MRT-Gerät mit 3–8 Millionen US\$ an, andere Quellen sprechen von 2 Mio. US\$ für das eigentliche Gerät und den Operationstisch [77], oder auch von Gesamtkosten in Höhe von 12,5 Mio. AU\$ (knapp 10 Mio. €) [78]. Die Hamburger Heidberg-Klinik gibt die Kosten pro Operation in der Größenordnung von 1000 € an [79]. Demgegenüber berichten Makary *et al.* [72] über Kosten von 13 500–45 000 US\$ pro Operation mit einem Niedrigfeld-Gerät (Anschaffungskosten des Geräts: 1–1,6 Mio. US\$). Die Kosten stiegen über den Beobachtungszeitraum stark an, da die Auslastung des Geräts *sank*.

Operationen mit intraoperativem MRT dauern im Mittel um die 7 h. Das ist 1 bis 2 h länger als konventionelle Operationen [76, 80]. Die verlängerte Operationsdauer führte laut Archer *et al.* [80] jedoch nicht zu zusätzlichen Komplikationen.

**Vollständige Resektion:** Die vorgestellten Studien zu neuen intraoperativen Diagnosemethoden beurteilen das verbleibende Tumolvolumen im postoperativen MRT mit dem Surrogatmarker „Kontrastmittel-aufnehmendes Volumen“. Das führt zu einer systematisch zu optimistischen Einschätzung der Methoden, bei denen Medikamente oder das Kontrastmittel die Blut-Hirn-Schranke durchschreiten müssen (vgl. [74]). Letztlich müsste das im Patienten verbliebene Tumolvolumen im Vergleich mit dem im MRT nicht erkennbaren Tumolvolumen, was nach der obenstehenden Abschätzung in der Größenordnung von einigen  $\text{cm}^3$  liegt, diskutiert werden und den Studien sollten Langzeitbeobachtungen zur direkten Messung der erzielten Überlebenszeit (wie in [70]) folgen. Außer ALA, MRT- und CT-Kontrastmitteln sind auch andere neuartige Kontrastmittel betroffen, wie zum Beispiel die von Kircher *et al.* [81] beschriebenen Nanopartikel, die gleichzeitig als Kontrastmittel für MRT, photoakustische Bildgebung und Oberflächenverstärkte Raman-Spektroskopie (engl. *Surface Enhanced Raman Spectroscopy*, SERS) dienen.

Da die Raman-Spektroskopie eine markerfreie Messmethode ist, ist sie von diesem Problem nicht betroffen.

### 3 Schwingungsspektroskopie

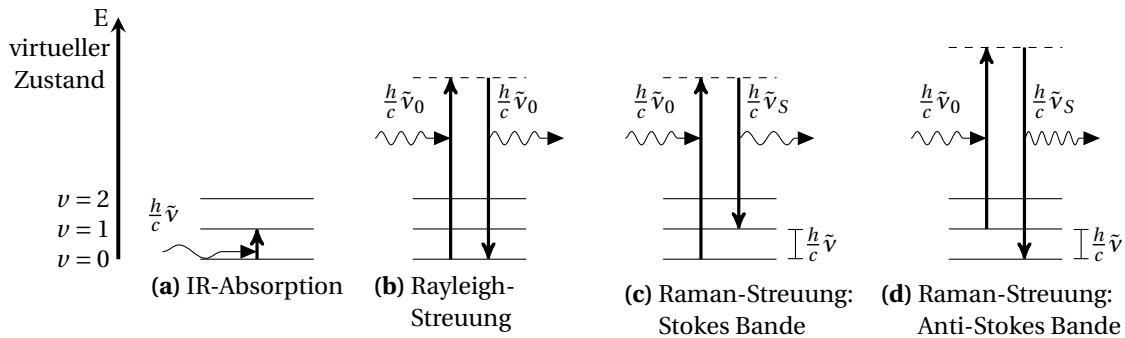
Die Atome eines Moleküls schwingen relativ zueinander. Alle möglichen Schwingungen lassen sich auf die sogenannten Normalschwingungen oder Normalmoden zurückführen, bei denen jeweils alle Atome mit der gleichen Frequenz und in Phase schwingen. Ein Molekül aus  $n$  Atomen hat  $3n - 6$  (bei linearen Molekülen  $3n - 5$ ) verschiedene Normalmoden. Die zur Anregung eines Moleküls um einen Schwingungszustand erforderliche Energie liegt in der Größenordnung unterhalb von  $50 \frac{\text{kJ}}{\text{mol}}$ . Das entspricht elektromagnetischer Strahlung mit Wellenlängen  $\lambda > 2,5 \mu\text{m}$  oder Wellenzahlen  $\tilde{\nu} = \frac{1}{\lambda} < 4000 \text{ cm}^{-1}$ . Die genaue Anregungsenergie für die einzelnen Schwingungen ist vom gesamten Molekül und seiner Umgebung abhängig. Daher ist das Schwingungsspektrum einer Substanz für diese charakteristisch. Wie ein Mensch durch seinen Fingerabdruck identifiziert werden kann, kann auch eine Substanz durch ihr Schwingungsspektrum mit Hilfe einer Spektrenbibliothek identifiziert werden.

Das Schwingungsspektrum wird durch Wechselwirkungen zwischen den Molekülen beeinflusst. Für die Identifikation oder Quantifizierung von Substanzen reichen Reinsubstanzspektren daher nicht aus. Stattdessen werden Spektren der entsprechenden chemischen Spezies, also der Reinsubstanz in einer geeigneten Matrix und im selben Aggregatzustand benötigt. Das Spektrum einer Mischung von chemischen Spezies setzt sich aus den Spektren der einzelnen Spezies in der Mischung zusammen. Daher kann auch eine definierte Mischung anhand ihres Schwingungsspektrums identifiziert werden.

An verschiedenen Schwingungen sind die einzelnen Atome des Moleküls in unterschiedlichem Maße beteiligt. Auch wenn bei einer Normalschwingung alle Atome in Phase schwingen, so führen sie doch Bewegungen mit unterschiedlicher Amplitude aus. Große Unterschiede in der Bindungsstärke (Mehrfachbindungen) und stark unterschiedliche Atommassen (Wasserstoff oder Halogene in einem organischen Molekül) führen zu einer mechanischen Entkopplung der Schwingungen dieser Atomgruppen vom Rest des Moleküls. Dadurch ist die Anregungsenergie dieser Schwingungen nahezu unabhängig vom Rest des Moleküls. Für viele funktionelle Gruppen können daher Schwingungsenergien tabelliert werden, die sogenannten charakteristischen Banden.

Schwingungsspektren können also nicht nur zur Identifikation einer Substanz verwendet werden, sie geben auch eine Einordnung in die Substanzklassen. Damit sind auch Summenparameter (Gesamtprotein, Gesamtlipide, vorherrschende Proteinfaltung) recht einfach zugänglich.

Die Anregungsenergie von Schwingungen, die mit einer Änderung des Dipolmoments  $\mu$  des Moleküls einhergehen, kann direkt im mittleren Infrarot gemessen werden, also im Wellenlängenbereich von ca.  $\lambda = 2,5 \mu\text{m}$  bis  $25 \mu\text{m}$ , oder  $\tilde{\nu} = \frac{1}{\lambda} = 400$  bis  $4000 \text{ cm}^{-1}$ . Nur Licht, dessen Wellenlänge der Anregungsenergie einer Schwingung entspricht, wird absorbiert (Abbildung 3.1a). Daher treten auch nur bei Wellenlängen, die der Anregungsenergie von Schwingungen der Moleküle in der Probe entsprechen, Absorptionsbanden auf. Infrarot- (IR)-Spektren sind zum Beispiel als Transmissionsspektren messbar. Unter der Annahme, dass Streu- und Reflexionsverluste vernachlässigbar sind, gilt das Lam-



**Abbildung 3.1** Energieschema der IR-Absorption, Rayleigh- und Raman-Streuung.

bert-Beersche Gesetz: die Extinktion ist der Konzentration der betrachteten Spezies proportional. Damit kann das Extinktionsspektrum einer Substanz als Linearkombination der Extinktionsspektren der einzelnen Spezies, gewichtet mit ihren Konzentrationen, beschrieben werden.

Schwingungen, bei denen sich die Polarisierbarkeit  $\alpha$  der Elektronenhülle des Moleküls ändert, können mit der Raman-Spektroskopie gemessen werden. Elektromagnetische Wellen können mit der Elektronenhülle eines Moleküls interagieren und sie polarisieren. Die Elektronenhülle wirkt dann als Hertzscher Dipol und sendet ihrerseits elektromagnetische Wellen aus. Dabei geht die Richtung der einfallenden Welle verloren, das Licht wird gestreut. Dieser Streuprozess ist instantan. Die Sendeleistung  $P_S$  dieses Hertzschens Dipols ist [82]:

$$P_S \sim P_0 v_S^4 \left( \frac{\partial \alpha}{\partial q} \right)_{q=q_0}^2 \quad (3.1)$$

und ist also proportional zur Anregungsleistung  $P_0$  und zum Quadrat der Veränderung der Polarisierbarkeit  $\alpha$  aufgrund der Schwingung (Normalkoordinate  $q$ ) um die Ruhelage  $q_0$ . Sie wächst weiterhin mit der vierten Potenz der Frequenz des Streulichts  $v_S$ .

Weit überwiegend entsteht Streustrahlung derselben Wellenlänge. Diese elastische Streuung von Photonen an Molekülen wird als Rayleigh-Streuung bezeichnet (Abb. 3.1b). Die Intensität der Rayleigh-Streustrahlung liegt etwa 5 Größenordnungen unter der Intensität der Anregungsstrahlung. Relaxiert das Molekül in einen anderen Schwingungszustand, so resultiert eine Energiedifferenz, die Streustrahlung hat eine andere Wellenlänge Abbildung 3.1c und (d). Die Energiedifferenz zwischen Anregungs- und Streustrahlung entspricht dabei der Anregungsenergie der Schwingung. Aus historischen Gründen wird in der Schwingungsspektroskopie statt der Energie  $E$  oder der Frequenz  $\nu$  die Wellenzahl  $\tilde{\nu} = \frac{\nu}{c} = \frac{1}{\lambda}$  benutzt, die in  $\text{cm}^{-1}$  angegeben wird. Daher werden Raman-Spektren als Intensität  $I$  der Streustrahlung über der Wellenzahl des Schwingungsübergangs  $\tilde{\nu}$  aufgetragen. Soll im Folgenden die Differenz zwischen den Wellenzahlen von Anregungslaser und beobachteter Streustrahlung betont werden, so wird der Differenzoperator verwendet:  $\Delta \tilde{\nu} = \tilde{\nu}_0 - \tilde{\nu}_S$  [83]. Raman-gestreute Photonen werden mit einer ungefähr um weitere fünf Größenordnungen geringeren Intensität beobachtet. Als Faustregel kann man also sagen, dass die Raman-Streustrahlung acht bis zehn Größenordnungen schwächer als die Anregungsstrahlung ist [82, 84]. In Abbildung 3.1c wird eine Schwingung angeregt. Es

entsteht längerwellige Streustrahlung (*Stokes-Strahlung*). Abbildung 3.1d verdeutlicht, dass auch Streustrahlung mit kürzerer als der Anregungswellenlänge entsteht. Hier wird ein Molekül aus einem angeregten Schwingungszustand auf das virtuelle Energieniveau gebracht und relaxiert in den Schwingungsgrundzustand. Diese höherenergetische Raman-Streustrahlung heißt *Anti-Stokes-Strahlung*.

Gleichung 3.1 betrachtet zunächst nur ein einzelnes Molekül. Die Gesamtleistung der Raman-Streustrahlung einer makroskopischen Probe ergibt sich als Summe der Sendeleistungen aller Teilchen der entsprechenden Spezies im Anregungsvolumen. Dabei muss berücksichtigt werden, dass die Anregungsleistung in Abhängigkeit vom Ort variiert. Für Proben mit vergleichbaren optischen Eigenschaften und homogener räumlicher Verteilung der betrachteten Spezies ist die beobachtete Raman-Intensität aber proportional zur Konzentration (die Proportionalitätskonstante ist in dieser Formulierung gerätespezifisch). Damit kann auch das Raman-Spektrum einer Substanz als Linearkombination der Raman-Spektren  $I_i(\tilde{\nu})$  der einzelnen Spezies  $i$  mit ihren Konzentrationen  $c_i$  beschrieben werden:

$$I(\tilde{\nu}) = \sum_{\forall i} c_i I_i(\tilde{\nu}) \quad (3.2)$$

Aus den Unterschieden im physikalischen Prinzip hinter der Infrarot- (IR-) und Raman-Spektroskopie folgen unterschiedliche Auswahlregeln:  $\left(\frac{\partial\mu}{\partial q}\right)_{q=0} \neq 0$  für IR beziehungsweise  $\left(\frac{\partial\alpha}{\partial q}\right)_{q=0} \neq 0$  für die Ramanstreuung. Das führt dazu, dass im Extremfall punktsymmetrischer Moleküle die einzelnen Normalmoden entweder IR- oder Raman-aktiv sind, aber nicht beides. Raman- und IR-Spektroskopie liefern daher komplementäre Informationen. Auch für komplexe Moleküle wie sie in biologischen Proben vorherrschen, ist dieselbe Bande meist in einer der beiden Spektroskopiearten intensiver.

## 3.1 Praktische Aspekte

Sowohl Raman- als auch IR-Spektren enthalten trotz der unterschiedlichen Auswahlregeln jeweils eine genügend große Zahl an Banden, um die Identität einer Substanz oder die Substanzklasse zu bestimmen. Ob Raman- oder IR-Spektroskopie geeigneter ist, hängt von der genauen Fragestellung ab.

IR-Spektren werden mit Michelson-Interferometern gemessen (sogenannte Fourier-Transform- oder FT-IR-Spektroskopie). Die interferometrische Messung erlaubt eine hohe Wellenzahlgenauigkeit, ohne dass ein Spalt zu Intensitätsverlusten führt. Transmissionsmessungen von trockenen Gewebe-Dünnschnitten liefern in kurzer Zeit FTIR-Spektren mit sehr gutem Signal-Rausch-Verhältnis. Mit Array-Detektoren werden dabei direkt orts aufgelöste sogenannte IR-Images gemessen [85].

Für eine Anwendung *in vivo* bietet die Raman-Spektroskopie die entscheidenden Vorteile. Wässrige oder feuchte Proben können problemlos vermessen werden und faseropische Sonden ermöglichen räumliche Flexibilität.

Die interferometrische Messtechnik findet in der Raman-Spektroskopie nur bei Anregung mit Wellenlängen oberhalb von 830 nm (typisch: 1064 nm) Anwendung. Für Anre-

gungswellenlängen bis 830 nm werden Gitterspektrometer eingesetzt, da der Spektralbereich unterhalb von 1050 nm mit Si-basierten CCD- (*Charge Coupled Device*) Detektoren gemessen werden kann. Das entspricht einem Wellenzahlunterschied von  $3200\text{ cm}^{-1}$  ausgehend von  $\lambda_0 = 785\text{ nm}$  beziehungsweise von  $2500\text{ cm}^{-1}$  bei  $\lambda_0 = 830\text{ nm}$ . Hierbei projiziert ein Gitter die einzelnen Wellenlängen entlang der längeren Richtung der CCD („nebeneinander“). Die andere, kürzere Richtung („übereinander“) bildet zunächst den Spalt ab und die zusammengehörenden Pixel werden aufsummiert. Diese Richtung kann jedoch auch genutzt werden, um räumliche Informationen (mehrere Fasern oder auch Linienscans) simultan aufzunehmen. Außerdem können mehrere Spektralbereiche übereinander projiziert werden, so dass höher spektral aufgelöste Spektren simultan gemessen werden können [82, 85].

Mit der Raman-Spektroskopie kann bei Anregung mit sichtbarem oder ultraviolettem Licht eine bessere laterale Auflösung als bei der IR-Spektroskopie erreicht werden, da die Auflösungsgrenze proportional zur Wellenlänge ist. Für die in dieser Arbeit untersuchten Fragestellungen spielt dies allerdings keine Rolle, da die Neurochirurgen größere Messvolumina (ca. 1 oder sogar 3 mm Durchmesser, gern einige Millimeter in die Tiefe) bevorzugen würden [64, 86].

**Wahl der Anregungswellenlänge für die Raman-Spektroskopie:** Die Intensität der Rayleigh- und Raman-Streustrahlung steigt mit der vierten Potenz der Frequenz der Streustrahlung (Gl. 3.1). Daher wird das Raman-Spektrum umso intensiver, je kürzer die Anregungswellenlänge ist. Andererseits fluoreszieren viele Biomoleküle. Die Fluoreszenz-Strahlung ist etwa 6 bis 8 Größenordnungen intensiver als die Raman-Streustrahlung [82], so dass Fluoreszenz eine Raman-Messung empfindlich stört oder ganz unmöglich macht. Bei biologischen Proben kann Fluoreszenz durch Verwendung einer langwelligeren Anregung in der Regel verringert werden.

Die  $\nu_S^4$ -Abhängigkeit der Raman-Intensität (Gl. 3.1) bedeutet auch, dass sich die Intensitäten der einzelnen Raman-Banden nicht nur absolut mit der Anregungswellenlänge ändern, sondern auch die Intensitäten der einzelnen Normalmoden zueinander. Außerdem kann Resonanzverstärkung auftreten. In Extremfällen treten Verstärkungsfaktoren von bis zu 6 Größenordnungen auf [82]. Die Resonanzverstärkung betrifft unterschiedliche Normalmoden unterschiedlich stark: die Verstärkung ist für Schwingungen, die den Molekülteil mit dem elektronisch angeregten Chromophor betreffen, größer als für Schwingungen, die hauptsächlich weit entfernt vom Chromophor lokalisiert sind. Daher kann sich das beobachtete Raman-Spektrum abhängig von der Anregungswellenlänge sehr stark ändern. Bei 785 nm Anregungswellenlänge tritt eine solche Resonanzverstärkung zum Beispiel für Hämoglobin oder Melanin auf. Bei einer Anregung mit 514 nm erfahren Cytochrom C und Carotine Resonanzverstärkung. Praktisch wird für eine Anwendung daher die Anregungswellenlänge in aller Regel fest gewählt.

Das Spektrum der Fluoreszenz-Emission ist konstant gegenüber kleinen Änderungen der Anregungswellenlänge. Raman-Banden treten demgegenüber immer bei einer bestimmten Energiedifferenz von der Anregungswellenlänge auf. Daher lassen sich Fluoreszenzemission und andere, bezüglich der Wellenlänge konstante, Beiträge zum gemessenen Spektrum vom Raman-Spektrum trennen, indem die Anregungswellenlänge verändert wird [87, 88]. Allerdings kann auch elastisch gestreutes Licht des Anregungslasers

einen Untergrund erzeugen, der sich nicht durch Modulation der Anregungswellenlänge abtrennen lässt. Ein solcher Streulicht-Untergrund tritt besonders bei stark streuenden Proben und Spektrometern mit kleiner Blendenzahl und zu geringer optischer Dichte des Interferenzfilters auf [82, 89].

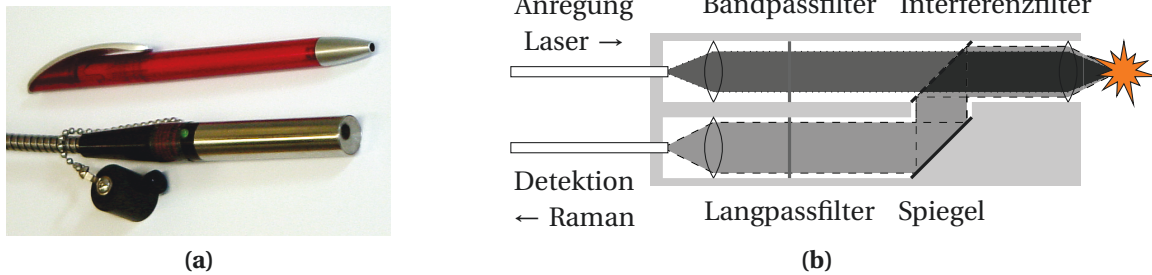
Aus statistischer Sicht kann die Messung der Lichtintensität als Poisson-Prozess beschrieben werden (sogenannter *shot noise*). Das bedeutet, dass die Varianz des Signals (Rauschen) der mittleren Intensität entspricht [90]. Das Rauschen auf dem Messsignal wächst also mit der Gesamtintensität. Während der systematische Beitrag des Untergrundsignals korrigiert werden kann, verbleibt das zusätzliche Rauschen auf dem Raman-Spektrum. Letztlich sollte die Anregungswellenlänge also so gewählt werden, dass möglichst wenig Untergrundsignal entsteht. Typische Lichtquellen für die Raman-Spektroskopie von biologischen Proben sind Diodenlaser mit  $\lambda_0 = 785 \text{ nm}$  oder  $830 \text{ nm}$ . Das führt zu Raman-Spektren im nahen Infrarot. Auch wenn in diesem Spektralbereich (und mit dieser langwelligen Anregung) verglichen mit kürzerwelliger Anregung nur wenig Fluoreszenz auftritt, so tragen Biomoleküle wie NAD(P)H, Flavine (FAD) oder Porphyrine zu einem Fluoreszenzuntergrund bei. Shao, Zheng und Huang [91] nutzen diese Autofluoreszenz zur Erkennung von Kolonkarzinomen und -vorstufen *in vivo*. Dort sind auch Emissionsspektren verschiedener Biomoleküle im Spektralbereich von 810 bis 1000 nm ( $\Delta\tilde{\nu} = 390 - 2740 \text{ cm}^{-1}$  gegenüber Anregung mit 785 nm) gegeben.

Für biologische Proben ist weiterhin von Bedeutung, dass Wasser in diesem Spektralbereich nur schwach absorbiert. Allerdings liegt bei 970 nm ( $\Delta\tilde{\nu} = 2430$  gegenüber einer Anregung mit 785 nm) eine Absorptionsbande mit einem Extinktionskoeffizienten  $\epsilon_{970} \approx 4 \cdot 10^{-3} \text{ mol}^{-1} \text{ cm}^{-1}$ . Die Extinktion von 1 mm reinem Wasser ist also ungefähr 0,022 [92]. Die in dieser Arbeit untersuchten Proben zeigen jedoch im Spektralbereich von  $\tilde{\nu} = 1800 - 2800 \text{ cm}^{-1}$  keine nutzbaren Banden, so dass bei Anregung mit 785 nm der Spektralbereich um das Absorptionsmaximum zwischen ca. 915 und 1005 nm nicht benötigt wird. Bei einer optischen Weglänge von 1 mm in Hirngewebe werden im betrachteten Spektralbereich höchstens etwa 3,5 % des Raman-Streulichts absorbiert [92]. Damit ist die Absorption im Hinblick auf das geringere Messsignal unkritisch.

Absorption spielt im Hinblick auf *in vivo* Anwendungen allerdings noch eine zweite wichtige Rolle: die absorbierte Energie führt zu einer Erwärmung des untersuchten Gewebes. Im Sinne der Anwendungssicherheit der Raman-Sonden muss nachgewiesen werden, dass das Gewebe durch die Messung nicht thermisch geschädigt wird. Hierfür ist der Anregungslaser verantwortlich. Da die Raman-Streustrahlung viele Größenordnungen schwächer ist, trägt sie nicht nennenswert zur Erwärmung bei. Auch in dieser Hinsicht sind Wasser bzw. weiße und graue Substanz unkritisch. Probleme können jedoch bei geronnenem Blut und stark pigmentierten Tumoren, zum Beispiel schwarzen Melanomen auftreten.

**Faseroptische Sonden:** Im sichtbaren Spektralbereich und nahen Infrarot können Faseroptiken zur Raman-Spektroskopie verwendet werden. Faseroptische Sonden sind besonders wegen der möglichen Anwendung in der Diagnostik während einer (minimal invasiven oder endoskopischen) Operation interessant. Einen Überblick über Sondenkonfigurationen und Materialien im Hinblick auf *in vivo* Anwendungen geben Latka *et al.* [93] und Utzinger und Richards-Kortum [94].





**Abbildung 3.2** Die faseroptische Sonde. (a) Foto mit Größenvergleich und (b) optischer Aufbau [95].

Abbildung 3.2 zeigt den Aufbau der verwendeten Sonde. Die Filter im Sondenkopf verhindern, dass das Raman-Spektrum der Faser das Probenspektrum überlagert. Am Ende der Anregungsfaser wird die bereits entstandene Raman-Streustrahlung mit einem Bandpass-Filter entfernt. Vor dem Eintritt des Streulichts in die Detektionsfaser wird die Rayleigh-Streustrahlung mit einem Interferenzfilter entfernt, da sonst das Raman-Spektrum der Detektionsfaser das Probenspektrum überlagern würde.

Bei ungefilterten Sonden muss das Raman-Signal der optischen Faser vom Rohsignal subtrahiert werden. Auch hier gilt, dass das Rauschen auf dem Spektrum von der Intensität des gemessenen Rohsignals abhängt. Gefilterte Sonden führen also gegenüber einer nachträglichen Korrektur im Rahmen der Datenvorbehandlung zu einem wesentlich besseren Signal-Rausch-Verhältnis. Eine Ausnahme stellt der Spektralbereich oberhalb von  $2400\text{ cm}^{-1}$  dar. Dieser Spektralbereich enthält keine störenden Raman-Signale der Faser, so dass eine Diagnostik auf Basis ungefilterter Sonden unter Verwendung der C – H-Valenzschwingungen vorgeschlagen wurde. Dies demonstrieren Koljenović *et al.* [96] neben anderen Tumoren auch an der Unterscheidung zwischen vitalem und nekrotischem Gewebe eines Glioblastoms. In den Kapiteln 16.2.4 und 17.2.2 wird daher untersucht, ob ein Grading der Astrozytomgewebe bzw. eine Unterscheidung zwischen Astrozytomen und Lymphomen ausschließlich anhand der C – H-Streckschwingungen möglich ist, so dass ungefilterte Raman-Sonden verwenden könnten.

Zusätzlich zu den Filtern enthält die verwendete Sonde eine Frontlinse, so dass der Arbeitsabstand 5 mm beträgt. Die Linse erhöht außerdem die Sammeleffizienz der Sonde. Die Größe der in Abbildung 3.2 gezeigten Sonde ist im Wesentlichen durch die Filter und Linsen bestimmt. Kleinere Bauarten sind möglich, stehen bislang aber nur in Kleinstserien kommerziell zur Verfügung oder sind in Entwicklung. Besonders vielversprechend ist dabei die Möglichkeit, Filter in Form von Faser-Bragg-Gittern direkt in die Fasern einzuschreiben. Obwohl dies nur für *single-mode* Fasern möglich ist, deren Sammeleffizienz gering ist, können mit Bündeln von *single-mode* Fasern Sammeleffizienzen wie für *multi-mode* Fasern realisiert werden [CB9, CB10]. Auch Linsen können in Fasern realisiert werden, dazu werden Fasern mit radialem Gradienten der Brechzahl (engl. *Gradient Refractive INdex, GRIN*) verwendet.

Verglichen mit spektroskopischen Messungen unter dem Mikroskop ist das mit einer faseroptischen Sonde gemessene Probenvolumen wesentlich größer. Mikrospektroskopische Messungen nutzen meist Objektive mit einer hohen numerischen Apertur und

einer hohen Vergrößerung. Das Messvolumen liegt dann in der Größenordnung von wenigen  $\mu\text{m}$  und subzelluläre Auflösung ist möglich. Demgegenüber haben faseroptische Sonden eine wesentlich geringere numerische Apertur und die Anregung erfolgt mit einem divergenten Strahl (ohne Frontlinse) bzw. mit einem wesentlich größeren Fokusbereich. So hat die in dieser Arbeit verwendete Sonde (Abb. 3.2a) eine numerische Apertur von 0,2 und einen Fokusbereich von etwa  $60\ \mu\text{m}$ . In Hirngewebe ergibt sich ein Messvolumen in der Größenordnung von  $10^6\ \mu\text{m}^3$  oder  $10^3$  Zellen, vgl. Kap. 10).

Die Sammeleffizienz einer faseroptischen Sonde ist aufgrund der kleineren numerischen Apertur wesentlich geringer als die eines mikrospektroskopischen Aufbaus. Für Messungen *in vivo* während einer Hirntumoroperation kann allerdings kein normales Mikroskopobjektiv verwendet werden. Operationsmikroskope haben einen extrem langen Arbeitsabstand und eine noch geringere numerische Apertur.

Das Gewebe darf durch den Laser weder thermisch noch photochemisch geschädigt werden. Da für eine Anwendung während der Operation die Messzeit möglichst kurz sein soll, wird eine Anregungsleistung angestrebt, die so hoch wie für das Gewebe sicher möglich ist. Wird die Anregungsleistung sollte auf ein größeres Gewebevolumen verteilt, so kann die Gesamtanregungsleistung auch erhöht werden. Damit sollte die räumliche Auflösung der Sonde der Arbeitsgenauigkeit der Chirurgen angepasst, aber nicht höher sein.

## 3.2 Raman- und IR-Spektroskopie von primären Hirntumoren

Biologische Gewebe sind sehr komplexe, aber wohldefinierte Substanzgemische: die Konzentrationen der einzelnen chemischen Spezies in einem Gewebe können nicht beliebig schwanken. Unterschiede in der biochemischen Zusammensetzung führen zu unterschiedlichen Raman-Spektren. Daher können biologische Gewebe anhand ihrer Raman-Spektren identifiziert werden. Viele Krankheiten gehen mit einer Veränderung der biochemischen Zusammensetzung der betroffenen Gewebe einher und sind prinzipiell einer schwingungsspektroskopischen Diagnostik zugänglich. Zum Beispiel enthalten Astrozytome weniger Lipide als normales Hirngewebe. Das zeigen auch die Schwingungsspektren [97]. Im Unterschied zur Identifikation einer reinen Substanz müssen bei biologischen Proben aber krankheitsbedingte Veränderungen im Vergleich zur natürlichen Variation betrachtet werden. Das leistet die statistische (chemometrische) Auswertung.

Aufgrund der charakteristischen Banden können die Schwingungsspektren von Geweben auch im Hinblick auf die Substanzklassen interpretiert werden. So können recht einfach Aussagen über Unterschiede in den Lipid- oder Proteinkonzentrationen, oder auch über die vorherrschende Sekundärstruktur der Proteine getroffen werden.

Sowohl FTIR- als auch Raman-Spektroskopie sind etablierte Methoden in der chemischen Analytik. In den letzten fünfzehn Jahren wurden beide Methoden verstärkt auf medizinische Fragestellungen hin angewendet [CB4, CB5, 98, 99]. Die Referenzdiagnostik erfolgt an konventionell gefärbten Parallelschnitten. Daher bietet es sich an, zunächst (ungefärbte) Gefrierschnitte zu untersuchen. Getrocknete Gefrierschnitte sind sehr gut für die IR-Spektroskopie geeignet. Dementsprechend gibt es wesentlich mehr Untersuchungen [98] mit größeren Patientenzahlen [100–102] an Hirntumoren mittels Infrarotspektroskopie als mittels Raman-Spektroskopie.

Eine Operationsbegleitende *Attenuated Total Reflection* (ATR)-IR-Spektroskopie ist anhand von Quetschpräparaten beschrieben [103, 104]. Dabei blieb jedoch das Problem ungelöst, dass keine orts aufgelöste Referenzdiagnose verfügbar ist. Folglich werden in den genannten Artikeln unüberwachte, deskriptive Methoden angewendet. Für eine Klassifikation, die die Tumorränder unbekannter Proben anzeigen soll, sind jedoch orts aufgelöste Referenzdiagnosen aufgrund der erheblichen Heterogenität von Gliomproben erforderlich, wie in Kapitel 14 dargelegt wird.

Auch Gajjar *et al.* [102] nutzen ATR-IR-Spektroskopie, allerdings an Gewebeschnitten, die aus in Paraffin eingebetteten Proben hergestellt wurden. Da große Sammlungen von in Paraffin eingebetteten Proben existieren, können größere Patientenzahlen recht einfach realisiert werden. Allerdings werden sowohl bei der Einbettung als auch beim Auswaschen des Paraffins aus dem Schnitt vor der Spektroskopie unpolare Lösungsmittel verwendet, so dass erhebliche Anteile der Lipide aus der Probe ausgewaschen werden. Das ist für die Untersuchung von Hirntumoren besonders problematisch, da Lipide ein Hauptbestandteil von Hirngewebe sind [105], und für Hirntumore sowohl Änderungen im Gesamtlipidgehalt als auch in der Zusammensetzung der Lipide beschrieben sind [CB12, 97, 100, 105–107]. Diese Einschränkung wird besonders an der starken Überlappung der beobachteten Lipid-zu-Protein-Intensitätsverhältnisse von Gajjar *et al.* [102] im Vergleich zu den in [101] gezeigten Daten für Gefrierschnitte, die nicht mit Lösungsmitteln oder Einbettmedien behandelt wurden, deutlich.

Aber auch Raman-Spektroskopie kann an Gewebeschnitten durchgeführt werden. Gajjar *et al.* [102] haben parallel zu den ATR-IR-Messungen auch Raman-Messungen an den Gewebeschnitten untersucht, die allerdings derselben problematischen Präparation unterzogen wurden. Lineare Diskriminanzanalyse (LDA) auf der Basis der zweiten Hauptkomponente von Clustermittelwertspektren trennt Tumorgewebe von nekrotischem Gewebe [108]. Koljenović *et al.* [96] schlagen vor, ausschließlich die C – H-Streckschwingungen zur Klassifikation zu benutzen. Bergner [109], [110] und [111] untersucht primäre Hirntumore von 5 Patienten mittels Raman-Mikroskopie, sowie sekundäre Hirntumore (Metastasen) an 22 Patienten [110, 112] (auch IR-Imaging). Leslie *et al.* [113] beschreiben eine Serie von binären Klassifikationsmodellen an Hirntumorproben von 44 Patienten (Pädiatrie). Die Diplomarbeit von Marx [114] behandelt Raman-Spektroskopie an den auf CaF<sub>2</sub> präparierten Parallelschnitten zu den hier untersuchten Bulkproben (Kap. 9.1).

Eine Stärke der Schwingungsspektroskopie sind Aussagen über die Biochemie der untersuchten Proben. Krafft *et al.* [101] untersucht IR-Spektren von 56 Gliomproben (7 Astrozytome °II, 9 Astrozytome °III, 40 Glioblastome). In Astrozytomen und Glioblastomen wurden niedrigere Lipidgehalte (bezogen auf den Proteingehalt) gefunden als in normalem Hirngewebe, und auch die Zusammensetzung der Lipidfraktion ist unterschiedlich [97, 101, 105]. Unterschiedliche Lipidgehalte in verschiedenen Hirnregionen und in normalem Gewebe beziehungsweise einem Astrozytom in Mäusen wurden mittels kohärenter anti-Stokes Raman-Spektroskopie (CARS) anhand der C – H-Streckschwingungen dargestellt [115]. CARS an humanen Biopsieproben von Hirntumoren wurde von Meyer *et al.* [116] vorgestellt. Meyer *et al.* [117] vergleicht nichtlineare optische Bildgebung an Schweinehirn mit Raman-Mapping und IR-Imaging. Noreen *et al.* [118, 119] untersuchen Proteinfaltung und Kollagengehalt anhand von IR-Images im Mausmodell.

Banerjee und Zhang [120] beschreiben die Raman-Spektren einzelner Astrozyten und Astrozytom-Zellen in Zellkultur. Das Wachstum von Glioblastomen in Kollagen wurde

### 3 Schwingungsspektroskopie

mit Hilfe von konfokaler Mikroskopie und CARS untersucht [121]. Ein Glioblastom-Tiermodell (C6) wurde in einer Reihe von Veröffentlichungen mittels IR- [106, 122–124] und Raman-Spektroskopie [125, 126] untersucht. Stelling *et al.* [104] untersuchen ATR-IR-Spektren von Glioblastom-Zellkultur, Mausmodell und humane Proben (alles feuchte Proben) auf kollagenartige Proteine. Wehbe *et al.* [127] untersuchen die Blutgefäße in hochgradigen Gliomen, sowohl im Mausmodell (U87) als auch in humanen Proben. Studien an durch Implantation von Krebszelllinien induzierten Gliomen im Tiermodell untersuchen ausschließlich hochgradige Tumore, bei der Zelllinie U87 [128] handelt es sich um ein humanes Glioblastom, C6 ist ein chemisch induziertes Glioblastommodell in der Ratte [129, 130].

Schockgefrorene und für die Messung wieder aufgetaute Proben gleichen in ihren Raman-Spektren frischen Proben [131]. Allerdings dominiert Blut die Raman-Spektren einer frischen und nativen Hirntumorprobe weniger stark als die einer Probe, die einen Tag bei  $-80^{\circ}\text{C}$  gelagert war [132]. Messungen *in vivo* in der Aorta von Kaninchen wurden nicht durch die spektrale Signatur von Hämoglobin gestört [133].

Raman-Spektren von Hirntumor-Bulkproben in physiologischer Kochsalzlösung mit Phosphatpuffer (engl. *Phosphate Buffered Saline, PBS*) von fünf Patienten (davon ein Astrozytom und zwei Glioblastome) wurden hinsichtlich ihres Lipid-, Protein- und Wassergehaltes untersucht [132] während Bergner [109] bei einer frischen humanen Astrozytom-Bulkprobe keine verwertbaren Raman-Signale erhielt. In beiden Fällen handelt es sich um beschreibende Untersuchungen, prädiktive Modelle wurden nicht erstellt.

Messungen *in vivo* am Menschen sind gegenwärtig noch nicht vertretbar. Kirsch *et al.* [134] demonstrieren aber die Raman-spektroskopische Erkennung von Metastasen im Mausmodell *in vivo* und Beljebbar *et al.* [126] Messungen von C6-Glioblastomen *in vivo* in der Ratte. In beiden Fällen handelt es sich aber nicht um ein Grading der Gewebe, sondern ausschließlich um die Unterscheidung von höchstgradigem Tumor von normalem Hirngewebe.

Zusammenfassend lässt sich sagen, dass die vorliegende Arbeit erstmals ein prädiktives Grading von Astrozytomgeweben auf der Basis von Raman-Spektren, die durch eine faseroptische Sonde aufgenommen wurden, an humanen Bulkproben vorstellt.

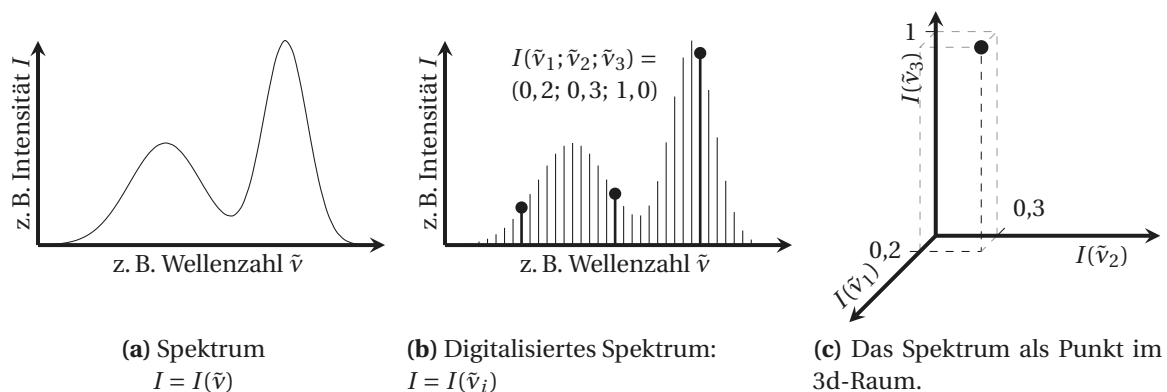
Kammer [135] unterscheidet anhand von IR-Spektren der Parallelschnitte zu den hier untersuchten Proben zwischen Astrozytomen und Lymphomen. Weitere Raman- oder IR-spektroskopische Untersuchungen von PZNSL sind in der Literatur nicht beschrieben. Einzig Bergner [109] untersucht neben anderen primären und sekundären Hirntumoren auch ein B-Zell-Lymphom, möglicherweise ein PZNSL. Es gibt aber einige schwingungsspektroskopische Studien von Lymphomen. Andrus und Strickland [136] untersuchten KBr-Tabletten mit Lymphom-Proben und korrelierten die Malignität mit DNA- und RNA-Signalen. Lymphozyten zweier Patienten mit akuter lymphatischer Leukämie wurden im Verlauf der Chemotherapie mittels IR-Spektroskopie charakterisiert und die Zellzahlen verfolgt [137, 138].

# 4 Chemometrische Datenanalyse

## 4.1 Spektren als multivariate Daten

Aus spektroskopischer Sicht ist die gemessene Raman-Intensität eine stetige Funktion der Wellenzahl des Schwingungsübergangs :  $I(\tilde{\nu})$ . Bereits bei der Messung mit der CCD-Kamera eines Gitter-Spektrometers wird die Wellenzahl-Achse jedoch diskretisiert, also als ein Vektor von Intensitäten, die zu bestimmten Wellenzahlen gehören, aufgenommen ( $I_i(\tilde{\nu}_i)$ ). Der Spektren-Vektor wird nun als ein Punkt in einem hochdimensionalen Raum beschrieben. Die Dimensionalität entspricht dabei der Anzahl der Elemente im Vektor, also zum Beispiel der Anzahl der Messkanäle des Raman-Spektrometers. Abbildung 4.1 veranschaulicht dies für ein Spektrum von nur 3 Messpunkten an verschiedenen Wellenzahlen, das als Punkt im 3-dimensionalen Raum dargestellt wird. Analog ist ein Vektor von  $p$  Elementen (ein Spektrum mit  $p$  Datenpunkten) ein Punkt im  $p$ -dimensionalen Raum. Der gesamte Datensatz wird dann als Punktwolke aufgefasst. Diese Sicht auf den Datensatz liegt den meisten multivariaten Verfahren zu Grunde, viele Modelle lassen sich so in einer Matrixschreibweise darstellen. Die Datenmatrix  $\mathbf{X}^{(n \times p)}$  hat  $n$  Zeilen, die jeweils ein Spektrum (oder allgemeiner: Objekt) enthalten. Die  $p$  Spalten sind die Variaten, also zum Beispiel Messgrößen wie  $I(\tilde{\nu}_1), \dots, I(\tilde{\nu}_p)$ .

Die Variaten werden in aller Regel von den multivariaten Verfahren als voneinander unabhängig und beliebig behandelt. Statt Intensitäten an verschiedenen Wellenlängen könnten auch Konzentration und pH-Wert aufgetragen sein. Gegebenenfalls können daher neben dem Spektrum weitere Variate vorliegen. Diese sind dann entweder aus den Messgrößen abgeleitet, oder sie stellen zusätzliche Informationen zu den Spektren dar, zum Beispiel die Gewebeart, von der das Spektrum gemessen wurde. Dies gibt Flexibilität: weitere Messgrößen (auch von anderen Methoden) können leicht in die Auswertung



**Abbildung 4.1** Ein Spektrum als Punkt im hochdimensionalen Raum. Das Spektrum (a) liegt im Rechner diskretisiert vor (b). Punkte: Ein Spektrum mit 3 Wellenlängen  $I(\tilde{\nu}_1; \tilde{\nu}_2; \tilde{\nu}_3) = (0,2; 0,3; 1,0)$ . Ein solcher Vektor mit 3 Elementen kann als Punkt im 3-dimensionalen Raum dargestellt werden (c).

einbezogen werden. Andererseits hat diese Flexibilität auch ihren Preis. Spezielle Eigenschaften spektroskopischer Datensätze können bei einer solchen Auswertung nicht oder nur sehr eingeschränkt genutzt werden.

Die Dimensionen eines spektroskopischen Datensatzes sind geordnet. Sie entstehen letztlich durch Diskretisierung einer physikalischen Größe wie der Wellenlänge. Aus spektroskopischer Sicht sollten Spektren so hoch aufgelöst gemessen werden, dass das als Vektor gespeicherte Spektrum eine gute Näherung der zugrunde liegenden stetigen Funktion ist. Dann sind benachbarte Datenpunkte des Spektrums einander ähnlich. Eine starke Korrelation benachbarter Variate ist ein Qualitätsmerkmal der Spektren, weil sie auf ein gutes Signal-Rausch-Verhältnis und eine gute Abbildung der stetigen Banden hinweist. Die Punktwolke eines solchen Datensatzes ist auf einen kleinen Teil des Datenraumes um die Raumdiagonale konzentriert.

### 4.2 Datenvorbehandlung

Zu Beginn der chemometrischen Modellierung werden die Rohdaten in der Regel einer sogenannten Vorbehandlung unterzogen, die die Unterschiede in Bezug auf die jeweilige Fragestellung der Auswertung hervorhebt und/oder störende Unterschiede anderer Ursache verringert. Die Grenzen zwischen Vorbehandlung und Modellierung sind fließend. Vorbehandlung und Modellierung müssen aufeinander abgestimmt sein und einer *gemeinsamen* Validierung unterzogen werden.

Viele Datenvorbehandlungsmethoden sollen störende physikalische Effekte im Datensatz korrigieren. Solchen Methoden liegen daher physikalische Modelle mit bestimmten Annahmen zugrunde, zum Beispiel die Entstehung einer Basislinie im Raman-Spektrum durch Ausläufer von Fluoreszenzbanden. Allerdings lässt sich der genaue Einfluss der betrachteten Effekte oft nicht ermitteln. Daher werden Näherungen verwendet. Die meisten Datenvorbehandlungsmethoden mit physikalisch-spektroskopischem Hintergrund beziehen sich auf ganz bestimmte Phänomene, die in den Spektren auftreten können, wie zum Beispiel die Basislinienkorrektur zum Entfernen des erwähnten Fluoreszenzuntergrunds oder eine Intensitätsnormierung, die Abweichungen in der Gesamtintensität aufgrund geringer Schwankungen im Fokus oder der Intensität des Anregungslasers ausgleichen soll.

Andere Vorbehandlungsmethoden haben einen mathematischen (besonders numerischen und statistischen) Hintergrund. Sie bringen die Daten in eine Form, die für die weitere Datenanalyse besser geeignet ist. Das ermöglicht genauere numerische Lösungen und/oder der Auswertungsalgorithmus konvergiert schneller. Diese Datenvorbehandlung sollte trotzdem spektroskopisch sinnvoll sein. Anderenfalls können die in der weiteren Auswertung gebildeten Modelle möglicherweise nur schwer oder gar nicht interpretiert werden.

**Verringern der Anzahl verwendeter Variaten (Dimensionsreduktion) – Schneiden des Spektralbereiches:** Messkanäle, die kein Signal tragen, können das chemometrische Modell nicht verbessern. Oft verschlechtern sie sogar das Modell, da mehr Parameter geschätzt werden und die Gesamtunsicherheit wächst (Kap. 5.1). Spektralbereiche, die nicht zur Problemlösung beitragen können, sollten also aus der Auswertung

ausgeschlossen werden. Dabei ist besonders *a-priori*-Wissen von Vorteil, also biochemisches und spektroskopisches Vorwissen. Weiterhin können statt der einzelnen Messkanäle der Rohspektren integrale Intensitäten verschiedener Banden gezielt als Variate genutzt werden.

Häufig wird aber auch eine *datengesteuerte Variablenselektion* durchgeführt. Informationen aus dem Datensatz selbst werden zum Auswählen herangezogen. Zum Beispiel können Modelle, die unterschiedliche Spektralbereiche verwenden, anhand der erreichten Modellqualität verglichen werden. Aber auch univariate (bezüglich der Datenmatrix  $\mathbf{X}$ ) Maße werden als Kriterium herangezogen, zum Beispiel die Korrelation der einzelnen Messkanäle mit der Abhängigen. Damit die datengesteuerte Variablenreduktion erfolgreich ist, muss der Datensatz genügend Proben enthalten, um zwischen wichtigen und unwichtigen Variaten zu unterscheiden. Diese Einschränkung besteht bei *a-priori*-Wissen für die Modellbildung nicht, spektroskopisches und biologisches Wissen ist also besonders effektiv.

**Zentrieren:** Viele chemometrische Modelle nutzen Linearkombinationen der Intensitäten an verschiedenen Wellenzahlen (vgl. Gleichung 3.2)  $\mathbf{Y} = \beta\mathbf{X} + \boldsymbol{\varepsilon}$  (sogenannte bilineare Modelle). Dabei ist zunächst kein konstanter Term (Achsenabschnitt) vorgesehen. In der Praxis muss aber oft ein Achsenabschnitt berücksichtigt werden. Der Achsenabschnitt kann in das lineare Modell eingeführt werden, indem die Daten um eine konstante Variate  $x_0$  erweitert werden. Dann beschreibt der zu dieser Variate gehörende Koeffizient  $\beta_0$  den Achsenabschnitt:  $Y = \beta_1 x_1 + \beta_0 x_0 + \boldsymbol{\varepsilon}$ . Eine Alternative ist das Zentrieren des Datensatzes. Mittelwertzentrieren verschiebt den Mittelpunkt des Datensatzes auf den Koordinatenursprung, so dass kein Achsenabschnitt benötigt wird.

Das Zentrieren von Raman- oder IR-Spektren bedeutet, dass der Datensatz in Differenzspektren bezogen auf die mittlere Zusammensetzung umgerechnet wird. Allerdings kann aus chemischer Sicht ein anderes Bezugsspektrum sinnvoller sein. Bei spektroskopischen Messungen kann zum Beispiel auf das Spektrum der Probenmatrix [CB11], auf Blind- oder Leerwert-Spektren [CB10] oder das Mittelwertspektrum einer Kontrollgruppe zentriert werden [CB2]. Das Spektrum, auf das zentriert wird, muss hinreichend genau bekannt sein, weil dessen zufälliger Fehler zu einem systematischen Fehler der zentrierten Spektren wird.

Numerisch ist das Zentrieren der Daten für viele Auswertemethoden vorteilhaft. Manche Algorithmen konvergieren schneller und genauer, wenn die Daten um den Koordinatenursprung herum liegen. Andere Algorithmen zentrieren die Daten automatisch. Schließlich gibt es auch Auswerteverfahren, die auf nichtnegative Daten ausgelegt sind. Daher muss der Auswertalgorithmus in der Entscheidung über zentrieren oder nicht immer berücksichtigt werden.

**Skalieren:** Unterschiedlich große Wertebereiche in den Variaten bewirken bei vielen multivariaten Methoden eine unterschiedliche Gewichtung der Variaten. Das (Varianz-)skalieren der Datenmatrix soll dies verhindern. Dazu wird jede Variate (Spalte) der zentrierten Datenmatrix durch ihre Varianz geteilt. Jede Variate der varianzskalierten Datenmatrix hat also Varianz 1. Dieses Vorgehen ist Standard, wenn unterschiedliche physikalische Größen in einem Datensatz zusammen ausgewertet werden. Letztlich liegt hier die

Annahme zu Grunde, dass alle Variaten auch tatsächlich Signal tragen, und der Datensatz keine uninformativen Variaten enthält.

Bei spektroskopischen Datensätzen haben bereits alle Variaten dieselbe physikalische Einheit. Damit ist eine Skalierung weitgehend unnötig. Darüberhinaus ist das Rauschen auf den Spektren ggf. nicht proportional zur Signalhöhe (z. B. Extinktionsspektren, aber auch Raman-Spektren, die um einen hohen spektralen Hintergrund korrigiert wurden), sondern annähernd konstant. Wird ein solcher Datensatz varianzskaliert, so entstehen Variate mit extrem großen Rauschen, weil Variate, die kein oder nur wenig Signal zeigen, stark verstärkt werden. Bei Raman-Spektren ist das Rauschen nicht konstant, sondern hat in erster Näherung eine Varianz proportional zur beobachteten Intensität der Streustrahlung. Dass die Varianz eine gute Näherung an die Variation des zur Auswertung benötigten Signals ist, gilt nach der Untergrundkorrektur nur noch für Spektralbereiche in denen das Rohsignal hinreichend intensiv war, so dass der subtrahierte Untergrund nicht ins Gewicht fällt.

Spektroskopische Daten sollten nur in begründeten Fällen varianzskaliert werden.

### 4.3 Bilineare Modelle

Außerdem kann der Datensatz in eine geeignetere Darstellung projiziert werden. Korrelierte Messkanäle können so mit Hilfe von Linearkombinationen zusammengefasst werden. In der Biospektroskopie übliche Methoden sind die Hauptkomponentenanalyse (engl. *Principal Component Analysis, PCA*) und die *Partial-Least-Squares- (PLS-)Regression*, aber auch während einer linearen Diskriminanzanalyse (LDA) erfolgt ein solcher Projektionsschritt.

Diese auch als *bilineare Modelle* zusammengefassten Methoden zerlegen die Datenmatrix  $\mathbf{X}$  aus  $n$  Zeilen (Spektren) und  $p$  Spalten (Variate, Wellenlängen, Messkanäle) jeweils nach unterschiedlichen Kriterien in zwei neue Matrizen:

$$\mathbf{X}^{(n \times p)} = \mathbf{C}^{(n \times m)} \mathbf{S}^T(m \times p) \quad (4.1)$$

Diese neuen Matrizen haben je nach Methode traditionell unterschiedliche Namen:  $\mathbf{C}$  kann Konzentrationen oder Häufigkeiten (engl. *abundance*) enthalten, wird aber oft einfach englisch als *Scores-Matrix* bezeichnet,  $\mathbf{S}$  zum Beispiel als Reinkomponentenspektren, *loadings*, Koeffizientenmatrix, latente Variablen oder Komponenten. Wichtig ist, dass diese bilineare Form der physikalischen Sicht auf einzelnen Spezies in  $\mathbf{S}$ , gewichtet mit der jeweiligen Konzentration in  $\mathbf{C}$ . Die Anzahl  $m$  an unterschiedlichen (latenten) Spektren, die in das bilineare Modell eingeht, heißt daher auch der *chemische Rang* der Datenmatrix  $\mathbf{X}$ .

Meist werden durch diese Projektion unkorrelierte Anteile der Datenmatrix (Rauschen) abgetrennt ( $m < p$ ). Dann ist die Projektion auch eine Form der datengesteuerten Variablenreduktion. Die eigentliche Modellierung wird dann im Koordinatensystem der PCA oder PLS durchgeführt. Das soll helfen, die Modelle zu stabilisieren, da das Modell dann weniger Koeffizienten schätzen muss. Allerdings müssen natürlich auch die Koeffizienten für die entsprechende Projektion geschätzt werden.



### 4.3.1 Partial Least Squares (PLS) Regression und Principal Component Analysis (PCA)

Die PLS-Regression [139–142] ist eine Regressions- oder Kalibriertechnik. Die Daten werden zunächst in ein niedrigerdimensionales Koordinatensystem transformiert. Dieses Koordinatensystem wird durch die sogenannten *latenten Variablen* aufgespannt. Die latenten Variablen sind die Richtungen im Datenraum, die die höchste Korrelation oder Kovarianz mit den abhängigen Variaten aufweisen. Damit ist die PLS-Regression die analoge Kalibriertechnik zur LDA und der PCA. Aufgrund dieser engen Verwandtschaft wird PLS als Vorbehandlungsmethode für die LDA empfohlen [143].

Die PCA unterscheidet sich von der PLS dadurch, dass sie keine Referenzinformationen berücksichtigt und nur die Matrix  $\mathbf{X}$  der unabhängigen Variaten nutzt. Die PCA zerlegt die Kovarianz- oder Korrelationsmatrix der unabhängigen Variaten,

$$(n - 1) \text{COV}(\mathbf{X}) = \mathbf{X}_z^T \mathbf{X}_z \quad (4.2)$$

$$(n - 1) \text{COR}(\mathbf{X}) = \mathbf{X}_{z,s}^T \mathbf{X}_{z,s} \quad (4.3)$$

dabei ist  $\mathbf{X}_z$  die zentrierte und  $\mathbf{X}_{z,s}$  die zentrierte und varianzskalierte Datenmatrix.

Die PLS-Regression nutzt die Informationen sowohl der unabhängigen als auch der abhängigen Variaten, arbeitet ansonsten aber ähnlich. Einflüsse auf die unabhängigen Variaten, die sich nicht auf die Abhängigen auswirken, werden unterdrückt. Statt der Kovarianz-Matrix der unabhängigen Variaten wird die Matrix

$$\mathbf{S}_z^T \mathbf{S}_z = \mathbf{X}_z^T \mathbf{Y}_z \mathbf{Y}_z^T \mathbf{X}_z \quad (4.4)$$

zerlegt. Analog zur PCA entsteht so eine Projektion sowohl der unabhängigen Datenmatrix  $\mathbf{X}$ , als auch der Abhängigen  $\mathbf{Y}$ . Da bei dieser Projektion die neuen Koordinaten nach ihrer Wichtigkeit für die nachfolgende lineare Regression geordnet sind, werden nur die ersten  $m$  neuen Richtungen (latente Variablen) weiterverwendet, wobei der *Hyperparameter*<sup>(a)</sup>  $m$  vom Nutzer anzugeben ist. Im neuen Koordinatensystem wird nun eine lineare Regression durchgeführt. Damit sind PCA und PLS auch Regularisierungstechniken [145]: Die Einschränkung der Anzahl an latenten Variablen entspricht der Einschätzung, für die Modellierung seien nur  $m$  chemische Spezies (oder Substanzklassen) bedeutsam. Die PLS-Regression ist eine sehr robuste Kalibriertechnik. Sie kann gut mit Daten umgehen, bei denen die relevante Information über viele Variate verteilt ist und ist daher für die Auswertung von spektroskopischen Daten gut geeignet.

Wenn sich das quantitative Analysenproblem in guter Näherung linear beschreiben lässt, dann ist die Anzahl der benötigten latenten Variablen die Anzahl an Substanzen (bzw. an unterschiedlichen Spektren). Werden statt der absoluten Konzentration Anteile modelliert, so reicht eine latente Variable weniger zur Beschreibung aus. Die Vorhersagen der PLS-Regression sind dann in der Summe konstant (1 oder 100 %). Allerdings können einzelne Komponenten mit Anteilen  $< 0$  oder  $> 1$  vorhergesagt werden. Das weist entweder auf eine ungenaue Modellierung hin (nichtlineare Anteile, Wechselwirkungen zwischen den einzelnen Spezies), oder es ist ein Ausdruck der zufälligen Unsicherheit.

<sup>(a)</sup> *Hyperparameter* [144] steuern das Verhalten des gesamten Modells

Die Ergebnisse einer PLS-Regression können wie alle Regressionsergebnisse mit einem Grenzwert in Klassen umgerechnet werden. Das entspricht einer qualitativen Analyse: errechnet das Kalibriermodell eine Analytkonzentration oberhalb der Nachweisgrenze, so gilt der Analyt als vorhanden. Aber auch Klassifikationsprobleme werden mit der sogenannten *Partial Least Squares Discriminant Analysis* (PLS-DA) bearbeitet. Die Klassenzugehörigkeiten werden dabei als feste Level der Abhängigen, zum Beispiel  $y = +1$  und  $y = 0$  kodiert. Weitere Klassen können in Form von zusätzlichen Spalten der  $Y$ -Matrix aufgenommen werden. Nun wird eine PLS-Regression gerechnet, deren Ergebnisse mit einem Grenzwert in die Klassen umgerechnet werden. Der Vorteil der PLS-DA gegenüber der LDA ist dabei, dass diese Rechnung bereits mit wenigen Proben numerisch stabil ist. Die Inversion der Kovarianzmatrix bei der LDA ist dagegen numerisch instabil (siehe unten), so dass mehr Proben benötigt werden. Gegenüber der PCA werden weniger Komponenten (Richtungen) benötigt, um die Abhängigen  $Y$  zu modellieren. Andererseits ist die PLS-Regression eine Regressionsmethode. Die PLS versucht daher, die Klassen auf jeweils einen Punkt ( $Y = +1$  bzw.  $Y = 0$ ) zu projizieren. Das ist für Klassifikationsprobleme oft nicht sinnvoll, und „echte“ Klassifikationsmethoden wie die LDA sind besser als Regressionsmodelle darauf zugeschnitten, innerhalb der Klassen eine gewisse Streuung zuzulassen. Ein guter Kompromiss ist, die PLS-Regression zunächst als Projektionsmethode in einen niedrigerdimensionalen Raum ( $m \ll p$ ) zu nutzen, und in diesem dann statt einer linearen Regression eine LDA durchzuführen. Diese Modelle werden als PLS-LDA bezeichnet.

Die PLS-Regression ist eine etablierte Methode in der chemometrischen Kalibrierung. Auf Raman-Spektren von Hirngewebe wurde sie z. B. zur Bestimmung des Wassergehaltes [146] und der Lipide [107] angewendet. Neben der LDA wird auch PLS-DA zu Klassifikation von biospektroskopischen Daten eingesetzt, zum Beispiel zur Erkennung von Lymphknotenbefall bei Mamma-Karzinomen [147] und zur Erkennung des Primärtumors von Hirnmetastasen [109, 112].

### 4.4 Klassifikation

In der medizinischen Diagnostik teilen Tests das Untersuchungsobjekt (einzelne Zellen, Gewebe oder auch ganze Patienten) in bereits bekannte Gruppen wie „Gliom“ oder „normales Gewebe“ ein. Aus analytisch-chemischer Sicht handelt es sich dabei um qualitative Analysen. In der Chemometrie spricht man von Klassifikationsproblemen.

Chemometrische Algorithmen modellieren Funktionen in dem durch die Messkanäle aufgespannten Datenraum. Im Fall einer Klassifikation handelt es sich dabei um Trennflächen<sup>(b)</sup>. Das Berechnen der Modellparameter wird als Modellbildung oder *Training* bezeichnet. Dementsprechend heißt der dazu verwendete Datensatz *Trainingsdatensatz*. Der Trainingsdatensatz für ein Klassifikationsmodell besteht aus Beispieldaten (Spektren), deren Klassenzugehörigkeit bekannt ist (*Referenz* oder Gold-Standard). Die Referenzdaten gelten normalerweise als sicher richtig.

Grundsätzlich können alle Regressions- oder Kalibriermodelle in Klassifikationsmodelle umgewandelt werden: die metrische Ausgabe des Modells wird dann mit Grenz-

---

<sup>(b)</sup> Genauer: um  $(p - 1)$ -dimensionale Gebilde, die den  $p$ -dimensionalen Datenraum unterteilen.

werten in Klassenbezeichnungen umgerechnet. Das wird im Folgenden als *Dichotomisierung* oder *Härten* bezeichnet. Auf diese Weise wird zum Beispiel aus einer Analytkonzentration oberhalb der Nachweisgrenze „Analyt vorhanden“. In der klinischen Chemie werden Konzentration und Normalbereich für den Analyten angegeben, die Diagnose stellt dann der Arzt in der Zusammenschau aller Befunde. Die hier betrachteten Szenarien der Raman-Diagnostik funktionieren ähnlich. Die Raman-spektroskopische Klassifikation stellt dem Chirurgen einen weiteren Befund zur Verfügung, der in die Entscheidung über das Entfernen des Gewebestücks einfließen kann.

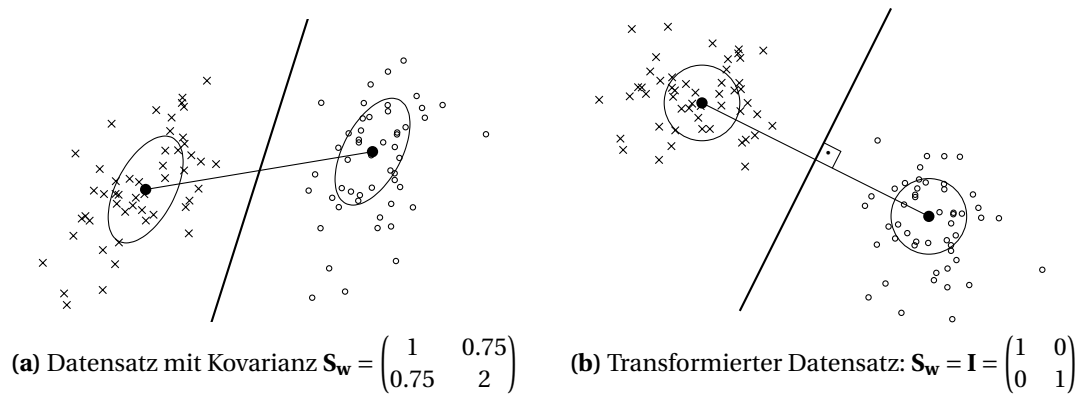
Im Unterschied zu einem Regressionsmodell, das metrische Abhängige als Funktion der Objekte (Spektren) modelliert, ordnet ein Klassifikationsmodell die Objekte (Spektren) vorher festgelegten Klassen zu. Die Abhängige ist also nominal oder ordinal skaliert. Die Klassen sind vorher festgelegte Gruppen von Objekten. Das Klassifikationsmodell soll diese Klassen möglichst gut trennen. Dazu stehen verschiedene Methoden zur Verfügung. Das Ziel ist meist die Zuordnung neuer Spektren durch das einmal gebildete Modell, die sogenannte *Vorhersage*. Dazu werden *prädiktive* Modelle erstellt. Im Unterschied zur Clusteranalyse ist Klassifikation ein überwachtes Verfahren: die Modellbildung (*training*) nutzt immer sowohl die Spektren als auch deren Klassenzugehörigkeit.

Die Einteilung in Klassifikations- und Regressionsmethoden ist insofern willkürlich, als Klassifikationsprobleme in Regressionsprobleme abgebildet werden können: ein 2-Klassen-Problem kann mit einer Regression, z. B. mit  $Y = +1$  für die eine und  $Y = -1$  für die andere Klasse, behandelt werden. Dann kann die Vorhersage der Klassenzugehörigkeit mit Hilfe des Vorzeichens des vorhergesagten  $Y$  oder allgemeiner mit einem Grenzwert für das vorhergesagte  $\hat{Y}$  getroffen werden. Eine Alternative kodiert die zweite Klasse als  $Y = 0$  oder „nicht Klasse 1“.

Neben der Klassenzugehörigkeit ist oft noch die sogenannte *a posteriori Klassenzugehörigkeitswahrscheinlichkeit*  $\Pr(K = k|X)$  (engl. *class posterior probability*) gesucht. Sie gibt die Wahrscheinlichkeit an, dass das beobachtete Spektrum von einem Objekt der betreffenden Klasse stammt. *A-posteriori*-Wahrscheinlichkeit im Unterschied zur *a-priori*-Wahrscheinlichkeit  $\Pr(k = k)$  bedeutet, dass die Beobachtung (das konkret vorliegende Spektrum  $X$ ) berücksichtigt ist. Die *a-priori*-Wahrscheinlichkeit gibt den Anteil der Objekte der Klasse  $G$  in der Grundgesamtheit an, das heißt, ohne eine Beobachtung zu berücksichtigen. In der Medizin heißt die *a-priori*-Wahrscheinlichkeit des Vorliegens einer Krankheit *Prävalenz* oder *Inzidenz*, falls es sich um die Erstdiagnose handelt. Prävalenz und Inzidenz beziehen sich immer auf eine bestimmte Patientenpopulation. Für konkrete Aufgabenstellungen in der medizinischen Diagnostik muss deshalb die Prävalenz der relevanten Patientenpopulation berücksichtigt werden. Im vorliegenden Fall der operationsbegleitenden Diagnostik ist nicht die Prävalenz der einzelnen Tumore entscheidend. Die relevante Grundgesamtheit besteht aus allen Geweben, bei denen der Chirurg das entsprechende Diagnosewerkzeug nutzen würde.

#### 4.4.1 Lineare Diskriminanzanalyse (LDA)

Die lineare Diskriminanzanalyse (LDA) geht auf Arbeiten der 1920er und 1930er Jahren zurück, insbesondere Mahalanobis Arbeit über generalisierte Distanzmaße in der Statistik [148] und die von Fisher [149]. Die LDA maximiert die Varianz zwischen den Klassen unter Berücksichtigung der Kovarianzstruktur innerhalb der Klassen [145, 150] und ist



**Abbildung 4.2** Die lineare Diskriminanzanalyse maximiert den Abstand zwischen den vorgegebenen Klassen (Kreise und Kreuze) unter Berücksichtigung der Kovarianzstruktur der Daten (Ellipsen) (a). Die Daten werden so transformiert, dass die Kovarianz innerhalb der Klassen die Einheitsmatrix  $\mathbf{I}$  wird (b). Im neuen Koordinatensystem ist die Richtung maximaler Distanz zwischen den Klassen die Verbindungslinie der Klassenmittelpunkte (Punkte). Sie ist gleichzeitig der Normalenvektor der Trennebene (dick). Im Koordinatensystem der Originaldaten (a) schneidet die Trennebene die Verbindungslinie der Klassenmittelpunkte aufgrund der Kovarianz zwischen den Variaten unter einem anderen Winkel. (nach [CB5])

ein überwachtes Analogon zur PCA, sowie das Klassifikationsanalogon zur PLS [143].

Die LDA transformiert die Daten zunächst so, dass die gemeinsame Kovarianzmatrix aller Klassen  $\mathbf{S}_w$  Abbildung 4.2a zur Einheitsmatrix  $\mathbf{I}$  wird (b). Die elliptische Kovarianzstruktur (Abb. 4.2a) geht in eine sphärische Struktur über (b). Eine solche Koordinatentransformation liegt auch der Mahalanobis-Distanz zu Grunde [150]. Wenn die gemeinsame Kovarianzmatrix innerhalb der Klassen  $\mathbf{S}_w$  sphärisch ist, gibt die Richtung zwischen den Klassenmittelwerten den Normalenvektor der optimalen Trennebene an. Der Schnittpunkt der Trennebene mit dieser Verbindungslinie wird anhand der *a-priori*-Wahrscheinlichkeit der Klassenzugehörigkeit für die untersuchten Klassen festgelegt. Oft wird der Schnittpunkt so gewählt, dass die Wahrscheinlichkeit einer Fehlklassifikation für beide Klassen gleich ist. Die Summe der Fehlklassifikationswahrscheinlichkeiten ist dann minimal. Eine Kostenfunktion kann die unterschiedlichen Fehlklassifikationsmöglichkeiten aber verschieden stark gewichten und den Schnittpunkt entsprechend verschieben.

Die Diskriminanzfunktionen sind die Eigenvektoren der Matrix  $\mathbf{S}_w^{-1}\mathbf{S}_b$  ( $\mathbf{S}_b$  ist die Kovarianzmatrix zwischen den Klassen). Der Ausdruck  $\mathbf{S}_w^{-1}\mathbf{S}_b$  kann als Signal-Rausch-Verhältnis bezüglich der Klassifikation interpretiert werden. Ein LDA-Modell zur Unterscheidung von  $n_g$  Klassen nutzt höchstens  $n_g - 1$  Diskriminanzfunktionen. Damit beinhaltet die LDA eine Projektion aus dem  $p$ -dimensionalen Datenraum in einen  $(n_g - 1)$ -dimensionalen Raum, in dem die Datenpunkte als *scores-plot* visualisiert werden können.

Mit Hilfe der Annahme, dass die Daten multivariat normalverteilt sind, kann die LDA auch die *a-posteriori*-Wahrscheinlichkeit für die Klassenzugehörigkeit vorhersagen: für jede Klasse wird die Wahrscheinlichkeit ausgerechnet, dass eine Probe mit dem beobachteten Spektrum zu dieser Klasse gehört. Das entspricht der *a-posteriori*-Wahrscheinlichkeit einer Einklassifikation. Meist wird LDA aber auf geschlossene Systeme angewendet (Kap. 4.5). Dann wird diese absolute *a-posteriori*-Wahrscheinlichkeit noch mit der Summe dieser Wahrscheinlichkeiten für alle Klassen ins Verhältnis gesetzt. Die-

ses Verfahren wird als *softmax* bezeichnet.

Die Annahmen, dass

1. die Klassen multivariat normalverteilt sind und
2. eine gemeinsame Kovarianzmatrix besitzen,

sind für reale Datensätze oft nicht erfüllt. Unterschiedliche Kovarianzstrukturen der einzelnen Klassen können im Rahmen einer quadratischen Diskriminanzanalyse berücksichtigt werden. Trotzdem wird dies selten genutzt, da quadratische Trennflächen mindestens ein quadratisches Wachstum der Probenzahl  $n$  mit der Anzahl der Variaten  $p$  (Messpunkte pro Spektrum) erfordern [22]. In der Praxis zeigt die lineare Diskriminanzanalyse zwar eine gewisse Empfindlichkeit gegen Ausreißer [23], liefert ansonsten aber in aller Regel gute Ergebnisse [145]. Barker und Rayens [143] zeigen, dass die PLS aufgrund ihrer engen Verwandtschaft mit der LDA zur Datenreduktion vor der LDA sehr gut geeignet ist. Die PLS-LDA kann besser als eine PCA-LDA Komponenten abtrennen, die zwar viel Varianz bewirken, aber nicht bei der Klassifikation helfen.

LDA ist eine etablierte Klassifikationsmethode und wurde in verschiedenen Studien zur Klassifikation von Hirntumoren angewendet [CB5–CB7, 100, 108, 151–153]. Kammer [135] nutzt *random forests* zur Variablenselektion und bildet dann LDA- und (harte) logistische Regressionsmodelle. Beide Modelle erreichten dieselbe Vorhersagequalität. PLS-LDA wurde zur Klassifikation von Raman-Spektren von Gebärmutterhals-Karzinomen [154] und zur Unterscheidung zwischen nativen und denaturierten gefrieretrockneten Proteinen [155] anhand von ATR-IR-Spektren, sowie für schwingungsspektroskopische Fragestellungen bezüglich Futter- [156] und Nahrungsmitteln [157] beschrieben.

#### 4.4.2 Logistische Regression

Da die *a-posteriori*-Wahrscheinlichkeit eine metrische Größe ist, kann sie mit Hilfe einer Regression modelliert werden. Eine lineare Funktion der Form  $Y = \beta X$  ist jedoch nicht geeignet, da  $Y$  Werte von  $-\infty$  bis  $+\infty$  annehmen kann ( $Y \in \mathbb{R}$ ). Deshalb wird eine Transformation benötigt, die den Wertebereich auf  $[0, 1]$  einschränkt. Dazu kann  $Y$  mit einer Sigmoid-Funktion multipliziert werden, zum Beispiel der *logistischen Funktion*:

$$f(x) = \frac{1}{1 + e^{-x}} \quad (4.5)$$

$$= \frac{e^x}{1 + e^x} \quad (4.6)$$

Diese Funktion nimmt bei  $x = -\infty$  den Wert 0 an, und bei  $x = +\infty$  den Wert 1. Bei  $(0; 0,5)$  liegt der Wendepunkt der Funktion (Abb. 4.3).

Die Umkehrfunktion der logistischen Funktion ist die *Logit-Funktion*:

$$L(x) = \ln\left(\frac{x}{1-x}\right) \quad (4.7)$$

Wenn  $p$  die Wahrscheinlichkeit ist, dass die Probe der betrachteten Klasse angehört, dann sind  $\frac{p}{1-p}$  die zugehörigen Chancen (engl. *odds*)<sup>(c)</sup>. Die Chancen haben also einen

<sup>(c)</sup> Chancen werden oft mit Doppelpunkt angegeben, können aber auch als Bruch oder Dezimalzahl ge-

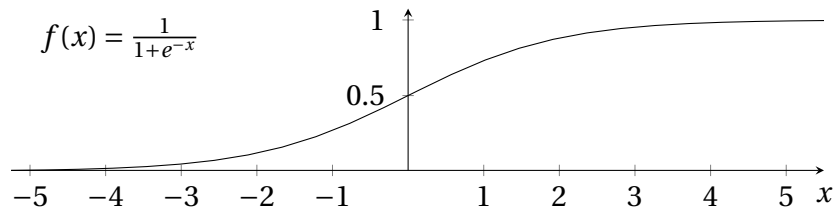


Abbildung 4.3 Die logistische Funktion

Wertebereich zwischen  $0 = 1 : \infty$  bis  $\infty = \infty : 1$ . Der Logit entspricht dem natürlichen Logarithmus der Chancen.

Damit ist das mathematische Modell der logistischen Regression:

$$L(p) = \ln\left(\frac{p}{1-p}\right) = \beta X + \varepsilon \quad (4.8)$$

Dieses Modell ist der LDA insofern ähnlich, als der Logarithmus der Chancen für 2-Klassen-Probleme dieselbe Form annimmt [158].

Allerdings sind die Residuen  $\varepsilon$  in dieser Situation nicht normalverteilt: in aller Regel gehört eine Probe entweder der einen oder der anderen Klasse an. Damit sind die Residuen entweder 0 oder 1. Normalverteilte Residuen sind aber eine Voraussetzung für die Methode der kleinsten Quadrate zum Bestimmen der Parameter eines linearen Modells. Logistische Regressionsmodelle werden daher mit Maximum-Likelihood-Methoden<sup>(d)</sup> geschätzt [145, 150, 159].

Theoretisch ist die logistische Regression dadurch weniger anfällig gegenüber Ausreißern als die lineare Diskriminanzanalyse. Die LDA macht stärkere Annahmen über die den Klassen zugrunde liegende Verteilung als die logistische Regression. Wenn die Annahmen erfüllt sind, sollten LDA-Modelle effektiver sein und weniger Trainingsproben benötigen, um dieselbe Modellqualität wie ein logistisches Regressionsmodell zu erreichen. Efron [158] zeigt, dass die LDA der logistischen Regression bei unbegrenzter Anzahl an Trainingsproben *relativ* überlegen ist. Allerdings ist der *absolute* Unterschied in der Modellqualität vernachlässigbar [160], weil die großen relativen Unterschiede bei sehr kleinen absoluten Fehlern auftreten. In der Praxis sind die Ergebnisse der logistischen Regression mit denen der LDA vergleichbar [145].

Die logistische Funktion dient auch als sigmoide Funktion in künstlichen neuronalen Netzen (engl. *artificial neuronal net*, ANN). Dadurch bieten künstliche neuronale Netze eine bequeme Methode zum Schätzen logistischer Regressionsmodelle [150]. Ein logistisches Regressionsmodell entspricht einem neuronalen Netz ohne verdeckte Schicht. Künstliche neuronale Netze arbeiten problemlos mit mehreren abhängigen Variaten, können also einfach auch zur Beschreibung von Mehrklassenproblemen verwendet werden. Geschlossene Systeme werden auch hier mit dem *softmax* abgebildet. Weiterhin lassen sich die Modelle durch verdeckte Schichten erweitern.

---

geschrieben werden. Wenn die Chancen  $3 : 1 = \frac{3}{1} = 3$  für ein Ereignis stehen, entspricht das einer Wahrscheinlichkeit von  $\frac{3}{3+1} = \frac{3}{4} = 0,75 = 75\%$ , dass das Ereignis eintritt.

<sup>(d)</sup> Für normalverteilte Residuen ist die Methode der kleinsten Quadrate auch die Maximum-Likelihood-Methode.

Die logistische Regression ist eine wichtige Methode in der medizinischen Diagnostik<sup>(e)</sup>. Silva Martinho *et al.* [161] messen Raman-Spektren von feuchten Biopsieproben des Gebärmutterhalses. Eine logistische Regression der Hauptkomponenten soll Gebärmutterhals-Karzinome diagnostizieren, wobei allerdings Entzündungen nicht von Tumoren unterschieden wurden. Robichaux-Viehoever *et al.* [162] klassifizieren *in vivo* aufgenommene Raman-Spektren des Gebärmutterhalses mit einer zweistufigen logistischen Regression. Hauptkomponenten zwei bis vier von Raman-Spektren von Sehnen wurden als Eingangsvariante für eine logistische Regression zur Diagnostik von Sehnenerkrankungen genutzt [163]. Nijssen *et al.* [164] nutzen die Ergebnisse einer Clusteranalyse (basierend auf den ersten 100 Hauptkomponenten) als Eingangsvariablen für eine logistische Regression zur Diagnostik von Basalzell-Karzinomen. Majumder *et al.* [165] unterscheiden verschiedene Brustgewebe und -tumore. Aus den Raman-Spektren werden nichtlineare Variante berechnet. Die eigentliche Klassifikation nimmt dann ein logistisches Regressionsmodell vor. Die Raman-Diagnostik war Klassifikationsmodellen auf der Basis von Fluoreszenzspektren und diffusen Reflexionsspektren sowie der Kombination dieser beiden Methoden überlegen. Dieselbe chemometrische Methode wurde auch zur Klassifikation von Raman-Spektren von Hauttumoren eingesetzt [166].

#### 4.4.3 Weitere Klassifikationsmethoden

Eine Vielzahl weiterer Klassifikationsmethoden ist bekannt. Hier seien noch *random forests* und Supportvektor-Maschinen (engl. *support vector machines*, SVM) erwähnt.

*Random forests* [145] sind Ensembles von Entscheidungsbäumen. Entscheidungsbäume (engl. *decision trees*) teilen die Daten anhand einer Serie von Vergleichen der Intensitäten einzelner Banden auf. Sie sind stark nichtlinear, leider aber auch sehr instabil gegenüber Änderungen in den Trainingsdaten. Daher werden zur Stabilisierung Ensemble solcher Entscheidungsbäume ausgewertet, die solche Änderungen in den Trainingsdaten simulieren. Ensemble können auch für andere Modelle als Entscheidungsbäume genutzt werden. Kapitel 4.6 betrachtet diese Methode zur Stabilisierung näher.

SVM [167, 168] konzentrieren sich noch stärker als logistische Regressionsmodelle auf die Klassengrenzen: die Klassengrenzen werden ausschließlich durch eine geringe Anzahl an Stützpunkten (engl. *support vectors*) beschrieben. Die oben beschriebenen Projektionen durch PCA-, PLS- und LDA-Modelle *reduzieren* die Anzahl der Dimensionen des Datensatzes, um einerseits Störsignale (Rauschen) abzutrennen und andererseits die Korrelation zwischen den einzelnen Variaten zu verringern. Dabei bleibt eine lineare Trennbarkeit der Klassen durch eine (Hyper)ebene im ursprünglichen oder reduzierten Datenraum erhalten. Der umgekehrte Weg hin zu höherer Dimensionalität kann helfen, wenn die Klassen im ursprünglichen Datenraum nicht linear trennbar sind. Dazu werden neue Variante durch nichtlineare Transformation der ursprünglichen Variaten erzeugt.

Bei der mathematischen Formulierung der SVM werden weniger die Koordinaten der einzelnen Datenpunkte als vielmehr ausschließlich das Skalarprodukt zwischen den verschiedenen Datenpunkten benötigt. Dieses Skalarprodukt kann durch das Skalarprodukt der nichtlinear transformierten Daten ersetzt werden, ohne dass die möglicherweise sehr großen nichtlinearen Datenmatrizen erzeugt werden müssen. Dieser bequeme Über-

<sup>(e)</sup> Pubmed verzeichnet über 80 000 Artikel zu „logistic regression and diagnosis“

gang zu nichtlinearen Modellen ist als *Kernel-Trick* bekannt. Er lässt sich auf alle Modelle anwenden, die *ausschließlich* mit dem Skalarprodukt zwischen Datenpunkten formuliert werden können. Für SVM ist die Kernel-Formulierung die Standardformulierung, aber auch für PCA, PLS [169], LDA [170] und logistische Regression [171] stehen Kernel-Algorithmen zur Verfügung, die einen entsprechenden Übergang zu nichtlinearen Modellen ermöglichen.

Im Zusammenhang mit schwingungsspektroskopischen Untersuchungen an Hirntumoren wurden SVM mit linearem und Gaußschem Kern von Bergner [109, 110, 112] verwendet, während Sattlecker *et al.* [172] und Leslie *et al.* [113] ausschließlich Gaußsche Kerne nutzen. SVM nehmen direkt eine harte Zuordnung zu den Klassen vor. Die besonders für das Astrozytom-Grading wichtigen Klassenzugehörigkeitswahrscheinlichkeiten sind zunächst nicht zugänglich. Mit `libSVM` können sie jedoch über eine nachgeschaltete logistische Regression berechnet werden [173]. In der vorliegenden Arbeit wird direkt die logistische Regression verwendet.

Bei SVM wird oft ein sogenannter Kostenhyperparameter benutzt, der einzelne Fehlklassifikationen im Bereich der Klassengrenze zulässt. SVM mit radialer Basisfunktion haben außerdem einen Hyperparameter für den Radius des Kerns. Beim Training von SVM werden diese Hyperparameter meist optimiert, indem mit einer Rastersuche (engl. *grid search*) das beste Modell gesucht wird. Diese Suche nach dem Maximum bewirkt generell ein extremes Ansteigen der benötigten Probenzahl (Kap. 4.8.5). Für SVM ist die Optimierung aber besonders problematisch, da die Modelle sprunghaft auf Änderungen der Hyperparameter reagieren [167, 168]. Cawley und Talbot [144] zeigen, dass diese Optimierungsstrategie zu einer erheblichen Überanpassung führen kann.

### 4.4.4 Modellqualität: zufällige und systematische Fehler

Klassifikationsmodelle unterliegen als Ergebnis einer statistischen Auswertung von Messdaten sowohl systematischen als auch zufälligen Fehlern. Zusätzlich kann bei der Klassifikation ein weiterer Fehler auftreten, wenn die gewählten Klassen tatsächlich nicht vollständig trennbar sind. Dieser *Bayes-Fehler* ist der minimale Fehler, der bei idealer Datenlage (unverrauschte Daten der Grundgesamtheit) und korrekter Modellierung der zugrunde liegenden Gesetzmäßigkeiten immer noch besteht. Er entsteht, wenn die Messmethode keine perfekte Trennung der Klassen erlaubt. Systematische Fehler (engl. *bias*) *bias* bedeuten, dass das gebildete Modell das zugrunde liegende Problem nur unzureichend oder fehlerhaft beschreibt, also Einflussgrößen nicht oder falsch berücksichtigt oder Näherungen bei der Modellierung der Abhängigkeiten durchgeführt werden. Demgegenüber entsteht der zufällige Fehler aus Rauschen und anderen zufälligen Einflüssen auf die Messdaten, die zur Schätzung der Parameter verwendet werden.

### 4.4.5 Erforderliche Trainingsprobenzahl

Lineare Modelle schätzen in der Regel für jede Dimension  $p$  einen Parameter, nichtlineare Modelle noch mehr. Bei einer gegebenen Komplexität des Modells hängt daher die Unsicherheit bei der Bestimmung der Modellparameter von der Dichte der Datenpunkte im hochdimensionalen Datenraum ab. Für Teile des Datenraums, die nur sehr dünn mit Proben besetzt sind, muss daher mit großen Ungenauigkeiten gerechnet werden. In die-



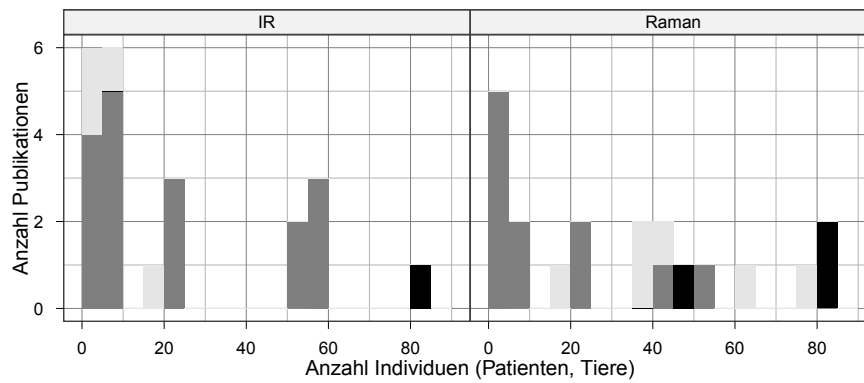
sen Gebieten dominieren wenige Proben große Teile des Datenraumes. Je nach Art des gebildeten Modells können sie auch das gesamte Modell stark beeinflussen (vgl. [145]). Das betrifft spektroskopische Datensätze aufgrund der Konzentration um die Raumdiagonale besonders. Andersherum formuliert führt eine steigende Anzahl an Variaten zu einem extrem steigenden Bedarf an Proben, weil das Volumen des durch die Variaten aufgespannten Raumes exponentiell steigt. Um die Probedichte aufrecht zu erhalten, sind auch exponentiell steigende Probenzahlen notwendig. Dieser Zusammenhang ist auch als *Fluch der Dimensionalität* (engl. *curse of dimensionality*) bekannt [145]. Eine gute Einführung bieten auch die „Frequently Asked Questions“ zu künstlichen neuronalen Netzen [174].

Der exponentiell mit der Anzahl der Variaten wachsende Probenbedarf lässt sich durch die Wahl geeigneter Randbedingungen einschränken. Für lineare Klassifikationsmodelle genügt auch eine linear wachsende Trainingsprobenzahl von mindestens 3 bis 5 statistisch unabhängigen Proben (Patienten) pro Variate in jeder Klasse [21–25]. Datengesteuerte Vorbehandlung wie eine Dimensionsreduktion mit PLS oder PCA vor der „eigentlichen“ Modellbildung bedeutet zusätzlich zu den  $m$  Variaten für die eigentliche Klassifikation Freiheitsgrade für die Projektion. Deshalb können solche Techniken zwar das Datenvolumen stark reduzieren, aber der Probenbedarf sinkt nur wenig [21].

Spektroskopische Datensätze haben oft hunderte bis tausende Variate. Die Anzahl der Spektren reicht von wenigen einzeln aufgenommenen Spektren bis zu mehreren Millionen Spektren bei IR-Imaging Datensätzen. In der vorliegenden Arbeit wurden mehrere zehntausend Spektren aufgenommen. Die einzelnen Spektren eines Patienten sind allerdings untereinander ähnlicher als zu Spektren anderer Patienten, sie sind nicht statistisch unabhängig voneinander. Ein zweites Spektrum desselben Patienten trägt also weniger als ein Spektrum einer Probe eines neuen Patienten zur Gesamtinformation bei. Deshalb liegt der Stichprobenumfang im statistischen Sinn, der *effektive* Stichprobenumfang, zwischen der Patientenzahl und der Anzahl der Spektren. Für IR-Images von Gliomen gibt es Hinweise, dass innerhalb der Probe eines Patienten jede gemessene Gewebeart merklich Information beiträgt, während weitere Spektren der Images (also desselben Gewebes) kaum zusätzliche Information bereitstellen [CB6].

Damit wären nach den Empfehlungen [21–25] hunderte oder tausende Patienten für die Modellbildung in typischen biospektroskopischen Studien erforderlich. Meist stehen jedoch weniger als 100 Patienten zur Verfügung. Abbildung 4.4 zeigt die Anzahl an Patienten (bzw. Tieren) der in der Literatur veröffentlichten Studien zur Schwingungsspektroskopie von Gliomen. Bei [CB2] und [CB3] handelt es sich um das in dieser Arbeit vorgestellte Astrozytom-Grading, [135] und [114] untersuchen Parallelschnitte zu den hier vorgestellten Bulkproben, die im Rahmen der Präparation der Bulkproben (Kap. 9.1) mit angefertigt wurden. Der empfohlene Stichprobenumfang wird damit in aller Regel um mindestens zwei, häufiger fast drei Größenordnungen unterschritten. Hinzu kommt die Schwierigkeit, gleichmäßige Probenzahlen für alle Klassen zu erhalten. So stellen Glioblastome mit über  $\frac{1}{4}$  der Proben (mehrere hundert) in der für diese Arbeit genutzten Hirntumorsammlung die weitaus größte Gruppe. Demgegenüber stehen 9 Kontrollproben von Patienten ohne Hirntumor und Proben von 8 Patienten mit PZNSL.

Die Abhängigkeit der Leistungsfähigkeit eines chemometrischen Modells von der Trainingsprobenzahl heißt *Lernkurve* [145]. Grundsätzlich hängt die Lernkurve stark davon ab, wie schwierig die jeweilige Problemstellung ist. Im Mittel steigt die Leistungsfähig-



**Abbildung 4.4** Anzahl der untersuchten Individuen in den Publikationen [CB2, CB3, CB6, CB7, CB12, 71, 96, 100–106, 108, 110–115, 118, 122, 124–127, 132, 134–138, 151, 152, 161–164, 166, 172]. Bei 4 Publikationen [106, 115, 122, 125] ist die Anzahl der Tiere unklar. Bei 3 weiteren Publikationen [101, 151, 172] ist nur die Anzahl der Proben bekannt, aber nicht die Patientenzahl. Publikationen zu den im Rahmen der vorliegenden Arbeit präparierten Proben [CB2, CB3, 114, 135], sind schwarz markiert. Publikationen, die sich nicht mit Hirntumordiagnostik beschäftigen, sind hellgrau. Bei Publikationen über Hirntumore (dunkelgrau) wurden außer Gliomen auch andere primäre und sekundäre Hirntumore gezählt.

keit aber mit steigender Probenzahl, während gleichzeitig die Streuung um diesen Mittelwert sinkt. Insgesamt sind diese Zusammenhänge komplex: In der Praxis schränken die extrem niedrigen Patientenzahlen in vielen biospektroskopischen Studien die mögliche Modellkomplexität ein. Deshalb werden bei steigender Trainingsprobenzahl auch komplexere Modelle gebildet. In der Literatur werden verschiedene Möglichkeiten zur Planung der erforderlichen Trainingspatientenzahlen vorgeschlagen, basierend auf Modellannahmen [22, 24, 25, 175, 176] oder durch Extrapolation von Lernkurven [177, 178].

Kapitel 5 und [CB4] untersuchen diesen Zusammenhang für eine Raman-spektroskopische Unterscheidung verschiedener Zelllinien und vergleichen die notwendigen *Trainings*-Probenzahlen mit den notwendigen *Test*-Probenzahlen.

## 4.5 Erweiterte Klassifikationskonzepte: Einklassen- und weiche Klassifikation

Meist bedeutet Klassifikation die Zuordnung der Proben oder Spektren  $x$  zu *genau einer* der  $n_g$  vorher festgelegten Klassen:

$$x \mapsto G \in \{1, \dots, n_g\} \quad (4.9)$$

Eine alternative Darstellung beschreibt die Klassenzugehörigkeit als Vektor mit  $n_g$  Elementen, wobei das  $G$ te Element 1 ist und alle anderen Null sind:

$$x \mapsto g \in \{0, 1\}^{n_g} \begin{cases} 1 & \text{für Element } G, \\ 0 & \text{sonst.} \end{cases} \quad (4.10)$$

Dass jede Probe zu genau einer Klasse gehören soll, wird dann als Summe ausgedrückt:

$$\sum_{j=1}^{n_g} g_j = 1 \quad (4.11)$$

Für einen ganzen Datensatz von  $n$  Spektren können diese Vektoren in eine Matrix  $\mathbf{G}^{(n \times n_g)}$  untereinandergeschrieben werden. Diese Darstellung wird im Folgenden verwendet.

In der medizinischen Diagnostik entspricht dieses Vorgehen der Differentialdiagnose: ein Tumor ist entweder ein Gliom oder ein Lymphom, kann aber nicht gleichzeitig Gliom *und* Lymphom sein. Dieses Klassifikationskonzept kann nun dahingehend erweitert werden, dass eine Probe mehreren Klassen gleichzeitig angehören kann. Ein Beispiel aus der medizinischen Diagnostik wäre ein Patient, der sowohl einen Tumor als auch Hepatitis hat. In diesem Fall werden die Klassen unabhängig voneinander modelliert und man spricht von *Einklassen-Klassifikation* (engl. *one-class classifier*) [167, 179, 180] oder einem *offenen Klassifikationssystem* (engl. *open world*). Entsprechend schließen sich die Klassen eines *geschlossenen* Klassifikationssystems (engl. *closed world*) gegenseitig aus. Es gilt also Gleichung 4.11. Sicherlich die bekannteste Einklassen-Klassifikationsmethode in der Chemometrie ist engl. *Soft Independent Modeling of Class Analogies* (SIMCA)<sup>(f)</sup> [167, 181]. Auch SIMCA-Modelle zur Erkennung des Primärtumors von Hirnmetastasen mittels IR-Imaging sind beschrieben [151, 182].

Die Klassifikationsprobleme in der vorliegenden Arbeit betreffen die Differentialdiagnostik, also geschlossenen Klassifikationssysteme: ein Tumor kann nicht gleichzeitig Lymphom *und* Astrozytom sein. Genausowenig kann ein Gewebe sowohl *ganz* normales Gewebe als auch *ganz* Astrozytomgewebe sein. Das gemessene Gewebavolumen kann aber sehr wohl *Anteile* von Astrozytomgewebe als auch *Anteile* von normalem Gewebe enthalten.

Eine zweite Erweiterung der Klassifikation, die solche anteiligen Klassenzugehörigkeiten beschreibt, ist die *weiche* (engl. *soft*) Klassifikation. Analog zum Übergang von der harten oder scharfen (engl. *crisp*) zur unscharfen (engl. *fuzzy*) Clusteranalyse wird der Grad der Klassenzugehörigkeit durch eine metrische Zugehörigkeit angegeben:

$$x \mapsto g \in [0, 1]^{n_g} \quad (4.12)$$

In der Fernerkundung ist der Begriff der *weichen* (engl. *soft*) Klassifikation bereits etabliert, so dass er für die vorliegende Arbeit übernommen wird. Außerdem bezieht sich *unscharf* auf die Theorie unscharfer Mengen (engl. *fuzzy set theory*) [183], also auf Uneindeutigkeit. Die anteiligen Klassenzugehörigkeiten können aber sowohl Uneindeutigkeit als auch Wahrscheinlichkeiten abbilden. Für die chemometrische Klassifikation bietet sich außerdem die Interpretation als Substanzgemisch an.

Die unscharfe (*fuzzy*) Clusteranalyse ist in der chemometrischen Auswertung von biospektroskopischen Daten ein etabliertes Verfahren [CB5, CB11, 184, 185], das die Gruppenzugehörigkeit der Spektren als Anteile ausdrückt. Im Gegensatz dazu sind anteilige Klassenzugehörigkeiten in der chemometrischen Klassifikation bislang unüblich, ob-

<sup>(f)</sup> Das „soft“ in SIMCA bezieht sich auf das *a priori* Wissen im Modellansatz (siehe unten) und hat erstmal nichts mit weicher Klassifikation im Sinne der 2. Erweiterung zu tun.

wohl Kuske *et al.* [186] eine unscharfe  $k$ -nächste-Nachbarn-Klassifikation vorstellen.

Für die Hirntumordiagnostik können anteilige (oder weiche) Klassenzugehörigkeiten drei unterschiedliche Effekte abbilden:

1. Es kann sich um eine Wahrscheinlichkeit handeln, mit der die betrachtete Probe zu einer Klasse gehört. Damit kann sowohl eine von einem Pathologen geäußerte Unsicherheit bei der Diagnose als auch uneinheitliche Diagnosen innerhalb Panels von Pathologen beschrieben werden (vgl. [187, 188]).
2. Für eine operationsbegleitende Diagnostik sollte die räumliche Auflösung an die Arbeitsgenauigkeit des Chirurgen von ca. 1 mm angepasst sein (Kap. 3.1).  $1 \text{ mm}^3$  entspricht etwa einer Million Astrozyten. Astrozytome wachsen infiltrativ, so dass häufig gemischte Zellpopulationen im Messvolumen eines Spektrums vorliegen. Die weiche Klassenzugehörigkeit kann Anteile unterschiedlicher Zellpopulationen (also eine Mischung) im Messvolumen widerspiegeln.
3. Astrozytome neigen außerdem zu weiterer Entdifferenzierung. Daher treten auch Diagnosen wie „Diese Zellen sind in der Entdifferenzierung von Astrozytom °II zu Astrozytom °III“ auf. Dies entspricht dem Konzept der Unschärfe in der Theorie der unscharfen Mengen.

Eine unsichere Diagnose kann gegebenenfalls so lange verbessert werden, bis keine Unsicherheit (1.) mehr vorliegt. Gemischte Zellpopulationen im Messvolumen (2.) könnten vermieden werden, indem die Messungen mit einer räumlichen Auflösung auf Einzelzellniveau durchgeführt werden. Für eine intra-operative Diagnostik ist dies jedoch nicht erstrebenswert, solange nicht auch die Therapie auf Einzelzellniveau erfolgen kann. Grund dafür ist, dass die Messzeit zur Aufnahme eines Raman-Spektrums recht lang ist. Das Gewebe darf bei der Diagnostik aber nicht geschädigt werden (weder thermisch noch photochemisch). Dies begrenzt die mögliche Anregungsleistung, die aber absolut höher sein kann, wenn sie sich auf ein größeres Volumen verteilt. Dadurch lässt sich mit wachsendem Messvolumen die Messzeit pro Spektrum verringern. Hinzu kommt, dass bei den gegenwärtig möglichen Messzeiten pro Spektrum nur einzelne Raman-Spektren aufgenommen werden können. Ist dabei das tatsächlich gemessene Volumen sehr klein, kommt es zu einer extremen Unterabtastung (engl. *undersampling*), das heißt, aus einem ohne Not klein gewählten Messvolumen muss der Chirurg auf ein sehr viel größeres Gewebevolumen extrapolieren. Aufgrund der Heterogenität der Astrozytome (Kap. 2.1) ist dies ungünstig.

Inwieweit auch der dritte Punkt eine Rolle spielt, hängt von der Tumorart ab. Die in dieser Arbeit untersuchten Astrozytome wachsen infiltrativ, so dass Effekt 2. bei einer angestrebten räumlichen Auflösung, die der Arbeitsgenauigkeit des Chirurgen entspricht, unvermeidlich ist. Die Entdifferenzierung der Astrozytome erfolgt in kleineren Schritten als das Grading der WHO suggeriert (Kap. 2.1), dies spiegelt die Unschärfe entsprechend Punkt 3. wider. Die detaillierte histologische Begutachtung der präparierten Parallelschnitte (Kap. 14) zeigt, dass diese Unschärfe auch in praktisch relevanter Häufigkeit beobachtet wird.

In der chemometrischen Literatur kommt „weich“ mit verschiedenen Bedeutungen vor. Der Unterschied zwischen harten und weichen Modellen kann sich auf a-priori-Wissen in den Modellen beziehen: harte Modelle nutzen Ansätze, die aus starken physikochemischen Annahmen abgeleitet werden, zum Beispiel die Ordnung der Reaktion in

Kinetiken. Weiche Modelle machen weniger oder schwächere Annahmen und modellieren empirische Näherungen [181, 189]. Ein Beispiel für diese Verwendung ist das *Soft Independent Modeling of Class Analogies* (SIMCA). Demgegenüber verwenden Varmuza und Filzmoser [190] *weich* synonym mit Einklassen-Klassifikation. Bei Breton [180] erlauben weiche Klassifikationsmodelle eine Überlappung der Klassen im *Datenraum*, was näher am Sprachgebrauch von SVM mit weichem Rand (engl. *soft margin*) [168] ist. In der vorliegenden Arbeit bezieht sich *weich* auf Klassenzugehörigkeiten (engl. *labels*). Proben mit weicher Klassenzugehörigkeit werden *uneindeutig* genannt, um Verwechslungen mit mechanisch weichen Proben zu vermeiden.

Weiche Klassenzugehörigkeiten können auf drei Ebenen auftreten:

**Weiche Vorhersagen** sind weit verbreitet, zum Beispiel die *a-posteriori*-Wahrscheinlichkeiten der linearen oder quadratischen Diskriminanzanalyse oder der logistischen Regression, die „Stimm-Verhältnisse“ der  $k$  nächsten Nachbarn ( $k$ NN) oder von *random forests*. Sie werden oft als Zwischenergebnis behandelt und mit Hilfe eines Grenzwerts *gehärtet*.

**Weiche Referenz beim Modelltraining:** Verschiedene etablierte Methoden wie die logistische Regression, ANN oder PLS-DA können anteilige Klassenzugehörigkeiten beim Training verarbeiten. Bislang wurde diese Möglichkeit aber nicht genutzt. Alle für diese Arbeit relevanten Anwendungen der logistischen Regression [161–164, 166] oder PLS-DA [109, 110, 112, 172] nutzen harte Modelle zur schwingungsspektroskopischen Tumordiagnostik.

**Weiche Referenz bei der Validierung:** Eine leistungsfähige Möglichkeit, die Qualität von Klassifikationsaussagen über uneindeutige Proben im Rahmen einer Validierung zu messen, fehlte bislang. In der Fernerkundungsliteratur hat sich zwar eine Vorgehensweise herauskristallisiert [191–194], die jedoch die Modellqualität systematisch überschätzt, weil nur der bestmögliche Fall betrachtet wird. Ein solcher optimistischer Bias ist für die Validierung einer medizinischen Diagnostik nicht akzeptabel.

Das Härten von metrischen *Scores* oder Konzentrationen in Klassen führt immer zu einem Informationsverlust [195]: Zwei Patienten, deren *Scores* knapp über und knapp unter dem Grenzwert liegen, erscheinen genauso unterschiedlich wie zwei Patienten mit extremen Werten. Dabei sind sich die ersten beiden Patienten auf der metrischen Skala sehr ähnlich, möglicherweise untereinander ähnlicher als demjenigen der beiden Patienten mit extremem *Score*, der in dieselbe Klasse eingeordnet wird. Hinzu kommt, dass die Wahl des Grenzwerts in mancher Hinsicht willkürlich ist. Insbesondere in der medizinischen Forschung hat diese Praxis der Dichotomisierung zu Problemen und entsprechender Kritik geführt [195–197].

Bislang werden in aller Regel entweder harte Referenzzuordnungen erzwungen und/oder Grenzfälle aus der Modellierung und Validierung ausgeschlossen. Das ist in mehrfacher Hinsicht problematisch. Der angesprochene Informationsverlust schlägt sich in einer erhöhten Varianz (Rauschen) nieder. In der Biospektroskopie werden oft harte, scheinbar eindeutige, Referenzzuordnungen erzwungen: Der Pathologe wird gebeten, sich auf *eine* Klasse festzulegen, auch wenn er die Probe als Grenzfall oder gemischte Zellpopulation beschreibt. Gelingt das, so enthält der ausführliche Befund oft deutlich detailliertere

Informationen und spiegelt die Uneindeutigkeit oder gegebenenfalls die Heterogenität der Probe wider (Kap. 14). Ist es dem Pathologen nicht möglich, sich auf eine Klasse festzulegen, so wird die Probe aus der Studie ausgeschlossen. Wenn mehrere Pathologen zu unterschiedlichen Ergebnissen kommen, wird ähnlich vorgegangen. Entweder die Mehrheitsmeinung wird als Referenzzuordnung genommen (was die Information über die Uneinigkeit entfernt) oder der Fall wird ausgeschlossen.

Damit verringert sich entweder die Probenbasis (oft drastisch, so wurden in [187]  $\frac{1}{3}$  aller Patienten wegen uneindeutiger Histologie ausgeschlossen) oder unangemessene Referenzinformationen werden verwendet. Der Ausschluss von Grenzfällen aus der Modellbildung führt dazu, dass künstlich leichtere Klassifikationsprobleme bearbeitet werden. Der Ausschluss der Grenzfälle aus der Validierung führt dazu, dass Probleme bei der Einordnung von Grenzfällen nicht erkannt werden können. Beim Grading der Astrozytome ist das besonders kritisch, da die Grenzfälle ja die Zielproben des Verfahrens sind: eine operationsbegleitende Diagnostik wird *hauptsächlich* für Grenzfälle benötigt.

Die Modellbildung mit anteiligen Klassenzugehörigkeiten in den Referenzdaten wird in Kapitel 12.3 (S. 105) beschrieben, Diskussion und Ergebnisse für das Grading der Astrozytome in Kapitel 16 (S. 119). Die Modelle für die Differentialdiagnostik zwischen Astrozytomen und Lymphomen werden in Kapitel 17 (S. 133) vorgestellt.

Die Validierung mit weichen Referenzdaten wird in Kapitel 4.9 eingeführt und in Kapitel 8 detaillierter behandelt. Die Erweiterung der Klassifikation wurde bezogen auf Grenzfälle beim Modelltraining in [CB2] und für Validierungsproben in [CB1] veröffentlicht.

### 4.6 Robustheit, Modellstabilität und Aggregation

**Robustheit (engl. *ruggedness*):** Messsignale können instabil sein: Das gemessene Spektrum einer unveränderlichen Probe kann sich verändern. Raman-Spektren ändern sich zum Beispiel bei Änderungen der Laserintensität und -wellenlänge, der Detektorempfindlichkeit (Auftreten toter Pixel) und der Justierung (Spalt- und Gitterposition) des Spektrometers. Alle diese Veränderungen können zufällig oder systematisch (Drift) sein. Die Stärke und Häufigkeit solcher Veränderungen hängt nicht zuletzt von der Bauart der verwendeten Messtechnik ab. Ein robustes Messgerät wird auch unter ungünstigen Umgebungsbedingungen (Temperaturänderungen, mechanische Belastung, ...) praktisch unveränderte Signale liefern. Auf der nächsten Ebene des analytischen Prozesses kann auch ein chemometrisches Modell mehr oder weniger robust sein, also mehr oder weniger stark auf die veränderten Messsignale reagieren.

Robuste Modelle können also Veränderungen in den Spektren (derselben Probe) gut kompensieren. Das kann durch eine gezielte Vorbehandlung wie zum Beispiel Intensitätsnormierung oder Basislinienkorrektur oder auch durch Anpassen der Kalibrierung geschehen. Auch das „eigentliche“ chemometrische Modell kann Veränderungen in den Daten kompensieren. Wie gut, hängt stark davon ab, inwieweit die zur Modellbildung genutzten Daten die Bandbreite an Veränderungen abdeckt.

Sattlecker *et al.* [147] haben verschiedene Klassifikationsmodelle auf ihre Robustheit gegenüber verschiedenen Veränderungen in den Spektren untersucht. Die Modelle waren sehr robust gegenüber zufälligem Rauschen, systematische Veränderungen gegenüber den Trainingsdaten führten aber zu deutlichen Einbrüchen in der Vorhersagequa-

lität. Das entspricht insofern den Erwartungen, als die Modelle in den Trainingsdaten Rauschen „kennengelernt“ hatten, aber ansonsten nur gute Spektren zum Training verwendet wurden und die untersuchten systematischen Störungen ausdrücklich nur die Testdaten betrafen. Die Datenvorbehandlung umfasste keine spezifischen Maßnahmen zum Kompensieren der untersuchten Störungen (z. B. weder Basislinienkorrektur noch Intensitätsnormierung).

**Stabilität:** Chemometrische Modelle werden aus Messdaten errechnet. Deshalb unterliegen sie wie andere Messergebnisse auch zufälligen und gegebenenfalls systematischen Unsicherheiten. Zu wenige Proben bei der Modellbildung führen zu einer großen zufälligen Unsicherheit der Modellparameter. Die Modelle sind *instabil*. Die ermittelten Modellparameter schwanken stark, wenn der Trainingsdatensatz nur wenig verändert wird. Möglicherweise verändern sich auch die Vorhersagen stark.

In bestimmten Grenzen ist es beim Modelltraining aber möglich, systematische Fehler zugunsten von niedrigeren zufälligen Fehlern (oder umgekehrt) in Kauf zu nehmen. Dies wird im Englischen als *bias-variance-tradeoff* bezeichnet [145]. In der vorliegenden Situation mit extrem wenigen Patienten im Verhältnis zur Anzahl der Variaten dominiert die zufällige Unsicherheit oft über systematische Fehler. Es kann daher vorteilhaft sein, bewusst systematische Fehler in Kauf zu nehmen, wenn die Varianz dadurch stärker reduziert werden kann. So kann ein vereinfachtes Modell bei geringen Probenzahlen besser sein als ein komplexeres Modell, dessen (viele) Parameter aber nur sehr ungenau geschätzt werden können [145]. Bei Klassifikationsvorhersagen können sich zufällige Fehler nicht wie bei der Messung einer metrischen Größe aufheben, weil es sozusagen nur eine Richtung für Klassifikationsfehler gibt: die richtige (harte) Klassenzugehörigkeit liegt immer am Rand des möglichen Wertebereichs. Deshalb bedeutet ein zufälliger Fehler immer auch systematisch schlechte Vorhersagen.

Kapitel 7 zeigt, wie die Stabilität sowohl der Modellparameter als auch der Vorhersagen sehr einfach im Rahmen einer  $k$ -fachen Kreuzvalidierung gemessen werden kann. Diese Vorgehensweise wurde in [CB6] veröffentlicht. Unabhängig davon wurde dieselbe Idee zur Messung der Stabilität der Vorhersagen wenig später von Dixon *et al.* [198] ebenfalls vorgestellt.

**Aggregation:** Instabile Modelle können durch *Modellaggregation* stabilisiert werden [145]. Die Aggregation kann sowohl auf der Ebene der Modelle als auch bei den Modellvorhersagen geschehen. Viele lineare Modelle können sehr einfach aggregiert werden. Die Parameter des aggregierten Modells sind die Mittelwerte der Modellparameter der einzelnen Untermodelle. Die zufällige Unsicherheit der Modellparameter kann über Fehlerfortpflanzung in die zufällige Unsicherheit der Ergebnisse umgerechnet werden.

Die Aggregation der Modellvorhersagen ist flexibler: Aussagen von ganz unterschiedlichen Verfahren können zu einer Aussage kombiniert werden. Die Unsicherheit der aggregierten Aussage wird aus der Verteilung der einzelnen Aussagen abgeschätzt. Metrische Vorhersagen können genau wie die Ergebnisse von Wiederholungsbestimmungen zusammengefasst werden. Üblich sind Mittelwert und Standardabweichung (Abb. 7.1 S. 75 Kreise unten). Klassifikationsergebnisse, die nominalskaliert vorliegen, werden oft in Form einer Mehrheitsentscheidung zusammengefasst (Abb. 7.1 Kreise oben). Hier kann die

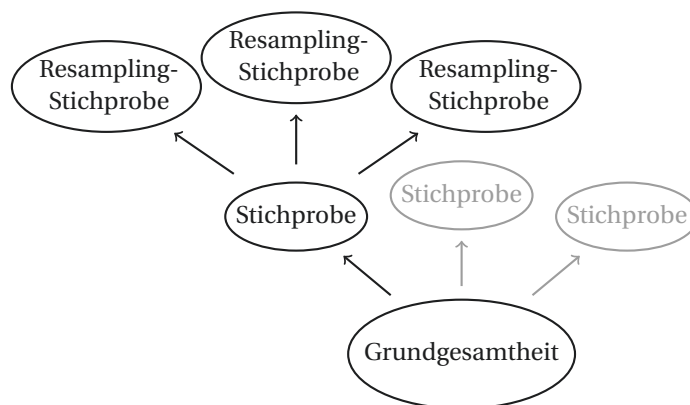
Unsicherheit in den „Stimmverhältnissen“ ausgedrückt werden. Viele Klassifikationsmodelle produzieren metrische Ergebnisse, die erst in einem zweiten Schritt in die eigentlichen Klassen umgewandelt werden. Diese Ergebnisse können vor der Umwandlung in nominalskalierte Klassen wie die Ergebnisse von Regressionsmodellen behandelt und nach der Aggregation umgewandelt werden.

Bootstrap-Aggregation (*bagging*) von Modellen wurde von Breiman [199] eingeführt. Seitdem ist Modellaggregation zu einer Standardtechnik in der Mustererkennung geworden. Das Ensemble an Modellen zur Aggregation wird dabei durch *bootstrap-Resampling* (s. u.) erzeugt. Durch *bootstrapping* sowohl der Trainingsproben als auch der Variaten entsteht so aus vielen Entscheidungsbäumen ein *random forest*.

Zur chemometrischen Auswertung spektroskopischer Datensätze wurde diese Technik bislang jedoch nur sehr selten angewandt. Bagging wurde von Menze, Petrich und Hamprecht [200] zur Erkennung von boviner spongiformer Enzephalitis (BSE) und von Sattlecker *et al.* [172] zur Unterscheidung von normalem Brustgewebe, *Carcinoma in situ* und invasiven Karzinomen auf der Basis ihrer IR-Spektren genutzt. Kammer [135] nutzt *random forests* zur Variablenselektion. Die *k*-fache Kreuzvalidierung kann ebenfalls zur chemometrischen Modellaggregation genutzt werden, sowohl für quantitative Bestimmungen [201] als auch für Klassifikationsmodelle [CB6]. Kapitel 7 diskutiert den Zusammenhang zwischen Modellstabilität und iterierter Kreuzvalidierung. Kapitel 7.1 zeigt dann eine Analogie zwischen Wiederholungsmessungen, aggregierten Modellen und aufgepolsterter (engl. *bolstered*) Validierung auf.

### 4.7 Resampling

Resampling-Verfahren werden verwendet, um die statistische Sicherheit von Aussagen abzuschätzen, obwohl echte Mehrfachmessungen nicht möglich sind. Die Idee hinter diesem Vorgehen ist, dass der vorliegende Datensatz wie eine Grundgesamtheit behandelt wird, aus der (viele) Stichproben gezogen werden. Aus der Variation zwischen solchen Stichproben wird die Sicherheit der Aussage bezüglich neuer echter Stichproben extrapoliert. Abbildung 4.5 verdeutlicht das Vorgehen. Resampling-Verfahren werden in



**Abbildung 4.5** Resampling. Aus der Grundgesamtheit (unten) werden Stichproben gezogen (Mitte). Dieser Schritt wird beim Resampling simuliert, indem weitere Stichproben (die Resampling-Stichproben, oben) aus der zur Verfügung stehenden Stichprobe gezogen werden.



der Validierung sehr häufig eingesetzt. Wie oben erläutert kann aber auch die Modellbildung in Form von aggregierten Modellen von diesen Strategien profitieren.

Unter *Resampling* versteht man also Strategien, die aus einer gegebenen Stichprobe eine weitere Stichprobe ziehen. Das kann entweder als *Ziehen mit Zurücklegen* oder als *Ziehen ohne Zurücklegen* erfolgen. Die Unterscheidung in Ziehen mit oder ohne Zurücklegen bezieht sich auf das Ziehen *einer* Stichprobe. Weitere Resampling-Stichproben werden wieder aus dem gesamten Datensatz gezogen. Im ersten Fall spricht man von *Bootstrap*<sup>(g)</sup>-*Resampling* [203]. *Jackknife*-Methoden ziehen ohne Zurücklegen. Der Begriff *Jackknife* im engeren Sinne bezeichnet den Spezialfall, immer genau eine Probe auszulassen (engl. *leave-one-out*). Johnson [202] gibt eine kurze Einführung in Bootstrap-Techniken, Efron und Tibshirani [204] eine detaillierte Diskussion.

Der *Jackknife* wird in der Regel *vollständig* durchgeführt. Das heißt, bei einem Probenumfang von  $n$  Proben werden  $n$  Stichproben gezogen, bei denen jeweils eine Probe ausgelassen wurde. Weitere Kombinationen von  $n - 1$  Probe sind nicht möglich. Wenn die Probenzahl wie die Anzahl der Patienten in dieser Arbeit eher klein ist, ist es besser, mehr als eine Probe auszulassen, dann sind mehr Kombinationen möglich. Allgemein gibt es beim Ziehen ohne Zurücklegen von  $m$  Proben aus einer Grundgesamtheit von  $n$  Proben  $\binom{n}{m} = \frac{n!}{(n-m)!m!}$  mögliche Stichproben. Beim Bootstrap werden mit Zurücklegen in aller Regel  $m = n$  Proben gezogen. Hier gibt es  $\binom{n+m-1}{m} = \frac{(n+m-1)!}{(n-1)!m!}$  verschiedene Möglichkeiten. Im Durchschnitt sind  $1 - \frac{1}{e} \approx 63\%$  der Proben in der Stichprobe vertreten. Außer beim *leave-one-out* sind in aller Regel zu viele Kombinationen möglich, als dass sie alle ausgewertet werden könnten. Daher wird normalerweise nur eine festgelegte Anzahl an zufällig gebildeten Kombinationen betrachtet. Üblich sind Größenordnungen von  $10^2$  bis  $10^3$  Resampling-Stichproben. Bei sehr kleinen Stichproben können bereits hier Wiederholungen auftreten. Das ist besonders zu beachten, wenn die Anzahl der möglichen Kombinationen durch Randbedingungen wie Stratifizierung eingeschränkt ist.

Ziehen ohne Zurücklegen ist in der chemometrischen Validierung häufig und wird meist im Rahmen einer Kreuzvalidierung durchgeführt. Bei der  $k$ -fachen Kreuzvalidierung werden die Proben auf  $k$  möglichst gleich große Teildatensätze aufgeteilt. Nun werden  $k$  Modelle gebildet, bei denen reihum je ein Teildatensatz ausgelassen wird. Dadurch gehen die Proben mit gleichem Gewicht in die Auswertung ein. Während das *leave-one-out*-Resampling nach  $k = n$  Modellen vollständig ist, kann und sollte die gesamte Prozedur bei  $k < n$  mit einer neuen zufälligen Aufteilung der Proben auf die Teildatensätze wiederholt werden [145, 205].

Beim Ziehen mit Zurücklegen ist eine gleiche Gewichtung aller Proben erst nach einer großen Anzahl von Bootstrap-Stichproben annähernd der Fall. Ziehen mit Zurücklegen entspricht andererseits dem Bestreben, die untersuchte Grundgesamtheit möglichst unverändert zu lassen.

---

<sup>(g)</sup> *Bootstrap* ist englisch für Stiefelriemen. Ein englischer Münchhausen hat sich nicht an den eigenen Haaren, sondern an seinen Stiefelriemen aus einem See gezogen [202].

## 4.8 Validierung

Die Qualität eines chemometrischen Modells wird im Rahmen einer Validierung gemessen. Das wird oft auch als *Testen* des Modells bezeichnet. Dabei berechnet das Modell Vorhersagen für Proben. Diese Vorhersagen werden dann mit den bekannten Referenzinformationen für die jeweiligen Proben verglichen. Dazu wird zum Einen eine Strategie zum Gewinnen der Testdaten benötigt, das Validierungsschema. Zum Anderen braucht es auch geeignete Qualitätsmaße oder Kenngrößen, die die verschiedenen Aspekte der Leistungsfähigkeit des untersuchten Modells beschreiben.

### 4.8.1 Validierungsschemata

Für Testproben müssen also ebenfalls Referenzinformationen vorliegen. Daher stellt sich die Frage, welche Proben zur Modellbildung und welche zur Validierung eingesetzt werden sollen. Verschiedene Validierungsschemata geben unterschiedliche Antworten auf diese Frage. Da die Validierung letztlich eine Messung ist, unterliegen auch diese Ergebnisse systematischen und zufälligen Fehlern. Mit welchem systematischen Fehler gerechnet werden muss, entscheidet im Wesentlichen das Validierungsschema. Die Varianz (zufälliger Fehler) der Validierungsergebnisse hängt entscheidend von der Anzahl der Testproben ab. Diese wird ebenfalls durch die Wahl des Validierungsschemas beeinflusst.

Im folgenden Abschnitt werden verschiedene Validierungsschemata und Literaturangaben über deren Auswirkungen auf Bias und Varianz der Validierungsergebnisse [205–209] vorgestellt. Kapitel 6 und [CB8] vergleicht diese Validierungsschemata dann in Bezug auf typische biospektroskopische Datensätze.

### Statistisch unabhängiger Testdatensatz (*Hold-out-Validierung*)

Meist soll die Qualität des Modells bezüglich der Vorhersage neuer, zum Zeitpunkt der Modellbildung unbekannter, Proben gemessen werden (engl. *generalization error*). Dazu können Modellvorhersagen für neue, unbekannte Patienten mit Referenzinformationen für diese Proben verglichen werden. Dieses Verfahren ist die einzige Möglichkeit, die Qualität der Modellvorhersagen für unbekannte Proben ohne systematischen Fehler (engl. *bias*)<sup>(h)</sup> zu messen.

Dennoch handelt es sich bei der Validierung mit einem eigens vorrätig gehaltenen Testdatensatz (engl. *hold-out*) nicht unbedingt um die beste Methode zur Modellvalidierung, wenn insgesamt nur wenige Proben vorhanden sind [205, 206]. Testdaten sind immer auch geeignete Trainingsdaten und könnten also auch zur Verbesserung des Modells dienen. Die Validierung mit einem statistisch unabhängigen Testdatensatz ist in der Praxis daher dasselbe wie das Teilen aller vorhandenen Daten in Test- und Trainingsdatensatz. Damit steht aber nur ein Teil der Daten für die Modellbildung zur Verfügung und ebenfalls nur ein Teil der Daten kann zum Testen genutzt werden. Steigt die Lernkurve

---

<sup>(h)</sup> Im Folgenden wird der Ausdruck *Bias* für systematische Fehler bei der Messung der Modellqualität verwendet, um Verwechslungen zwischen dem zu messenden Fehler des Modells und dem systematischen Fehler bei der Validierung zu vermeiden.

zwischen der Anzahl der beim *Hold-out* verwendeten Trainingsproben und der Gesamtzahl an Proben mit geeigneter Referenzinformation an, so bedeutet das, dass das Modell nicht so gut ist, wie es unter Einbeziehen aller Proben in den Trainingsdatensatz wäre. Zusätzlich ist die Messung der Modellqualität unsicherer, als würden alle Proben als Testproben genutzt.

Die Anwendung prädiktiver Modelle ist fast immer eine zeitliche Extrapolation: ein Modell wird erstellt und *danach* zur Vorhersage von später gemessenen Proben verwendet. Deshalb soll die Validierung in aller Regel nicht nur die Qualität der Modellvorhersagen für unbekannte Proben messen, sondern die Qualität der Vorhersagen für *zukünftige* unbekannte Proben. Dazu ist ein statistisch auch zeitlich unabhängiger Testdatensatz erforderlich. Dieser sollte möglichst zeitnah zu den Analysen gemessen werden. Im Idealfall werden die Analysen von Testproben begleitet [208]. Anders ist eine mögliche Drift nicht erkennbar. Natürlich kann das Modell im Nachhinein mit den neuen Testmessungen angepasst werden (engl. *model update*).

In Bezug auf die Fähigkeit des Modells, *unbekannte* Proben einzuordnen, ist das einmalige Teilen des Datensatzes also uneffektiv [205, 206]. Das gilt aufgrund der geringen Patientenzahlen auch und besonders für biospektroskopische Datensätze (Kap. 6 und [CB8]). In Bezug auf die Fähigkeit, *zukünftige* Proben einzuordnen, ist ein entsprechend später gemessener, zusätzlicher unabhängiger Testdatensatz jedoch unverzichtbar [208]. In Kapitel 5.2 und [CB3] wird gezeigt, dass die erforderliche Testpatientenzahl für eine solche Validierung für typische biospektroskopische Klassifikationsprobleme oft größer ist als die erforderliche Trainingspatientenzahl. Damit ist es ratsam, einen solchen Nachweis mit einem unabhängigen Testdatensatz von vornherein als Validierungsstudie zu planen, die dann auch den Einfluss der zeitlichen Extrapolation mitmisst.

## Resubstitution

Analog zum Vorgehen bei der Residuenanalyse einer Kalibrierung könnte man nach der Modellbildung die Referenzinformation mit der Vorhersage für die Trainingsproben vergleichen (engl. *resubstitution* oder *autoprediction*). Dann stünden alle Proben für die Modellbildung zur Verfügung und auch der Test nutzt die maximal mögliche Probenzahl. Genau genommen messen die Kennwerte dann allerdings nicht die Qualität der Vorhersage unbekannter Proben, sondern die Anpassungsfähigkeit des Modells an die Trainingsdaten (engl. *goodness of fit*). Diese Anpassungsfähigkeit ist im Kontext der Klassifikation biospektroskopischer Daten jedoch praktisch bedeutungslos. Die Zahl an statistisch unabhängigen Proben, also Patienten, ist im Verhältnis zur Anzahl der verwendeten Variaten sehr klein. Die Anzahl der Modellparameter ist daher im Verhältnis zur Probenzahl sehr groß. Das Modell hat viele Freiheitsgrade und kann sich an die wenigen Proben gut anpassen. Daher leiden die Modelle meist eher an Überanpassung (engl. *overfitting*) als an zu geringer Anpassungsfähigkeit.

Resubstitution liefert scheinbar sehr gute Kennwerte. Vorhersagen für neue Proben erreichen diese Qualität jedoch nur in den seltensten Fällen. Da die Ergebnisse systematisch besser scheinen, als sie wirklich sind, spricht man von einem *optimistischen Bias*.

### 4.8.2 Resampling-basierte Validierung

Die in beschriebenen Resampling-Strategien helfen weiter. Zunächst wird aus *allen* Proben ein Modell gebildet, im Folgenden als das *große Modell* bezeichnet. Dessen Qualität wird nun auf der Grundlage von Resampling-Stichproben geschätzt. Eine Stichprobe wird aus dem eigentlichen Datensatz gezogen. In der Regel wird der Stichprobenumfang so gewählt, dass fast alle Proben des Datensatzes in dieser Resampling-Stichprobe enthalten sind. Ein neues *Unter-* oder *Surrogatmodell* (vgl. [210]) wird mit dieser Resampling-Stichprobe als Trainingsdaten gebildet. Da nur wenige Proben nicht in der Stichprobe sind, sind die Trainingsdaten also annähernd dieselben wie die Trainingsdaten des großen Modells. Einige Proben wurden jedoch nicht in die Modellbildung des Untermodells einbezogen. Diese sind gegenüber dem Surrogatmodell statistisch unabhängig und werden als Testproben genutzt. Die Anzahl der Testproben ist bislang sehr klein. Die Validierungsergebnisse unterliegen daher einer sehr großen Varianz. Deshalb werden *viele* Resampling-Stichproben gezogen, und viele Surrogatmodelle mit den jeweils von *diesem* Modell unabhängigen Proben getestet. Die Ergebnisse aller dieser Tests werden dann gemeinsam ausgewertet. Obwohl solche Iterationen oder Wiederholungen für alle Resampling-Verfahren empfohlen werden, ist das in der Praxis bislang nur für die *bootstrap*-Verfahren üblich.

***k*-fache und *Leave-one-out*-Kreuzvalidierung:** Bei der Kreuzvalidierung (Abb. 4.6 auf Seite 46) werden die Resampling-Stichproben ohne Zurücklegen gezogen. Damit entspricht die Kreuzvalidierung konzeptionell dem *Jackknife* (Kap. 4.7, S. 40). Arlot und Celisse [207] geben eine ausführliche Übersicht.

Die Proben werden in  $k$  Unterdatensätze aufgeteilt (üblich sind  $k$  zwischen 5 und 10), von denen jeder bei einer Resampling-Stichprobe ausgelassen wird.  $k$  Surrogatmodelle werden gebildet und mit den jeweils ausgelassenen Proben getestet. Jede Probe ist also exakt ein Mal Testprobe. Diese ganze Prozedur kann und sollte mit neuen zufälligen Aufteilungen der Proben auf die Unterdatensätze wiederholt werden (Iterationen).

Da den Surrogatmodellen nur jeweils  $\frac{1}{k}$  der Proben fehlen, sind sie dem großen Modell normalerweise sehr ähnlich. Der pessimistische Bias ist daher gering. Ohne Iterationen ist die Varianz-Unsicherheit allerdings oft sehr viel größer als der Bias. Deshalb sollte die  $k$ -fache Kreuzvalidierung iteriert werden [205], was in Kapitel 6 und [CB8] auch für spektroskopische Daten gezeigt wird.

Kapitel 7 zeigt, wie aus den Ergebnissen der iterierten Kreuzvalidierung zusätzlich sehr einfach Aussagen über die Stabilität der Modelle abgeleitet werden können und welche Aussagen sich daraus über die mögliche Reduktion der Varianzunsicherheit ergeben.

Ein Sonderfall der  $k$ -fachen Kreuzvalidierung ist die *Leave-one-out*-Validierung. Hier gilt  $k = n$ , es wird also immer genau eine Probe ausgelassen. Wie schon beim *Leave-one-out-Jackknife* sind auch hier nur  $n$  Modelle mit  $n$  Tests möglich. Die Varianz kann nicht weiter verringert werden. Deshalb ist die *leave-one-out*-Validierung nicht zu empfehlen.

Eine Besonderheit betrifft den systematischen Fehler der *leave-one-out* Validierung. Die Trainingsdaten stimmen bis auf eine einzige Probe mit denen des großen Modells überein. Deshalb sind die Ergebnisse der *leave-one-out* Validierung meistens fast ohne Bias. Wenn die Klassifikationsmodelle bei wenigen Trainingsproben stark auf die Anteile der einzelnen Klassen an den Trainingsdaten reagieren, kann aber auch ein sehr star-

ker pessimistischer Bias auftreten. Da die Testprobe immer aus einer Klasse stammt, die gegenüber dem Gesamtdatensatz unterrepräsentiert ist, können solche Modelle unverhältnismäßig oft eine falsche Klasse voraussagen [205]. Das wurde für spektroskopische Daten auch in [CB8] beobachtet.

**Aufteilen in Trainings- und Testdatensatz (Set-Validierung):** Auch das Aufteilen der Daten in einen Trainings- und einen Testdatensatz (engl. *set validation*) kann als Ziehen ohne Zurücklegen aufgefasst werden. Damit können und sollten auch bei dieser Validierungsstrategie Iterationen durchgeführt werden.

Die Set-Validierung unterscheidet sich vom Reservieren eines eigenständigen Testdatensatzes (*hold-out*) dadurch, dass beim *hold-out* das gebildete Modell als das tatsächlich zu untersuchende Modell betrachtet wird, während hier das gebildete Modell als Surrogatmodell aufgefasst wird und die Validierungsergebnisse auf das endgültige Modell aus allen Proben übertragen werden.

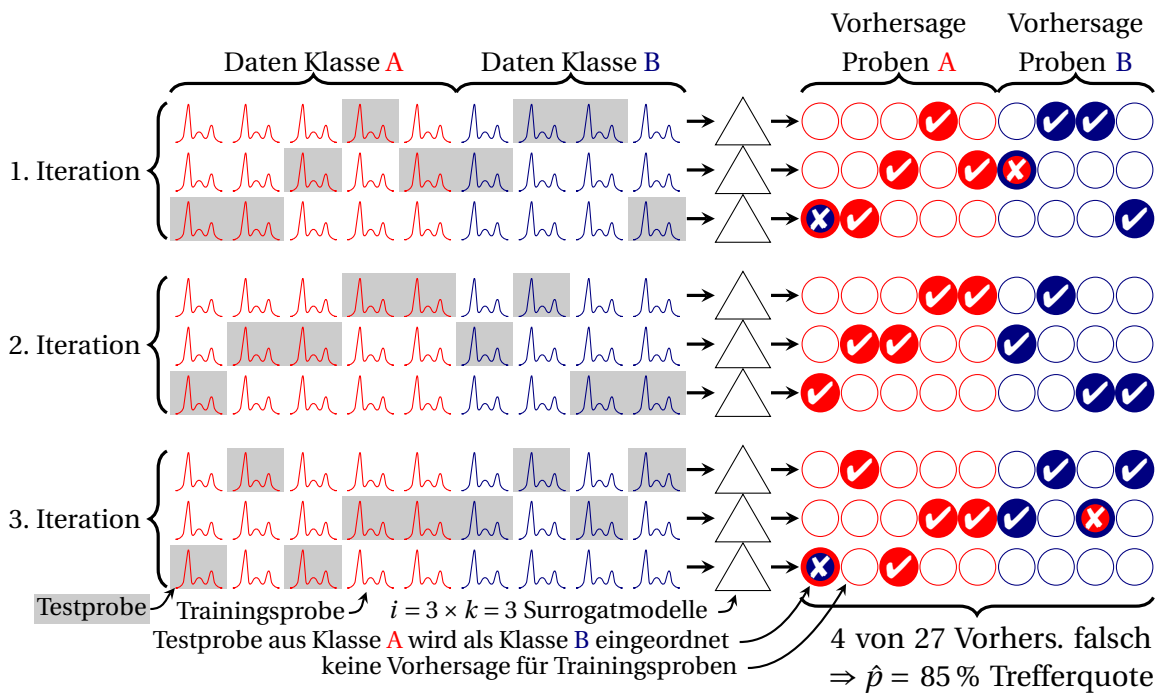
**Out-of-bootstrap-Validierung:** Die *out-of-bootstrap*- (oob) Validierung zieht mit Zurücklegen. Meistens werden genau so viele Proben gezogen, wie im Gesamtdatensatz vorhanden sind. Im Durchschnitt bleibt dann  $\frac{1}{e}$  der Proben, also ein gutes Drittel, übrig und kann zum Testen verwendet werden. Die *out-of-bootstrap*-Validierung wird meist mit 100–1000 Iterationen durchgeführt. Die *out-of-bootstrap*-Validierung hat meistens einen größeren pessimistischen Bias als die  $k$ -fache Kreuzvalidierung ( $k$  ist in der Regel größer als 3). Die Varianz ist allerdings aufgrund der größeren Testprobenzahl geringer (außerdem erfolgt in aller Regel eine größere Zahl an Iterationen).

Es gibt eine Reihe von Varianten zur *out-of-bootstrap*-Validierung, die den pessimistischen Bias korrigieren sollen. Der *.632bootstrap* mischt den *out-of-bootstrap*-Fehler mit dem Resubstitutionsfehler in den Anteilen  $1 - \frac{1}{e}$  und  $\frac{1}{e}$ . Wenn die Surrogatmodelle allerdings überangepasst sind, so dass der Resubstitutionsfehler 0 wird, kann ein optimistischer Bias des *.632bootstrap* auftreten (Kap. 6). Dieses Problem ist bekannt und hat zu einer Verbesserung, dem *.632+bootstrap*, geführt. Dabei wird versucht, das Ausmaß an Überanpassung abzuschätzen, und der Resamplingfehler entsprechend niedriger gewichtet [145]. Kim [211] beobachtete allerdings auch beim *.632+bootstrap* in bestimmten Situationen einen optimistischen Bias (unter anderem auch in einer Situation mit Resubstitutions-Fehlerrate von ca. 15 %).

In der genannten Studie [211] werden verschiedene Resampling-Verfahren zur Validierung von Klassifikationsmodellen mit jeweils gleicher Anzahl an Surrogatmodellen verglichen. Dabei schnitt der *.632+bootstrap* in manchen Simulationen etwas besser, in anderen etwas schlechter als die iterierte  $k$ -fache Kreuzvalidierung ab.

### 4.8.3 Validierung von aggregierten Modellen

Auch aggregierte Modelle können und müssen mit statistisch unabhängigen Testproben validiert werden [CB6, 212]. Werden die Untermodelle mit Resampling-Trainingsdatensätzen gebildet, so werden immer einige Proben aus dem Training des Untermodells ausgeschlossen. Diese Proben können insofern zum Testen verwendet werden, als sie von diesem Untermodell statistisch unabhängig sind. Werden genügend viele Untermodelle berechnet, so gibt es für jede Probe mehrere Testvorhersagen. Diese werden nicht einzeln



**Abbildung 4.6**  $k$ -fache Kreuzvalidierung. Jede Probe (Kästchen) ist in jeder Iteration exakt einmal Testprobe (hellgrauer Hintergrund links). Bei  $i$  Iterationen werden insgesamt  $k \cdot i$  Modelle gebildet. Die  $n \cdot i$  Modellvorhersagen (gefüllte Kreise rechts) werden mit der jeweiligen wahren Klassenzugehörigkeit (rot oder blau) verglichen (✓= richtig, ✗= falsch). Hier sind 4 von 27 Vorhersagen falsch, die Trefferquote ist  $\frac{23}{27} = 85\%$ . (nach [CB6])

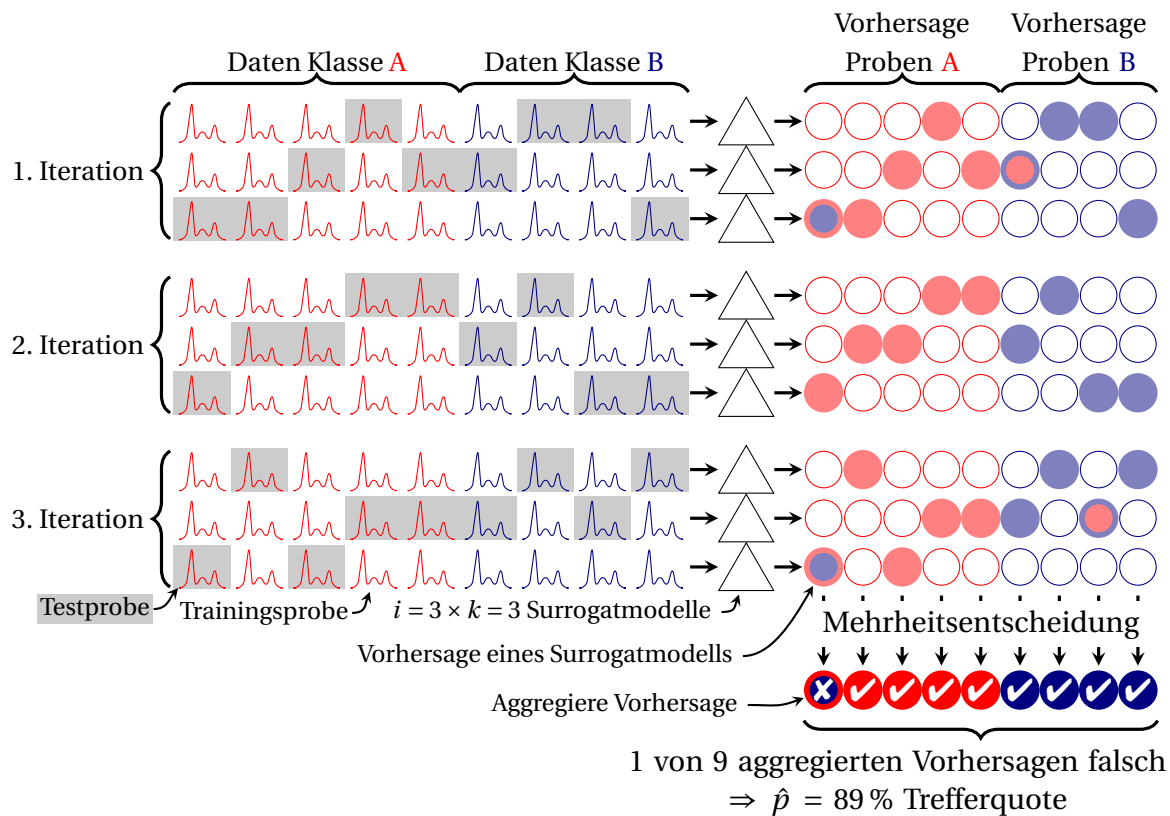
mit der Referenz verglichen, sondern zuvor aggregiert (Abb. 4.7). Der Vergleich erfolgt nun zwischen den aggregierten Vorhersagen und der Referenz.

Der Vergleich zwischen Abbildung 4.6 und Abbildung 4.7 verdeutlicht, wieso das Ensemble-Modell insgesamt bessere Vorhersagen liefert: bei insgesamt 27 Tests wird zunächst viermal die falsche Klasse vorhergesagt (83 % richtige Vorhersagen). Aggregieren der Vorhersagen (Abb. 4.7) korrigiert aber zwei dieser Fehler mit einer  $\frac{2}{3}$ -Mehrheit für die richtige Klasse. In einem Fall wird die falsche Klasse mehrheitlich vorhergesagt. Damit ist liegt ein Fehler bei neun aggregierten Vorhersagen vor, also 89 % Richtige. Die Aggregation ist also ein Filter, der einzelne zufällig verteilte falsche Vorhersagen herausfiltert. Bei systematisch falsch eingeordneten Proben wird die Vorhersage hingegen noch schlechter. Die Aggregation mit Mehrheitsentscheidung ist also ein Filter, der die Vorhersagen härtet [CB6].

Auch hier gilt wieder, dass die getesteten Ensemble-Modelle im Schnitt etwas schlechter sind als das große aggregierte Modell: sie enthalten weniger Untermodelle und die Vorhersagen unterliegen einer größeren Varianz.

#### 4.8.4 Statistische Unabhängigkeit von Trainings- und Testdaten

Wichtig ist also, dass die zum Testen verwendeten Spektren statistisch unabhängig vom getesteten Modell sind. Das bedeutet, dass sie keinerlei Einfluss auf die Modellbildung hatten. Test- und Trainingsdatensatz müssen strikt voneinander getrennt werden. Test-



**Abbildung 4.7**  $k$ -fache Kreuzvalidierung eines Ensemble-Modells. Jede Probe (Kästchen) ist in jeder Iteration exakt einmal Testprobe (hellgrauer Hintergrund links). Jede Probe ist also insgesamt von  $i$  Modellen statistisch unabhängig. Die  $i$  Modellvorhersagen (grauer und schwarze Kreise rechts) werden aggregiert (unten). Die aggregierte Vorhersage wird mit der jeweiligen wahren Klassenzugehörigkeit (graue oder schwarze Spektren links) verglichen ( $\checkmark$ = richtig,  $\times$ = falsch). Nur noch eine von neun Vorhersagen ist falsch (Trefferquote  $\frac{8}{9} = 89\%$ ). (vgl. [CB6])

daten bleiben nur dann unabhängig vom Modell, wenn sie in *keiner* Weise in die Berechnung der Modelle einfließen. In der Praxis sind wichtige Ursachen für „Lecks“ zwischen Test- und Trainingsdaten, dass die Struktur der Daten nicht berücksichtigt wird oder dass Testdaten in die Berechnung von Hyperparametern einfließen.

**Hierarchische Datenstruktur:** In dieser Arbeit sind von einem Patienten gegebenenfalls mehrere Proben verfügbar, von jeder Probe können mehrere Messungen angefertigt werden und jede Messung besteht aus hunderten von Spektren. In diesem Fall muss die Trennung zwischen Trainings- und Testdaten auf der obersten Ebene, also in Trainingspatienten und Testpatienten erfolgen. Tabelle 5.1 (S. 65) zeigt eindrucksvoll für Raman-Spektren von einzelnen Zellen, wie wichtig dies ist: Bei der Erkennung unbekannter Spektren von bekannten Batches werden 90 % der MCF-7-Zellen korrekt eingeordnet. Für unbekannte Batches raten die Modelle innerhalb der Tumorklassen, die Sensitivität für die MCF-7-Zellen bricht auf 30 % ein.

**Datengesteuerte Vorbehandlung,** also Vorbehandlungsschritte, die die Daten mehrerer oder aller (Trainings-)Patienten einbeziehen, müssen für Trainings- und Testpatienten

ten getrennt ausgeführt werden. Typische Beispiele sind Zentrieren der Daten, Varianzskalierung oder auch PCA- PLS- oder ähnliche Projektionsschritte, wobei das Mittelwertspektrum, die Varianzen oder die Projektion aus mehreren Patienten berechnet wird. Die Testpatienten bleiben nur dann unabhängig vom Modell, wenn sowohl Datenvorbehandlung als auch die „eigentliche“ Modellbildung nur aus den Trainingspatienten berechnet und die jeweiligen Transformationen dann auf die Testpatienten nur angewendet werden. Anderenfalls gehen die Testpatienten letztlich mit in die Modellbildung ein. Die Größe eines so entstehenden optimistischen Bias ist stark von der Art des betroffenen Vorbehandlungsschrittes abhängig. Zentrieren führt möglicherweise nur zu geringfügigen Fehlern, während ein Datenleck bei der PCA- und PLS-Vorbehandlung gerade bei geringen Patientenzahlen dazu führen kann, dass die Anzahl der erkannten Fehler um eine Größenordnung zu gering ausfällt.

**Datengesteuerte Modellselektion bzw. -optimierung:** Eine Reihe von Modellen wird aus den Trainingsdaten gebildet. Dabei werden Hyperparameter variiert. Diese Modelle werden getestet und das beste Modell wird ausgewählt. Aufgrund des Auswahl-schrittes sind nun die Testdaten in die Bildung des endgültigen (optimalen) Modells mit eingegangen. Typische Beispiele hierfür sind die datengesteuerte Auswahl von Spektralbereichen, die Bestimmung der Anzahl an Hauptkomponenten oder latenten Variablen bei der PLS, die Optimierung der Hyperparameter einer SVM oder die Festlegung eines optimalen Grenzwerts mit Hilfe der engl. *Receiver Operating Curve* (ROC) (vgl. Kap. 4.8.5, S. 54). Eine Validierung mit unabhängigen Testdaten erfordert hier, dass weitere Testpatienten eigens für die Validierung des optimierten Modells bereitgehalten werden. Dieses Vorgehen wird in der Literatur auch als geschachtelte (engl. *nested*) oder doppelte (engl. *double*) Validierung bezeichnet [144, 145, 167, 190, 213]. Kapitel 5.2.1 befasst sich näher mit diesen Fragestellungen. Werden dieselben Testergebnisse sowohl zur Modellbildung als auch zur Angabe der letztendlich erreichten Modellqualität verwendet, so muss mit einem *erheblichen* optimistischen Bias gerechnet werden [16, 144, 214–219].

**Abhängigkeit der Referenzdaten von den Spektren:** Besonders bei orts aufgelösten Messungen ist es sehr mühsam, Referenzinformationen auf die einzelnen Spektren zu übertragen. Eine automatisierte Zuordnung ist daher wünschenswert. Wichtig ist jedoch, dass die spektrale Information bei der Übertragung höchstens für Trainingsproben zu Hilfe genommen wird, aber *nie* für Testproben. Nur dann ist die Unabhängigkeit von Testspektren und Referenzinformation garantiert. Kapitel 11 (S. 95) diskutiert diese Abhängigkeit genauer.

#### 4.8.5 Kenngrößen für die Qualität der Vorhersagen

Außer einer *Messmethode* (Validierungsschema) zur Bestimmung der Qualität der Klassifikationsmodelle benötigt die Validierung noch geeignete *Maße* oder Kenngrößen, die die Modellqualität beschreiben.



## Die Zuordnungsmatrix $\mathbf{Z}$

Der Vergleich zwischen Referenzinformation und Vorhersage wird bei Klassifikationsmodellen oft als Zuordnungsmatrix (engl. *confusion matrix*) oder Kontingenztabelle (engl. *contingency table*) tabelliert. Diese stellt die Referenzinformation  $R$  in den Zeilen der Vorhersage  $P$  (Spalten) gegenüber (Abbildung 4.8a). Das heißt, jedes Testobjekt wird im Element  $\mathbf{Z}_{R,P}$  gezählt. Manchmal ist es bequemer, dies als Funktion zu formulieren:  $Z(R, P)$ .

$$\mathbf{Z}_{i,j} = \begin{cases} 1 & \text{wenn } i = R \wedge j = P, \\ 0 & \text{sonst.} \end{cases} \quad \text{für eine Probe (Spektrum)} \quad (4.13)$$

Dabei steht  $\wedge$  für den UND-Operator, der nur dann 1 ergibt, wenn sowohl die Referenzklassenzugehörigkeit  $r_i$  als auch die vorhergesagte Klassenzugehörigkeit  $p_j$  1 ist. Ansonsten ergibt der UND-Operator 0. Mit dieser Codierung der Wahrheitswerte „falsch“ mit 0 und „richtig“ mit 1 kann statt des UND-Operators also auch die Multiplikation zur Berechnung verwendet werden. Mit der in Gleichung 4.10 (Kap. 4.5) eingeführten Darstellung der Klassenzugehörigkeiten als Vektor oder Matrix lässt sich das elegant als Matrixmultiplikation ausdrücken:  $\mathbf{Z} = \mathbf{R}^{T(n_g \times n)} \mathbf{P}^{n \times n_g}$ .

Im Folgenden werden außerdem einige Abkürzungen für Summen über bestimmte Teile der Zuordnungsmatrix  $\mathbf{Z}$  genutzt. Dabei werden alle Elemente addiert, die den in den Indices angegebenen Bedingungen entsprechen. So ist  $\sum \mathbf{Z}_{i,P}$  die Summe über alle Zeilen der Spalte  $P$ :  $\sum_{i=1}^{n_g} \mathbf{Z}_{i,P}$ .  $\sum \mathbf{Z}_{i \neq G, P}$  steht abkürzend für  $\sum_{i \in \{1, \dots, n_g\} | i \neq G} \mathbf{Z}_{i,P}$ , die Summe über alle Zeilen außer  $G$  der Spalte  $P$  und  $\sum_n$  steht für die Summe über alle  $n$  Zeilen (Spektren, Proben) der Klassenzugehörigkeitsmatrix.

Die Zuordnungsmatrix des gesamten Tests erhält man als die Summe aller Zuordnungsmatrizen der einzelnen Proben:

$$\mathbf{Z}_{i,j} = \sum_n Z(r_i, p_j) = \sum_n r_i \wedge p_j. \quad (4.14)$$

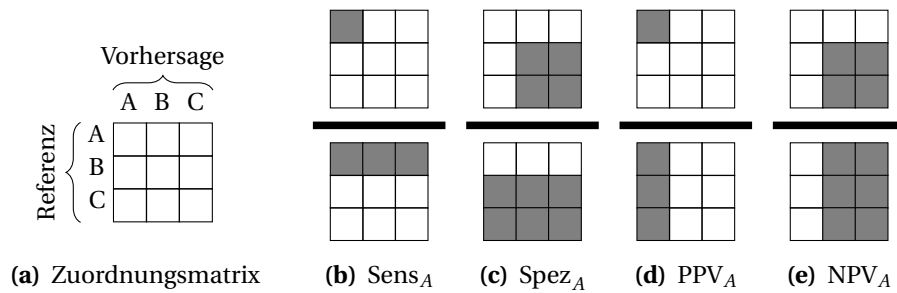
Zuordnungsmatrizen mehrerer Tests (z. B. der einzelnen Surrogatmodelle einer Resampling-Validierung) werden oft addiert. Das ist erlaubt, wenn die Modelle hinreichend ähnlich sind. Umgekehrt kann man die Zuordnungsmatrix eines Tests mit mehreren Testproben als die Summe der Zuordnungsmatrizen der einzelnen Proben auffassen.

Die Zuordnungsmatrix liefert einen sehr detaillierten Überblick über die Leistungsfähigkeit des Klassifikationsmodells.

## Von der Zuordnungsmatrix abgeleitete Kenngrößen

Aus der Zuordnungsmatrix lassen sich verschiedene Kenngrößen ableiten, die schnell einen Überblick über bestimmte Aspekte der Qualität der Vorhersagen des Modells liefern [159, 220, 221]. Diese Kenngrößen sind Verhältnisse der Summen verschiedener Teile der Zuordnungsmatrix.

Im Folgenden werden insbesondere Kenngrößen verwendet, die zur Beurteilung von diagnostischen Tests in der Medizin eingesetzt werden [159, 220, 222, 223]. Andere Dis-



**Abbildung 4.8** Schematische Darstellung der Zuordnungsmatrix (a) und wichtiger Kenngrößen für geschlossene Mehrklassen-Klassifikation (b) – (e). Die Kenngrößen sind Bruchteile von bestimmten Bereichen der Zuordnungsmatrix. Die Teile der Zuordnungsmatrix, die für den Zähler und Nenner der jeweiligen Kenngröße (bezogen auf Klasse  $G = A$ ) aufsummiert werden, sind schattiert. (nach [CB3, CB6])

ziplinen nutzen dieselben Kenngrößen unter anderem Namen. Die Kenngrößen werden zunächst für binäre Klassifikationsprobleme mit den Klassen „krank“ (Test auf die Krankheit positiv) und „nicht krank“ (Test negativ) definiert. Sie beantworten Fragen wie:

**Trefferquote:** Welcher Anteil der Testergebnisse ist richtig?

**Fehlerquote:** Welcher Anteil der Testergebnisse ist falsch?

**Sensitivität  $Sens_G$ :** Welchen Anteil der Proben, die tatsächlich von Patienten mit Krankheit  $G$  stammen, erkennt der Test als „krank“?  
Die Sensitivität entspricht der Trefferquote für eine bestimmte Klasse. Sie wird auch als Richtig-Positiv-Quote bezeichnet.

**Spezifität  $Spez_G$ :** Welchen Anteil der Proben, die tatsächlich von Patienten ohne Krankheit  $G$  stammen, erkennt der Test als „nicht krank“?  
Bei der geschlossenen Klassifikation gilt Spezifität = 1 – Falsch-Negativ-Quote.

**Positiver Vorhersagewert oder positiver prädiktiver Wert  $PPV_G$ :** Wenn der Test „krank“ ergibt, wie häufig liegt tatsächlich Krankheit  $G$  vor?  
Der positive Vorhersagewert wird auch *Relevanz* und Englisch *precision*<sup>(i)</sup> genannt.

**Negativer Vorhersagewert oder negativer prädiktiver Wert  $NPV_G$ :** Wenn der Test „nicht krank“ ergibt, wie häufig liegt Krankheit  $G$  tatsächlich nicht vor?

**Positives Wahrscheinlichkeitsverhältnis  $LR^+$ :** Wieviel häufiger ist ein positives Testergebnis für tatsächlich kranke Personen, als für Personen, die tatsächlich nicht Krankheit  $G$  haben?

**Negatives Wahrscheinlichkeitsverhältnis  $LR^-$ :** Wieviel häufiger ist ein negatives Testergebnis für tatsächlich kranke Personen, als für Personen, die tatsächlich nicht Krankheit  $G$  haben?

<sup>(i)</sup> Diese *precision* hat nichts mit der Präzision im Sinne von geringer Varianz zu tun.

Die Vorhersagewerte sind inverse Größen zu Sensitivität und Spezifität. Sensitivität und Spezifität beschreiben die Verteilung der Diagnose oder Vorhersage als Funktion der wahren Klasse. Demgegenüber geben die Vorhersagewerte die Verteilung der wahren Klassen als Funktion der Diagnose an. Insbesondere die Begriffe *Sensitivität* und *Spezifität* unterscheiden sich also von der in der analytischen Chemie üblichen Definition [224]. Die Spezifität eines Klassifikationsmodells bzw. eines medizinischen Tests ist ein quantitativer Begriff und entspricht daher eher der Selektivität in der analytischen Chemie [225]. Allerdings mit dem wichtigen Unterschied, dass mangelnde Spezifität [225] nicht zwingend auf Interferenzen zwischen Analyten zurückgehen muss, sondern insbesondere auch durch Unterschiede zwischen Individuen entstehen kann.

Für Anwender einer Diagnostik oder eines Klassifikationsmodells sind die Vorhersagewerte normalerweise wichtiger als Sensitivität und Spezifität: Ärzte und Patienten wollen wissen, ob (oder mit welcher Wahrscheinlichkeit) der Patient tatsächlich krank ist, nachdem die Diagnose „krank“ ergeben hat. Die Vorhersagewerte sind von der Häufigkeit der jeweiligen Klassen abhängig (siehe auch [226]), im Kontext medizinischer Diagnostik also der Prävalenz oder der Inzidenz. Die Prävalenz einer Krankheit kann jedoch in verschiedenen Patientenpopulationen sehr unterschiedlich sein. Unterschiede zwischen den relativen Häufigkeiten der Klassen in den Testdaten und der Prävalenz der Klassen in der Zielpopulation der Diagnostik sollten also korrigiert werden. Das betrifft auch alle Kenngrößen, die mehrere Klassen zusammenfassen, also zum Beispiel Trefferquote, Fehlerquote<sup>(j)</sup> oder auch die zufällige Übereinstimmung, die in die Berechnung der  $\kappa$ -Statistik einfließt.

Das positive Wahrscheinlichkeitsverhältnis gibt an, wie stark ein positives Testergebnis die Chancen (Kap. 4.4.2) des Patienten ändert, tatsächlich die getestete Krankheit zu haben. Je größer es ist, desto stärker verschieben sich die Chancen aufgrund eines positiven Testergebnisses in Richtung „krank“.  $LR^+$  und  $LR^-$  können alle Werte zwischen 0 und  $\infty$  annehmen. Ein Wahrscheinlichkeitsverhältnis  $> 1$  bedeutet, dass die Chancen steigen, dass die Krankheit vorliegt. Wahrscheinlichkeitsverhältnisse  $< 1$  senken die Chancen, dass die Krankheit vorliegt.  $LR^+$  sollte also möglichst groß und  $LR^-$  möglichst klein sein.

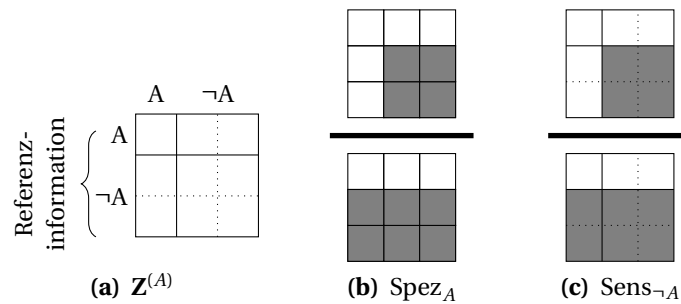
Die Wahrscheinlichkeitsverhältnisse sind ein Maß für den Zugewinn an Information durch den Test. Dieser Informationsgewinn ist unabhängig von der Prävalenz der Krankheit in der betrachteten Patientenpopulation [222, 223, 227], aber unterschiedlich, je nachdem, ob der Test positiv ( $LR^+$ ) oder negativ ( $LR^-$ ) ausfällt:

$$LR_G^+ = \frac{\text{Sens}_G}{1 - \text{Spez}_G} \quad (4.15)$$

$$LR_G^- = \frac{1 - \text{Sens}_G}{\text{Spez}_G} \quad (4.16)$$

Ein Beispiel: eine Krankheit habe Prävalenz (*a-priori*-Wahrscheinlichkeit)  $\frac{1}{3}$  in der untersuchten Patientenpopulation, jeder dritte Patient ist wirklich krank. Die Chance, dass ein Patient die Krankheit hat, ist  $1 : 2 = \frac{1}{2}$ . Auf jeden kranken Patienten kommen 2 Pati-

<sup>(j)</sup> Diese Quoten werden oft auch als „Raten“ bezeichnet, obwohl sie keinen zeitlichen Bezug haben und die Einheiten in Zähler und Nenner gleich sind.



**Abbildung 4.9** (a) – (c) Erweiterung der Kenngrößen auf Mehrklassen-Probleme: Konstruktion der *Dummy*-Klasse  $\neg G$  (nicht Klasse  $G$ ) am Beispiel von Klasse  $A$ . (nach [CB1])

enten, die die Krankheit nicht haben. Nun wird ein diagnostischer Test mit Sensitivität und Spezifität von jeweils 80 % durchgeführt. Ist das Testergebnis auf die Krankheit positiv (Vorhersage „krank“), so erhöhen sich die Chancen, dass der Patient tatsächlich krank ist, um den Faktor  $LR^+ = \frac{0,80}{0,20} = 4$  auf  $\frac{1}{2} \cdot 4 = 2 = \frac{2}{1} = 2 : 1$ . In dem Teil der Patientenpopulation mit positiven Testergebnissen kommen also 2 tatsächlich kranke Patienten auf jeden falsch-positiven Patienten (Test sagt „krank“, obwohl der Patient gar nicht diese Krankheit hat). Durch das positive Testergebnis steigt die Wahrscheinlichkeit, dass der Patient tatsächlich krank ist, auf  $\frac{2}{3}$  (positiver Vorhersagewert bei einer Prävalenz von  $\frac{1}{3}$ ). Ist das Testergebnis negativ, so verringert sich die Chance, dass der Patient tatsächlich krank ist, um  $\frac{1}{LR^-}$  auf  $\frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8} = 1 : 8$ . Unter den Patienten mit negativen Testergebnis kommt also ein tatsächlich kranker Patient (mit falsch-negativer Diagnose) auf 8 Patienten, die tatsächlich nicht diese Krankheit haben. Die Wahrscheinlichkeit, dass der Patient krank ist, obwohl das Testergebnis negativ ausgefallen ist, beträgt also  $\frac{1}{9} \approx 11\%$  (negativer Vorhersagewert bei Prävalenz  $\frac{1}{3}$ ).

Die Spezifität kann in der Regel auf Kosten der Sensitivität erhöht werden und umgekehrt. Dementsprechend bedeutet ein positives Testergebnis mehr oder weniger Informationszuwachs (und ein negatives Testergebnis weniger oder mehr). Das spiegelt sich im Unterschied zwischen  $LR^+$  und  $LR^-$  wider.

Die Definitionen aus der medizinischen Diagnostik beziehen sich zunächst darauf, ob eine bestimmte Krankheit vorliegt oder nicht. Das entspricht einer Einklassen-Klassifikation. Im Rahmen dieser Arbeit werden diese Begriffe auch auf Mehrklassenprobleme (engl. *multi-class classification*) angewendet. Dazu wird statt der Klassenzugehörigkeit „krank“ im Fall des binären medizinischen Tests jeweils eine bestimmte Klasse  $G$  eingesetzt und alle anderen Klassen werden zu einer *Dummy*-Klasse „nicht  $G$ “ ( $\neg G$ , nicht Krankheit  $G$ ) zusammengefasst (Abb. 4.9). Medizinisch beschreiben geschlossene Mehrklassen-Klassifikationen zum Beispiel eine Differentialdiagnostik, die mehrere Alternativen betrachtet. Offene Mehrklassen-Klassifikationen beschreiben Situationen, in denen sich mehrere Krankheiten gegenseitig nicht ausschließen.

Im Unterschied zur binären Klassifikation ist die Einklassen-Klassifikation darauf ausgelegt, dass die „negativ“-Klasse schlecht definiert (engl. *ill defined*) ist: „nicht Krankheit  $G$ “ kann alles mögliche bedeuten. Das Zusammenfassen aller anderen Klassen zu „nicht  $G$ “ hat wichtige Konsequenzen für die Kennwerte, die sich auf „nicht  $G$ “ beziehen, also insbesondere Spezifität und negativen Vorhersagewert. Das wird am Beispiel einer

Differentialdiagnostik besonders deutlich. Die in Kapitel 17 vorgestellten Modelle unterscheiden normales Hirngewebe von Astrozytom- und Lymphomgewebe. Die Spezifität für die Erkennung der Astrozytome fasst nun normales und Lymphomgewebe zusammen. Die Erkennung dieser beiden Gewebetypen ist nicht nur unterschiedlich schwierig (die Raman-Spektren von normalem Hirngewebe unterscheiden sich stärker von den Raman-Spektren von Tumorgeweben als deren Spektren untereinander), sondern auch von wesentlich unterschiedlicher Wichtigkeit: es ist klar, dass jeder Patient auch normales Hirngewebe hat. Insofern bedeutet „normales Gewebe“ nur, dass weitere Messungen für die Differentialdiagnostik notwendig sind. Demgegenüber bedeutet „Lymphom“ den Ausschluss eines Astrozytoms. Damit ist für die Differentialdiagnostik die Verwechslung zwischen Astrozytomen und Lymphomen sowohl wahrscheinlicher (weil die Spektren ähnlicher sind), als auch schwerwiegender als wenn das jeweilige Tumorgewebe für normal gehalten wird. Normalgewebe im Testdatensatz verwässert also wichtige Aspekte der Spezifität bei der Erkennung der jeweiligen Tumorart. Zusätzlich muss bei der Interpretation von Spezifitäten (und negativen Vorhersagewerten) bei Mehrklassenproblemen immer beachtet werden, dass bereits zufällige Zuordnungen zu hohen Spezifitäten führen können<sup>(k)</sup>. Selbstverständlich müssen auch beobachtete Sensitivitäten und positive Vorhersagewerte im Vergleich zur trivial erreichbaren Übereinstimmung betrachtet werden. Die trivial erreichbaren „positiven“ Kennwerte Sensitivität und positiver Vorhersagewert sinken aber bei Mehrklassenproblemen. Demgegenüber steigen die trivial erreichbaren Werte für die „negativen“ Kenngrößen Spezifität und negativer Vorhersagewert. Vernachlässigen der trivial erreichbaren Kenngrößen beeinflusst die Interpretation von Sensitivität und positivem Vorhersagewert bei der Mehrklassenklassifikation also nicht so gravierend wie die von Spezifität und negativem Vorhersagewert (vgl. auch [CB3]).

Mit der Dummy-Codierung ergeben sich aus den oben angeführten Fragen direkt folgende Berechnungsvorschriften [159, 220, 222, 223]:

$$\text{Sens}_G = \frac{\mathbf{Z}_{G,G}}{\sum_n r_G} \quad (4.17)$$

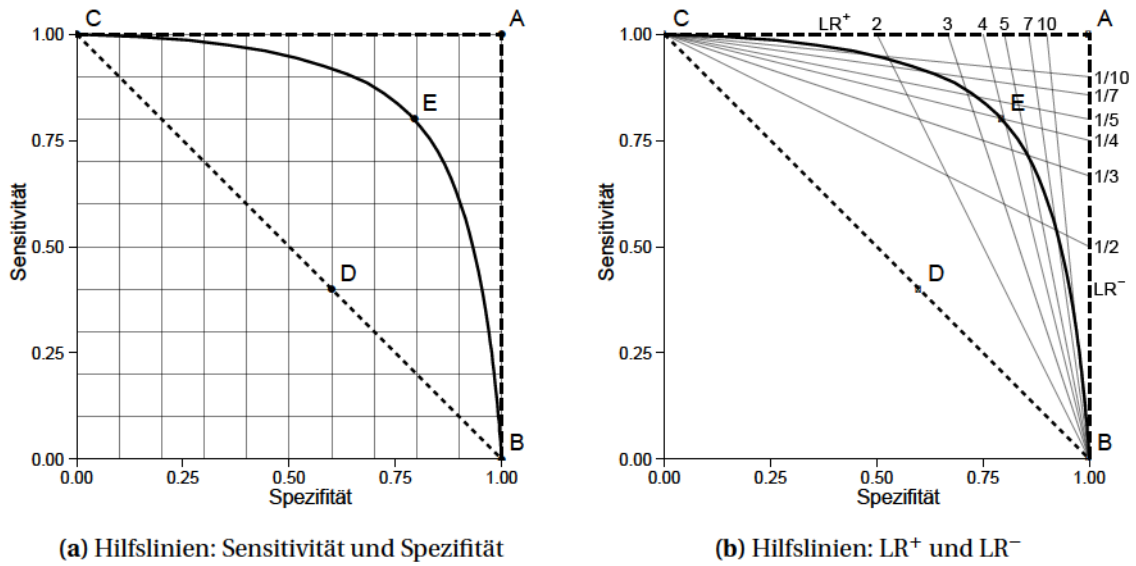
$$\text{Spez}_G = \frac{\mathbf{Z}_{G,G}}{\sum_n p_G} \quad (4.18)$$

$$\text{PPV}_G = \frac{\sum \mathbf{Z}_{i=G,j=G}}{\sum \mathbf{Z}_{i,j=G}} \quad (4.19)$$

$$\text{NPV}_G = \frac{\sum \mathbf{Z}_{i \neq G,j \neq G}}{\sum \mathbf{Z}_{i,j \neq G}} \quad (4.20)$$

Die Klassenzugehörigkeit  $g_{\neg G}$  der Dummy-Klasse  $\neg G$  („nicht Klasse  $G$ “) ist dabei  $g_{\neg G} = 1 - g_G$  (Abb. 4.9). Bei geschlossenen (engl. *closed world*) Klassifikationsmodellen erlaubt die Summenbedingung Gleichung 4.11 zusätzlich die Berechnung als Summe der Klas-

<sup>(k)</sup> Ansätze wie die  $\kappa$ -Statistik korrigieren die beobachteten richtigen Zuordnungen um die für Raten oder andere triviale Modelle erwartete Übereinstimmung. In der Praxis ist allerdings oft unklar, wieviele richtige Vorhersagen trivial erreicht werden können. Insbesondere hängt die erwartete Übereinstimmung durch Raten von der Prävalenz der einzelnen Klassen in der Anwendungssituation ab.



**Abbildung 4.10** Dasselbe Sensitivitäts-Spezifitäts-Diagramm, (a) mit Hilfslinien zum Ablesen von Sensitivität und Spezifität, (b) mit Hilfslinien zum Ablesen der Wahrscheinlichkeitsverhältnisse.

senzugehörigkeiten der anderen Klassen:  $g_{-G} = 1 - g_G = \sum_{g \neq G} g_g$ . Wie die Zuordnungsmatrix können auch die Kenngrößen als Funktion formuliert werden.

### Spezifitäts-Sensitivitäts-Diagramm und Receiver-Operating-Curve

Die *Receiver Operating Curve* (ROC) stellt  $1 - \text{Spezifität}$  auf der Abszisse der Sensitivität auf der Ordinate gegenüber.

In dieser Arbeit wird das Spezifitäts-Sensitivitäts-Diagramm (Abb. 4.10) verwendet, so dass die Abszisse direkt die Spezifität angibt. Dabei handelt es sich also um eine gespiegelte Form der ROC, die in allen wichtigen Charakteristika der ROC entspricht.

Jedes Klassifikationsmodell kann in diesen Diagrammen mit seiner Sensitivität und Spezifität eingetragen werden, z. B. Abbildung 4.10 Punkt E. Ein *ideales* (perfektes) Modell erreicht 100 % Sensitivität und gleichzeitig 100 % Spezifität, wird also in der (1; 1)-Ecke aufgetragen (Abb. 4.10 Punkt A). Zusätzlich gibt es immer zwei triviale Klassifikationsmodelle, die die betrachtete Klasse immer vorhersagen (Abb. 4.10 Punkt C) beziehungsweise immer ablehnen (Punkt B). Sie haben also 100 % Sensitivität bei 0 % Spezifität und umgekehrt. Wird die betrachtete Klasse zufällig (aber mit einer bestimmten Wahrscheinlichkeit  $p$ ) vorhergesagt, so entstehen Modelle, deren Sensitivität  $p$  und deren Spezifität  $1 - p$  ist. Sie bilden die Verbindungsgerade zwischen den beiden trivialen Modellen (Abb. 4.10 Punkt D und gepunktete Linie).

Viele Klassifikationsalgorithmen errechnen die Klassenzugehörigkeit zunächst als metrischen *Score*. Das ist im Fall der logistischen Regression das Odds-Ratio oder die Klassenzugehörigkeitswahrscheinlichkeit, bei der LDA die LDA-Scores (Koordinaten im neuen Koordinatensystem) oder wiederum die Klassenzugehörigkeitswahrscheinlichkeit. Die Umwandlung in eine harte Klassenvorhersage erfolgt dann anhand eines Grenzwertes zu dem ein entsprechender Arbeitspunkt im Spezifitäts-Sensitivitäts-Diagramm gehört.

Wird der Grenzwert variiert, so verändern sich Sensitivität und Spezifität. Im Spezifitäts-Sensitivitäts-Diagramm entsteht eine Kennlinie (durchgezogene Linie durch die Punkte B–E–C). Die Kennlinie für das ideale Modell ist gestrichelt (Punkte B–A–C), die für das zufällig ratende Modell gepunktet (durch B–D–C).

Genau genommen sind die harten Modelle neue Klassifikationsmodelle. Ein einfacher Grenzwert ist nur *eine* Möglichkeit, harte Vorhersagen zu erhalten. Eine beliebte Erweiterung dieses Konzepts ist zum Beispiel, bei mittleren *Scores* die Aussage als unsicher zu verweigern (engl. *reject*).

Das Sensitivitäts-Spezifitäts-Diagramms hilft bei der Wahl des Arbeitspunkts. Beliebte Kriterien sind minimaler Abstand zum idealen Modell, maximaler Abstand zur zufälligen Vorhersage,  $LR^+$  und  $LR^-$ , die mindestens benötigte Sensitivität und Spezifität oder das angestrebte Verhältnis zwischen falsch-negativ-Quote und falsch-positiv-Quote. Die so ermittelten Arbeitspunkte können sehr unterschiedlich sein [228]. Welche Kriterien sinnvoll sind, hängt auch von der Anwendung des Tests ab. Die Bestimmung des *optimalen* Arbeitspunkts anhand der Kennlinie ist eine datengesteuerte Modellselektion (Kap. 5.2.1).

Da die hier verwendeten Klassifikationsmodelle *a-posteriori*-Wahrscheinlichkeiten vorhersagen, ist  $\frac{1}{n_g}$  (eins durch die Anzahl der Klassen) ein „natürlicher“ Grenzwert, der sich ohne datengesteuerte Optimierung ergibt. Wenn die *a-posteriori*-Wahrscheinlichkeit gut kalibriert ist, sind an diesem Arbeitspunkt Sensitivität und Spezifität gleich.

### Messunsicherheit bei der Bestimmung der Kenngrößen

Die angesprochenen Kenngrößen werden aus der Häufigkeit der jeweiligen Ereignisse geschätzt. Der Test eines Klassifikationsmodells kann als Bernoulli-Experiment aufgefasst und mit Hilfe der Binomialverteilung beschrieben werden. Ist der wahre Wert der Kenngröße  $p$  und es werden  $n$  Tests durchgeführt, so ist die Wahrscheinlichkeit  $\Pr(k)$ ,  $k$  Ereignisse zu beobachten:

$$\Pr(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad (4.21)$$

Für Erwartungswert  $E(k)$  und Varianz  $s^2(k)$  gilt

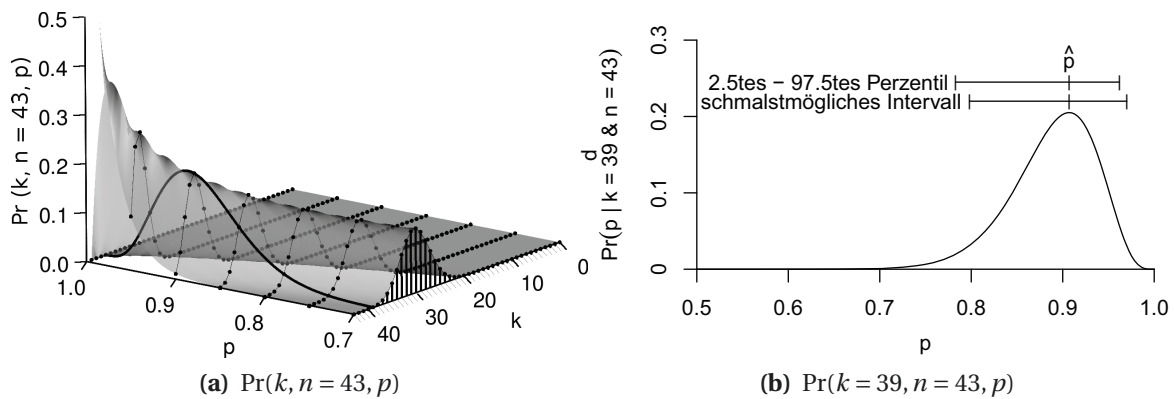
$$E(k) = np \quad \text{und} \quad (4.22)$$

$$s^2(k) = np(1-p) \quad (4.23)$$

$k$  und  $n$  sind die Summen aus der Zuordnungsmatrix, die im Zähler und Nenner der jeweiligen Kenngröße stehen (vgl. Abb. 4.8 und 4.9). Der Stichprobenumfang  $n$  im Nenner des Bruchs ist also bei den klassenspezifischen Kenngrößen nur ein Teil der Gesamtprobenzahl.

Die gesuchte Kenngröße  $p$  wird aus den beobachteten Ereignissen geschätzt:

$$\hat{p} = \frac{k}{n}$$



**Abbildung 4.11** Die Wahrscheinlichkeit  $\Pr(k, n, p)$ , bei  $n$  Tests  $k$  Ereignisse zu beobachten. (a) Für jedes  $k$  folgt  $\Pr$  einer Binomialverteilung (Punkte). Aussagen über  $p$  lassen sich anhand der Schnittebene für das beobachtete  $k$  treffen (Linie). (b) Schnitt entlang  $k = 39$  mit zwei 95 % Konfidenzintervallen für  $p$ : oben 2,5tes bis 97,5tes Perzentil, unten das schmalstmögliche Intervall.

Konfidenzintervalle für  $p$  können nun entweder direkt aus der Binomialverteilung<sup>(1)</sup> berechnet werden. Das Konfidenzintervall  $[u, o]$  mit dem Konfidenzniveau  $1 - \alpha$  ist

$$(n + 1) \int_u^o \Pr(k, n, p) dp = 1 - \alpha \quad \text{mit der Normierung} \quad (4.24)$$

$$c \int_0^1 \Pr(k, n, p) dp = 1 \quad (4.25)$$

beschrieben. Den Normierungsfaktor  $c = n + 1$  erhält man durch Integration über alle  $p$ . Das Integral muss dann 1 ergeben, da auf jeden Fall irgendein  $p$  vorliegt. Abbildung 4.11 veranschaulicht das Vorgehen: Für den Stichprobenumfang  $n$  wird zunächst für alle  $p \in [0, 1]$  die Wahrscheinlichkeit  $\Pr(p | k, n)$  berechnet,  $k$  Ereignisse zu beobachten (Abb. 4.11). Die Grenzen des schmalstmöglichen Konfidenzintervalls werden von  $u = o = \frac{k}{n} = \hat{p}$  ausgehend berechnet. Das Intervall wird immer in die Richtung mit der höheren Wahrscheinlichkeit erweitert, bis das gesuchte Vertrauenslevel  $1 - \alpha$  erreicht ist. Dieses Intervall ist das *schmalstmögliche* Konfidenzintervall zu den gegebenen Parametern  $k, n$  und  $\alpha$  (Abb. 4.11b, unteres Konfidenzintervall). Alternativ können die Intervallgrenzen so bestimmt werden, dass  $\int_0^u \Pr(k) dp = \frac{\alpha}{2}$  und  $\int_o^1 \Pr(k) dp = \frac{\alpha}{2}$  (Abb. 4.11b, oberes Konfidenzintervall). Zentrierte Konfidenzintervalle sind nur für bestimmte Kombinationen von  $k, n$  und  $\alpha$  möglich. Einseitige Intervalle erhält man von  $u = 0$  oder  $o = 1$  ausgehend.  $\Pr(k)$  kann mit Hilfe der Beta-Verteilung ausgedrückt werden [229]. Ross [229] diskutiert diese Herleitung der Konfidenzintervalle unter der Bezeichnung „Integration der Bayesschen *a-posteriori*-Wahrscheinlichkeit“. Diese Methode ist auch für  $\hat{p}$  nahe 0 oder 1 und für kleine  $n$  genau. Weitere Methoden zur Berechnung von Konfidenzintervallen für Anteile,

<sup>(1)</sup> Die für die Näherung als Normalverteilung mit Mittelwert  $p$  und Varianz  $\frac{p(1-p)}{n}$  erforderlichen Patientenzahlen von  $n > 5 \max(p, 1 - p)$  werden in der Biospektroskopie für Modelle mit einer erstrebenswerten Qualität (z. B.  $p > 90\%$ ) in der Regel nicht erreicht.



insbesondere bei kleinen Stichprobenumfängen, finden sich in [230, 231].

Die Bayessche Berechnung der *a-posteriori*-Wahrscheinlichkeit gibt immer die Veränderung gegenüber einer *a-priori*-Wahrscheinlichkeit an. Die hier vorgestellte Berechnung nutzt implizit die Gleichverteilung als *a-priori*-Wahrscheinlichkeit (Bayes-Laplace *a-priori*-Wahrscheinlichkeit). Häufig wird stattdessen der sogenannte *Jeffrey's prior* genutzt. Der Unterschied ist aber praktisch selbst bei den in der Biospektroskopie typischen extrem kleinen Stichprobenumfängen bedeutungslos.

Bei der Anwendung dieser Konfidenzintervalle in der Biospektroskopie treten zwei Probleme auf:

1. Zum Einen ist die Zahl der Tests  $n$ , also der statistisch relevante Stichprobenumfang, in der Biospektroskopie oft nicht bekannt.

Das ist immer dann der Fall, wenn von einem Patienten mehrere Spektren im Datensatz sind. Die Patienten sind auf jeden Fall statistisch unabhängig voneinander. Die zusätzlichen Spektren jedes Patienten tragen durchaus zum Gesamtinformationsgehalt bei, jedoch nicht in dem Umfang wie ein Spektrum eines weiteren Patienten. Der *effektive Stichprobenumfang* liegt zwischen der Anzahl an Patienten und der Anzahl an Spektren.

Die Anzahl an Patienten kann aber als konservative Grenze genutzt werden.

2. Zum Anderen läßt diese Berechnung den Anteil der zufälligen Unsicherheit außer Acht, der aus den zufälligen Unterschieden zwischen dem validierten Modell und den Surrogatmodellen stammt (Instabilität). Das Konfidenzintervall bezieht sich also ausschließlich auf die zufällige Unsicherheit aufgrund der begrenzten Anzahl an Testpatienten.

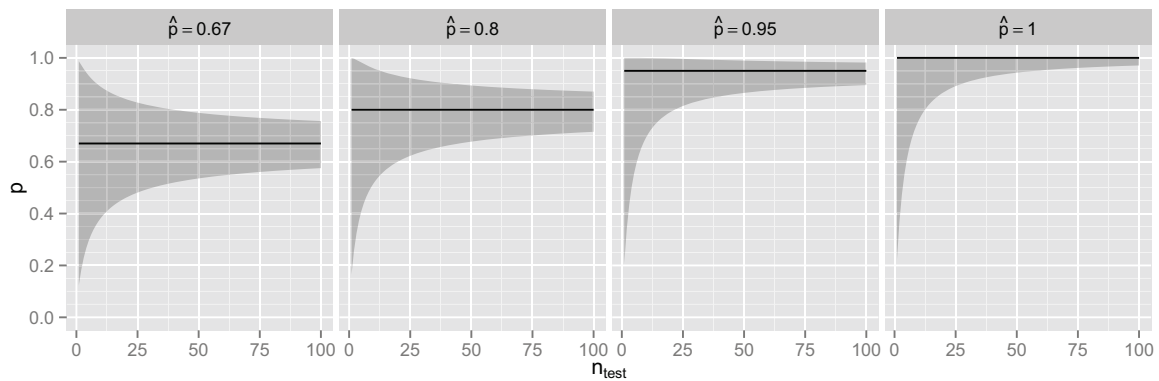
Die zusätzliche Unsicherheit aufgrund der (In)stabilität der Modelle kann im Rahmen einer Resampling-Validierung abgeschätzt werden (vgl. Kap. 7).

Nichtparametrische Ansätze zur Bestimmung der Konfidenzintervalle [159] nutzen Wiederholungsmessungen. Die Iterationen bei  $k$ -facher Kreuzvalidierung oder Bootstrap-Validierung sind allerdings keine echten Wiederholungsmessungen (vgl. [232]). Damit stellt sich wieder das Problem, dass die effektive Anzahl an Wiederholungsmessungen unbekannt ist.

## Notwendige Testpatientenzahl

Wie bei der Modellbildung bewirkt eine kleine Zahl von Proben auch bei der Validierung eine große zufällige Unsicherheit. Die zufällige Unsicherheit beim Testen hängt nach Gleichung 4.23 im Gegensatz zur zufälligen Unsicherheit bei der Modellbildung von der *absoluten* Testpatientenzahl in Nenner der jeweiligen Kenngröße ab. Damit ist die Ungenauigkeit in der Regel für die Sensitivität der Klasse mit der kleinsten Testpatientenzahl am größten. Verglichen mit typischen Varianzunsicherheiten bei der Validierung von Regressionsmodellen (Kalibrierungen) ist die Varianzunsicherheit bei der Messung dieser Anteile sehr groß.

Abbildung 4.12 veranschaulicht die Breite des Konfidenzintervalls für verschiedene beobachtete Sensitivitäten in Abhängigkeit von der Anzahl der Testproben. Angenommen, 19 von 20 Testpatienten der fraglichen Klasse wurden erkannt. Dann ist die Punktschätzung der Sensitivität 95 %, was als sehr gutes Ergebnis bewertet würde. Allerdings



**Abbildung 4.12** 95 % Konfidenzintervall bei der Bestimmung der Sensitivität  $p$  in Abhängigkeit der Testprobenzahl  $n$ , wenn  $\hat{p} = 67, 80, 95$  oder  $100$  % Sensitivität beobachtet werden. (nach [CB3])

reicht das 95 %-Konfidenzintervall von knapp  $100$  % bis knapp unter  $80$  %. Bei vielen als nicht zu schwierig eingeschätzten biospektroskopischen Klassifikationsproblemen würden  $80$  % Sensitivität als schlechte Modellqualität eingeordnet werden. Unter der Annahme, dass eine solche sehr gute Sensitivität von  $95$  % beobachtet wird, sind aber etwa  $100$  Testpatienten notwendig, um das untere Ende des Konfidenzintervalls auf  $90$  % („gut“) zu heben. Damit die Breite des  $95$  % Konfidenzintervalls unter  $10$  %-Punkte sinkt, hätten mehr als  $150$  Proben (Patienten) getestet werden müssen. Soll das Konfidenzintervall nicht breiter als  $5$  %-Punkte sein, wären über  $600$  Testproben erforderlich.

Bei der Abschätzung der notwendigen Testpatientenzahl kann insbesondere die Interpretation der Grenzen des Konfidenzintervalls im Hinblick auf praktisch relevante Modellqualitäten helfen.

Testergebnisse werden häufig für Modellvergleiche verwendet: es soll nachgewiesen werden, dass ein Modell besser als ein anderes ist oder aus mehreren Modellen soll das beste herausgesucht werden. Aufgrund der großen Varianz (Gl. 4.23) der gemessenen Anteile lassen sich Klassifikationsmodelle nur schwer anhand der harten Kenngrößen wie Sensitivität, Spezifität oder auch der Fehlerrate vergleichen. Fleiss, Levin und Paik [223] geben eine Übersicht über notwendige Testprobenzahlen zum Unterscheiden zweier Modelle. Ein Beispiel. Zwei Modelle mit einer wahren Sensitivität  $p = 75$  % und  $85$  % sollen als unterschiedlich erkannt werden. Diese Werte könnten den Validierungsergebnissen für die logistischen Regressionsmodelle des Astrozytom-Gradings entnommen sein, zum Beispiel bei der Beurteilung, ob zusätzlich zum Spektralbereich der  $\nu$ CH-Streckschwingungen auch der Fingerprintbereich zur Erkennung von normalem Gewebe notwendig ist (Abb. 16.8 auf Seite 125). Für Konfidenzniveau  $1 - \alpha$  und Teststärke  $1 - \beta$  werden die üblichen  $95$  % und  $80$  % gewählt. Sind die Testproben für die beiden Modelle voneinander unabhängig, so sind für jedes Modell  $270$ , also insgesamt  $540$  unabhängige Testproben für die betrachtete Klasse nötig [223]. Können dieselben Testproben verwendet werden, so kann ein gepaarter Test durchgeführt werden (McNemar-Test) [223], der mit weniger Testproben auskommt. Die bessere Unterscheidungskraft des gepaarten Tests beruht darauf, dass die Ergebnisse zunächst in zwei Gruppen unterschieden werden können: diejenigen Proben, bei denen beide Modelle dieselbe Vorhersage liefen und diejenigen Proben, für die sich die Vorhersagen unterscheiden. Erstere Proben tra-

gen nicht zur Unterscheidung der Modelle bei. Ob ein Modell besser ist als das andere kann allein anhand der zweiten Gruppe entschieden werden. Im besten Fall macht das bessere Modell keine Fehler, die das schlechtere Modell nicht auch macht. Eine Abschätzung der erforderlichen Probenzahl nach [233] ergibt, dass dann bereits mit 76 Proben erkannt werden kann, dass das bessere Modell tatsächlich besser ist. Im schlechtesten Fall verteilen sich die falschen Vorhersagen ausschließlich auf Proben, die nur das eine oder nur das andere Modell falsch vorhersagt. Dann sind 311 Proben erforderlich, um das bessere Modell als besser zu erkennen. Gegenüber dem ungepaarten Test könnten also 40 bis 85 % der Testproben eingespart werden. Mit weniger als 76 unabhängigen Testproben kann die Verbesserung aber auf keinen Fall nachgewiesen werden. Die 15 Patienten, von denen eindeutig normales Hirngewebe verfügbar ist, sind also selbst im Idealfall um einen Faktor 5 zu wenige, um auch nur einen einzigen gepaarten Modellvergleich zu erlauben.

Dieses Beispiel bezieht sich auf einen einmaligen Vergleich zwischen zwei Modellen. Insbesondere bei der Optimierung von Hyperparametern, zum Beispiel im Rahmen einer Rastersuche, werden aber viele Modelle verglichen. Aus statistischer Sicht handelt es sich um multiple Tests. Daher sollten von vornherein nicht zu viele Modellvergleiche geplant werden. Das Risiko eines Fehlers erster Art steigt sonst sehr stark an.

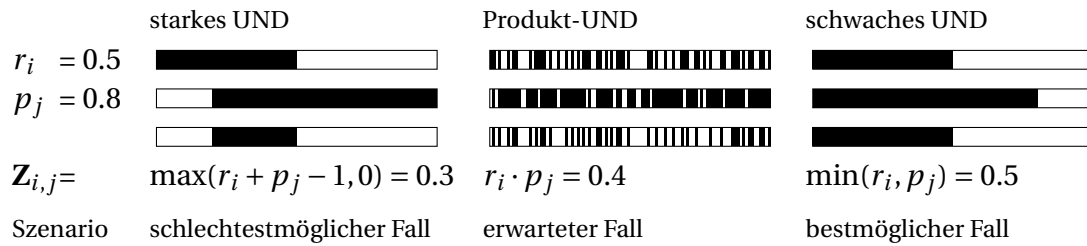
## 4.9 Validierung mit weicher Referenz

Weiche Klassifikation spielt in der Fernerkundung eine große Rolle. Dort tritt häufig das „Problem der gemischten Pixel“ auf [234]: ein Pixel eines Satellitenbildes kann zum Beispiel sowohl Wald als auch Weideland beinhalten.

Verschiedene Methoden zur Konstruktion einer weichen Zuordnungsmatrix sind in der Literatur beschrieben [191, 192, 235]. Die Zuordnungsmatrix eines Tests zählt jede Probe in Zeile  $R$  und Spalte  $P$ . Dies kann als die Schnittmenge der Referenzklasse  $R$  und der Vorhersageklasse  $P$  aufgefasst werden. Mit anderen Worten: in  $Z_{R,P}$  werden die Proben gezählt, die Element von  $R$  UND Element von  $P$  sind. Damit kann die Zuordnungsmatrix mit Hilfe der Operatoren  $\cap$  (Schnittmenge oder Konjunktion) beziehungsweise  $\wedge$  (UND) konstruiert werden. Die Zuordnungsmatrix (Gl. 4.14) wird im Folgenden auf weiche Klassenzugehörigkeiten sowohl für die Vorhersage als auch die Referenz erweitert. Dazu wird zunächst ein UND-Operator benötigt, der nicht nur 0 und 1 als Eingabewerte akzeptiert, sondern den gesamten Wertebereich von 0 bis 1.

Geeignete UND-Operatoren [236–238] sind das Minimum, auch bekannt als *schwache Konjunktion* (vorgeschlagen von Łukasiewicz [239] und Gödel [236]), das Produkt (vorgeschlagen von Reichenbach [240]) und die *starke Konjunktion*  $\max(x + y - 1, 0)$  nach Łukasiewicz [239].

Die Bedeutung dieser Operatoren kann anhand einer Situation illustriert werden, in der die Uneindeutigkeit durch geringe räumliche Auflösung entsteht (Abb. 4.13). Angenommen, im Messvolumen befindet sich eine Anzahl Zellen und die Hälfte dieser Zellen seien Tumorzellen, die andere Hälfte normale Zellen. Das Klassifikationsmodell berechnet einen Anteil von 0,8 an Tumorzellen. Im besten Fall hat das Modell die Hälfte der Zellen, die tatsächlich Tumorzellen sind, richtig erkannt. Dann ist die Überlappung (Konjunktion) zwischen Referenz und Vorhersage 0,5. Im schlechtesten Fall ordnet das Modell



**Abbildung 4.13** Die „weichen“ UND-Operatoren: hypothetische hochaufgelöste Szenarios, die in niedriger Auflösung jeweils einer Referenzzugehörigkeit von 0,5 (obere Zeile) und einer vorhergesagten Zugehörigkeit von 0,8 (mittlere Zeile) zur schwarzen Klasse entsprechen. In jeder Spalte wird die Überlappung mit dem klassischen Boole’schen UND berechnet: für jede der (aufgelösten) Positionen ist  $Z_{i,j, pos} = r_{i, pos} \wedge p_{j, pos}$ , die weiche Konjunktion ist der Anteil der Positionen, für die  $r_i$  UND  $p_j$  der schwarzen Klasse angehören. (nach [CB1])

den normalen Zellen die Klasse „Tumor“ zu. Da es aber insgesamt einen Tumoranteil von 0,8 vorhergesagt hat, besteht trotzdem eine Überlappung von 0,3. Die tatsächliche Überlappung liegt irgendwo zwischen diesen beiden Grenzen. Wo genau, hängt von der wahren räumlichen Verteilung der Tumorzellen und der räumlichen Verteilung der Vorhersagen ab. Wenn beide gleichverteilt sind, ist die Wahrscheinlichkeit, dass eine Zelle sowohl wirklich Tumorzelle ist und auch als solche vom Klassifikationsmodell erkannt wurde,  $0,5 \cdot 0,8 = 40\%$ . Damit ist die erwartete Überlappung 0,4.

Die **schwache Konjunktion** ist der Standard-UND-Operator für unscharfe Mengen [183, 238]. Dies wird von Binaghi *et al.* [191] zur Konstruktion einer weichen Zuordnungsmatrix übernommen:

$$Z^{\text{schwach}}(r_i, p_j) = \min(r_i, p_j) \tag{4.26}$$

Das Minimum gibt die größtmögliche Übereinstimmung zwischen Referenz und Vorhersage an. Während bei der klassischen Zuordnungsmatrix (Gleichung 4.13)  $\sum Z_{i,j}$  für jede Probe gleich 1 ist, kann sie bei  $Z^{\text{schwach}}$  deutlich größer sein.

Die **starke Konjunktion**

$$Z^{\text{stark}}(r_i, p_j) = \max(r_i + p_j - 1, 0) \tag{4.27}$$

wurde von Pontius und Connors [193] zur Berechnung der Zuordnungsmatrix von weichen Klassifikationsmodellen vorgestellt. Sie entspricht der kleinstmöglichen Übereinstimmung zwischen Referenz und Vorhersage.

Sowohl  $Z^{\text{schwach}}$  als auch  $Z^{\text{stark}}$  fehlen zwei Eigenschaften, die Pontius und Cheuk [192] und Silván-Cárdenas und Wang [194] für die Zuordnungsmatrix wünschen. Erstens entsprechen die Zeilen- und Spaltensummen nicht den Referenz- und vorhergesagten Klassenzugehörigkeitsvektoren. Zweitens erzeugt eine Vorhersage, die die Referenz genau reproduziert (ideales Modell) keine Diagonalmatrix. Die ideale Modellqualität ist daher schwerer erkennbar als bei der harten Zuordnungsmatrix. In der Fernerkundungsliteratur werden verschiedene Ansätze zur „Reparatur“ dieser beiden Eigenschaften verfolgt [192–194]. Alle diese Vorschläge ziehen zunächst die Übereinstimmung, also die Diagonale, ab und verteilen dann die restlichen Anteile der Referenz- und vorhergesagten Klassenzugehörigkeiten. Wie für  $Z^{\text{schwach}}$  und  $Z^{\text{stark}}$  haben auch bei den so entstandenen

Kompositmatrizen Diagonal- und Außerdiagonalelemente unterschiedliche Interpretationen (Kap. 8.1).

Das **Produkt** ist der UND-Operator in Reichenbachs Wahrscheinlichkeitslogik [240]

$$Z^{\text{prod}}(r_i, p_j) = r_i \cdot p_j \quad (4.28)$$

und ist auch für weiche Zuordnungsmatrizen eingeführt [235] und diskutiert worden [192–194].

Werden die Klassenzugehörigkeiten als Wahrscheinlichkeiten interpretiert, so ist  $Z^{\text{prod}}$  die erwartete Übereinstimmung zwischen zwei unabhängigen Prozessen, die die Referenz- und vorhergesagten Zugehörigkeiten gleichverteilt bestimmen. In der Interpretation als Mischung folgt  $Z^{\text{prod}}$  aus dem Informationsverlust durch geringe (zum Beispiel räumliche) Auflösung [192]. Angenommen, bei hoher Ortsauflösung sind harte Referenz und Vorhersage verfügbar. Nun geht die Information über die Position verloren (die hochaufgelösten Daten werden gemischt). Die erwartete Zuordnungsmatrix (auf die jeweilige Anzahl an Proben normiert) ist  $Z^{\text{prod}}$  (Abb. 4.13).

Die Interpretation der anteiligen Klassenzugehörigkeit als Mischung legt eine Behandlung analog zu Regressionsfehlermaßen nahe. In der Regression messen Residuen die Abweichung der Vorhersage von der Referenz. Analog zur Berechnung der Residuen einer Regression  $\varepsilon = \hat{y} - y$ , kann die beobachtete Zuordnungsmatrix  $Z^{\text{prod}}(r, p)$  mit der „idealen“ Zuordnungsmatrix  $Z^{\text{prod}}(r, p = r)$  verglichen werden. Das wird auch von Lewis und Brown [235] für  $Z^{\text{prod}}$  vorgeschlagen:

$$\Delta^{\text{prod}}(r, p) = Z^{\text{prod}}(r, p) - Z^{\text{prod}}(r, r) \quad (4.29)$$

Wie bei der Regression unterscheidet auch hier das Vorzeichen von  $\Delta$  Fehler durch Über- oder Unterschätzen der tatsächlichen Klassenzugehörigkeit. Daher können die Beträge von  $\Delta$  zu einem Gesamtfehler aufsummiert oder gemittelt werden (engl. *mean absolute error*, MAE). Bei geschlossenen Klassifikationssystemen bedeutet jedes Unterschätzen einer Klasse, dass andere Klassenzugehörigkeiten im gleichen Maß überschätzt werden, die Zeilensummen von  $\Delta$  sind 0 [235]. In Kapitel 8.2.2 wird dieses Vorgehen mit dem in der Regression üblicheren mittleren quadratischen Fehler (engl. *mean squared error*, MSE) verglichen.

Die Berechnung von Kenngrößen wie Sensitivität und Spezifität aus diesen verallgemeinerten Zuordnungsmatrizen ist schwieriger. Gómez, Biging und Montero [241] entwickeln entsprechende Vorschriften für die in [191] vorgestellte Zuordnungsmatrix. Diese Verfahren können allerdings nur mit harten Referenzdaten angewendet werden. Silván-Cárdenas und Wang [194] berichten, dass in der Praxis nur die Diagonale von  $Z^{\text{schwach}}$  genutzt wird. Das ist auch in der neueren Literatur [242, 243] der Fall. Daher ergeben sich aus den Kompositmatrizen [192–194] dieselben verallgemeinerten Ausdrücke für Sensitivität (engl. auch *producer's accuracy*) und positiven Vorhersagewert (engl. auch *user's accuracy*) wie für  $Z^{\text{schwach}}$  (Sensitivität und positiver Vorhersagewert: [191, 192, 194, 242]). Ausdrücke für Spezifität und negativen Vorhersagewert werden nicht gegeben, da die Interpretation der Außerdiagonalelemente unklar ist [194]. Für  $Z^{\text{prod}}$  geben Lewis und Brown [235] nur einen Ausdruck für die Gesamtfehlerquote.

In der Anwendung wird letztlich die Konstruktion nach [191] mit geringfügigen Ab-

wandlungen [192–194] favorisiert. Alternativen [193, 235] wurden vorgestellt, konnten sich aber nicht durchsetzen. Insgesamt wird das Problem der Validierung mit weichen Referenzdaten auch innerhalb der Fernerkundungsliteratur weiterhin als ungelöst eingestuft [194, 242, 243]. Silván-Cárdenas und Wang [194] sprechen von einem „*urgent need to investigate the appropriateness of the existing operators and to develop new reliable ones.*“

In Kapitel 8 und [CB1] wird ein einheitlicher Rahmen zur Interpretation der drei grundlegenden Vorschläge entwickelt und gezeigt, wie der komplementäre Satz von Operatoren  $Z^{\text{schwach}}$ ,  $Z^{\text{prod}}$  und  $Z^{\text{stark}}$  genutzt werden kann. Darauf aufbauend werden dann Ausdrücke für Sensitivität, Spezifität, positiven und negativen Vorhersagewert hergeleitet. Kapitel 8.3 diskutiert die so erhaltenen Größen im Hinblick auf systematischen und zufälligen Fehler (Bias und Varianz), und in Kapitel 8.4 wird die Implementierung als R-Paket vorgestellt.

## **Teil II**

# **Entwicklungen und Untersuchungen zur Validierung von chemometrischen Modellen im Rahmen der Dissertation**

## 5 Planung der erforderlichen Patientenzahlen

Dieses Kapitel fasst die in [CB3] vorgestellten Ergebnisse zur notwendigen Trainings- und Testprobenzahl zusammen. Der Grundgedanke dabei ist, dass es in der Praxis nicht ausreicht, ein *gutes* Modell zu trainieren: die Modellqualität muss auch (anhand von statistisch unabhängigen Testproben) nachgewiesen werden. Das bedeutet, dass bei der Planung des Stichprobenumfangs für eine Studie sowohl die notwendige Trainingsprobenzahl als auch die notwendige Testprobenzahl abgeschätzt werden müssen.

### 5.1 Abschätzung der notwendigen Trainingsprobenzahl durch Messen der Lernkurve

Ausgangspunkt ist die Frage, inwieweit der Bedarf an Trainingsproben aus einer z. B. durch iterierte Kreuzvalidierung erstellten Lernkurve abgeschätzt werden kann. Damit das funktioniert, muss die mit Hilfe eines kleinen Datensatzes (hier: bis zu 25 unabhängige Patienten/Zellkultur-Batches pro Klasse) aus einem Vorversuch erstellte Lernkurve die tatsächlich erreichbare Modellqualität (und ihre Streubreite) halbwegs verlässlich wiedergeben.

Dazu wurden PLS-LDA-Modelle auf der Basis von Raman-Spektren einzelner Zellen von 3 verschiedenen Tumorzelllinien (OCI-AML, MCF-7 und BT-20) sowie normalen Erythrozyten und Leukozyten untersucht. Die verschiedenen Zelltypen bilden eine Reihe von Klassifikationsproblemen mit sehr unterschiedlichem Schwierigkeitsgrad: aufgrund der Resonanzverstärkung des Hämoglobins (vgl. [244]) sind die Raman-Spektren von roten Blutkörperchen sehr einfach zu erkennen. Auch normale Leukozyten werden von Klassifikationsmodellen in der Regel sehr gut erkannt. Die Unterscheidung der drei Tumorzelllinien ist dagegen für unbekannte Zellkulturbatches und Messtage sehr viel schwieriger [CB3]. Der Datensatz umfasst mehrere Messtage und verschiedene Zellkultur-Batches (biologische Replikate). Diese Information wurde für die vorliegende Untersuchung bewusst ignoriert und die Daten stattdessen behandelt, als wären die einzelnen Spektren statistisch unabhängig. Dadurch wirken Batches und Messtage wie unbekannte Einflussgrößen. Das ist insofern zulässig, da hier *keine* Aussagen über die Zelllinien bzw. ihre Unterscheidbarkeit mit Hilfe von Raman-Spektren getroffen werden sollen, sondern lediglich eine Bandbreite an unterschiedlich schwierigen Klassifikationsproblemen einerseits und eine genügend große Anzahl an Spektren als „unabhängiger“ Testdatensatz andererseits benötigt wird. Insgesamt stehen etwa 2500 Spektren zur Verfügung, die recht gleichmäßig über die Klassen verteilt sind (Tab. 5.1).

Zusätzlich wurden aus den gemessenen Daten für jede Klasse Mittelwertspektrum und Kovarianzmatrix geschätzt und multivariat normalverteilt simulierte Datensätze mit diesen Charakteristika erzeugt. Hierbei sind die einzelnen Spektren tatsächlich statistisch unabhängig voneinander. Außerdem können beliebige Mengen Spektren simuliert werden, so dass genaue Referenzmessungen der jeweils erzielten Sensitivität möglich sind.



## 5.1 Abschätzung der notwendigen Trainingsprobenzahl durch Messen der Lernkurve

Klasse	Zelltyp	n <sub>Spektren</sub>	Sensitivität	
			PLS-LDA	PLS-LDA nach Batches
rbc	Erythrozyten	372	0.99 (0.96 – 0.99)	0.97 (0.96 – 0.98)
leu	Leukozyten	569	0.97 (0.96 – 0.97)	0.87 (0.84 – 0.90)
mcf	MCF-7 Brustkrebs	558	0.91 (0.90 – 0.92)	0.31 (0.24 – 0.42)
bt	BT-20 Brustkrebs	532	0.75 (0.74 – 0.76)	0.38 (0.32 – 0.45)
oci	OCI-AML3 Leukämie	518	0.89 (0.88 – 0.90)	0.30 (0.23 – 0.17)

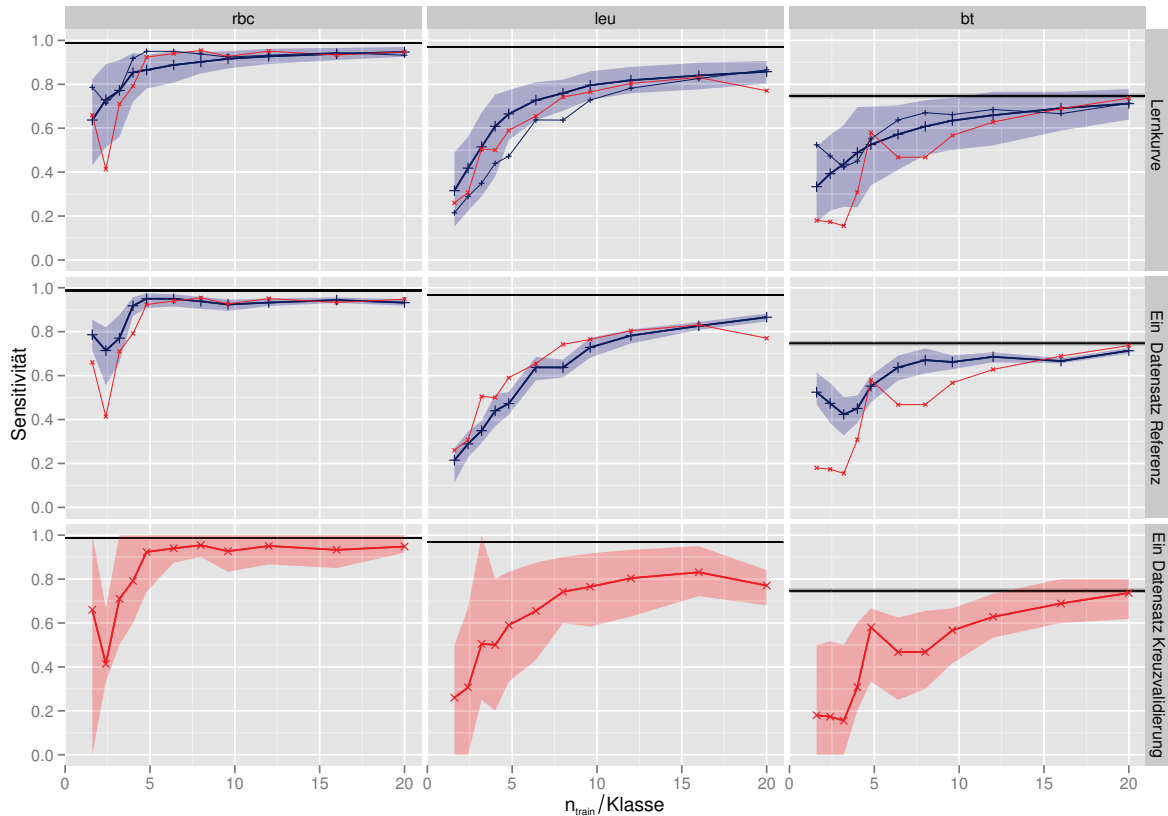
**Tabelle 5.1** Charakteristika des Datensatzes zur Lernkurvenstudie: Die angegebenen Sensitivitäten (Mittelwert, 5. und 95. Perzentil über die Iterationen) wurden mit einer 100× iterierten 5-fachen Kreuzvalidierung gemessen. Die letzte Spalte bezieht sich auf die Vorhersage von unbekanntem Batches und Messtagen, die vorletzte Spalte auf die innerhalb des Datensatzes für unbekannte Spektren der Trainingsbatches erreichbare Sensitivität. (nach [CB3])

Andererseits ist die Datenstruktur natürlich stark vereinfacht: Cluster in den Daten (weder Batches noch Messtage, noch möglicherweise vorliegende unbekannte Einflüsse). Hier werden daher nur die Ergebnisse für die Messdaten präsentiert. Die Ergebnisse für die simulierten Daten sowie weitere Details zu den Modellen können in [CB3] nachgelesen werden. Simulierte Datensätze und die echten Messdaten verhalten sich in allen wesentlichen Aspekten gleich.

Abbildung 5.1 zeigt das Verhalten der Modelle in der oberen Zeile: die Lernkurve steigt wie erwartet an, aber die tatsächliche Sensitivität schwankt stark. Diese Lernkurven wurden gemessen, indem 100 Mal ein Datensatz der entsprechenden Größe aus den Raman-Daten gezogen wurde. Gemessen wurde die Sensitivität mit allen übrigen Spektren. In der Praxis steht aber nur der begrenzte Datensatz mit 25 Patienten (Batches) pro Klasse zur Verfügung. Für diesen muss dann die Lernkurve abgeschätzt werden (mittlere Zeile). Da insgesamt nur wenige Patienten bzw. Batches vorhanden sind, sind auch die daraus gezogenen kleineren Datensätze untereinander sehr viel ähnlicher: die mittlere Sensitivität weicht deutlich von der mittleren Sensitivität in der oberen Zeile ab, und die Streuung ist sehr viel niedriger: die Modelle werden dem Modell aus allen 25 Spektren pro Klasse schnell sehr ähnlich. Tatsächlich kann jedoch auch die mittlere Zeile in der Praxis nicht gemessen werden, da auch die Testdaten nur aus dem kleinen Datensatz kommen können. Wird die Lernkurve durch iterierte Kreuzvalidierung innerhalb des kleinen Datensatzes gemessen, so ergibt sich das Bild in der unteren Zeile: die tatsächlichen Sensitivitäten aus der mittleren Zeile sind zusätzlich mit der Varianzunsicherheit aufgrund der begrenzten Anzahl unterschiedlicher Testspektren überlagert.

Für die sehr einfach erkennbaren Erythrozyten (rbc) gibt auch die mit Kreuzvalidierung gemessene Lernkurve dieses wieder. Allerdings könnte man alternativ feststellen: Klassen, deren Spektren sich so extrem von allen anderen Spektren unterscheiden, sind unproblematisch und bereits mit 5–10 unabhängigen Patienten lassen sich akzeptable Modellqualitäten erhalten. Insofern ist es für solche extrem einfach erkennbaren Klassen unnötig, die Lernkurve zu messen. Die Leukozyten sind erfahrungsgemäß [245, 246] ebenfalls gut erkennbar. Die mittels Kreuzvalidierung gemessene Lernkurve gibt den Verlauf der tatsächlichen Lernkurve gut wieder. Bei einer Extrapolation würde allerdings wohl die etwas verschlechterte beobachtete Qualität des letzten Messpunktes stören. Am

## 5 Planung der erforderlichen Patientenzahlen



**Abbildung 5.1** Messen der Lernkurve. Die schwarzen horizontalen Linien geben die beste erreichte Sensitivität an. Die dickeren Kurven geben jeweils den Mittelwert an, schattiert ist das 5. bis 95. Perzentil über die 100 Iterationen. Zur besseren Vergleichbarkeit sind die Mittelwerte der darunter liegenden Zeilen jeweils dünn in den darüberliegenden Zeilen eingetragen. Messungen mit den Referenz-Testdaten sind blau gekennzeichnet, Messungen mit iterierter Kreuzvalidierung im „kleinen“ Datensatz rot.

(obere Zeile) Die eigentliche Lernkurve: mittlere Sensitivität und 5. bis 95. Perzentil gemessen für unabhängige Ziehungen von  $n$  Spektren aus dem gesamten Datensatz, gemessen mit allen am Training unbeeiligten Spektren.

(mittlere Zeile) Steht nur ein kleiner Datensatz zur Verfügung, aus dem die noch kleineren Trainingsdatensätze gezogen werden, so kann die Lernkurve in diesem Datensatz deutlich abweichen. Zusätzlich ist die Streuung kleiner und reduziert sich schneller, da die Modelle dem gesamten kleinen Datensatz sehr schnell ähnlicher werden.

(untere Zeile) In der Praxis kommt zu dieser Abweichung noch die Varianz-Unsicherheit beim Testen: auch die Testproben müssen aus dem kleinen Datensatz gestellt werden, so dass in dem Szenario hier auch mit einer Resampling-Validierung nie mehr als 25 Spektren als Testspektren fungieren können.

wichtigsten sind in der Praxis Lernkurven wie für die MCF-7 Zelllinie, da diese Klasse selbst für Trainingsbatches deutlich schwerer zu erkennen ist. Hier weicht das Verhalten der gemessenen Lernkurve im unteren Bereich sehr stark von der echten Lernkurve in der oberen Zeile ab und das ist auch nur teilweise auf das konkrete Verhalten des kleinen Datensatzes zurückzuführen.

Weiterhin wurde abgeschätzt, wie gut PLS-LDA-Modelle für den gegebenen Datensatz überhaupt werden können. Dazu wurde eine  $100 \times$  iterierte 5-fache Kreuzvalidierung von PLS-LDA-Modellen mit 10 latenten Variablen für den gesamten Datensatz verwendet. Diese Ergebnisse sind in Abbildung 5.1 als schwarze horizontale Linien eingetragen. Interessant ist, dass alle drei Lernkurven für die MCF-7 Zelllinie diese Grenze erreichen, während das für die Erkennung der Erythrozyten und insbesondere der Leukozyten nicht der Fall ist. Das bedeutet, dass die MCF-7 Zellen im vorliegenden Datensatz entweder nur mit komplexeren Modellen oder sogar gar nicht von den anderen Klassen (besonders BT-20 und OCI) trennbar ist. Im Zusammenhang mit der Abschätzung des erforderlichen Stichprobenumfangs für das Modelltraining ergeben sich, dass aus den gemessenen Lernkurven nicht verlässlich auf die erreichbare Modellqualität geschlossen werden kann. Der gewählte Modellansatz erlaubt für die Leukozyten mit mehr Trainingsproben eine deutlich bessere Erkennung als die gemessene Lernkurve suggeriert. Hingegen suggeriert die gemessene Lernkurve für die MCF-7 Zelllinie, dass bessere Ergebnisse erreichbar wären, als es mit den betrachteten Modellen tatsächlich der Fall ist. Das bedeutet auch, dass die gemessene Lernkurve nicht zuverlässig anzeigt, dass das Klassifikationsproblem mit dem erweiterten Datensatz und dem gewählten Modellansatz nur eingeschränkt gelöst werden kann.

Praktisch ist eine Abschätzung der benötigten Trainingsprobenzahl durch Messen der Lernkurve in Situationen mit den für die Biospektroskopie typischen sehr geringen Anzahlen an Patienten oder Batches also nicht oder nur sehr schlecht möglich. Dabei ist auch zu bedenken, dass 25 unabhängige Batches pro Klasse zwar nur  $\frac{1}{15}$  Probe pro Variate ist, andererseits aber eine Studie mit 25 Patienten pro Klasse und/oder 125 unabhängigen Patienten insgesamt gegenwärtig in der Biospektroskopie einen herausragend großen Stichprobenumfang hätte (vgl. Abb. 4.4 S. 34). Hinzu kommt, dass sich die Anzahl an verfügbaren unabhängigen Proben und die Modellkomplexität gegenseitig bedingen. Für die hier vorliegenden Daten ist klar, dass die Trennung der Tumorzelllinien komplexere Modelle benötigt. Diese können bei den betrachteten kleinen Probenzahlen aber nicht gebildet werden: die Lernkurve in der oberen Zeile zeigt, dass selbst die hier gewählten stark regularisierten Modelle instabil sind.

Allerdings bleibt ebenfalls festzuhalten, dass die *Messung* der Sensitivität (untere Zeile) einer noch größeren Varianz unterliegt.

## 5.2 Abschätzen der notwendigen Testprobenzahl

Die Validierung soll nachweisen, dass das chemometrische Modell den gestellten Ansprüchen genügt. Das heißt auch, dass grobe Vorstellungen über die erwartete und/ oder notwendige Qualität existieren. Um diese Qualität zu erreichen, muss eine ausreichende Anzahl an Trainingsproben vorhanden sein. Um nachzuweisen, dass das Modell tatsächlich die gewünschte Qualität erreicht, müssen ausreichend Proben getestet werden, so

dass das Konfidenzintervall schmal genug wird. Obwohl die letztlich erreichte Intervallbreite natürlich von der beobachteten Sensitivität abhängt, lassen sich für den Idealfall  $\hat{p} = 1$  (oder 0) und für den schlechtestmöglichen Fall  $\hat{p} = 0,5$  Grenzen ableiten. Insbesondere die Grenzen für den Idealfall sind hier insofern bedeutsam, als sie das absolute Minimum an notwendigen Proben angeben. Soll das Konfidenzintervall zum Beispiel nicht breiter als 0,1 sein bzw. das untere Ende des Konfidenzintervalls mindestens 90 % Sensitivität erreichen, so sind selbst im Idealfall (keine falsch-negative Vorhersage) 26 statistisch unabhängige Testproben erforderlich. Werden „nur“ 95 % beobachtete Sensitivität erwartet, so sind bereits 116 Testproben erforderlich. Bei der *hold-out*-Validierung müssen zusätzlich zu den 116 Testpatienten ausreichend Trainingsproben bereitgestellt werden, was bei einer Resampling-Validierung vermieden wird. Dieser Überlegung liegt aber die Annahme der Resamplingvalidierung zu Grunde, dass alle Surrogatmodelle hinreichend ähnlich sind. Nur dann dürfen die Surrogatmodelle wie ein Modell behandelt und die Ergebnisse *gepoolt* werden. Sind die Modelle instabil, so entspricht das einer weiteren Varianzquelle. Das aus der Anzahl der unterschiedlichen und unabhängigen Testproben mit Hilfe der Binomialverteilung berechnete Konfidenzintervall ist dann zu schmal.

Für die Zusammenschau der benötigten Trainings- und Testdaten ist wichtig, dass bereits die im Idealfall erforderlichen 26 Testpatienten der *größten* Trainingsprobenzahl entspricht, die im vorigen Abschnitt diskutiert wurde. Damit ist der Nachweis der erreichten Modellqualität in den in der Biospektroskopie typischen Situationen mit sehr geringen Anzahlen an Patienten oder Zellkulturbatches zur Zeit der limitierende Faktor: während die Modellbildung durch Verwendung restriktiverer Modelle und von Modell-Ensembles in gewissen Grenzen an Situationen mit viel weniger Proben als Variate angepasst werden kann, besteht bei den notwendigen Testprobenzahlen keine vergleichbare Möglichkeit.

### 5.2.1 Modellvergleiche und Optimierung, Festlegen von Hyperparametern

In Kapitel 4.8.5 wurde dargelegt, dass die gegenwärtig üblichen Optimierungsverfahren zum Festlegen von Hyperparametern um Größenordnungen mehr Patienten benötigen, als zur Zeit für biospektroskopische Studien üblicherweise zur Verfügung stehen. Der Grund für diese extrem großen notwendigen Patientenzahlen ist letztlich die hohe Varianzunsicherheit bei der Messung von Anteilen.

Für den Vergleich zweier gegebener Modelle ließe sich der Testprobenbedarf also reduzieren, wenn Kenngrößen mit geringerer Varianzunsicherheit verwendet werden. Kapitel 8.3 zeigt, dass dies zumindest mit einigen der weichen Kenngrößen in vielen praktisch wichtigen Situationen erreicht werden kann.

Die meist verwendeten Such- und Optimierungsalgorithmen machen implizit zwei Annahmen, die bei der datengesteuerten Optimierung von Hyperparametern oft verletzt sind:

- Das verwendete Zielfunktional muss garantieren, dass am Maximum auch tatsächlich optimale Modelle vorliegen,
- und es muss im Suchraum hinreichend glatt sein (stetig, oft wird auch stetig differenzierbar angenommen).

Für die Optimierung von Klassifikationsmodellen werden die sogenannten *strictly pro-*

*per scoring rules* (engl., sinngemäß: streng richtige Bewertungsregeln), zum Beispiel *Brier's score* [247], als Zielfunktional benötigt. *Strictly proper scoring rules* vergleichen die vorhergesagte Klassenzugehörigkeitswahrscheinlichkeit mit der tatsächlich vorliegenden Klasse. Sie erreichen genau dann ihr (eindeutiges) Maximum, wenn die vorhergesagte Klassenzugehörigkeitswahrscheinlichkeit mit der Häufigkeit des Auftretens der einzelnen Klassen übereinstimmt [248]. Harte Vorhersagen entsprechen dabei 0 bzw. 100 % vorhergesagter Klassenzugehörigkeitswahrscheinlichkeit. Wichtig ist, dass die üblicherweise verwendeten harten Kenngrößen wie Sensitivität, Spezifität usw. *nicht* geeignet sind: sie sind keine *strictly proper scoring rules* [248]. Insbesondere zeigen sie stetige Veränderungen des Modells nicht stetig, sondern sprunghaft an (vgl. Beispiel S. 82).

Dass das Zielfunktional im Suchraum hinreichend glatt ist, bedeutet auch, dass die Varianzunsicherheit bei der Bestimmung der Modellqualität sehr viel kleiner sein muss als die Änderung der wahren Modellqualität in Abhängigkeit der Hyperparameter. Das ist bei den vorliegenden Patientenzahlen nicht gegeben.

Die Ergebnisse verschiedener, in Wirklichkeit unterschiedlich leistungsfähiger, Modelle können *zufällig* gut oder sogar perfekt erscheinen. Die Varianzunsicherheit bei der Messung der Modellqualität steigt mit zunehmender Instabilität der Modelle. Damit steigt mit wachsender Modellkomplexität auch das Risiko, dass die Optimierung nur die Varianz beim Messen der Modellqualität „abschöpft“. Dazu kommt, dass eine Optimierung nicht sinnvoll zwischen mehreren Modellen unterscheiden kann, die alle perfekt erscheinen. Die Varianzunsicherheit kann also zu einer Überanpassung des optimierten Modells führen, da auch mit einer Resampling-Validierung alle Modelle letztlich nur gegen dieselben  $n$  Proben getestet werden können. Die optimierten Modelle werden dann im Mittel, also systematisch, zu komplex.

Eine unabhängige Validierung des optimierten Modells, zum Beispiel im Rahmen einer geschachtelten Validierung, liefert ein realistischeres Bild der erreichten Leistungsfähigkeit. Allerdings ist der optimistische Bias, also der Unterschied zwischen der wesentlich besser erscheinenden inneren Messung der Modellqualität und den äußeren Messergebnissen, nur ein Symptom der Überanpassung. Wird ein starker optimistischer Bias zwischen der inneren und der äußeren Validierung beobachtet, ist es also auch fraglich, ob die Optimierung überhaupt in dem Sinne erfolgreich war, dass gute Modelle erzielt werden. Die äußere Messung der erreichten Modellqualität ist aber nicht beeinträchtigt, solange die Testdaten vom betrachteten Modell unabhängig sind.

Als Beispiel seien hier die Ergebnisse der internen *leave-one-out* Messungen des Modelloptimierers und der äußeren  $40 \times 5$ -fachen Kreuzvalidierung für eines der drei Astrozytome  $^{\circ}\text{II}$  aus [CB6] gegenübergestellt (Abb. 5.2). Trainiert wurde mit dem Mittelwertspektrum der Messung, so dass die interne *leave-one-out*-Messung auf der Ebene der FTIR-Images unabhängig ist. Die Testdaten bestehen aus  $32 \times 32$  Pixel FTIR-Images. Ein genetischer Algorithmus optimiert die Auswahl einiger weniger (8) spektraler Regionen, aus denen dann ein LDA-Modell gebildet wird [249]. Dabei handelt es sich um eine sehr aggressive Optimierungsstrategie, die jedes Surrogatmodell der äußeren Kreuzvalidierung optimiert, indem es das (scheinbar) beste aus insgesamt 4760 Modellen herausucht, die aus den entsprechenden Trainingsdaten gebildet werden. Abbildung 5.2a zeigt die inneren *leave-one-image-out*-Vorhersagen für Image 45 der  $40 \times 4$  Surrogatmodelle mit Image 45 im Trainingsdatensatz: 159 von 160 Modellen ( $> 99\%$ ) ordnen das Image richtig der Klasse Astrozytom  $^{\circ}\text{II}$  zu. Demgegenüber sind nur 51 % der Zuordnun-

gen richtig, wenn das Image nicht mit zur Optimierung herangezogen wurde (Abb. 5.2b). Das verdeutlicht eindrucksvoll, dass die Optimierung intern die Modelle praktisch perfekt an den Trainingsdatensatz (über)angepasst hat, obwohl auch intern im Optimierer die jeweilige innere Testprobe nicht zum Training des LDA-Modells verwendet wurde. Auf unbekannte Proben angewendet, die auch nicht zur Auswahl der spektralen Regionen genutzt wurden, bricht die Vorhersagequalität zusammen (vgl. auch [16]).

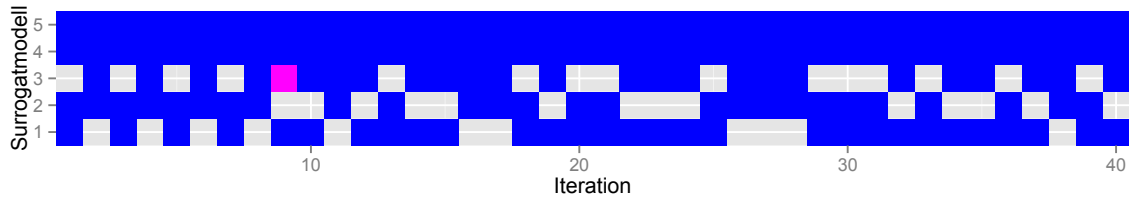
Eine datengesteuerte Optimierung der Hyperparameter sollte diese Varianzunsicherheit also berücksichtigen. Mögliche Lösungswege sind im Ausblick (S. 153) skizziert.

Das Ziel der Optimierung ist es, einen Satz an Hyperparametern zu finden, der zu guten Modellen für das vorliegende Problem führt. Eine Alternative zur angesprochenen datengesteuerten Vorgehensweise besteht darin, die Hyperparameter aufgrund von externen Kenntnissen und Informationen über die Problemstellung festzulegen. In dieser Betrachtungsweise werden dann auch die Wahl des Klassifikationsalgorithmus und die Entscheidung, welche Vorbehandlungsschritte in welcher Reihenfolge durchgeführt werden, zu Hyperparametern der Modellierung.

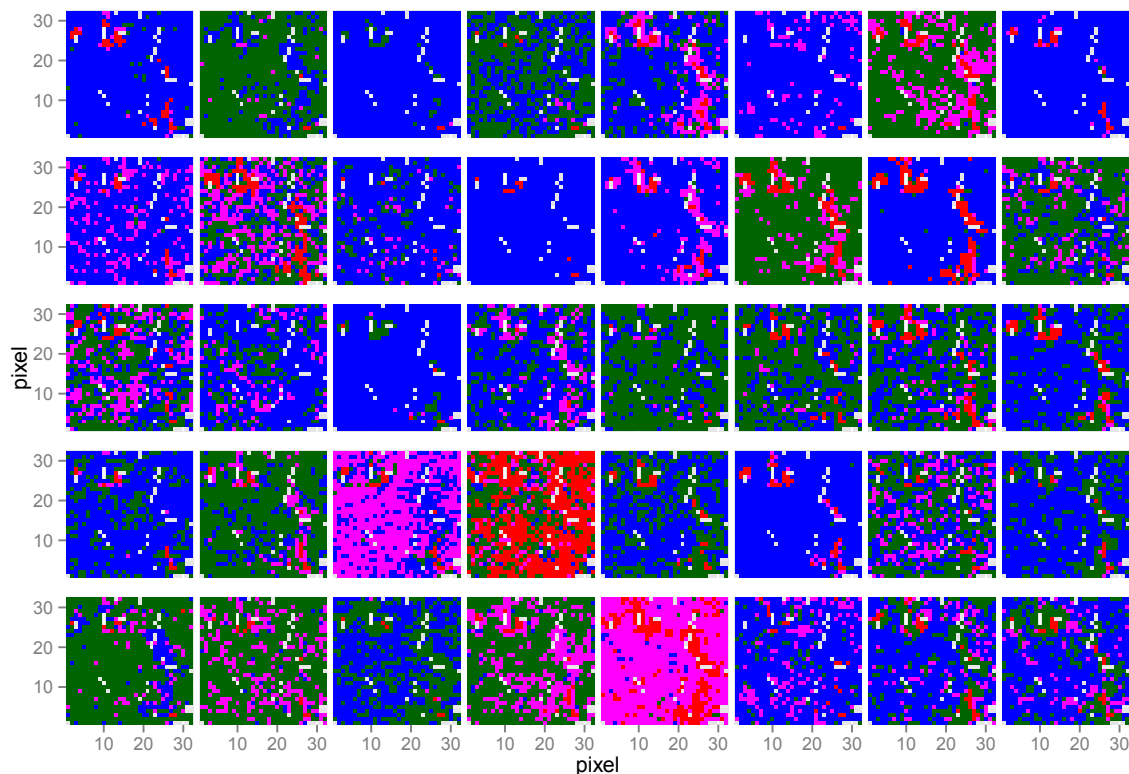
Suchalgorithmen, insbesondere Heuristiken, können immer nur erfolgreich sein, wenn sie an das gegebene Problem angepasst sind. Es lässt sich zeigen, dass es keinen Suchalgorithmus gibt, der allgemein besser als andere Suchalgorithmen ist. Im Englischen heißt eine Sammlung solcher Ergebnisse *no free lunch theorems* [250, 251]. Sie bedeuten, dass der Erfolg auch von Klassifikationsmodellen erheblich davon abhängt, wie gut die Klassifikationsmethode an das Problem und den Datensatz allgemein angepasst sind. Anpassung an den Datensatz meint hier weniger die einzelnen Werte (diese Anpassung nimmt ja das Modell vor), sondern Metainformationen über die Daten wie die Anzahl der Patienten, das Signal-Rausch-Verhältnis oder dass es sich um Schwingungsspektren handelt und welche Modelle oder Regularisierungsstrategien für Schwingungsspektren geeignet sind. Aber auch, ob eine Klassifikation durchgeführt werden soll oder eine quantitative Bestimmung; zum Beispiel kann Intensitätsnormierung für eine Klassifikation helfen, irrelevante Signaländerungen zu reduzieren, während eine Quantifizierung möglicherweise aber genau diese Signaländerung braucht.

Eine Schlussfolgerung aus den *no free lunch*-Theoremen ist also, dass solches externes Wissen für die Modellierung extrem wertvoll ist. Unabhängig vom konkret vorliegenden Datensatz kann so der Suchbereich für die Hyperparameter stark eingegrenzt werden. Damit sinken auch die Anzahl der zu vergleichenden Modelle und das Risiko, dass eine Überanpassung durch Abschöpfen der Varianzunsicherheit eintritt. Wie stark der Suchbereich eingegrenzt werden kann, hängt aber auch stark vom gewählten Algorithmus ab: Haben die Hyperparameter physikalische, chemische oder biologische Bedeutung, so kann das externe Wissen gut und einfach eingebracht werden. Ist die Interpretation der Hyperparameter abstrakt („Modellkomplexität“ statt „Anzahl der unabhängig modellierten Spezies“), kann existierendes Wissen nur schwer umgesetzt werden.

Die in biospektroskopischen Studien typischen Patientenzahlen reichen nicht aus, um auch nur einzelne Modellvergleiche auf einer statistisch soliden Basis durchzuführen. Sofern möglich, sollte also vorerst auf eine datengesteuerte Optimierung verzichtet und die Hyperparameter einschließlich der Auswahl der Klassifikationsalgorithmen, ihrer Hyperparameter und der Datenvorbehandlung aufgrund von spektroskopischem Wissen festgelegt werden (vgl. Kap. 12.1, S. 97 und 12.3, S. 105). Wenn überhaupt, sollte eine datengesteuerte Optimierung nur mit allergrößter Vorsicht durchgeführt werden.



(a) Interne *leave-one-out* Messung des Optimierers. Da insgesamt nur 3 Images von Astrozytomen °II zur Verfügung standen, wurde die Kreuzvalidierung stratifiziert, so dass die ersten drei Surrogatmodelle innerhalb jeder Iteration jeweils ein Astrozytom °II testen und die Surrogatmodelle 4 und 5 keines.



(b) Externe 40× iterierte 5-fache Kreuzvalidierung. Die 40 Modelle erzeugen sehr unterschiedliche Vorhersagen für *dieselben* Daten. Die Vorhersagequalität ist wesentlich schlechter, als der Optimierer intern schätzt. (Nach [CB6])

**Abbildung 5.2** Optimistischer Bias der inneren Messung der Modellqualität, die der Optimierer zur Selektion verwendet. Der genetische Optimierer nutzt *leave-one-out* Messungen der Modellqualität, um das beste Modell zu finden. (a) Im Rahmen der äußeren Validierung der so erhaltenen Modelle wurde Image 45 160 mal als Trainingsprobe verwendet. Die interne Messung der Modellqualität des Optimierers ergab 159 richtige von 160 Vorhersagen (> 99 %). (b) Die äußere Kreuzvalidierung testet dieses Image 40 mal mit Modellen, die unter Ausschluss von Image 45 trainiert und optimiert wurden und zeigt, dass die Ergebnisse der Validierung sehr instabil sind und insgesamt wesentlich schlechter (51 % richtige Vorhersagen) als der Optimierer aufgrund der inneren Messung annimmt.

## 6 Vergleich der Validierungsschemata für biospektroskopische Daten

Im Folgenden wird zunächst das in [CB8] gefundene Verhalten der in Kapitel 4.8.1 eingeführten Validierungsschemata für typische Patientenzahlen in der Biospektroskopie kurz vorgestellt.

Der Resampling-Validierung liegen drei wichtige Annahmen zu Grunde:

1. Die Surrogatmodelle haben dieselbe Qualität wie das große Modell.
2. Die Surrogatmodelle haben untereinander dieselbe Qualität, so dass die Validierungsergebnisse zusammengefasst werden dürfen.
3. Die Proben (genauer: ihre Messungen) sind untereinander äquivalent, Drift tritt nicht auf (vgl. [208]).

Die erste Annahme ist sehr oft verletzt. Wenn die Lernkurve zwischen der Trainingsprobenzahl der Surrogatmodelle und der des großen Modells steigt, sind die Surrogatmodelle im Mittel entsprechend schlechter als das große Modell. Resampling-basierte Validierung hat daher einen (leichten) pessimistischen Bias (vgl. [CB3, CB8]).

Ist zusätzlich die zweite Annahme verletzt, so unterliegen die Validierungsergebnisse einer größeren Varianz als erwartet: zur Varianz aufgrund der beschränkten Zahl an Testproben kommt noch ein Beitrag aus den zufälligen Abweichungen der Surrogatmodelle vom großen Modell (vgl. [CB3, CB6]). Dieser Aspekt wird in Kapitel 7 behandelt.

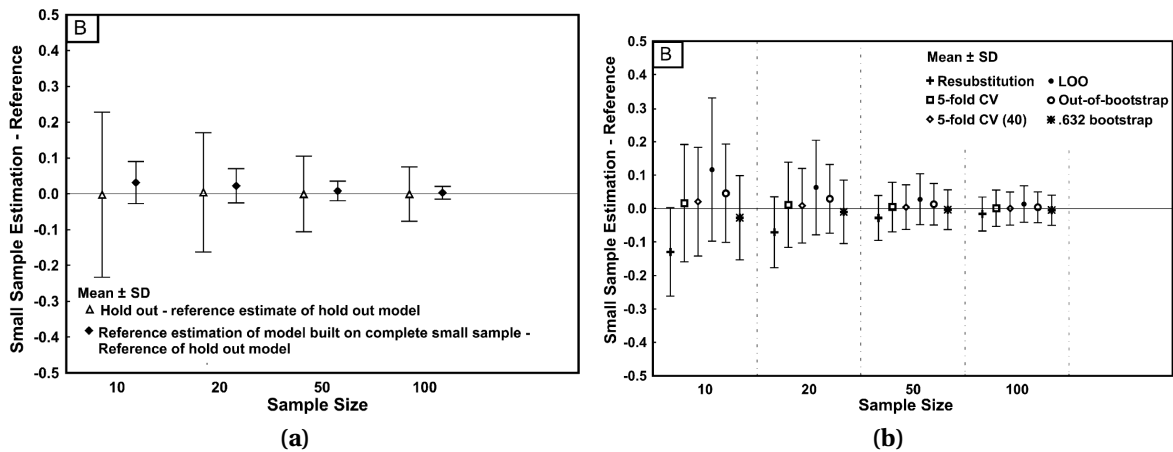
Eine mögliche zeitliche Drift der Messungen wird bei der Resampling-Validierung nicht berücksichtigt. Tritt eine solche Drift auf, so wird die Vorhersagequalität mit der Zeit immer schlechter. Die Resampling-Validierung hat dann einen optimistischen Bias.

Da spektroskopische Datensätze in aller Regel viele Variate haben, neigen die gebildeten Modelle zur Überanpassung und der Resubstitutionsfehler (Abbildung 6.1b Kreuze) ist erwartungsgemäß nahe 0 und damit wesentlich zu optimistisch.

Abbildung 6.1a zeigt, dass die Validierung mit einem eigens reservierten Testdatensatz (*hold-out*, offene Dreiecke) tatsächlich als einziges der untersuchten Validierungsschemata ohne Bias ist. Der Bias wurde mit einem zusätzlichen großen Testdatensatz gemessen, der in der Praxis natürlich nicht zur Verfügung steht. Allerdings unterliegt die *hold-out*-Validierung von allen verglichenen Methoden der größten Varianz. Der Gesamtfehler bei der Validierung mit unabhängigem Testdatensatz ist in den betrachteten Situationen mit sehr kleinen Stichprobenumfängen größer als der Gesamtfehler einer iterierten Kreuzvalidierung oder einer *out-of-bootstrap*-Validierung. Damit aussagekräftige Ergebnisse bei einer *hold-out*-Validierung überhaupt möglich sind, muss also eine hinreichend große Anzahl an Testpatienten eingeplant werden (vgl. [CB3]). Eine Validierung mit einem eigens vorrätig gehaltenen, unabhängigen Testdatensatz sollte daher als eigenes auch zeitlich abgegrenztes Experiment durchgeführt werden, so dass auch eine mögliche Drift gemessen werden kann.

Entsprechend der extrem geringen Stichprobenumfänge wurden sehr restriktive Modelle verwendet. Daher sind die Modelle trotz der geringen Trainingsprobenzahl recht





**Abbildung 6.1** Systematischer und zufälliger Fehler der verschiedenen Validierungsschemata für simulierte Spektrendaten (aus [CB8]). (a) Validierung mit einem eigens reservierten Testdatensatz. (b) Resubstitution und Resampling-Schemata.

stabil (Abbildung 6.1a gefüllte Rauten).

Die Gesamtgenauigkeit von iterierter  $k$ -facher Kreuzvalidierung (Abb. 6.1b offene Rauten bzw. offene Quadrate) und *out-of-bootstrap*-Validierung (Abb. 6.1b offene Kreise) war für spektroskopische Daten vergleichbar, wenn die Kreuzvalidierung mit genügend vielen Iterationen durchgeführt wird. Der Gesamtfehler für die iterierte  $k$ -fache Kreuzvalidierung und die *out-of-bootstrap*-Validierung war praktisch gleich. Die Anzahl der Iterationen wurde so gewählt, dass immer dieselbe Anzahl an Surrogatmodellen berechnet wurde. Das entspricht auch der Beobachtung von Kim [211]. Für den *.632bootstrap* (Abb. 6.1b Sternchen) wurde ein optimistischer Bias bei weiter reduzierter Varianz beobachtet. Die Ursache dafür ist, dass der Resubstitutionsfehler bei der spektroskopischen Klassifikation aufgrund von Überanpassung (engl. *overfitting*) oft bis auf 0 sinkt. Dann entspricht der *.632bootstrap*-Fehler also pauschal 63,2 % des *out-of-bootstrap*-Fehlers und auch die Varianz sinkt entsprechend – allerdings ohne dass der Resubstitutionsfehler brauchbare Information beigetragen hätte. Problematisch ist, dass der *.632bootstrap* Überanpassung für biospektroskopische Daten also nur schlecht anzeigt. Für die Klassifikation spektroskopischer Datensätze mit nur geringen Patientenzahlen gibt es zum *.632+bootstrap* keine detaillierten Untersuchungen und Erfahrungen. Allerdings lagen die Resubstitutionsfehler durchgehend nahe 0 (extrem optimistischer Bias). In dieser Situation gewichtet der *.632+bootstrap* den Resubstitutionsfehler sehr niedrig, so dass der *.632+bootstrap*-Fehler in den normalen *out-of-bootstrap* übergeht. In [CB8] wurde daher auf die Untersuchung des *.632+bootstrap* verzichtet.

## 7 Messung der Modellstabilität mit iterierter Kreuzvalidierung

Wesentliche Ergebnisse dieses Abschnitts wurden in [CB6] veröffentlicht.

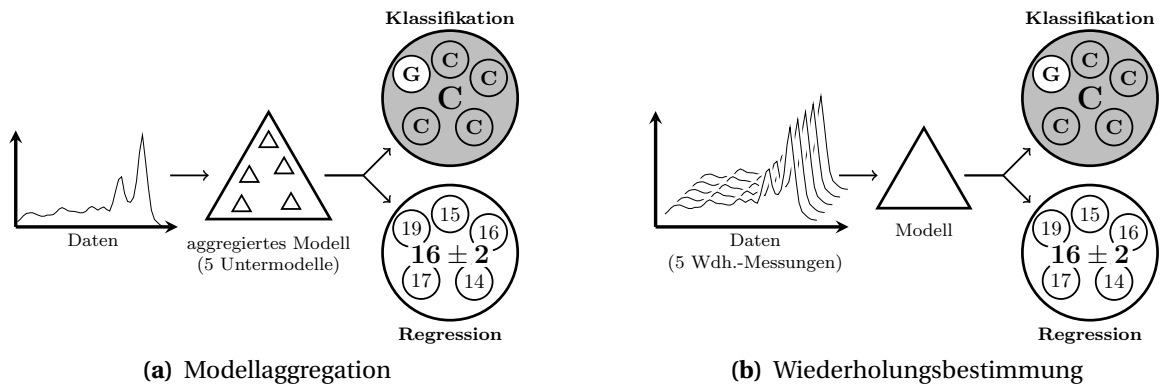
Resampling-Datensätze werden in verschiedenen Situationen unterschiedlich interpretiert: In der Einleitung (Kap. 4.7) wurden Resampling-Stichproben als einfach erhältliche Näherungen für echte neue Datensätze eingeführt. Im Rahmen einer Resampling-Validierung werden die Resampling-Stichproben andererseits als Näherung an den vorliegenden Datensatz betrachtet.

Tatsächlich unterscheiden sich die Resampling-Datensätze natürlich sowohl vom gesamten Datensatz als auch untereinander ein wenig, genauer: durch Auslassen bzw. Austausch einzelner Proben. Daher können Resampling-Datensätze auch als gestörte Versionen des Ausgangsdatensatzes aufgefasst werden. Deshalb kann mit Hilfe der Surrogatmodelle gemessen werden, wie die Modelle auf kleine Störungen in den Trainingsdaten reagieren, also wie *stabil* die Modell gegenüber geringfügig veränderten Trainingsdaten sind.

Die Reaktion des Modells kann anhand der Modellparameter oder der Vorhersagen diskutiert werden. Es kann vorkommen, dass die Vorhersagen stabil sind, obwohl sich viele Modellparameter stark verändern. Das ist zum Beispiel bei korrelierten Eingangsvariaten möglich. Die Korrelation bewirkt, dass verschiedene Kombinationen der Messkanäle dieselben Vorhersagen ergeben. Die Auswirkungen von Änderungen an den Modellparametern sind also möglicherweise schwer zu interpretieren. Bei schwingungsspektroskopischen Datensätzen können Modelle zum Beispiel unterschiedliche Schwingungen derselben funktionellen Gruppe oder Substanzklasse nutzen. Beim Grading von Gliomen anhand von IR-Spektren nutzen z. B. manche Modelle die symmetrische und andere die antisymmetrische Phosphatstretschwingung [CB6]. Die Modelle waren Surrogatmodelle im Rahmen einer iterierten Kreuzvalidierung, unterschieden sich also nur dadurch, dass einzelne Trainingsproben durch andere ausgetauscht wurden.

Die Variabilität der Modellparameter kann auch bei vielen deskriptiven Modellen gemessen werden. Bei prädiktiven Modellen ist es oft wichtiger, weniger die Stabilität der Modellparameter als vielmehr die Stabilität der Vorhersagen zu messen.

Die Stabilität der Vorhersagen lässt sich sehr einfach im Rahmen einer iterierten  $k$ -fachen Kreuzvalidierung mitberechnen. Geeignete Maße sind zum Beispiel die Standardabweichung oder Streubreite der  $i$  Vorhersagen, die über  $i$  Iterationen erhalten werden [CB6]. Die  $i$  Surrogatmodelle, bei denen eine gegebene Probe Testprobe ist, unterscheiden sich nur durch Austausch von maximal  $\frac{n}{k} - 1$  Trainingsproben. Während jeder Iteration wird jede Probe genau einmal getestet. Damit sind die Validierungsergebnisse für die einzelnen Iterationen mit jeweils genau denselben Testproben errechnet, die  $k$  Surrogatmodelle unterscheiden sich jedoch durch Austausch von einigen Trainingsproben. Das bedeutet, dass die Streuung zwischen den Iterationen ausschließlich auf Instabilität der Vorhersagen der Surrogatmodelle zurückzuführen ist: wären die Vorhersagen stabil,



**Abbildung 7.1** (a) Modellaggregation: Mehrere Untermodelle (kleine Dreiecke) verarbeiten die Daten einer Messung (links). Mehrere Aussagen (kleine Kreise) werden aggregiert (großer Kreis). (b) Bei Wiederholungsbestimmungen trifft ein Modell Aussagen für mehrere Messungen. (Nach [CB6])

also immer gleich, so wären auch die Validierungsergebnisse für alle Iterationen gleich. Iterationen bei der  $k$ -fachen Kreuzvalidierung reduzieren also ausschließlich den Anteil der Varianzunsicherheit, der durch die Instabilität der Vorhersagen entsteht, nicht aber die Varianzunsicherheit, die von der insgesamt begrenzten Anzahl an unterschiedlichen Proben (Patienten, Batches) herrührt.

Wird die Stabilität als Streubreite von Kenngrößen über die Iterationen einer Kreuzvalidierung ausgedrückt, so muss bei der Interpretation beachtet werden, dass bereits innerhalb jeder Iteration  $k$  Surrogatmodelle *gepoolt* wurden.

Praktisch bedeutet das, dass bei der Validierung von Klassifikationsmodellen die Stabilität der Vorhersagen gemessen werden kann und sollte. Stellt sich nach einigen Iterationen heraus, dass die Vorhersagen stabil sind, so kann auf weitere Iterationen verzichtet werden. Sind die Vorhersagen instabil, so kann einerseits die Genauigkeit der Validierungsergebnisse durch weitere Iterationen verbessert werden. Andererseits können aus den Surrogatmodellen direkt Ensemble-Modelle gebildet und validiert werden.

## 7.1 Wiederholungsmessungen, Modellaggregation und bolstered error estimation

Bei einer quantitativen Analyse wird die Streubreite der Analysenergebnisse durch Wiederholungsbestimmungen gemessen. Die Unsicherheit wird reduziert, indem der Mittelwert der Wiederholungsbestimmungen verwendet wird (Abb. 7.1b).

Die Modellaggregation geht analog vor, nur wird anstelle der Varianz auf den Messdaten die Varianzunsicherheit des Modells bzw. der Modellvorhersagen reduziert. Abbildung 7.1a zeigt ein aggregiertes oder *Ensemble*-Modell mit seinen Untermodellen. Die aggregierte Vorhersage unterscheidet sich von Wiederholungsbestimmungen dadurch, dass bei der Wiederholungsbestimmung dasselbe chemometrische Modell auf Wiederholungsmessungen angewendet wird. Demgegenüber fasst das aggregierte Modell die Aussagen verschiedener Modelle für dieselben Messwerte zusammen.

Auch bei der Messung der Qualität der Vorhersagen tritt Varianz aufgrund der begrenzten Anzahl an Testproben auf. Die entsprechende Reduktion der Varianz durch Aggrega-

tion der Ergebnisse für mehrere Surrogatmodelle im Rahmen einer Resampling-Validierung (gegenüber einer einfachen *hold-out*- oder *Set*-Validierung) ist so selbstverständlich, dass die Analogie zur Modellaggregation gar nicht auffällt.

Es gibt aber noch eine weitere Methode in der Modellvalidierung, die der Verwendung von Wiederholungsmessungen ähnelt: die sogenannte *bolstered error estimation* (engl., gepolsterte Fehlerschätzung) [252]. Dabei wird jeder Datenpunkt zum Beispiel mit einer multivariaten Normalverteilung „aufgepolstert“. In der Analogie entspricht das Ergebnis Wiederholungsmessungen, die ein wenig um einen Mittelwert streuen. In der Biospektroskopie werden oft viele Spektren von jeder Probe (Patient, Batch) aufgenommen. Die Varianz zwischen den einzelnen Batches und Patienten ist oft größer als die zwischen den einzelnen Spektren eines Batches oder Patienten. Deshalb steht automatisch sozusagen das gemessene Äquivalent zur gepolsterten Fehlermessung zur Verfügung.

Bei der chemometrischen Modellierung und Validierung von biospektroskopischen Daten treten also auf drei Ebenen Varianzunsicherheiten auf, für die jeweils mindestens eine Methode zur Reduktion der Varianzunsicherheit zur Verfügung steht.

## 8 Weiche Kenngrößen für die Qualität von Klassifikationsmodellen

Dieses Kapitel entspricht bis auf die Untersuchung der Varianzeigenschaften der neuen Kenngrößen in Abschnitt 8.3 der Veröffentlichung [CB1]. Ein einheitlicher Rahmen für bekannte Kenngrößen wie Sensitivität, Spezifität und die Vorhersagewerte auch für Daten mit anteiliger Klassenzugehörigkeit wird entwickelt (Kap. 8.2) und ein enger Zusammenhang mit Kenngrößen für die Qualität von Regressionsmodellen wird aufgezeigt (Kap. 8.2.1). Schließlich wird insbesondere die Varianz der neuen Kenngrößen mit der der „klassischen“ Kenngrößen verglichen (Kap. 8.3) und die im Rahmen dieser Arbeit entstandene Implementierung kurz vorgestellt (Kap. 8.4).

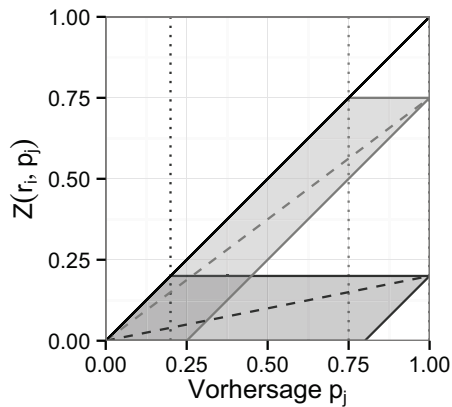
### 8.1 Die Zuordnungsmatrix bei weicher Klassifikation

**Symmetrie der Operatoren:**  $Z^{\text{schwach}}$  und  $Z^{\text{stark}}$  (Kap. 4.9, Gl. 4.26 und 4.27) sind zueinander punktsymmetrisch um  $(p = \frac{1}{2}; Z = \frac{1}{2}r)$ , da  $\max(r_i + p_j - 1, 0) = r_j - \min(r_i, 1 - p_j)$  (Abb. 8.1). Die Diagonale der schwachen Zuordnungsmatrix  $Z^{\text{schwach}}$  gibt die bestmögliche Modellqualität an, die mit den beobachteten Vorhersagen und der Referenz vereinbar ist. Entsprechend geben die Außerdiagonalelemente die schlechtestmögliche Vorhersagequalität für die entsprechende Misklassifikation  $R \mapsto P$ .  $Z^{\text{stark}}$  verhält sich gegensätzlich.

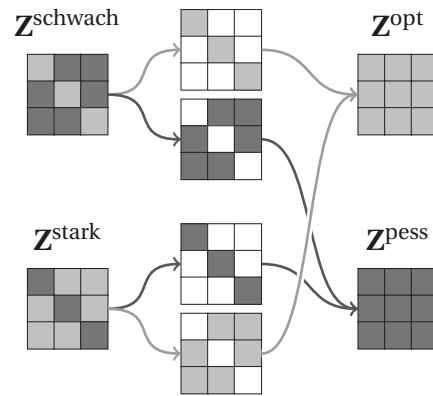
**Bias:** Die Eigenschaften von  $Z^{\text{schwach}}$  und  $Z^{\text{stark}}$  wurden bislang nicht im Hinblick auf den resultierenden Bias für die Kenngrößen interpretiert. Wird ausschließlich die Diagonale von  $Z^{\text{schwach}}$  genutzt, so sind die Kenngrößen systematisch zu optimistisch, da nur der bestmögliche Fall zur Berechnung der Kenngrößen herangezogen wird. Obwohl der komplementäre pessimistische Operator  $Z^{\text{stark}}$  in [193] erwähnt wird, wird auch dort nur die optimistische Diagonale von  $Z^{\text{schwach}}$  mit dem optimistischen Bias verwendet. Ich habe keine Publikation gefunden, die zum Messen der Modellqualität die optimistischen Ergebnisse der Diagonalen von  $Z^{\text{schwach}}$  mit den entsprechenden Werten aus  $Z^{\text{stark}}$  komplementiert.

Ein Qualitätsmaß mit einem konstruktionsbedingt optimistischen Bias überschätzt systematisch die Modellqualität. Die in dieser Arbeit entwickelten Klassifikationsmodelle zielen auf eine Diagnostik, die letztlich über das Entfernen oder Verbleiben von Gewebe im Gehirn entscheidet. Für solche sensiblen Anwendungen ist ein systematisch zu optimistisches Qualitätsmaß ungeeignet. Aber auch für viele andere chemometrische Anwendungen, nicht nur im Bereich der medizinischen Diagnostik, sind systematisch zu optimistische Qualitätsmaße inakzeptabel.

Zunächst könnten  $Z^{\text{schwach}}$  und  $Z^{\text{stark}}$  entsprechend Abbildung 8.2 rekombiniert werden, so dass Zuordnungsmatrizen mit einheitlicher Interpretation als optimistisch und



**Abbildung 8.1** Verhalten der drei Operatoren zur Berechnung der Zuordnungsmatrix  $Z(r_i, p_j)$  für Referenzzugehörigkeiten (Punkte) von  $r_i = 1$  (schwarz), 0,75 (hellgrau), und 0,2 (dunkelgrau):  $Z^{\text{schwach}}$  (obere Grenze der Parallelogramme, durchgezogen),  $Z^{\text{stark}}$  (untere Grenze der Parallelogramme, gepunktet) und  $Z^{\text{prod}}$  (gestrichelt). Bei einer harten Referenz  $r_i = 1$  sind alle drei Operatoren der vorhergesagten Zugehörigkeit  $p_j$  gleich. (nach [CB1])



**Abbildung 8.2** Rekombination von  $Z^{\text{schwach}}$  und  $Z^{\text{stark}}$  als  $Z^{\text{opt}}$  und  $Z^{\text{pess}}$ . Die Diagonale von  $Z^{\text{schwach}}$  und die Außerdiagonalelemente von  $Z^{\text{stark}}$  messen die bestmögliche Modellqualität (hellgrau), während die Diagonale von  $Z^{\text{stark}}$  und die Außerdiagonalelemente von  $Z^{\text{schwach}}$  die schlechtestmögliche Modellqualität angeben. Die Zuordnungsmatrizen  $Z^{\text{opt}}$  und  $Z^{\text{pess}}$  geben also konsistent die best- und schlechtestmögliche Modellqualität an, die im Einklang mit der Vorhersage und der Referenz steht. (nach [CB1])

pessimistisch entstehen. Diese beiden Matrizen spannen direkt den Bereich an möglichen Modellqualitäten auf, die im Einklang mit den beobachteten Testresultaten stehen. Aufgrund der Uneindeutigkeit der Referenz kann die tatsächliche Modellqualität nicht genauer eingegrenzt werden.

$Z^{\text{prod}}$  lässt sich wie in der Einleitung dargelegt als Erwartungswert interpretieren, insofern liegt kein Bias vor.

Die Zeilen- und Spaltensummen der Produkt-basierten Zuordnungsmatrix  $Z^{\text{prod}}$  verhalten sich wie die Zeilen und Spaltensummen der harten Zuordnungsmatrix. Das gilt sowohl für geschlossene als auch für offene Klassifikationssysteme. Die Zeilensummen sind  $\sum_n (r \cdot \sum p)$ , die Spaltensummen  $\sum_n (p \cdot \sum r)$ . Die Gesamtsumme über alle Elemente ist  $\sum_n \sum_n p \cdot \sum r$ . Für geschlossene Klassifikationssysteme sind Zeilen- und Spaltensummen und die Gesamtsumme gleich der Gesamtzahl an Proben. Wie bei den anderen weichen Zuordnungsmatrizen außer  $Z^{\text{opt}}$  und  $Z^{\text{schwach}}$  ist  $Z^{\text{prod}}$  allerdings keine Diagonalmatrix, wenn die Vorhersage gleich der Referenz ist. Das kann als Ausdruck der verbleibenden Unsicherheit über die (nicht aufgelöste) Verteilung der Klassen, also die in Referenz und Vorhersage ausgedrückte Unsicherheit oder Uneindeutigkeit, interpretiert werden.

## 8.2 Berechnung der Kenngrößen bei weicher Klassifikation

Die Kenngrößen für harte Referenz und Vorhersage (Gln. 4.17–4.20) wurden so definiert, dass zu ihrer Berechnung nur einzelne Diagonalelemente der Zuordnungsmatrix bzw. Zuordnungsmatrix mit Dummy-Klasse benötigt werden.  $Z^{\text{schwach}}$ ,  $Z^{\text{stark}}$  und  $Z^{\text{prod}}$  können daher direkt eingesetzt werden, um die best- und schlechtestmögliche Gren-

ze sowie den Erwartungswert für die jeweilige Kenngröße zu berechnen. Das vermeidet alle Probleme damit, dass Zeilen- oder Spaltensummen nicht bei allen diesen Zuordnungsmatrizen mit den Referenz- oder vorhergesagten Klassenzugehörigkeitsvektoren übereinstimmen. Insbesondere können Spezifität und negativer Vorhersagewert so nicht über 100 % steigen. Außerdem gelten diese Berechnungen sowohl für offene als auch für geschlossene Klassifikationssysteme.

Best- und schlechtestmöglicher Fall beziehen sich hier ausschließlich auf die Unsicherheit, die aus der Uneindeutigkeit der Referenz erwächst. Die Kenngrößen unterliegen also zusätzlich der Unsicherheit durch das gewählte Validierungsschema (systematische Fehler, instabile Surrogatmodelle) und die begrenzte Anzahl an Testproben.

Abbildung 8.3 zeigt die Kenngrößen für die drei UND-Operatoren. Jedes Paar aus  $Z^{\text{schwach}}$  (optimistisch) und  $Z^{\text{stark}}$  (pessimistisch) bringt für ein Viertel des Definitionsbereichs keine Information: wenn Referenz und Vorhersage zu uneindeutig sind, sind sie mit allen möglichen Werten der Kenngröße vereinbar (Intervallbreite in Abb. 8.3).

Für das Produkt-UND vereinfachen sich die Ausdrücke für die Kenngrößen zu gewichteten Mittelwerten der Vorhersage oder der Referenz:

$$\text{Sens}^{\text{prod}}(r, p) = \frac{1}{\sum_n r} p \quad (8.1)$$

$$\text{Spez}^{\text{prod}}(r, p) = \frac{1}{\sum_n r} (1 - p) \quad (8.2)$$

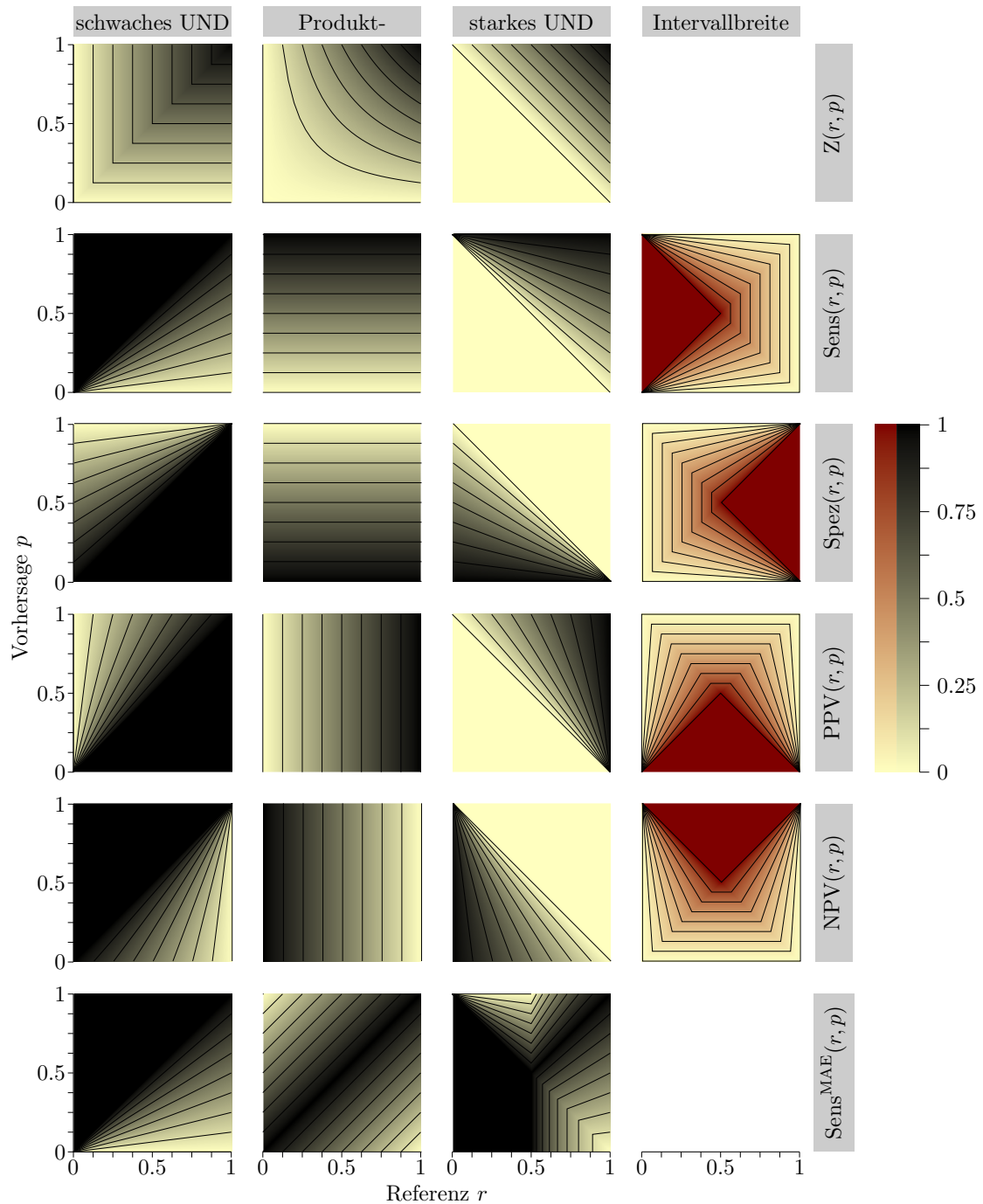
$$\text{PPV}^{\text{prod}}(r, p) = \frac{1}{\sum_n p} r \quad (8.3)$$

$$\text{NPV}^{\text{prod}}(r, p) = \frac{1}{\sum_n p} 1 - r \quad (8.4)$$

### 8.2.1 Abweichung von der idealen Zuordnungsmatrix

Die Möglichkeit, Fehler der Vorhersage analog zum Vorgehen für Regressionsmodelle zu behandeln, wurde in der Einleitung bereits für den Produkt-UND -Operator eingeführt (Kap. 4.9). Mit  $Z^{\text{schwach}}$  werden alle vier Kenngrößen 1, wenn die Vorhersage gleich der Referenz ist. Das heißt, dass  $Z^{\text{schwach}}$  immer schon Abweichungen von der Referenz anzeigt. Allerdings nur zu niedrige Vorhersagen, zu hohe Vorhersagen werden nicht angezeigt. Die Differenz  $Z^{\text{stark}}(r, p) - Z^{\text{stark}}(r, r)$  erzeugt ein komplizierteres Muster (Abb. 8.3, unterste Zeile). Demgegenüber vereinfachen sich die auf  $Z^{\text{prod}}$  basierenden Kenngrößen zu gewichteten Analoga der bekannten Regressionskenngrößen:  $\Delta$  (Gl. 4.29) misst einen Fehler. Die Kenngrößen Gleichung 8.1 – 8.4 beschreiben aber eine Übereinstimmung mit

## 8 Weiche Kenngrößen für die Qualität von Klassifikationsmodellen



**Abbildung 8.3** Die Werte der weichen Kenngrößen (2. bis 5. Zeile) für die drei UND-Operatoren nach Gl. (4.26) – (4.28) (Spalten) für eine Probe (Spektrum) als Funktion von Referenz- und vorhergesagter Zugehörigkeit. In der 1. Zeile sind die Werte der weichen UND-Operatoren (also der Zuordnungsmatrix-Funktion  $Z$ ) aufgetragen, in der 2. bis 5. Zeile die Kenngrößen entsprechend Gleichungen 8.12 – 8.15. Die Symmetrie (vgl. Abb. 8.4) wird in dieser Darstellung besonders deutlich. Die 4. Spalte (rot) gibt die Breite des Intervalls zwischen best- (schwaches UND) und schlechtestmöglichem Fall (starkes UND) wieder. Im Dreieck zwischen der Seite, auf der die harte Kenngröße nicht definiert ist und dem Zentrum der Eingabewerte ( $r = 0,5; p = 0,5$ ) deckt das Intervall den gesamten Wertebereich von 0 bis 1 ab. Dort kann keine Aussage über das Modell gemacht werden.  $Z^{\text{prod}}$  ähnelt für kleine Werte  $Z^{\text{schwach}}$  und für hohe Werte  $Z^{\text{stark}}$ . Die letzte Zeile zeigt die MAE-basierte Sensitivität nach Gl. (8.7).



der Referenz. Das kann mit Hilfe von  $1 - |\Delta|$  ausgedrückt werden:

$$\text{Sens}^{\text{MAE}}(r, p) = 1 - \frac{\sum_n |\Delta^{\text{prod}}(r, p)|}{\sum_n r} \quad (8.5)$$

$$= 1 - \frac{\sum_n |Z^{\text{prod}}(r, p) - Z^{\text{prod}}(r, r)|}{\sum_n r} \quad (8.6)$$

$$= 1 - \sum_n \frac{r}{\sum_n r} |p - r| \quad (8.7)$$

$\text{Sens}^{\text{MAE}}$  ist also der mittlere absolute Fehler (engl. *mean absolute error, MAE*), gewichtet mit den Klassenzugehörigkeiten der Referenz (vgl. Gl. 4.17).

Die Spezifität beschreibt den Rest der Residuen, der den Proben zugeordnet wird, die nicht zur jeweiligen Klasse gehören:

$$\text{Spez}^{\text{MAE}}(r, p) = 1 - \sum_n \frac{1 - r}{\sum_n 1 - r} |p - r| \quad (8.8)$$

Die Vorhersagewerte verfolgen den inversen Gedanken und verteilen die Residuen nach den *vorhergesagten* Klassenzugehörigkeiten:

$$\text{PPV}^{\text{MAE}}(r, p) = 1 - \sum_n \frac{p}{\sum_n p} |p - r| \quad (8.9)$$

$$\text{NPV}^{\text{MAE}}(r, p) = 1 - \sum_n \frac{1 - p}{\sum_n 1 - p} |p - r| \quad (8.10)$$

### 8.2.2 Absoluter und quadrierter Fehler

Anstelle der gewichteten MAE können auch die quadrierten Fehler genutzt werden, zum Beispiel der wRMSE, die gewichtete Wurzel aus dem mittleren quadrierten Fehler (engl. *weighted root mean squared error*)

$$\text{Sens}^{\text{RMSE}}(r, p) = 1 - \sqrt{\frac{\sum_n r}{\sum_n r} (p - r)^2} \quad (8.11)$$

Der MAE kommt der üblichen Zählung der Fehler von Klassifikationsmodellen näher, der RMSE ist für Regressionsmodelle gebräuchlicher. Der mittlere quadrierte Fehler MSE für weiche Vorhersagen und harte Referenz ist auch unter dem Namen *Brier's score* bekannt [247]. Im Gegensatz zu den harten aus der Zuordnungsmatrix abgeleiteten Kenngrößen ist *Brier's score* eine *strictly proper scoring rule* (vgl. Kap. 5.2.1) [248]. MSE-basierte Kenngrößen sind deshalb als Zielfunktional für die Modelloptimierung geeignet, zumal sie oft auch einer geringeren Varianz unterliegen (Kap. 8.3).

MAE und RMSE hängen eng miteinander zusammen. Allgemein gilt für  $n$  Vorhersagen  $\text{MAE} \leq \text{RMSE} \leq \sqrt{n} \text{MAE}$ . Bei der hier genutzten Formulierung der Klassenzugehörigkeit

mit  $p$  und  $r$  jeweils  $\in [0,1]$  kann eine einzelne Vorhersage nie um mehr als 1 von der Referenz abweichen. Damit wird  $0 \leq \text{MAE} \leq 1$  und  $\text{MAE} \leq \text{RMSE} \leq \sqrt{\text{MAE}}$ . Diese Obergrenzen werden auch nur für harte Referenzdaten erreicht, bei weicher Referenz liegen sie entsprechend niedriger.

$\Delta^{\text{prod}}$  und die davon abgeleiteten Kenngrößen zählen sowohl zu niedrige als auch zu hohe Vorhersagen. Das erzeugt das beabsichtigte Verhalten für die Kenngrößen, die sich auf einzelne Klassen beziehen. Kenngrößen wie die Trefferquote, die mehr als eine Klasse zusammenfassen, zählen Fehlzuordnungen damit aber gegebenenfalls mehrfach. Bei geschlossenen Klassifikationssystemen werden für jede unterschätzte Klassenzugehörigkeit andere Klassenzugehörigkeiten im gleichen Maß überschätzt. Daher gilt für klassenübergreifende Kenngrößen  $\text{MAE} \leq 2$  und  $\text{RMSE} \leq \sqrt{2}$ . Hier ist es gegebenenfalls sinnvoll, die Kenngrößen so zu normieren, dass jeder Fehler nur einmal gezählt wird.

Bei der Einklassen-Klassifikation sind die einzelnen Fehlzuordnungen unabhängig. Im Extremfall könnte ein Spektrum allen Klassen zugeordnet werden, obwohl es keiner einzigen wirklich angehört. Für die Obergrenzen gilt  $\text{MAE} \leq n_g$  und  $\text{RMSE} \leq \sqrt{n_g}$ .

### 8.3 Vergleich der weichen und harten Kenngrößen

Mit diesen Definitionen können Sensitivität und Spezifität auch bei der Validierung von weichen Klassifikationsmodellen angegeben werden. Auch weiche Sensitivitäten und Spezifitäten können im Spezifitäts-Sensitivitäts-Diagramm (Kap. 4.8.5) eingetragen werden. Allerdings ist ein ganzes Modell durch eine *einzig*e Sensitivität und Spezifität gekennzeichnet, die an die Stelle der Kennlinie tritt.

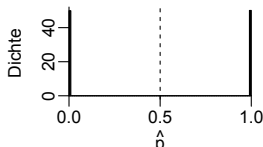
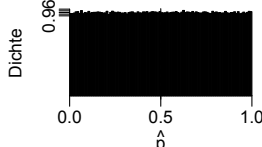
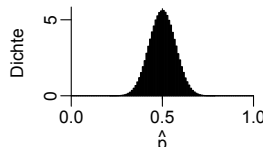
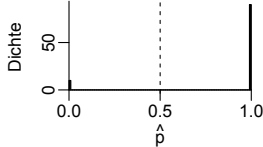
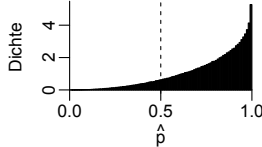
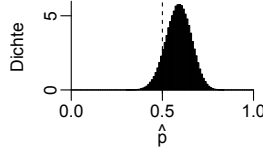
Die weichen Kenngrößen reagieren wesentlich empfindlicher und kontinuierlicher (sie sind stetig differenzierbar) auf Abweichungen zwischen Vorhersage und Referenz als die harten Kenngrößen. Bei den harten Kenngrößen ist es egal, um wieviel der Grenzwert überschritten wird. Die hier vorgestellten weichen Kenngrößen messen hingegen den *Abstand* der Vorhersage zur Referenz.

Ein Beispiel verdeutlicht den Unterschied. Ein Zwei-Klassen-Modell sage die Klassenzugehörigkeitswahrscheinlichkeit vorher. Die Vorhersagen sind also Werte zwischen 0 und 1. Das betrachtete Modell liefere nun für Klasse „A“ immer die Vorhersage „0,6 A, 0,4 B“ und für alle Proben aus Klasse „B“ „0,4 A, 0,6 B“. Wird der Grenzwert zur Berechnung der harten Klassenzugehörigkeit auf 0,5 gesetzt, so erhält man ein ideales Modell mit 100 % Sensitivität und 100 % Spezifität. Die weichen Kenngrößen sind deutlich niedriger: die Sensitivitäten und Spezifitäten erreichen für beide Klassen nur 0,6. Das drückt aus, dass es bessere Vorhersagen geben könnte. Ideal würde jede Probe mit 100 %iger Wahrscheinlichkeit der richtigen Klasse zugeordnet. Andersherum kann diese niedrigere weiche Sensitivität als Hinweis aufgefasst werden, dass das Risiko einer Fehlklassifikation bei neuen Proben relativ groß sein kann, schließlich sind die vorhergesagten Klassenzugehörigkeitswahrscheinlichkeiten recht nahe an 0,5.

Die harten Sensitivitäts-Spezifitäts-Kurven entstehen durch systematisches Variieren des Grenzwerts (Kap. 4.8.5). Dabei wird nur aufgezeigt, welche Sensitivitäts-Spezifitäts-Wertepaare realisiert werden können. Welchen Wert der zur Klassifikation verwendete Score annimmt, spielt dabei zunächst keine Rolle, obwohl diese Information zusätzlich eingezeichnet werden kann. Die hier verwendete Klassenzugehörigkeiten haben je-

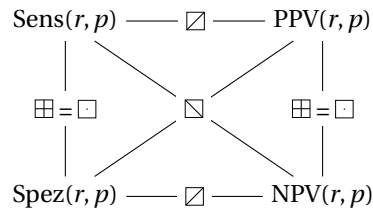
### 8.3 Vergleich der weichen und harten Kenngrößen

**Tabelle 8.1** Varianz der weichen Kenngrößen (Sensitivität), berechnet über  $10^4$  Ziehungen von jeweils 100 Proben, mit der den Histogrammen entsprechenden Verteilung. Modelle a–c raten die Klassen (klassische Sensitivität von 50%), Modelle d–f haben klassische Sensitivitäten von 90%. Dabei sind die Vorhersagen von Modellen a und d hart, also immer entweder 0 oder 1, die anderen vier Modelle erzeugen weiche Vorhersagen. Die Vorhersagen der Modelle c und f nutzen nicht den gesamten Wertebereich der Klassenzugehörigkeiten aus. Ihre Vorhersagen sind um mittelgroße Werte konzentriert.

Operator	Harte Modelle		Weiche Modelle				
	(a)	(b)	(c)	(d)	(e)	(f)	
							
	$\bar{x} \pm s(x)$	$\frac{s^2(x)}{s^2(x^{\text{hart}})}$	$\bar{x} \pm s(x)$	$\frac{s^2(x)}{s^2(x^{\text{hart}})}$	$\bar{x} \pm s(x)$	$\frac{s^2(x)}{s^2(x^{\text{hart}})}$	
Sens <sup>(hart)</sup>	0,50 ± 0,050		0,50 ± 0,050		0,50 ± 0,050		
Sens <sup>MAE</sup>	0,50 ± 0,050	1,00	0,50 ± 0,029	0,33	0,50 ± 0,007	0,02	
Sens <sup>MSE</sup>	0,50 ± 0,050	1,00	0,67 ± 0,030	0,36	0,75 ± 0,007	0,02	
Sens <sup>RMSE</sup>	0,29 ± 0,036	0,51	0,42 ± 0,026	0,27	0,50 ± 0,007	0,02	
	$\bar{x} \pm s(x)$	$\frac{s^2(x)}{s^2(x^{\text{hart}})}$	$\bar{x} \pm s(x)$	$\frac{s^2(x)}{s^2(x^{\text{hart}})}$	$\bar{x} \pm s(x)$	$\frac{s^2(x)}{s^2(x^{\text{hart}})}$	
Sens <sup>(hart)</sup>	0,90 ± 0,030		0,90 ± 0,030		0,90 ± 0,030		
Sens <sup>MAE</sup>	0,90 ± 0,030	1,00	0,78 ± 0,019	0,39	0,59 ± 0,007	0,05	
Sens <sup>MSE</sup>	0,90 ± 0,030	1,00	0,92 ± 0,013	0,18	0,83 ± 0,006	0,04	
Sens <sup>RMSE</sup>	0,69 ± 0,048	2,62	0,71 ± 0,022	0,52	0,58 ± 0,007	0,05	

doch absolute Bedeutung: errechnet die LDA eine *a-posteriori*-Wahrscheinlichkeit von 0,6, so bedeutet das: das Modell stuft die Wahrscheinlichkeit, dass eine Probe mit den gegebenen Messdaten zur betrachteten Klasse gehört, mit 60 % ein. Das heißt aber, dass für Scores, die *a-posteriori*-Wahrscheinlichkeit darstellen, der Grenzwert der Sensitivität entsprechen sollte. Das wird von den weichen Kenngrößen berücksichtigt. Mit anderen Worten, schlecht *kalibrierte* Modelle, deren vorhergesagte *a-posteriori*-Wahrscheinlichkeiten also systematisch von den in der Validierung beobachteten Sensitivitäten abweichen, werden durch die weichen Kenngrößen ebenso „bestraft“, wie Modelle, bei denen Proben zufällig falsch zugeordnet werden.

Dichotomisierung oder Härten einer weichen Vorhersage beeinflusst die Varianz, der die Kenngrößen unterliegen. Der genaue Einfluss hängt von der konkreten Situation ab. Tabelle 8.1 zeigt einige Beispielsituationen für die Sensitivität. Für den Grenzfall, dass das Modell hart zwischen den Klassen entscheidet (a und d), sind Sens<sup>wMAE</sup> und Sens<sup>wMSE</sup> jeweils genau gleich der klassischen, harten Sensitivität. Auch ihre Varianz ist gleich. Die wRMSE-basierte Sensitivität fällt niedriger aus und zeigt eine andere Varianz, bei



**Abbildung 8.4** Symmetriebeziehungen zwischen den Kenngrößen.  $\oplus = \ominus$ :  $(r; p) \mapsto (1 - r; 1 - p)$  horizontal und vertikal spiegeln oder Punktspiegelung,  $\boxplus$ : Spiegelung an der 1. Winkelhalbierenden  $(r; p) \mapsto (p; r)$ , und  $\boxminus$ : Spiegelung an der 2. Winkelhalbierenden  $(r; p) \mapsto (1 - p; 1 - r)$ . Die Symmetrieelemente beziehen sich auf das Zentrum des Wertebereichs (0.5; 0.5) und die Symbole nutzen kartesische Koordinaten. (nach [CB1])

schlechten Modellen (a) eine niedrigere, bei besseren Modellen ist sie größer (d).

Modelle b, c, e und f erreichen „klassisch“ dieselbe Sensitivität wie Modelle a bzw. d, allerdings sind die Vorhersagen bei Modell b zwischen Klassenzugehörigkeit 0 und 1 gleichverteilt, also sehr weich. Die Vorhersagen von Modell e umfassen ebenfalls den gesamten Bereich an möglichen Klassenzugehörigkeiten zwischen 0 und 1. Die Modelle c und f produzieren hingegen Vorhersagen, die auf einen relativ schmalen Wertebereich beschränkt sind. Für alle Modelle mit weichen Vorhersagen ist die Varianz aller drei weichen Kennwerte wesentlich geringer als die der dichotomisierten klassischen Sensitivität. An Modell f wird deutlich, dass die weichen Kenngrößen gegenüber der klassischen Sensitivität stärker berücksichtigen, dass die Klassenzugehörigkeitswahrscheinlichkeiten durchweg zu gering (wenn auch überwiegend auf der richtigen Seite des Grenzwerts) vorhergesagt werden. Auffällig ist auch, dass der Dichotomisierungsschritt die Varianz um einen Faktor 20 erhöht.

Die wMAE- und wMSE-basierten Kennwerte entsprechen also für harte Vorhersagen genau den klassischen Kenngrößen. Sind die Vorhersagen hingegen weicher, so erlauben die weichen Kennwerte eine differenziertere Beurteilung der Modellqualität und unterliegen außerdem einer wesentlich geringeren Varianz.

## 8.4 Implementierung

Die entwickelten Maße für die Qualität von Klassifikationsmodellen wurden in R [253] als Paket *softclassval* implementiert und unter der GNU Public License (<http://www.gnu.org/licenses/gpl.html>) veröffentlicht. Die Homepage des Projekts ist <http://softclassval.r-project.org>. Stabile Versionen sind auch über das Comprehensive R Archive Network (CRAN, <http://www.cran.r-project.org/web/packages/softclassval/index.html>) erhältlich.

Aufgrund der Symmetriebeziehungen zwischen den vier Kenngrößen (Abb. 8.3 und 8.4) genügt eine Funktion, um alle vier Kenngrößen auszudrücken. Diese und die Klassenzugehörigkeiten beziehen sich auf jede einzelne Klasse. Wird vereinfachend der Klas-

senindex weggelassen, so sind die Funktionen zu den Kenngrößen:

$$\text{Sens}(r, p) = \frac{\sum_n Z(r, p)}{\sum_n r} \quad (8.12)$$

$$\text{Spez}(r, p) = \text{Sens}(1 - r, 1 - p) \quad (8.13)$$

$$\text{PPV}(r, p) = \text{Sens}(p, r) \quad (8.14)$$

$$\text{NPV}(r, p) = \text{Sens}(1 - p, 1 - r) \quad (8.15)$$

R Pakete werden einer Reihe von formalen Prüfungen unterzogen, die eine gute Qualität des Programmcodes sicherstellen sollen. Beispiele für solche formalen Prüfungen sind, dass alle exportierten Funktionen dokumentiert und alle Beispielrechnungen ausführbar sein müssen. Diese formale Überprüfung findet automatisch jedesmal statt, wenn auf dem Entwicklungsserver `http://r-forge.r-project.org` oder dem CRAN aus dem Quellcode des Pakets die installierbaren Pakete gebaut werden. Ist diese Prüfung nicht erfolgreich, so wird das Paket von diesen Servern nicht zum Installieren bereitgestellt. Außerdem kann die formale Prüfung manuell mit dem Kommando `R CMD check` ausgelöst werden. Die Ergebnisse dieser Prüfungen für *softclassval* sind im Anhang (Kap. C.1.1) abgedruckt.

Für statistische Berechnungen ist es darüberhinaus besonders wichtig, die Korrektheit der Berechnungen zu überprüfen. Im Unterschied zum Beispiel zu Gerätesteuerungen sind hier Fehler wahrscheinlicher, die nicht zu einer offensichtlichen Fehlfunktion wie dem Abbruch des Programms, unerlaubten Wertebereichen oder anderweitig zumindest unplausiblen Ergebnissen führen. Fehler in statistischen Berechnungen führen oft zu schwer erkennbaren fehlerhaften Ergebnissen. *Unit-Tests* erlauben, einzelne Funktionen mit vorgegebenen Eingabewerten auszuführen und die Ergebnisse mit Referenzergebnissen zu vergleichen. *softclassval* nutzt dies, indem Beispielsituationen durchgerechnet und die Ergebnisse mit manuell berechneten Referenzwerten überprüft werden. Insgesamt enthält *softclassval* etwa doppelt so viele Zeilen solcher Unit-Tests als „eigentlichen“ Programmcode. Unit-Tests werden auch im Rahmen der automatisierten formalen Überprüfung ausgeführt, so dass diese nur dann erfolgreich ist, wenn auch alle Unit-Tests erfolgreich waren. Mit der Funktion `softclassval.unittest()` können die Unit-Tests auch gezielt ausgeführt werden. Die Ergebnisse der Unit-Tests sind ebenfalls im Anhang C.1.2 ersichtlich.



**Teil III**

**Experimente,  
Materialien und Methoden**

## 9 Proben

Die untersuchten Tumorproben stammen aus der Probensammlung des Projekts molekulare Endospektroskopie der Volkswagen-Stiftung. Die Gewebe wurden im Rahmen von Tumorresektionen in der Neurochirurgie des Universitätsklinikums Carl-Gustav Carus der TU Dresden entnommen. Die Proben wurden sofort nach der Entnahme in flüsigem Stickstoff schockgefroren und bei  $-80^{\circ}\text{C}$  aufbewahrt.

Tumorproben werden am Rand des Tumors genommen, da nekrotische Gewebe im Rahmen der normalen histologischen Begutachtung keine Rückschlüsse mehr auf die Ursprungszellen des Tumors erlauben. Auch für die hier diskutierte Fragestellung bezüglich intraoperativer Diagnostik *in vivo* werden Proben des Tumorrandes benötigt, da die Diagnostik bei der Erkennung des Tumorrandes helfen soll. Auch bei Lagerung bei  $-80^{\circ}\text{C}$  muss damit gerechnet werden, dass die Proben altern. Bestimmte immunhistochemische Färbungen und genetische Untersuchungen funktionieren bereits nach 1 bis 2 Jahren nicht mehr zuverlässig [254]. Auf Anraten der Neuropathologin, Fr. PD Dr. Geiger [255], wurden aus der Hirntumorprobensammlung Astrozytome und Glioblastome präpariert, die nicht länger als 2 bzw. 1 Jahr (Glioblastome) gelagert waren.

Die Sammlung beinhaltet Gewebe von neun Kontrollproben von Patienten ohne Hirntumor, die alle präpariert wurden. Drei dieser Proben gehören zu den ältesten Proben der Sammlung überhaupt. Sie waren bei der Messung ca. 7 Jahre alt. Diese Proben werden in Kapitel 14 (S. 114) genauer diskutiert. Die sechs neueren Kontrollproben waren bei der Präparation etwa  $\frac{1}{2}$  bis 1 Jahr gelagert. Insgesamt standen 8 Proben von Lymphomen zur Verfügung, die bis zu 5 Jahre gelagert waren und alle präpariert wurden. Dabei handelt es sich in allen Fällen um diffuse, großzellige Non-Hodgkin-Lymphome der B-Zell-Reihe.

Die Abbildungen 12.3 und 12.4 (S. 104 bzw. 106) geben eine Übersicht über die Lagerdauer, Alters- und Geschlechterverteilung der Patienten für die Datensätze zum Grading der Astrozytome beziehungsweise der Differentialdiagnostik zwischen Astrozytomen und Lymphomen.

### 9.1 Präparation

Von den Proben habe ich Gefrierschnitte von  $7\ \mu\text{m}$  Schnittdicke auf Glas (für die Referenzdiagnose) und von  $14\ \mu\text{m}$  Dicke auf Raman-geeigneten  $\text{CaF}_2$ -Objektträgern<sup>(a)</sup> angefertigt. Die  $\text{CaF}_2$ -Objektträger laden sich leicht elektrostatisch auf, so dass die Gefrierschnitte auf den Objektträger springen und zusammengefaltet und damit unbrauchbar werden. Das lässt sich vermeiden, indem der Objektträger mit deionisiertem Wasser und Zellstoff feucht abgewischt wird. Die  $14\ \mu\text{m}$  Schnitte auf  $\text{CaF}_2$  wurden in den Diplomarbeiten von M. Kammer [135] und R. Marx [114] mit IR- und Raman-Spektroskopie untersucht.

---

<sup>(a)</sup> Crystal GmbH, Berlin, Vacuum-UV-Qualität, optisch geschliffen



Einbettmedien erleichtern das Schneiden von Proben mit dem Mikrotom. Sie stabilisieren die Probe mechanisch. Ein Objekt mit gleichmäßiger Härte und günstiger Form entsteht. Zusätzlich sinkt die Gefahr von Gefrierartefakten. Außerdem dient das Gefriermedium als Kleber, um die Probe auf dem Probenhalter auszurichten und zu befestigen. Das empfohlene Polyethylenglykol (PEG)-Gefriermedium dringt jedoch nach dem Auftauen in die Probe ein Kapitel 13 (S. 112) und trägt zu den Raman-Spektren bei. Daher wurden die Proben zunächst mit einem Tropfen Wasser auf den Probenträger geklebt. Viele der so präparierten Gefrierschnitte zeigten jedoch Gefrierartefakte. Die Gefrierartefakte konnten deutlich reduziert werden, indem die zu schneidende Probe unmittelbar vor dem Schneiden aufgetaut und erneut in flüssigem Stickstoff direkt auf den Probenhalter für das Gefriermikrotom schockgefroren wurde.

Die Gefrierartefakte könnten durch Antauen der Probe durch den Wassertropfen entstanden sein. Aber auch Schnitte von dickeren Proben, bei denen ein Antauen bis an die Oberfläche ausgeschlossen werden kann, wiesen Gefrierartefakte auf. Eine weitere mögliche Ursache für die Gefrierartefakte ist eine Havarie der Gefriertruhe, bei der alle Proben bis auf etwa  $-10^{\circ}\text{C}$  aufgewärmt wurden. Weiterhin wurden von manchen Proben, besonders den alten Kontrollproben, bereits mehrfach Schnitte präpariert. Dabei wurden die Proben immer mindestens auf  $-20^{\circ}\text{C}$  erwärmt und hinterher wieder abgekühlt. Einige Proben wurden bei vorhergehenden Präparationen ganz aufgetaut (vgl. Kap. 15).

Die  $7\mu\text{m}$ -Schnitte für die Referenzdiagnose wurden mit 4 % Formalin in 50 % Ethanol fixiert (Anh. A.2) und mit Methylenblau gefärbt (Anh. A.3). Eine detaillierte Referenzdiagnose aller Schnitte wurde von der Neuropathologin des Universitätsklinikums der TU Dresden gestellt (Kap. 11). Der etwa 3 mm dicke Rest der Probe (*Bulkprobe*) wurde gefroren vom Probenhalter gelöst und bis zur Raman-Messung (Kap. 10) in Aluminium-Folie bei  $-80^{\circ}\text{C}$  transportiert und aufbewahrt.

**Präparation einiger Gefrierschnitte *nach* der Raman-Messung:** Einige der ersten Proben wurden *nach* der Raman-Messung wieder schockgefroren. Davon wurden weitere  $7\mu\text{m}$  -Schnitte für die Referenzdiagnose präpariert und gefärbt. Anhand dieser Proben wurde untersucht, ob und wie Referenzdiagnosen der Messprobe (anstelle von Parallelschnitten) präpariert werden können. Der Schwerpunkt lag dabei weniger auf der Frage, inwieweit das Gewebe des Parallelschnittes repräsentativ für das gemessene Gewebe ist, sondern auf der gegebenenfalls besseren Übertragbarkeit der detaillierten Diagnose. Dazu muss die Präparation nicht nur histologisch auswertbare Schnitte ermöglichen, sondern die Probe darf sich auch nicht verformen. Diese beiden Ziele erwiesen sich als praktisch nicht vereinbar.

Einige Proben wurden mit Schnellgefrierspray<sup>(b)</sup> eingefroren. Schnellgefrierspray wird oft indirekt angewendet. Die Probe wird auf den Probenhalter des Gefriermikrotoms gelegt, der von unten durch die Schnellkühlung gekühlt wird. Ein vorgekühlter Metallblock wird auf die Probe gedrückt, und das Ganze mit Schnellgefrierspray besprüht. Diese Technik verformt die Probe allerdings sehr stark und ist hier daher nicht anwendbar.

Die Proben wurden stattdessen direkt auf dem zur Messung verwendeten  $\text{CaF}_2$  -Fenster schockgefroren. Wurde das Gefrierspray direkt auf die Probe gerichtet, so verformte

<sup>(b)</sup> Solidifix-Kältespray, Propan-Butan-Gemisch, Carl Roth GmbH + Co. KG, Karlsruhe

sie sich stark, weil die aufgetauten Hirngewebe sehr weich sind. Wurde das Gefrierspray aufgetropft, so fro die Probe nicht schnell genug ein und es entstanden Gefrierartefakte. Auch ein Einbetten der Probe in PEG half nicht weiter: die Proben waren ebenfalls wesentlich zu langsam gefroren und verformten sich zudem beim Aufbringen des recht zähflüssigen Einbettmittels.

Schockfrieren der gemessenen Bulkproben in flüssigem Stickstoff war wesentlich erfolgreicher. Die Proben konnten ohne große Verformung erfolgreich eingefroren werden. Der Temperaturschock beansprucht das  $\text{CaF}_2$ -Fenster aber sehr stark. Daher wurde entschieden, dass diese Präparation so oft durchgeführt wird, bis das erste  $\text{CaF}_2$ -Fenster springt. Das war bei der vierten Probe der Fall. Für die Raman-Spektroskopie sind sehr reine  $\text{CaF}_2$ -Objektträger erforderlich, die entsprechend teuer sind. Ein Fenster von 16 mm Durchmesser kostet ca. 50 €. Eine Alternative zum  $\text{CaF}_2$  ist Quarz. Quarz verträgt den Temperaturschock problemlos, erzeugt aber ein höheres Untergrundsignal im Raman-Spektrum. Da die feuchten Gewebeproben nur ein schwaches Raman-Signal liefern, verschlechtert sich das Signal-Rausch-Verhältnis deutlich. Daher waren Quarzfenster für diese Arbeit keine Alternative.

Das zweite Ziel dieser erneuten Referenzdiagnose war die Kontrolle, ob während der mehrstündigen Raman-Messungen in der feuchten Kammer sichtbare Veränderungen an den Proben auftreten. Soweit die Qualität der Schnitte eine Beurteilung erlaubte, fand die Neuropathologin keine Hinweise auf ein Verderben der Probe während der etwa sechsstündigen Messzeit.

## 10 Raman-Messungen

Die Raman-Spektren wurden mit einem  $f/1.8$  Spektrometer der Firma Kaiser<sup>(a)</sup> mit hintergrundbeleuchtetem CCD-Detektor<sup>(b)</sup> mit Flüssigstickstoff-Kühlung und einer faser-optischen Sonde<sup>(c)</sup> gemessen. Die Arbeitstemperatur der CCD-Kamera beträgt  $-85\text{ }^{\circ}\text{C}$ . Die Anregung erfolgte über einen Multimode-Laser<sup>(d)</sup> bei  $785\text{ nm}$  mit ca.  $75\text{ mW}$  Anreizungsleistung gemessen unter dem  $\text{CaF}_2$ -Fenster. Der Arbeitsabstand der Sonde beträgt  $5\text{ mm}$ , der Fokus-Durchmesser ca.  $60\text{ }\mu\text{m}$ . Die numerische Apertur ist etwa  $0,2$ . Die Sonde ist in der Tiefe recht unempfindlich: die Halbwertsbreite des Si-Signals beträgt circa  $800\text{ }\mu\text{m}$ . Für Schweinehirn erhält man das beste Signal mit dem Fokus etwa  $250$  bis  $300\text{ }\mu\text{m}$  in der Probe. Damit ergibt sich ein Messvolumen in der Größenordnung von  $10^6\text{ }\mu\text{m}^3$  oder etwa tausend Zellen.

Shim und Wilson [131] empfehlen, Gewebe-Bulkproben vor dem Schockfrieren in Gefriermedium einzubetten. Vor der Messung soll das Gefriermedium mit physiologischer Kochsalzlösung mit Phosphatpuffer (PBS) abgewaschen werden und die Messung soll in PBS stattfinden. Auch in der Machbarkeitsstudie zur Erkennung von Hirnmetastasen (Mausmodell) in Bulkproben [CB12] wurde eine solche Präparation gewählt. Für das Astrozytom-Grading ist diese Form der Präparation jedoch nicht praktikabel, da besonders bei den Glioblastomproben das Gewebe oft so weit zerstört ist, dass es in PBS suspendiert. Wird das Gefriermedium andererseits nicht abgewaschen, so dringt es ins Gewebe ein und stört die Spektren, wie in Kapitel 13 ausführlicher gezeigt wird. Die hier untersuchten Proben wurden daher *ohne* Gefriermedium schockgefroren.

Die Raman-Messungen erfolgten von der Schnittfläche her an der Bulkprobe. Dazu wurde die Probe auf einem  $\text{CaF}_2$ -Fenster aufgetaut. Zur Dokumentation wurde ein Bild der aufgetauten Probe mit einem handelsüblichen Scanner aufgenommen. Für die Messungen habe ich eine feuchte Kammer gebaut (Abb. 10.1), in der die Probe weder austrocknet, noch in Puffer suspendieren kann. Die hängende Anordnung stellt eine ebene Probenoberfläche sicher, so dass die Sonde während der Messung nicht nachfokussiert werden musste. Die Gesamtmesszeit wurde auf  $6\text{ h}$  begrenzt, damit die Probe nicht verdirbt. Der Zeitbedarf vom Auftauen der Probe bis zum Beginn der Messung lag bei  $\frac{1}{2} - 1\text{ h}$ .

Die Raman-Messungen wurden mit Hilfe einer zu diesem Zweck unter Matlab erstellten grafischen Oberfläche durchgeführt<sup>(e)</sup>. Gegenüber der Gerätesoftware HoLoGRAM bietet diese RamanGUI folgende Vorteile:

---

(a) Kaiser Optical Systems Inc., Ann Arbor, MI, USA

(b) Roper Scientific Inc., Princeton Instruments, NJ, USA

(c) RamanProbe, Inphotonics, MA, USA

(d) Toptica Photonics Inc., NY, USA

(e) Die Oberfläche kann auch für Messungen unter dem Mikroskop verwendet werden und wurde für die Studie von Lattermann *et al.* [CB13] um eine Nachführung des Fokus in der Tiefe erweitert

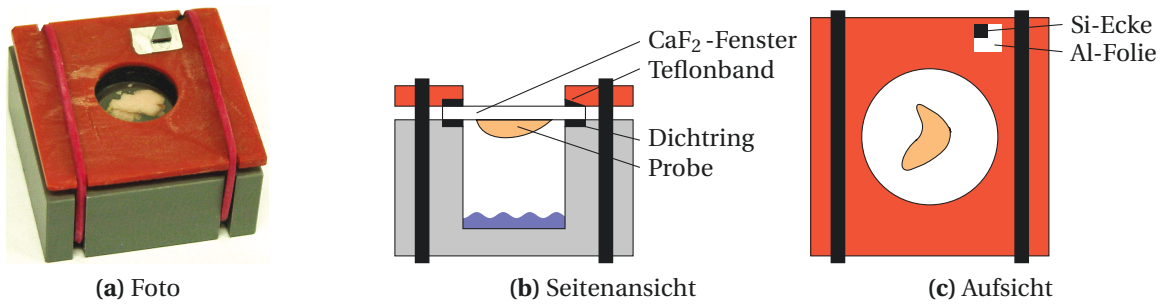


Abbildung 10.1 Feuchte Kammer für Raman-Messungen

- Messraster für beliebig geformte Flächen. Ein rechteckiges Messraster bedeutet bei den eher runden Gewebeproben etwa die  $1\frac{1}{2}$ -fache Messzeit.
- Messkoordinaten für die faseroptische Sonde werden unter dem Mikroskop erzeugt, indem die Probe „umfahren“ wird.
- Auch wenn eine kleine Verschiebung auftritt oder die Probe nur ungenau umfahren wurde, wird so die gesamte Probe gemessen, da ein zusätzlicher „Sicherheitsabstand“ angegeben werden kann. Spektren des Fenstermaterials werden dann ebenfalls während der Probenmessung mit aufgenommen.

Abbildung 10.2 veranschaulicht die Arbeitsschritte, eine ausführlichere Beschreibung mit Bildschirmfotos ist im Anhang C.2. Die RamanGUI läuft unter MatLab<sup>(f)</sup> und speichert die Messdaten in MatLab-Dateien. Die RamanGUI sucht vor jeder Messung die Ecke des Si-Stückes. Dabei wird auch automatisch bei jeder Messung die Wellenzahlkalibrierung überprüft. Das Spektrometer ist mit einem festen Gitter ausgestattet und sehr stabil, während der Messreihen musste nicht neu kalibriert werden. Die Spektren wurden als Rohsignal, also Counts über Pixeln, aufgenommen. Die Umrechnung in relative Wellenzahlen

<sup>(f)</sup> The Mathworks, Natick, USA

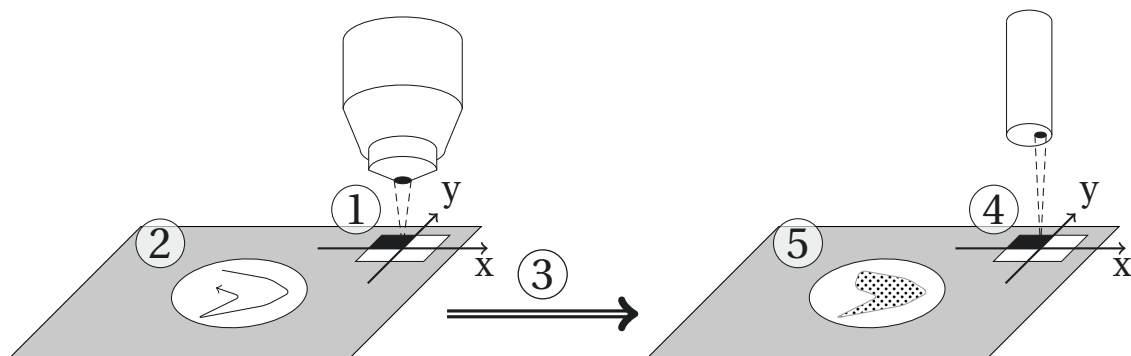


Abbildung 10.2 Raman-Messung mit der Steuersoftware RamanGUI. (1) Der Koordinatenursprung des Messrasters wird auf die Ecke des Si-Stückes gelegt. (2) Umfahren der Probe. Der Umriss des Messrasters wird dabei aufgenommen. Das Messraster wird durch Schrittweite und einen Sicherheitsabstand vom Probenrand definiert. (3) Die Probe wird unter der faseroptischen Sonde gefahren. (4) Die im ersten Schritt festgelegte Ecke wird mit Hilfe des Raman-Signals aufgesucht. Das Koordinatensystem wird wieder auf Null gesetzt. Da die feuchte Kammer auf dem Mikroskoptisch fixiert ist, genügt ein Punkt zur Synchronisation der Koordinaten. (5) Die Koordinaten des Messrasters werden abgearbeitet.

und die Intensitätskalibrierung waren Teil der Datenvorbehandlung (Kap. 12.1).

Die Raman-Spektren wurden in der Regel als  $2 \times 5$  Akkumulationen à 2 s mit der Option „Cosmic Ray Filter“ aufgenommen<sup>(g)</sup>. Sieben Proben wurden versehentlich ohne Cosmic Ray Filter, also mit einer Gesamtblendungszeit von 10 statt 20 s gemessen. In diesen Fällen wurden Spikes bei der Datenvorbehandlung entfernt. Bei einigen Proben wurde die Gesamtmesszeit von 20 s als  $2 \times 10$  s realisiert.

Da die Klassifikationsanwendung bei dieser Arbeit im Vordergrund steht, wurde die Messzeit pro Spektrum kurz gehalten und mehr Spektren pro Probe aufgenommen. 20 s pro Spektrum sind noch zu lang für die Anwendung während der Operation. Allerdings stand keine optimale Konfiguration des Messgerätes zur Verfügung. Die Anregungsfaser der Sonde hatte nur den halben Durchmesser vom Ausgang des Anregungslasers, so dass der größte Teil der Intensität bereits beim Einkoppeln in die Anregungsfaser verloren ging. Andererseits hat die verwendete Sonde einen Fokusbereich von etwa 60 µm, so dass aufgrund der thermischen Belastung der Probe die Anregungsleistung nicht wesentlich oberhalb der gemessenen 70 mW liegen sollte. Eine passende Sondenkonfiguration ermöglicht die Aufnahme von Raman-Spektren mit vergleichbarem Signal-Rausch-Verhältnis in wenigen Sekunden (zum Beispiel die InPhotonics Low Cost Probe mit 150 µm Fokusbereich: etwa 4 – 5 s).

---

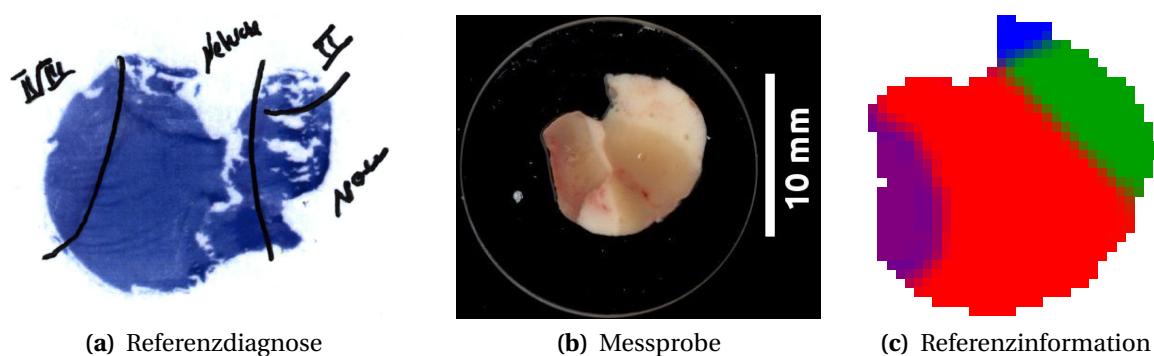
<sup>(g)</sup> Mit der Geräteeinstellung „Cosmic Ray Filter“ werden zwei Spektren entsprechend der Blendungseinstellungen gemessen. Daher verdoppelt sich die effektive Messzeit. Das resultierende Spektrum ist das Mittelwertspektrum dieser beiden Messungen.

## 11 Referenzdiagnose

Die Parallelschnitte wurden nach dem Färben der Neuropathologin des Universitätsklinikums Dresden, Fr. PD Dr. Geiger, zur histologischen Begutachtung vorgelegt. Als Ergebnis der mikroskopischen Untersuchung zeichnete sie auf Ausdrucken von Scans der Parallelschnitte die Bereiche mit den verschiedenen Geweben ein (Abb. 11.1a).

Diese detaillierte Beurteilung der einzelnen Proben verfolgt zwei Ziele. Zum Einen stellt sie sicher, dass das untersuchte Gewebe tatsächlich Tumor- beziehungsweise normales Gewebe ist. Zum Anderen zielt diese Arbeit auf ein Grading *innerhalb eines Tumors*. Auch die Referenzinformation muss also Gewebeunterschiede innerhalb der einzelnen Proben widerspiegeln. Die Tumore sind heterogen und die Tumorränder während der Operation nur schwer zu erkennen. Hinzu kommt, dass Tumorproben am Rand des Tumors genommen werden. Daher ist mit einem relativ großen Probennahmefehler zu rechnen: die Probe kann unbeabsichtigt statt Tumorgewebe normales Gewebe oder größere Bereiche mit morphologisch niedriggradigerem Gewebe enthalten als es der Diagnose des Patienten entspricht. Die histologische Begutachtung stellt fest, welches Gewebe tatsächlich in der Probe vorliegt.

An Referenzinformationen für ein Tumorgrading innerhalb eines Tumors werden allerdings noch weitere Anforderungen gestellt. Normalerweise hat ein Tumor genau einen Grad. Dieser ist histologisch durch das höchstgradige Gewebe definiert, das irgendwo in diesem Tumor existiert – sei es auch noch so klein. Allerdings sind die Astrozytome sehr heterogene Tumore. Selbst Glioblastome haben oft Bereiche, die morphologisch niedriggradigen Astrozytomen ähneln (vgl. Tab. 14.1, S. 114). Bei der chirurgischen Entfernung des Tumors wird oft ein Resektionsrand zwischen dem hochgradigen und dem (morphologisch) niedriggradigen Gewebe angestrebt. Eine Diagnostik, die dem Chirurgen wäh-



**Abbildung 11.1** Referenzdiagnose. (a) Scan des mit Methylenblau gefärbten Parallelschnittes mit eingezeichneter Referenzdiagnose. (b) Die Probe unmittelbar vor der Raman-Messung (durch den Objektträger gesehen). (c) Die Referenzdiagnose auf die einzelnen Punkte des Messrasters übertragen: grün für normales Gewebe (N), blau: niedriggradiges Astrozytomgewebe (A°II) und rot hochgradiges Astrozytomgewebe (A°III+). Mischfarben stehen für die Mischungen der Klassen, z. B. der violette Bereich für  $\frac{1}{2}$  A°II und  $\frac{1}{2}$  A°III+.

rend der Operation beim Festlegen des Resektionsrandes unterstützen soll, muss also Unterschiede innerhalb des Tumorgewebes eines Tumors erkennen. Als Referenz für ein operationsunterstützendes Grading wird daher eine detailliertere Einstufung als der normale Tumorgrad benötigt.

Im Folgenden bezeichnet der Tumorgrad eines Gewebes daher immer Gewebe, das morphologisch dem am weitesten entdifferenzierten Gewebe eines Tumors dieses Grades ähnelt. Zwischen diesem Konzept und dem normalen Grading eines Tumors bestehen wichtige Unterschiede. Beispielsweise kann ein Gewebe eines hochgradigen Astrozytoms morphologisch wie Astrozytomgewebe zweiten Grades aussehen. Dennoch ist unklar, ob es sich auch biochemisch wie ein Astrozytom zweiten Grades verhält. Raman-Spektren messen nicht die Morphologie, sondern die biochemische Zusammensetzung des Gewebes. Unterschiede in den Raman-Spektren fallen also möglicherweise nicht genau mit Unterschieden in der Morphologie zusammen. Die Unterschiede zwischen dem Tumorgrad des Patienten und dem detaillierten Tumorgrad (einer bestimmten Region der Probe) werden durch die Heterogenität der Proben verstärkt. Sobottka *et al.* [71] demonstrieren diese Heterogenität von Hirntumorproben in einer IR-Imaging Studie anhand von 54 Proben von 6 Patienten.

**Weiche Referenzinformationen:** Die Ergebnisse der histologischen Begutachtung spiegeln die in Kapitel 2.1 (S. 7) und 4.5 (S. 36) eingeführte Mischung der Klassen und gegebenenfalls die Unsicherheit bei der Diagnose als weiche Referenzinformation wider, zum Beispiel in Abbildung 11.1a die Region links mit °II/III.

Unschärfe Referenzdiagnosen wie °II/III wurden mit Anteilen von jeweils 0,5 in den entsprechenden Klassen eingetragen. Eine andere unscharfe Diagnose lautet „einzelne Tumorzellen“. Hier wurde der Anteil des jeweiligen Tumors mit 5% angesetzt. Formulierte die Neuropathologin eine Unsicherheit bei der Diagnose (entweder Gewebe A oder Gewebe B), so wurde dieses mit Anteilen von 0,5 codiert.

**Übertragung der Referenzdiagnose auf die Spektren:** Unüberwachte chemometrische Verfahren wie die Clusteranalyse können ohne Referenzinformationen Strukturen in den Spektren aufzeigen. Solche Strukturen können mit den Referenzdiagnosen verglichen werden, so dass eine sehr einfache Übertragung der Referenzinformationen möglich ist. Dieses Vorgehen ist eng verwandt mit der sogenannten halbüberwachten Modellbildung (Kap. 19, S. 150). Dabei gehen aber Informationen aus den Spektren in die Referenzinformation ein. Die Referenzinformationen sind nicht mehr unabhängig von den Spektren. Das kann dazu führen, dass Klassen einfacher trennbar erscheinen, als sie in Wirklichkeit sind. Die Cluster sind ja eine Unterteilung der Daten in Gruppen, die aufgrund ihrer Spektren trennbar sind. Damit ist *garantiert*, dass auch eine geeignete Klassifikation die Daten entlang der Clustergrenzen trennen kann. Genauer gesagt sind die Clustergrenzen besonders einfach auffindbare Trennflächen. Unklar bleibt jedoch, wie gut die Clustergrenzen mit den histologischen Grenzen und der Verteilung der verschiedenen biochemischen Substanzen im Gewebe übereinstimmen. Letzteres muss im Rahmen einer Validierung des Klassifikationsmodells immer überprüft werden. Dazu müssen die Referenzinformationen zu den Validierungsproben aber unabhängig von den Spektren sein, da sonst ein optimistischer Bias entstehen kann. Wird die Referenz-

information mit Hilfe einer Clusteranalyse auf die Spektren übertragen, so kann der Datensatz also höchstens zur Modellbildung herangezogen werden, keinesfalls aber für die Modellvalidierung.

In der medizinischen Forschung sind Doppelblindstudien der Standard. Diese Technik soll vermeiden, dass der Patient durch unbewusste Signale des Forschers beeinflusst wird. Im Rahmen des chemometrischen Trainings der Klassifikationsmodelle kann zwar angenommen werden, dass die Klassifikationsmodelle selbst deterministisch aus den Trainingsdaten folgen, aber die Übertragung der Referenzdiagnose auf die Spektren kann subjektiv beeinflusst sein. Aus diesem Grund wurden die Referenzdiagnosen *ohne* Zuhilfenahme der Spektren auf das Messraster übertragen: nur die Ergebnisse der detaillierten histologischen Untersuchung (Abb. 11.1a), der Scan der Probe (Abb. 11.1b) und die Koordinaten der einzelnen Messpunkte wurden angezeigt (Pixel in Abb. 11.1c). Die Übertragung erfolgte im Rahmen eines Forschungspraktikums durch Tobias Schulz und Ben Schüppel.

In einigen Fällen konnte die genaue Orientierung der Probe nicht nachvollzogen werden. Das war besonders bei den großen Proben der Fall, die den gesamten Querschnitt des Gefrierröhrchens ausfüllten und deshalb rund waren. Hochgradige Tumorproben verformen sich oft stark beim Auftauen, da das Gewebe mechanisch beeinträchtigt (fast flüssig) ist. In diesen Fällen gibt die Referenzinformation die Flächenanteile der unterschiedlichen Gewebe am gefärbten Schnitt wieder und ist bei allen betroffenen Spektren gleich. Die Unsicherheit bei der Übertragung der Referenzdiagnosen auf die Spektren ist eine Hauptquelle der Unsicherheit im Datensatz. Etwa  $\frac{1}{8}$  der Schnitte ist stark von diesem Problem betroffen, bei weiteren 20 % der Schnitte sind die Auswirkungen geringer, da der größte Teil der Probe homogen ist.

Der Begriff weiche Spektren bezeichnet im Folgenden Spektren, die nicht eindeutig zu einer einzigen Klasse zugeordnet werden können. Dementsprechend gehören harte Spektren eindeutig zu einer der Klassen.



## 12 Datenanalyse

**Datenbank:** Die Spektrendaten wurden in eine PostgreSQL<sup>(a)</sup>-Datenbank eingetragen und auf dem Dateiserver (Anh. C.3) abgelegt. Die Datenbank enthält auch digitale anonymisierte Patientenbriefe, Bilder der Bulkprobe vor der Raman-Messung, Scans der detaillierten histologischen Diagnose sowie weitere spektroskopische Messungen.

Struktur und Nomenklatur dieser Datenbank spiegeln die Hierarchie der Proben wieder: von einem Patienten liegen gegebenenfalls mehrere Proben vor. Von einer Probe wurden gegebenenfalls mehrere Schnitte präpariert. Von einem Schnitt können mehrere Messungen vorliegen, die ihrerseits aus vielen Spektren bestehen. Von dieser Hierarchie wurden die Ebenen Probe, Schnitt und Messung in der Datenstruktur abgebildet. Da von den meisten Patienten nur eine Probe vorliegt, wird jeder Probe eine Patientenummer zugeordnet und an dieser Stelle auf die Normalform verzichtet.

Die Bulkproben haben immer mindestens die Schnittnummer 2, da die 15  $\mu\text{m}$ -Schnitte auf  $\text{CaF}_2$  als Schnitt 1 gezählt wurden. Außerdem gibt es von diesen Bulkproben nie mehr als eine Messung, da sie nach ca. 6 h in der feuchten Kammer entsorgt wurden, während von den Gefrierschnitten mehrere IR- und Raman-Messungen vorliegen können. Die Messungen der Bulkproben werden im Folgenden als *Probe.Schnitt* eindeutig benannt.

**Software zur chemometrischen Datenauswertung:** Die gesamte statistische Auswertung wurde in R[253] durchgeführt. Dabei wurde zur optimierten und parallelisierten Berechnung von linearer Algebra die Bibliothek OpenBLAS [256] zusammen mit dem von Simon Fuller unter meiner Betreuung erstellten R-Paket *OpenBlasThreads* [CB15] zur Feinkontrolle der Parallelisierung genutzt. Die graphische Darstellung nutzt außer den Standardpaketen auch *lattice* [257] und *ggplot2* [258]. Zur Datentransformation wurden auch *plyr* [259] und *reshape2* [260] sowie das im Rahmen der Entwicklung von *softclassval* [CB1] entstandene Paket *arrayhelpers* genutzt. Viele Grafiken und Tabellen wurden mit *knitr* [261–263] erstellt und dann in die vorliegende Arbeit eingebunden. Die von der RamanGUI erstellten Matlab-Dateien wurden mit Hilfe des Pakets *R.matlab* [264] in Reingelesen und in ein *hyperSpec* [CB14]-Objekt umgewandelt.

### 12.1 Vorbehandlung der Raman-Spektren

**Schneiden des Spektralbereiches:** Als erster Datenvorbehandlungsschritt wurden uninformative Spektralbereiche entfernt. Der Interferenzfilter (Abb. 3.2) unterdrückt Licht mit  $\Delta\tilde{\nu} < 150 \text{ cm}^{-1}$ . Am langwelligen Rand des Spektrums nimmt die Quantenausbeute der CCD-Kamera stark ab. In beiden Fällen resultiert ein geringes Signal und damit ein niedriges Signal-Rausch-Verhältnis. Der kurzwellige Spektralbereich unterhalb von  $750 \text{ cm}^{-1}$  zeigt außerdem ein sehr hohes Untergrundsignal, was das (Raman-) Signal-

<sup>(a)</sup> The PostgreSQL Global Development Group, <http://www.postgresql.org>

Rausch-Verhältnis weiter beeinträchtigt. Die Raman-Spektren von Hirngewebe enthalten keine Banden im Spektralbereich oberhalb der Carbonyl-Banden  $\nu_{C=O}$  bis zum Beginn der C – H-Valenzschwingungsbanden  $\nu_{C-H}$ . Zunächst wurden daher die Spektralbereiche unterhalb von  $750\text{ cm}^{-1}$  und zwischen  $1850\text{ cm}^{-1}$  und  $2500\text{ cm}^{-1}$  entfernt.

**Entfernen von Spikes:** Bei sieben Messungen war der *Cosmic Ray Filter* beim Messen ausgeschaltet. Die Spikes in den Spektren wurden manuell als NA markiert, also gelöscht. Spikes können als sehr intensive und scharfe Signale, die nur in einzelnen Spektren auftreten, gut aufgefunden werden.

**Zusammenführen der Akkumulationen:** Akkumulationen wurden als Wiederholungsmessungen an derselben Tischposition realisiert. Diese wurden nun zu je einem Spektrum für jede Position zusammengeführt. Dazu wurden die Spektren einer multiplikativen Signalkorrektur (engl. *multiplicative signal correction, MSC*) unterzogen. Die MSC beschreibt Spektren durch ein konstantes Untergrundsignal  $b$  und ein um den Faktor  $a$  skaliertes Referenzspektrum  $I^{Ref}$ .

$$I_i = aI_i^{Ref} + b \quad (12.1)$$

Das Spektrum wird dann um die beiden Einflüsse  $a$  und  $b$  korrigiert:

$$I_i^{korr.} = \frac{1}{a}(I_i - b) \quad (12.2)$$

Die MSC nimmt also an, dass eigentlich alle betrachteten Spektren gleich sind. Sie sind nur durch einen (konstanten) Untergrund und einen Skalierungsfaktor (zum Beispiel aufgrund unterschiedlicher optischer Pfade) gestört. Diese Annahme ist für die Wiederholungsmessungen an derselben Position sinnvoll.

Schließlich wurde das Mittelwertspektrum der MSC-korrigierten Spektren gebildet.

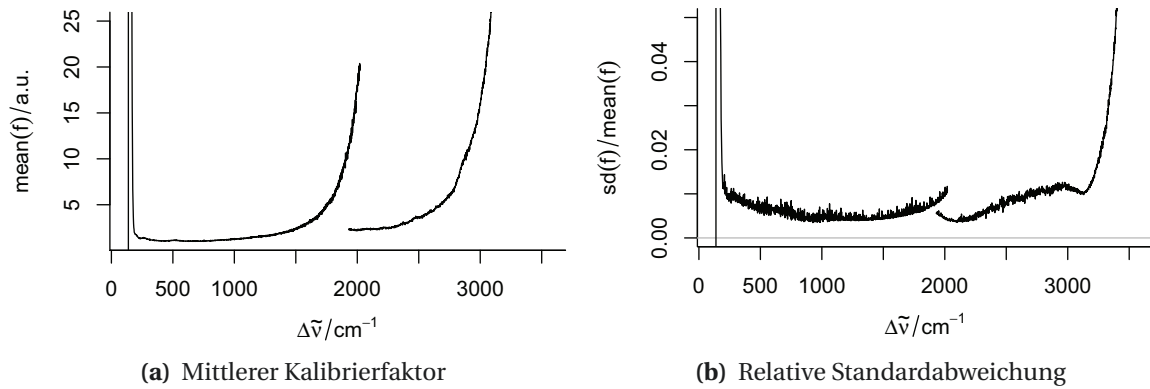
**Dunkelstromkorrektur und Intensitätskalibrierung:** Probleme mit dem Kalibrierlicht<sup>(b)</sup> führten dazu, dass die Intensitätskorrektur der Gerätesoftware nicht verwendet werden konnte. Die Intensitätskalibrierung wurde daher im Rahmen der Datenvorbehandlung durchgeführt.

Der CCD-Detektor wird mit einer Vorspannung betrieben. Dies führt auch ohne Licht zu einem Signal. Dieser Dunkelstrom von 24,1 Counts/s war über den Spektralbereich und über die Zeit konstant und wurde zuerst abgezogen.

In mehreren Messreihen wurden insgesamt über 4500 Spektren des Kalibrierlichts aufgenommen. Mit Hilfe des zertifizierten Spektrums der Lichtquelle wurden die Kalibrierfaktoren errechnet. Abbildung 12.1 zeigt den Mittelwert (a) und die relative Standardabweichung (b) in Abhängigkeit von der relativen Wellenzahl. Nach der Dunkelstrom-Korrektur wurden die Spektren mit den Kalibrierfaktoren in Abbildung 12.1a multipliziert.

Diese Intensitätskalibrierung korrigiert die Wellenlängenabhängigkeit der Detektorempfindlichkeit, den Einfluss der Projektionsgeometrie im Spektrometer und das Trans-

<sup>(b)</sup> Die Kalibrierlampe schaltete sich in unregelmäßigen Abständen selbständig aus.



**Abbildung 12.1** Intensitätskalibrierung. Die Messkanäle des Spektrums werden jeweils mit einem Faktor multipliziert, um das intensitätskalibrierte Spektrum zu erhalten. (a) zeigt den spektralen Verlauf dieser Faktoren, die relative Standardabweichung ist in (b) aufgetragen. Das Spektrometer bildet die Spektralbereiche bis ca.  $2000\text{ cm}^{-1}$  und ab ca.  $2000\text{ cm}^{-1}$  Raman-Verschiebung auf getrennte Bereiche (übereinander) der CCD-Kamera, so dass die Kalibrierfunktion zwei Abschnitte hat.

missionsprofil von Sonde und Spektrometeroptik. Auf weitergehende Korrekturen, zum Beispiel der  $\nu_s^4$ -Abhängigkeit von der Anregungswellenlänge, wurde verzichtet, da sie für die vorliegende Anwendung nicht sinnvoll sind (Kap. 3.1).

**Basislinien-Korrektur:** Raman-Spektren zeigen oft eine Basislinie. Zu diesem Untergrundsignal tragen zum Beispiel der Dunkelstrom der CCD, Streulicht [89], Küvettenmaterial und auch schwache Ausläufer von Fluoreszenzbanden bei. Eine Basislinienkorrektur soll diesen Untergrund herausrechnen. Dunkelstrom und Spektrum des Küvettenmaterials (bzw. hier des  $\text{CaF}_2$ -Fensters) können gemessen werden. Daher kann der Dunkelstrom direkt abgezogen werden (siehe oben). Beim Küvettenmaterial ist der Anteil am gemessenen Spektrum variabel und muss für die Korrektur abgeschätzt werden. Für die durch Streulicht und Fluoreszenz verursachten Untergrundsignale ist in der Regel kein genaues Spektrum bekannt, sondern nur, dass sie eine sehr breite spektrale Signatur haben. In der Praxis werden sie daher durch Näherungen, zum Beispiel Polynome, beschrieben.

Die Spektren zeigten besonders im niedrigen Wellenzahlbereich von  $750$  bis  $1850\text{ cm}^{-1}$  einen hohen Untergrund. Die Funktion `hyperSpec::spc.fit.poly.below` beschreibt die Basislinie des Spektrums durch ein Ausgleichspolynom. Geeignete Spektralbereiche für die Stützstellen des Polynoms werden automatisch gesucht. Im niedrigen Wellenzahlbereich von  $755$  bis  $1850\text{ cm}^{-1}$  wurde eine quadratische Funktion gewählt. Der Spektralbereich der C–H-Valenzschwingungen  $\nu_{C-H}$  wurde um eine Gerade zwischen  $2625$  und  $3100\text{ cm}^{-1}$  korrigiert. Die  $\text{CaF}_2$ -Bande bei  $320\text{ cm}^{-1}$  liegt außerhalb des betrachteten Spektralbereichs. Außer dieser Bande weist das Raman-Spektrum des verwendeten Fenstermaterials keine Struktur auf, so dass es nicht gesondert korrigiert werden muss.

**Glättende Interpolation:** Der Datenpunktabstand der Spektren variiert in Abhängigkeit der relativen Wellenzahl, weil ein Gitterspektrometer verwendet wurde. Spektren mit konstantem Datenpunktabstand können mit einer Interpolation erzeugt werden. In die-

sem Schritt kann außerdem in gewissen Grenzen das Signal-Rausch-Verhältnis auf Kosten der spektralen Auflösung erhöht werden. Der Datenpunktabstand nach der Interpolation sollte die halbe spektrale Auflösung nicht unterschreiten, weitere Datenpunkte tragen nicht zum Informationsgehalt des Spektrums bei. Bei der Modellbildung müssen jedoch für jeden Datenpunkt Parameter geschätzt werden. Daher sollte die spektrale Auflösung letztlich nicht höher sein, als für die chemometrische Klassifikation notwendig ist. Mit `hyperSpec::spec.loess` wurden die Spektren auf eine neue Wellenzahl-Achse umgerechnet, die von 755 bis 1800  $\text{cm}^{-1}$  und von 2800 bis 3025  $\text{cm}^{-1}$  reicht und einen Datenpunktabstand von 5  $\text{cm}^{-1}$  hat. Der Glättungsparameter wurde entsprechend so eingestellt, dass die spektrale Auflösung der interpolierten Spektren etwa 10  $\text{cm}^{-1}$  beträgt.

**Spektren mit zu geringer Intensität und Ausreißer:** Die Messungen wurden mit einem Sicherheitsabstand von zwei bis drei Spektren um die Probe herum durchgeführt. Dadurch enthalten die Daten auch Spektren niedriger Intensität, die nicht von den Proben stammen. Diese Spektren wurden aus dem Datensatz entfernt. Der Grenzwert für die Intensität wurde für jede Probe einzeln festgelegt. Er kann meist sehr einfach aus der Verteilung der Gesamtintensität im Spektralbereich von 2800 bis 3025  $\text{cm}^{-1}$  abgelesen werden: die Werte zeigen am Übergang vom leeren Objektträger zur Probe eine Stufe.

Bei einigen Spektren (besonders von koaguliertem Blut) war der Detektor gesättigt. Außerdem zeigten einzelne Spektren das Spektrum der Laborbeleuchtung. Auch diese Spektren wurden aus der weiteren Auswertung ausgeschlossen.

**Intensitätsnormierung:** Für eine erfolgreiche Normierung muss die Basislinie in dem Spektralbereich, der zum Berechnen der Normierungsfaktoren benutzt wird, bekannt bzw. vollständig korrigiert sein. Anderenfalls würde der vorhandene Untergrund unter dem Raman-Signal den Normierungsfaktor beeinträchtigen. Außerdem sollte das Signal-Rausch-Verhältnis hinreichend gut sein, also auch bei allen Spektren eine genügend große Intensität in diesem Spektralbereich vorliegen.

Diese Bedingungen sind im Spektralbereich von 2900 bis 3025  $\text{cm}^{-1}$  erfüllt. Die Spektren zeigen um die C – H-Valenzschwingungsbanden eine wohldefinierte Basislinie und haben zwischen 2900 und 3025  $\text{cm}^{-1}$  eine ähnliche Form sowie ein sehr gutes Signal-Rausch-Verhältnis (im Mittel 29, während das mittlere Signal-Rausch-Verhältnis im Spektralbereich von 755 bis 1800  $\text{cm}^{-1}$  nur bei etwa 6 lag) Daher wurde eine Flächennormierung vorgenommen, so dass die mittlere Intensität in diesem Spektralbereich 1 wird:

$$I^{norm}(\Delta\nu_i) = n \frac{I(\Delta\nu_i)}{\sum_{\Delta\nu_i=2900\text{ cm}^{-1}}^{3025\text{ cm}^{-1}} I(\Delta\nu_i)} \quad (12.3)$$

mit der Anzahl der Datenpunkte  $n$  im Spektralbereich von 2900 bis 3025  $\text{cm}^{-1}$ .

Die Normierung bewirkt, dass die Intensitäten zwischen 2900 und 3025  $\text{cm}^{-1}$  linear abhängig werden. Hohe Korrelation oder gar Kollinearität zwischen den Variaten können besonders in der LDA zu instabilen Modellen führen. Um diese lineare Abhängigkeit zu brechen, wurde die letzte Variate nach der Normierung aus dem Datensatz entfernt.

**Zentrieren:** Die verwendete LDA-Routine nimmt immer selbst eine Zentrierung vor. Die logistischen Regressionsmodelle zeigen mit oder ohne vorherige Zentrierung der Daten (auf normales graues Gewebe) keine praktisch bedeutsamen Unterschiede auf. Daher wurde auf eine Zentrierung verzichtet.

**Skalieren:** Wie in Kapitel 4.2 erläutert, ist eine Varianzskalierung von spektroskopischen Datensätzen in der Regel nicht sinnvoll. Für die Modellbildung mit `multinom` wird empfohlen, Eingangsdaten mit einem Wertebereich in der Größenordnung von 0 bis 1 zu verwenden [150, 265]. Aufgrund der Normierung ist das jedoch bereits annähernd der Fall. Daher wurden die Daten nicht zusätzlich skaliert.

**Ausschluss zweier Messungen:** Eine LDA mit den einzelnen Messungen als Klassen überprüft, ob sich bestimmte Messungen stark von allen anderen unterscheiden. Die ersten beiden Diskriminanzfunktionen trennen sechs Messungen (farbige Punkte in Abb. 12.2a) vom Rest der Daten (2d-Histogramm: graues 2d-Histogramm). Abbildung 12.2b zeigt die Spektren dieser Messungen (einzelne Punkte) im Vergleich zu den Spektren aller anderen Messungen.

Zwei dieser Bulkproben (Messung 1227.2 und 1633.2<sup>(c)</sup>) separieren sich auch in einer Hauptkomponentenanalyse (in den Hauptkomponenten zwei und drei, nicht gezeigt) von den restlichen Spektren. Besonders im Spektralbereich zwischen  $1400\text{ cm}^{-1}$  und  $1800\text{ cm}^{-1}$  weisen sie ein extrem schlechtes Signal-Rausch-Verhältnis und auch nach der Basislinienkorrektur noch einen sehr hohen Untergrund auf. Die Spektren der anderen vier in der Ausreißer-LDA auffälligen Messungen sind den Spektren aller anderen Proben wesentlich ähnlicher. Daher wurden die beiden Messungen der Proben 1227 und 1633 von der weiteren Auswertung ausgeschlossen. Bei den Proben 46, 53 und 60 handelt es sich um die alten Kontrollproben, die zu den ältesten Proben der Sammlung gehören. Da insgesamt nur neun Kontrollproben zur Verfügung standen, wurden diese drei Proben *nicht* ausgeschlossen. Sie werden in Kapitel 15 ausführlicher untersucht.

## 12.2 Einteilung und Charakterisierung der Datensätze

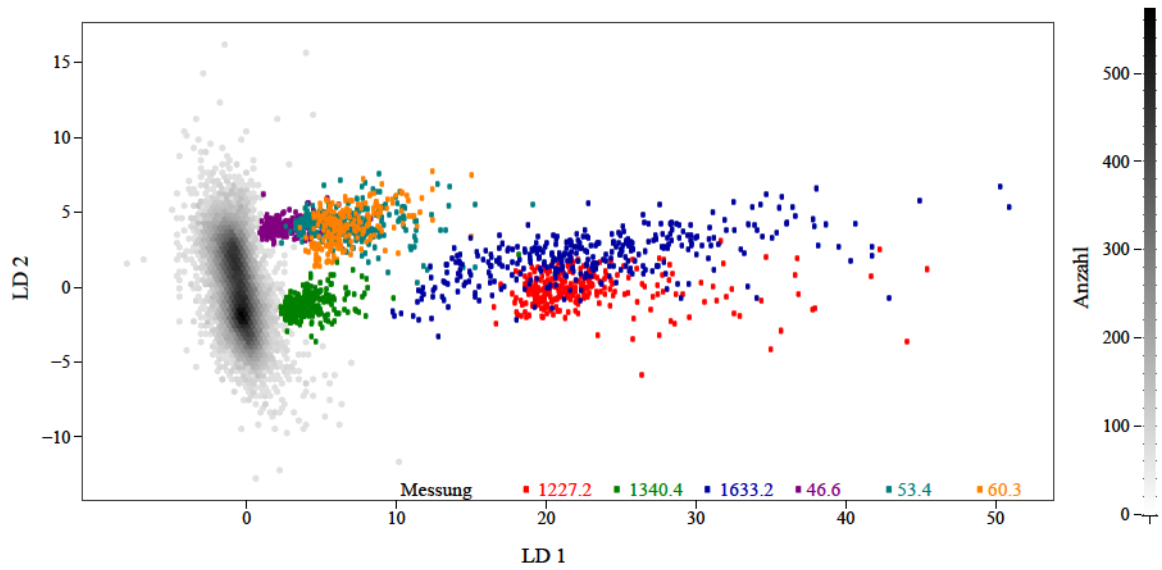
Diese Arbeit betrachtet zwei unterschiedliche diagnostische Fragestellungen: das Grading von Astrozytomen und die Differentialdiagnostik von Astrozytomen und Lymphomen.

### 12.2.1 Grading von Astrozytomen

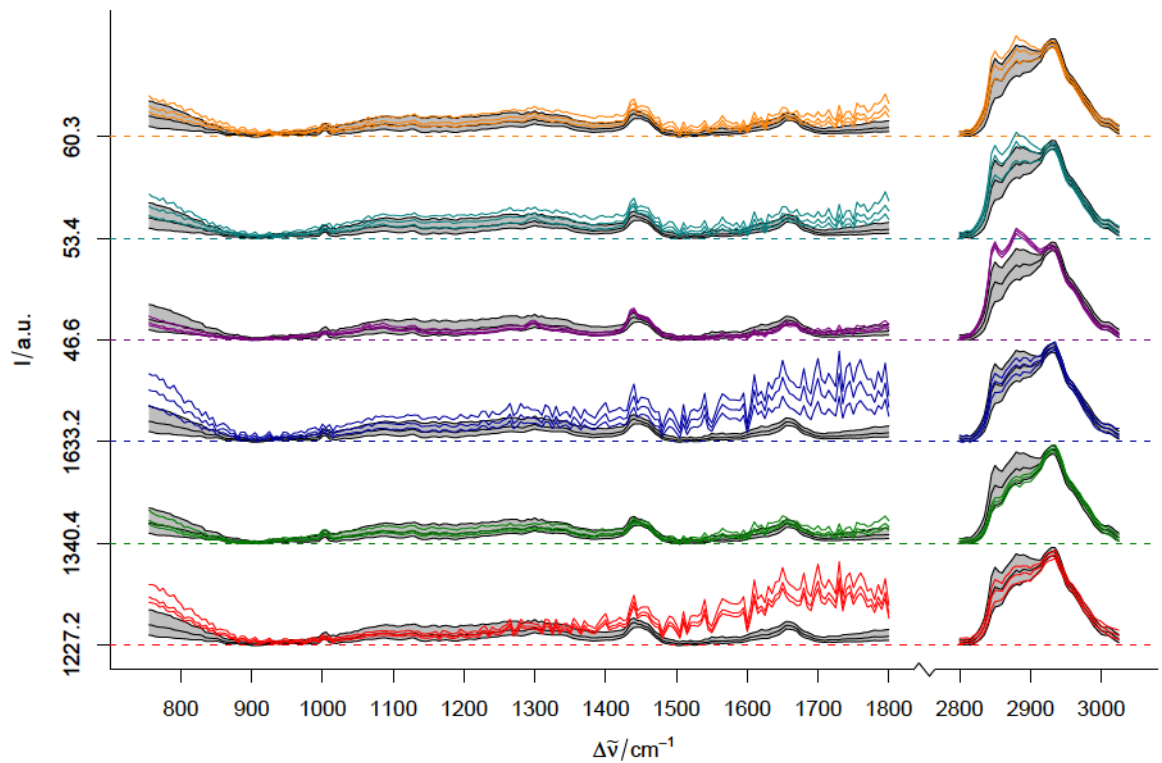
Der Erfolg einer Astrozytomoperation hängt stark davon ab, ob der gesamte Tumor entfernt werden kann [17, 18]. Astrozytome wachsen allerdings infiltrativ. Es gibt also keine scharfe Grenze zwischen Tumor und normalem Gewebe. Weiterhin entsteht um den Tumor herum oft ein Ödem, so dass Tumor- und normales Gewebe visuell nur sehr schwer unterscheidbar sind. Um normales Gewebe nicht zu gefährden, zielt die chirurgische

---

<sup>(c)</sup> Die Nummerierung folgt dem Schema *Probe.Schnitt*.



(a) In einer LDA zum Auftrennen der einzelnen *Messungen* erscheinen sechs Messungen auffällig (farbige Punkte: jeder Punkt entspricht einem Spektrum). Sie separieren sich im Koordinatensystem der ersten beiden LDA-Scores deutlich von allen anderen Messungen (2d-Histogramm in grau).



(b) Die Spektren der auffälligen Messungen (Median, 16. und 84. Perzentil). In grau die entsprechenden Perzentile aller anderen Spektren.

**Abbildung 12.2** Ausreißerkontrolle. (a) Sechs Messungen (farbige Punkte) unterscheiden sich stark von allen anderen Messungen (als 2d-Histogramm: graue Wolke). (b) Der Fingerabdruck-Bereich der Proben 1227 und 1633 hat ein extrem schlechtes Signal-Rausch-Verhältnis. Die Messungen von Probe 1227 und 1633 (jeweils Schnitt 2) werden aus der weiteren Auswertung ausgeschlossen.

**Tabelle 12.1** Zusammensetzung des Datensatzes Astro

Klasse	Weiche Referenz		davon mit harter Referenz	
	Patienten	Spektren	Patienten	Spektren
Normal	34	15401	15	7290
davon Kontrolle	9	4902	9	4902
Astrozytom °II	45	18668	16	4168
Astrozytom °III+	52	21342	27	8279
Insgesamt	78	36391	52	19737

Entfernung von Astrozytomen daher oft nur auf den malignen Teil des Tumors. Die Astrozytome sind in dieser Arbeit daher im Datensatz Astro in drei Klassen unterteilt, die die chirurgischen Bedürfnisse widerspiegeln:

**N:** Gewebe, die nicht zum Tumor gehören. Außer normaler grauer und weißer Substanz enthält diese Klasse auch geringe Anteile weicher Hirnhaut und Gliosen. Der Einfachheit halber wird diese Klasse als normal bezeichnet.

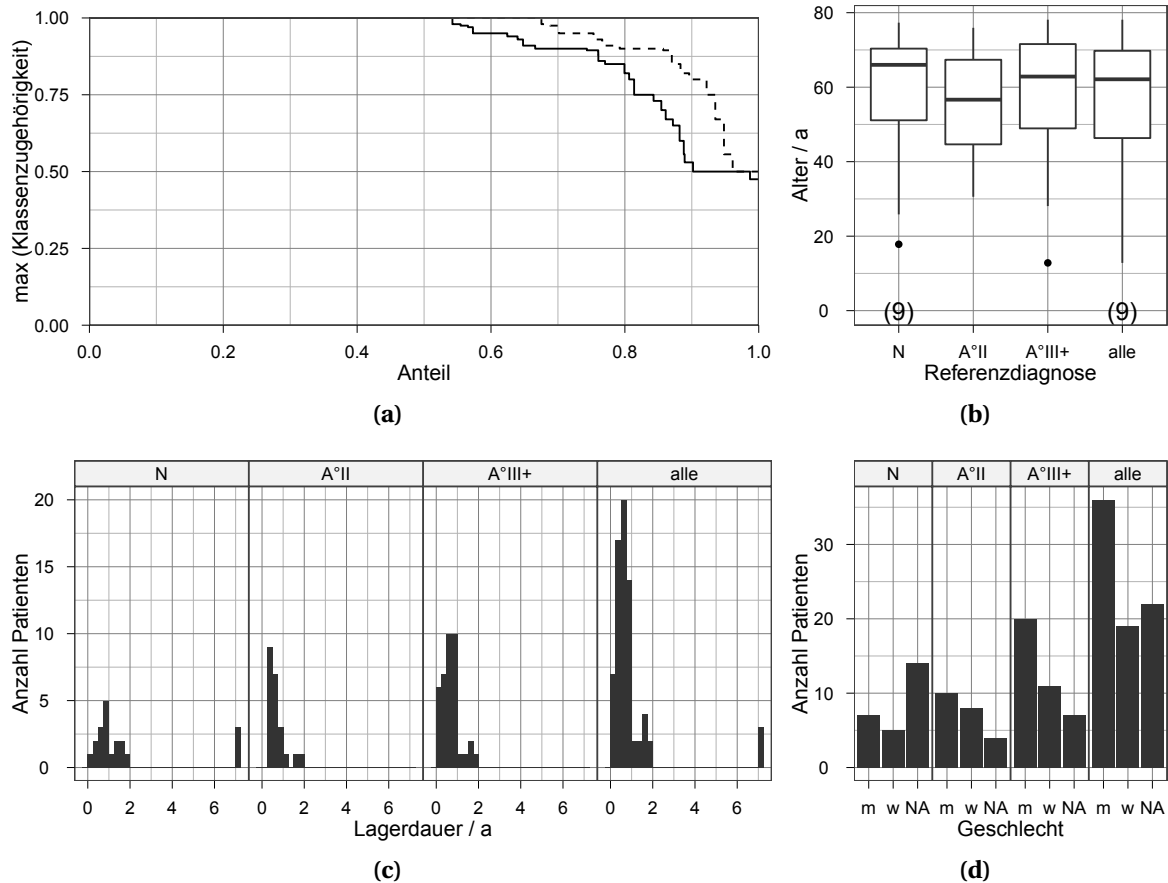
**A °II:** niedriggradiges Astrozytomgewebe

**A °III+:** hochgradiges Astrozytom- und Glioblastomgewebe und Nekrosen.

Im Folgenden werden die Farben grün für normal, blau für niedriggradiges Astrozytomgewebe und rot für hochgradiges Astrozytomgewebe verwendet.

Tabelle 12.1 und Abbildung 12.3 geben einen Überblick über die Zusammensetzung des Datensatzes. Abbildung 12.3a zeigt die Klassenzugehörigkeit der vorherrschenden Klasse  $r_{max}$  jedes Spektrums, also die Zeilenmaxima der Matrix mit den Referenzklassenzugehörigkeiten  $\mathbf{G}_{Ref}$ , über dem Anteil der Spektren (durchgezogen), die mindestens diesen Wert erreichen. Gestrichelt ist der Anteil der Patienten, bei denen mindestens 1 Spektrum  $r_{max}$  erreicht. Von gut 36 000 Spektren von 78 Patienten können 54 % (von 52 Patienten, also etwa bei  $\frac{2}{3}$  der Patienten) eindeutig einer der drei Klassen zugeordnet werden. Die Kontrollproben stellen etwa  $\frac{2}{3}$  der harten Spektren von Normalgewebe, jedoch nur knapp  $\frac{1}{3}$  der Spektren mit Anteilen normalen Gewebes. Niedriggradiges Astrozytomgewebe ist mit nur wenig über 4000 harten Spektren die kleinste Klasse. Dazu kommen allerdings noch  $3\frac{1}{2}$  mal so viele weiche Spektren mit niedriggradigem Anteil. Damit ist die Erkennung von niedriggradigem Astrozytomgewebe voraussichtlich extrem schwierig.

Das Alter der Patienten ist in allen drei Klassen (entsprechend der Referenzzuordnung) ungefähr gleich, wobei ein gewisser Trend zu steigendem Alter mit steigender Malignität zu beobachten ist. Dieser Trend fällt wesentlich geringer aus als der entsprechende Trend der Malignität des Tumors des Patienten bei Erstdiagnose mit dem Alter (siehe [26]). Das ist insofern zu erwarten, als ein großer Anteil der morphologisch niedriggradigen Tumorgewebe von Patienten mit malignen Gliomen stammt. Gliome treten häufiger bei Männern als bei Frauen (etwa im Verhältnis 4 : 3) auf [26]. Auch im vorliegenden Datensatz stammen mehr Proben von Männern (36) als von Frauen (19) (Abb. 12.3d). Allerdings ist bei 22 der 78 Patienten das Geschlecht unbekannt. Das ist immer dann der Fall, wenn außer den histologischen Befunden nicht noch der Patientenbrief zur Verfügung stand. Von den neun Kontrollproben ist weder das Alter noch das Geschlecht der Patienten bekannt.



**Abbildung 12.3** Der Datensatz Astro. (a) Anteil der Patienten (gestrichelt) und Spektren (durchgezogen), bei denen laut Referenzlabel die höchste Klassenzugehörigkeit mindestens den auf der Ordinate aufgetragenen Wert erreicht. (b) Alter der Patienten bei OP. (c) Lagerdauer der Proben bei  $-80^{\circ}\text{C}$ . (d) Geschlechterverteilung. Von den neun Kontrollproben ist weder das Alter noch das Geschlecht der Patienten bekannt. Das Geschlecht des Patienten ist nur bekannt, wenn außer den histologischen Befunden auch der Patientenbrief der Neurochirurgie vorlag. Für die Diagramme (b) – (d) wurden für die einzelnen Klassen alle Patienten berücksichtigt, die mindestens ein Spektrum mit einer Klassenzugehörigkeit von mehr als 5% zu der jeweiligen Klasse haben.



**Tabelle 12.2** Zusammensetzung des Datensatzes Lymph

Klasse	Weiche Referenz		davon mit harter Referenz	
	Patienten	Spektren	Patienten	Spektren
Normal	37	16217	16	7478
davon Kontrolle	9	4902	9	4902
Astrozytom	66	29101	55	20990
Lymphom	8	2550	5	1922
Insgesamt	86	39129	74	30390

### 12.2.2 Differentialdiagnostik von Lymphomen und Astrozytomen

Die Unterscheidung zwischen Lymphomen und Astrozytomen ist histologisch oft nur schwer zu treffen, da die Diagnostik an sehr kleinen Biopsieproben (der Durchmesser der Biopsienadel beträgt etwa 1 mm) mit entsprechend großer Probennahmeunsicherheit erfolgt. Hierbei ist die diagnostische Fragestellung kein Grading, sondern die Unterscheidung, ob es sich um ein Astrozytom oder um ein Lymphom handelt.

Dieses Potential wird anhand des Datensatzes Lymph mit den Klassen

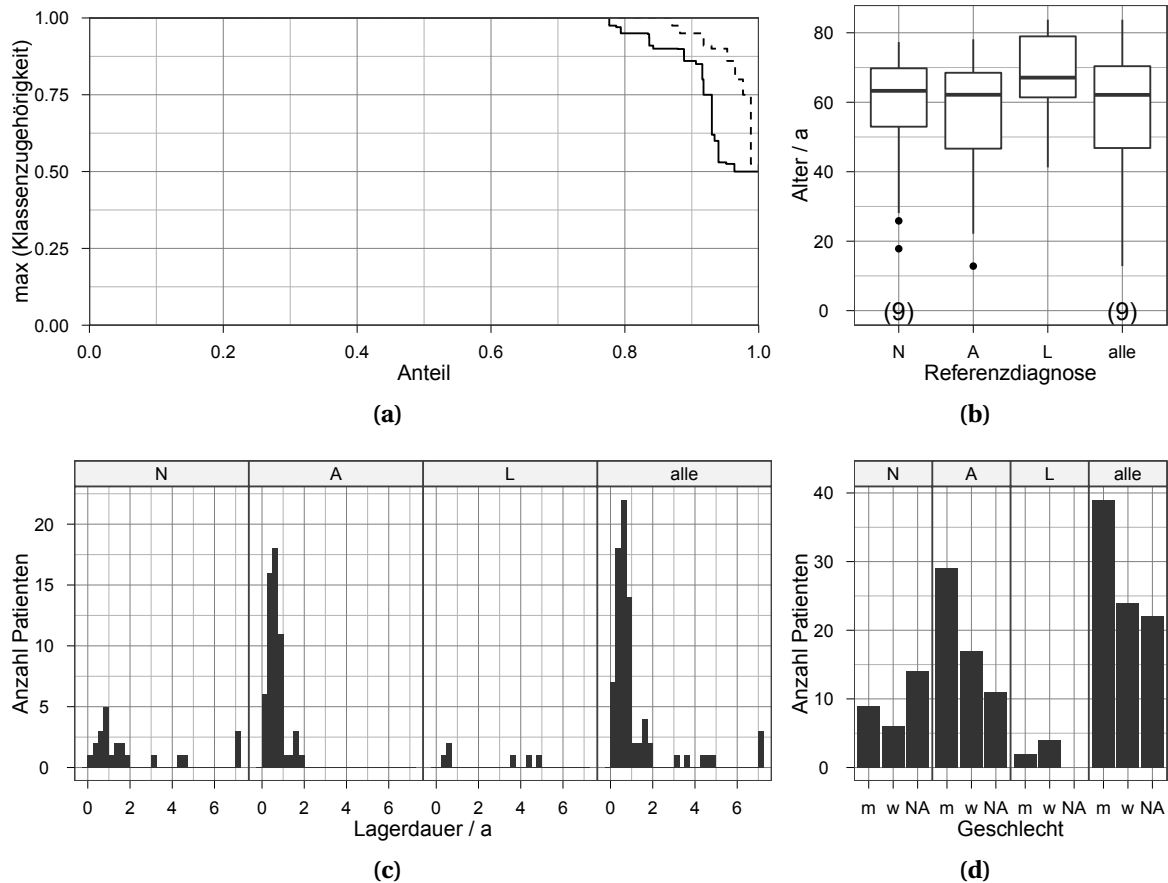
- N:** nicht-tumoröses Gewebe (wie beim Datensatz Astro),
- A:** Gewebe aus Astrozytomen (inklusive nekrotischem Gewebe) und
- L:** Gewebe von Lymphomen (inklusive nekrotischem Gewebe)

untersucht. Der Datensatz enthält zusätzlich zu den Proben des Astro-Datensatzes 8 Lymphome. Die drei Klassen werden in den Farben grün für normal, orange für Astrozytome und lila für Lymphome dargestellt.

Tabelle 12.2 und Abbildung 12.4 fassen den Datensatz zusammen. Von insgesamt 39 000 Spektren von 86 Patienten haben 78 % eine harte Referenz. Eine Lymphomprobe enthält eindeutig normales Gewebe, weitere zwei Proben tragen anteilig normales Gewebe bei. Die Astrozytome dominieren den Datensatz: sie stellen  $\frac{3}{4}$  der Patienten und 70 % der harten Spektren. Demgegenüber haben nur 5 Lymphomproben Spektren, die eindeutig Lymphomgewebe sind. Mit weniger als 2000 eindeutigen Lymphomspektren und auch nur 600 weiteren Spektren mit Lymphom-Anteilen liegt ihr Anteil bei unter 7 %. Damit ist dieser Datensatz noch deutlich unausgewogener zusammengesetzt als der Astro-Datensatz.

## 12.3 Klassifikationsmodelle

Wie in Kapitel 5.2.1 dargelegt, benötigt eine datengesteuerte Optimierung von Klassifikationsmodellen immer ausreichend genaue Schätzungen der erreichten Modellqualität in einer inneren Validierungsschleife, um einen Unterschied zwischen den Modellen überhaupt feststellen zu können. Aufgrund der beschränkten Patientenzahl kann in der vorliegenden Arbeit nicht davon ausgegangen werden, dass das auch nur annähernd der Fall ist (vgl. [CB3]). Deshalb wurden in der vorliegenden Arbeit alle Entscheidungen zur Modellierung aufgrund von spektroskopischen und chemometrischen Kenntnissen ge-



**Abbildung 12.4** Der Datensatz Lymph. (a) Anteil der Patienten (gestrichelt) und Spektren (durchgezogen), bei denen laut Referenzlabel die höchste Klassenzugehörigkeit mindestens den auf der Ordinate aufgetragenen Wert erreicht. (b) Alter der Patienten bei OP. (c) Lagerdauer der Proben bei  $-80^{\circ}\text{C}$ . (d) Geschlechterverteilung. Von den neun Kontrollproben ist weder das Alter noch das Geschlecht der Patienten bekannt. Das Geschlecht des Patienten ist nur bekannt, wenn außer den histologischen Befunden auch der Patientenbrief der Neurochirurgie vorlag. Für die Diagramme (b) – (d) wurden für die einzelnen Klassen alle Patienten berücksichtigt, die mindestens ein Spektrum mit einer Klassenzugehörigkeit von mehr als 5% zu der jeweiligen Klasse haben.

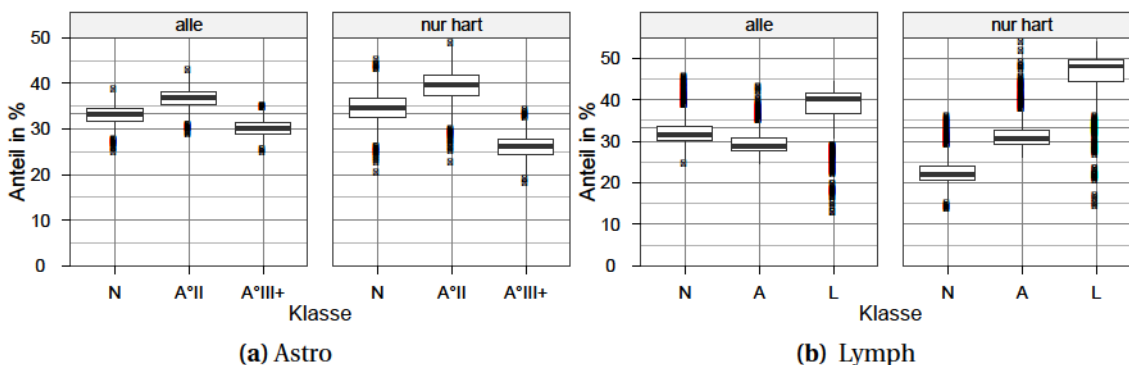
**Tabelle 12.3** Wichtung der Spektren für die Modellbildung. Die Chance, dass das jeweilige Spektrum in den Trainingsdatensatz gezogen wird, errechnet sich aus der Summe der mit den Wichtungsfaktoren gewichteten anteiligen Klassenzugehörigkeiten.

	Astro			Lymph		
	N	A °II	A °III+	N	A	L
Anteil an allen Spektren des Datensatzes in %	35	24	31	34	60	5
Wichtungsfaktor	1,5	2	1	2	1	16

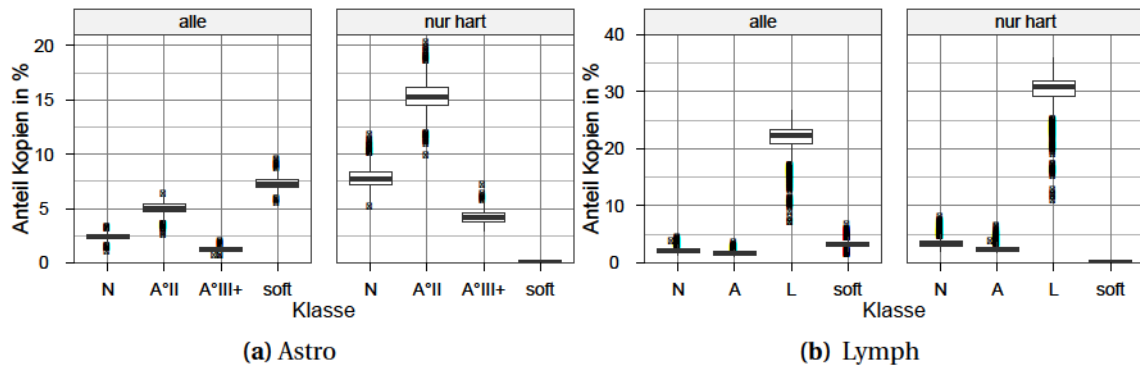
treffen. Diese Entscheidungen sind also unabhängig vom konkret vorliegenden Datensatz, wohl aber unter Einbeziehung von Erfahrungen mit ähnlichen Datensätzen ([CB6–CB8, CB12, 16]), gefallen.

Um einen Modell-Selektions-Bias aufgrund der Auswahl des scheinbar besten Modells [215, 216] zu vermeiden, wird hier nicht nur ein optimales Modell vorgestellt. Wie in den Kapiteln 4.8.5 und 5.2.1 dargelegt wurde, können die beobachteten Unterschiede in der Modellqualität allenfalls als Hinweis gewertet werden. Die verfügbaren Patientenzahlen erlauben keine datengesteuerte Modellselektion. Anhang B listet die Validierungsergebnisse für *alle* berechneten Modelle auf.

**Zusammenstellung der Trainingsdaten:** Aus allen Spektren der Trainingspatienten wurden für jedes Klassifikationsmodell 10 000 Spektren mit Zurücklegen gezogen. Die Anzahl der Patienten ist in den einzelnen Klassen extrem unterschiedlich (siehe Tab. 12.1 und 12.2). Daher wurde die Chance, dass das Spektrum zur Modellbildung benutzt wird, wie folgt angepasst: Zunächst werden die anteiligen Klassenzugehörigkeiten mit dem in Tabelle 12.3 angegebenen Wichtungsfaktoren multipliziert. Diese drei Gewichte ergeben zusammengezählt die Chance für das Ziehen des jeweiligen Spektrums. Die Wichtungsfaktoren wurden so gewählt, dass die jeweils unterrepräsentierten Klassen A °II und L jeweils etwas mehr als ein Drittel der Trainingspektren stellen (Abb. 12.5). Da die Spektren *mit Zurücklegen* gezogen werden, enthalten die Datensätze unterschiedlich große Anteile an kopierten Spektren (Abb. 12.6).



**Abbildung 12.5** Zusammensetzung der Trainingsdaten der  $8 \cdot 126 = 1008$  Modelle. (a) Astrozytomgrading, (b) Unterscheidung zwischen Astrozytomen und Lymphomen. Die weichen Spektren wurden entsprechend ihrer Klassenzugehörigkeit gewichtet.



**Abbildung 12.6** Kopien in den Trainingspektren der  $8 \cdot 126 = 1008$  Modelle. (a) Astrozytomgrading, (b) Unterscheidung zwischen Astrozytomen und Lymphomen. Die in den Datensätzen unterrepräsentierten Klassen werden bewusst höher gewichtet (Abb. 12.5). Dadurch enthalten die jeweiligen Trainingsdatensätze hohe Anteile an Spektralkopien. soft bezeichnet keine eigene Klasse, sondern die Spektren mit weicher Referenzinformation.

**Logistische Regression (LR) :** Die Funktion `multinom` aus dem R-Paket `nnet` [150] berechnet die logistischen Regressionsmodelle. Sie verarbeitet auch anteilige Klassenzugehörigkeiten in den Referenzdaten. Die Modelle sind als neuronales Netz ohne Zwischenschicht implementiert. Als Verbindungsfunktion (engl. *link function*) zwischen den Eingabewerten (Spektrum) und den Ausgabewerten (Klassenzugehörigkeitswahrscheinlichkeiten) wird die logistische Funktion verwendet.

Um die logistischen Regressionsmodelle besser mit den LDA-Modellen vergleichen zu können, wurde jeweils ein zusätzliches logistisches Regressions-Modell trainiert, das dieselben 10000 harten Trainingspektren wie das LDA-Modell verwendet. Dieses Modell heißt LR-crisp, das logistische Regressions-Modell, das auch weiche Spektren zum Training benutzt, LR-soft.

**Lineare Diskriminanzanalyse (LDA):** Die LDA ist eine Klassifikationsmethode, mit der bereits viel Erfahrung im Bereich der schwingungsspektroskopischen Hirntumor-Diagnostik vorhanden ist (Kap. 4.4.1). Sie dient daher als etablierte Vergleichsmethode.

Die LDA Modelle wurden mit der Funktion `lda` aus dem R-Paket `MASS` [150] berechnet. Die LDA benötigt harte Trainingsdaten. Spektren mit anteiligen Klassenzugehörigkeiten wurden aus dem Trainingsdatensatz für die LDA ausgeschlossen, da die LDA keine anteiligen Klassenzugehörigkeiten in den Referenzdaten verarbeiten kann. Diese Modelle heißen im folgenden LDA.

**Klassifikation im Spektralbereich von 2800 bis 3050  $\text{cm}^{-1}$ :** Eine Raman-Diagnostik, die nur den Spektralbereich der C – H-Streckschwingungen von 2800 bis 3050  $\text{cm}^{-1}$  nutzt, könnte mit ungefilterten Sonden auskommen. Diese sind wesentlich einfacher und kostengünstiger herzustellen und zu miniaturisieren als gefilterte Sonden (Kap. 3.1).

Die logistischen Regressions-Modelle LR-highwn nutzen nur den Spektralbereich zwischen 2800 und 3050  $\text{cm}^{-1}$ . Sie wurden mit denselben Spektren wie die LR-soft-Modelle trainiert. Die Differentialdiagnostik der Lymphome ist im Gegensatz zum Grading der Astrozytome letztlich ein hartes Klassifikationsproblem. Daher wurden für diese Anwen-

dung auch LDA-Modelle gebildet, die ausschließlich den Spektralbereich der C – H-Streck-schwingungen nutzen (LDA-highwn).

**PLS zur Datenreduktion:** Da insgesamt wenige Patienten einer großen Anzahl an Variaten gegenüberstehen, ist eine weitere Datenreduktion auf nochmals weniger Variate vor der LDA oder LR möglicherweise vorteilhaft (Kap. 4.3.1). Deshalb wurden jeweils mit genau denselben Trainingspektren zu den LR-soft- und LDA-Modellen noch logistische Regressionsmodelle und LDA-Modelle im Koordinatensystem der ersten 25 latenten Variablen einer PLS gebildet, das entspricht einer Reduktion der Variaten um einen Faktor 10. Zur Berechnung der PLS wurde die Funktion `pls` aus dem Paket `pls` [266] genutzt. Die Anzahl an latenten Variablen habe ich aufgrund meiner Erfahrung mit biospektroskopischen Datensätzen festgelegt. Diese Modelle heißen PLS-LDA und PLS-LR.

### 12.3.1 Aggregierte Modelle

Wie in Kapitel 4.6 (S. 38) beschrieben, können instabile Vorhersagen durch Mehrfachbestimmungen stabilisiert werden. Das können entweder Vorhersagen mehrerer Modelle für dasselbe Spektrum oder Vorhersagen für mehrere Spektren sein. Da die beiden Aggregationsmodi auf unterschiedliche Varianzquellen (Instabilität der Modelle und Rauschen auf den Spektren) zielen, können sie auch kombiniert werden.

**Modellensembles:** Aus den Surrogatmodellen der  $126 \times$  iterierten Kreuzvalidierung (siehe unten) wurden Ensemble-Vorhersagen gewonnen. Als Aggregationsfunktion wurde der Mittelwert von jeweils 9 der vorhergesagten Klassenzugehörigkeitswahrscheinlichkeiten genutzt. So entstanden 14 Vorhersagen von Ensemble-Modellen die ausschließlich Vorhersagen für Testproben aggregieren.

**Aggregieren mehrerer Spektren einer Messung beim Lymph-Datensatz:** Während das intraoperative Gliom-Grading ein Ergebnis für jedes einzelne Spektrum benötigt, zielt die Differentialdiagnostik zwischen Astrozytomen und Lymphomen auf den Gesamttumor des Patienten. Daher können die Vorhersagen von mehreren Spektren einer Messung zusammengefasst werden.

Hierbei spielt die Erkennung von normalem Gewebe nur insofern eine Rolle, als dieses Gewebe nicht zur Unterscheidung von Astrozytomen von Lymphomen beitragen kann. Daher werden alle Spektren, die mit einem Anteil von mehr als 25 % zum normalen Gewebe zugeordnet wurden, zurückgewiesen. Sie können bei der Differentialdiagnostik zwischen Astrozytomen und Lymphomen nicht helfen. Als nächstes wurden die vorhergesagten Anteile an Astrozytom- und Lymphomgewebe normiert, so dass sie in der Summe 1 ergeben (*softmax*). Analog zum Ausschluss von Spektren, die normalem Gewebe zugeordnet werden, könnten auch Spektren zurückgewiesen (engl. *reject*) werden, bei denen die Einordnung als Lymphom oder Astrozytom nicht eindeutig genug ausfällt (vgl. [CB6]). Hier wurde auf diese Möglichkeit allerdings verzichtet, um nicht noch einen weiteren Hyperparameter einzuführen.

Die aggregierte Vorhersage wurde als Mittelwert der Klassenzugehörigkeitswahrscheinlichkeiten von 9 Tumorspektren berechnet. Als Referenzinformation dient hier die Tumorart des Patienten, die die Zielgröße der Differentialdiagnostik ist.

## 12.4 Modellvalidierung

Die Patientenzahl reichte nicht aus, um eine aussagekräftige Validierung mit einem eigens reservierten Validierungsdatensatz durchzuführen: letztlich wären dazu in jeder Klasse mehr Testpatienten notwendig gewesen, als in den kleinen Klassen (besonders Lymphome) überhaupt Proben zur Verfügung standen. Alle Modelle wurden daher einer 126 mal<sup>(d)</sup> iterierten 8-fachen Kreuzvalidierung unterzogen. Die dafür notwendige Berechnung von jeweils 1008 Modellen ist ein hochgradig einfach parallelisierbares (engl. *embarrassingly parallel*) Problem. Diese Rechnungen wurden mit Hilfe des R-Pakets `snow` [267] parallelisiert. Sensitivitäten und Spezifitäten wurden mit `softclassval` [CB3] (Kap. 8.4) berechnet.

Die Spektren eines Patienten sind *nicht* statistisch unabhängig voneinander. Im Rahmen dieser Arbeit wird daher auf der Ebene der Patienten zwischen Trainings- und Testpatienten unterschieden. Da der Datensatz `Lymph` insgesamt nur Proben von 8 Lymphompatienten enthält, wurden jedem der 8 Kreuzvalidierungssets ein Lymphompatient zugeteilt. Alle anderen Patienten wurden mittels ziehen ohne Zurücklegen ohne weitere Stratifizierung aufgeteilt.

Dabei wurden die Surrogatmodelle für die unterschiedlichen Klassifikationsmethoden mit jeweils denselben Trainingspatienten gebildet und denselben Testpatienten getestet. Aus den Spektren der Trainingspatienten wurden jeweils zwei Trainingsspektrensätze zufällig gezogen: einer nur aus den Spektren mit eindeutiger Referenzzuordnung für alle Modelle, die nicht mit weichen Referenzdaten trainiert werden können (LDA-, LR-crisp- und PLS-LDA), und ein zweiter aus allen Spektren, also mit eindeutiger oder weicher Referenzzuordnung, für die LR-soft-, LR-highwn- und PLS-LR-Modelle.

Ein Modell ist immer auch von allen Spektren abhängig, die in die Berechnung von Vorbehandlungsschritten eingeflossen sind, zum Beispiel, indem sie zur Berechnung des Mittelwertspektrums oder einer PLS-Projektion herangezogen wurden. *Resampling*-Validierungsmethoden berechnen sehr viele Surrogatmodelle. Dabei ist es vorteilhaft, soviel wie möglich Vorbehandlungsschritte *vor* der Berechnung der Surrogatmodelle für alle Spektren gemeinsam durchzuführen, so dass sie nur einmal gerechnet werden müssen. Auf diese Weise vorgezogen werden können aber nur die Vorbehandlungsschritte vor dem *ersten* datengesteuerten Rechenschritt [CB10]. Da hier weder mittelwertzentriert noch varianzskaliert wurde, betrifft dies nur die Modellbildung. Die PLS-Projektion wird dabei als Teil der eigentlichen Modellbildung aufgefasst: die PLS-Projektion wurde immer für die jeweiligen Trainingsdaten neu berechnet, und dann auf die entsprechenden Testdaten angewendet.

Alle aggregierten Vorhersagen fassen nur Vorhersagen von Surrogatmodellen zusammen, bei denen der jeweilige Patient Testpatient war. Daher sind auch die aggregierten Vorhersagen statistisch unabhängig von den Trainingsdaten.

---

<sup>(d)</sup> Aus 126 unabhängigen Vorhersagen für jede Probe lassen sich 14 Ensemble-Vorhersagen berechnen, die jeweils 9 Vorhersagen aggregieren.

# **Teil IV**

## **Ergebnisse und Diskussion**

## 13 Einfluss des Einbettmediums

Einbettmedien erleichtern das Schneiden von Proben mit dem Mikrotom. Sie stabilisieren die Probe mechanisch. Ein Objekt mit gleichmäßiger Härte und günstiger Form entsteht. Zusätzlich sinkt die Gefahr von Gefrierartefakten.

Das Ziel dieser Arbeit sind chemometrische Modelle zur Beschreibung und Klassifikation nativer Gewebe. Bei der Verwendung von Einbettmedien muss also der Nachweis erbracht werden, dass sie nicht ins Gewebe eindringen oder die Spektren nicht verändern. Shim und Wilson [131] empfehlen die Verwendung von PEG, das vor der Messung abgewaschen werden kann. Bei den Tumorproben besteht aber die Gefahr, dass dabei Teile der Probe suspendieren oder aus beschädigten Zellen wasserlösliche Bestandteile ausgewaschen werden. Daher ist das Abwaschen von Gefriermedium ungünstig.

Abbildung 13.1 zeigt dieselbe Probe (Schweinehirn) vor und nach Überschichten mit Gefriermedium<sup>(a)</sup>. Das Gefriermedium besteht aus PEG (Referenzspektrum vgl. [268]) und Wasser.

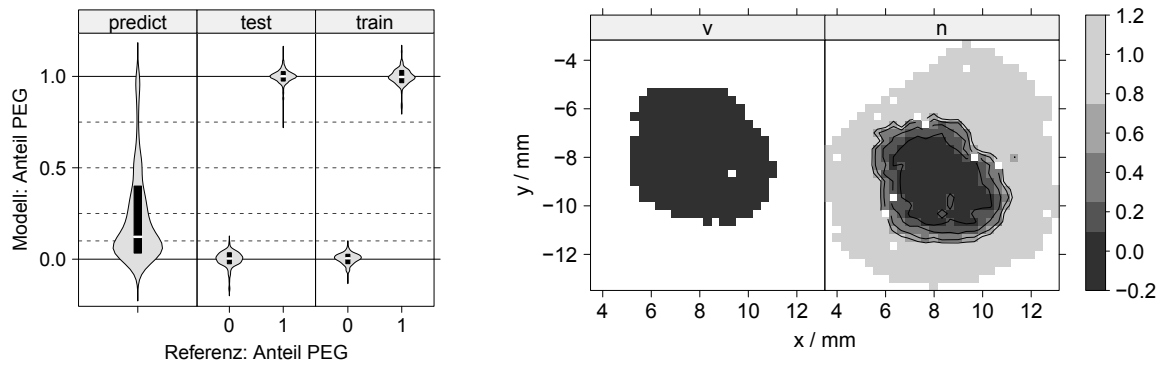
Ein PLS-Regressions-Modell mit drei latenten Variablen modelliert den Gehalt an PEG. Als Trainingspektren wurden aus der Messung vor dem Überschichten und dem nur PEG enthaltenden Rand der zweiten Messung jeweils 200 Spektren zufällig ausgewählt. Abbildung 13.1a zeigt die Verteilung der Ergebnisse des Modells (von rechts nach links) für die Trainingspektren (train), die restlichen Spektren der Probe vor Überschichten und des übrigen PEG-Randes (test) und für die mit PEG überschichtete Probe (predict).

Die Zuordnung durch das PLS-Modell ergibt in den Randbereichen der Probe deutliche Anteile von Polyethylenglykol (Abb. 13.1b; n), während vor dem Überschichten richtig kein PEG gefunden wird (v).  $\frac{1}{2} - 3\frac{1}{2}$ h nach Aufbringen des Gefriermediums ist die 10 % Linie des PEG-Gehalts mindestens 1 mm im Inneren des Gewebes. PEG ist kein geeignetes Einbettmedium für die hier beschriebenen Proben und Experimente.

---

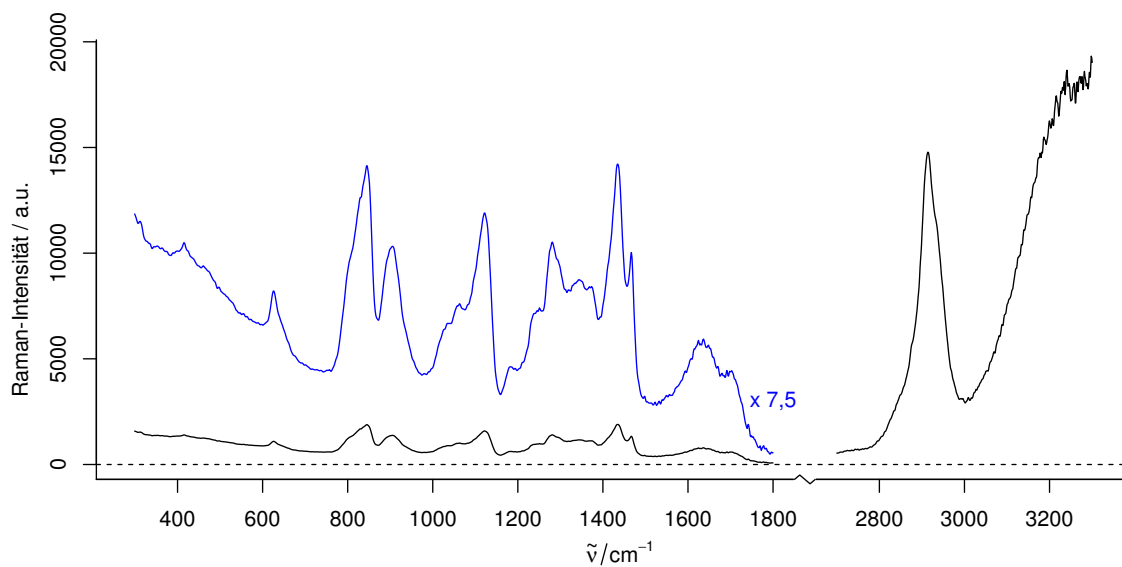
<sup>(a)</sup> Jung Tissue Freezing Medium, Leica Microsystems Nussloch GmbH, D-69226 Nussloch





(a) Verteilung der Modellvorhersagen

(b) Die Probe vor (v) und nach (n) Überschichten mit PEG



(c) Spektrum von Jung Tissue Freezing Medium (PEG in Wasser)

**Abbildung 13.1** Eindringen von PEG in die Probe. (a) Violinen-Diagramm der Modellvorhersagen für die Probe nach Überschichten mit PEG (predict), sowie die Test- und Trainingsdaten des Modells. Gestrichelt die den Konturlinien (10, 15, 50 und 75 %) in (b) entsprechenden Werte. (b) Das PLS-Regressions-Modell findet erhöhte PEG-Konzentrationen bis gut 1 mm in die Probe hinein. (c) Raman-Spektrum des verwendeten Gefriermediums.

## 14 Detaillierte Referenzdiagnosen

Insgesamt wurden detaillierte Referenzdiagnosen zu 104 Schnitten von Tumorproben erhalten. Tabelle 14.1 stellt die Diagnose des Patienten der detaillierten Histologie des Schnittes (vgl. Kap. 11) gegenüber.

Die dreieckige Form der Tabelle ist eine Folge der Unterschiede zwischen dem normalen Grading eines Tumors und der hier betrachteten detaillierten Diagnose und der Heterogenität der Astrozytome. Es können morphologisch niedriggradigere Gewebe, als es der Diagnose des Patienten entspricht (Kap. 11), auftreten. Die detaillierte Diagnose steht insofern in vollkommener Übereinstimmung mit der Diagnose der Patienten, als bei keinem der über 100 begutachteten Schnitte Anzeichen für eine mögliche höhere Malignität des Tumors gefunden wurden.

Auf der hervorgehobenen „Diagonalen“ der Tabelle werden Schnitte gezählt, deren dominierende Morphologie dem Tumorgrad des Patienten entspricht. Unterhalb dieser Diagonalen finden sich alle Proben, die größtenteils aus niedriggradigerem Gewebe bestehen, als der Diagnose des Patienten entspricht. Insgesamt bestehen gut 10 % der Tumorschnitte überwiegend aus normalem Gewebe. Hinzu kommen über 20 % der Schnitte mit entweder einer Mischung aus normalem Gewebe mit Tumorzellen oder bei denen die Neuropathologin nicht sicher sagen konnte, ob überhaupt Tumorzellen vorliegen. Das gilt auch für die Lymphom-Proben. Dieser hohe Anteil an normalem Gewebe unterstreicht die Notwendigkeit einer intraoperativen Diagnostik.

**Tabelle 14.1** Gegenüberstellung der Diagnose des Tumors des Patienten mit der detaillierten histologischen Diagnose des Gewebstückes. Der Tumorgrad des Patienten (Zeilen) wurde den Diagnosebriefen entnommen und anhand der üblichen Routine bestimmt (Schnellschnitte, ausführlichere Begutachtung, gegebenenfalls Rücksprache mit Tumorreferenzzentrum Bonn). Er wird verglichen mit dem den Schnitt dominierenden Gewebe (Spalten).

Diagnose (Patienten)	Dominierendes Gewebe des Schnittes ( $\geq 50\%$ der Schnittfläche)										
	Summe	Normal	Unsicher <sup>a</sup>	Rand <sup>b</sup>	Astrozytome					Nekrose	Lymphom
				II	II-III	III	III-IV	IV			
Kontrolle	9	<b>9</b>									
Astro. °II	7	3	1	1	<b>2</b>						
Astro. °III	20	2	2	3	4	4	<b>5</b>				
Glio.	60 <sup>c</sup>	4	3	9	5	4	12	6	<b>6</b>	<b>8</b>	
Lymphom	8	2		1						<b>5</b>	
Summe	104	20	6	14	11	8	17	6	6	8	5

<sup>a</sup> Pathologe war sich nicht sicher, ob das Gewebe Tumorgewebe war

<sup>b</sup> Tumorzellen infiltrieren normales Gewebe

<sup>c</sup> 3 Proben zu heterogen, kein dominierendes Gewebe

Nur  $\frac{1}{10}$  der Glioblastom-Proben weist überwiegend Tumorgewebe vierten Grades auf, und nur knapp  $\frac{1}{4}$  besteht überwiegend aus Gewebe, das diese Diagnose etablieren würde<sup>(a)</sup>. Letzterer Anteil ist bei allen Gliomgraden etwa gleich. Klassifikationsmodelle können auch dann erfolgreich trainiert werden, wenn ein gewisser Anteil der Trainingsdaten falsche Referenzlabel hat. Allerdings ist bei einem Anteil von nur 25 % korrekten Referenzinformationen keine erfolgreiche Modellbildung zu erwarten. Ein Klassifikationsmodell, das mit der Diagnose des Patienten trainiert und validiert wird, kann sehr wohl gute Vorhersagequalität erzielen. Allerdings wird es ein Grading des Patienten vornehmen, und nicht die Gewebe innerhalb der einzelnen Tumore unterscheiden. Die operationsbegleitende Diagnostik zielt aber auf die Unterscheidung innerhalb des Tumors. Der Tumorgrad des Patienten ist also *kein* geeigneter Surrogatmarker für die Morphologie von Gliomgeweben und eine detaillierte histologische Begutachtung der untersuchten Proben ist daher für das Erstellen der hier diskutierten intraoperativen Klassifikationsmodelle unbedingt notwendig.

Etwa  $\frac{1}{4}$  bis  $\frac{1}{3}$  der Schnitte sind von Gewebe dominiert, das in der histologischen Begutachtung zwischen den WHO-Graden eingeordnet wird, gemischte Zellpopulationen enthält (Rand, II–III, III–IV) oder bei dem unsicher ist, ob Tumorgewebe vorliegt. Diese Schnitte bestehen also *hauptsächlich* aus Gewebe, das nur mit anteiligen Klassenzugehörigkeiten adäquat beschrieben werden kann. Damit wird die Notwendigkeit von weichen Klassifikationsmethoden für eine repräsentative Beschreibung der Hirntumorproben auch aus dieser Perspektive deutlich.

**Unsicherheit der histologischen Ergebnisse:** Aus analytisch-chemischer Sicht ist das normale Grading des Tumors eine schwierige Fragestellung (engl. *ill-posed*). Das Auffinden des am stärksten entdifferenzierten Gewebes ist eine sehr störanfällige Aufgabe<sup>(b)</sup>, zumal trotz der bekannt heterogenen Gewebe oft nur wenige und kleine Proben gewonnen werden können. Daher ist mit einer großen Unsicherheit aufgrund der Probennahme zu rechnen. Probennahmefehler sind ein bekanntes Problem in der analytischen Chemie. Die Probennahmefehler überragen den Analysenfehler oft um ein bis zwei Größenordnungen [208, 269].

Die ungünstige Problemstellung beim Tumorgrading spiegelt sich in der teilweise hohen Variation der Befunde verschiedener Histologen wieder, zum Beispiel bei der Unterscheidung zwischen Astrozytomen mit oder ohne oligodendroglialer Komponente [270]. Allerdings differieren die Befunde von spezialisierten Neuropathologen weitaus weniger als die von chirurgischen Pathologen (engl. *surgical pathologists*) [271].

Bei der Beurteilung der Variation zwischen den Aussagen mehrerer Pathologen muss allerdings beachtet werden, dass zwei Bearbeitungsvorschriften beim Grading die Vari-

---

<sup>(a)</sup> Da die Diagnose für den Tumorgrad des Patienten auch aus kleinen Anteilen höhergradigem Gewebes gestellt wird, heißt das aber ausdrücklich *nicht*, dass der Tumorgrad nicht korrekt erkannt würde (siehe auch [71]). Belastbare Aussagen, wie häufig aus einer solchen Probe der Tumorgrad des Patienten histologisch korrekt bestimmt werden kann, lassen sich aus den vorliegenden Daten schon aufgrund der Probennahme-Prozedur nicht ableiten: während der Operation muss zunächst sichergestellt sein, dass die für das Tumorgrading notwendigen Proben genommen und in die Histologie gegeben werden. Erst danach können möglicherweise weitere Proben für Forschungszwecke gewonnen werden.

<sup>(b)</sup> Das Maximum sammelt sozusagen Rauschen.

anzunsicherheit erhöhen. Beide Effekte betreffen auch Studien über die Varianz der Befunde verschiedener Pathologen, werden bei der hier verwendeten detaillierten Diagnose jedoch weitgehend vermieden.

Zum Einen ist das die schon erwähnte Vorschrift, dass der Tumorgrad durch das differenzierteste Gewebe, und sei es noch so wenig, definiert ist. Die detaillierte histologische Diagnose ist von diesem Problem weitaus weniger betroffen. Im Gegensatz zur Diagnose des Patienten wird nicht von dem vorliegenden Gewebe auf einen ganzen Tumor extrapoliert, und der gesuchte Tumorgrad bezieht sich nur auf das vorliegende Gewebe.

Die zweite Quelle für Varianz ist das Erzwingen eines harten Befundes. Zu den in dieser Arbeit untersuchten Proben standen anonymisierte Patientenbriefe und Histologiebefunde für den Patienten zur Verfügung. Diese enthalten die Diagnose für den jeweiligen Patienten, gegebenenfalls auch weitere Ergebnisse aus dem Tumorreferenzzentrum in Bonn. Auffällig ist, dass die ausführliche Beurteilung des Tumorreferenzzentrums und des örtlichen Neuropathologen oft deutlich geringere Unterschiede aufweist als die harten Diagnosen (WHO-Grad des Tumors). In der Praxis lauten viele Befunde beispielsweise „mit Anzeichen von“ oder „Grad II bis III“ (bzw. „Grad III bis IV“). Als ein Beispiel sei der histologische Befund für Probe 1204 zitiert:

Beurteilung: Der Befund entspricht Anteilen eines malignen Astrozytoms (mindestens WHO-Grad III) mit Verdacht auf Übergang in ein Glioblastom.

Kommentar: Es handelt sich hier um einen malignen, überwiegend astrozytär differenzierten Tumor. Es finden sich einzelne Endothelproliferate, jedoch keine sicher abgrenzbaren Nekrosen, so dass hier aus Gründen der Nomenklatur zunächst von einem anaplastischen Astrozytom (WHO-Grad III) ausgegangen wird. Es besteht jedoch der Verdacht auf Übergang in ein Glioblastom, insbesondere bei entsprechenden radiologischen Befunden. Anhaltspunkte für eine Karzinometastase bestehen nicht.

Werden solche weichen Befunde in harte Klassen übersetzt, so kann die Varianz der Befunde sehr stark steigen. Dieser Effekt betrifft vermutlich hauptsächlich mittlere Tumorgrade.

Auch in der von Kendall *et al.* [187] veröffentlichten Studie waren die Pathologen am häufigsten bei den mittleren Diagnose-Klassen uneinig [CB2]. Insgesamt wurde  $\frac{1}{3}$  der Proben aufgrund uneindeutiger Befunde aus der Auswertung ausgeschlossen. Auch beim Astro-Datensatz dieser Arbeit liegt für  $\frac{1}{3}$  der Patienten kein einziges Spektrum mit harter Referenz vor.

Diese Beobachtungen deuten darauf hin, dass das Härten der Diagnosen tatsächlich zusätzliche Varianz verursachen könnte und weiche Klassifikationsmodelle möglicherweise einen erheblichen Anteil an ungenauen Referenzdaten vermeiden.

## 15 Einfluss der Probenlagerung

Insgesamt standen für die Auswertung 9 Proben von Kontroll-Patienten zur Verfügung. Die Proben 46, 53 und 60 gehören zu den ältesten Proben der Sammlung. Sie waren zum Zeitpunkt der Präparation bereits 7 Jahre bei  $-80\text{ °C}$  gelagert. Diese Proben waren bei der Voruntersuchung der Spektren auffällig vom Großteil der anderen Proben und von allen anderen Kontrollproben separiert (Kap. 12.1, S. 101 und Abb. 12.2, S. 102).

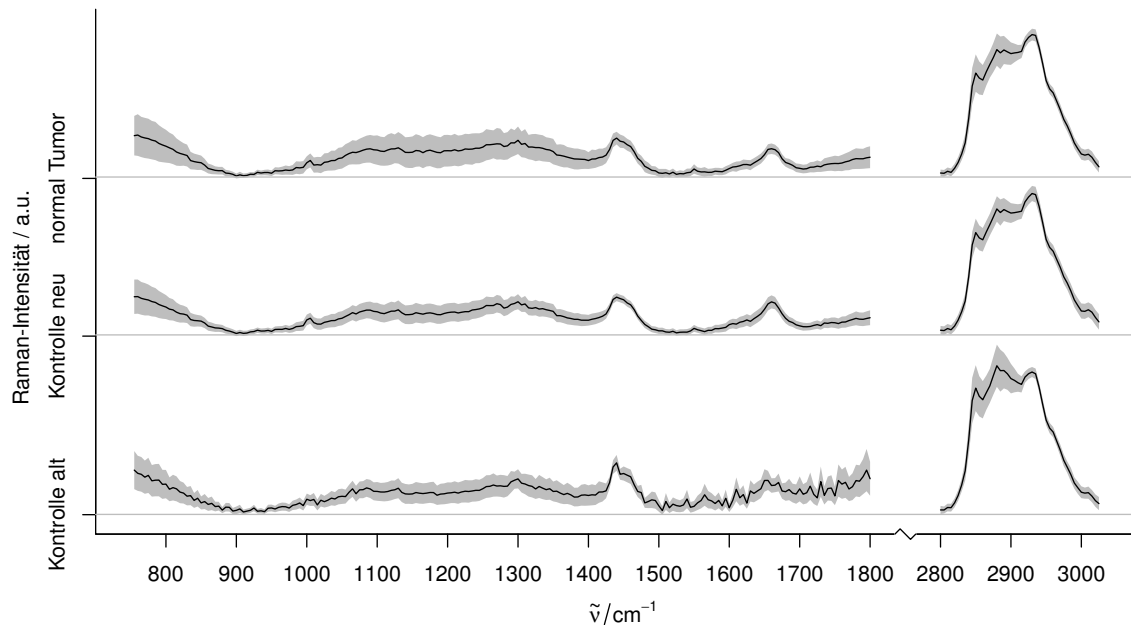
Die Spektren der älteren Kontrollproben (Abb. 15.1 unten) unterscheiden sich klar von den Spektren der neuen Kontrollproben (Abb. 15.1 Mitte). Die spektrale Signatur im Bereich der C – H-Valenzschwingungen ist verändert. Das Signal-Rausch-Verhältnis im Fingerprint-Bereich ist oberhalb von  $1450\text{ cm}^{-1}$  niedriger und auch die Ringschwingung der aromatischen Aminosäuren bei  $1005\text{ cm}^{-1}$  ist schwächer. Auch um die  $1650\text{ cm}^{-1}$  wird eine deutlich geringere Intensität beobachtet. Dort liegt außer der Amid-I-Bande bei auch die Deformationsschwingung von Wasser. So dass dieses wesentlich schwächere Signal dort möglicherweise auf eine Gefriertrocknung der Proben hinweist. Da die Messungen in der feuchten Kammer erfolgten ist das jedoch unwahrscheinlich. Gefriertrocknung könnte hier nur dann beobachtet werden, wenn das gefriergetrocknete Gewebe über die Dauer der Messungen von jeweils etwa 6 Stunden das verlorene Wasser nicht wieder aufnimmt. Die Unterschiede in den  $\nu_{\text{CH}}$ -Banden zwischen  $2800$  und  $2925\text{ cm}^{-1}$  und im Bereich der CH-Deformationsschwingung bei  $1460\text{ cm}^{-1}$  liegt vermutlich an einem höheren Anteil weißer Substanz in diesen Proben. Die neueren Proben stammen weitgehend aus der Hirnrinde (graue Substanz). Die neuen Kontrollproben enthalten ebenfalls weiße Substanz, aber in geringeren Anteilen als die alten Kontrollproben. Weiße Substanz enthält weniger Wasser als graue Substanz [272], wodurch ein möglicher Unterschied im Wassergehalt der Proben erklärt werden kann. In der LDA-Projektion in Abbildung 12.2a auf Seite 102 sind die Spektren von weißer Substanz der neueren Kontrollproben allerdings unauffällig.

Die gemessenen Gliomproben waren alle weniger als 2 Jahre alt (Kap. 9.1). Die Spektren der von der Neuropathologin als normal eingestuften Gewebe dieser Proben (Abb. 15.1 oben) unterscheiden sich nur geringfügig von denen der neueren Kontrollproben.

Vier der Lymphomproben waren zum Zeitpunkt der Präparation zwischen  $3\frac{1}{2}$  und 5 Jahren gelagert. Ihre Spektren unterscheiden sich aber nicht von den Spektren der neueren Proben, die LDA zur Ausreißerkontrolle (Abb. 12.2a, S. 102) projiziert diese Proben zu den neueren Proben.

Die alten Kontrollproben sind diejenigen Proben der gesamten Sammlung, von denen am häufigsten Schnitte präpariert wurden. Jede dieser Proben ist also mehrmals auf mindestens  $-20\text{ °C}$  erwärmt worden, teilweise wurden sie sogar aufgetaut. Demgegenüber wurden die Lymphomproben erstmals für die vorliegende Arbeit präpariert. Abgesehen von einer Havarie der Tiefkühltruhe, bei der alle Proben auf etwa  $-15\text{ °C}$  erwärmt wurden, wurden sie also im Unterschied zu den Kontrollproben *durchgehend* bei  $-80\text{ °C}$  gelagert.

Drei mögliche Erklärungen für die Unterschiede zwischen den Spektren der alten Kon-



**Abbildung 15.1** Einfluss der Probenlagerung auf die Spektren der Kontrollproben. Die Spektren (Mittelwert  $\pm$  eine Standardabweichung) der alten Kontrollproben nach etwa 7 Jahren Lagerdauer und vielfacher Präparation unterscheiden sich deutlich von denen der neuen Kontrollproben (unter einem Jahr gelagert und entsprechend seltener präpariert) und des normalen Gewebes in Tumorproben.

trollproben und allen anderen Spektren sind damit:

- Lagerzeiten von 7 Jahren führen zu drastischen Veränderungen, die nach 5 Jahren noch nicht auftreten.
- Normales Gewebe verändert sich anders (schneller) als Lymphomgewebe oder
- wesentlich anders als die reine Lagerdauer bei  $-80^\circ\text{C}$  ist, wie oft (und wie lange) die Proben höheren Temperaturen ausgesetzt sind.

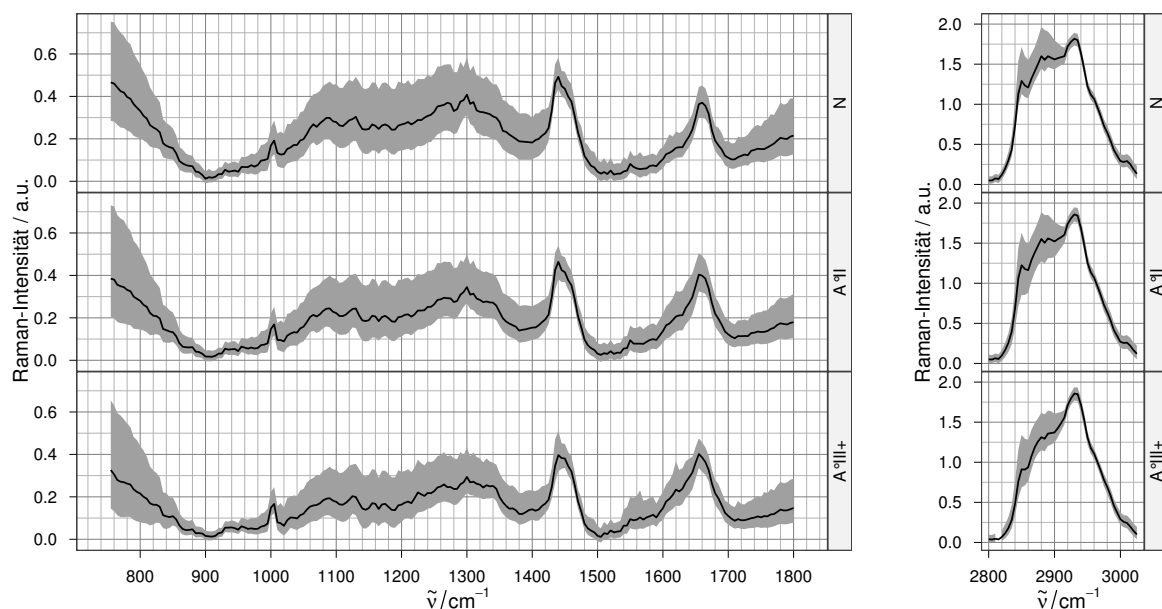
Bei Raumtemperatur zeigen die Hirnproben bereits nach wenigen Stunden deutliche Veränderungen: nach 8 Stunden in der feuchten Kammer riechen sie und nach 12 Stunden verfärben sie sich grünlich. Schweinehirn ist im Kühlschrank nur wenige Tage haltbar. Bei  $-20^\circ\text{C}$  sind die Hirnproben einerseits hart genug, dass lipidreiches Gewebe (weiße Substanz) nicht schmiert. Andererseits sind sie noch nicht so hart und spröde, dass sie beim Schneiden im Gefriermikrotom brechen, wie es bei tieferen Temperaturen der Fall ist. Die mechanischen Eigenschaften der Proben ändern sich im Temperaturbereich zwischen etwa  $-25$  und  $-15^\circ\text{C}$  deutlich. Entsprechend muss auch damit gerechnet werden, dass bei diesen Temperaturen Veränderungen sehr viel schneller ablaufen als bei  $-80^\circ\text{C}$ .

Damit ist die plausibelste Erklärung für die Unterschiede in den Spektren, dass die Temperaturzyklen eine wesentlich größere Rolle spielen als die Lagerdauer bei  $-80^\circ\text{C}$ . Für zukünftige Untersuchungen wäre es daher günstig, die Proben von vornherein in kleinere Teile zu portionieren, die jeweils für eine Präparation ausreichen.

## 16 Tumorgrading der Astrozytome

Die Raman-Spektren der drei Klassen sind in Abbildung 16.1 dargestellt. Mittelwert und Perzentile wurden aus allen Spektren gewichtet nach dem Anteil der jeweiligen Klasse berechnet. Das Signal-Rausch-Verhältnis der Spektren ist im Fingerabdruck-Bereich unterhalb von  $1800\text{ cm}^{-1}$  wesentlich schlechter (im Mittel etwa 6) als im Bereich der CH-Valenzschwingungsbanden zwischen  $2800$  und  $3050\text{ cm}^{-1}$  (im Mittel 29). Grund ist der hohe Untergrund der Rohspektren. Lediglich die Bande bei ca.  $1650\text{ cm}^{-1}$  (Wasser und Amid I), die CH-Deformationsschwingungen bei  $1440\text{ cm}^{-1}$  (Lipide) und die aromatische Ring-Atmungsschwingung bei  $1005\text{ cm}^{-1}$  (besonders Phenylalanin, also Proteine; diese Bande wird im folgenden als Phenylalanin-Bande bezeichnet) sind erkennbar. Im Gegensatz dazu sind die CH-Valenzschwingungsbanden sehr stark ausgeprägt. Die C-H-Streckschwingungsbanden von  $\text{CH}_2$  ( $\nu_s$  bei  $2850\text{ cm}^{-1}$  und  $\nu_{as}$  bei  $2930\text{ cm}^{-1}$ ) und von  $\text{CH}_3$  ( $\nu_s$  bei  $2880\text{ cm}^{-1}$  und  $\nu_{as}$  bei  $2960\text{ cm}^{-1}$ ) sind sehr intensiv. Normale weiße Substanz enthält sehr viel Lipide und weist daher besonders intensive  $\text{CH}_2$ -Signale auf (84. Perzentilspektrum des normalen Gewebes). Demgegenüber haben die Tumore deutlich verringerte Lipidgehalte (siehe auch [100]), die sich auch in den Spektren zeigen. Bei den hochgradigen Astrozytomgeweben zeichnen sich außerdem im 84. Perzentilspektrum spektrale Beiträge von Blut ab, besonders die Hämoglobin-Banden um  $1600\text{ cm}^{-1}$ .

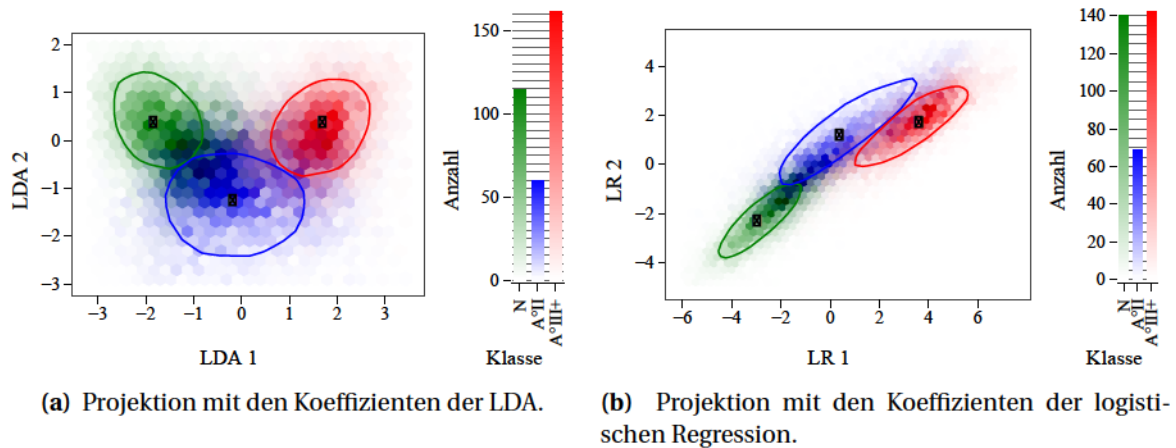
Die Normierung der Spektren auf die mittlere Intensität im Spektralbereich von  $2900$  bis  $3025\text{ cm}^{-1}$  (Kap. 12.1) normiert die Spektren praktisch auf ihren Proteingehalt: Subtrahiert man von den normierten Spektren das Mittelwertspektrum von normaler grauer



**Abbildung 16.1** Die Spekten des „Astro“ Datensatzes: gewichtetes Median-, 16. und 84. Perzentilspektrum.

Substanz, so verschwinden sowohl die Amid-I-Bande ( $1655\text{ cm}^{-1}$ ) als auch die Phenylalanin-Bande fast vollständig (vgl. [CB1, CB2]).

## 16.1 Deskriptive Modelle



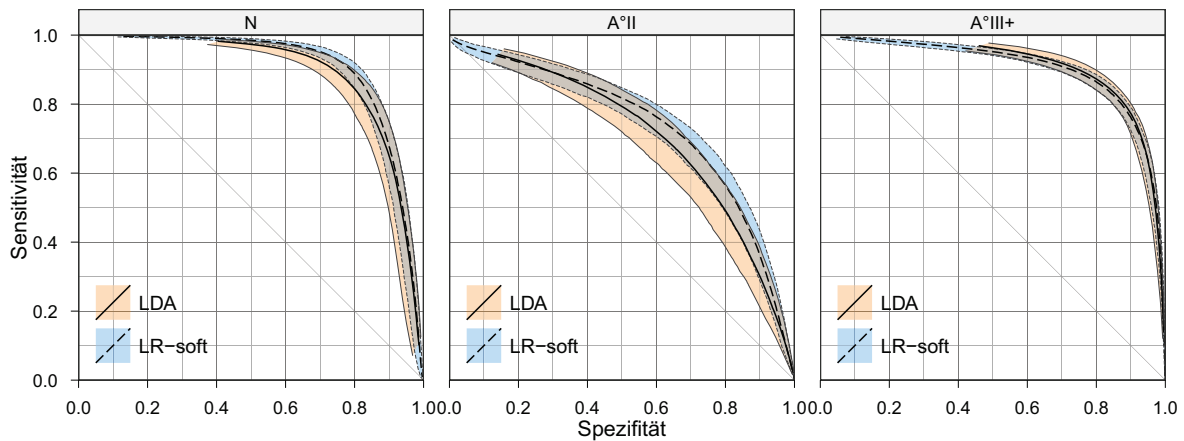
**Abbildung 16.2** Projektion der Astro-Daten als 2d-Histogramme. Die durchgezogenen Konturen enthalten jeweils 50 % der harten Spektren der jeweiligen Klasse, die weißen Punkte markieren den zweidimensionalen Median der harten Spektren.

Sowohl LDA als auch logistische Regressionsmodelle berechnen für  $n_g$  Klassen  $n_g - 1$  Linearkombinationen aus den einzelnen Intensitäten jedes Spektrums. Das entspricht bei Drei-Klassen-Modellen einer Projektion der Spektren in zwei Dimensionen. Abbildung 16.2 zeigt die entsprechenden zweidimensionalen Histogramme [273] aller Spektren mit einem LDA-Modell aus allen harten Spektren (a) und einer logistischen Regression des gesamten Datensatzes (b). Die Konturlinie ist eine Verallgemeinerung des Interquartilsabstandes: sie umschließt 50 % der jeweiligen Spektren. Dabei werden für die einzelnen Klassen nur Spektren betrachtet, die mindestens zur Hälfte zur jeweiligen Klasse gehören. Die Anzahl der Spektren berücksichtigt die anteiligen Klassenzugehörigkeiten.

Das zweidimensionale Quantil entsteht durch „Abschälen“ der konvexen Hülle der Punkte: von den betrachteten Punkten werden schrittweise immer diejenigen entfernt, die die konvexe Hülle bilden. Das zweidimensionale Quantil ist diejenige Kontur, die den entsprechenden Anteil der Punkte enthält. Analog ist der zweidimensionale Median der Median der Koordinaten der zuletzt übrigen (innersten) konvexen Hülle.

Folgen die einzelnen Klassen multivariaten Normalverteilungen mit derselben Kovarianzmatrix (Kap. 4.4.1), so werden die Klassen in Kreise mit demselben Durchmesser projiziert. Die Histogramme zeigen deutlich, dass der Datensatz sehr viele weiche Spektren aus Tumorrandbereichen (Mischung aus A°II und N) enthält. Aber auch ohne die weichen Spektren überlappen die Klassen.





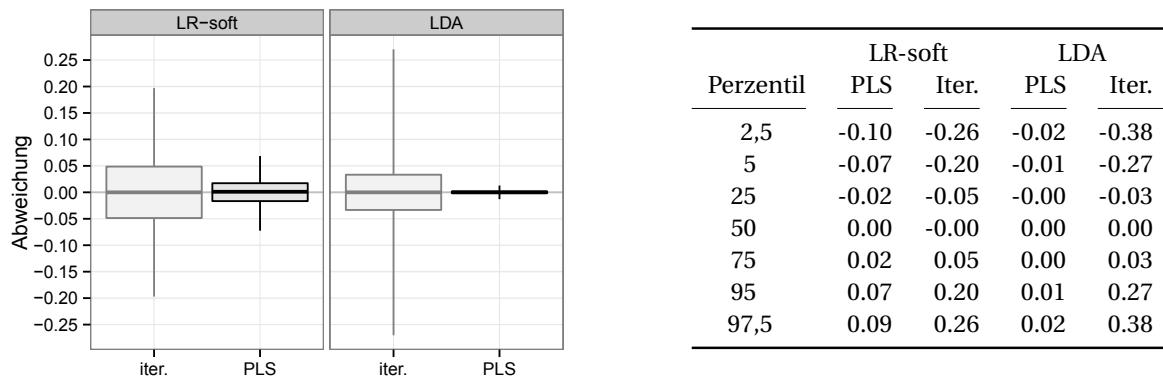
**Abbildung 16.3** Spezifitäts-Sensitivitäts-Diagramm. Vorhersagequalität der LDA (durchgezogen) und LR-soft Modelle (gestrichelt). Eingetragen sind jeweils Median, 5. und 95. Perzentil der über die 126 Iterationen der Kreuzvalidierung beobachteten Leistung. Sensitivität und Spezifität beziehen sich auf die Vorhersage der einzelnen Spektren.

## 16.2 Prädiktive Modelle

Die nur mit harten Spektren trainierten LDA-Modelle sind der weichen logistischen Regression LR-soft bei der Vorhersage der harten Spektren leicht unterlegen. Abbildung 16.3 zeigt die Spezifitäts-Sensitivitäts-Diagramme dieser beiden Modelle im Vergleich. Die LR-soft-Modelle sind für alle Klassen etwas besser als die LDA-Modelle, aber die beobachtete Bandbreite der Spezifitäts-Sensitivitäts-Kurven über die 126 Iterationen der Kreuzvalidierung überlappen.

Diese Verbesserung ist auf die zusätzlichen Trainingspektren des LR-soft-Modells zurückzuführen. Das logistische Regressionsmodell LR-soft nutzt dieselben Trainingspektren wie das LDA-Modell und ist jenem leicht unterlegen (Anhang B.1.2, S. 169). Parametrische Modelle sind nicht-parametrischen Modellen überlegen, wenn die Annahmen über die Verteilung stimmen [145, 159]. Sie erreichen dann mit weniger Proben dieselbe Qualität. Die Grundannahmen der LDA bezüglich der multivariaten Normalverteilung sind zwar nicht erfüllt, aber die LDA ist im Allgemeinen recht robust, solange die Verteilung nicht zu endlastig (engl. *heavily tailed*) ist. Das könnte hier der Vorteil für die LDA gegenüber der logistischen Regression mit denselben Trainingsdaten sein.

Das LR-soft-Modell hat gegenüber den Modellen LDA und LR-soft zwei Vorteile. Zum Einen enthält der Trainingsdatensatz sehr viele Spektren aus der Nähe der Klassengrenze (Abb. 16.2). Zum Anderen umfassen die Trainingsdaten auch  $1\frac{1}{2}$  mal so viele Patienten. Allerdings sind auch Spektren, bei denen die Referenzdiagnose nicht eindeutig auf das Messraster übertragen werden konnte, weich gelabelt. Deshalb haben die weichen Spektren vermutlich einen höheren Anteil an ungenauen Referenzinformationen, was sich nachteilig auf das LR-soft Modell auswirken sollte. Insgesamt überwiegt jedoch der Vorteil durch die größere Trainingsdatenbasis, die Modelle sagen auch die harten Spektren mit eindeutiger Referenz besser vorher als die LDA-Modelle.



**Abbildung 16.4** Abweichung der Vorhersagen der „PLS-LR“- und „PLS-LDA“ von den Modellen ohne PLS-Dimensionsreduktion (Box: 25. bis 75. Perzentil und Median, Whisker: 5. bis 95. Perzentil). Zum Vergleich Abweichungen zwischen den Vorhersagen von zufällig ausgewählten Iterationen. Die Unterschiede in den Vorhersagen mit und ohne PLS als Datenvorbehandlung sind sehr viel kleiner als die Unterschiede zwischen den einzelnen Iterationen der Kreuzvalidierung. Damit ist eine Vorbehandlung mit PLS unnötig.

### 16.2.1 PLS-LDA- und PLS-LR-Modelle

Die PLS-LDA-Modelle sind von den LDA-Modellen ohne PLS-Datenreduktion praktisch ununterscheidbar (Abb. 17.7): 90 % der Vorhersagen der PLS-LR-Modelle differieren nicht mehr als  $\pm 0,07$  von entsprechenden Vorhersagen der LR-soft-Modelle. Zum Vergleich: vergleicht man die Vorhersagen der einzelnen Iterationen der Kreuzvalidierung für jeweils dasselbe Spektrum, so liegen 90 % in einem Bereich von  $\pm 0,2$ . Die Wurzel aus der mittleren quadrierten Abweichung zwischen verschiedenen Iterationen ist 2,7mal so groß wie die Wurzel aus der mittleren quadrierten Abweichung zwischen PLS-LDA und LR.

Die PLS-LDA- und die LDA-Modelle sind sich noch ähnlicher: 90 % der Vorhersagen der PLS-LDA-Modelle liegen innerhalb von  $\pm 0,014$  um die entsprechenden Vorhersagen der LDA-Modelle. Allerdings zeigen die LDA-Modelle eine größere Schwankungsbreite zwischen den Iterationen:  $\pm 0,27$ .

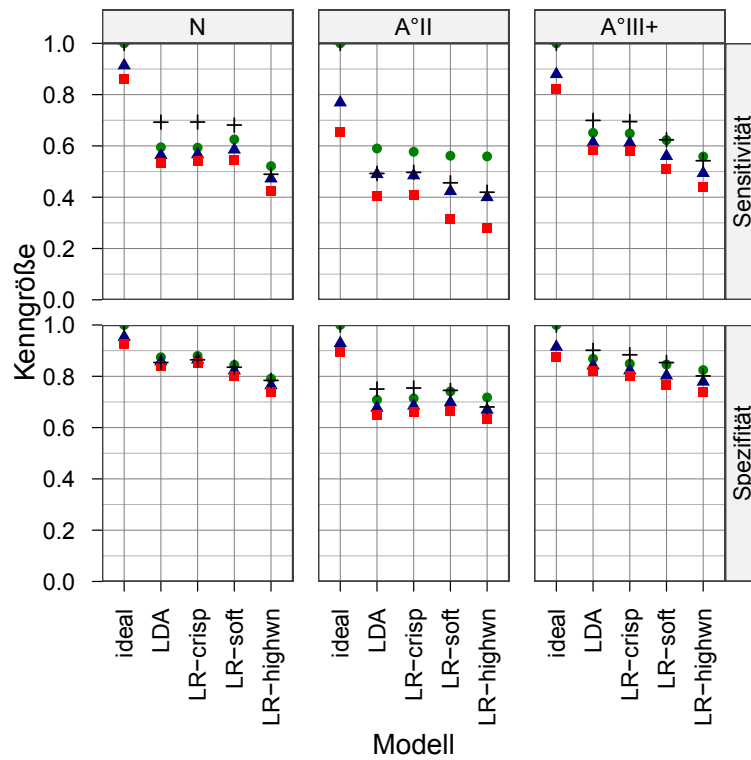
Damit ist die Datenreduktion auf 25 latente Variablen mittels PLS als Vorbehandlung für den Astro-Datensatz in dieser Arbeit nicht notwendig.

### 16.2.2 Weiche Kenngrößen

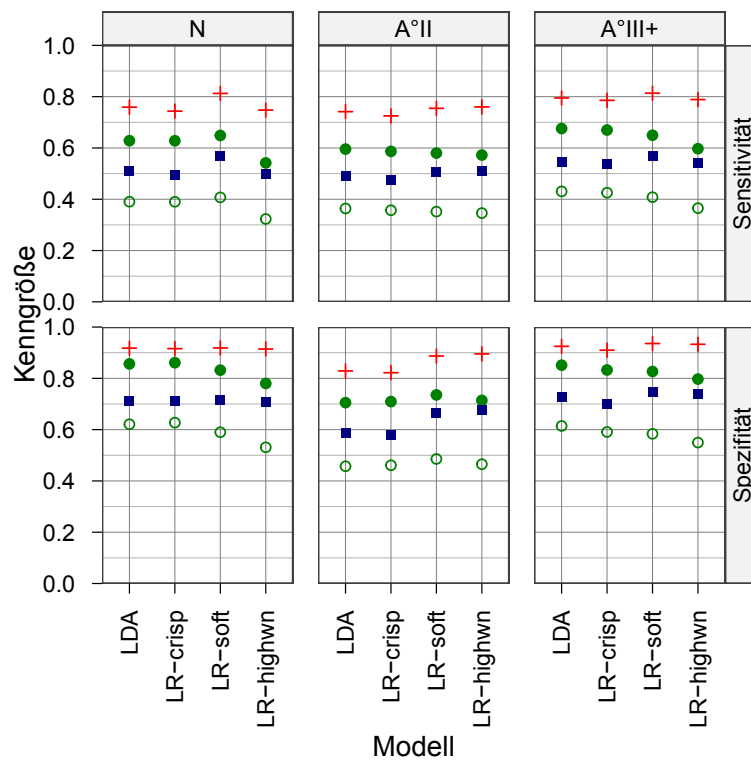
Die weichen Kenngrößen (Abb. 16.5) reagieren nicht nur empfindlicher auf Abweichungen der Vorhersage, sie erlauben auch die Beurteilung der Modellqualität bezüglich der Zuordnung der Spektren mit uneindeutiger Referenz (weiche Spektren).

Die UND-basierten Sensitivitäten und Spezifitäten der LR-soft-Modelle bezüglich der harten Spektren sind etwas schlechter als die der beiden harten Modelle (Abb. 16.5a). Allerdings erreichen sie für weiche Spektren sowohl höhere Sensitivitäten als auch höhere Spezifitäten.

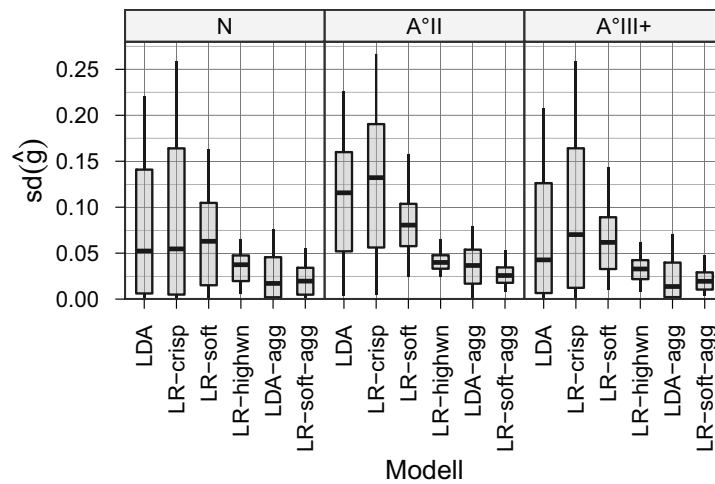
Die MSE- (engl. *mean squared error*) basierten Kennwerte (Abb. 16.5b) der LR-soft-Modelle fallen hingegen besser aus. Wie in Kapitel 8.2.2 (S. 81) dargelegt, erlaubt der Vergleich zwischen MAE (engl. *mean absolute error*) und RMSE (engl. *root mean squared error*) Aussagen über die Verteilung der Abweichungen des Modells von der Referenz.



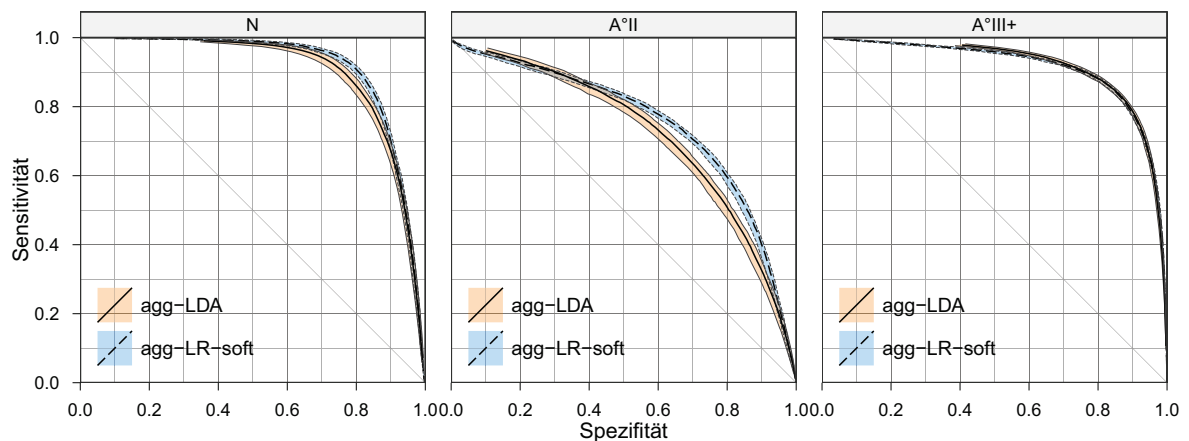
(a) UND-Operatoren.

(b)  $\Delta$ -Operatoren.

**Abbildung 16.5** Weiche Sensitivität und Spezifität für das Astrozytom-Grading, berechnet (a) mit den UND-Operatoren: ■ starkes UND, ▲ Produkt-UND, ● schwaches UND und + die UND-Operatoren nur auf die Spektren mit eindeutiger Referenzzuordnung angewendet. Ideal gibt die bestmöglichen Werte an, die erhalten werden, wenn die Vorhersage genau der Referenz entspricht. (b) Weiche Sensitivität und Spezifität berechnet mit den  $\Delta$ -Operatoren: ● 1 - wMAE, ■ 1 - wRMSE, ○ 1 - wRMSE (untere Grenze für 1 - wRMSE), und + 1 - wMSE. Gezeigt ist jeweils der Median über die 126 Iterationen der Kreuzvalidierung.

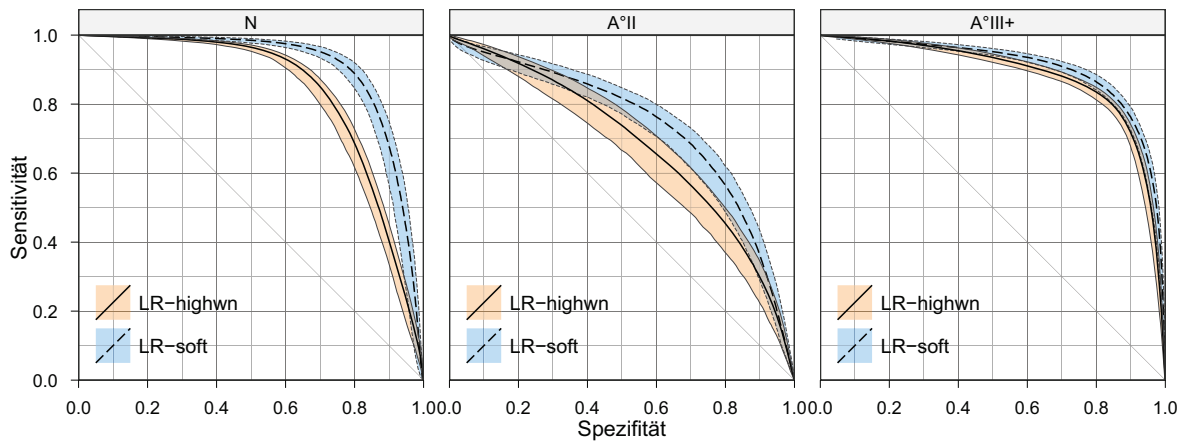


**Abbildung 16.6** Standardabweichung der Vorhersagen über die 126 Iterationen der Kreuzvalidierung für die drei Klassen beim Astrozytom-Grading. Die Boxen zeigen Median, erstes und drittes Quartil, die Whisker reichen zum 5. und 95. Perzentil bezüglich aller Spektren.



**Abbildung 16.7** Spezifitäts-Sensitivitäts-Diagramm. Vorhersagequalität der LR-soft- (gestrichelt) und LDA Ensemble-Modelle (durchgezogen). Die Vorhersagen von jeweils 9 Surrogatmodellen wurden gemittelt.

Die RMSE-Werte der LR-soft-Modelle sind nur wenig schlechter als ihre MAE-basierten Pendanten. Bei den harten Modellen liegen die RMSE-Werte dagegen in der Regel etwa auf halber Strecke zwischen dem MAE und der daraus folgenden Obergrenze des RMSE. Das bedeutet, dass die LR-soft-Modelle dieselbe absolute Abweichung auf mehr Spektren verteilen, während die LDA und LR-crisp-Modelle bei weniger Spektren größere Vorhersagefehler machen. Die LR-soft-Modelle modellieren die Übergänge zwischen den Klassen also tatsächlich weicher als die ausschließlich mit harten Trainingspektren gebildeten Modelle. Im Hinblick auf die operationsbegleitende Diagnostik sind viele kleine Abweichungen besser als seltenere, aber grobe Fehler.



**Abbildung 16.8** Spezifitäts-Sensitivitäts-Diagramm: Vorhersagequalität der LR-highwn- und LR-soft-Modelle. Die LR-highwn-Modelle erkennen alle drei Klassen schlechter als die LR-soft-Modelle. Am ausgeprägtesten ist der Einbruch in der Modellqualität bei der Sensitivität für normales Gewebe: Bei gleicher Spezifität macht LR-highwn bis zu dreimal so viele Fehler wie LR-soft. Auch bezüglich der niedriggradigen Tumore ist die Vorhersagequalität der LR-highwn sichtlich schlechter als die der LR-soft-Modelle

### 16.2.3 Modellstabilität und Aggregation

Der Interquartilsabstand der beobachteten Kennlinien (Abb. 16.3) für die 126 Iterationen ist gering, wobei die Vorhersagen der einzelnen Modelle durchaus Varianz aufweisen (Abb. 16.6). Die Vorhersagen der beiden harten Modelle LDA und LR-crisp sind besonders bei weichen Spektren und Spektren niedriggradigen Gewebes instabil. Die Vorhersagen bei normalem und hochgradigem Tumorgewebe sind meist deutlich stabiler. Das liegt daran, dass die Abwesenheit dieser beiden Gewebearten beim jeweils anderen Gewebe sehr gut erkannt wird. Die stabilsten Vorhersagen der LR-soft Modelle sind etwas instabiler als die stabilsten Vorhersagen der LDA (unteres Quartil). Andererseits sind die weniger stabilen Vorhersagen der LR-soft-Modelle (oberes Quartil und 95. Perzentil) stabiler als die entsprechenden Perzentile der Vorhersagen der LDA.

Die Variation der Vorhersagen hat aber auf die Kennlinien nur sehr geringe Auswirkungen, sie fällt größtenteils beim Umwandeln der vorhergesagten Klassenzugehörigkeitswahrscheinlichkeiten in harte Zuordnungen weg. Entsprechend bewirkt eine Aggregation der Vorhersagen der 126 Modelle im Median nur geringe Verbesserungen (Abb. 16.7). Auch die weichen Kenngrößen verändern sich praktisch nicht.

### 16.2.4 Klassifikation mit den CH-Valenzschwingungsbanden

Der Spektralbereich zwischen  $2800$  und  $3050\text{ cm}^{-1}$  ist leider nicht zum Grading der Astrozytome geeignet. Obwohl ein sehr gutes Signal-Rausch-Verhältnis (im Mittel 29) vorliegt, ist die Erkennung aller drei Klassen schlechter als bei den Modellen, die zusätzlich den Fingerabdruck-Bereich zwischen  $750$  bis  $1800\text{ cm}^{-1}$  nutzen (Abb. 16.8). Besonders die wesentlich schlechtere Sensitivität für normales Gewebe ist ein Ausschlusskriterium für dieses Modell, da die Schonung von normalem Gewebe Vorrang vor dem vollständigen Entfernen des Tumors hat. Die LR-highwn-Modelle sind noch weicher als die LR-soft-Modelle und sagen generell mittlere Anteile an allen drei Klassen voraus.

Insbesondere die schlechte Sensitivität für normales Gewebe ist unerwartet, da die Spektren der drei Klassen im betrachteten Spektralbereich deutliche Unterschiede aufweisen (Abb. 16.1 auf Seite 119). Allerdings überlappen sie auch in diesem Spektralbereich. Die anderen Modelle können zusätzlich kleine Veränderungen über weite Abschnitte des Fingerabdruck-Bereichs nutzen, die den LR-highwn-Modellen nicht zur Verfügung stehen.

Die Überlappung der Spektren der verschiedenen Gewebe kann durch die Normierung (Kap. 12.1, S. 100) verstärkt sein. Andererseits kann bei diesem Datensatz nicht auf die Normierung verzichtet werden. Sowohl wechselnde Abstände zwischen Probenoberfläche und Sonde als auch Schwankungen im Wassergehalt der Proben (durch Gefrier-trocknung einerseits und Kondenswasser, das beim Auftauen auf der Probe kondensiert andererseits) führen zu Schwankungen im Gesamtsignal, die nichts mit dem Gewebe zu tun haben. Diese Einflüsse werden durch die Normierung reduziert.

### 16.2.5 Spektroskopische Interpretation der prädiktiven LDA-Modelle

LDA-Modelle projizieren die Spektren zunächst in einen niedrigdimensionalen Raum, der durch die  $(n_g - 1)$  Diskriminanzfunktionen aufgespannt wird (LDA Scores, Abb. 16.2a). Alle 1008 im Rahmen der  $126 \times$  iterierten 8-fachen Kreuzvalidierung erstellten LDA-Modelle ähneln dem in Kapitel 16.1 (S. 120) vorgestellten LDA-Modell auf der Basis aller Patienten: Die Klassenmittelwerte bilden in der LDA-Projektion ungefähr rechtwinklige Dreiecke mit der Hypothense entlang der 1. Diskriminanzfunktion. Entsprechend bildet die 1. Diskriminanzfunktion ca.  $70 \pm 2,7\%$  der Varianz zwischen den Klassen ab (die über die Klassen gemittelte Varianz innerhalb der Klassen ist durch die LDA-Projektion in alle Richtungen 1).

Wären die Veränderungen in den Spektren, die zur Einordnung als hochgradiges Gliomgewebe herangezogen werden, ausschließlich größere Veränderungen derselben Art, die zur Unterscheidung von Gewebe mit A°II Morphologie von normalem Gewebe herangezogen werden, so würden die Klassen entlang einer Linie angeordnet. Das ist jedoch nicht der Fall. Die Verbindungslinien N—A°II und A°II—A°III+ bilden praktisch einen rechten Winkel ( $86,4 \pm 3^\circ$ ). Dieses Muster kann zustande kommen, wenn kontinuierliche Veränderungen mit der Malignität (1. Diskriminanzfunktion) auftreten und niedriggradige Gewebe zusätzlich durch weitere Veränderungen gekennzeichnet sind, die bei den hochgradigen Geweben nicht vorliegen. Eine andere Erklärungsmöglichkeit sind Veränderungen, die die Tumorgewebe von normalem Gewebe unterscheiden und davon unabhängige zusätzliche Veränderungen bei hochgradigen Tumoren. Zwischen diesen Möglichkeiten kann auf der Basis der hier diskutierten Modelle allerdings nicht unterschieden werden.

Die zugeordneten Klassenzugehörigkeiten und *a-posteriori*-Wahrscheinlichkeiten ändern sich nicht, wenn diese Projektion gedreht oder gespiegelt wird. Für die spektroskopische Interpretation in diesem Abschnitt wurden die Modelle so gedreht, dass der Mittelpunkt der niedriggradigen Astrozytome rechts (positive Scores auf der 1. rotierten Achse) des Mittelpunkts der Spektren von normalem Gewebe lag. Der Mittelpunkt der hochgradigen Astrozytome lag bei allen Modellen oberhalb (positive Scores auf der 2. rotierten Achse) dieser beiden Klassen, so dass keine Spiegelung notwendig war. Dadurch werden die Koeffizienten einerseits interpretierbar in Hinsicht auf die biochemischen

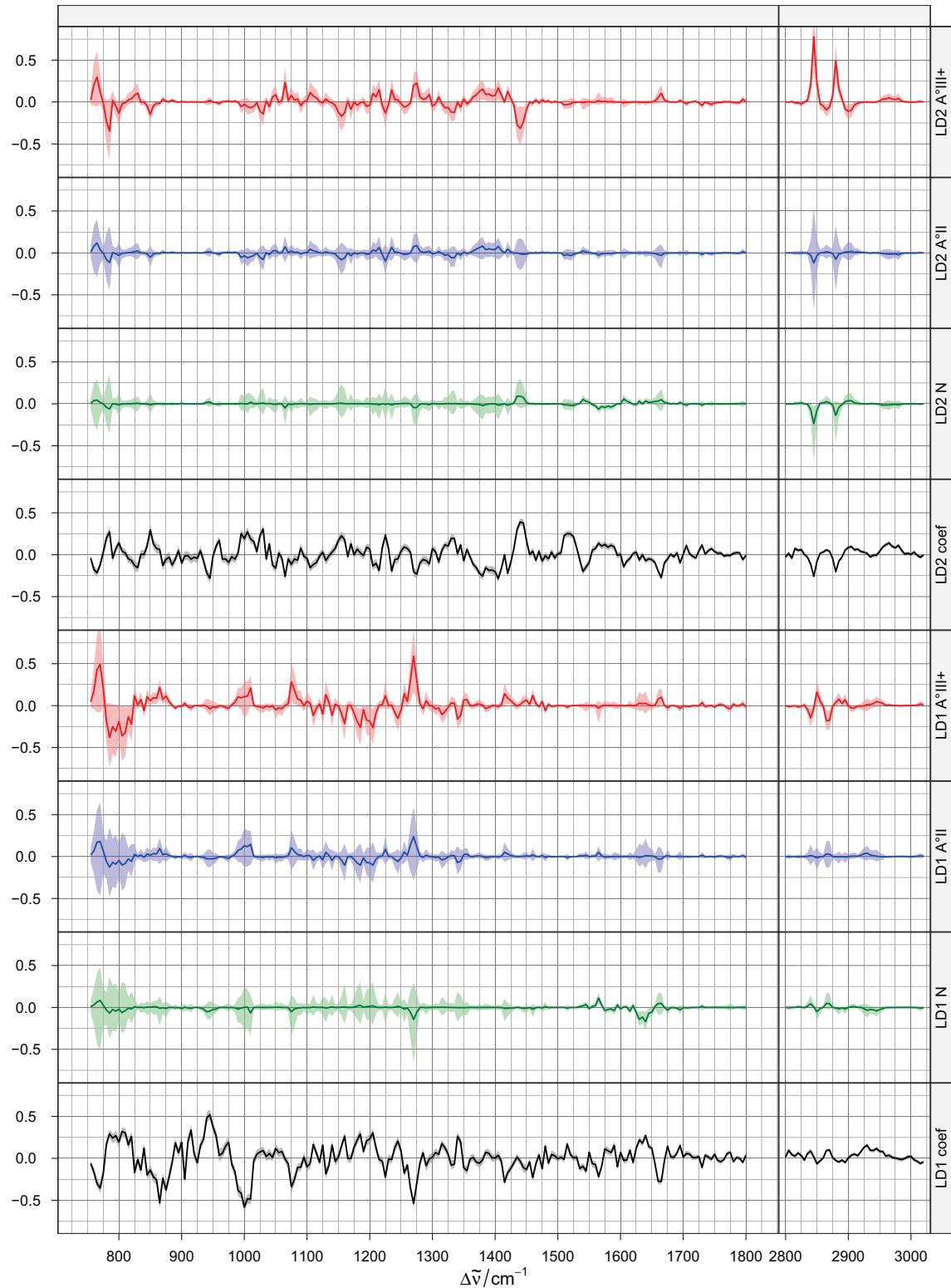
Veränderungen in den Tumorgeweben, die zur Einordnung als Tumor bzw. als hoch- oder niedriggradiges Gewebe herangezogen werden. Andererseits werden zufällige *unwesentliche* Unterschiede zwischen den Modellen unterdrückt. Die Stabilität der Modelle kann so besser beurteilt werden, als das bei direktem Betrachten der Koeffizienten möglich ist.

Die gedrehten LDA-Modelle können analog der Loading-Vektoren einer PCA oder der latenten Variablen einer PLS spektroskopisch interpretiert werden. Die Koeffizientenvektoren *coef* in Abbildung 16.9 geben die Gewichtung an, mit der die einzelnen Messwellenlängen zur Einstufung beitragen. Der Einfluss der einzelnen Koeffizienten auf den letztlich erzielten LDA-Score hängt aber nicht nur von Vorzeichen und Betrag des Koeffizienten, sondern auch von der Intensität des Spektrums an der jeweiligen Wellenlänge ab. Abbildung 16.9 zeigt farbig die Beiträge der einzelnen Wellenlängen zu den endgültigen Scores.

Bei der spektroskopischen Interpretation ist außerdem zu beachten, dass es sich bei den Koeffizienten- und Beitragsspektren nicht um echte Differenzspektren handelt. Die Schwingungsspektren zeigen viele Banden, die sich gegenseitig bedingen. So müssen beispielsweise für Alkylketten sowohl symmetrische ( $\nu_s$ ) und antisymmetrische ( $\nu_{as}$ )  $\text{CH}_2$ -Streckschwingungen als auch Deformationsschwingungen ( $\delta$ ,  $\rho$ ,  $\tau$ ,  $\omega$ ) und diverse Gerüstschwingungen der C – C-Ketten auftreten. Alle diese Banden gehen in Differenzspektren ein. Die Koeffizientenmuster eines chemometrischen Modells können jedoch einzelne dieser Banden ausmessen, andere aber komplett ignorieren. Die Modellkoeffizienten können weiterhin den Untergrund der Spektren (anstatt von Schwingungsbanden) ausmessen. Insgesamt sind die Koeffizientenspektren und die spektralen Beiträge zur jeweiligen Einstufung daher nicht geeignet, um die biochemische Zusammensetzung der Proben zu untersuchen. Sie können aber auf bereits bekannte Veränderungen in der biochemischen Zusammensetzung hin untersucht werden.

Die in dieser Arbeit verwendeten Polynome ersten und zweiten Grades entsprechen einer sehr restriktiven Basislinienkorrektur, die den größten Teil der Untergrundsignale entfernt. Besonders unterhalb von  $900\text{ cm}^{-1}$  bleiben jedoch unkorrigierte Reste der Untergrundsignale (Abb. 16.1). Die Klassifikationsmodelle können einerseits eine implizite Basislinienkorrektur vornehmen, beispielsweise kann die Höhe eines Signals abzüglich einer linearen Untergrundkorrektur durch das Koeffizientenmuster  $-\frac{1}{2}, 1, -\frac{1}{2}$  ausgemessen werden. Für prädiktive LDA-Modelle von IR-Spektren von Gliomen hat eine lineare Basislinienkorrektur im Bereich von  $1000 - 1800\text{ cm}^{-1}$  verglichen mit Modellen ohne Basislinienkorrektur nicht zu einer Verbesserung der Vorhersagen geführt [16]. Andererseits nimmt der Untergrund der Spektren mit steigender Malignität zu. Autofluoreszenz bei kürzeren Wellenlängen wurde von Majumder *et al.* [274] im Hinblick auf mögliche Hirntumordiagnostik untersucht. Daher kann der noch vorhandene Untergrund auch direkt Informationen zur Dignität des Gewebes beitragen.

Die Koeffizienten der hier diskutierten prädiktiven LDA-Modelle sind unterhalb von  $900\text{ cm}^{-1}$  betragsmäßig groß bei wechselnden Vorzeichen. Das heißt, dass der Untergrund der Spektren von den Modellen verwendet wird. Allerdings liegen in diesem Spektralbereich auch Raman-Signale von Substanzen mit tumorbedingten Konzentrationsänderungen in derselben Richtung, die die Koeffizienten ausmessen. Sowohl für die DNS-Bande bei  $785\text{ cm}^{-1}$  als auch das Signal von Glykogen bei  $850\text{ cm}^{-1}$  wird eine erhöhte Intensität erwartet. Diese Banden werden von den LDA-Modellen als Hinweis auf hochgradiges Tumorgewebe gewertet. Auf der anderen Seite messen die Modelle eine Abnahme



**Abbildung 16.9** Die gedrehten LDA-Modelle zur Vorhersage des Astrozytom-Grades. Gezeigt sind Median und Quartile der spektralen Beiträge der einzelnen Klassen (farbig), sowie die Koeffizienten der beiden gedrehten Diskriminanzfunktionen (schwarz). Die Koeffizienten (coef) wurden um einen Faktor 10 verkleinert.



der Intensität bei  $865\text{ cm}^{-1}$ , einer typischen Raman-Bande von Phosphatidylethanolamin [97] und Beljebbar *et al.* [106] berichten eine Abnahme von Phosphatidylethanolamin für maligne Astrozytome.

Im Bereich zwischen  $2825$  und  $3025\text{ cm}^{-1}$  nehmen die Intensitäten der  $\nu_{\text{CH}}$  Streckschwingungen mit steigender Malignität ab. Die Banden bei  $2850$  ( $\nu_s\text{CH}_2$ ) und  $2885\text{ cm}^{-1}$  ( $\nu_s\text{CH}_2$  und  $\nu_s\text{CH}_3$  in Fermi-Resonanz) sind typisch für Lipide in Hirngewebe [97, 105, 275]. Die dazugehörigen antisymmetrischen Schwingungen um  $2930$  und  $2960\text{ cm}^{-1}$  überlappen mit den (CH)-Streckschwingungen anderer Komponenten (Proteine, DNS, RNS, Glykogen) [96, 244]. Ungesättigte Lipide zeigen außerdem die  $\nu=\text{CH}$  Streckschwingungen bei  $3010$ – $3015\text{ cm}^{-1}$ . Dieser Spektralbereich geht kaum in die erste rotierte Diskriminanzfunktion ein, wohl aber in die Erkennung hochgradiger Gewebe mit der zweiten Diskriminanzfunktion. Im Median stammt die Hälfte des Scores für die Einstufung als  $\text{A}^\circ\text{III}+$  aus diesem Spektralbereich. Viele Spektren von  $\text{A}^\circ\text{II}$ -Gewebe erreichen hier sogar höhere Intensitäten als normales graues Hirngewebe, während das untere Quartil den  $\text{A}^\circ\text{III}+$ -Spektren ähnlich ist (Abb. 16.1). Dasselbe gilt auch für die C–C-Streckschwingung bei  $1065\text{ cm}^{-1}$ . Die LDA-Modelle messen also die bekannte Verringerung des Lipidgehalts bei den Astrozytomen [100]. Allerdings nutzen die Koeffizienten der zweiten Diskriminanzfunktion hohe Intensitäten der  $\text{CH}_2$ -Deformationsschwingungen von Lipiden bei  $1440\text{ cm}^{-1}$  als Indikation für  $\text{A}^\circ\text{III}+$ . Diese Bande zeigt jedoch keine konsistente Erhöhung in den zentrierten Spektren der hochgradigen Tumore, so dass der Einfluss auf die Scores letztlich gering ist. Ebenfalls interessant ist, dass die Modelle die typische Torsionsschwingung der  $\text{CH}_2$ -Ketten bei  $1295\text{ cm}^{-1}$  ebenfalls nicht nutzen, obwohl die Bande in den zentrierten Spektren präsent ist.

Den größten Beitrag zur Unterscheidung zwischen normalem und Tumorgewebe liefern die Deformationen  $\delta=\text{CH}$  von ungesättigten Fettsäuren bei  $1270\text{ cm}^{-1}$ . Hochgradige Gewebe erhalten dort etwa  $\frac{3}{4}$  ihres gesamten Scores entlang der ersten Diskriminanzfunktion, niedriggradige etwa die Hälfte. Außerdem nutzen beide rotierten Diskriminanzfunktionen die Bande bei  $1665\text{ cm}^{-1}$ . Die Streckschwingungen  $\nu\text{C}=\text{C}$  ungesättigter Fettsäuren ( $1660\text{ cm}^{-1}$ ) tragen jedoch kaum zur Erkennung von Tumorgewebe bei. Aber sie werden zur Erkennung von hochgradigem Tumorgewebe herangezogen. Cholesterolester haben ebenfalls eine Bande in diesem Spektralbereich ( $1670\text{ cm}^{-1}$ ), während unverestertes Cholesterol ein Signal bei  $1675\text{ cm}^{-1}$  zeigt [97]. Dort sind die Koeffizienten beider rotierter Diskriminanzfunktionen praktisch 0. Die zentrierten Spektren weisen auf abnehmenden Gehalt an ungesättigten Lipiden und Cholesterolestern mit steigender Malignität hin. Aber auch die Spektren von normalem weißen Gewebe zeigen wenig ungesättigte Lipide oder Cholesterolester. Diese Befunde stimmen mit den Ergebnissen von Köhler *et al.* [105] und Beljebbar *et al.* [106] überein. Dort [106] wurde in einem Tiermodell in Glioblastomgewebe verglichen mit dem umgebenden normalen grauen Gewebe ein höherer Gehalt an Ölsäure, aber weniger Cholesteryloleat gefunden. Der gesamte Gehalt an Ölsäure und Oleat war im Tumor geringer als im grauen Gewebe. Die hier vorgestellten Modelle messen eher den Gesamtgehalt an Ölsäure und Oleat. Die Koeffizientenmuster unterscheiden nicht zwischen ungesättigten Fettsäuren und ihren Cholesterolestern, dazu wäre ein Vorzeichenwechsel in den Koeffizienten bei  $1665\text{ cm}^{-1}$  notwendig.

Da sowohl die  $\delta=\text{CH}$ - als auch die  $\nu\text{C}=\text{C}$ -Banden im Bereich der Amid III und Amid I Banden der Proteine ( $1225$  bis  $1300$  bzw.  $1645$ – $1675\text{ cm}^{-1}$ ) liegen, muss die Interpreta-

tion im Hinblick auf erniedrigte Gehalte an ungesättigten Lipiden gegen Veränderungen im Proteingehalt abgewogen werden. Zwei Argumente sprechen hier für einen erniedrigten Lipidgehalt. Erstens zeigen die zentrierten Spektren, dass die Banden schmal sind. Zweitens löscht die Zentrierung und Normierung der Spektren praktisch die gesamte spektrale Signatur der Proteine aus den Spektren aus. Die Normierung kann durch den Lipidgehalt der Gewebe beeinflusst sein, da sowohl  $\nu_{as}CH_2$ - als auch  $\nu_{as}CH_3$ -Streckschwingungen in dem Spektralbereich liegen, der zur Berechnung der Normierungsfaktoren genutzt wurde. Die für die Normierung berechneten Intensitäten können daher höher sein, als es ausschließlich dem Proteingehalt entspricht. Allerdings haben die untersuchten Tumore niedrigere Lipid-zu-Protein-Verhältnisse als normales graues Gewebe [CB12, 105, 106]. Nach Normierung und Zentrierung sollten in diesem Fall also positive Reste der spektralen Signatur von Proteinen bleiben. Die von den Modellen genutzten Differenzen bei  $1270$  und  $1665\text{ cm}^{-1}$  sind jedoch negativ. Tatsächlich zeigt sich bei  $1005\text{ cm}^{-1}$  (Phenylalanin) ein kleiner positiver Rest bei den hochgradigen Tumorgeweben, der zur Unterscheidung zwischen normalem und Tumorgewebe genutzt wird. Das Signal an dieser Stelle ist ungewöhnlich breit für die scharfe Phenylalanin-Bande. Das kann durch das Downsampling im Rahmen der glättenden Interpolation verursacht sein, da die verwendete Interpolationsmethode die integrale Intensität und nicht die Signalarhöhe erhält.

Koljenović *et al.* [108] fanden einen hohen Glykogegehalt (Referenzspektrum siehe [244]) in vitalem Glioblastomgewebe. Die zweite rotierte Diskriminanzfunktion nutzt Banden bei  $850$ ,  $1090$ , und  $1340\text{ cm}^{-1}$  zur Erkennung von hochgradigem Tumorgewebe. Allerdings ist der Beitrag dieser Spektralbereiche aufgrund der negativen Basislinie im Median auch für hochgradige Gewebe negativ, wirkt also der Einstufung als hochgradiges Gewebe entgegen.

Weiterhin werden zur Unterscheidung zwischen normalem und Tumorgewebe wenig intensive, aber breite Beiträge um  $1635\text{ cm}^{-1}$  genutzt. Das entspricht der Deformationsschwingung von Wasser. Köhler *et al.* [105] berichten einen erhöhten Wassergehalt in Glioblastomproben gegenüber normalem weißen und grauen Gewebe. Der Wassergehalt von Geweben ist definiert und Veränderungen können zur Diagnostik herangezogen werden. Bei den in dieser Arbeit untersuchten Proben kann er jedoch durch verschiedene Einflüsse verändert sein. Zum Einen kann Gefriertrocknung den Wassergehalt der Probe verringern. Andererseits kondensiert beim Auftauen der Probe Luftfeuchtigkeit, so dass die Feuchte erhöht wird. Nach dem Auftauen wurden die Proben zunächst gescannt, bevor die feuchte Kammer geschlossen wurde. In dieser Zeit verliert die Probe Feuchtigkeit. Je nach Wetterlage ist die relative Luftfeuchte der Raumluft sehr unterschiedlich, in der Heizperiode lag sie in den Räumen oft unterhalb von  $25\%$ , so dass innerhalb weniger Minuten die Probenoberfläche leicht antrocknete. Ein willkommener Nebeneffekt dieser oberflächlichen Trocknung ist, dass die Probe besser auf dem  $CaF_2$ -Fenster fixiert wird. Dadurch sinkt die Gefahr, dass die Probe auf dem Fenster verrutscht oder gar am Rand der feuchten Kammer herabfließt. Nach der Messung, also nach einigen Stunden in der feuchten Kammer, war keine oberflächliche Austrocknung mehr zu erkennen.

Für eine intraoperative Diagnostik beeinflussen jedoch zwei weitere Effekte den Wassergehalt im Messvolumen. Zum Einen treten im normalen Gewebe um den Tumor oft Ödeme auf und die Operation kann zu weiteren Schwellungen führen. Der Wassergehalt des Gewebes kann daher nur zur Diagnostik herangezogen werden, wenn auch mit Kon-

trollgewebe aus Ödemen trainiert wurde. Zum Anderen haben zwar die Gewebe selbst einen definierten Wassergehalt, während der Operation wird aber physiologische Kochsalzlösung zum Spülen und Anfeuchten der offengelegten Gewebe verwendet, so dass auf der Oberfläche der Gewebe eine undefinierte Schichtdicke von Wasser mit in die Raman-Spektren eingehen kann.

Die dritten Quartile der Raman-Spektren der hochgradigen Tumorgewebe (Abb. 16.1, S. 119) zeigen die typische spektrale Signatur von Hämoglobin, die bei Anregung mit 785 nm resonanzverstärkt ist. Insgesamt waren die Proben von hochgradigen Tumoren blutiger als die der niedriggradigen Tumore und die Kontrollproben, aber auch eine der Kontrollproben enthielt viel Blut. Trotz der intensiven charakteristischen Signatur und der Korrelation mit der Malignität trägt Hämoglobin nicht zur Klassifikation bei (Abb. 16.9). Das ist gut, denn ähnlich wie beim Wassergehalt ist das Auftreten von Hämoglobin-Signalen für eine intraoperative Diagnostik nicht verwertbar. Die Klassifikationsmodelle dürfen weder durch das Auftreten von Blut, noch durch den Oxidationszustand des Hämoglobins (siehe [244]) „verwirrt“ werden. Hinzu kommt, dass Änderungen im Oxidationszustand sowie Abbauprodukte von Hämoglobin nach Lagerung zu einer veränderten spektralen Signatur führen. Krafft *et al.* [132] haben dies an einer Hirntumorprobe gezeigt. Auch Raman-Messungen *in vivo* in Blutgefäßen zeigen eine erfreulich geringe spektrale Signatur von Blut [133].

### 16.2.6 Diskussion

Im Hinblick auf eine Anwendung der Raman-Spektroskopie *in vivo* während einer Astrozytomoperation erweist sich die logistische Regression mit weichen Trainingsdaten als vielversprechend. Die logistischen Regressionsmodelle, die im Gegensatz zur LDA auch Zugriff auf weiche (nicht komplett und eindeutig einer der Klassen zugehörige) Spektren haben, sind besonders bei der Einstufung solcher weichen Spektren im Vorteil.

Fast die Hälfte des Astro-Datensatzes sind weiche Spektren. Die weichen Spektren beinhalten die Übergangsbereiche zwischen den Tumorgaden und die Tumorränder. Daher handelt es sich um Testproben, die dem Einsatzgebiet einer intraoperativen Diagnostik entsprechen. Damit sind die LR-soft-Modelle den harten Modellen beim Astrozytom-Grading insgesamt überlegen. Sie zeigen auch geringere Unterschiede zwischen wMAE und wRMSE, also absolutem und quadriertem Fehler. Das heißt, dass die logistische Regression viele Spektren mit geringerem Fehler einordnet, während die LDA bei der Verarbeitung bei weniger Spektren größere Fehler macht. Das schlägt sich auch in den Sensitivitäts-Spezifitäts-Diagrammen als Vorteil der logistischen Regression nieder.

Die logistischen Regressionsmodelle für das Astrozytom-Grading erreichen Sensitivitäten von etwa 87 %, 67 % und 81 % bei Spezifitäten von 82 %, 71 % und 86 % für die harten Spektren von normalem Gewebe, niedriggradigem und hochgradigem Astrozytomgewebe, wenn ein Grenzwert von  $\frac{1}{3}$  zum Härten der Vorhersage angewendet wird. Die weichen Sensitivitäten für diese Klassen liegen bei 0,57, 0,51 und 0,65 wRMSE und die weichen Spezifitäten erreichen 0,70, 0,67 und 0,83 wRMSE (jeweils Median).

Die detaillierten Referenzdiagnosen für die einzelnen präparierten Schnitte zeigen, dass die Proben sehr heterogen sind (Kap. 14, S. 114). Das führt letztlich dazu, dass der vorgestellte Datensatz Anteile an Spektren mit ungenauer Referenzinformation enthält. Klassifikationsmodelle können auch mit einem gewissen Anteil falscher Referenzinfor-

mationen erfolgreich gebildet werden. Die Klassen erscheinen lediglich weniger separiert als sie in Wirklichkeit sind. Allerdings ist fraglich, ob die Fehlreferenzierung in realen Datensätzen den betrachteten mathematischen Modellen folgt [276]. Insbesondere kann die Fehlreferenzierung eine systematische Komponente haben (bestimmte Proben einer Klasse werden einer bestimmten anderen Klasse zugeordnet), die die Modellbildung beeinträchtigt. Auf jeden Fall leidet aber die Validierung unter Testdaten mit falscher Referenzinformation: wird ein (gutes) Modell getestet, so erscheint es schlechter, als es in Wirklichkeit ist.

Diese Situation spielt bei den hier diskutierten Daten durchaus eine Rolle, da bei etwa  $\frac{1}{3}$  der Messungen unterschiedlich große Anteile der verschiedenen Gewebe auf dem Messraster nicht genau lokalisiert werden konnten.

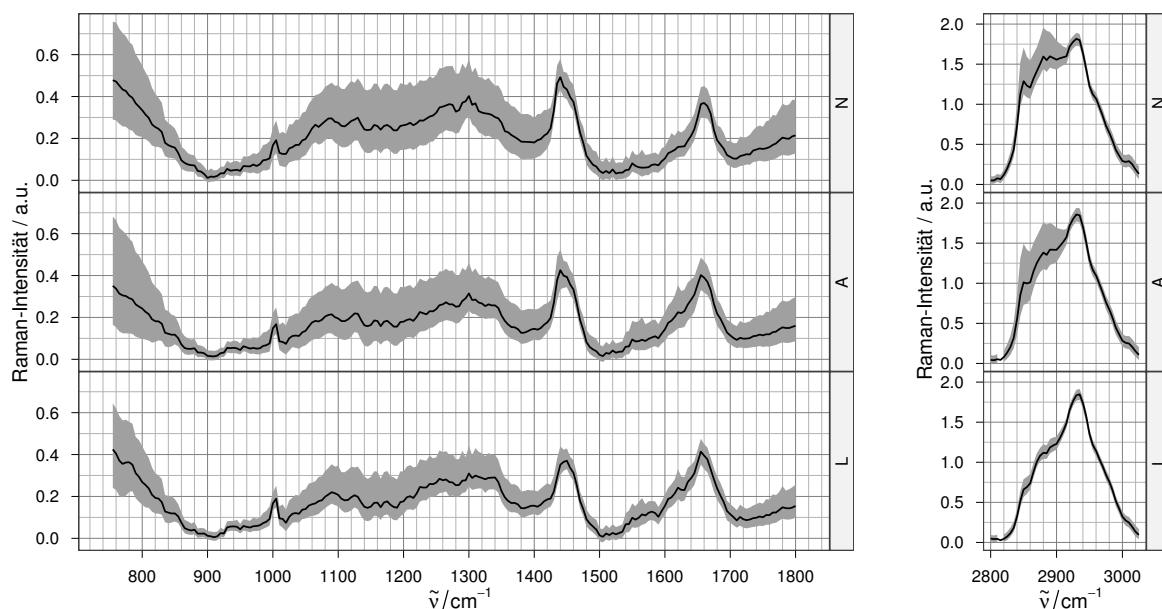
Die Vorhersagen der einzelnen Klassifikationsmodelle sind für die meisten Proben jeweils sehr ähnlich. Das gilt sowohl für die Iterationen der Kreuzvalidierung als auch, in etwas geringerem Maße, für die LDA-, LR-crisp- und LR-soft-Modelle. Die Vorhersagen der Modelle sind also relativ stabil gegenüber veränderten Trainingspatienten. Bestimmte Spektren (von bestimmten Proben) werden aber stabil falsch eingeordnet. Das kann einerseits daran liegen, dass die Modelle statt der einzelnen Gewebearten eine Kovariate vorhersagen, die aus den Spektren zugänglich ist, aber nicht perfekt mit der Gewebeart, genauer: mit der Morphologie der Gewebe, korreliert. Das wäre insofern nicht erstaunlich, als die morphologischen Änderungen nicht gleichzeitig mit den relativ kontinuierlichen biochemischen Veränderungen auftreten [46, 48, 49, 277]. Andererseits führen auch falsche Referenzinformationen in den Trainingsdaten zu so einer Ergebnisstruktur, solange die Modelle die Klassen korrekt abbilden. Mit den vorliegenden Daten kann der Einfluss dieser beiden Fehlerquellen nicht unterschieden werden. Dazu wäre eine Anzahl an homogenen Proben aus der Nähe der Klassengrenzen erforderlich.

## 17 Differentialdiagnostik von Lymphomen und Astrozytomen

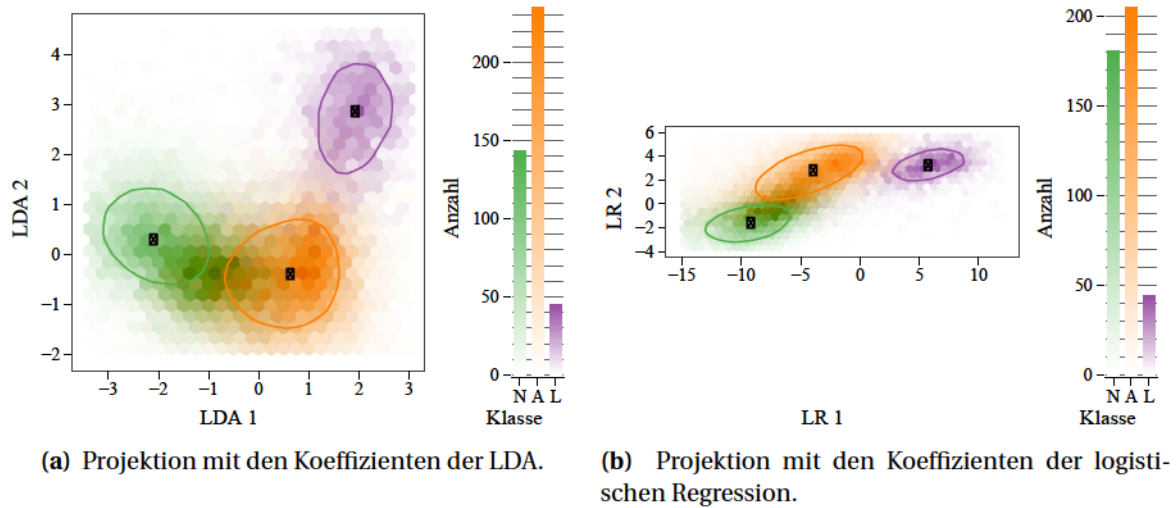
Verglichen mit den Unterschieden zwischen den Astrozytomgraden, unterscheiden sich die Raman-Spektren der Lymphome (Abb. 17.1) stärker von den Spektren der anderen beiden Klassen. Besonders im Bereich der CH-Valenzschwingungen zeigen sie wesentlich geringere Intensitäten. Wie die hochgradigen Astrozytome waren auch die Lymphomproben recht blutig und das schlägt sich in den Spektren nieder.

### 17.1 Deskriptive Modelle

Abbildung 17.2 zeigt die zweidimensionalen Histogramme aller Spektren des Lymph-Datensatzes mit einer LDA-Projektion (a) und einer Projektion mit den Koeffizienten eines logistischen Regressionsmodells des gesamten Datensatzes (b). Hier zeigt sich noch deutlicher als in den entsprechenden Histogrammen des Astro-Datensatzes (Abb. 16.2, S. 120) die große Zahl weicher Spektren, die Anteile von Astrozytom- und normalem Gewebe enthalten. Die Lymphome werden sowohl vom normalen Gewebe als auch von den Astrozytomen deutlich abgetrennt. Im Gegensatz zu den Astrozytomproben stammen alle Lymphomproben von hochmalignen Tumoren. Die 800 Spektren mit anteilig normalem und Lymphomgewebe sind zu wenige, um gegenüber den 8000 weichen Spektren mit Astrozytom- und normalen Anteilen aufzufallen, auch wenn es sich dabei um 30 % aller Spektren von den Lymphomproben handelt.



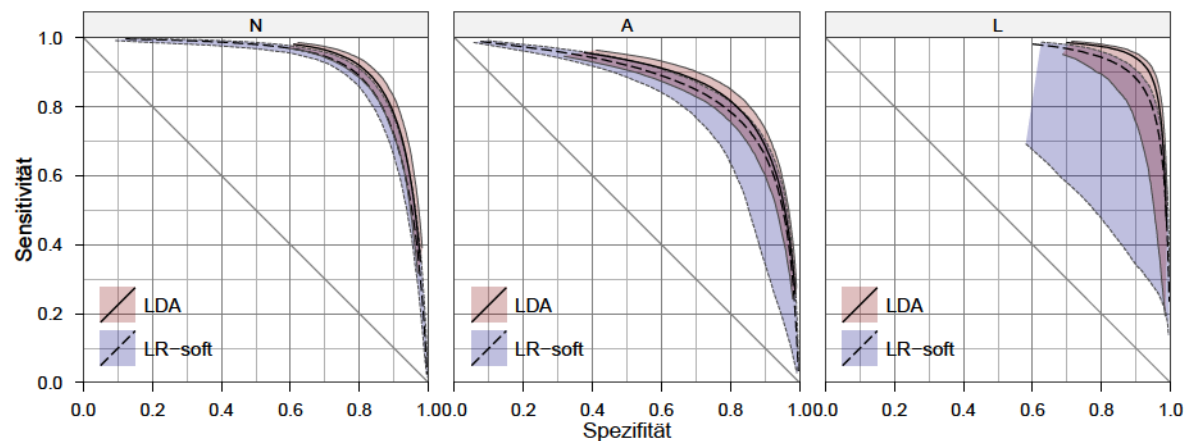
**Abbildung 17.1** Die Spekten des Lymph-Datensatzes: gewichtetes Median-, 16. und 84. Perzentilspektrum.



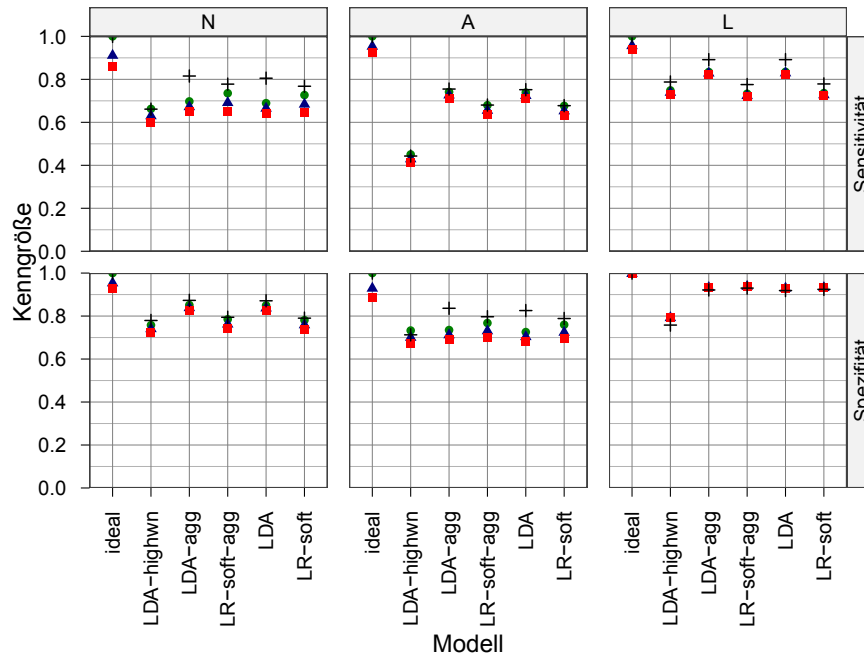
**Abbildung 17.2** Projektion der Lymph-Daten als 2d-Histogramme. Die durchgezogenen Konturen enthalten jeweils 50 % der harten Spektren der jeweiligen Klasse, die weißen Punkte markieren den zweidimensionalen Median der harten Spektren.

## 17.2 Prädiktive Modelle

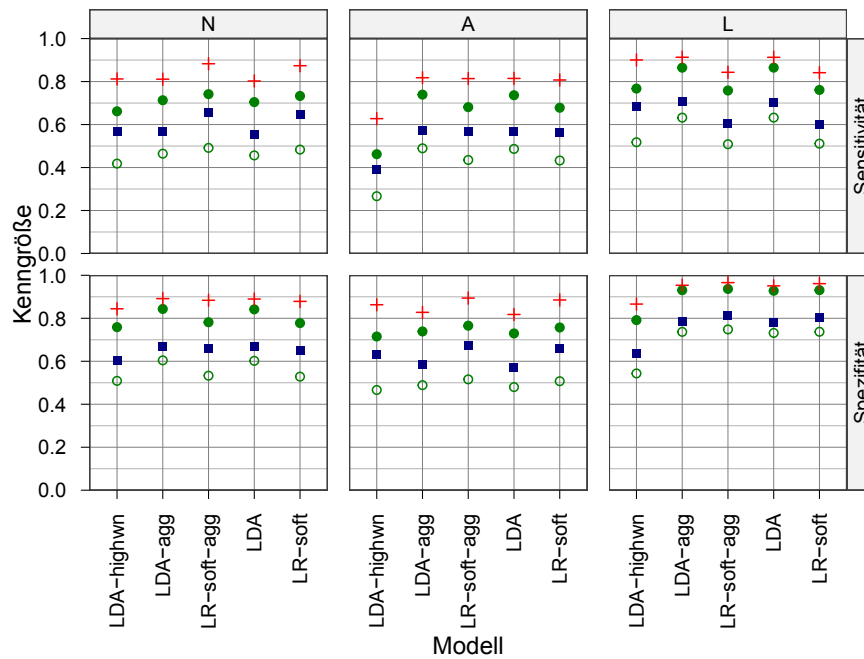
Bei der Unterscheidung zwischen Astrozytomen, Lymphomen und normalem Gewebe ist die LDA leistungsfähiger als die logistische Regression. Die LDA-Modelle erreichen sowohl bessere Spezifitäts-Sensitivitäts-Kennlinien für die harten Spektren (Abb. 17.3) als auch bessere weiche Sensitivitäten und Spezifitäten (Abb. 17.4a und 17.4b). Sowohl LDA als auch logistische Regression sind instabil, aber die logistische Regression ist von diesem Problem wesentlich stärker betroffen als die LDA. Obwohl die Erkennung von allen drei Klassen im Median nur wenig schlechter ist als bei der LDA, kommen einzelne Modelle vor, die insbesondere die Lymphome sehr viel schlechter erkennen (Abb. 17.3).



**Abbildung 17.3** Spezifitäts-Sensitivitäts-Diagramm. Vorhersagequalität der LDA und LR-soft Modelle. Eingetragen sind jeweils Median, 5. und 95. Perzentil der über die 126 Iterationen der Kreuzvalidierung beobachteten Kenngrößen. Sensitivität und Spezifität beziehen sich auf die Vorhersage der einzelnen Spektren, also der detaillierten histologischen Referenz.



(a) UND-Operatoren.



(b)  $\Delta$ -Operatoren.

**Abbildung 17.4** Weiche Sensitivität und Spezifität für die Unterscheidung von Lymphomen und Astrozytomen, berechnet (a) mit den UND-Operatoren: ■ starkes UND, ▲ Produkt-UND, ● schwaches UND und + die UND-Operatoren nur auf die Spektren mit eindeutiger Referenzzuordnung angewendet. Ideal gibt die bestmöglichen Werte an, die erhalten werden, wenn die Vorhersage genau der Referenz entspricht.) (b) Weiche Sensitivität und Spezifität berechnet mit den  $\Delta$ -Operatoren: ● 1 - wMAE, ■ 1 - wRMSE, ○ 1 - wMAE (untere Grenze für 1 - wRMSE), und + 1 - wMSE. Gezeigt ist jeweils der Median über die 126 Iterationen der Kreuzvalidierung.

Sowohl LDA als auch logistische Regression haben im Verhältnis zum mittleren absoluten Fehler einen großen quadrierten Fehler. Das heißt, viele Spektren werden komplett falsch eingeordnet.

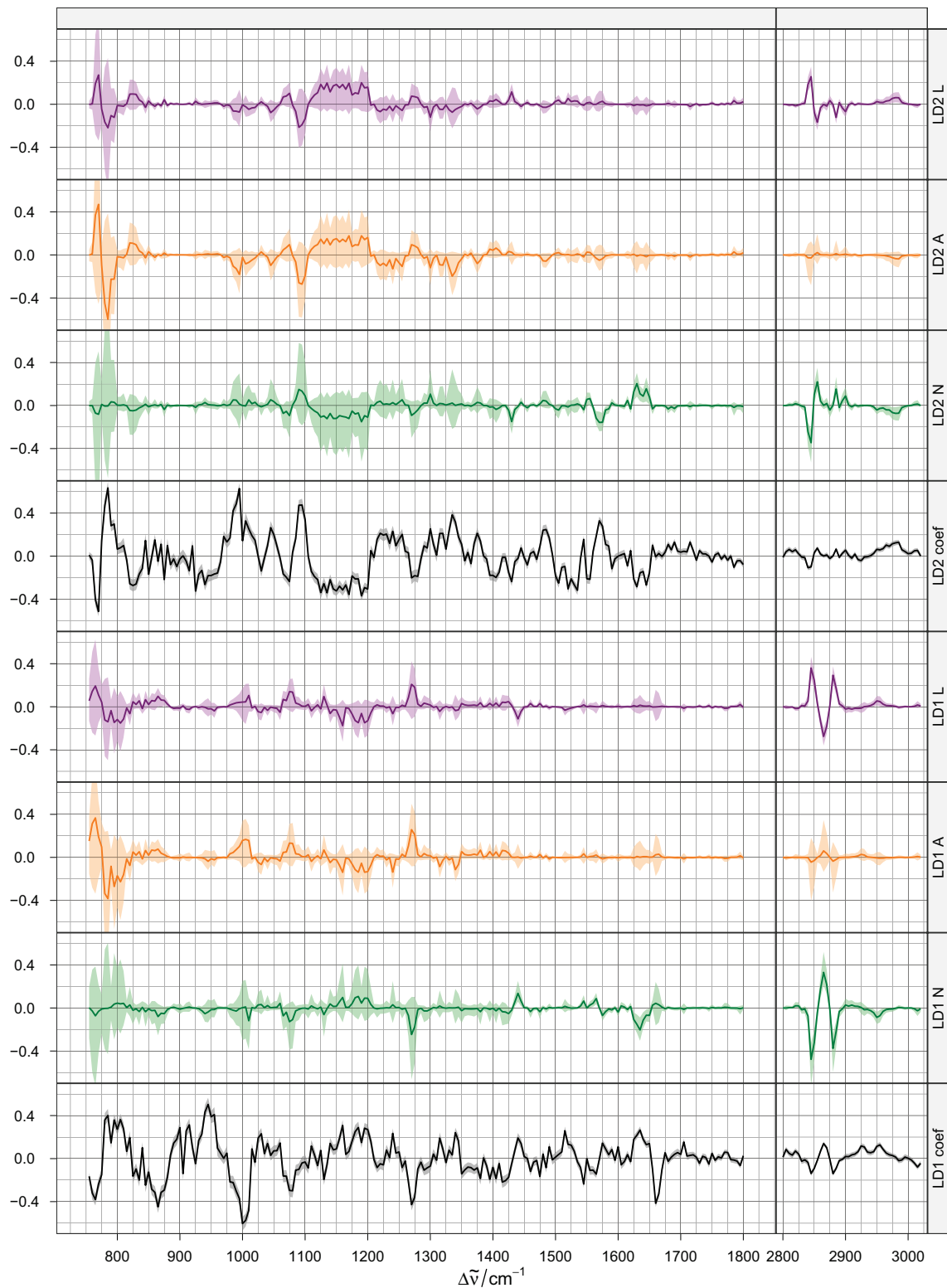
### 17.2.1 Spektroskopische Interpretation

Eine spektroskopische Interpretation der LDA-Modelle zur Unterscheidung von Lymphomen und Astrozytomen kann hier nur sehr eingeschränkt durchgeführt werden. Zum Einen muss berücksichtigt werden, dass nur von 8 Lymphom-Patienten überhaupt Proben zur Verfügung standen und davon 3 Proben große Anteile an normalem und gliotischem Gewebe enthielten. Zum Anderen ist im Gegensatz zu den Gliomen in der Literatur noch nicht im Detail über Raman-Spektren von primären Lymphomen des Zentralnervensystems berichtet worden, so dass die Koeffizientenmuster nicht auf ihre Übereinstimmung mit gezielten deskriptiven Studien zu den schwingungsspektroskopischen Unterschieden zwischen Lymphomen und Astrozytomen überprüft werden können. Bergner [109] beschreibt *ein* B-Zell-Lymphom mittels Raman-Mikrospektroskopie.

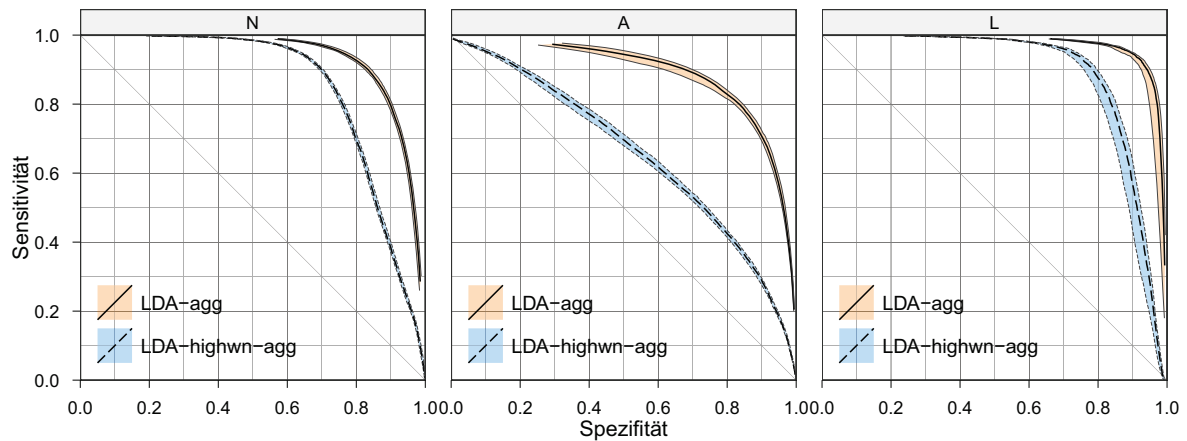
Die Koeffizienten der ersten gedrehten Diskriminanzfunktion ähneln denen der ersten gedrehten Diskriminanzfunktion beim Astrozytom-Grading. Insbesondere die Koeffizienten um 950, 1000, 1270 und 1625 bis 1675  $\text{cm}^{-1}$  weisen dasselbe Muster auf und die spektralen Beiträge der Astrozytome und Lymphome sind praktisch nicht zu unterscheiden. Im Bereich der C – H-Valenzschwingungen ist bei den Modellen zur Unterscheidung von Astrozytomen und Lymphomen allerdings ein w-förmiges Muster stärker ausgeprägt. Dieses hat negative Koeffizienten bei 2845 und 2880  $\text{cm}^{-1}$  und etwa gleichgroße positive Koeffizienten bei 2865  $\text{cm}^{-1}$ . Ein ähnliches Koeffizientenmuster hilft beim Grading der Astrozytome, die hochmalignen Gewebe von niedriggradigen Astrozytomen und normalen Hirngeweben abzutrennen. Dort war allerdings der Vorzeichenwechsel weniger stark ausgeprägt. Normales Gewebe erhält hier insgesamt negative Scores, während Lymphome positive Scores erhalten. Dieses Koeffizientenmuster misst die symmetrischen  $\nu_s\text{CH}_2$ - und  $\nu_s\text{CH}_3$ -Streckschwingungsbanden von Lipiden aus [97, 105, 275] und erzeugt starke negative Scores, wenn intensive und gut separierte Banden vorliegen. Verglichen mit dem Koeffizientenmuster beim Astrozytomgrading wird hier dem Minimum zwischen den Bandenmaxima mehr Bedeutung zugemessen. Das heißt, dass dieses Koeffizientenmuster größeres Gewicht auf eine gute Trennung der Banden legt. Das ist bei den normalen Hirngeweben der Fall. Auch in vielen Astrozytomspektren sind diese Banden identifizierbar, allerdings mit geringerer Intensität als bei den normalen Geweben. Demgegenüber zeigen die Lymphom-Spektren hier keine scharfen Banden. Die Lymphome erhalten entlang der ersten gedrehten Diskriminanzfunktion insgesamt etwas höhere Scores als die Astrozytome. Möglicherweise misst diese Funktion die Malignität des Gewebes. Bei den Lymphomen handelt es sich durchweg um hochmaligne Tumorproben, während die Astrozytomproben ja einen weiten Bereich an unterschiedlich malignen Geweben umfassen.

Dieser Unterschied wird auch von der zweiten gedrehten Diskriminanzfunktion aufgegriffen, die zu messen scheint, wie steil sich die steigende Flanke der Bande bei 2840  $\text{cm}^{-1}$  vom darunterliegenden, weniger strukturierten Signal abhebt. Unterhalb von 1800  $\text{cm}^{-1}$  sind die spektralen Beiträge der Astrozytome und Lymphome sehr ähnlich. Die Lymphomspektren erhalten aber über einen recht breiten Bereich mit unstrukturiert nega-





**Abbildung 17.5** Die gedrehten LDA-Modelle zur Unterscheidung zwischen Astrozytomen und Lymphomen. Gezeigt sind Median und Quartile der spektralen Beiträge der einzelnen Klassen (farbig), sowie die Koeffizienten der beiden gedrehten Diskriminanzfunktionen. Die Koeffizienten (coef) wurden um einen Faktor 10 verkleinert.

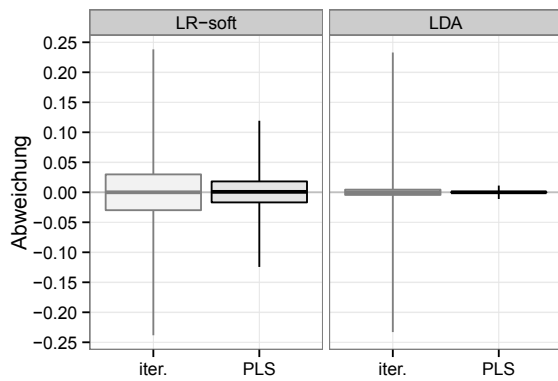


**Abbildung 17.6** Spezifitäts-Sensitivitäts-Diagramm. Vorhersagequalität der LDA- und LDA-high-Ensemblemodelle. Aggregiert wurden die Vorhersagen von jeweils 9 Surrogatmodellen, bei denen der jeweilige Patient *nicht* am Modelltraining beteiligt war. Eingetragen sind jeweils Median, 5. und 95. Perzentil der über die 126 Iterationen der Kreuzvalidierung beobachteten Kenngrößen. Sensitivität und Spezifität beziehen sich auf die Vorhersage der einzelnen Spektren, also der detaillierten histologischen Referenz.

tiven Koeffizienten zwischen  $1125$  und  $1200\text{ cm}^{-1}$  positive Scores, denen möglicherweise bestimmte Raman-Signale unterschiedlich stark entgegenwirken. Neben den C – H-Streckschwingungen der Lipide fallen besonders stark positive Koeffizienten bei  $1095\text{ cm}^{-1}$  auf. Das entspricht einer typischen DNS-Bande, die der antisymmetrischen Streckschwingung der Phosphatbrücken der DNS zugeordnet wird [CB11, 111, 275, 278]. Die spektralen Beiträge schwanken hier allerdings stark zwischen den einzelnen Spektren und im Median erreichen die normalen Gewebe hier leicht positive, die Lymphome etwas weniger negative Scores als die Astrozytome. Eine weitere typische DNS-Bande liegt bei  $785\text{ cm}^{-1}$  [CB11, 111], sie wird auf die Ringatmungsschwingung von Cytosin (laut De Gelder *et al.* [278] bei  $792\text{ cm}^{-1}$ ) zurückgeführt. Auch hier hat die 2. gedrehte Diskriminanzfunktion stark positive Koeffizienten. Bezüglich der spektralen Beiträge erhalten allerdings auch hier wieder sowohl Lymphome als auch Astrozytome im Median negative Scores. Allerdings wirken diese der Einstufung als Lymphom bei den Astrozytomen im Median stärker entgegen als bei den Lymphomen. Möglicherweise werden hier aber auch unkorrigierte Reste des Untergrunds unter den Raman-Spektren abgetastet.

### 17.2.2 Klassifikation mit den CH-Valenzschwingungsbanden

Auch Biopsienadeln mit faseroptischer Raman-Sonde könnten wesentlich einfacher realisiert werden, wenn kein Filter am Sondenkopf erforderlich ist (Kap. 3.1, S. 16). Die LDA-highwn-Modelle bleiben aber stark hinter den Modellen zurück, die auch den Spektralbereich von  $755$  bis  $1800\text{ cm}^{-1}$  nutzen. Obwohl der Spektralbereich der C – H-Valenzschwingungen ein sehr viel besseres Signal-Rausch-Verhältnis aufweist und die Koeffizienten der gedrehten LDA-Modelle hier wichtige Beiträge zur Einstufung der Spektren abgreifen, haben die LDA-highwn-Modelle erhebliche Schwierigkeiten beim Erkennen der Astrozytome. Dementsprechend ist die Zuordnung als normal oder Lymphom wenig spezifisch. Aggregation der Vorhersagen von jeweils neun Modellen Abbildung 17.6 stabilisiert die Vorhersagen zwar, führt aber zu keiner nennenswerten Verbesserung.



Perzentil	LR-soft		LDA	
	PLS	Iter.	PLS	Iter.
2,5	-0.18	-0.34	-0.02	-0.38
5	-0.12	-0.24	-0.01	-0.23
25	-0.02	-0.03	-0.00	-0.00
50	0.00	-0.00	0.00	-0.00
75	0.02	0.03	0.00	0.00
95	0.12	0.24	0.01	0.23
97,5	0.18	0.34	0.02	0.38

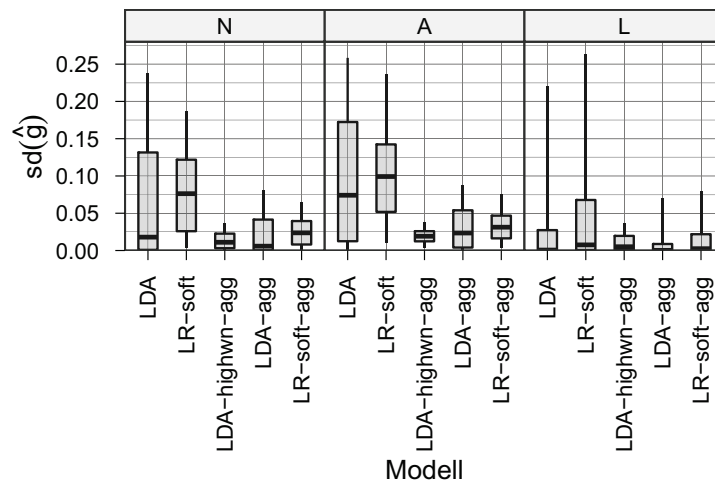
**Abbildung 17.7** Abweichung der vorhergesagten Klassenzugehörigkeiten der PLS-LR- und PLS-LDA-Modelle von den Modellen ohne PLS-Dimensionsreduktion (Box: 1. bis 3. Quartil und Median, Whisker: 5. bis 95. Perzentil). Zum Vergleich Abweichungen zwischen den Vorhersagen zufällig ausgewählter Iterationen. Die Unterschiede in den Vorhersagen mit und ohne PLS als Datenvorbehandlung sind kleiner als die Unterschiede zwischen den einzelnen Iterationen der Kreuzvalidierung. Wie beim Grading der Astrozytome besteht praktisch kein Unterschied zwischen den Vorhersagen der LDA-Modelle mit und ohne PLS-Vorbehandlung. Bei den LR-soft-Modellen treten hingegen deutliche Unterschiede auf, die allerdings immer noch kleiner sind als die Unterschiede zwischen den einzelnen Iterationen der Kreuzvalidierung.

### 17.2.3 PLS-LDA- und PLS-LR-Modelle

Die PLS-LDA-Modelle sind auch für diese Anwendung den LDA-Modellen ohne PLS-Vorprojektion praktisch gleich (Abb. 17.7). 90 % der Vorhersagen der PLS-LDA-Modelle differieren nicht mehr als  $\pm 0,012$  von entsprechenden Vorhersagen der LDA-Modelle, zwischen den Iterationen der Kreuzvalidierung differieren die einzelnen LDA-Modelle um  $\pm 0,23$ . Zwischen den LR-soft-Modellen mit und ohne PLS-Vorprojektion sind die Unterschiede allerdings größer: 90 % der Vorhersagen liegen um bis zu  $\pm 0,12$  auseinander. Aber auch hier sind die Differenzen zwischen den Iterationen noch deutlich größer ( $\pm 0,24$ ). Die PLS-LR-Modelle schneiden etwas besser ab als die LR-soft-Modelle (Anh. B.2), bleiben den LDA-Modellen jedoch unterlegen. Insgesamt teilen sie das Problem der Instabilität der LR-soft-Modellen ohne PLS-Vorbehandlung.

### 17.2.4 Aggregation von 9 Modellen

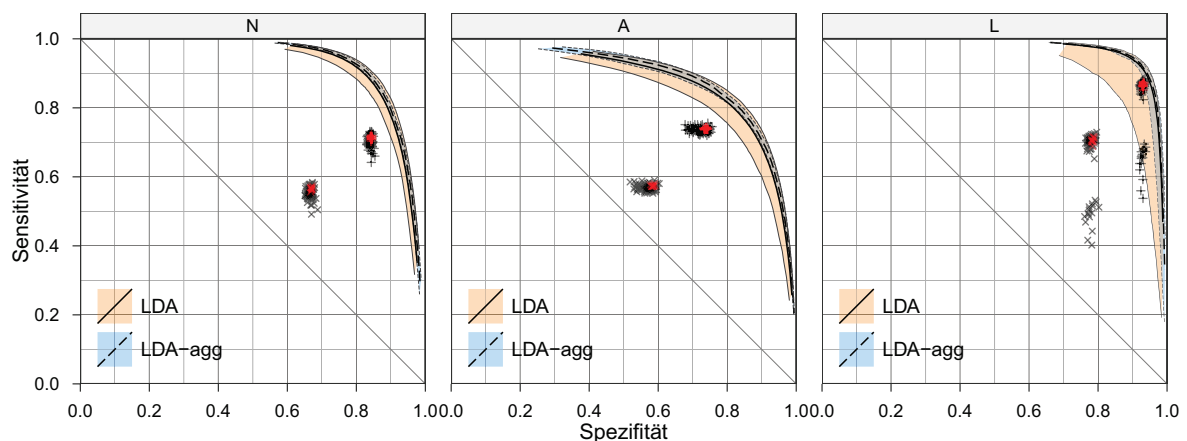
Die Standardabweichung der Vorhersagen der einzelnen im Rahmen der iterierten Kreuzvalidierung gebildeten Surrogatmodelle ist hoch (Abb. 17.8), die Vorhersagen sind also instabil. Das zeigt auch die große Bandbreite an beobachteter Modellqualität im Spezifitäts-Sensitivitäts-Diagramm (Abb. 17.3). Diese Instabilität ist zu erwarten, da insgesamt nur von 8 Patienten Lymphomproben verfügbar waren, von denen auch nur bei 5 Proben Bereiche eindeutig als Lymphom gekennzeichnet werden konnten. Von den anderen 3 Proben bestand eine weit überwiegend aus Leptomeninx, die möglicherweise auch Tumorzellen enthält, und zwei hauptsächlich aus gliotischem Gewebe mit jeweils eingesprenkelten Anteilen Tumor, die jedoch aufgrund der Verformung der Bulkprobe nicht genau lokalisiert werden konnten. Damit hatten die LDA-Modelle nur Trainingspektren von 4 oder 5 Patienten zur Beschreibung der Lymphome zur Verfügung. Die logistische Regression hat zusätzlich die Spektren mit Lymphom-Anteilen zur Verfügung, so dass



**Abbildung 17.8** Standardabweichung der Vorhersagen über die 126 Iterationen der Kreuzvalidierung für die drei Klassen der Unterscheidung von Lymphomen von Astrozytomen. Die Boxen zeigen Median, erstes und drittes Quartil, die Whisker reichen zum 5. und 95. Perzentil bezüglich aller Spektren.

einerseits die Probenbasis fast  $1\frac{1}{2}$ mal so viele Lymphom-Proben umfasst. Andererseits ist der Tumoranteil an der Leptomeninx-Probe sehr gering und bei den beiden anderen Proben ist nur die Referenzzuordnung weich (aufgrund der Ungenauigkeit bei der Lokalisierung), an sich liegen aber zwei unterschiedliche Gewebe vor. Letztlich reichen diese Proben mit geringen Lymphom-Anteilen also nicht aus, um den gegenüber der LDA größeren Bedarf an Trainingsproben für die logistische Regression zu decken.

Tatsächlich tritt bei der LDA ein gewisser Anteil an Modellen auf, die Lymphome deutlich weniger sensitiv als die Mehrheit der Modelle erkennen (Abb. 17.9, schwarze Kreuze). Die entsprechenden Modelle schneiden bei der logistischen Regression noch schlechter ab (Abb. 17.10, schwarze Kreuze). Damit liegt eine Situation vor, in der Modellaggrega-



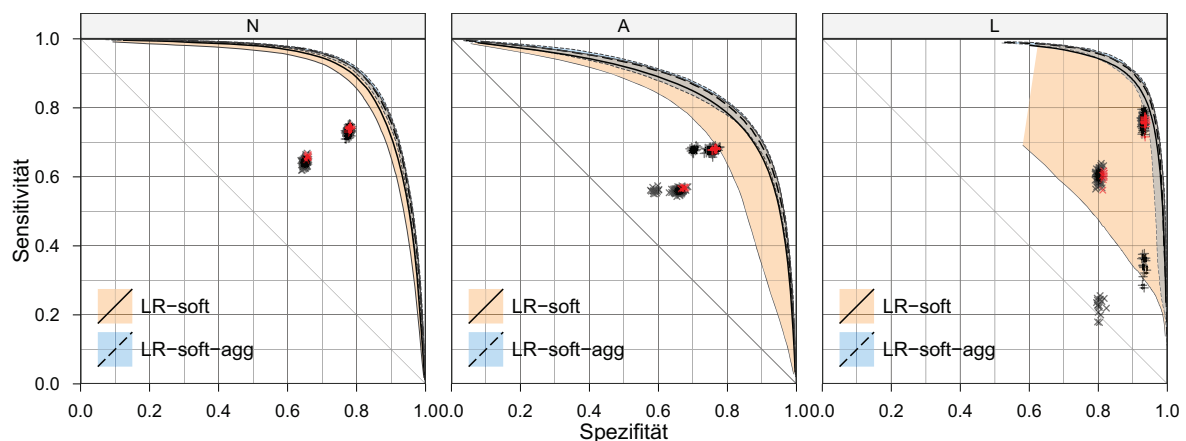
**Abbildung 17.9** Spezifitäts-Sensitivitäts-Diagramm. Vorhersagequalität der LDA Ensemble-Modelle im Vergleich zu den nicht aggregierten Modellen. Aggregiert wurden die Vorhersagen von jeweils 9 Surrogatmodellen. Eingetragen sind jeweils Median, 5. und 95. Perzentil der über die 126 Iterationen der Kreuzvalidierung beobachteten Kenngrößen. Weiche Sensitivität und Spezifität sind als wMAE (x) und wRMSE (+) eingetragen, schwarz die 126 einzelnen Modelle, rot die Ensemble-Modelle. Sensitivität und Spezifität beziehen sich auf die Vorhersage der einzelnen Spektren, also der detaillierten histologischen Referenz.

tion helfen kann, die Vorhersagen zu stabilisieren. Da die meisten Modelle gute Vorhersagen treffen (Mediane in (Abb. 17.3)), ist zu erwarten, dass die Modellaggregation die Vorhersagen nicht nur stabilisiert, sondern tatsächlich besonders schlechte Vorhersagen so vermieden werden. Das ist auch der Fall (rote Kreuze in Abb. 17.9 und 17.10).

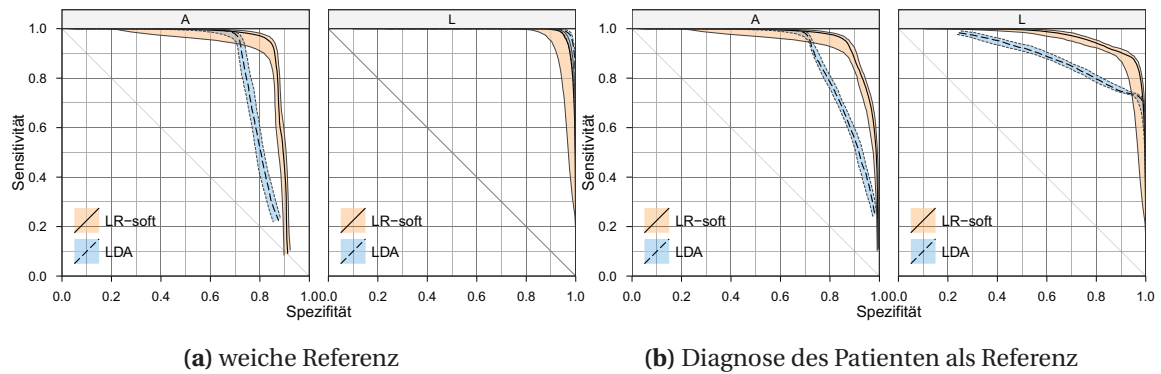
Da hier der Median über jeweils 9 Vorhersagen gebildet wird, kann sich der MAE praktisch nur dann verbessern, wenn die Referenzzuordnung zwischen den Vorhersagen der Surrogatmodelle liegt, also grundsätzlich nur für weich gelabelte Spektren. Diese machen bei den Proben von Lymphom- und Astrozytompatienten etwa  $\frac{1}{4}$  der Spektren aus. Wird über Spektren mit eindeutiger Referenz, also die restlichen  $\frac{3}{4}$  der Spektren, unter Mittelwertbildung aggregiert, so ändert sich der MAE überhaupt nicht. Entsprechend sind auch bei Bildung des Medians kaum Veränderungen zu erwarten. Tatsächlich ändern sich die mittleren und medianen wMAE-basierten Kennzahlen kaum (Anh. B.2).

Hingegen verringern sich wMSE und wRMSE auch für die Vorhersage von Spektren mit eindeutiger Referenzzuordnung, wenn durch die Aggregation einzelne sehr falsche Vorhersagen vermieden werden. Im Gegenzug werden natürlich auch einzelne sehr gute Vorhersagen auf eine mittlere Vorhersagequalität zurückfallen.

Bei der logistischen Regression vermeidet die Aggregation die sehr schlechten Vorhersagen: Minimum und 5. Perzentil der beobachteten Sensitivität über die 126 Iterationen steigen beim wMAE von 0,39 bzw. 0,33 auf 0,58 und 0,61. Beim wRMSE ist die Steigerung von 0,19 und 0,22 auf 0,49 und 0,52 noch viel drastischer. Im Gegenzug sinken der mediane und maximale beobachtete wMAE von 0,76 und 0,81 auf 0,74 und 0,77 leicht ab, beim wRMSE sinkt nur das Maximum (von 0,65 auf 0,63), Median und 95. Perzentil steigen leicht. Bei Astrozytom- und normalem Gewebe ist die Tendenz ähnlich, die Steigerungen bei den unteren Perzentilen ist aber weniger ausgeprägt – diese waren auch bei den einfachen Modellen nicht sehr viel schlechter als der Median. Dieses Muster ergibt sich auch bei den Spezifitäten, wobei hier die Änderungen geringer ausfallen. Bei der LDA folgen die Veränderungen derselben Tendenz, sind aber insgesamt weniger aus-



**Abbildung 17.10** Spezifitäts-Sensitivitäts-Diagramm. Vorhersagequalität der LR-soft-Ensemble-Modelle im Vergleich zu den nicht aggregierten Modellen. Aggregiert wurden die Vorhersagen von jeweils 9 Surrogatmodellen. Eingetragen sind jeweils Median, 5. und 95. Perzentil der über die 126 Iterationen der Kreuzvalidierung beobachteten Kenngrößen. Weiche Sensitivität und Spezifität sind als wMAE ( $\times$ ) und wRMSE ( $+$ ) eingetragen, schwarz die 126 einzelnen Modelle, rot die Ensemble-Modelle. Sensitivität und Spezifität beziehen sich auf die Vorhersage der einzelnen Spektren, also der detaillierten histologischen Referenz.



**Abbildung 17.11** Spezifitäts-Sensitivitäts-Diagramm. Vorhersagequalität der LDA und LR-soft Ensemble-Modelle zur Unterscheidung von Astrozytomen und Lymphomen. Aggregiert wurden die Vorhersagen von jeweils 9 Surrogatmodellen. Spektren mit vorhergesagten Anteilen von über 25 % normalem Gewebe werden zurückgewiesen und über 9 der gültigen Vorhersagen aggregiert. (a) zeigt Sensitivität und Spezifität dieser Vorhersagen gegenüber der detaillierten histologischen Referenz, (b) bezieht sich auf die Differentialdiagnose für den Patienten. Eingetragen sind jeweils Median, 5. und 95. Perzentil der über die 126 Iterationen der Kreuzvalidierung beobachteten Kenngrößen.

geprägt. Diese Modelle waren aber von Anfang an weniger instabil. Insgesamt erreichen die aggregierten LDA-Modelle etwas bessere Kennwerte als die aggregierten logistischen Regressionsmodelle.

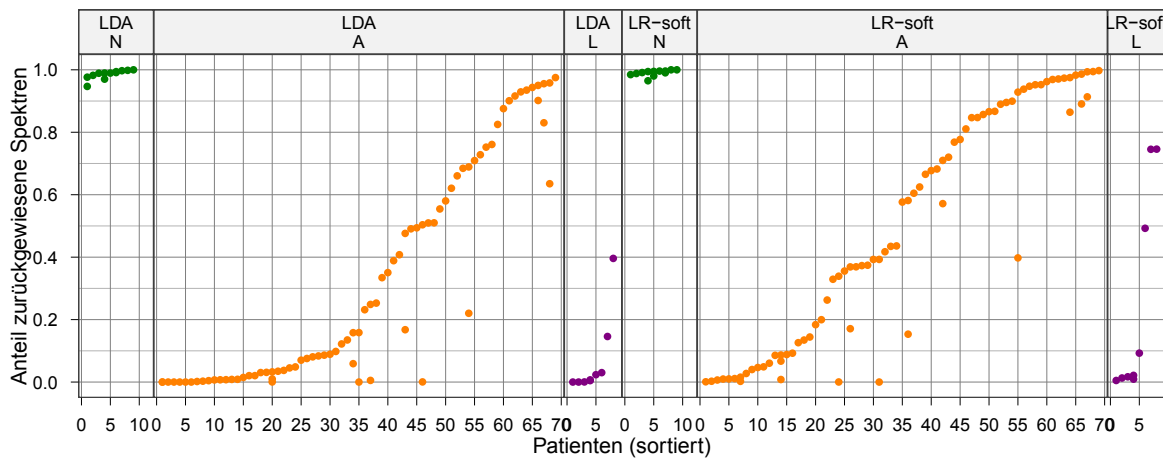
### 17.2.5 Aggregation über mehrere Spektren

Für die Differentialdiagnostik von Astrozytomen und Lymphomen ist die Vorhersagequalität für die weichen Spektren eher unwichtig: es ist höchst unwahrscheinlich, dass ein Patient gleichzeitig ein ZNS-Lymphom und ein Astrozytom hat. Normales Gewebe muss nur insofern erkannt werden, als diese Spektren keine Differentialdiagnostik erlauben. Hier gibt ausschließlich die Qualität der Vorhersagen für die Klassen Astrozytom und Lymphom für die Anwendung den Ausschlag, die bei den LDA-Modelle etwas besser aussieht.

Die Unterscheidung von Astrozytomen und Lymphomen erfolgt im Hinblick auf eine Anwendung der Raman-Spektroskopie bei Biopsien. Dabei kommt es nicht darauf an, ob *ein bestimmtes* Spektrum die korrekte Diagnose liefert: Bei Bedarf können mehrere Spektren aufgenommen werden, um eine sichere Differentialdiagnose zu stellen. In einem ersten Schritt können Spektren von normalem Gewebe als uninformativ zurückgewiesen werden. Dann können solange als Tumor erkannte Spektren gesammelt werden, bis die Entscheidung zwischen Lymphom und Astrozytom sicher möglich ist.

Als Annäherung an einen solchen Arbeitsablauf wurden zweistufige Modelle erstellt. In einem ersten Schritt werden alle Spektren ausgeschlossen, bei denen die vorhergesagte Zugehörigkeit zu normalem Gewebe mehr als 25 % erreicht, normales Gewebe hilft bei der Differentialdiagnostik nicht. In der zweiten Stufe werden Astrozytome von Lymphomen unterschieden. Diese Differentialdiagnostik ist wesentlich besser als die Vorhersagen der ersten Stufe (Abb. 17.11).

Wird aus diesen Vorhersagen mit einem Grenzwert von 0,5 die letztendliche Vorhersage bestimmt, so ergibt sich folgendes Bild: Von allen Proben der Kontrollpatienten wer-

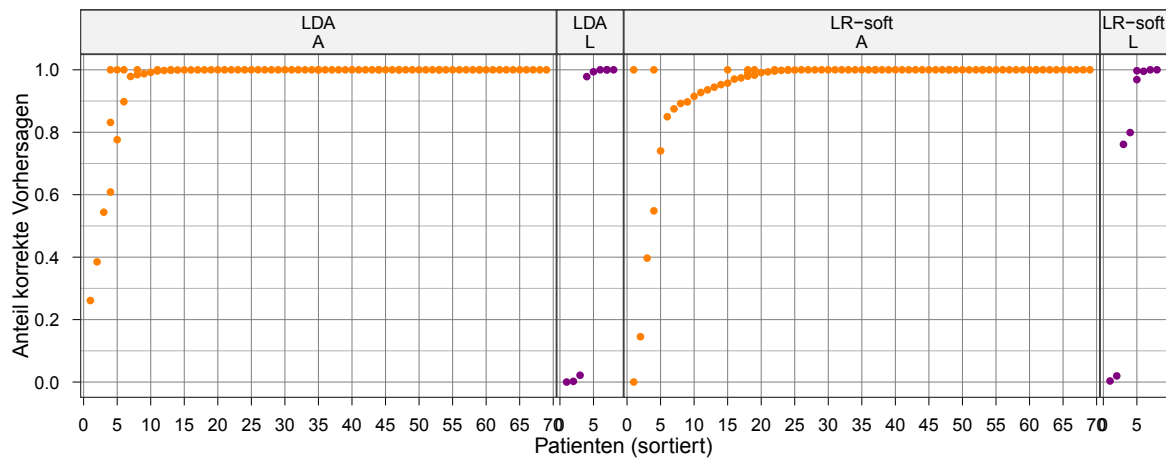


**Abbildung 17.12** Anteil zurückgewiesener Spektren für die verschiedenen Patienten. Die Patienten wurden jeweils nach aufsteigenden Anteilen an zurückgewiesenen Spektren sortiert. Bei einigen Patienten wurden mehrere Proben präpariert und gemessen.

den mindestens 96 % der Spektren zurückgewiesen (Abb. 17.12). Bei den Lymphompatienten werden von den drei Messungen mit großen Anteilen an normalem oder gliotischem Gewebe 52, 75 und 76 % der Spektren von den logistischen Regressionsmodellen zurückgewiesen, bei den restlichen fünf Messungen sind es jeweils unter 10 %. Die LDA weist von zwei der Proben mit hohem Anteil an normalem Gewebe 17 und 37 % zurück, bei allen anderen Proben sind es Anteile im unteren einstelligen Prozentbereich. Bei den Astrozytomproben werden Anteile von 0 bis 100 % zurückgewiesen. Die LDA-Modelle weisen auch bei den Astrozytomen weniger Spektren zurück. Insgesamt korreliert der Anteil an zurückgewiesenen Spektren mit dem Anteil an normalem Gewebe in der Probe, liegt aber oft höher. Hingegen ist keine Korrelation zwischen dem Anteil an zurückgewiesenen Spektren und dem Anteil an richtigen Zuordnungen zu Astrozytomen oder Lymphomen erkennbar (Abb. B.1, S. 186).

### 17.2.6 Diskussion

Abbildung 17.13 zeigt den Anteil der korrekten Zuordnungen nach Patienten aufgeschlüsselt. Die logistische Regression erkennt 2 der 9 Lymphomproben (2 von 8 Patienten) nicht als solche. Dabei handelt es sich um die beiden Proben mit viel gliotischem Gewebe. Die LDA erkennt auch den dritten Patienten mit viel Leptomeninx nicht. Bei diesem erkennt die logistische Regression in 80 % der Tests, dass es sich um einen Lymphompatienten handelt. Die logistische Regression erkennt bei einem weiteren Patienten nur in  $\frac{3}{4}$  der Tests das Lymphom, dieser wird von der LDA aber durchweg korrekt erkannt. 59 bzw. 53 der 81 Astrozytom-Proben werden von der logistischen Regression bzw. der LDA in mindestens 95 % der Tests korrekt erkannt. Diese stammen von 54 bzw. 49 der 69 Astrozytom-Patienten. Die LDA erkannte bei 4 dieser Patienten beide Proben, bei 4 weiteren eine von zwei Proben in mindestens 95 % der Tests korrekt. Bei der logistischen Regression waren es 5 Patienten, bei denen beide Proben in mindestens 95 % der Tests erkannt wurden und bei weiteren 6 Patienten gab es zusätzliche Proben, die in weniger als 95 % der Tests erkannt wurden. Vier dieser Proben (von 3) Astrozytom-Patienten enthalten laut Referenzdiagnose kein, 6 weitere Proben (von 6 Patienten) nur bis zu 5 % Tumorge-



**Abbildung 17.13** Anteil korrekter Vorhersagen bei der Differentialdiagnostik zwischen Lymphomen und Astrozytomen für die verschiedenen Patienten. Die Patienten wurden jeweils nach aufsteigenden Anteilen an korrekten Zuordnungen sortiert. Bei einigen Patienten wurden zwei oder drei Proben präpariert und gemessen.

webe. Von diesen werden richtig die allermeisten Spektren zurückgewiesen. Die wenigen zur Unterscheidung von Lymphomen und Astrozytomen akzeptierten Spektren werden zum Teil aber weit überwiegend richtig als Astrozytompatient beurteilt.

5 von 8 korrekt zugeordneten Patienten entspricht einer Sensitivität für Lymphome von 63 %, allerdings reicht das 95 %-Konfidenzintervall von etwa 30 bis 86 %, umfasst also mehr als den halben Wertebereich. 54 korrekt erkannte von 69 Astrozytom-Patienten entspricht 78 %. Hierbei ist das 95 %-Konfidenzintervall mit etwa  $\pm 10$  %punkten schon deutlich schmaler.

Bei dem gewählten Grenzwert von 0,5 für die Entscheidung zwischen Astrozytom und Lymphom sind die Modelle sehr sensitiv, aber wenig spezifisch den Astrozytomen gegenüber. Oft wird ein solcher Grenzwert anhand der Spezifitäts-Sensitivitäts-Kurve optimiert, zum Beispiel so, dass der Abstand vom idealen Modell möglichst gering ist, oder dass die Sensitivität gleich der Spezifität sein soll. Letzteres Kriterium sollte bei 50 % vorhergesagter *a-posteriori*-Wahrscheinlichkeit liegen. Das ist hier nicht der Fall, Sensitivität und Spezifität sind erst bei einem Grenzwert von 0,65 für die Diagnose Astrozytom durch die logistischen Regressionsmodelle (LDA: 0,79) gleich. Das heißt, dass die Modelle nicht gut kalibriert sind: die vorhergesagte Klassenzugehörigkeitswahrscheinlichkeit stimmt nicht mit der tatsächlichen Wahrscheinlichkeit, dass es sich um eine Probe der jeweiligen Klasse handelt, überein. Diese Abweichung wird von den wRMSE- und wMAE-basierten Kenngrößen im Gegensatz zum Spezifitäts-Sensitivitäts-Diagramm berücksichtigt. Die auch in Abbildung 17.11a erkennbare generell niedrige Spezifität der Modelle (bezogen auf die detaillierte histologische Referenz) den Astrozytomen gegenüber verstärkt sich so zu extrem niedrigen wRMSE von im Median 0,50 und 0,36. Diese Abweichungen werden möglicherweise durch die extrem unterschiedlichen Probenzahlen (mit)verursacht: 81 Astrozytomproben bzw. 69 Astrozytompatienten stehen nur 9 Lymphomproben (8 Patienten) gegenüber.

Für die hier betrachtete Anwendung sollte eine solche Festlegung jedoch nicht im Spezifitäts-Sensitivitäts-Diagramm geschehen, sondern besser im entsprechenden Dia-



gramm, das die gegenseitige Abhängigkeit von positivem und negativem prädiktivem Wert aufträgt. Da die relative Häufigkeit von Lymphomen und Astrozytomen bei der Nadelbiopsie aber unbekannt ist, können die prädiktiven Werte nicht ausgerechnet werden. Die Prävalenz bzw. Inzidenz der verschiedenen Tumore in der Bevölkerung kann hier nicht verwendet werden, sie kann erheblich von der Prävalenz der unterschiedlichen Tumore bei der Biopsie abweichen.

Lymphome im ZNS werden normalerweise nicht operiert. Mit nur 8 Patienten oder weniger als 0,5 % der Proben sind Lymphome daher in der Hirntumor-Sammlung extrem selten. Diese Proben können auch nur einen kleinen Teil der Lymphomgewebe repräsentieren, die bei der Biopsie zur Differentialdiagnostik zwischen Astrozytomen und Lymphomen auftreten: Nur wenige Lymphome entwickeln sich so, dass eine Operation notwendig und sinnvoll ist. Das muss bei der Interpretation der Ergebnisse berücksichtigt werden. Die Modelle zur Einordnung der einzelnen Spektren unterscheiden sehr gut: die wMAE-basierte Sensitivität und Spezifitäten von Astrozytomen gegenüber Lymphomen sowohl der LDA- als auch der logistischen Regressionsmodelle liegen im Median deutlich über 90 %.

Im Gegensatz zur logistischen Regression können LDA-Modelle Proben ablehnen, die *keiner* der Klassen ähnlich sind. Das ist bei der Unterscheidung zwischen Lymphomen und Astrozytomen insofern wichtig, als bei einer Biopsie nicht nur diese beiden Gewebe auftreten können. Die Differentialdiagnostik bei Verdacht auf ein PZNSL umfasst eine ganze Reihe anderer Krankheiten wie zum Beispiel multiple Sklerose oder Toxoplasmose [55], die sich auch auf die biochemische Zusammensetzung des untersuchten Gewebes auswirken können. Bei einer Raman-Biopsie können daher sehr unterschiedliche Spektren auftreten, die weder zu Astrozytomen noch zu Lymphomen gehören und mehr oder weniger stark von den Spektren normalen Gewebes abweichen. LDA kann Proben als „keiner der vorgegebenen Klassen ähnlich“ zurückweisen, was in dieser Situation sehr hilfreich sein kann. Voraussetzung dafür ist natürlich, dass die Spektren hinreichend unterschiedlich von denen der betrachteten Tumore sind.

Drei Lymphomproben sind von Gewebe dominiert, das nicht zum Tumor gehört. Die deskriptive LDA ordnet diese Gewebe nicht der Klasse N, sondern sehr eindeutig den Astrozytomen zu. In der deskriptiven logistischen Regression erscheinen sie in der Nähe der Astrozytomgewebe, jedoch eindeutig von ihnen getrennt und eher auf halber Entfernung zwischen den normalen Geweben und den Lymphomgeweben. Bei den Proben 381 und 388 handelt es sich dabei überwiegend um gliotisches Gewebe. Gliosen sind Narbengewebe im Hirn, die von Gliazellen gebildet werden. Sie können unter anderem als Reaktion auf ein Lymphom, nach Verletzungen oder Operationen vorkommen [45]. Da der vorgestellte Datensatz insgesamt nur sehr wenige Spektren von gliotischem Gewebe enthält, konnten diese nicht als eigene Klasse modelliert werden. Für die Zukunft stellt dies aber eine wichtige Erweiterung des Modells dar, zumal Gliosen bei weiteren Erkrankungen wie multipler Sklerose auftreten, die bei Verdacht auf ein PZNSL ebenfalls ausgeschlossen werden müssen. Auch als Reaktion auf absterbendes Tumorgewebe zum Beispiel bei einer Strahlentherapie kommt es zur Bildung von Gliosen.

## 18 Zusammenfassung

Die vorliegende Arbeit untersucht das Potential der Raman-Spektroskopie zur intraoperativen Hirntumordiagnostik. Zwei konkrete Fragestellungen wurden betrachtet: das Grading von Astrozytomen und die Unterscheidung (Differentialdiagnostik) zwischen Astrozytomen und Lymphomen. Astrozytome werden, so möglich, mit einer Operation entfernt. Dabei ist der Tumorrand für die Neurochirurgen nur schwer und ungenau erkennbar. Eine Diagnostik *in vivo* während der Operation könnte dabei helfen. Demgegenüber werden Lymphome nicht operiert, sondern mit einer Chemotherapie oder kombinierter Chemo- und Strahlentherapie behandelt. Für die Therapieentscheidung muss allerdings bekannt sein, ob es sich um ein Astrozytom oder ein Lymphom handelt. Diese Differentialdiagnose wird anhand von Biopsieproben gestellt. Trotz Stereonavigation kann aber nicht immer eine geeignete Biopsieprobe gewonnen werden. Eine ramangestützte Diagnostik könnte hier sicherstellen, dass die Biopsieprobe tatsächlich von Tumorgewebe genommen wird. In letzter Konsequenz könnte sie die Entnahme einer Biopsieprobe sogar ganz ersetzen. Das Astrozytom-Grading befasst sich also mit der Erkennung von normalen, niedriggradigen und hochgradigen Geweben innerhalb eines Tumors, während die Unterscheidung von Astrozytomen und Lymphomen auf eine Diagnostik für den gesamten Tumor zielt.

Mit Hilfe einer faseroptischen Sonde wurden Raman-Spektren von frisch aufgetauten Proben in einer feuchten Kammer gemessen. Zusätzlich wurden Mikrotomschnitte für eine histologische Begutachtung angefertigt. Die Begutachtung zeigt, dass die Proben sehr heterogen sind. Viele Proben sind zudem von Gewebe dominiert, das morphologisch niedriggradigerem Tumorgewebe ähnelt, als es dem Tumor des Patienten entspricht. Große Anteile der Proben sind auch von Infiltration geprägt, also gemischten Zellpopulationen und Zellpopulationen, die im Prozess der Entdifferenzierung waren. In keinem Gewebeschnitt wurden Hinweise auf höhere Malignität gefunden, als es dem Tumorgrad des Patienten entsprach.

Die Auswertung der Spektrendaten erfolgte mit chemometrischen Klassifikationsmethoden, der linearen Diskriminanzanalyse LDA und der logistischen Regression. Klassifikationsmethoden unterscheiden definierte Gruppen innerhalb der Datensätze, die sogenannten Klassen. Dazu werden sowohl die Spektren als auch die Ergebnisse der histologischen Untersuchung der Proben als Trainingsdaten verwendet. So entstehen Modelle, die neue, unbekannte Spektren einer der Klassen zuordnen (Vorhersage).

Das ist zunächst nur für Spektren möglich, die eindeutig zu einer der Klassen gehören. Spektren, die nicht eindeutig in die Klassen eingeordnet werden können, mussten bisher von der weiteren Auswertung ausgeschlossen werden. Das ist besonders für das Astrozytom-Grading in mehrfacher Hinsicht problematisch: Astrozytome wachsen infiltrativ. Sie neigen außerdem zur Entdifferenzierung und durchlaufen dabei molekularbiologisch mehrere Schritte [30, 46, 47]. Die damit verbundene Uneindeutigkeit bezüglich der Klassen ist unvermeidlich. Astrozytome lassen sich ohne diese uneindeutigen Bereiche nicht repräsentativ beschreiben. Ein Astrozytom-Grading während der Opera-

tion soll dem Neurochirurgen bei der Entscheidung, wie weit ein Tumor entfernt wird, helfen. Gerade am Rand des Tumors herrschen aber infiltrierte Gewebe vor. Die Einstufung solcher gemischten Zellpopulationen und auch homogener Zellpopulationen mit uneindeutiger Dignität ist das eigentliche Ziel der Astrozytom-Diagnostik – bei eindeutigen Geweben würde eine Raman-Diagnostik nicht oder nur selten benötigt. Werden Gewebe ausgeschlossen, die nicht eindeutig einer der Klassen zugeordnet werden können, so entsteht ein *scheinbar* einfaches Klassifikationsproblem: die schwierigen Fälle werden ignoriert. Auch wenn möglicherweise dennoch ein funktionierendes Klassifikationsmodell gebildet werden kann, ist es ungünstig, genau diejenigen Proben zu ignorieren, die Beispiele der gesuchten Klassengrenze sind. Hierin unterscheidet sich die lineare Diskriminanzanalyse, die nur eindeutige Trainingsdaten verarbeiten kann von der logistischen Regression. Die uneindeutigen Klassenzugehörigkeiten wurden als Anteile modelliert und direkt beim Training der hier vorgestellten logistischen Regressionsmodelle genutzt.

Bei der Beurteilung der Leistungsfähigkeit der gebildeten Modelle sind die Grenzfälle unverzichtbar. Da die Grenzfälle die Zielproben für die Astrozytom-Diagnostik sind, muss die Qualität des Modells auch für Vorhersagen solcher Grenzfälle gemessen werden. Für die klassischen Kenngrößen für die Qualität von Klassifikationsmodellen wie Sensitivität, Spezifität, positive und negative prädiktive Werte wurde in der vorliegenden Arbeit wurde ein einheitlicher Rahmen entwickelt, der eine konsistente Validierung ermöglicht, bei der sowohl die Vorhersage des Modells als auch die Referenzinformation anteilige Klassenzugehörigkeiten aufweisen können. Die so erhaltenen Kenngrößen bilden einerseits die Uneindeutigkeit der Referenz ab. Man erhält einen möglichen Bereich von der bestmöglichen bis zur schlechtestmöglichen Übereinstimmung. Auch eine erwartete Übereinstimmung kann berechnet werden. Aus dieser wurde eine weitere Reihe von Kenngrößen abgeleitet, die in engem Zusammenhang mit den aus der Regression bekannten mittleren absoluten und mittleren quadrierten Fehlern stehen. Letztere Kenngrößen zeigen unter anderem sehr einfach an, ob viele Vorhersagen des Modells um kleine Beträge von der Referenz abweichen, oder ob weniger, aber grob falsche Vorhersagen vorliegen. Weiterhin wurden die neuen „weichen“ Kenngrößen im Hinblick auf Bias und Varianz interpretiert. Kenngrößen mit optimistischem Bias können gezielt durch Angabe der entsprechenden Kenngröße mit pessimistischem Bias komplementiert werden. Die weichen Kenngrößen zeigen in verschiedenen praktisch relevanten Situationen eine wesentlich geringere Varianz als ihre klassischen Analoga. Insbesondere stehen die auf dem mittleren quadrierten Fehler der erwarteten Übereinstimmung basierenden Kenngrößen in engem Zusammenhang mit Brier's score, einer *proper scoring rule*.

Um die Aussagekraft eines chemometrischen Modells für unbekannte Proben zu messen, muss das Modell Zuordnungen für Proben treffen, die die Modellbildung in keiner Weise beeinflusst haben, die also *statistisch unabhängig* sind. Diese Zuordnungen werden dann mit der Referenzzuordnung (Histologie) verglichen. Testdaten müssen mindestens dieselben Anforderungen erfüllen wie Trainingsdaten. Praktisch wird daher ein Datensatz in Test und Trainingsdaten geteilt. Da die einzelnen Spektren eines Patienten möglicherweise untereinander ähnlicher sind als zu Spektren anderer Patienten, sind sie nicht statistisch unabhängig. Deshalb muss die Teilung Test- und Trainingspatienten unterscheiden. Weiterhin dürfen patientenübergreifende Datenvorbehandlung und eine eventuell erfolgende Modelloptimierung ausschließlich Trainingspatienten nutzen. Die

Übertragung der Referenzinformationen auf die einzelnen Spektren erfolgte verblindet, also ohne dass spektrale Informationen zugänglich waren.

Zu wenige Trainingsproben bedeuten, dass die Qualität der chemometrischen Modelle leidet. Insbesondere können die Modelle instabil werden. Zu wenige Testproben bedeuten, dass die Qualität des Modells nur sehr ungenau gemessen werden kann. Um diese Effekte auszugleichen oder wenigstens die Modellstabilität zu messen, wird der Datensatz daher nicht nur einmal, sondern viele Male in Test- und Trainingspatienten geteilt (*re-sampling*). Dabei muss die gesamte Modellbildung jedesmal neu durchgeführt werden. Im Rahmen der vorliegenden Arbeit wurden insgesamt 1008 solche Teilungen in Form einer 126× iterierten 8-fachen Kreuzvalidierung durchgerechnet.

In der Zusammenschau folgt, dass es bei den derzeit typischen Klassifikationsproblemen (und Probenzahlen) in der schwingungsspektroskopischen Diagnostik oftmals einfacher ist, ein gutes Modell zu trainieren, als nachzuweisen, dass das Modell tatsächlich gut ist. Der Nachweis, dass ein Modell tatsächlich besser ist als ein anderes, ist gegenwärtig in der Biospektroskopie nicht zu führen.

Die logistischen Regressionsmodelle für das Astrozytom-Grading erreichen Sensitivitäten von etwa 87 %, 67 % und 81 % bei Spezifitäten von 82 %, 71 % und 86 % für normales, niedriggradiges und hochgradiges Gewebe. Aus den beobachteten Sensitivitäten und Spezifitäten (nach der „klassischen“ Rechenmethode) können mit Hilfe der Anzahl an unabhängigen Testproben Konfidenzintervalle für die Sensitivitäten und Spezifitäten berechnet werden. Im vorliegenden Datensatz ist allerdings der Stichprobenumfang im statistischen Sinne unbekannt, er liegt zwischen der Anzahl an Testpatienten und der Anzahl an Testspektren. Diese beiden Größen unterscheiden sich jedoch um zwei Zehnerpotenzen. Allerdings kann die Anzahl an Testpatienten zu einer konservativen Abschätzung genutzt werden. Dann ergeben sich für das Astrozytom-Grading 95 %-Konfidenzintervalle für die Sensitivitäten von etwa 62 – 96 %, 43 – 84 % und 63 – 91 % bei Spezifitäten von 67 – 91 %, 57 – 83 % und 70 – 94 % für normales, niedriggradiges und hochgradiges Gewebe. Hinzu kommt noch ein Beitrag aufgrund der Instabilität der Surrogatmodelle. Dieser ist hier aber praktisch insofern von untergeordneter Bedeutung, als bereits die ausschließlich aus der Anzahl an unabhängig getesteten Patienten berechneten Konfidenzintervalle so breit sind, dass sie von guter bis sehr guter Vorhersagequalität bis zu wesentlich zu geringen Vorhersagequalitäten reichen. Insbesondere die Sensitivität für niedriggradige Astrozytomgewebe reicht von „schlechter als raten“ (43) bis 84 %. Letzteres wäre in Anbetracht der objektiv schwierigen Abgrenzung der niedriggradigen Gewebe eine gute Vorhersagequalität.

Bei der Unterscheidung der Lymphome von den Astrozytomen erkannte die LDA 53 von 81 Astrozytom-Proben (von 69 Patienten) in über 95 % der Tests (65 %; 95 %-Konfidenzintervall 55 – 75 %) und 5 von 8 Lymphom-Patienten in über 95 % der Tests (63 %; 95 %-Konfidenzintervall 30 – 86 %). Die logistische Regression erkannte 59 von 81 Astrozytom-Proben in über 95 % der Tests (73 %; 95 %-Konfidenzintervall 63 – 82 %) und 4 von 8 Lymphom-Patienten in über 95 % der Tests (50 %; 95 %-Konfidenzintervall 21 – 79 %), sowie zwei weitere in 75 und 80 % der Tests. Drei der nicht bzw. ungenau erkannten Lymphome sind Proben, die nur wenig Tumorgewebe enthalten. Auch unter den Astrozytomproben waren drei ohne Anteile von Tumorgewebe.

Die Breite dieser Konfidenzintervalle verdeutlicht, dass die Bestimmung der Modellqualität oft wesentlich mehr Proben benötigt, als das Training guter Modelle erfordert.

## 19 Ausblick

**Probenpräparation:** Der nächste Schritt in Richtung Diagnostik *in vivo* könnte die Messung frischer Proben direkt nach Entnahme sein. Diese Messungen sind in mehrfacher Hinsicht erforderlich, bevor *In-vivo*-Modelle gebildet werden können. Auf der einen Seite ist ungeklärt, inwieweit die Proben auch durch kurze Lagerung verändert werden. Für eine einzelne Hirntumorprobe beschreiben Krafft *et al.* [132] Veränderungen zwischen den Spektren der frischen Probe und nach einem Tag bei  $-80\text{ }^{\circ}\text{C}$ , die auf Änderungen im Oxidationsstatus und möglichen Abbau von Hämoglobin zurückgeführt werden. Wie in Kapitel 16.2.5 diskutiert, sollte die spektrale Signatur von Blut jedoch nicht in die Einstufung der Gewebe einfließen.

Auch wenn diese Veränderungen minimal sind, können sie die Klassifikationsmodelle beeinflussen, da die Modelle ja kleine Veränderungen über weite Spektralbereiche nutzen. Auf der anderen Seite sind gute Referenzinformationen bei Messungen *in vivo* praktisch nicht zugänglich. So beschreiben Stelling *et al.* [103] und Stelling *et al.* [104] ATR-IR Messungen an Quetschpräparaten von frisch entnommenem Gewebe. Als Referenz wird allerdings nur die Diagnose für den Patienten verwendet. Durch die mechanische Präparation lassen sich örtliche Unterschiede nicht auf ein Messraster übertragen. Detaillierte histologische Gutachten können grundsätzlich nur dann auf Quetschpräparate übertragen werden, wenn die Probe homogen ist.

Der Schwerpunkt der Probennahme muss aber in den heterogenen Bereichen der Tumore liegen, da dort das Anwendungsgebiet einer Raman-basierten Diagnostik während der Operation liegt. Auch wenn das Gewebe direkt nach der Messung entnommen und histologisch begutachtet wird, ist die Probennahmeunsicherheit daher erheblich. Andererseits würde sie die tatsächlichen Verhältnisse der Gewebe im Anwendungsgebiet der Diagnostik besser widerspiegeln als Proben, die von homogen wirkenden Bereichen des Tumors entnommen werden. Um die Unsicherheit beim Übertragen der detaillierten histologischen Referenz auf die Messdaten zu verringern, sollten möglichst kleine Proben präpariert werden.

Ein Mess- und Präparationsprotokoll für frische Proben könnte wie folgt aussehen: Eine Probe wird entnommen, geteilt und sofort gemessen. Die Probe sollte dabei Körpertemperatur haben. Viele tierische Fette erstarren bereits bei Raumtemperatur. Wenn Lipide kristallisieren, verändern sich auch die Raman-Spektren mit dem Aggregatzustand. Die einzelnen Teilproben werden direkt nach Abschluss der Raman-Messung in Formalin fixiert und der normalen histologischen Bearbeitung unterzogen. Alternativ könnten die gemessenen Proben auch schockgefroren werden. Da die Raman-Messung bereits abgeschlossen ist, ist Gefriermedium unproblematisch.

Eine wichtige Frage beim Übergang zu Messungen *in vivo* ist, ob sich *ex-vivo*- und *in-vivo*-Messungen in einem Modell ergänzen können, oder ob ein ganz neues Modell für die *in vivo* gemessenen Daten erforderlich ist. Ein erster Hinweis ließe sich zum Beispiel aus den Vorhersagen der hier vorgestellten Modelle für frische Proben ableiten. Theoretisch sollte die Intensitätskalibrierung einen Transfer der Modelle von einer Sonde auf ei-

ne neue Sonde erlauben. Praktisch sind die Ergebnisse von Kalibrierungstransfers jedoch oft nicht zufriedenstellend (siehe unten). Daher sollten die Messungen an frischen Proben mit einem Aufbau durchgeführt werden, der auch *in vivo* verwendet werden kann.

**Spektrometer und Ramansonde:** Neurochirurgen operieren auf etwa 1 mm genau. Damit ist die laterale Auflösung der hier verwendeten Sonde wesentlich höher als notwendig. Mit einer auf diese Anwendung optimierten Sonde könnte also die Anregungsleistung auf noch wesentlich größere Gewebevolumina verteilt werden, so dass mit insgesamt mehr Anregungsleistung in kürzerer Zeit gleich gute oder bessere Spektren aufgenommen werden. Die hier vorgestellten chemometrischen Modelle nutzen eine spektrale Auflösung von  $10 \text{ cm}^{-1}$ , während das Spektrometer mit einer Auflösung von  $4 \text{ cm}^{-1}$  arbeitet. Geringere spektrale Auflösungen erlauben größere Spaltbreiten im Spektrometer und so lichtstärkere Spektrometerkonfigurationen. Auf diese Weise könnte die erforderliche Messzeit pro Spektrum weiter verringert werden. Bergholt *et al.* [279] beschreiben ein endoskopisches Raman-System zur Aufnahme von Spektren von Speiseröhren *in vivo*. Dabei wird dasselbe Spektrometer (incl. CCD) wie in der vorliegenden Arbeit genutzt. Bei einer spektralen Auflösung von  $9 \text{ cm}^{-1}$  wurden die verwendeten Spektren mit einer Messzeit von 500 ms bei ca. 30 mW Laserleistung auf der Probe aufgenommen.

**Halbüberwachte Modellbildung:** Gewebeproben sind oft viel einfacher erhältlich als ihre Referenzdiagnose. Sogenannte halbüberwachte (engl. *semi-supervised*) Verfahren können neben Trainingspektren mit Referenzinformation zusätzlich Spektren ohne Referenzinformation nutzen. Ist eine gewisse Datenbasis an gelabelten Spektren vorhanden, so können zusätzlich ungelabelte Spektren bei der Modellbildung helfen. Wichtig ist, dass die Spektren in Bezug auf die Klassifikationsaufgabe tatsächlich Information beitragen [280]. Der Erfolg dieser Strategie hängt sehr stark davon ab, ob die Annahmen über die Struktur der Daten stimmen oder nicht [281]. Wenn die Modellannahmen stimmen, können die ungelabelten Trainingsdaten wertvolle Informationen zum Modell beitragen. Abweichungen zwischen Modellannahmen und der Verteilung der Daten können jedoch dazu führen, dass die ungelabelten Daten die Modelle verschlechtern [282, 283]. Berget und Næs [276] weisen darauf hin, dass die Verteilung gelabelter Trainingsdaten oft von der der einfacher erhältlichen ungelabelten Daten abweicht, zum Beispiel, weil Routinemessungen eines (industriellen) Prozesses verrauschter sind oder auch mehr Ausreißer enthalten als sorgfältig präparierte und gemessene Trainingsproben.

Xu, White und Schuurmans [284] zerlegen die Residuen bei halbüberwachter Klassifikation in Anteile, die nur von den Spektren beziehungsweise sowohl von den Spektren als auch von den Referenzinformationen abhängen. Die ungelabelten Spektren helfen, ersteren Anteil an den Residuen zu verringern, nicht aber letzteren. Klassifikationsmodelle können daher nur in gewissen Grenzen durch ungelabelte Daten verbessert werden.

Die sogenannte *Cluster-Annahme* ist eine der wichtigsten grundlegenden Annahmen von halbüberwachten Klassifikationsverfahren. Sie besagt, dass die Klassengrenzen mit niedrigen Probendichten einhergehen [280]. Das ist beim Astrozytomgrading nicht der Fall, wohl aber bei der Unterscheidung zwischen Astrozytomen und Lymphomen. Für das Astrozytomgrading würden dementsprechend eher Ansätze für eine halbüberwachte Regression benötigt, während halbüberwachte Ansätze bei der Differentialdiagnostik

zwischen Astrozytomen und Lymphomen helfen könnten.

Lloyd *et al.* [188] nutzen halbüberwachte Modelle für eine Raman-spektroskopische Diagnostik von Speiseröhrenkrebs. Dabei gehen Proben, die von mehreren Pathologen übereinstimmend bewertet wurden, als gelabelte Trainingsproben und Proben, für die die Pathologen keinen Konsens erreichen konnten, als ungelabelte Trainingsproben in das Modell ein. Leider gehen die Grenzfälle nicht in die Validierung des Modells ein. Für die Testproben mit Konsens-Diagnose wird die Vorhersage jedoch verbessert, wenn nicht zu viele ungelabelte Trainingsdaten verwendet werden. Im größten Datensatz sind die ungelabelten Daten zwischen den gelabelten Daten der beiden Klassen konzentriert. Das legt nahe, dass die ungelabelten Daten nicht derselben Verteilung folgen wie die gelabelten Daten (vgl. [276] und Abb. 16.2) und die Cluster-Annahme verletzt ist.

Für die Validierung von halbüberwachten Modellen werden aber weiterhin gelabelte Testproben benötigt, die zudem für die Anwendung repräsentativ sein müssen. Da gegenwärtig für die diagnostischen Fragestellungen der Biospektroskopie die Anzahl der Testpatienten die kritischere Größe ist, können halbüberwachte Ansätze gegenwärtig nur in beschränktem Maße helfen.

Die Variation der Spektren von Patient zu Patient ist eine wichtige Varianzquelle bei der biospektroskopischen Diagnostik. Daher könnte die intraoperative Diagnostik insbesondere von Spektren des jeweiligen Patienten profitieren, die während der Operation von Geweben aufgenommen werden. Der Neurochirurg könnte Spektren von Geweben aufnehmen, für die zwar keine histologische Diagnose vorliegt, die aber entweder ohne Referenzdiagnose oder mit der jeweiligen Einschätzung des Chirurgen in das Modell einbezogen werden. Auch eine Aggregation der Vorhersagen von überwachten, patientenübergreifenden Modellen und unüberwachten oder halbüberwachten Modellen für den jeweiligen Patienten ist denkbar.

Weitergedacht könnte dieser Ansatz soweit verfolgt werden, dass ein patientenspezifisches Modell ausschließlich aus den vom aktuellen Patienten gemessenen Spektren gebildet wird. Ein solcher Ansatz wäre unabhängig von der Frage der Übertragbarkeit von Modellen auf andere Geräte (siehe unten). Dazu wären allerdings Modelle erforderlich, die auf eine Achse steigender Malignität projizieren, die dann mit den Patientendaten angepasst wird. Das bedeutet auch, dass die als bekannt einstuftbar gemessenen Spektren tatsächlich sicher richtig vom Chirurgen eingeschätzt werden müssen. Außerdem muss (patienten- und geräteübergreifend) geklärt werden, wo genau die angestrebte Grenze zwischen den einzelnen Gewebearten verläuft. Für die Differentialdiagnostik zwischen Lymphomen und Astrozytomen ist eine Anpassung der Modelle an den Patienten wohl nur bezüglich des normalen Gewebes möglich, da ja nicht eine Tumorgrenze innerhalb des Patienten gefunden werden soll, sondern die Art des Tumors. Beispielspektren für andere Tumore können praktisch immer nur von anderen Patienten kommen, so dass diese Fragestellung grundsätzlich patientenübergreifend ist.

**Testdaten und Validierung:** Auch wenn die Modellbildung von Spektren ohne Referenzinformation profitieren kann und durch falsch gelabelte Spektren nicht oder nur geringfügig beeinträchtigt wird, benötigt die Modellvalidierung Testdaten mit genauen und richtigen Referenzinformationen. Daher sind die Anforderungen an Testdaten letztlich höher als an Trainingsdaten. Zur Zeit wird die Genauigkeit der Validierungsergebnisse

durch die geringe Anzahl an Testpatienten stark beeinträchtigt (Kap. 5.2 und [CB3]). Zum Einen ist es also unumgänglich, die Probenbasis auf hinreichende Anzahlen an Testpatienten zu erweitern. Zum Anderen könnten aber die in dieser Arbeit entwickelten weichen Kenngrößen für die Modellqualität auch bei eindeutigen (harten) Referenzdaten die benötigte Anzahl an Testpatienten reduzieren, weil sie eine geringere Varianzunsicherheit aufweisen [CB1]. Hierzu sind weitere Erfahrungen mit den neu eingeführten Kenngrößen notwendig.

Modellvorhersagen sind in aller Regel eine Extrapolation: Ein einmal gebildetes Modell soll *in der Zukunft* angewendet werden. Im Rahmen der Validierung muss daher die Leistungsfähigkeit des Modells bei der Vorhersage nicht nur *neuer*, sondern auch *zukünftiger, neuer* Proben gemessen werden (Kap. 4.8.1, S. 42). Diese Problematik der Drift ist in der industriellen Analytik bekannt. An Testproben, die zwischen den Analysenmessungen bestimmt werden, wird die aktuelle Genauigkeit der Modellvorhersagen überprüft. Da die Testproben immer auch als Trainingsproben verwendet werden können, können die Testdatensätze *nach abgeschlossener* Validierung auch zum Anpassen des Modells verwendet werden (engl. *model update*), so dass die Probenbasis des Modells kontinuierlich wächst. Eine Validierung im Hinblick darauf, wie lange einmal gebildete Kalibriermodelle ohne erneute Kalibrierung angewendet werden können, erfordert Testdaten, die in entsprechendem zeitlichen Abstand nach den Trainingsproben gemessen werden [208]. Es ist sinnvoll, dies als eigene Validierungsstudie eines bereits gebildeten Modells durchzuführen. Dann kann die tatsächliche Unabhängigkeit der Validierung sehr einfach garantiert werden. Eine solche Validierungsstudie kann auch verblindet durchgeführt werden, indem weder derjenige, der Spektren misst und auswertet, noch der Pathologe die Ergebnisse des jeweils anderen kennen. Die beiden Vorhersagen werden dann erst in einer letzten Stufe zusammengeführt. Aufgrund des großen Aufwandes (notwendige Patientenzahlen) ist das jedoch erst dann effektiv, wenn die Instrumentierung für Operationsbedingungen optimiert ist und das Modell bereits im Rahmen einer Resampling-Validierung eine für die Praxis ausreichende Leistungsfähigkeit gezeigt hat.

**Übertragbarkeit der Modelle (engl. *calibration transfer*):** Ein verwandtes, ebenfalls noch ungelöstes [285, 286] Problem ist die Übertragung von chemometrischen Modellen auf andere Raman-Spektrometer bzw. die Verwendbarkeit von chemometrischen Modellen nach dem Austausch einzelner Komponenten. Dazu muss das Gerät sowohl in der spektralen (Wellenzahl-) Achse als auch in der Intensität kalibriert sein. Insbesondere die Kalibrierung der spektralen Achse muss sehr genau sein, da Methoden wie die PLS-Regression gegebenenfalls auf Unterschiede aufgrund einer Dejustierung um Bruchteile der spektralen Auflösung reagieren. Aktuelle Publikationen zum Transfer von chemometrischen Modellen behandeln zum Beispiel Mischungen dreier bekannter Komponenten [287] oder die Nutzung von Spektrenbibliotheken zur Substanzidentifikation [288]. Die Übertragung von Modellen auf wesentlich andere Anregungswellenlängen ist naturgemäß aufwändig (vgl. S. 15), selbst wenn die Intensitäten nur der  $\tilde{\nu}^4$ -Abhängigkeit unterliegen und keine Resonanzverstärkung auftritt [289].

Die beschriebenen Ansätze betrachten die Unterschiede in der Kalibrierung zwischen verschiedenen Messgeräten als mess- und korrigierbar, also als systematische Abweichung. Betrachtet man hingegen eine Vielzahl an (baugleichen) Messgeräten, dann kön-



nen deren Unterschiede als Streuung modelliert werden. Damit ist möglicherweise ein Kompromiss zwischen einer praktikabel durchführbaren Kalibrierung der Geräte und dem Einbeziehen der verbleibenden Unsicherheit in die Modelle eine gangbare Lösung. In dieser Sichtweise kann ein chemometrisches Modell an ein Messgerät überangepasst sein, und diese Überanpassung sollte vermieden werden.

**Modelloptimierung:** Bei den hier vorgestellten Modellen wurde keinerlei datengetriebene Optimierung vorgenommen. Datengesteuerte Modelloptimierung bedeutet, dass die Trainingsdaten weiter in „innere“ Trainings- und Testdaten geteilt werden müssen. Das war hier nicht sinnvoll. Die entsprechenden Datensätze wären zu klein geworden, als dass die erwarteten Unterschiede auch nur annähernd sicher hätten erkannt werden können. Systematische Empfehlungen zur Vorbehandlung chemometrischer Datensätze sind Gegenstand aktueller Forschung [290]. Die Optimierung besonders von Klassifikationsmodellen ist ein ungelöstes Problem [144]. Die im Rahmen dieser Arbeit entwickelten Sensitivitäts- und Spezifitätsmaße auf Basis des wMSE könnten nicht zuletzt aufgrund der niedrigeren Varianz einen wertvollen Beitrag zur Lösung dieser Probleme bieten.

Gegenwärtig werden zur Modelloptimierung oft Suchstrategien aus der numerischen Optimierung verwendet, zum Beispiel genetische Algorithmen [CB7, 16, 291] oder auch Rastersuche [112]. Diese Strategien nutzen in der Regel die maximale beobachtete Modellqualität und berücksichtigen weder die systematische (typisch: steigender optimistischer Bias mit steigender Komplexität der Modelle) noch die zufällige Unsicherheit beim Messen der jeweiligen Modellqualität, was zu erheblichen Problemen bei der Optimierung führen kann [CB7, 144]. Eine ähnliche Situation ist in der chemischen Prozessoptimierung häufig: auch dort sind die Messwerte des Zielfunktional (z. B. Ausbeute) mit einer Varianzunsicherheit behaftet. Deshalb werden dort zum Beispiel *response-surface*-Methoden [286, 292] zusammen mit den entsprechenden experimentellen Designs [286, 293] angewendet, die das Zielfunktional aus verrauschten Messdaten als glatte Funktion der Einflussgrößen modellieren. Für die Optimierung von chemometrischen Modellen müssen weitere Besonderheiten berücksichtigt werden, insbesondere die steigende Varianzunsicherheit mit steigender Modellkomplexität.

Die hier vorgestellten Modelle benutzen implizit verschiedene Hyperparameter oder können bei ausreichenden Trainingspatientenzahlen in verschiedene Richtungen erweitert werden. Dazu ist aber eine geeignete Kenngröße zur Beurteilung der Modellqualität erforderlich. Hier ist der Schritt von den beschriebenen Sensitivitäten und Spezifitäten hin zu prädiktiven Werten unumgänglich. Sensitivität und Spezifität beschreiben, wie häufig Gewebe, die tatsächlich der betrachteten Klasse angehören, als solche erkannt werden. In der Anwendung einer medizinischen Diagnostik ist aber wichtiger, wie häufig die getroffenen Vorhersagen tatsächlich richtig sind. Diese Frage kann nur beantwortet werden, wenn bekannt ist, wie häufig die Diagnostik auf welches Gewebe angewendet wird. Fehleinschätzungen der tatsächlichen Prävalenz führen zu erheblichen Fehleinschätzungen bezüglich der Aussagekraft von diagnostischen Tests [294, 295]. Letztlich muss der Chirurg einen Arbeitspunkt auf der Kurve aus positivem und negativem prädiktiven Wert wählen, die als Analogon zur Spezifitäts-Sensitivitäts-Kurve die Prävalenz der verschiedenen Gewebe berücksichtigt. In diese Wahl wird auch einfließen, wie schwerwiegend die einzelnen Verwechslungen zwischen den Geweben sind.

Die aggregierten Modelle zur Unterscheidung von Lymphomen und Astrozytomen benutzen implizit einen weiteren Hyperparameter, den Mindestanteil der Klasse „normal“, der in der ersten Klassifikationsstufe zum Ausschluss der Spektren führt. Außerdem bietet es sich an, die Differentialdiagnose erst ab einer Mindestzahl und/oder einem Mindestprozentsatz an in der ersten Stufe akzeptierten Spektren zu stellen. Alle diese Hyperparameter sollten letztlich einer Optimierung unterzogen werden.

**Biochemie der untersuchten Tumore:** Schwingungsspektroskopie kann wichtige Einblicke in die krankheitsbedingten biochemischen Veränderungen von Geweben liefern. Die Experimente für solche Untersuchungen unterscheiden sich allerdings von denen in dieser Arbeit grundsätzlich: Eine operationsbegleitende Diagnostik soll mit kürzestmöglichen Messzeiten, also sehr niedrigem Signal-Rausch-Verhältnis, eine Einordnung des Gewebes erreichen. Demgegenüber zielt das Studium krankheitsbedingter biochemischer Veränderungen auf Detailinformationen in den Spektren, die deshalb mit wesentlich besserem Signal-Rausch-Verhältnis aufgenommen werden sollten. Solche Studien sind in der Literatur mehrfach beschrieben (Kap. 3.2). Dennoch ist bislang zum Beispiel unbekannt, welche Untergruppen die Glioblastome aus schwingungsspektroskopischer Sicht bilden. Klinisch und histologisch sind verschiedene Untergruppen bekannt, zum Beispiel genetische Variationen bei Glioblastomen [50].

Für diese Fragestellungen bietet sich auch die Kombination der Schwingungsspektroskopie mit anderen Methoden, besonders der Massenspektrometrie, an. Die durch Schwingungsspektroskopie und Massenspektrometrie zugänglichen Informationen ergänzen sich insofern hervorragend, als Summenparameter wie zum Beispiel Gesamtlipid- oder Gesamtproteingehalt, aber auch vorherrschende Proteinfaltung, einfach über die Schwingungsspektren zugänglich sind. Demgegenüber erlaubt die Massenspektrometrie, gezielt bestimmte Moleküle wie zum Beispiel einzelne Proteine zu verfolgen, während die Summenparameter nicht oder nur sehr schwer aus den Massenspektren abgeleitet werden können. Eine erste Studie, die Raman-Spektroskopie und Massenspektrometrie an Mäusehirn vergleicht, ist [296].

Ein andere interessante Frage ist, ob bzw. inwieweit Gewebe eines höhergradigen Tumors, die morphologisch niedriggradig erscheinen, auch in den Schwingungsspektren niedriggradigen Tumoren ähneln. Eng damit verwandt ist die Frage, in wie weit sich morphologisch normales Gewebe in der Nähe des Tumors einerseits von niedriggradigem Tumorgewebe und andererseits von entfernteren normalen Geweben unterscheidet. In diesem Zusammenhang bietet sich auch der Vergleich mit den Spektren der Kontrollproben an, die ja erst nach dem Tod des Spenders gewonnen wurden.

Die extrem geringe Zahl von Lymphomproben (acht Patienten, davon drei Proben mit weit überwiegend gliotischem oder normalem Gewebe) erlaubt zur Zeit nur sehr vorsichtige Aussagen über den Unterschied zwischen Astrozytomen und Lymphomen. Für die Differentialdiagnostik zwischen Lymphomen und Astrozytomen muss daher die Messung weiterer Lymphomproben an erster Stelle stehen. Dies sollte auch von Experimenten zur Untersuchung der biochemischen und schwingungsspektroskopischen Besonderheiten der Lymphome begleitet werden, da hierzu erst sehr wenig bekannt ist (Kap. 3.2).

## Danksagung

An dieser Stelle möchte ich mich ganz besonders bei Prof. Reiner Salzer bedanken. Sie haben mir diese hochinteressanten Themengebiete zur eigenständigen Bearbeitung überlassen. Ihre unermüdliche Fähigkeit, neue hochinteressante Fragestellungen zu entdecken und neue Möglichkeiten zu erahnen, hat mich immer wieder tief beeindruckt.

Genauso wichtig war und ist Prof. Jürgen Popp für die vorliegende Arbeit. Jürgen, Du hast dieser Arbeit mit Deinem „Asylangebot“, sie in Jena einzureichen, eine neue Perspektive gegeben.

Als Leiter des Projekts „molekulare Endospektroskopie“ der Volkswagen-Stiftung hat PD Dr. Christoph Krafft es mir überhaupt erst ermöglicht, die tumordiagnostischen Fragestellungen zu untersuchen: obwohl ich nie dem Projekt angehört habe, durfte ich die Probensammlung nutzen und alle benötigten Proben präparieren. Für einen Besuch in Deinem Büro sollte man genug Zeit mitbringen: Du hast immer viele interessante, neue Artikel, Proben, Fragestellungen, Erkenntnisse oder Ideen und gibst dieses Wissen auch gern weiter. Christoph, Du hast mir viel beigebracht über die Raman-Spektroskopie von Bioproben. Dafür und für die andauernde Zusammenarbeit möchte ich an dieser Stelle ganz herzlich danken. Zur FTIR-Spektrensammlung des Projekts haben viele Leute beigetragen. Insbesondere Dr. Wolfram Steller hat unermüdlich FTIR-Images aufgenommen. Ohne diese hätten die Publikationen [CB7] und [CB6] so nicht geschrieben werden können.

Das Stipendium der Deutschen Telekom-Stiftung war prägend und wichtig für diese Arbeit. Das Stipendium hat letztlich die experimentelle Seite der Arbeit auf die Raman-Spektroskopie mit einer faseroptischen Sonde festgelegt. Die Stipendiaten-Treffen habe ich in guter Erinnerung: durch die Workshops und auch als gute Gelegenheit, über den wissenschaftlichen Tellerrand hinauszuschauen, hochkarätige Vorträge über andere Fachgebiete zu hören und mit den Kollegen aus den anderen Fächern ungezwungen in Kontakt zu kommen. Ohne die finanzielle Unterstützung durch das Stipendium hätte ich viele Konferenzen nicht besuchen können. Auch die Besuche bei der Arbeitsgruppe von Prof. Hamprecht am Institut für wissenschaftliches Rechnen in Heidelberg, die Teilnahme an Forschungsseminaren dieser Arbeitsgruppe und der Forschungsaufenthalt bei Prof. Sutherland in der Neurochirurgie des Foothills Hospital der Universität Calgary/Kanada wäre ohne die mit dem Stipendium verbundenen Reisemittel nicht möglich gewesen.

Manchmal ist es gut, weit weg zu fahren, um Antworten auf Fragen zu finden, die es auch in der Nähe gegeben hätte: I learned a lot not only about the brain but also about the practical needs for diagnostic tools for intraoperative use in neurosurgery from Prof. Sutherland. I'm very grateful for the opportunity to visit a surgery during which an intraoperative MRI was taken. The possibility to actually witness these procedures helped me tremendously in understanding not only the general need for better tools but also to have a better grasp for practical issues. Considering the busy neurosurgery work-schedule, I'd like to thank you the more for taking the time to explain and discuss neurosurgeons'

needs and to understand what potential Raman spectroscopy holds.

Auch in Dresden durfte ich eine neurochirurgische Operation besuchen. Noch viel wichtiger als der Besuch einer einzelnen Operation ist aber, dass Prof. Gabriele Schackert, Dr. Matthias Kirsch und Dr. Stephan Sobottka die Tumorprobensammlung jahrelang unermüdlich mit neuen Proben und uns mit Wissen über die Hirntumore versorgt haben.

Ich möchte mich auch ganz herzlich bei der Neuropathologin PD Dr. Kathrin Geiger bedanken. Sie haben mir nicht nur die Möglichkeit gegeben, in ihrem Institut am Gefriermikrotom meine Proben zu präparieren. Viel mehr haben Sie sich – oft vor dem offiziellen Arbeitsbeginn, weil tagsüber keine Zeit für sowas ist – viel Zeit für die Proben genommen. Ich habe viel gelernt, über die Tumore, die sich oft nicht an das halten, was die WHO an Graden festlegt; über die Präparation von Gefrierschnitten, die Lagerung von Proben und auch wie man Proben manchmal doch noch retten kann.

Bei einer Reihe von Studenten möchte ich mich für ihre Unterstützung bedanken: im Rahmen von Forschungspraktika haben Benjamin Schumm, Andre Mirtschink, Mario Leonhardt und Martin Kammer für verschiedene Proben von Si bis hin zu Hirngewebe vom Schwein untersucht: bis aus welcher Tiefe das Signal stammt, wie tief unter der Probenoberfläche der Fokus liegen sollte, welche Schärfentiefe und welche laterale Auflösung die Sonde und verschiedene Mikroskopobjektive haben. Martin Kammer hat die auf CaF<sub>2</sub> präparierten Parallelschnitte zu den hier vorgestellten Bulkproben im Rahmen seiner Diplomarbeit untersucht [135], während Mario Leonhardt angefangen hat, die räumliche Heterogenität der Proben zu quantifizieren. Romy Marx hat im Rahmen ihres Forschungspraktikums bei der Durchführung der hier vorgestellten Messungen mitgewirkt. Tobias Schulz und Ben Schüppel haben, ebenfalls als Forschungspraktikanten, die von Dr. Geiger erstellten detaillierten Referenzdiagnosen auf die Messpunkte übertragen. Ohne sie hätte ich keine verblindete Referenz gehabt.

Frau Herzog hat die Datenbank mit vielen Daten per Hand gefüttert und unermüdlich Patientendaten auf papierenen Kopien der Patientenbriefe geschwärzt, bis Friederike Schlemmer die Informationsbeschaffung revolutioniert und uns mit digitalen Patienten- und Diagnosebriefen versorgt hat. Auch Frank Drescher und Gennadi Gudi haben mitgeholfen, dass wir bei über 1800 Proben nicht den Überblick verlieren. Gennadi hat mich in die Kunst der Probenpräparation am Gefriermikrotom eingeführt, ein paar weitere Tricks habe ich später bei Nicole Peukert und Angelo Città gelernt. Matthias Gestrich aus der Werkstatt danke ich für seine tatkräftige Hilfe, aber auch für Material und das Ausborgen von Werkzeug. Ein sozusagen literarischer Dank gilt Benita Göbel, die mich mit viel Literatur versorgt, die ich sonst nur sehr schwer bekommen könnte.

Prof. Valter Sergio gilt in Bezug auf diese Arbeit ein besonderer Dank. Obwohl meine Arbeit an den Materialwissenschaften der Universität Triest *nichts* mit Klassifikationsmodellen zu tun gehabt hat, hat er mir doch immer wieder die Freiheit gegeben, über die weichen Kenngrößen nachzudenken. Nur so konnte aus der Lösung für das unmittelbar anstehende Problem, wie ich die Qualität meiner Klassifikationsmodelle für die weichen Proben messen kann, das hier vorgestellte, sehr viel allgemeinere Konzept wachsen. Valter, es ist mir eine Ehre, Dich zu kennen: nur sehr wenige Wissenschaftler sind klar und deutlich integer in ihrer Forschung wie Du und thematisieren ethische Aspekte und Konsequenzen. Nur sehr wenige Chefs vertrauen ihren Leuten so wie Du. Danke!

Ich möchte mich auch bei allen Kollegen, Freunden und Bekannten bedanken, die diese Arbeit korrekturgelesen haben. Besonders hervorheben möchte ich an dieser Stelle

außerdem Dr. Ute Neugebauer, Dr. Christian Matthäus und Friederike Schlemmer. Ute und Christian, eure Hinweise waren immer auf den Punkt gebracht und hilfreich. Ute, bei Dir möchte ich mich auch nochmal für den Datensatz, den ich in [CB3] verwendet habe, sehr bedanken. Friederike, Deine Hinweise, Erklärungen (und Dein Bücherregal) zu medizinischen Aspekten waren für mich unverzichtbar. Und wen sonst hätte ich zu  $\LaTeX$  und TikZ fragen sollen?

Irgendwann liegen die Nerven blank: ich möchte mich deshalb bei allen bedanken, die mir immer mal wieder Mut zugesprochen oder auch Wogen geglättet haben. Ich möchte mich auch bei allen Kollegen und Freunden bedanken, die mir nicht nur beigebracht haben, wie man in der Forschung (über)lebt, sondern viel wichtiger: wieviel Spaß Wissenschaft macht. Ich arbeite gern mit euch zusammen. Ute, Christian, Michael, Clara, Kokila, Jan, Iwan, Marcel, Isa, Sara, Sebastian, Tania, Katharina, Norbert, Larysa, Gennadi, Renate, Alois, Daniela, Mike, Leo, Friederike, Wolfram, Thomas, Thomas, Thomas *et al.*: Danke!

Die wichtigen Fundamente werden natürlich viel früher gelegt. Ich möchte mich deshalb bei meiner Familie bedanken. Ohne die generelle Atmosphäre von der Überschlagsrechnung zur Frage „Kann das überhaupt stimmen?“ oder ohne „brockeln“<sup>(a)</sup>, ohne Pipetten und Mikroskop im Haus, Sonnenlinien betrachten im Spektroskop und Leuchtstoffröhren, die sowohl mit Kurzwelle als auch mit Weidezaun leuchten: ohne diese oder ähnliche Ausprägungen der Neugier, der Experimentierfreudigkeit und des handwerklichen Könnens, des gesunden Menschenverstands und der nötigen Skepsis, ist so eine Arbeit überhaupt denkbar?

---

<sup>(a)</sup> analog zu „googeln“: suche im Lexikon



**Teil V**  
**Anhang**

# A Protokolle zur Probenpräparation

## A.1 Präparationsprotokoll Gefrierschnitte

**Temperatur:** Kopf- und Kammertemperatur jeweils  $-20^{\circ}\text{C}$ .

**Durchführung:**

1. Größere Proben gefroren teilen: für die weiteren Arbeiten wurde ein etwa 3 mm dickes Stück verwendet.
2. 3 mm-Probe auftauen und auf Probenhalter des Gefriermikrotoms schockfrieren ( $\text{N}_{2\text{fl}}$ )
3. Gegebenenfalls anfertigen von  $14\ \mu\text{m}$  dicken Schnitten auf  $\text{CaF}_2$
4. Anfertigen von  $7\ \mu\text{m}$  dicken Schnitten auf Glas für die Referenzdiagnose
5. Ablösen der verbleibenden Probe (gefroren) und Lagerung bei  $-80^{\circ}\text{C}$

## A.2 Fixierung der Dünnschnitte

**Lösung:** Ethanolische Formalin-Lösung zur Fixierung

4 % Formalin  
50 % Ethanol  
46 % Wasser

**Durchführung:** Trockene Schnitte 30 s fixieren.

## A.3 Färbung mit Methylenblau

nach [297, Seite 558]

**Lösung:** LÖFFLERS Methylenblau-Lösung

30 ml Methylenblau gesättigt in Ethanol  
1 ml KOH 1 %  
99 ml  $\text{H}_2\text{O}$

**Durchführung:** Fixierte Schnitte 15 s färben, 5 s abtropfen lassen. Wässern.



## B Details zu den Modellen

### B.1 Details zum Astrozytomgrading

#### B.1.1 Tabellen Astrozytomgrading

Kenngröße	Klasse	astro/LR-soft					wMAE	wRMSE	wRMAE	wMSE	
		(hard)	(crisp)	weak	prd	strong					
Sensitivität	N	$\bar{x}$	0.758	0.682	0.626	0.586	0.548	0.649	0.567	0.408	0.813
		$s(x)$	0.023	0.017	0.012	0.012	0.012	0.012	0.012	0.010	0.010
		0 %	0.676	0.629	0.578	0.540	0.500	0.602	0.524	0.369	0.773
		5 %	0.717	0.654	0.607	0.568	0.529	0.630	0.547	0.392	0.795
		50 %	0.756	0.681	0.625	0.585	0.546	0.649	0.567	0.407	0.813
		95 %	0.798	0.709	0.644	0.604	0.565	0.666	0.586	0.422	0.829
		100 %	0.823	0.724	0.658	0.617	0.578	0.680	0.601	0.435	0.841
	A°II	$\bar{x}$	0.438	0.455	0.562	0.424	0.316	0.579	0.504	0.351	0.754
		$s(x)$	0.029	0.016	0.011	0.010	0.009	0.009	0.009	0.007	0.009
		0 %	0.363	0.415	0.528	0.396	0.291	0.557	0.480	0.335	0.730
		5 %	0.385	0.429	0.543	0.407	0.300	0.563	0.488	0.339	0.737
		50 %	0.440	0.456	0.562	0.423	0.316	0.580	0.505	0.352	0.755
		95 %	0.483	0.480	0.579	0.439	0.330	0.593	0.518	0.362	0.767
		100 %	0.494	0.485	0.583	0.441	0.334	0.601	0.525	0.368	0.775
	A°III+	$\bar{x}$	0.709	0.624	0.622	0.560	0.511	0.649	0.568	0.408	0.814
		$s(x)$	0.012	0.009	0.008	0.007	0.008	0.007	0.007	0.006	0.006
		0 %	0.683	0.605	0.603	0.544	0.494	0.630	0.552	0.392	0.800
		5 %	0.689	0.610	0.610	0.548	0.498	0.639	0.558	0.399	0.805
50 %		0.710	0.624	0.623	0.560	0.512	0.650	0.568	0.408	0.814	
95 %		0.728	0.638	0.635	0.572	0.523	0.660	0.579	0.417	0.823	
100 %		0.737	0.647	0.641	0.580	0.532	0.668	0.586	0.424	0.829	
Spezifität	N	$\bar{x}$	0.878	0.835	0.844	0.822	0.802	0.832	0.714	0.590	0.918
		$s(x)$	0.006	0.006	0.004	0.004	0.004	0.004	0.007	0.005	0.004
		0 %	0.859	0.817	0.829	0.808	0.788	0.818	0.691	0.573	0.905
		5 %	0.866	0.825	0.836	0.815	0.793	0.824	0.702	0.580	0.911
		50 %	0.878	0.835	0.844	0.822	0.802	0.832	0.715	0.590	0.919
		95 %	0.886	0.842	0.850	0.828	0.807	0.838	0.722	0.597	0.923
		100 %	0.900	0.853	0.854	0.833	0.811	0.841	0.729	0.601	0.926
	A°II	$\bar{x}$	0.866	0.746	0.740	0.698	0.665	0.735	0.664	0.485	0.887
		$s(x)$	0.010	0.008	0.006	0.006	0.006	0.006	0.006	0.006	0.004
		0 %	0.842	0.727	0.725	0.682	0.648	0.720	0.649	0.471	0.877
		5 %	0.852	0.733	0.730	0.688	0.654	0.726	0.653	0.476	0.880
		50 %	0.865	0.745	0.741	0.698	0.665	0.735	0.664	0.486	0.887
		95 %	0.883	0.760	0.751	0.708	0.674	0.744	0.674	0.494	0.894
		100 %	0.891	0.767	0.756	0.714	0.681	0.750	0.679	0.500	0.897

## B Details zu den Modellen

Kenngröße	Klasse		astro/LR-soft					wMAE	wRMSE	wRMAE	wMSE
			(hard)	(crisp)	weak	prd	strong				
	A°III+	$\bar{x}$	0.930	0.852	0.845	0.802	0.767	0.826	0.746	0.583	0.935
		$s(x)$	0.007	0.006	0.005	0.005	0.005	0.005	0.006	0.006	0.003
		0 %	0.904	0.829	0.831	0.787	0.752	0.813	0.729	0.568	0.927
		5 %	0.916	0.842	0.836	0.792	0.758	0.817	0.733	0.573	0.929
		50 %	0.931	0.854	0.846	0.802	0.767	0.827	0.747	0.584	0.936
		95 %	0.940	0.860	0.852	0.808	0.774	0.833	0.754	0.591	0.940
		100 %	0.942	0.867	0.858	0.812	0.778	0.838	0.762	0.598	0.943

Kenngröße	Klasse		astro/LDA					wMAE	wRMSE	wRMAE	wMSE
			(hard)	(crisp)	weak	prd	strong				
Sensitivität	N	$\bar{x}$	0.723	0.691	0.594	0.563	0.532	0.627	0.509	0.389	0.758
		$s(x)$	0.035	0.029	0.020	0.020	0.020	0.019	0.019	0.016	0.018
		0 %	0.626	0.614	0.541	0.510	0.479	0.577	0.461	0.349	0.710
		5 %	0.660	0.639	0.556	0.526	0.496	0.592	0.475	0.362	0.725
		50 %	0.724	0.693	0.595	0.564	0.533	0.628	0.509	0.390	0.759
		95 %	0.779	0.736	0.625	0.594	0.564	0.658	0.538	0.415	0.787
		100 %	0.817	0.763	0.644	0.613	0.582	0.678	0.556	0.432	0.803
	A°II	$\bar{x}$	0.498	0.490	0.589	0.488	0.403	0.594	0.490	0.363	0.740
		$s(x)$	0.032	0.025	0.016	0.015	0.015	0.014	0.015	0.011	0.015
		0 %	0.386	0.393	0.544	0.444	0.357	0.552	0.447	0.331	0.695
		5 %	0.439	0.442	0.559	0.459	0.375	0.568	0.465	0.343	0.713
		50 %	0.503	0.493	0.590	0.490	0.404	0.596	0.492	0.364	0.742
		95 %	0.541	0.524	0.611	0.508	0.424	0.614	0.511	0.378	0.761
		100 %	0.566	0.543	0.614	0.514	0.429	0.623	0.522	0.386	0.771
	A°III+	$\bar{x}$	0.730	0.698	0.651	0.613	0.582	0.676	0.547	0.431	0.794
		$s(x)$	0.017	0.015	0.012	0.011	0.012	0.011	0.010	0.009	0.009
		0 %	0.641	0.627	0.610	0.571	0.540	0.636	0.512	0.397	0.762
		5 %	0.706	0.675	0.631	0.594	0.563	0.660	0.531	0.417	0.780
50 %		0.732	0.699	0.651	0.613	0.583	0.676	0.547	0.431	0.795	
95 %		0.755	0.721	0.671	0.632	0.602	0.694	0.564	0.446	0.810	
100 %		0.770	0.733	0.678	0.639	0.610	0.700	0.569	0.452	0.814	
Spezifität	N	$\bar{x}$	0.875	0.853	0.873	0.857	0.840	0.856	0.712	0.620	0.917
		$s(x)$	0.010	0.009	0.006	0.006	0.006	0.006	0.010	0.007	0.006
		0 %	0.846	0.822	0.851	0.837	0.821	0.837	0.681	0.597	0.898
		5 %	0.857	0.835	0.862	0.846	0.829	0.844	0.693	0.606	0.905
		50 %	0.876	0.854	0.874	0.858	0.841	0.856	0.713	0.621	0.918
		95 %	0.889	0.866	0.882	0.864	0.848	0.863	0.726	0.630	0.925
		100 %	0.899	0.870	0.886	0.869	0.853	0.867	0.732	0.635	0.928
	A°II	$\bar{x}$	0.787	0.749	0.707	0.676	0.650	0.705	0.585	0.457	0.828
		$s(x)$	0.019	0.015	0.011	0.011	0.012	0.011	0.011	0.010	0.009
		0 %	0.726	0.701	0.673	0.643	0.617	0.671	0.545	0.427	0.793
		5 %	0.754	0.722	0.685	0.654	0.627	0.686	0.566	0.440	0.812
		50 %	0.790	0.750	0.708	0.677	0.651	0.705	0.586	0.457	0.829
		95 %	0.812	0.769	0.722	0.692	0.666	0.721	0.602	0.472	0.841
		100 %	0.833	0.791	0.733	0.703	0.678	0.729	0.610	0.480	0.848

## B.1 Details zum Astrozytomgrading

Kenngröße	Klasse	astro/LDA					wMAE	wRMSE	wRMAE	wMSE	
		(hard)	(crisp)	weak	prd	strong					
	A°III+	$\bar{x}$	0.923	0.900	0.868	0.841	0.819	0.851	0.724	0.614	0.924
		$s(x)$	0.009	0.008	0.007	0.007	0.007	0.007	0.010	0.008	0.006
		0 %	0.890	0.870	0.841	0.814	0.792	0.827	0.689	0.584	0.903
		5 %	0.907	0.883	0.853	0.827	0.805	0.837	0.704	0.597	0.912
		50 %	0.924	0.901	0.869	0.842	0.820	0.851	0.726	0.614	0.925
		95 %	0.935	0.912	0.879	0.851	0.829	0.861	0.739	0.627	0.932
		100 %	0.939	0.913	0.883	0.856	0.834	0.867	0.744	0.635	0.934

Kenngröße	Klasse	astro/LR-highwn					wMAE	wRMSE	wRMAE	wMSE	
		(hard)	(crisp)	weak	prd	strong					
Sensitivität	N	$\bar{x}$	0.469	0.488	0.521	0.472	0.423	0.541	0.497	0.323	0.747
		$s(x)$	0.020	0.009	0.006	0.006	0.007	0.006	0.006	0.004	0.006
		0 %	0.411	0.460	0.500	0.451	0.400	0.522	0.477	0.309	0.727
		5 %	0.435	0.471	0.509	0.461	0.412	0.530	0.486	0.315	0.736
		50 %	0.472	0.490	0.521	0.473	0.424	0.542	0.498	0.323	0.748
		95 %	0.497	0.501	0.530	0.481	0.432	0.550	0.506	0.329	0.756
		100 %	0.505	0.504	0.534	0.485	0.436	0.554	0.509	0.332	0.759
	A°II	$\bar{x}$	0.298	0.418	0.559	0.400	0.279	0.572	0.509	0.346	0.759
		$s(x)$	0.025	0.011	0.007	0.006	0.006	0.006	0.007	0.005	0.007
		0 %	0.225	0.390	0.536	0.382	0.259	0.554	0.490	0.332	0.740
		5 %	0.245	0.398	0.546	0.388	0.267	0.559	0.496	0.336	0.746
		50 %	0.301	0.420	0.559	0.400	0.279	0.572	0.510	0.346	0.760
		95 %	0.337	0.432	0.569	0.409	0.287	0.580	0.518	0.352	0.768
		100 %	0.343	0.437	0.573	0.414	0.294	0.583	0.522	0.354	0.771
A°III+	$\bar{x}$	0.632	0.542	0.558	0.494	0.439	0.597	0.540	0.365	0.789	
	$s(x)$	0.013	0.007	0.005	0.005	0.005	0.005	0.005	0.004	0.004	
	0 %	0.594	0.523	0.543	0.480	0.425	0.582	0.526	0.353	0.775	
	5 %	0.606	0.529	0.548	0.484	0.430	0.588	0.533	0.358	0.782	
	50 %	0.633	0.542	0.559	0.494	0.439	0.597	0.540	0.365	0.789	
	95 %	0.652	0.552	0.566	0.502	0.448	0.605	0.547	0.371	0.795	
	100 %	0.656	0.557	0.569	0.503	0.450	0.607	0.550	0.373	0.797	
Spezifität	N	$\bar{x}$	0.881	0.784	0.791	0.765	0.738	0.780	0.707	0.531	0.914
		$s(x)$	0.005	0.004	0.002	0.002	0.002	0.002	0.003	0.003	0.002
		0 %	0.866	0.775	0.786	0.759	0.732	0.775	0.697	0.526	0.908
		5 %	0.871	0.778	0.787	0.761	0.734	0.776	0.701	0.527	0.911
		50 %	0.881	0.784	0.791	0.765	0.738	0.780	0.707	0.531	0.914
		95 %	0.889	0.789	0.795	0.769	0.743	0.784	0.712	0.536	0.917
		100 %	0.891	0.792	0.797	0.771	0.744	0.785	0.714	0.537	0.918
	A°II	$\bar{x}$	0.896	0.680	0.717	0.669	0.631	0.714	0.676	0.465	0.895
		$s(x)$	0.008	0.005	0.004	0.004	0.004	0.003	0.004	0.003	0.002
		0 %	0.870	0.669	0.705	0.657	0.619	0.702	0.663	0.454	0.887
		5 %	0.881	0.672	0.711	0.663	0.625	0.708	0.671	0.460	0.892
		50 %	0.897	0.680	0.718	0.669	0.632	0.714	0.676	0.465	0.895
		95 %	0.907	0.687	0.724	0.675	0.638	0.719	0.682	0.470	0.899
		100 %	0.912	0.694	0.725	0.677	0.640	0.722	0.685	0.472	0.901

## B Details zu den Modellen

Kenngröße	Klasse		astro/LR-highwn					wMAE	wRMSE	wRMAE	wMSE
			(hard)	(crisp)	weak	prd	strong				
	A°III+	$\bar{x}$	0.933	0.801	0.824	0.779	0.740	0.797	0.740	0.549	0.932
		$s(x)$	0.006	0.004	0.003	0.003	0.003	0.003	0.004	0.003	0.002
		0 %	0.914	0.790	0.815	0.770	0.731	0.787	0.727	0.539	0.926
		5 %	0.921	0.794	0.819	0.774	0.735	0.792	0.733	0.544	0.929
		50 %	0.933	0.801	0.824	0.779	0.740	0.797	0.741	0.550	0.933
		95 %	0.941	0.806	0.829	0.783	0.744	0.802	0.745	0.555	0.935
		100 %	0.945	0.809	0.830	0.784	0.745	0.803	0.747	0.556	0.936

Kenngröße	Klasse		astro/LR-crisp					wMAE	wRMSE	wRMAE	wMSE
			(hard)	(crisp)	weak	prd	strong				
Sensitivität	N	$\bar{x}$	0.718	0.694	0.594	0.567	0.541	0.628	0.495	0.390	0.745
		$s(x)$	0.030	0.026	0.019	0.018	0.018	0.018	0.016	0.015	0.017
		0 %	0.635	0.613	0.551	0.525	0.498	0.582	0.457	0.353	0.705
		5 %	0.669	0.655	0.564	0.538	0.511	0.599	0.467	0.367	0.716
		50 %	0.717	0.693	0.593	0.567	0.541	0.628	0.494	0.390	0.743
		95 %	0.774	0.741	0.626	0.599	0.573	0.658	0.525	0.416	0.774
		100 %	0.804	0.772	0.652	0.625	0.600	0.686	0.544	0.439	0.792
	A°II	$\bar{x}$	0.502	0.494	0.575	0.483	0.406	0.584	0.474	0.355	0.723
		$s(x)$	0.031	0.026	0.016	0.015	0.015	0.014	0.014	0.011	0.015
		0 %	0.398	0.407	0.522	0.435	0.362	0.546	0.437	0.326	0.683
		5 %	0.444	0.443	0.543	0.451	0.373	0.560	0.450	0.337	0.697
		50 %	0.506	0.497	0.577	0.485	0.407	0.587	0.476	0.357	0.725
		95 %	0.546	0.530	0.595	0.504	0.428	0.603	0.492	0.370	0.742
		100 %	0.567	0.552	0.608	0.514	0.438	0.617	0.506	0.381	0.756
A°III+	$\bar{x}$	0.722	0.693	0.648	0.611	0.581	0.671	0.537	0.426	0.786	
	$s(x)$	0.016	0.015	0.011	0.011	0.011	0.010	0.010	0.009	0.009	
	0 %	0.655	0.635	0.608	0.571	0.540	0.633	0.504	0.394	0.754	
	5 %	0.697	0.669	0.631	0.594	0.563	0.655	0.522	0.413	0.772	
	50 %	0.725	0.695	0.649	0.612	0.582	0.670	0.537	0.425	0.786	
	95 %	0.746	0.714	0.665	0.627	0.597	0.686	0.554	0.440	0.801	
	100 %	0.765	0.737	0.681	0.643	0.613	0.699	0.564	0.452	0.810	
Spezifität	N	$\bar{x}$	0.880	0.863	0.878	0.863	0.849	0.859	0.708	0.625	0.915
		$s(x)$	0.011	0.010	0.007	0.007	0.007	0.006	0.011	0.008	0.006
		0 %	0.845	0.828	0.851	0.839	0.825	0.838	0.671	0.597	0.892
		5 %	0.858	0.840	0.864	0.849	0.835	0.846	0.686	0.608	0.901
		50 %	0.880	0.864	0.879	0.865	0.850	0.861	0.710	0.627	0.916
		95 %	0.894	0.875	0.888	0.873	0.858	0.869	0.722	0.638	0.923
		100 %	0.902	0.884	0.892	0.878	0.863	0.873	0.729	0.643	0.927
	A°II	$\bar{x}$	0.786	0.753	0.712	0.684	0.660	0.710	0.577	0.461	0.821
		$s(x)$	0.016	0.014	0.011	0.011	0.011	0.010	0.010	0.009	0.009
		0 %	0.733	0.711	0.683	0.655	0.631	0.679	0.543	0.433	0.791
		5 %	0.759	0.729	0.692	0.664	0.640	0.692	0.560	0.445	0.806
		50 %	0.787	0.754	0.714	0.685	0.661	0.709	0.578	0.461	0.822
		95 %	0.811	0.774	0.727	0.698	0.675	0.725	0.593	0.475	0.834
		100 %	0.836	0.800	0.742	0.715	0.692	0.736	0.603	0.486	0.843

## B.1 Details zum Astrozytomgrading

Kenngröße	Klasse	astro/LR-crisp					wMAE	wRMSE	wRMAE	wMSE	
		(hard)	(crisp)	weak	prd	strong					
	A°III+	$\bar{x}$	0.907	0.884	0.849	0.822	0.801	0.833	0.700	0.591	0.910
		$s(x)$	0.010	0.009	0.008	0.008	0.008	0.007	0.010	0.009	0.006
		0 %	0.872	0.853	0.818	0.792	0.771	0.807	0.664	0.560	0.887
		5 %	0.892	0.872	0.834	0.808	0.786	0.820	0.682	0.576	0.899
		50 %	0.907	0.884	0.849	0.823	0.801	0.833	0.700	0.591	0.910
		95 %	0.921	0.898	0.861	0.834	0.812	0.844	0.716	0.605	0.919
		100 %	0.925	0.904	0.864	0.837	0.816	0.852	0.721	0.616	0.922

Kenngröße	Klasse	astro/ideal					wMAE	wRMSE	wRMAE	wMSE	
		(hard)	(crisp)	weak	prd	strong					
Sensitivität	N	$\bar{x}$	1	1	1	0.914	0.861	1	1	1	1
	A°II	$\bar{x}$	1	1	1	0.769	0.653	1	1	1	1
	A°III+	$\bar{x}$	1	1	1	0.880	0.823	1	1	1	1
Spezifität	N	$\bar{x}$	1	1	1	0.953	0.925	1	1	1	1
	A°II	$\bar{x}$	1	1	1	0.929	0.893	1	1	1	1
	A°III+	$\bar{x}$	1	1	1	0.915	0.874	1	1	1	1

Kenngröße	Klasse	astro/PLS-LR					wMAE	wRMSE	wRMAE	wMSE	
		(hard)	(crisp)	weak	prd	strong					
Sensitivität	N	$\bar{x}$	0.751	0.673	0.622	0.581	0.542	0.644	0.566	0.403	0.811
		$s(x)$	0.026	0.018	0.013	0.013	0.013	0.013	0.013	0.010	0.011
		0 %	0.665	0.619	0.573	0.534	0.494	0.596	0.520	0.364	0.769
		5 %	0.705	0.641	0.601	0.561	0.521	0.624	0.543	0.386	0.791
		50 %	0.753	0.674	0.622	0.581	0.541	0.643	0.565	0.403	0.811
		95 %	0.794	0.703	0.640	0.599	0.560	0.663	0.584	0.419	0.827
	100 %	0.824	0.718	0.655	0.613	0.573	0.676	0.600	0.431	0.840	
	A°II	$\bar{x}$	0.438	0.455	0.565	0.425	0.316	0.581	0.508	0.353	0.757
		$s(x)$	0.030	0.016	0.011	0.010	0.009	0.009	0.010	0.007	0.009
		0 %	0.364	0.416	0.532	0.398	0.291	0.559	0.484	0.336	0.734
		5 %	0.390	0.428	0.547	0.410	0.301	0.566	0.491	0.341	0.741
		50 %	0.440	0.457	0.565	0.425	0.317	0.583	0.509	0.354	0.759
95 %		0.487	0.480	0.580	0.440	0.331	0.595	0.522	0.364	0.772	
100 %	0.492	0.486	0.587	0.443	0.334	0.604	0.530	0.371	0.779		
A°III+	$\bar{x}$	0.713	0.625	0.624	0.561	0.511	0.651	0.571	0.409	0.816	
	$s(x)$	0.011	0.008	0.007	0.007	0.007	0.007	0.006	0.006	0.005	
	0 %	0.690	0.607	0.606	0.545	0.494	0.632	0.556	0.393	0.803	
	5 %	0.693	0.613	0.612	0.549	0.499	0.641	0.561	0.401	0.807	
	50 %	0.713	0.625	0.625	0.562	0.512	0.652	0.571	0.410	0.816	
	95 %	0.732	0.637	0.636	0.572	0.523	0.661	0.582	0.418	0.825	
100 %	0.740	0.649	0.642	0.580	0.531	0.669	0.589	0.425	0.831		
Spezifität	N	$\bar{x}$	0.877	0.831	0.842	0.820	0.799	0.830	0.715	0.588	0.919
		$s(x)$	0.006	0.006	0.004	0.004	0.004	0.004	0.006	0.005	0.004
		0 %	0.859	0.814	0.828	0.807	0.787	0.817	0.692	0.572	0.905
		5 %	0.865	0.822	0.834	0.812	0.791	0.822	0.702	0.578	0.911
		50 %	0.878	0.832	0.842	0.820	0.799	0.830	0.716	0.588	0.919
		95 %	0.886	0.838	0.848	0.826	0.804	0.836	0.724	0.595	0.924
100 %	0.899	0.848	0.852	0.830	0.808	0.839	0.729	0.599	0.927		

## B Details zu den Modellen

Kenngröße	Klasse		astro/PLS-LR					wMAE	wRMSE	wRMAE	wMSE
			(hard)	(crisp)	weak	prd	strong				
A°II		$\bar{x}$	0.872	0.746	0.742	0.699	0.665	0.736	0.668	0.487	0.890
		$s(x)$	0.010	0.008	0.006	0.006	0.006	0.006	0.006	0.006	0.004
		0 %	0.850	0.729	0.726	0.683	0.649	0.721	0.654	0.472	0.880
		5 %	0.855	0.734	0.732	0.689	0.655	0.727	0.657	0.477	0.882
		50 %	0.872	0.747	0.742	0.699	0.666	0.737	0.669	0.487	0.890
		95 %	0.890	0.759	0.751	0.708	0.674	0.746	0.679	0.496	0.897
		100 %	0.897	0.767	0.756	0.713	0.681	0.751	0.684	0.501	0.900
A°III+		$\bar{x}$	0.932	0.851	0.846	0.801	0.766	0.826	0.749	0.583	0.937
		$s(x)$	0.007	0.007	0.005	0.005	0.005	0.005	0.006	0.006	0.003
		0 %	0.904	0.829	0.832	0.788	0.753	0.813	0.731	0.567	0.928
		5 %	0.917	0.838	0.834	0.790	0.755	0.816	0.735	0.571	0.930
		50 %	0.934	0.853	0.846	0.802	0.766	0.827	0.750	0.584	0.938
		95 %	0.942	0.859	0.853	0.808	0.773	0.833	0.757	0.592	0.941
		100 %	0.945	0.864	0.856	0.812	0.777	0.837	0.763	0.596	0.944

Kenngröße	Klasse		astro/PLS-LDA					wMAE	wRMSE	wRMAE	wMSE
			(hard)	(crisp)	weak	prd	strong				
Sensitivität	N	$\bar{x}$	0.723	0.690	0.594	0.563	0.532	0.627	0.509	0.389	0.758
		$s(x)$	0.035	0.029	0.020	0.020	0.020	0.020	0.019	0.016	0.018
		0 %	0.627	0.615	0.541	0.510	0.479	0.577	0.461	0.350	0.709
		5 %	0.661	0.639	0.556	0.526	0.496	0.592	0.475	0.361	0.725
		50 %	0.724	0.693	0.595	0.564	0.533	0.628	0.509	0.390	0.759
		95 %	0.780	0.736	0.626	0.594	0.564	0.658	0.539	0.415	0.787
		100 %	0.818	0.764	0.644	0.613	0.582	0.678	0.556	0.432	0.803
A°II		$\bar{x}$	0.499	0.490	0.589	0.488	0.403	0.594	0.490	0.363	0.740
		$s(x)$	0.032	0.025	0.016	0.015	0.015	0.014	0.015	0.011	0.015
		0 %	0.384	0.392	0.544	0.444	0.357	0.552	0.447	0.331	0.695
		5 %	0.438	0.442	0.560	0.459	0.375	0.569	0.465	0.343	0.713
		50 %	0.503	0.493	0.590	0.490	0.404	0.596	0.492	0.364	0.742
		95 %	0.542	0.524	0.611	0.508	0.423	0.614	0.511	0.378	0.761
		100 %	0.568	0.543	0.614	0.514	0.429	0.624	0.522	0.386	0.771
A°III+		$\bar{x}$	0.730	0.698	0.651	0.613	0.582	0.676	0.547	0.431	0.794
		$s(x)$	0.017	0.015	0.012	0.011	0.012	0.011	0.010	0.009	0.009
		0 %	0.641	0.627	0.610	0.571	0.540	0.636	0.512	0.397	0.762
		5 %	0.705	0.674	0.631	0.594	0.563	0.660	0.531	0.417	0.780
		50 %	0.733	0.699	0.651	0.613	0.583	0.676	0.547	0.431	0.795
		95 %	0.755	0.721	0.670	0.631	0.601	0.694	0.564	0.446	0.810
		100 %	0.769	0.733	0.678	0.640	0.610	0.700	0.569	0.453	0.814
Spezifität	N	$\bar{x}$	0.875	0.853	0.873	0.857	0.840	0.856	0.712	0.620	0.917
		$s(x)$	0.010	0.009	0.006	0.006	0.006	0.006	0.010	0.007	0.006
		0 %	0.846	0.822	0.851	0.837	0.821	0.837	0.681	0.597	0.898
		5 %	0.857	0.835	0.862	0.846	0.829	0.845	0.693	0.606	0.906
		50 %	0.876	0.854	0.874	0.858	0.841	0.856	0.713	0.621	0.918
		95 %	0.888	0.866	0.882	0.865	0.848	0.863	0.726	0.630	0.925
		100 %	0.899	0.870	0.886	0.869	0.853	0.867	0.732	0.635	0.928

## B.1 Details zum Astrozytomgrading

Kenngröße	Klasse	astro/PLS-LDA					wMAE	wRMSE	wRMAE	wMSE	
		(hard)	(crisp)	weak	prd	strong					
	A°II	$\bar{x}$	0.787	0.749	0.707	0.676	0.650	0.705	0.585	0.457	0.828
		$s(x)$	0.019	0.015	0.011	0.011	0.012	0.011	0.011	0.010	0.010
		0 %	0.725	0.701	0.673	0.643	0.617	0.671	0.544	0.427	0.792
		5 %	0.753	0.722	0.685	0.654	0.627	0.686	0.566	0.440	0.812
		50 %	0.790	0.750	0.708	0.677	0.650	0.705	0.586	0.457	0.829
		95 %	0.812	0.769	0.722	0.692	0.666	0.721	0.602	0.472	0.841
		100 %	0.833	0.792	0.733	0.703	0.678	0.729	0.610	0.480	0.848
	A°III+	$\bar{x}$	0.923	0.900	0.868	0.841	0.819	0.851	0.725	0.614	0.924
		$s(x)$	0.009	0.008	0.007	0.007	0.007	0.007	0.010	0.008	0.006
		0 %	0.891	0.871	0.842	0.814	0.792	0.827	0.689	0.584	0.903
		5 %	0.907	0.884	0.854	0.827	0.805	0.837	0.704	0.597	0.913
		50 %	0.924	0.902	0.869	0.842	0.820	0.851	0.726	0.614	0.925
		95 %	0.935	0.912	0.879	0.851	0.829	0.861	0.739	0.627	0.932
		100 %	0.939	0.913	0.883	0.856	0.834	0.867	0.744	0.635	0.934

Kenngröße	Klasse	astro/LR-soft-agg					wMAE	wRMSE	wRMAE	wMSE	
		(hard)	(crisp)	weak	prd	strong					
Sensitivität	N	$\bar{x}$	0.774	0.682	0.628	0.586	0.546	0.651	0.579	0.409	0.823
		$s(x)$	0.010	0.008	0.005	0.005	0.005	0.005	0.005	0.004	0.004
		0 %	0.758	0.667	0.621	0.579	0.539	0.644	0.573	0.403	0.817
		5 %	0.760	0.670	0.621	0.580	0.539	0.644	0.573	0.403	0.817
		50 %	0.776	0.683	0.629	0.587	0.546	0.652	0.580	0.410	0.824
		95 %	0.789	0.691	0.635	0.593	0.553	0.658	0.586	0.415	0.829
		100 %	0.791	0.692	0.636	0.594	0.553	0.658	0.588	0.415	0.830
	A°II	$\bar{x}$	0.440	0.455	0.569	0.424	0.311	0.584	0.513	0.355	0.762
		$s(x)$	0.013	0.007	0.004	0.004	0.004	0.004	0.004	0.003	0.004
		0 %	0.424	0.445	0.562	0.418	0.306	0.579	0.507	0.351	0.757
		5 %	0.424	0.446	0.562	0.418	0.306	0.579	0.508	0.351	0.758
		50 %	0.438	0.456	0.570	0.424	0.312	0.585	0.513	0.356	0.763
		95 %	0.461	0.465	0.574	0.429	0.316	0.589	0.518	0.359	0.768
		100 %	0.464	0.468	0.576	0.431	0.319	0.592	0.520	0.361	0.770
	A°III+	$\bar{x}$	0.716	0.624	0.625	0.560	0.509	0.653	0.576	0.411	0.820
		$s(x)$	0.004	0.003	0.002	0.002	0.002	0.002	0.002	0.001	0.001
		0 %	0.709	0.617	0.622	0.557	0.506	0.650	0.574	0.408	0.818
		5 %	0.710	0.620	0.623	0.558	0.506	0.650	0.574	0.408	0.819
		50 %	0.717	0.624	0.625	0.560	0.509	0.652	0.576	0.410	0.820
		95 %	0.721	0.627	0.628	0.563	0.512	0.655	0.578	0.413	0.822
		100 %	0.722	0.628	0.629	0.563	0.512	0.655	0.578	0.413	0.822
Spezifität	N	$\bar{x}$	0.882	0.835	0.845	0.822	0.800	0.833	0.723	0.591	0.923
		$s(x)$	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.001
		0 %	0.877	0.832	0.843	0.820	0.798	0.831	0.720	0.588	0.921
		5 %	0.878	0.832	0.843	0.820	0.798	0.831	0.720	0.589	0.921
		50 %	0.882	0.835	0.845	0.823	0.800	0.833	0.723	0.591	0.923
		95 %	0.887	0.839	0.848	0.825	0.803	0.835	0.727	0.594	0.925
		100 %	0.888	0.839	0.848	0.825	0.803	0.836	0.727	0.595	0.925

## B Details zu den Modellen

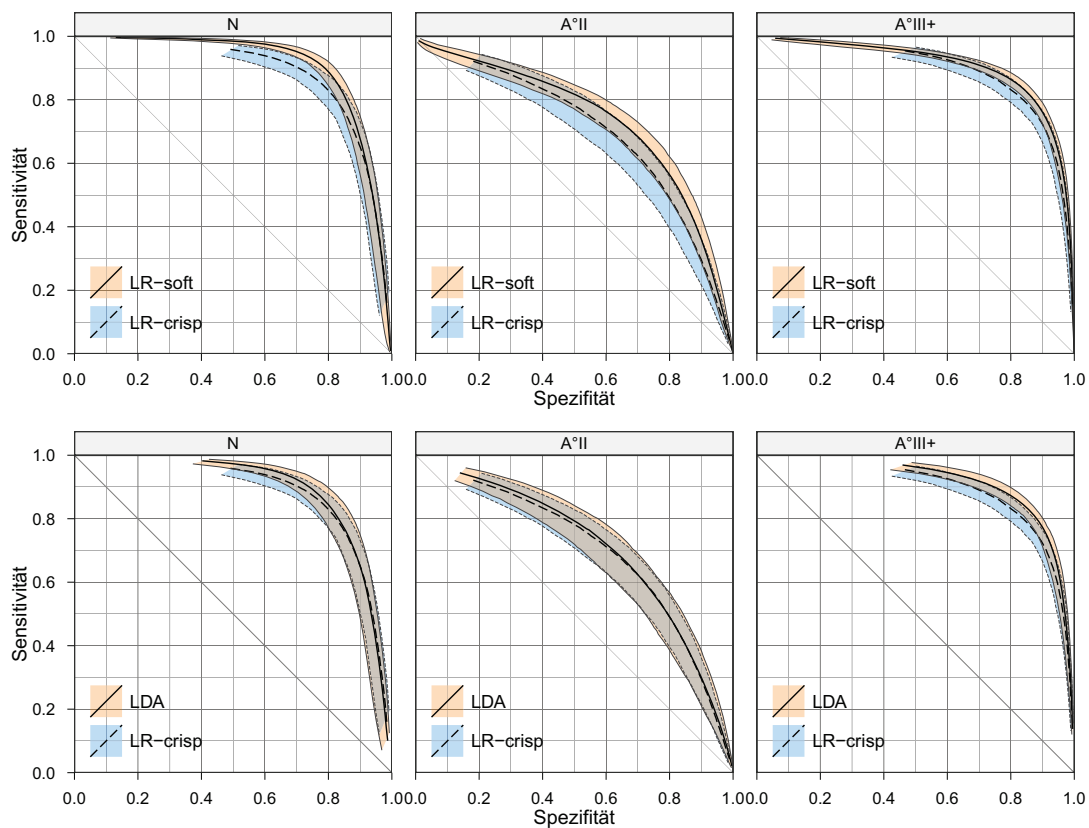
Kenngröße	Klasse		astro/LR-soft-agg					wMAE	wRMSE	wRMAE	wMSE
			(hard)	(crisp)	weak	prd	strong				
	A°II	$\bar{x}$	0.882	0.746	0.742	0.698	0.663	0.737	0.675	0.488	0.894
		$s(x)$	0.004	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.001
		0 %	0.876	0.740	0.738	0.694	0.659	0.733	0.671	0.483	0.892
		5 %	0.877	0.742	0.739	0.695	0.660	0.734	0.671	0.484	0.892
		50 %	0.882	0.746	0.743	0.698	0.664	0.738	0.675	0.488	0.894
		95 %	0.887	0.749	0.745	0.701	0.666	0.740	0.678	0.491	0.896
		100 %	0.889	0.749	0.746	0.701	0.667	0.741	0.679	0.491	0.897
	A°III+	$\bar{x}$	0.936	0.852	0.848	0.802	0.765	0.828	0.755	0.586	0.940
		$s(x)$	0.002	0.002	0.001	0.001	0.001	0.001	0.002	0.002	0.001
		0 %	0.932	0.849	0.846	0.800	0.764	0.826	0.752	0.583	0.938
		5 %	0.934	0.850	0.846	0.800	0.764	0.826	0.753	0.583	0.939
		50 %	0.937	0.853	0.848	0.802	0.765	0.829	0.755	0.586	0.940
		95 %	0.939	0.854	0.849	0.803	0.767	0.830	0.758	0.588	0.941
		100 %	0.939	0.854	0.850	0.803	0.767	0.831	0.758	0.589	0.941

Kenngröße	Klasse		astro/LDA-agg					wMAE	wRMSE	wRMAE	wMSE
			(hard)	(crisp)	weak	prd	strong				
Sensitivität	N	$\bar{x}$	0.735	0.691	0.598	0.563	0.528	0.631	0.527	0.393	0.776
		$s(x)$	0.016	0.014	0.009	0.009	0.009	0.009	0.009	0.007	0.008
		0 %	0.716	0.671	0.582	0.547	0.512	0.615	0.515	0.380	0.764
		5 %	0.718	0.671	0.586	0.552	0.516	0.619	0.515	0.383	0.765
		50 %	0.731	0.689	0.598	0.563	0.528	0.631	0.526	0.392	0.775
		95 %	0.761	0.711	0.612	0.578	0.542	0.645	0.540	0.404	0.789
		100 %	0.769	0.718	0.613	0.579	0.545	0.646	0.542	0.405	0.790
	A°II	$\bar{x}$	0.500	0.490	0.602	0.488	0.394	0.604	0.506	0.371	0.756
		$s(x)$	0.011	0.010	0.006	0.007	0.007	0.006	0.006	0.004	0.005
		0 %	0.485	0.475	0.593	0.479	0.385	0.596	0.498	0.364	0.748
		5 %	0.486	0.476	0.593	0.480	0.385	0.596	0.500	0.365	0.750
		50 %	0.502	0.492	0.601	0.487	0.393	0.604	0.507	0.371	0.757
		95 %	0.515	0.504	0.613	0.499	0.405	0.611	0.514	0.377	0.764
		100 %	0.518	0.504	0.614	0.500	0.406	0.612	0.514	0.377	0.764
A°III+	$\bar{x}$	0.740	0.698	0.656	0.613	0.579	0.681	0.560	0.435	0.806	
	$s(x)$	0.004	0.003	0.003	0.002	0.003	0.002	0.002	0.002	0.002	
	0 %	0.735	0.695	0.652	0.609	0.575	0.679	0.557	0.433	0.804	
	5 %	0.736	0.695	0.653	0.610	0.575	0.679	0.557	0.433	0.804	
	50 %	0.739	0.698	0.656	0.613	0.579	0.681	0.559	0.435	0.806	
	95 %	0.747	0.704	0.660	0.617	0.582	0.685	0.564	0.438	0.810	
	100 %	0.747	0.705	0.662	0.618	0.584	0.687	0.564	0.440	0.810	
Spezifität	N	$\bar{x}$	0.880	0.853	0.875	0.857	0.838	0.858	0.724	0.623	0.924
		$s(x)$	0.004	0.003	0.002	0.002	0.003	0.002	0.003	0.003	0.002
		0 %	0.874	0.847	0.872	0.853	0.834	0.854	0.720	0.618	0.921
		5 %	0.875	0.848	0.872	0.854	0.835	0.854	0.720	0.618	0.922
		50 %	0.881	0.854	0.875	0.856	0.838	0.858	0.724	0.623	0.924
		95 %	0.885	0.858	0.879	0.860	0.841	0.861	0.729	0.627	0.927
		100 %	0.887	0.858	0.879	0.860	0.841	0.861	0.729	0.627	0.927

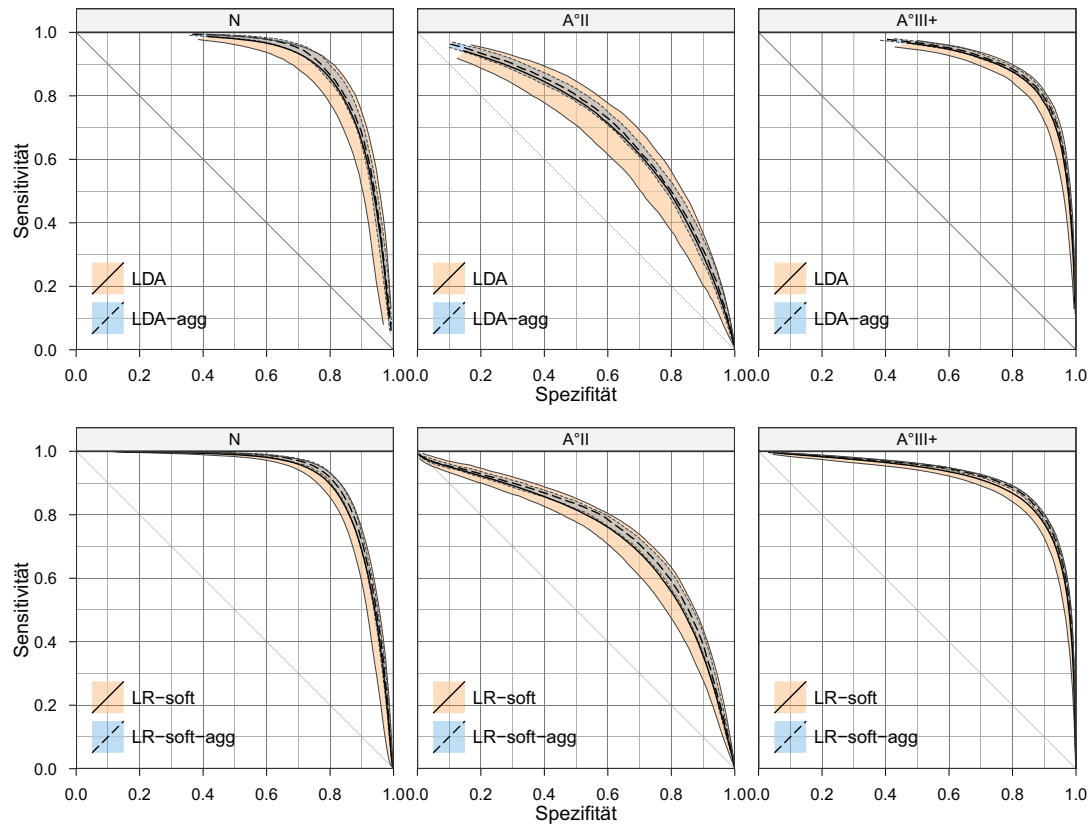


Kenngröße	Klasse		astro/LDA-agg					wMAE	wRMSE	wRMAE	wMSE
			(hard)	(crisp)	weak	prd	strong				
A°II	$\bar{x}$		0.800	0.749	0.711	0.676	0.647	0.711	0.604	0.462	0.843
		$s(x)$	0.007	0.005	0.004	0.004	0.004	0.004	0.004	0.003	0.003
		0 %	0.791	0.743	0.706	0.671	0.641	0.706	0.599	0.457	0.839
		5 %	0.792	0.743	0.706	0.671	0.642	0.706	0.599	0.458	0.839
		50 %	0.798	0.749	0.711	0.676	0.647	0.710	0.602	0.461	0.842
		95 %	0.809	0.756	0.717	0.682	0.653	0.716	0.610	0.467	0.848
		100 %	0.810	0.756	0.719	0.684	0.655	0.717	0.611	0.468	0.849
A°III+	$\bar{x}$		0.929	0.900	0.872	0.841	0.817	0.854	0.739	0.618	0.932
		$s(x)$	0.002	0.002	0.002	0.002	0.002	0.002	0.003	0.002	0.001
		0 %	0.924	0.898	0.868	0.837	0.813	0.851	0.733	0.614	0.929
		5 %	0.925	0.898	0.869	0.839	0.814	0.852	0.735	0.615	0.930
		50 %	0.929	0.900	0.872	0.841	0.817	0.854	0.739	0.618	0.932
		95 %	0.931	0.904	0.874	0.844	0.819	0.856	0.742	0.621	0.934
		100 %	0.933	0.904	0.875	0.845	0.820	0.857	0.742	0.622	0.934

### B.1.2 Weitere Spezifiäts-Sensitivitäts-Diagramme



## B Details zu den Modellen



## B.2 Details Differentialdiagnostik Astrozytom oder Lymphom

### B.3 „normale“ Modelle

		lymph/ideal									
Kenngröße	Klasse		(hard)	(crisp)	weak	prd	strong	wMAE	wRMSE	wRMAE	wMSE
Sensitivität	N	$\bar{x}$	1.000	1.000	1.000	0.911	0.859	1.000	1.000	1.000	1.000
	A	$\bar{x}$	1.000	1.000	1.000	0.953	0.925	1.000	1.000	1.000	1.000
	L	$\bar{x}$	1.000	1.000	1.000	0.956	0.941	1.000	1.000	1.000	1.000
Spezifität	N	$\bar{x}$	1.000	1.000	1.000	0.954	0.926	1.000	1.000	1.000	1.000
	A	$\bar{x}$	1.000	1.000	1.000	0.929	0.886	1.000	1.000	1.000	1.000
	L	$\bar{x}$	1.000	1.000	1.000	0.998	0.997	1.000	1.000	1.000	1.000

		lymph/LR-soft									
Kenngröße	Klasse		(hard)	(crisp)	weak	prd	strong	wMAE	wRMSE	wRMAE	wMSE
Sensitivität	N	$\bar{x}$	0.855	0.768	0.727	0.684	0.645	0.733	0.644	0.483	0.873
		$s(x)$	0.015	0.012	0.010	0.009	0.010	0.009	0.010	0.009	0.007
		0%	0.821	0.739	0.701	0.659	0.619	0.710	0.619	0.461	0.855
		5%	0.831	0.749	0.711	0.669	0.630	0.718	0.628	0.469	0.861
		50%	0.856	0.768	0.727	0.684	0.646	0.733	0.645	0.483	0.874
		95%	0.879	0.788	0.742	0.699	0.661	0.747	0.658	0.497	0.883
		100%	0.893	0.798	0.749	0.706	0.668	0.754	0.669	0.504	0.890

Kenngröße	Klasse	lymph/LR-soft					wMAE	wRMSE	wRMAE	wMSE		
		(hard)	(crisp)	weak	prd	strong						
Spezifität	A	$\bar{x}$	0.725	0.677	0.676	0.653	0.633	0.678	0.561	0.432	0.807	
		$s(x)$	0.008	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.005	0.005
		0 %	0.700	0.657	0.656	0.634	0.614	0.657	0.543	0.414	0.791	
		5 %	0.713	0.667	0.667	0.643	0.623	0.668	0.551	0.424	0.799	
		50 %	0.725	0.678	0.676	0.653	0.633	0.678	0.561	0.433	0.807	
		95 %	0.737	0.687	0.685	0.662	0.642	0.686	0.569	0.440	0.815	
		100 %	0.745	0.692	0.690	0.666	0.646	0.691	0.574	0.444	0.819	
	L	$\bar{x}$	0.743	0.719	0.682	0.673	0.668	0.706	0.553	0.470	0.783	
		$s(x)$	0.174	0.158	0.146	0.146	0.145	0.145	0.131	0.112	0.153	
		0 %	0.249	0.255	0.241	0.240	0.238	0.278	0.178	0.150	0.324	
		5 %	0.301	0.323	0.318	0.309	0.304	0.341	0.227	0.188	0.402	
		50 %	0.809	0.779	0.736	0.727	0.723	0.761	0.602	0.511	0.841	
		95 %	0.835	0.808	0.764	0.754	0.749	0.787	0.626	0.539	0.860	
		100 %	0.846	0.818	0.775	0.764	0.758	0.796	0.639	0.549	0.870	
	N	$\bar{x}$	0.835	0.790	0.781	0.758	0.738	0.777	0.651	0.528	0.878	
		$s(x)$	0.005	0.004	0.004	0.004	0.004	0.004	0.005	0.004	0.003	
		0 %	0.823	0.779	0.771	0.749	0.728	0.767	0.638	0.517	0.869	
		5 %	0.826	0.783	0.774	0.752	0.732	0.771	0.643	0.521	0.872	
		50 %	0.836	0.790	0.781	0.759	0.738	0.778	0.652	0.529	0.879	
		95 %	0.844	0.797	0.787	0.765	0.744	0.784	0.658	0.535	0.883	
		100 %	0.847	0.802	0.789	0.767	0.747	0.787	0.662	0.538	0.886	
	A	$\bar{x}$	0.852	0.778	0.753	0.717	0.686	0.750	0.652	0.501	0.878	
		$s(x)$	0.038	0.032	0.021	0.021	0.021	0.021	0.026	0.020	0.019	
		0 %	0.744	0.687	0.694	0.657	0.625	0.691	0.577	0.444	0.821	
5 %		0.758	0.700	0.702	0.667	0.635	0.701	0.589	0.453	0.831		
50 %		0.864	0.789	0.760	0.724	0.693	0.757	0.661	0.507	0.885		
95 %		0.885	0.805	0.773	0.736	0.706	0.769	0.676	0.520	0.895		
100 %		0.896	0.817	0.781	0.746	0.716	0.778	0.685	0.529	0.901		
L	$\bar{x}$	0.944	0.924	0.932	0.932	0.931	0.931	0.803	0.737	0.961		
	$s(x)$	0.004	0.004	0.003	0.003	0.003	0.003	0.005	0.005	0.002		
	0 %	0.934	0.915	0.925	0.924	0.924	0.924	0.791	0.724	0.956		
	5 %	0.937	0.918	0.928	0.927	0.927	0.926	0.793	0.728	0.957		
	50 %	0.944	0.924	0.932	0.932	0.931	0.931	0.803	0.737	0.961		
	95 %	0.948	0.930	0.937	0.936	0.936	0.935	0.810	0.745	0.964		
	100 %	0.955	0.934	0.942	0.941	0.941	0.940	0.823	0.755	0.969		

Kenngröße	Klasse	lymph/LR-crisp					wMAE	wRMSE	wRMAE	wMSE	
		(hard)	(crisp)	weak	prd	strong					
Sensitivität	N	$\bar{x}$	0.761	0.738	0.631	0.604	0.578	0.657	0.517	0.415	0.767
		$s(x)$	0.026	0.023	0.015	0.015	0.015	0.015	0.015	0.013	0.014
		0 %	0.686	0.666	0.590	0.564	0.540	0.619	0.479	0.383	0.728
		5 %	0.716	0.697	0.603	0.577	0.551	0.630	0.487	0.391	0.737
		50 %	0.765	0.740	0.631	0.605	0.579	0.659	0.519	0.416	0.769
		95 %	0.801	0.774	0.653	0.626	0.600	0.680	0.538	0.434	0.787
		100 %	0.820	0.790	0.666	0.638	0.612	0.691	0.551	0.445	0.798

## B Details zu den Modellen

Kenngröße	Klasse		lymph/LR-crisp					wMAE	wRMSE	wRMAE	wMSE	
			(hard)	(crisp)	weak	prd	strong					
	A	$\bar{x}$	0.817	0.796	0.788	0.773	0.759	0.779	0.607	0.530	0.846	
		$s(x)$	0.008	0.008	0.007	0.007	0.007	0.007	0.009	0.008	0.007	
		0 %	0.791	0.769	0.765	0.750	0.736	0.755	0.580	0.505	0.824	
		5 %	0.804	0.785	0.777	0.762	0.748	0.769	0.593	0.519	0.835	
		50 %	0.817	0.796	0.788	0.773	0.759	0.779	0.608	0.530	0.846	
		95 %	0.828	0.808	0.800	0.785	0.770	0.790	0.620	0.542	0.855	
		100 %	0.845	0.819	0.805	0.791	0.777	0.798	0.631	0.550	0.864	
	L	$\bar{x}$	0.731	0.726	0.679	0.676	0.675	0.711	0.504	0.471	0.742	
		$s(x)$	0.134	0.132	0.122	0.122	0.121	0.122	0.106	0.097	0.127	
		0 %	0.297	0.297	0.280	0.278	0.277	0.315	0.182	0.172	0.332	
		5 %	0.389	0.388	0.369	0.366	0.364	0.399	0.237	0.225	0.418	
		50 %	0.780	0.774	0.723	0.720	0.718	0.755	0.546	0.505	0.794	
		95 %	0.816	0.810	0.756	0.754	0.752	0.788	0.578	0.540	0.822	
		100 %	0.837	0.831	0.776	0.773	0.770	0.808	0.600	0.562	0.840	
	Spezifität	N	$\bar{x}$	0.897	0.880	0.867	0.853	0.840	0.853	0.694	0.617	0.906
			$s(x)$	0.006	0.005	0.005	0.005	0.005	0.005	0.008	0.006	0.005
			0 %	0.882	0.865	0.854	0.839	0.826	0.840	0.671	0.600	0.892
			5 %	0.887	0.872	0.859	0.845	0.831	0.844	0.680	0.605	0.897
50 %			0.897	0.881	0.868	0.854	0.841	0.854	0.695	0.618	0.907	
95 %			0.905	0.887	0.874	0.860	0.847	0.860	0.704	0.626	0.912	
100 %			0.918	0.896	0.879	0.866	0.852	0.866	0.716	0.634	0.919	
A		$\bar{x}$	0.764	0.744	0.655	0.632	0.610	0.669	0.516	0.425	0.765	
		$s(x)$	0.035	0.033	0.022	0.022	0.022	0.022	0.022	0.018	0.022	
		0 %	0.639	0.630	0.577	0.555	0.532	0.593	0.439	0.362	0.686	
		5 %	0.690	0.676	0.608	0.585	0.563	0.622	0.469	0.385	0.718	
		50 %	0.773	0.754	0.658	0.636	0.614	0.673	0.519	0.428	0.769	
		95 %	0.805	0.783	0.682	0.658	0.636	0.695	0.541	0.447	0.789	
		100 %	0.820	0.797	0.693	0.669	0.647	0.705	0.554	0.457	0.801	
L		$\bar{x}$	0.946	0.943	0.950	0.950	0.950	0.948	0.796	0.772	0.959	
		$s(x)$	0.004	0.004	0.003	0.003	0.003	0.003	0.007	0.007	0.003	
		0 %	0.935	0.933	0.941	0.941	0.941	0.940	0.774	0.754	0.949	
		5 %	0.939	0.935	0.944	0.944	0.944	0.942	0.785	0.760	0.954	
	50 %	0.947	0.943	0.950	0.950	0.950	0.948	0.797	0.772	0.959		
	95 %	0.952	0.949	0.955	0.955	0.955	0.953	0.808	0.784	0.963		
	100 %	0.957	0.954	0.959	0.959	0.959	0.957	0.814	0.793	0.965		

Kenngröße	Klasse		lymph/LR-highwn					wMAE	wRMSE	wRMAE	wMSE
			(hard)	(crisp)	weak	prd	strong				
Sensitivität	N	$\bar{x}$	0.740	0.635	0.661	0.611	0.567	0.663	0.606	0.420	0.845
		$s(x)$	0.018	0.009	0.006	0.006	0.006	0.006	0.006	0.005	0.005
		0 %	0.688	0.614	0.646	0.596	0.551	0.648	0.587	0.407	0.830
		5 %	0.707	0.620	0.651	0.602	0.556	0.653	0.595	0.411	0.836
		50 %	0.745	0.637	0.661	0.612	0.567	0.664	0.606	0.420	0.845
		95 %	0.762	0.647	0.669	0.620	0.576	0.672	0.614	0.427	0.851
		100 %	0.770	0.650	0.673	0.624	0.580	0.674	0.617	0.429	0.853

Kenngröße	Klasse		lymph/LR-highwn					wMAE	wRMSE	wRMAE	wMSE
			(hard)	(crisp)	weak	prd	strong				
Spezifität	A	$\bar{x}$	0.380	0.430	0.449	0.421	0.398	0.458	0.411	0.264	0.653
		$s(x)$	0.011	0.006	0.005	0.005	0.005	0.005	0.005	0.003	0.006
		0 %	0.349	0.415	0.436	0.409	0.386	0.444	0.398	0.255	0.638
		5 %	0.359	0.420	0.439	0.412	0.389	0.448	0.401	0.257	0.641
		50 %	0.381	0.431	0.449	0.422	0.399	0.458	0.411	0.264	0.653
		95 %	0.396	0.438	0.457	0.429	0.406	0.465	0.418	0.269	0.661
		100 %	0.411	0.448	0.466	0.438	0.414	0.473	0.427	0.274	0.672
	L	$\bar{x}$	0.806	0.665	0.633	0.623	0.615	0.655	0.599	0.415	0.836
		$s(x)$	0.113	0.058	0.054	0.054	0.054	0.054	0.053	0.043	0.048
		0 %	0.279	0.407	0.387	0.383	0.378	0.417	0.367	0.236	0.599
		5 %	0.549	0.545	0.521	0.511	0.505	0.544	0.484	0.324	0.733
		50 %	0.849	0.684	0.651	0.640	0.632	0.673	0.617	0.428	0.853
		95 %	0.884	0.715	0.679	0.668	0.661	0.701	0.642	0.453	0.872
		100 %	0.892	0.738	0.700	0.689	0.682	0.722	0.656	0.473	0.882
	N	$\bar{x}$	0.794	0.735	0.730	0.704	0.681	0.729	0.630	0.479	0.863
		$s(x)$	0.005	0.003	0.003	0.003	0.003	0.003	0.004	0.003	0.003
		0 %	0.781	0.727	0.722	0.697	0.674	0.721	0.618	0.472	0.854
		5 %	0.786	0.730	0.725	0.699	0.676	0.724	0.624	0.475	0.859
		50 %	0.795	0.735	0.730	0.704	0.681	0.729	0.630	0.479	0.863
		95 %	0.801	0.740	0.735	0.709	0.686	0.734	0.635	0.484	0.867
		100 %	0.807	0.742	0.737	0.711	0.687	0.735	0.637	0.486	0.868
	A	$\bar{x}$	0.840	0.691	0.727	0.685	0.650	0.713	0.662	0.465	0.886
		$s(x)$	0.018	0.011	0.008	0.008	0.008	0.008	0.008	0.007	0.006
		0 %	0.769	0.655	0.705	0.663	0.628	0.690	0.633	0.443	0.865
5 %		0.804	0.670	0.712	0.670	0.635	0.698	0.647	0.451	0.875	
50 %		0.844	0.693	0.728	0.686	0.651	0.715	0.664	0.466	0.887	
95 %		0.862	0.705	0.736	0.694	0.659	0.723	0.673	0.474	0.893	
100 %		0.868	0.708	0.742	0.700	0.666	0.728	0.677	0.478	0.895	
L	$\bar{x}$	0.815	0.777	0.807	0.806	0.806	0.805	0.677	0.559	0.896	
	$s(x)$	0.006	0.004	0.003	0.003	0.003	0.003	0.005	0.004	0.003	
	0 %	0.800	0.765	0.796	0.796	0.795	0.795	0.662	0.547	0.886	
	5 %	0.805	0.770	0.802	0.801	0.800	0.800	0.669	0.553	0.891	
	50 %	0.815	0.778	0.807	0.806	0.806	0.806	0.677	0.559	0.896	
	95 %	0.825	0.783	0.811	0.810	0.810	0.809	0.684	0.564	0.900	
	100 %	0.839	0.791	0.818	0.817	0.817	0.817	0.697	0.572	0.908	

Kenngröße	Klasse		lymph/LDA					wMAE	wRMSE	wRMAE	wMSE
			(hard)	(crisp)	weak	prd	strong				
Sensitivität	N	$\bar{x}$	0.825	0.804	0.688	0.664	0.641	0.704	0.555	0.456	0.802
		$s(x)$	0.024	0.022	0.016	0.016	0.016	0.015	0.016	0.014	0.014
		0 %	0.716	0.707	0.626	0.601	0.576	0.642	0.492	0.402	0.742
		5 %	0.787	0.771	0.661	0.637	0.613	0.679	0.528	0.433	0.777
		50 %	0.827	0.805	0.689	0.665	0.642	0.704	0.556	0.456	0.803
		95 %	0.857	0.836	0.711	0.687	0.664	0.726	0.577	0.476	0.821
		100 %	0.868	0.844	0.718	0.694	0.671	0.732	0.584	0.482	0.827

## B Details zu den Modellen

Kenngröße	Klasse		lymph/LDA					wMAE	wRMSE	wRMAE	wMSE	
			(hard)	(crisp)	weak	prd	strong					
	A	$\bar{x}$	0.776	0.752	0.739	0.724	0.711	0.736	0.569	0.487	0.814	
		$s(x)$	0.008	0.008	0.007	0.007	0.007	0.007	0.008	0.007	0.007	
		0 %	0.757	0.734	0.722	0.708	0.693	0.720	0.552	0.471	0.800	
		5 %	0.763	0.740	0.727	0.713	0.700	0.726	0.557	0.476	0.803	
		50 %	0.776	0.752	0.738	0.724	0.710	0.736	0.569	0.487	0.815	
		95 %	0.789	0.764	0.750	0.736	0.722	0.748	0.582	0.499	0.825	
		100 %	0.801	0.774	0.757	0.744	0.731	0.756	0.590	0.506	0.832	
	L	$\bar{x}$	0.879	0.860	0.806	0.800	0.796	0.835	0.675	0.602	0.889	
		$s(x)$	0.086	0.084	0.078	0.078	0.078	0.078	0.076	0.080	0.062	
		0 %	0.530	0.537	0.504	0.500	0.497	0.538	0.402	0.320	0.643	
		5 %	0.682	0.663	0.625	0.619	0.615	0.654	0.492	0.412	0.742	
		50 %	0.911	0.892	0.835	0.829	0.824	0.865	0.705	0.632	0.913	
		95 %	0.924	0.906	0.847	0.841	0.837	0.877	0.719	0.649	0.921	
		100 %	0.929	0.914	0.854	0.849	0.845	0.885	0.729	0.660	0.927	
	Spezifität	N	$\bar{x}$	0.881	0.871	0.849	0.836	0.824	0.841	0.666	0.601	0.889
			$s(x)$	0.005	0.005	0.005	0.005	0.005	0.004	0.007	0.005	0.004
			0 %	0.866	0.857	0.837	0.824	0.812	0.830	0.651	0.588	0.878
			5 %	0.873	0.863	0.841	0.828	0.816	0.834	0.655	0.592	0.881
50 %			0.882	0.871	0.850	0.837	0.825	0.842	0.668	0.602	0.890	
95 %			0.888	0.878	0.856	0.844	0.832	0.848	0.676	0.610	0.895	
100 %			0.899	0.887	0.862	0.850	0.838	0.854	0.688	0.618	0.903	
A		$\bar{x}$	0.841	0.821	0.723	0.701	0.681	0.727	0.571	0.477	0.816	
		$s(x)$	0.026	0.025	0.017	0.017	0.018	0.017	0.017	0.016	0.015	
		0 %	0.761	0.749	0.674	0.652	0.631	0.678	0.518	0.433	0.768	
		5 %	0.784	0.765	0.687	0.665	0.645	0.692	0.538	0.445	0.786	
		50 %	0.846	0.826	0.726	0.704	0.683	0.729	0.573	0.480	0.818	
		95 %	0.874	0.853	0.747	0.726	0.706	0.749	0.594	0.499	0.835	
		100 %	0.884	0.858	0.753	0.731	0.711	0.756	0.603	0.506	0.843	
L		$\bar{x}$	0.931	0.919	0.930	0.930	0.929	0.928	0.780	0.732	0.951	
		$s(x)$	0.005	0.005	0.004	0.004	0.004	0.004	0.008	0.007	0.003	
		0 %	0.920	0.908	0.921	0.921	0.921	0.920	0.761	0.717	0.943	
		5 %	0.924	0.912	0.924	0.924	0.924	0.923	0.768	0.722	0.946	
	50 %	0.931	0.919	0.930	0.929	0.929	0.928	0.778	0.732	0.951		
	95 %	0.940	0.927	0.936	0.936	0.936	0.934	0.792	0.744	0.957		
	100 %	0.944	0.933	0.940	0.940	0.940	0.938	0.802	0.752	0.961		

Kenngröße	Klasse		lymph/PLS-LR					wMAE	wRMSE	wRMAE	wMSE
			(hard)	(crisp)	weak	prd	strong				
Sensitivität	N	$\bar{x}$	0.865	0.765	0.726	0.682	0.641	0.731	0.649	0.482	0.877
		$s(x)$	0.015	0.012	0.010	0.009	0.010	0.009	0.010	0.009	0.007
		0 %	0.830	0.735	0.698	0.656	0.615	0.704	0.623	0.456	0.858
		5 %	0.840	0.747	0.710	0.667	0.626	0.716	0.632	0.467	0.865
		50 %	0.866	0.764	0.726	0.682	0.642	0.731	0.650	0.481	0.878
		95 %	0.887	0.782	0.741	0.696	0.656	0.746	0.664	0.496	0.887
		100 %	0.904	0.797	0.747	0.703	0.664	0.751	0.672	0.501	0.892

Kenngröße	Klasse	lymph/PLS-LR					wMAE	wRMSE	wRMAE	wMSE	
		(hard)	(crisp)	weak	prd	strong					
Spezifität	A	$\bar{x}$	0.722	0.667	0.667	0.642	0.622	0.669	0.561	0.425	0.807
		$s(x)$	0.008	0.006	0.006	0.006	0.006	0.006	0.006	0.005	0.005
		0 %	0.700	0.648	0.647	0.624	0.605	0.650	0.544	0.408	0.792
		5 %	0.709	0.656	0.657	0.632	0.612	0.658	0.550	0.415	0.798
		50 %	0.723	0.667	0.667	0.642	0.622	0.669	0.561	0.425	0.807
		95 %	0.736	0.677	0.677	0.653	0.632	0.678	0.570	0.433	0.815
		100 %	0.743	0.682	0.680	0.655	0.636	0.682	0.574	0.436	0.819
	L	$\bar{x}$	0.785	0.750	0.711	0.700	0.694	0.737	0.600	0.500	0.823
		$s(x)$	0.182	0.154	0.142	0.142	0.142	0.142	0.131	0.114	0.141
		0 %	0.273	0.301	0.283	0.280	0.278	0.321	0.222	0.176	0.394
		5 %	0.326	0.359	0.353	0.341	0.333	0.377	0.265	0.211	0.460
		50 %	0.856	0.808	0.762	0.753	0.747	0.789	0.648	0.541	0.876
		95 %	0.877	0.833	0.789	0.777	0.770	0.813	0.673	0.568	0.893
		100 %	0.881	0.839	0.798	0.785	0.778	0.820	0.678	0.575	0.896
	N	$\bar{x}$	0.840	0.789	0.780	0.757	0.736	0.777	0.656	0.528	0.882
		$s(x)$	0.005	0.004	0.004	0.004	0.004	0.004	0.005	0.004	0.003
		0 %	0.827	0.777	0.770	0.747	0.725	0.766	0.644	0.517	0.873
		5 %	0.831	0.783	0.774	0.751	0.730	0.771	0.647	0.521	0.876
		50 %	0.840	0.790	0.781	0.758	0.737	0.777	0.656	0.528	0.882
		95 %	0.849	0.796	0.787	0.764	0.743	0.783	0.664	0.534	0.887
		100 %	0.853	0.801	0.789	0.766	0.746	0.787	0.666	0.539	0.889
	A	$\bar{x}$	0.870	0.783	0.757	0.720	0.689	0.754	0.663	0.504	0.886
		$s(x)$	0.039	0.031	0.021	0.020	0.021	0.020	0.025	0.020	0.018
		0 %	0.761	0.693	0.699	0.661	0.629	0.696	0.590	0.448	0.832
5 %		0.772	0.707	0.708	0.671	0.639	0.707	0.605	0.459	0.844	
50 %		0.882	0.793	0.764	0.727	0.695	0.760	0.671	0.510	0.892	
95 %		0.902	0.810	0.777	0.740	0.708	0.774	0.687	0.524	0.902	
100 %		0.913	0.818	0.785	0.747	0.716	0.780	0.695	0.531	0.907	
L	$\bar{x}$	0.946	0.917	0.926	0.925	0.925	0.924	0.807	0.725	0.963	
	$s(x)$	0.004	0.004	0.003	0.003	0.003	0.003	0.006	0.006	0.002	
	0 %	0.936	0.907	0.918	0.917	0.917	0.916	0.793	0.711	0.957	
	5 %	0.939	0.911	0.921	0.920	0.920	0.919	0.797	0.716	0.959	
	50 %	0.946	0.917	0.926	0.925	0.925	0.924	0.807	0.725	0.963	
	95 %	0.951	0.922	0.931	0.930	0.930	0.929	0.815	0.734	0.966	
	100 %	0.959	0.929	0.936	0.935	0.935	0.934	0.829	0.744	0.971	

Kenngröße	Klasse	lymph/PLS-LDA					wMAE	wRMSE	wRMAE	wMSE	
		(hard)	(crisp)	weak	prd	strong					
Sensitivität	N	$\bar{x}$	0.825	0.804	0.689	0.664	0.641	0.704	0.555	0.456	0.802
		$s(x)$	0.025	0.022	0.016	0.016	0.016	0.015	0.016	0.014	0.014
		0 %	0.715	0.708	0.626	0.601	0.576	0.642	0.492	0.402	0.741
		5 %	0.787	0.771	0.661	0.636	0.613	0.679	0.527	0.433	0.777
		50 %	0.827	0.805	0.689	0.665	0.642	0.704	0.556	0.456	0.803
		95 %	0.857	0.836	0.712	0.687	0.664	0.726	0.577	0.477	0.821
		100 %	0.869	0.844	0.718	0.694	0.672	0.732	0.584	0.482	0.827

## B Details zu den Modellen

Kenngröße	Klasse		lymph/PLS-LDA					wMAE	wRMSE	wRMAE	wMSE	
			(hard)	(crisp)	weak	prd	strong					
	A	$\bar{x}$	0.776	0.752	0.739	0.724	0.711	0.737	0.570	0.487	0.815	
		$s(x)$	0.008	0.008	0.007	0.007	0.007	0.007	0.008	0.007	0.007	
		0 %	0.757	0.734	0.723	0.708	0.694	0.721	0.553	0.471	0.800	
		5 %	0.763	0.740	0.728	0.713	0.700	0.726	0.557	0.476	0.804	
		50 %	0.776	0.752	0.739	0.724	0.711	0.737	0.570	0.487	0.815	
		95 %	0.789	0.765	0.751	0.736	0.723	0.749	0.582	0.499	0.826	
		100 %	0.800	0.774	0.758	0.744	0.731	0.756	0.591	0.506	0.832	
	L	$\bar{x}$	0.879	0.859	0.805	0.800	0.795	0.835	0.675	0.602	0.889	
		$s(x)$	0.086	0.085	0.078	0.078	0.078	0.078	0.076	0.080	0.062	
		0 %	0.529	0.536	0.502	0.499	0.496	0.536	0.401	0.319	0.641	
		5 %	0.683	0.662	0.623	0.618	0.614	0.654	0.492	0.412	0.742	
		50 %	0.911	0.892	0.835	0.829	0.824	0.865	0.705	0.632	0.913	
		95 %	0.924	0.906	0.847	0.841	0.837	0.877	0.719	0.650	0.921	
		100 %	0.930	0.913	0.854	0.849	0.844	0.884	0.730	0.660	0.927	
	Spezifität	N	$\bar{x}$	0.881	0.871	0.849	0.836	0.825	0.841	0.667	0.601	0.889
			$s(x)$	0.005	0.005	0.005	0.005	0.005	0.004	0.007	0.005	0.004
			0 %	0.866	0.857	0.837	0.824	0.812	0.830	0.651	0.588	0.879
			5 %	0.873	0.864	0.841	0.829	0.817	0.834	0.655	0.593	0.881
50 %			0.882	0.871	0.850	0.837	0.825	0.842	0.668	0.602	0.890	
95 %			0.889	0.878	0.856	0.844	0.832	0.848	0.676	0.610	0.895	
100 %			0.899	0.887	0.862	0.850	0.838	0.854	0.688	0.618	0.903	
A		$\bar{x}$	0.841	0.821	0.723	0.701	0.681	0.727	0.571	0.477	0.816	
		$s(x)$	0.026	0.025	0.017	0.017	0.018	0.017	0.017	0.016	0.015	
		0 %	0.760	0.749	0.674	0.652	0.631	0.678	0.518	0.433	0.768	
		5 %	0.783	0.765	0.687	0.665	0.644	0.692	0.537	0.445	0.786	
		50 %	0.846	0.826	0.725	0.703	0.683	0.729	0.573	0.480	0.818	
		95 %	0.874	0.853	0.747	0.726	0.706	0.749	0.594	0.499	0.835	
		100 %	0.884	0.858	0.753	0.731	0.711	0.756	0.603	0.506	0.843	
L		$\bar{x}$	0.931	0.919	0.930	0.930	0.930	0.928	0.780	0.733	0.951	
		$s(x)$	0.005	0.005	0.004	0.004	0.004	0.004	0.008	0.007	0.003	
		0 %	0.920	0.908	0.922	0.921	0.921	0.920	0.761	0.717	0.943	
		5 %	0.924	0.912	0.925	0.924	0.924	0.923	0.769	0.723	0.946	
	50 %	0.931	0.919	0.930	0.930	0.929	0.928	0.779	0.732	0.951		
	95 %	0.940	0.927	0.936	0.936	0.936	0.935	0.792	0.744	0.957		
	100 %	0.944	0.933	0.940	0.940	0.940	0.938	0.803	0.752	0.961		

Kenngröße	Klasse		lymph/LDA-highwn					wMAE	wRMSE	wRMAE	wMSE
			(hard)	(crisp)	weak	prd	strong				
Sensitivität	N	$\bar{x}$	0.705	0.659	0.662	0.629	0.598	0.661	0.565	0.418	0.811
		$s(x)$	0.024	0.016	0.009	0.009	0.009	0.009	0.010	0.008	0.009
		0 %	0.595	0.593	0.622	0.590	0.559	0.620	0.518	0.384	0.768
		5 %	0.665	0.634	0.649	0.615	0.584	0.649	0.549	0.408	0.796
		50 %	0.707	0.661	0.663	0.630	0.599	0.661	0.566	0.418	0.812
		95 %	0.735	0.679	0.676	0.642	0.611	0.673	0.579	0.428	0.822
		100 %	0.747	0.692	0.684	0.651	0.620	0.681	0.585	0.436	0.827



Kenngröße	Klasse	lymph/LDA-highwn					wMAE	wRMSE	wRMAE	wMSE		
		(hard)	(crisp)	weak	prd	strong						
Spezifität	A	$\bar{x}$	0.441	0.443	0.450	0.429	0.411	0.462	0.389	0.266	0.627	
		$s(x)$	0.008	0.006	0.005	0.005	0.005	0.005	0.005	0.003	0.006	
		0 %	0.416	0.425	0.434	0.413	0.395	0.446	0.373	0.256	0.607	
		5 %	0.428	0.433	0.441	0.420	0.402	0.453	0.381	0.260	0.617	
		50 %	0.441	0.443	0.450	0.429	0.411	0.462	0.390	0.267	0.627	
		95 %	0.452	0.452	0.458	0.437	0.419	0.469	0.397	0.271	0.637	
		100 %	0.458	0.454	0.460	0.439	0.421	0.471	0.400	0.273	0.640	
	L	$\bar{x}$	0.892	0.770	0.732	0.722	0.715	0.751	0.669	0.503	0.888	
		$s(x)$	0.069	0.053	0.049	0.049	0.049	0.049	0.046	0.045	0.035	
		0 %	0.510	0.525	0.504	0.496	0.490	0.525	0.466	0.311	0.715	
		5 %	0.776	0.674	0.642	0.633	0.626	0.663	0.584	0.420	0.827	
		50 %	0.915	0.788	0.748	0.739	0.731	0.767	0.684	0.518	0.900	
		95 %	0.932	0.811	0.771	0.761	0.754	0.790	0.702	0.541	0.911	
		100 %	0.940	0.825	0.783	0.773	0.765	0.802	0.713	0.555	0.918	
	Spezifität	N	$\bar{x}$	0.800	0.779	0.758	0.740	0.724	0.758	0.604	0.508	0.844
			$s(x)$	0.004	0.004	0.004	0.004	0.004	0.004	0.005	0.004	0.004
			0 %	0.785	0.763	0.742	0.725	0.709	0.744	0.587	0.494	0.829
			5 %	0.793	0.773	0.752	0.734	0.718	0.753	0.597	0.503	0.837
			50 %	0.800	0.779	0.758	0.741	0.724	0.759	0.605	0.509	0.844
			95 %	0.807	0.785	0.763	0.746	0.730	0.764	0.611	0.514	0.849
			100 %	0.812	0.789	0.766	0.749	0.733	0.768	0.617	0.518	0.854
		A	$\bar{x}$	0.785	0.710	0.732	0.700	0.672	0.714	0.629	0.465	0.862
			$s(x)$	0.022	0.015	0.010	0.010	0.010	0.009	0.010	0.009	0.008
			0 %	0.701	0.666	0.702	0.671	0.643	0.684	0.589	0.438	0.831
5 %			0.744	0.680	0.712	0.679	0.651	0.695	0.611	0.448	0.849	
50 %			0.787	0.713	0.733	0.700	0.672	0.715	0.630	0.466	0.863	
95 %			0.812	0.729	0.744	0.712	0.685	0.726	0.643	0.477	0.873	
100 %			0.820	0.736	0.749	0.717	0.690	0.730	0.646	0.480	0.874	
L	$\bar{x}$	0.774	0.757	0.792	0.792	0.791	0.791	0.634	0.543	0.866		
	$s(x)$	0.006	0.004	0.003	0.003	0.003	0.003	0.005	0.004	0.004		
	0 %	0.755	0.743	0.781	0.780	0.780	0.780	0.617	0.531	0.853		
	5 %	0.764	0.749	0.787	0.786	0.785	0.785	0.624	0.537	0.859		
	50 %	0.775	0.758	0.793	0.792	0.792	0.791	0.634	0.543	0.866		
	95 %	0.783	0.764	0.797	0.797	0.796	0.796	0.641	0.548	0.871		
	100 %	0.789	0.767	0.799	0.799	0.798	0.798	0.648	0.551	0.876		

## B.4 Ensemble-Modelle

Kenngröße	Klasse	lymph-agg/LR-soft					wMAE	wRMSE	wRMAE	wMSE	
		(hard)	(crisp)	weak	prd	strong					
Sensitivität	N	$\bar{x}$	0.875	0.777	0.735	0.691	0.652	0.742	0.658	0.492	0.883
		$s(x)$	0.005	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.003
		0 %	0.867	0.770	0.724	0.681	0.641	0.732	0.648	0.482	0.876
		5 %	0.867	0.771	0.730	0.686	0.646	0.737	0.653	0.487	0.880
		50 %	0.874	0.778	0.735	0.691	0.652	0.741	0.658	0.491	0.883
		95 %	0.883	0.782	0.740	0.696	0.656	0.746	0.664	0.496	0.887
		100 %	0.886	0.785	0.741	0.697	0.657	0.747	0.664	0.497	0.887

## B Details zu den Modellen

Kenngröße	Klasse		lymph-agg/LR-soft					wMAE	wRMSE	wRMAE	wMSE	
			(hard)	(crisp)	weak	prd	strong					
	A	$\bar{x}$	0.731	0.680	0.679	0.655	0.634	0.681	0.568	0.435	0.814	
		$s(x)$	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.002	
		0 %	0.725	0.677	0.677	0.653	0.632	0.678	0.565	0.433	0.811	
		5 %	0.727	0.678	0.677	0.653	0.632	0.678	0.565	0.433	0.811	
		50 %	0.732	0.680	0.680	0.655	0.635	0.681	0.569	0.435	0.814	
		95 %	0.735	0.682	0.681	0.657	0.636	0.683	0.570	0.437	0.815	
		100 %	0.735	0.683	0.682	0.657	0.637	0.683	0.570	0.437	0.815	
	L	$\bar{x}$	0.807	0.775	0.732	0.723	0.718	0.757	0.602	0.508	0.842	
		$s(x)$	0.016	0.015	0.014	0.014	0.014	0.014	0.014	0.014	0.011	
		0 %	0.761	0.729	0.690	0.681	0.676	0.716	0.561	0.467	0.808	
		5 %	0.785	0.750	0.709	0.701	0.695	0.735	0.581	0.486	0.824	
		50 %	0.809	0.776	0.733	0.724	0.718	0.758	0.604	0.508	0.843	
		95 %	0.825	0.790	0.746	0.737	0.732	0.772	0.617	0.522	0.854	
		100 %	0.826	0.792	0.748	0.739	0.734	0.773	0.617	0.524	0.854	
	Spezifität	N	$\bar{x}$	0.840	0.794	0.784	0.762	0.741	0.781	0.658	0.532	0.883
			$s(x)$	0.002	0.002	0.002	0.001	0.001	0.002	0.002	0.002	0.001
			0 %	0.836	0.791	0.781	0.758	0.738	0.777	0.653	0.528	0.880
			5 %	0.837	0.792	0.782	0.759	0.738	0.779	0.655	0.530	0.881
50 %			0.841	0.794	0.785	0.762	0.741	0.782	0.659	0.533	0.884	
95 %			0.843	0.796	0.786	0.763	0.743	0.783	0.660	0.534	0.884	
100 %			0.843	0.796	0.786	0.763	0.743	0.783	0.660	0.534	0.884	
A		$\bar{x}$	0.880	0.796	0.767	0.730	0.699	0.764	0.674	0.515	0.893	
		$s(x)$	0.005	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.003	
		0 %	0.871	0.787	0.759	0.722	0.690	0.756	0.666	0.507	0.888	
		5 %	0.874	0.790	0.759	0.722	0.691	0.757	0.666	0.507	0.888	
		50 %	0.880	0.797	0.768	0.731	0.699	0.766	0.674	0.516	0.894	
		95 %	0.888	0.801	0.771	0.734	0.703	0.768	0.679	0.519	0.897	
		100 %	0.888	0.801	0.772	0.734	0.703	0.769	0.680	0.519	0.897	
L		$\bar{x}$	0.951	0.930	0.938	0.937	0.937	0.937	0.815	0.748	0.966	
		$s(x)$	0.001	0.001	0.001	0.001	0.001	0.001	0.002	0.001	0.001	
		0 %	0.948	0.928	0.937	0.936	0.936	0.935	0.812	0.746	0.965	
		5 %	0.949	0.928	0.937	0.936	0.936	0.935	0.813	0.746	0.965	
	50 %	0.951	0.930	0.938	0.937	0.937	0.937	0.815	0.748	0.966		
	95 %	0.952	0.931	0.939	0.938	0.938	0.938	0.817	0.750	0.967		
	100 %	0.952	0.931	0.939	0.939	0.938	0.938	0.818	0.750	0.967		

Kenngröße	Klasse		lymph-agg/LDA					wMAE	wRMSE	wRMAE	wMSE
			(hard)	(crisp)	weak	prd	strong				
Sensitivität	N	$\bar{x}$	0.839	0.815	0.698	0.673	0.649	0.713	0.565	0.464	0.811
		$s(x)$	0.007	0.006	0.005	0.005	0.005	0.004	0.004	0.004	0.004
		0 %	0.828	0.806	0.689	0.663	0.639	0.704	0.556	0.456	0.803
		5 %	0.830	0.807	0.692	0.666	0.642	0.707	0.559	0.459	0.806
		50 %	0.839	0.815	0.697	0.672	0.649	0.713	0.565	0.464	0.811
		95 %	0.850	0.824	0.704	0.679	0.656	0.719	0.572	0.470	0.816
		100 %	0.850	0.825	0.706	0.681	0.658	0.720	0.573	0.471	0.818

## B.5 Ensemble-Modell: 9 Modelle und 9 Spektren aggregiert

Kenngröße	Klasse	lymph-agg/LDA					wMAE	wRMSE	wRMAE	wMSE	
		(hard)	(crisp)	weak	prd	strong					
Spezifität	A	$\bar{x}$	0.780	0.755	0.742	0.727	0.713	0.740	0.574	0.490	0.819
		$s(x)$	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
		0 %	0.777	0.751	0.739	0.724	0.710	0.737	0.571	0.487	0.816
		5 %	0.777	0.752	0.739	0.724	0.710	0.738	0.572	0.488	0.817
		50 %	0.779	0.755	0.741	0.727	0.713	0.739	0.573	0.489	0.818
		95 %	0.784	0.759	0.745	0.730	0.717	0.743	0.579	0.493	0.822
		100 %	0.784	0.760	0.746	0.732	0.718	0.745	0.580	0.495	0.824
	L	$\bar{x}$	0.911	0.892	0.835	0.830	0.825	0.865	0.706	0.633	0.913
		$s(x)$	0.006	0.006	0.006	0.006	0.006	0.006	0.007	0.008	0.004
		0 %	0.897	0.877	0.821	0.816	0.811	0.852	0.689	0.615	0.903
		5 %	0.903	0.884	0.828	0.823	0.818	0.858	0.696	0.623	0.908
		50 %	0.910	0.892	0.834	0.829	0.824	0.865	0.705	0.632	0.913
		95 %	0.918	0.901	0.843	0.838	0.833	0.873	0.716	0.644	0.919
		100 %	0.919	0.901	0.843	0.838	0.834	0.874	0.716	0.645	0.919
	N	$\bar{x}$	0.883	0.873	0.851	0.838	0.826	0.843	0.670	0.604	0.891
		$s(x)$	0.002	0.001	0.001	0.001	0.002	0.001	0.002	0.002	0.001
		0 %	0.878	0.870	0.849	0.836	0.823	0.841	0.667	0.601	0.889
		5 %	0.880	0.871	0.849	0.836	0.823	0.841	0.667	0.602	0.889
		50 %	0.883	0.873	0.851	0.838	0.826	0.843	0.670	0.604	0.891
		95 %	0.885	0.875	0.853	0.840	0.828	0.845	0.672	0.606	0.893
		100 %	0.885	0.875	0.853	0.840	0.828	0.846	0.674	0.607	0.894
	A	$\bar{x}$	0.859	0.837	0.735	0.713	0.692	0.738	0.584	0.489	0.827
		$s(x)$	0.005	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.003
		0 %	0.851	0.831	0.728	0.705	0.684	0.732	0.577	0.482	0.821
5 %		0.852	0.831	0.730	0.707	0.685	0.733	0.579	0.483	0.822	
50 %		0.858	0.836	0.735	0.712	0.691	0.739	0.585	0.489	0.828	
95 %		0.868	0.844	0.741	0.719	0.698	0.743	0.590	0.493	0.832	
100 %		0.868	0.844	0.742	0.720	0.700	0.744	0.591	0.495	0.833	
L	$\bar{x}$	0.934	0.922	0.932	0.932	0.932	0.931	0.785	0.737	0.954	
	$s(x)$	0.002	0.002	0.001	0.001	0.001	0.001	0.003	0.002	0.001	
	0 %	0.931	0.919	0.930	0.930	0.930	0.929	0.780	0.733	0.952	
	5 %	0.932	0.919	0.931	0.930	0.930	0.929	0.781	0.733	0.952	
	50 %	0.934	0.922	0.932	0.932	0.932	0.931	0.784	0.737	0.954	
	95 %	0.936	0.924	0.934	0.934	0.934	0.932	0.789	0.740	0.955	
	100 %	0.937	0.925	0.935	0.934	0.934	0.933	0.789	0.741	0.956	

## B.5 Ensemble-Modell: 9 Modelle und 9 Spektren aggregiert

Kenngröße	Klasse	lymph/LR-soft, Referenz: Patient					wMAE	wRMSE	wRMAE	wMSE	
		(hard)	(crisp)	weak	prd	strong					
Sensitivität	A	$\bar{x}$	0.959	0.871	0.871	0.871	0.871	0.871	0.798	0.641	0.959
		$s(x)$	0.004	0.001	0.001	0.001	0.001	0.001	0.004	0.002	0.001
		0 %	0.953	0.868	0.868	0.868	0.868	0.868	0.792	0.637	0.957
		5 %	0.953	0.869	0.869	0.869	0.869	0.869	0.792	0.638	0.957
		50 %	0.959	0.872	0.872	0.872	0.872	0.872	0.799	0.642	0.960
		95 %	0.965	0.873	0.873	0.873	0.873	0.873	0.802	0.643	0.961
		100 %	0.967	0.873	0.873	0.873	0.873	0.873	0.803	0.644	0.961

## B Details zu den Modellen

Kenngröße	Klasse	lymph/LR-soft, Referenz: Patient					wMAE	wRMSE	wRMAE	wMSE		
		(hard)	(crisp)	weak	prd	strong						
Spezifität	L	$\bar{x}$	0.852	0.731	0.731	0.731	0.731	0.731	0.645	0.482	0.874	
		$s(x)$	0.016	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.014	0.011
		0 %	0.803	0.686	0.686	0.686	0.686	0.686	0.686	0.601	0.440	0.841
		5 %	0.828	0.712	0.712	0.712	0.712	0.712	0.712	0.624	0.464	0.859
		50 %	0.854	0.734	0.734	0.734	0.734	0.734	0.734	0.646	0.484	0.875
		95 %	0.869	0.747	0.747	0.747	0.747	0.747	0.747	0.660	0.497	0.885
		100 %	0.873	0.752	0.752	0.752	0.752	0.752	0.752	0.664	0.502	0.887
	A	$\bar{x}$	0.836	0.721	0.721	0.721	0.721	0.721	0.628	0.472	0.861	
		$s(x)$	0.015	0.014	0.014	0.014	0.014	0.014	0.013	0.013	0.010	
		0 %	0.793	0.679	0.679	0.679	0.679	0.679	0.679	0.589	0.433	0.831
		5 %	0.817	0.702	0.702	0.702	0.702	0.702	0.702	0.610	0.454	0.848
		50 %	0.838	0.723	0.723	0.723	0.723	0.723	0.723	0.629	0.474	0.862
		95 %	0.854	0.735	0.735	0.735	0.735	0.735	0.735	0.641	0.485	0.871
		100 %	0.860	0.739	0.739	0.739	0.739	0.739	0.739	0.643	0.489	0.873
L	$\bar{x}$	0.958	0.870	0.870	0.870	0.870	0.870	0.795	0.640	0.958		
	$s(x)$	0.004	0.001	0.001	0.001	0.001	0.001	0.004	0.002	0.001		
	0 %	0.952	0.867	0.867	0.867	0.867	0.867	0.867	0.790	0.636	0.956	
	5 %	0.952	0.868	0.868	0.868	0.868	0.868	0.868	0.790	0.637	0.956	
	50 %	0.958	0.871	0.871	0.871	0.871	0.871	0.871	0.796	0.640	0.958	
	95 %	0.964	0.872	0.872	0.872	0.872	0.872	0.872	0.799	0.642	0.960	
	100 %	0.966	0.872	0.872	0.872	0.872	0.872	0.872	0.800	0.642	0.960	

Kenngröße	Klasse	lymph/LDA, Referenz: Patient					wMAE	wRMSE	wRMAE	wMSE		
		(hard)	(crisp)	weak	prd	strong						
Sensitivität	A	$\bar{x}$	0.951	0.884	0.884	0.884	0.884	0.884	0.800	0.659	0.960	
		$s(x)$	0.004	0.002	0.002	0.002	0.002	0.002	0.004	0.003	0.002	
		0 %	0.945	0.880	0.880	0.880	0.880	0.880	0.880	0.794	0.654	0.957
		5 %	0.946	0.880	0.880	0.880	0.880	0.880	0.880	0.794	0.654	0.958
		50 %	0.950	0.884	0.884	0.884	0.884	0.884	0.884	0.800	0.659	0.960
		95 %	0.957	0.888	0.888	0.888	0.888	0.888	0.888	0.807	0.665	0.963
		100 %	0.958	0.888	0.888	0.888	0.888	0.888	0.888	0.808	0.666	0.963
	L	$\bar{x}$	0.733	0.694	0.694	0.694	0.694	0.694	0.694	0.531	0.447	0.780
		$s(x)$	0.002	0.006	0.006	0.006	0.006	0.006	0.006	0.005	0.006	0.004
		0 %	0.730	0.682	0.682	0.682	0.682	0.682	0.682	0.523	0.436	0.773
		5 %	0.730	0.685	0.685	0.685	0.685	0.685	0.685	0.523	0.439	0.773
		50 %	0.734	0.695	0.695	0.695	0.695	0.695	0.695	0.531	0.448	0.780
		95 %	0.735	0.704	0.704	0.704	0.704	0.704	0.704	0.537	0.456	0.786
		100 %	0.736	0.705	0.705	0.705	0.705	0.705	0.705	0.538	0.457	0.787
Spezifität	A	$\bar{x}$	0.712	0.677	0.677	0.677	0.677	0.677	0.515	0.432	0.764	
		$s(x)$	0.004	0.006	0.006	0.006	0.006	0.006	0.006	0.005	0.005	0.005
		0 %	0.706	0.669	0.669	0.669	0.669	0.669	0.669	0.507	0.425	0.757
		5 %	0.707	0.669	0.669	0.669	0.669	0.669	0.669	0.508	0.425	0.758
		50 %	0.713	0.678	0.678	0.678	0.678	0.678	0.678	0.514	0.433	0.764
		95 %	0.716	0.685	0.685	0.685	0.685	0.685	0.685	0.522	0.439	0.771
		100 %	0.717	0.689	0.689	0.689	0.689	0.689	0.689	0.522	0.443	0.772

### B.5 Ensemble-Modell: 9 Modelle und 9 Spektren aggregiert

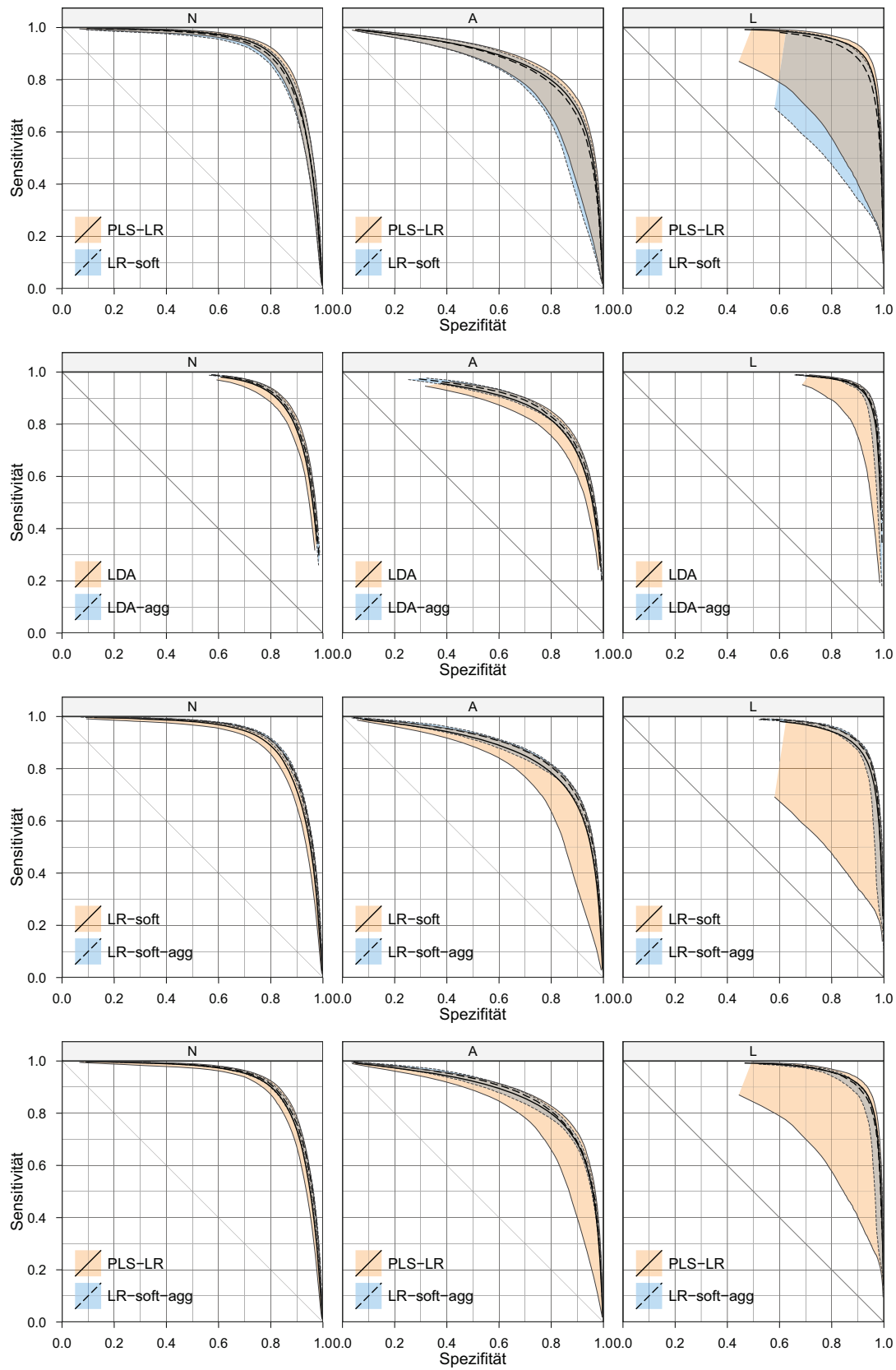
Kenngröße	Klasse		lymph/LDA, Referenz: Patient					wMAE	wRMSE	wRMAE	wMSE
			(hard)	(crisp)	weak	prd	strong				
	L	$\bar{x}$	0.950	0.884	0.884	0.884	0.884	0.884	0.800	0.659	0.960
		$s(x)$	0.004	0.002	0.002	0.002	0.002	0.002	0.004	0.003	0.002
		0 %	0.944	0.880	0.880	0.880	0.880	0.880	0.794	0.654	0.957
		5 %	0.946	0.880	0.880	0.880	0.880	0.880	0.794	0.654	0.957
		50 %	0.950	0.883	0.883	0.883	0.883	0.883	0.799	0.659	0.960
		95 %	0.957	0.888	0.888	0.888	0.888	0.888	0.807	0.665	0.963
		100 %	0.958	0.888	0.888	0.888	0.888	0.888	0.808	0.665	0.963

Kenngröße	Klasse		lymph/LR-soft, Referenz: detaillierte Histologie					wMAE	wRMSE	wRMAE	wMSE
			(hard)	(crisp)	weak	prd	strong				
Sensitivität	A	$\bar{x}$	0.956	0.870	0.876	0.871	0.866	0.864	0.787	0.631	0.955
		$s(x)$	0.005	0.002	0.002	0.002	0.001	0.002	0.004	0.002	0.002
		0 %	0.949	0.867	0.873	0.868	0.863	0.861	0.780	0.627	0.951
		5 %	0.951	0.868	0.873	0.868	0.864	0.861	0.781	0.628	0.952
		50 %	0.957	0.871	0.876	0.872	0.867	0.864	0.790	0.631	0.956
		95 %	0.963	0.872	0.877	0.873	0.868	0.866	0.790	0.634	0.956
		100 %	0.965	0.873	0.878	0.873	0.868	0.867	0.791	0.635	0.957
	L	$\bar{x}$	0.963	0.802	0.805	0.791	0.783	0.803	0.756	0.557	0.940
		$s(x)$	0.017	0.015	0.015	0.015	0.015	0.015	0.017	0.016	0.009
		0 %	0.910	0.756	0.759	0.745	0.737	0.759	0.703	0.509	0.912
		5 %	0.940	0.780	0.783	0.769	0.761	0.782	0.731	0.533	0.927
		50 %	0.967	0.804	0.807	0.793	0.784	0.805	0.759	0.558	0.942
		95 %	0.976	0.818	0.820	0.806	0.797	0.819	0.772	0.574	0.948
		100 %	0.978	0.823	0.826	0.811	0.802	0.824	0.779	0.580	0.951
Spezifität	A	$\bar{x}$	0.854	0.720	0.567	0.545	0.523	0.623	0.518	0.386	0.768
		$s(x)$	0.015	0.013	0.012	0.012	0.011	0.012	0.010	0.010	0.009
		0 %	0.808	0.680	0.531	0.510	0.490	0.588	0.492	0.358	0.742
		5 %	0.834	0.702	0.551	0.528	0.506	0.607	0.504	0.373	0.754
		50 %	0.855	0.721	0.567	0.546	0.525	0.624	0.520	0.387	0.769
		95 %	0.870	0.737	0.578	0.556	0.534	0.635	0.529	0.396	0.778
		100 %	0.871	0.738	0.579	0.557	0.536	0.637	0.529	0.397	0.778
	L	$\bar{x}$	0.956	0.870	0.870	0.868	0.867	0.870	0.795	0.639	0.958
		$s(x)$	0.005	0.002	0.001	0.001	0.001	0.001	0.004	0.002	0.001
		0 %	0.948	0.867	0.867	0.865	0.864	0.867	0.790	0.635	0.956
		5 %	0.950	0.867	0.868	0.866	0.865	0.868	0.790	0.636	0.956
		50 %	0.956	0.871	0.870	0.868	0.867	0.870	0.796	0.640	0.958
		95 %	0.962	0.872	0.871	0.870	0.869	0.872	0.799	0.642	0.960
		100 %	0.964	0.872	0.871	0.870	0.869	0.872	0.801	0.642	0.960

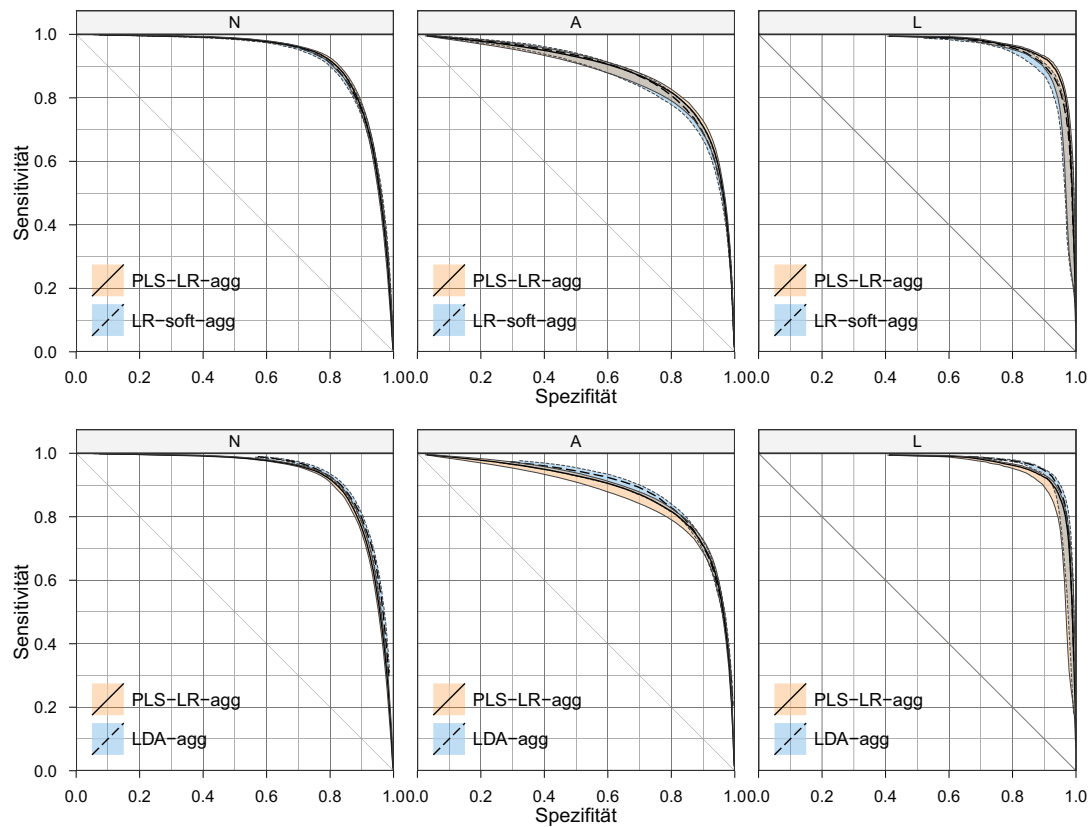
B Details zu den Modellen

Kenngröße	Klasse	lymph/LDA, Referenz: detaillierte Histologie									
		(hard)	(crisp)	weak	prd	strong	wMAE	wRMSE	wRMAE	wMSE	
Sensitivität	A	$\bar{x}$	0.942	0.875	0.882	0.877	0.872	0.863	0.772	0.630	0.948
		$s(x)$	0.004	0.003	0.003	0.003	0.003	0.003	0.004	0.003	0.002
		0 %	0.934	0.870	0.878	0.873	0.868	0.858	0.765	0.624	0.945
		5 %	0.936	0.871	0.878	0.873	0.868	0.859	0.765	0.624	0.945
		50 %	0.941	0.875	0.882	0.877	0.872	0.863	0.772	0.629	0.948
		95 %	0.949	0.879	0.886	0.881	0.876	0.867	0.778	0.635	0.951
		100 %	0.951	0.880	0.886	0.882	0.876	0.867	0.778	0.636	0.951
	L	$\bar{x}$	0.997	0.898	0.872	0.856	0.847	0.891	0.843	0.670	0.975
		$s(x)$	0.001	0.007	0.007	0.007	0.007	0.007	0.007	0.010	0.002
		0 %	0.995	0.883	0.858	0.841	0.832	0.878	0.829	0.650	0.971
		5 %	0.996	0.888	0.862	0.846	0.837	0.882	0.832	0.656	0.972
		50 %	0.997	0.898	0.872	0.856	0.847	0.891	0.842	0.670	0.975
		95 %	0.998	0.908	0.883	0.866	0.858	0.902	0.854	0.686	0.979
		100 %	0.998	0.909	0.883	0.866	0.858	0.902	0.854	0.687	0.979
Spezifität	A	$\bar{x}$	0.710	0.656	0.407	0.389	0.370	0.478	0.352	0.277	0.580
		$s(x)$	0.012	0.011	0.005	0.005	0.005	0.006	0.005	0.004	0.007
		0 %	0.691	0.641	0.401	0.383	0.364	0.469	0.343	0.271	0.568
		5 %	0.694	0.643	0.401	0.384	0.364	0.470	0.344	0.272	0.570
		50 %	0.710	0.654	0.405	0.387	0.368	0.477	0.353	0.277	0.581
		95 %	0.729	0.675	0.416	0.397	0.377	0.487	0.359	0.284	0.589
		100 %	0.733	0.676	0.417	0.398	0.378	0.488	0.361	0.284	0.592
	L	$\bar{x}$	0.944	0.878	0.885	0.883	0.883	0.883	0.800	0.658	0.960
		$s(x)$	0.004	0.003	0.002	0.002	0.002	0.002	0.004	0.003	0.002
		0 %	0.937	0.874	0.882	0.880	0.879	0.880	0.795	0.653	0.958
		5 %	0.938	0.874	0.882	0.880	0.879	0.880	0.795	0.654	0.958
		50 %	0.943	0.878	0.885	0.883	0.882	0.883	0.800	0.658	0.960
		95 %	0.951	0.882	0.889	0.887	0.887	0.887	0.808	0.664	0.963
		100 %	0.952	0.882	0.889	0.888	0.887	0.887	0.809	0.664	0.963

### B.5.1 Weitere Spezifiäts-Sensitivitäts-Diagramme



## B Details zu den Modellen



### B.5.2 Ergebnisse Harte Vorhersage

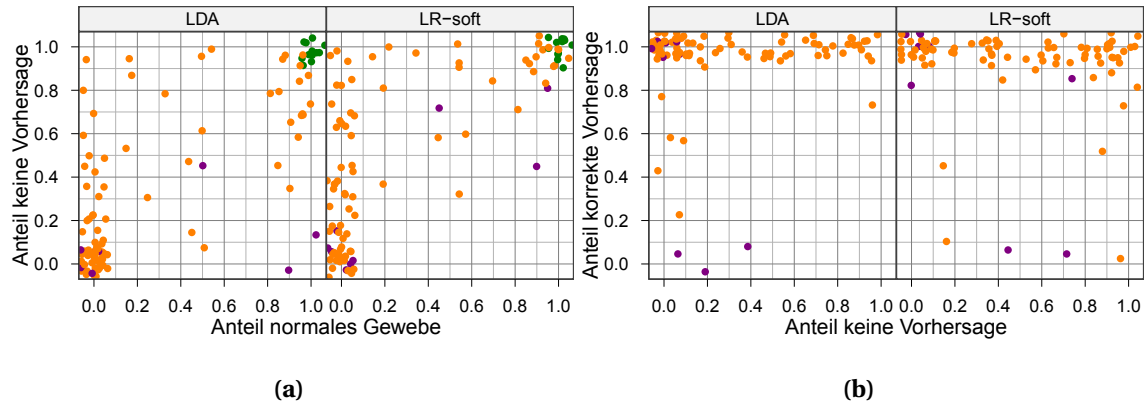
Patient	Messung	Diagnose	normales Gewebe Anteil	LR-soft			LDA						
				Anzahl ✓	Anzahl ✗	k. A.	Anteil ✓	Anteil k. A.	Anzahl ✓	Anzahl ✗	k. A.	Anteil ✓	Anteil k. A.
46	46.6	N	1.00	0	3668	1.00	0	3668	1.00				
53	53.4	N	1.00	17	4197	1.00	44	4170	0.99				
60	60.3	N	1.00	52	3336	0.98	60	3328	0.98				
1401	1401.5	N	1.00	20	4992	1.00	44	4968	0.99				
	1401.7	N	1.00	49	5047	0.99	30	5066	0.99				
1402	1402.3	N	1.00	217	5971	0.96	329	5859	0.95				
	1402.5	N	1.00	27	4803	0.99	114	4716	0.98				
1403	1403.3	N	1.00	114	5640	0.98	172	5582	0.97				
	1403.5	N	1.00	24	4666	0.99	49	4641	0.99				
1404	1404.3	N	1.00	92	7678	0.99	86	7684	0.99				
1405	1405.3	N	1.00	86	9658	0.99	28	9716	1.00				
1406	1406.2	N	1.00	0	8274	1.00	14	8260	1.00				
590	1579.2	A	0.00	5633	0	527	1.00	0.09	5930	0	230	1.00	0.04
720	1466.2	A	0.00	1209	0	4209	1.00	0.78	1840	0	3578	1.00	0.66
856	1624.2	A	0.00	2472	283	2117	0.90	0.43	4723	1	148	1.00	0.03
958	1605.2	A	0.20	814	0	3484	1.00	0.81	2175	0	2123	1.00	0.49
1204	1204.2	A	0.00	4066	71	749	0.98	0.15	4877	7	2	1.00	0.00
	1501.2	A	0.00	2719	0	3777	1.00	0.58	3224	0	3272	1.00	0.50
1218	1218.2	A	0.95	114	40	5096	0.74	0.97	653	0	4597	1.00	0.88
1220	1393.2	A	0.00	3510	0	1250	1.00	0.26	4350	0	410	1.00	0.09
1226	1485.2	A	0.00	5571	0	939	1.00	0.14	6481	0	29	1.00	0.00
1235	1235.2	A	0.90	393	27	3598	0.94	0.90	1969	0	2049	1.00	0.51
1240	1240.2	A	0.33	323	39	5476	0.89	0.94	1396	0	4442	1.00	0.76
1241	1484.2	A	0.00	5238	0	3386	1.00	0.39	7928	0	696	1.00	0.08
1266	1266.2	A	0.97	28	4	5638	0.88	0.99	1542	0	4128	1.00	0.73
1268	1327.4	A	0.60	0	44	3148	0.00	0.99	104	30	3058	0.78	0.96
	1327.5	A	0.95	496	0	4040	1.00	0.89	1655	0	2881	1.00	0.64



### B.5 Ensemble-Modell: 9 Modelle und 9 Spektren aggregiert

Patient	Messung	Diagnose	normales Gewebe Anteil	LR-soft						LDA					
				Anzahl			Anteil			Anzahl			Anteil		
				✓	✗	k. A.	✓	k. A.	✓	✗	k. A.	✓	k. A.		
1292	1292.2	A	0.95	12	0	4664	1.00	1.00	1475	0	3201	1.00	0.68		
1313	1313.2	A	0.00	7942	0	52	1.00	0.01	7980	0	14	1.00	0.00		
1316	1316.4	A	0.50	3471	23	2078	0.99	0.37	5042	64	466	0.99	0.08		
1320	1320.2	A	0.00	1721	3	3428	1.00	0.67	4791	0	361	1.00	0.07		
1322	1322.2	A	0.00	2366	2	488	1.00	0.17	2646	42	168	0.98	0.06		
	1399.2	A	0.00	2607	0	1523	1.00	0.37	3477	0	653	1.00	0.16		
1332	1332.3	A	1.00	833	0	5383	1.00	0.87	2610	0	3606	1.00	0.58		
1337	1337.2	A	0.00	6406	0	328	1.00	0.05	6596	0	138	1.00	0.02		
1338	1338.2	A	0.00	4157	0	267	1.00	0.06	4392	0	32	1.00	0.01		
1339	1339.2	A	0.00	9394	0	0	1.00	0.00	9394	0	0	1.00	0.00		
	1595.2	A	0.00	3715	0	2403	1.00	0.39	5149	0	969	1.00	0.16		
1340	1340.4	A	0.00	1109	34	2455	0.97	0.68	3106	7	485	1.00	0.13		
1343	1343.2	A	0.55	158	0	4938	1.00	0.97	2271	0	2825	1.00	0.55		
1360	1360.2	A	0.00	4985	0	13	1.00	0.00	4984	0	14	1.00	0.00		
1363	1363.2	A	0.91	51	4	3165	0.93	0.98	319	0	2901	1.00	0.90		
1368	1368.2	A	0.00	1405	0	7765	1.00	0.85	1603	1	7566	1.00	0.83		
1369	1498.2	A	0.00	2593	0	403	1.00	0.13	2903	0	93	1.00	0.03		
1375	1375.2	A	0.84	1223	0	4055	1.00	0.77	3428	0	1850	1.00	0.35		
1377	1377.3	A	0.00	1303	0	2169	1.00	0.62	1317	0	2155	1.00	0.62		
1378	1378.2	A	0.00	3367	0	343	1.00	0.09	3682	0	28	1.00	0.01		
1382	1382.3	A	0.22	334	0	8500	1.00	0.96	741	0	8093	1.00	0.92		
1383	1383.4	A	0.95	153	126	3627	0.55	0.93	1215	0	2691	1.00	0.69		
	1383.5	A	0.00	2825	0	1865	1.00	0.40	3657	0	1033	1.00	0.22		
1395	1395.2	A	0.00	4038	0	64	1.00	0.02	4102	0	0	1.00	0.00		
	1437.2	A	0.00	6064	26	14	1.00	0.00	6082	22	0	1.00	0.00		
1408	1408.3	A	0.00	169	0	6173	1.00	0.97	158	0	6184	1.00	0.98		
1412	1412.2	A	0.00	5886	0	246	1.00	0.04	6132	0	0	1.00	0.00		
1417	1417.2	A	0.00	3120	4742	1770	0.40	0.18	2268	6420	944	0.26	0.10		
1424	1424.2	A	0.00	4563	0	1	1.00	0.00	4540	0	24	1.00	0.01		
	1559.2	A	0.21	4081	0	2093	1.00	0.34	4639	0	1535	1.00	0.25		
1426	1426.2	A	0.00	5980	0	292	1.00	0.05	5968	0	304	1.00	0.05		
1432	1432.2	A	0.00	759	0	1033	1.00	0.58	1711	0	81	1.00	0.05		
1433	1433.2	A	0.00	441	26	3047	0.94	0.87	2081	0	1433	1.00	0.41		
1440	1440.5	A	0.48	329	0	6545	1.00	0.95	388	0	6486	1.00	0.94		
1442	1442.2	A	0.00	3864	193	2425	0.95	0.37	3095	2596	791	0.54	0.12		
1446	1446.2	A	0.00	650	0	5244	1.00	0.89	4529	0	1365	1.00	0.23		
1453	1453.3	A	0.86	1049	0	6679	1.00	0.86	1311	0	6417	1.00	0.83		
	1453.4	A	0.90	217	0	8477	1.00	0.98	387	0	8307	1.00	0.96		
1454	1454.2	A	0.00	413	2433	710	0.15	0.20	1369	2187	0	0.38	0.00		
1459	1459.2	A	0.95	569	0	3407	1.00	0.86	281	0	3695	1.00	0.93		
1461	1461.5	A	0.47	2742	0	3656	1.00	0.57	3352	0	3046	1.00	0.48		
	1461.6	A	0.00	696	31	1779	0.96	0.71	1873	213	420	0.90	0.17		
1462	1462.2	A	0.00	4770	0	4	1.00	0.00	4773	1	0	1.00	0.00		
1465	1465.2	A	1.00	438	0	8774	1.00	0.95	2678	0	6534	1.00	0.71		
1469	1469.3	A	0.10	338	0	6088	1.00	0.95	418	0	6008	1.00	0.93		
1470	1470.3	A	0.00	3058	64	224	0.98	0.07	1970	1267	109	0.61	0.03		
	1470.4	A	0.00	6081	71	50	0.99	0.01	5155	1045	2	0.83	0.00		
	1470.5	A	0.00	7800	0	740	1.00	0.09	8455	1	84	1.00	0.01		
1471	1471.2	A	0.00	3295	0	37	1.00	0.01	3332	0	0	1.00	0.00		
1486	1486.2	A	0.00	2088	0	4380	1.00	0.68	3172	0	3296	1.00	0.51		
1487	1487.2	A	0.07	3834	0	2746	1.00	0.42	4914	2	1664	1.00	0.25		
1490	1490.3	A	0.52	2076	193	3471	0.91	0.60	5194	113	433	0.98	0.08		
1496	1496.2	A	1.00	36	0	5648	1.00	0.99	284	0	5400	1.00	0.95		
	1496.3	A	1.00	696	0	7340	1.00	0.91	790	0	7246	1.00	0.90		
1503	1503.3	A	0.00	3382	90	98	0.97	0.03	3511	29	30	0.99	0.01		
1511	1560.2	A	0.00	3672	0	1802	1.00	0.33	4987	0	487	1.00	0.09		
1515	1515.2	A	0.00	4698	0	454	1.00	0.09	5107	0	45	1.00	0.01		
1516	1516.3	A	0.75	957	0	5301	1.00	0.85	1550	0	4708	1.00	0.75		
1535	1535.2	A	0.00	4850	44	706	0.99	0.13	5483	2	115	1.00	0.02		
1541	1541.3	A	0.00	294	52	3098	0.85	0.90	1754	0	1690	1.00	0.49		
1549	1549.2	A	0.00	4047	0	41	1.00	0.01	4088	0	0	1.00	0.00		
1591	1591.2	A	0.00	3506	0	2710	1.00	0.44	4138	0	2078	1.00	0.33		
1594	1594.2	A	0.00	1685	0	4335	1.00	0.72	3680	0	2340	1.00	0.39		
1601	1601.2	A	0.00	4118	0	40	1.00	0.01	4130	0	28	1.00	0.01		
1627	1627.2	A	0.00	3275	2	1917	1.00	0.37	5014	0	180	1.00	0.03		
1630	1630.3	A	0.00	1182	0	652	1.00	0.36	1807	0	27	1.00	0.01		

## B Details zu den Modellen



**Abbildung B.1** (a) Korrelation zwischen dem Anteil an zurückgewiesenen Spektren und dem Anteil an normalem Gewebe in der Probe. (b) Eine Korrelation zwischen dem Anteil an zurückgewiesenen Spektren und dem Anteil an korrekten Zuordnungen ist nicht erkennbar.

Patient	Messung	Diagnose	normales Gewebe Anteil	LR-soft			LDA						
				Anzahl ✓	Anzahl ✗	k. A.	Anzahl ✓	Anzahl ✗	k. A.	Anzahl ✓	Anzahl ✗	k. A.	
279	279.2	L	0.00	2617	0	267	1.00	0.09	2736	61	87	0.98	0.03
381	381.2	L	0.86	8	2344	2282	0.00	0.49	0	4525	109	0.00	0.02
388	388.3	L	0.49	648	163	2381	0.80	0.75	42	1886	1264	0.02	0.40
652	652.2	L	0.00	2981	0	15	1.00	0.01	2995	0	1	1.00	0.00
931	931.3	L	0.99	18	898	2682	0.02	0.75	7	3066	525	0.00	0.15
1422	1422.3	L	0.00	7513	22	165	1.00	0.02	7665	0	35	1.00	0.00
	1422.4	L	0.00	4998	162	48	0.97	0.01	5172	0	36	1.00	0.01
1434	1434.2	L	0.00	4628	1452	80	0.76	0.01	6122	38	0	0.99	0.00
1632	1632.2	L	0.00	1918	9	33	1.00	0.02	1960	0	0	1.00	0.00

✓: korrekte Vorhersage, ✗: falsche Vorhersage, k. A.: Vorhersage zurückgewiesen

# C Software

## C.1 Testprotokolle und Details zu softclassval

### C.1.1 Testprotokoll R Pakettest

```
$ R CMD check softclassval_1.0-20130317.tar.gz
* using log directory '/home/cb/Projekte/softclassval/softclassval.Rcheck'
* using R version 2.15.3 (2013-03-01)
* using platform: x86_64-pc-linux-gnu (64-bit)
* using session charset: UTF-8
* checking for file 'softclassval/DESCRIPTION' ... OK
* checking extension type ... Package
* this is package 'softclassval' version '1.0-20130317'
* checking package namespace information ... OK
* checking package dependencies ... OK
* checking if this is a source package ... OK
* checking if there is a namespace ... OK
* checking for executable files ... OK
* checking whether package 'softclassval' can be installed ... OK
* checking installed package size ... OK
* checking package directory ... OK
* checking for portable file names ... OK
* checking for sufficient/correct file permissions ... OK
* checking DESCRIPTION meta-information ... OK
* checking top-level files ... OK
* checking for left-over files ... OK
* checking index information ... OK
* checking package subdirectories ... OK
* checking R files for non-ASCII characters ... OK
* checking R files for syntax errors ... OK
* checking whether the package can be loaded ... OK
* checking whether the package can be loaded with stated dependencies ... OK
* checking whether the package can be unloaded cleanly ... OK
* checking whether the namespace can be loaded with stated dependencies ... OK
* checking whether the namespace can be unloaded cleanly ... OK
* checking loading without being on the library search path ... OK
* checking for unstated dependencies in R code ... OK
* checking S3 generic/method consistency ... OK
* checking replacement functions ... OK
* checking foreign function calls ... OK
* checking R code for possible problems ... OK
* checking Rd files ... OK
* checking Rd metadata ... OK
* checking Rd cross-references ... OK
* checking for missing documentation entries ... OK
* checking for code/documentation mismatches ... OK
* checking Rd \usage sections ... OK
* checking Rd contents ... OK
* checking for unstated dependencies in examples ... OK
* checking examples ... OK
* checking for unstated dependencies in tests ... OK
* checking tests ...
  Running 'tests.R'
  OK
* checking PDF version of manual ... OK
```

Die Unit-Tests verbergen sich hinter dem „OK“ in der vorletzten Zeile.

## C.1.2 Unit-Tests

```
> softclassval.unittest ()
              kind timing                time unit msg
test(.make01)  OK  0.001 2013-04-03 17:49:35
test(sens)     OK  0.037 2013-04-03 17:49:35
test(checkrp)  OK  0.006 2013-04-03 17:49:35
test(ppv)      OK  0.003 2013-04-03 17:49:35
test(harden)   OK  0.003 2013-04-03 17:49:35
test(confmat)  OK  0.006 2013-04-03 17:49:35
test(dev)      OK  0.001 2013-04-03 17:49:35
test(prd)      OK  0.000 2013-04-03 17:49:35
test(wRMAE)    OK  0.000 2013-04-03 17:49:35
test(strong)   OK  0.000 2013-04-03 17:49:35
test(weak)     OK  0.000 2013-04-03 17:49:35
test(and)      OK  0.001 2013-04-03 17:49:35
test(npv)      OK  0.004 2013-04-03 17:49:35
test(nsamples) OK  0.008 2013-04-03 17:49:35
test(luk)      OK  0.000 2013-04-03 17:49:35
test(wMAE)     OK  0.000 2013-04-03 17:49:35
test(hard)     OK  0.001 2013-04-03 17:49:35
test(hardclasses) OK 0.008 2013-04-03 17:49:35
test(factor2matrix) OK 0.001 2013-04-03 17:49:35
test(postproc) OK  0.000 2013-04-03 17:49:35
test(gdl)      OK  0.000 2013-04-03 17:49:35
Summary statistics on all tests run:

      OK  **FAILS**  **ERROR**  DEACTIVATED
      21         0         0         0
>
```

## C.2 Messprogramm RamanGUI

Die Messoberfläche RamanGUI (Abb. C.1 und `abfig:RamanGUI-screenshot2`) steuert die Messsoftware HoLoGRAM über deren DDE-Schnittstelle.

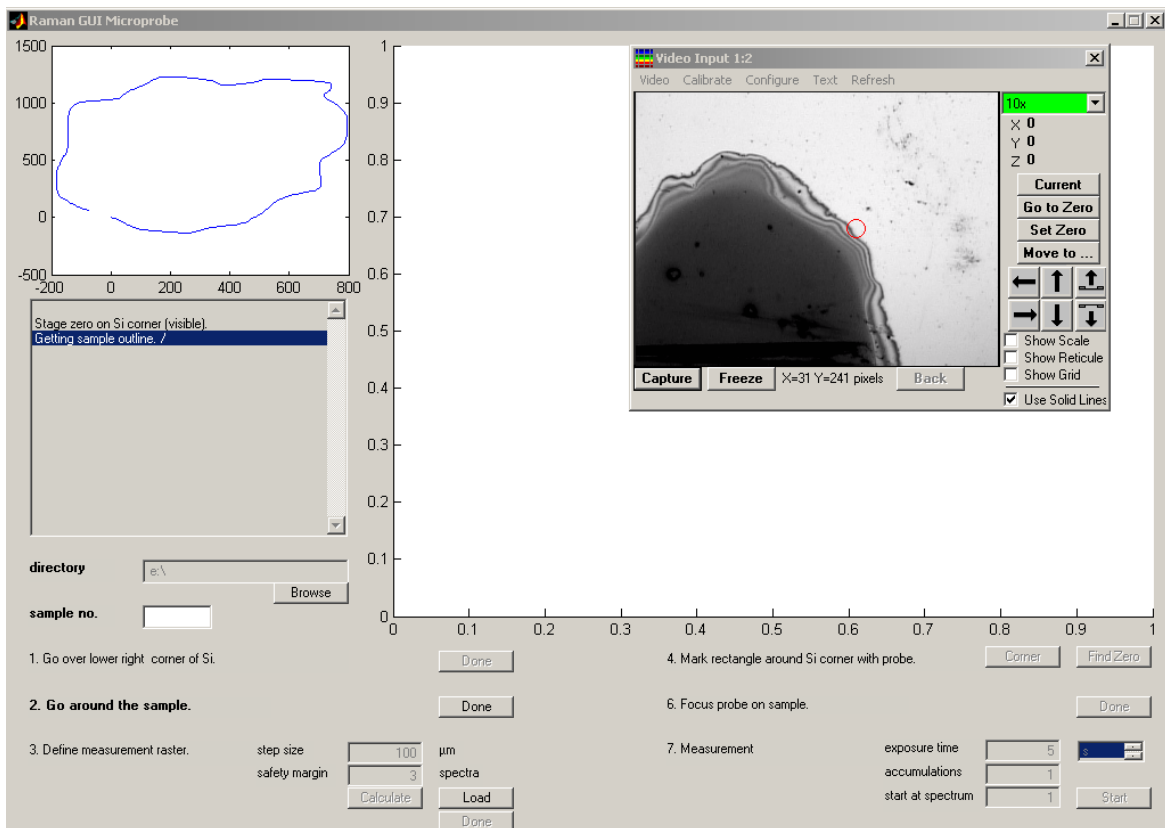
Leider steht dabei kein direkter Befehl zum Import des gemessenen Spektrums zur Verfügung. Deshalb wird das Spektrum zunächst abgespeichert und dann in Matlab eingelesen. Für alle Messungen in HoLoGRAM ist es extrem wichtig, die Anzeigeeinstellungen zu kontrollieren. Die verwendete Version von HoLoGRAM **speichert nur den aktuell eingestellten Spektralbereich ab!** Steht die Anzeige auf dem hohen oder niedrigen Wellenzahlbereich, so kann auch nur dieser Teil des Spektrums gespeichert werden (unabhängig davon, dass immer beide Bereiche gemessen werden). Die kompletten Rohspektren können nur gespeichert werden, wenn beide Spektralbereiche übereinander angezeigt werden. Wird das zusammengesetzte Spektrum angezeigt, kann nur das korrigierte, interpolierte und zusammengesetzte Spektrum, aber nicht die Rohdaten abgespeichert werden. Diese Regeln gelten auch für die automatisch von HoLoGRAM gespeicherten Spektrendateien.

Die Anzeigeeinstellungen können nicht per DDE geändert werden und stellen daher eine wichtige Fehlerquelle beim Benutzen der RamanGUI dar. Zur Zeit ist es jedoch die einzige Möglichkeit zum Import der gemessenen Spektren nach Matlab. Zum Auffinden des Koordinatenursprungs unter der faseroptischen Sonde benötigt die RamanGUI die Rohspektren.

Ändert sich die Wellenlängen-Kalibrierung des Gerätes (sehr selten, normalerweise nur nach dem Umzug des Gerätes bzw. nach dem Austausch des Fiberports und der entsprechenden Justierung des Spalts), so muss möglicherweise die Suchfunktion `findsiecke` angepasst werden.

**Dauer einer Messung:** Die Dauer einer Messung setzt sich aus verschiedenen Einzeldauern zusammen:

- Belichtungszeit, eingestellte „exposure time“ mal der Anzahl der Koadditionen



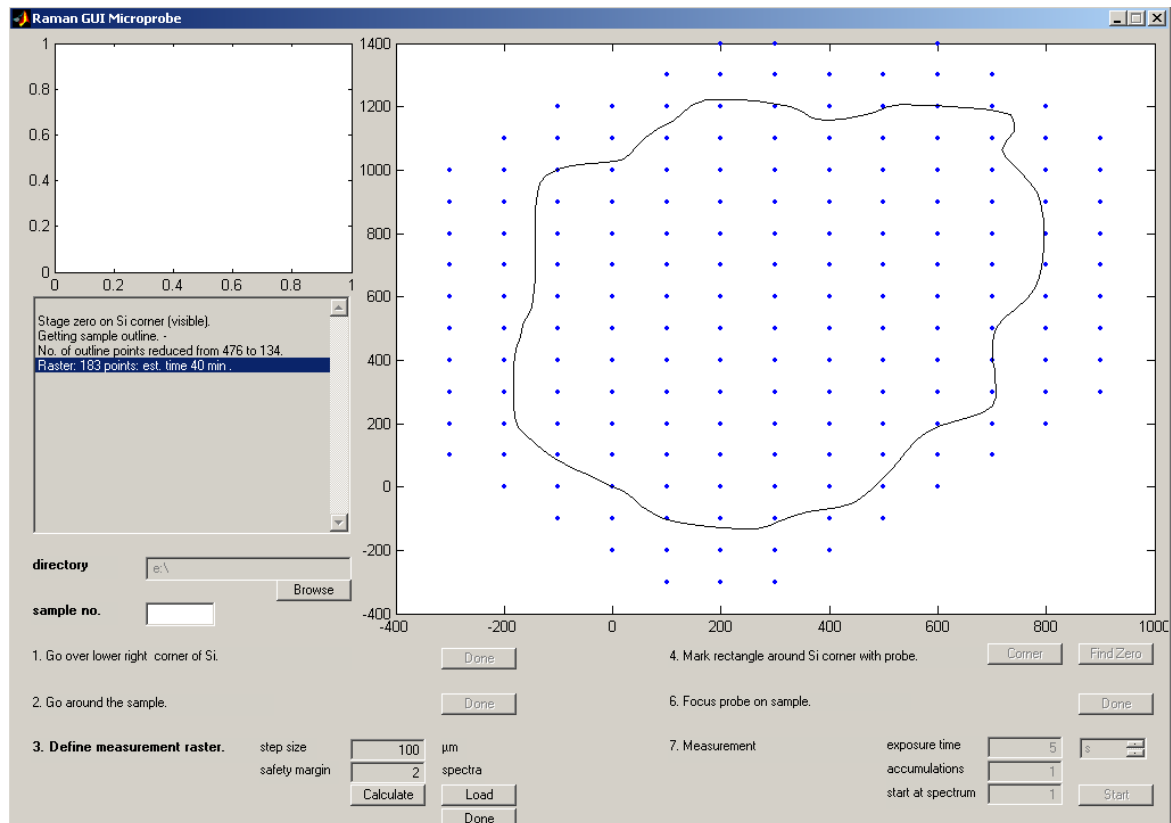
**Abbildung C.1** Bildschirmfoto der „RamanGUI“ beim Umfahren der Probe. Die Position des Mikroskopisches wird über das Kamerabild kontrolliert. Gleichzeitig zeigt der kleine Graph links oben die bereits aufgenommene Kurve an.

- „Cosmic Ray Filter“ bewirkt mindestens ein Verdoppeln der Messzeit (und die doppelte Belichtungszeit): jedes Spektrum wird zunächst zwei mal gemessen. Wird in einem der beiden Spektren ein *Spike* entdeckt, so wird das Spektrum verworfen, und ein weiteres Spektrum gemessen
- Zwischen den einzelnen Spektrenmessungen liegt eine „Reaktionszeit“ der Software. Diese betrug während der hier beschriebenen Messungen zwischen 6 und ca. 15 s. Defragmentieren der Festplatte konnte die Zeit etwas reduzieren.  
Eine Ursache für diese Wartezeiten wurde inzwischen von Martin Kammer gefunden: HoLoGRAM schreibt regelmäßig Leerzeichen in seine .ini-Datei. Diese Datei wird vor jeder Messung gelesen.

### C.3 Datenbank der Tumorproben

Um die Sammlung von Tumorproben und -spektren zu verwalten und die zu den Proben bekannten Informationen leichter zugänglich zu machen, wurde eine Datenbank erstellt. Im Sommer 2004 wurde im Institut beschlossen, die Messdaten von den Metainformationen getrennt abzulegen. Für die Messdaten wurde eine verbindliche Verzeichnis- und Namensstruktur festgelegt (im folgenden mit Dateisystem bezeichnet), die Metainformationen sind in einer relationalen Datenbank abgelegt.

Datenbank und Dateisystem wurden auf dem Rechner 141.30.8.131 untergebracht.



**Abbildung C.2** Bildschirmfoto der „RamanGUI“: Definition des Messrasters. Nachdem die Probe ganz umrandet ist, wird das Messraster festgelegt. Schrittweite und ein zusätzlicher „Sicherheitsabstand“ um die aufgenommene Hülle (hier 2 Spektren) können verändert werden. Das entsprechende Messraster wird berechnet und im großen Diagramm angezeigt. Im Meldungsfenster links erscheint die Anzahl der Messpunkte sowie eine grobe Schätzung der gesamten Messzeit.

Die Probensammlung ist in mehrere Ebenen gegliedert: von jedem Patienten können mehrere Proben vorhanden sein. Von den Proben werden Schnitte präpariert. Von jedem Schnitt können Messungen angefertigt werden. Diese 1 : n Beziehungen wurde in der Datenbank mit Tabellen abgebildet. Ausnahme sind die Ebenen Patient und Probe, die in einer Tabelle abgelegt sind, da nur von wenigen Patienten mehrere Proben vorliegen.

### C.3.1 Daten auf dem Dateiserver

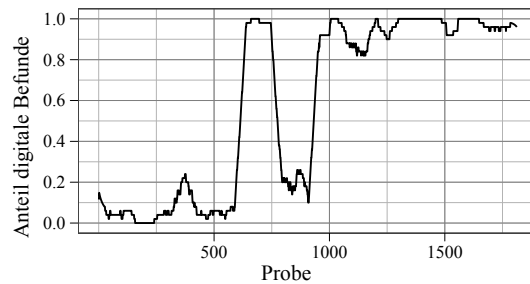
Im Dateisystem sind folgende Informationen zu finden:

**Befunde:** Die Befunde, die von der Neuropathologie an die Neurochirurgie gesendet wurden, für die Proben 600 – 700 und ab ca. 950 als anonymisierte Text-Datei gespeichert (Abbildung C.3).

**Bilddateien:** Aufnahmen der gefärbten (.mb = Methylenblau, .he = Hämatoxylin-Eosin) und und ungefärbten (.hf) Schnitte.

**Ortsaufgelöste Befunde:** Bilder von gefärbten Schnitten, auf denen die Diagnose für die verschiedenen Bereiche der Probe eingezeichnet ist (.diagnose).

**Rohspektren:** Rohdaten im Gerätespezifischen Format, ggf. in ein zip-Archiv gepackt und ggf. zusätzlich als Matlab-Datei.



**Abbildung C.3** Anteil an Proben mit digital vorliegenden Befunden. Laufender Mittelwert über jeweils 50 Proben.

**Messparameter:** Informationen zum Messraster (Koordinaten, Bild, Parameterdateien)

Das Dateisystem ist wie folgt aufgebaut:

tumordb/		
-- Pxxx/		<b>Probe und Patient</b>
-- Pxxx.diagnose.*.txt		Befunde bzgl. der OP
-- Pxxx_Syy/		<b>Schnitt</b>
-- Pxxx_Syy.*.mb.png		Bild Parallelschnitt
-- Pxxx_Syy.ot.mb.png		Bild Objektträger mit Parallelschnitten
-- Pxxx_Syy.diagnose.png		Bild mit eingetragener Diagnose
-- Pxxx_Syy.hf.jpg		Bild ungefärbter Schnitt
-- Pxxx_Syy_Mzzz/		<b>Messung</b>
-- Pxxx_Syy_Mzzz.mat		Matlab-Datei mit Messdaten
-- Pxxx_Syy_Mzzz.raster.*		Messraster
-- Pxxx_Syy_Mzzz.*		Parameterdateien zur Messung
-- Pxxx_Syy_Mzzz.spc.zip		gezippte Rohspektren
:	:	

### C.3.2 PostgreSQL-Datenbank

Die Metainformationen zu den Patienten, Proben, Schnitten und Messungen sind über eine relationale Datenbank verfügbar. Die Wahl fiel auf das Client-Server-System PostgreSQL. Mit Openoffice oobase ist eine MS Access ähnliche graphische Nutzeroberfläche sowohl unter Linux als auch unter Windows verfügbar (wahlweise über ODBC, JDBC oder den PostgreSQL-Treiber für oobase). Auch ein Kommandozeilen-Client (psql) ist für Linux und Windows verfügbar, über eine ssh-Verbindung kann auch von Windows-Rechnern psql auf dem Linux-Rechner ausgeführt werden. Sowohl für Matlab als auch für R existieren Schnittstellen-Pakete zum Zugriff auf PostgreSQL-Server.

Als Beispiel sei hier eine Abfrage in R gezeigt, die die Informationen zu allen Proben des Astro-Datensatzes in R importiert:

```
treiber <- dbDriver ("PostgreSQL")
tumordb <- dbConnect (treiber, dbname = "tumordb", host = "127.0.0.1",
                      user = "beleites", password = "Paßwort")

abfrage <- paste ("SELECT probe_id, pat_id, geburtsdatum, opdatum, geschlecht",
                  "FROM pat_proben",
                  "WHERE probe_id IN (", paste (unique (astro$Probe), collapse = ", "), ")")
)
```

## C Software

```
ergebnis <- dbSendQuery (tumordb, abfrage)
daten <- fetch (ergebnis, n = -1)
dbDisconnect(tumordb)
dbUnloadDriver(treiber)
```

Als Ergebnis erhält man einen `data.frame`:

```
> summary (daten)
  probe_id   pat_id   geburtsdatum   opdatum   geschlecht
Min.   : 46   Min.   : 46   Min.   :1928-02-12   Min.   :2005-07-14   Length:85
1st Qu.:1360 1st Qu.:1320 1st Qu.:1937-02-11 1st Qu.:2006-04-26   Class :character
Median :1426 Median :1401 Median :1944-06-09 Median :2006-07-25   Mode  :character
Mean   :1385 Mean   :1327 Mean   :1949-11-07 Mean   :2006-07-18
3rd Qu.:1490 3rd Qu.:1462 3rd Qu.:1961-04-29 3rd Qu.:2006-10-04
Max.   :1633 Max.   :1633 Max.   :1994-03-31 Max.   :2007-05-03
```

Die zentralen Tabellen der Datenbank sind `pat_proben`, `schnitte` und `messungen`. Die Tabelle `pat_proben` enthält die Informationen über die Probe und den Patienten. Analog sind Informationen zu den Schnitten in `schnitte` und zu den Messungen in `messungen` abgelegt. Dazu gibt es diverse Hilfstabellen, um Redundanzen zu vermeiden.

Die grundlegende Indizierung der Datenbank erfolgt über die Probennummer `probe_id`, die Nummer des Schnittes `schnitt_id` und die Nummer der Messung `messung_id`. Diese Felder bilden zusammen die Primärschlüssel der jeweiligen Tabellen und dienen als Fremdschlüssel für die Hilfstabellen mit den Bemerkungen, Diagnosen, usw.

Übersicht über die wichtigsten Tabellen der Tumorproben-Datenbank:

Tabelle	Beschreibung
<code>pat_proben</code>	Informationen zu den Proben und Patienten
<code>schnitte</code>	Informationen zu den Schnitten
<code>messungen</code>	Informationen zu den Messdaten
<code>bemerkungen_messung</code>	Bemerkungen zu den einzelnen Messungen
<code>bemerkungen_patprobe</code>	Bemerkungen zu den einzelnen Proben
<code>bemerkungen_schnitt</code>	Bemerkungen zu den einzelnen Schnitten
<code>diag_kuerzel</code>	Liste möglicher Diagnosen und der dazugehörige Buchstaben-Code
<code>kurzdiagnosen</code>	Diagnosen für die einzelnen Proben. Jede Probe kann mehrere Diagnosen haben. Rezidive und Verdachtsdiagnosen sind hier gekennzeichnet.
<code>probenmenge</code>	Ungefähre Menge der gefrorenen Probe

Tabelle `pat_proben`

Spalte	Beschreibung
<code>probe_id</code>	Probennummer
<code>pat_id</code>	Patientennummer (Nummer der ersten Probe dieses Patienten)
<code>diagnose</code>	Auszug aus dem Patientenbrief
<code>patientenbrief</code>	Dateiname
<code>quelle_institut</code>	hier immer „Neurochirurgie“
<code>quelle_biologisch</code>	hier immer „Hirn Mensch“
<code>beschreibung</code>	Weitere Beschreibung, z. B. Lokalisation
<code>geburtsdatum</code>	Geburtsdatum
<code>opdatum</code>	Operationsdatum (Datum der Probennahme)
<code>geschlecht</code>	Geschlecht

Tabelle `schnitt`



Spalte	Beschreibung
probe_id	Probennummer
schnitt_id	Nummer des Schnittes
buchstabe	Buchstabe auf dem CaF <sub>2</sub> -Objektträger
ot	Nummer des CaF <sub>2</sub> -Objektträgers
raman	Raman-geeigneter CaF <sub>2</sub> -Objektträger?
gefaerbt	gefärbter Schnitt
he_bild	Bild des H&E-gefärbten (Parallel)schnittes
hf_bild	Bild des ungefärbten Schnittes
dicke	Präparationsdicke
datum	Präparationsdatum
name	Wer hat präpariert?
diag_ort	Detaillierte Diagnose (Bild)
parallelschnitt_ort	In welcher Schachtel ist der Schnitt?
parallelschnitt_pos	z. B. Zimmernummer
nativ	aufgetaute dicke Probe (statt Gefrierschnitt)?

Tabelle messung

Spalte	Beschreibung
messung_id	Nummer der Messung
schnitt_id	Nummer des Schnittes
probe_id	Probennummer
geraet	Messgerät, verweist auf Tabelle
name	Wer hat gemessen?
datum	Datum der Messung
interferogramm	Interferogramm vorhanden?
raster_zeilen	bei ortsaufgelöster Messung
raster_spalten	bei ortsaufgelöster Messung
rasterdatei	bei ortsaufgelöster Messung
parameterdatei	Mess-/Geräteparameter
aufloesung	spektrale Auflösung
wz_start	gemessener Spektralbereich: Minimum
wz_ende	gemessener Spektralbereich: Maximum
n_messpunkte	Anzahl Punkte / Spektrum
n_spektren	Anzahl Spektren
zusammengesetzt	Map aus mehreren IR-Images?
rasternull_x	Koordinatenursprung Messraster bzgl. bild
rasternull_y	Koordinatenursprung Messraster bzgl. bild
bild	Bild mit Messraster

## Publikationen und Software zu dieser Arbeit

- [CB1] C. Beleites, R. Salzer und V. Sergo. „Validation of soft classification models using partial class memberships: An extended concept of sensitivity & Co. applied to grading of astrocytoma tissues“. *Chemom Intell Lab Syst* 122 (2013), S. 12–22. DOI: 10.1016/j.chemolab.2012.12.003.
- [CB2] C. Beleites, K. Geiger, M. Kirsch, S. B. Sobottka, G. Schackert und R. Salzer. „Raman spectroscopic grading of astrocytoma tissues: using soft reference information.“ *Anal Bioanal Chem* 400.9 (5/2011), S. 2801–2816. DOI: 10.1007/s00216-011-4985-4.
- [CB3] C. Beleites, U. Neugebauer, T. Bocklitz, C. Krafft und J. Popp. „Sample size planning for classification models.“ *Anal Chim Acta* 760 (1/2013), S. 25–33. DOI: 10.1016/j.aca.2012.11.007.
- [CB4] C. Beleites, A. Bonifacio, D. Codrich, C. Krafft und V. Sergo. „Raman spectroscopy and imaging: promising optical diagnostic tools in pediatrics.“ *Curr Med Chem* 20.17 (2/2013), S. 2176–2187. DOI: 10.2174/0929867311320170003.
- [CB5] C. Krafft, G. Steiner, C. Beleites und R. Salzer. „Disease recognition by infrared and Raman spectroscopy.“ *Journal of Biophotonics* 2.1-2 (2/2009), S. 13–28. DOI: 10.1002/jbio.200810024.
- [CB6] C. Beleites und R. Salzer. „Assessing and improving the stability of chemometric models in small sample size situations“. *Anal Bioanal Chem* 390.5 (3/2008), S. 1261–1271. DOI: 10.1007/s00216-007-1818-6.
- [CB7] C. Beleites, G. Steiner, M. G. Sowa, R. Baumgartner, S. Sobottka, G. Schackert und R. Salzer. „Classification of human gliomas by infrared imaging spectroscopy and chemometric image processing“. *Vib.Spec.* 38 (2005), S. 143–149.
- [CB8] C. Beleites, R. Baumgartner, C. Bowman, R. Somorjai, G. Steiner, R. Salzer und M. G. Sowa. „Variance reduction in estimating classification error using sparse datasets“. *Chemom Intell Lab Syst* 79 (2005), S. 91–100.
- [CB9] S. Dochow, M. Becker, R. Spittel, C. Beleites, S. Stanca, I. Latka, K. Schuster, J. Kobelke, S. Unger, T. Henkel, G. Mayer, J. Albert, M. Rothhardt, C. Krafft und J. Popp. „Raman-on-chip device and detection fibres with fibre Bragg grating for analysis of solutions and particles.“ *Lab Chip* 13.6 (2/2013), S. 1109–1113. DOI: 10.1039/c2lc41169e.
- [CB10] S. Dochow, C. Beleites, T. Henkel, G. Mayer, J. Albert, J. Clement, C. Krafft und J. Popp. „Quartz microfluidic chip for tumour cell identification by Raman spectroscopy in combination with optical traps.“ *Anal Bioanal Chem* 405.8 (3/2013), S. 2743–2746. DOI: 10.1007/s00216-013-6726-3.
- [CB11] A. Bonifacio, C. Beleites, F. Vittur, E. Marsich, S. Semeraro, S. Paoletti und V. Sergo. „Chemical imaging of articular cartilage sections with Raman mapping, employing uni- and multi-variate methods for data analysis.“ *The Analyst* 135.12 (12/2010), S. 3193–3204. DOI: 10.1039/c0an00459f.
- [CB12] C. Krafft, M. Kirsch, C. Beleites, G. Schackert und R. Salzer. „Methodology for fiber-optic Raman mapping and FTIR imaging of metastases in mouse brains“. *Anal Bioanal Chem* 389.4 (10/2007), S. 1133–1142.

- [CB13] A. Lattermann, C. Matthäus, N. Bergner, C. Beleites, B. F. Romeike, C. Krafft, B. R. Brehm und J. Popp. „Characterization of atherosclerotic plaque depositions by Raman and FTIR imaging.“ *J Biophotonics* 6.1 (1/2013), S. 110–121. DOI: 10.1002/jbio.201200146.
- [CB14] C. Beleites und V. Sergo. *hyperSpec: a package to handle hyperspectral data sets in R*. R package version 0.98-20120923. 2012. URL: <http://hyperspec.r-forge.r-project.org>.
- [CB15] S. Fuller und C. Beleites. *OpenBlasThreads: Miniature Package to Set Number of OpenBlasThreads*. R package version 0.1. 2012. URL: <http://github.com/simonfullernuim/OpenBlasThreads>.

## Referenzen

- [16] C. Beleites. „Chemometrische Auswertung von IR-Images und -Maps“. Diplomarbeit. Technische Universität Dresden, 2003.
- [17] I. Duran und J. J. Raizer. „Low-grade gliomas: management issues.“ *Expert Rev Anticancer Ther* 7.12 Suppl (12/2007), S15–S21. DOI: 10.1586/14737140.7.12s.S15.
- [18] R. Stupp, M. Reni, G. Gatta, E. Mazza und C. Vecht. „Anaplastic astrocytoma in adults“. *Crit Rev Oncol Hematol* 63.1 (7/2007), S. 72–80.
- [19] H. C. Diener, C. Weimar, P. Berlit, C. Elger, R. Gold, W. Hacke, A. Hufschmidt, H. Mattle, U. Meier, W. Oertel, H. Reichmann, E. Schmutzhard, C.-W. Wallesch und M. Weller. *Leitlinien für Diagnostik und Therapie in der Neurologie*. Hrsg. von H. C. Diener und C. Weimar. 5. Aufl. Thieme, Stuttgart, 2012. ISBN: 9783131324153. URL: <http://www.dgn.org/leitlinien-online.html>.
- [20] D. Ricard, A. Idbah, F. Ducray, M. Lahutte, K. Hoang-Xuan und J.-Y. Delattre. „Primary brain tumours in adults.“ *Lancet* 379.9830 (05/2012), S. 1984–1996. DOI: 10.1016/S0140-6736(11)61346-9.
- [21] B. Kowalski und S. Wold. „Pattern Recognition in Chemistry“. In: *Pattern Recognition and Reduction of Dimensionality*. Hrsg. von P. R. Krishnajah und L. N. Kanal. Bd. II. Handbook of Statistics. North-Holland, Amsterdam, 1982, S. 673–697.
- [22] A. Jain und B. Chandrasekaran. „Dimensionality and Sample Size Considerations in Pattern Recognition Practice“. In: *Handbook of Statistics*. Hrsg. von P. R. Krishnaiah und L. Kanal. Bd. II. Handbook of Statistics. North-Holland, Amsterdam, 1982. Kap. 39, S. 835–855.
- [23] C. J. Huberty. *Applied Discriminant Analysis*. John Wiley & Sons, Inc., New York, 1994.
- [24] H. M. Kalayeh und D. A. Landgrebe. „Predicting the required number of training samples.“ *IEEE Trans Pattern Anal Mach Intell* 5.6 (6/1983), S. 664–667.
- [25] S. Raudys und A. Jain. „Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners“. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13 (1991), S. 252–264. DOI: <http://doi.ieeecomputersociety.org/10.1109/34.75512>.
- [26] T. A. Dolecek, J. M. Propp, N. E. Stroup und C. Kruchko. „CBTRUS statistical report: primary brain and central nervous system tumors diagnosed in the United States in 2005–2009.“ *Neuro Oncol* 14 Suppl 5 (11/2012), S. v1–49. DOI: 10.1093/neuonc/nos218.
- [27] „Neurochirurgie“. In: Hrsg. von D. Moskopp und H. Wassmann. Schattauer, 2004. Kap. Intrakranielle Tumoren, S. 407–488.
- [28] Allgäuer, Atzinger, Bogdahn, Brügge, Collmann, Gruß, Hau, Heyder, Marienhagen, Kreuser, Röckelein, Schuierer und Ulrich. *Leitfaden des Tumorzentrums Regensburg — Projektgruppe ZNS-Tumore*. Techn. Ber. Tumorzentrum Regensburg e.V., D-93042 Regensburg: Tumorzentrum Regensburg e.V., 2002. URL: <http://www.uni-regensburg.de/Einrichtungen/Klinikum/Tumorzentrum/pdf/2002/457-495%20ZNS.pdf>.

- [29] P. Kleihues, D. N. Louis, B. W. Scheithauer, L. B. Rorke, G. Reifenberger, P. C. Burger und W. K. Cavenee. „The WHO classification of tumors of the nervous system.“ *J Neuropathol Exp Neurol* 61.3 (3/2002), S. 215–25, 215–25.
- [30] D. N. Louis, H. Ohgaki, O. D. Wiestler, W. K. Cavenee, P. C. Burger, A. Jouvett, B. W. Scheithauer und P. Kleihues. „The 2007 WHO classification of tumours of the central nervous system.“ *Acta Neuropathol. (Berl.)* 114.2 (8/2007), S. 97–109. DOI: 10 . 1007 / s00401 - 007 - 0243 - 4.
- [31] K. Poeck und W. Hacke. *Neurologie*. 12. Aufl. Springer, 2006. ISBN: 978-3-540-29997-4.
- [32] A. Deimling. *Gliomas*. @Recent Results in Cancer Research 171. Berlin: Springer, 2009. Online-Ressource. ISBN: 9783540312062. URL: <http://nbn-resolving.de/urn:nbn:de:1111-2009120869>.
- [33] L. C. U. Junqueira, J. Carneiro und M. Gratzl, Hrsg. *Histologie*. 6. Aufl. Springer, 2005. ISBN: 3-540-21965-X.
- [34] D. H. Lachance, P. Yang, D. R. Johnson, P. A. Decker, T. M. Kollmeyer, L. S. McCoy, T. Rice, Y. Xiao, F. Ali-Osman, F. Wang, S. M. Stoddard, D. J. Sprau, M. L. Kosel, J. K. Wiencke, J. L. Wiemels, J. S. Patoka, F. Davis, B. McCarthy, A. L. Rynearson, J. B. Worra, B. L. Fridley, B. P. O’Neill, J. C. Buckner, D. Il’yasova, R. B. Jenkins und M. R. Wensch. „Associations of high-grade glioma with glioma risk alleles and histories of allergy and smoking.“ *Am J Epidemiol* 174.5 (9/2011), S. 574–581. DOI: 10 . 1093/aje/kwr124.
- [35] S. Shete, F. J. Hosking, L. B. Robertson, S. E. Dobbins, M. Sanson, B. Malmer, M. Simon, Y. Marie, B. Boisselier, J.-Y. Delattre, K. Hoang-Xuan, S. El Hallani, A. Idbaih, D. Zelenika, U. Andersson, R. Henriksson, A. T. Bergenheim, M. Feychting, S. Lönn, A. Ahlbom, J. Schramm, M. Linnebank, K. Hemminki, R. Kumar, S. J. Hepworth, A. Price, G. Armstrong, Y. Liu, X. Gu, R. Yu, C. Lau, M. Schoemaker, K. Muir, A. Swerdlow, M. Lathrop, M. Bondy und R. S. Houlston. „Genome-wide association study identifies five susceptibility loci for glioma.“ *Nat Genet* 41.8 (8/2009), S. 899–904. DOI: 10 . 1038/ng . 407.
- [36] M. Wensch, R. B. Jenkins, J. S. Chang, R.-F. Yeh, Y. Xiao, P. A. Decker, K. V. Ballman, M. Berger, J. C. Buckner, S. Chang, C. Giannini, C. Halder, T. M. Kollmeyer, M. L. Kosel, D. H. LaChance, L. McCoy, B. P. O’Neill, J. Patoka, A. R. Pico, M. Prados, C. Quesenberry, T. Rice, A. L. Rynearson, I. Smirnov, T. Tihan, J. Wiemels, P. Yang und J. K. Wiencke. „Variants in the CDKN2B and RTEL1 regions are associated with high-grade glioma susceptibility.“ *Nat Genet* 41.8 (8/2009), S. 905–908. DOI: 10 . 1038/ng . 408.
- [37] R. B. Jenkins, M. R. Wensch, D. Johnson, B. L. Fridley, P. A. Decker, Y. Xiao, T. M. Kollmeyer, A. L. Rynearson, S. Fink, T. Rice, L. S. McCoy, C. Halder, M. L. Kosel, C. Giannini, T. Tihan, B. P. O’Neill, D. H. Lachance, P. Yang, J. Wiemels und J. K. Wiencke. „Distinct germ line polymorphisms underlie glioma morphologic heterogeneity.“ *Cancer Genet* 204.1 (1/2011), S. 13–18. DOI: 10 . 1016/j . cancergencyto . 2010 . 10 . 002.
- [38] M. Z. Braganza, C. M. Kitahara, A. Berrington de González, P. D. Inskip, K. J. Johnson und P. Rajaraman. „Ionizing radiation and the risk of brain and central nervous system tumors: a systematic review.“ *Neuro Oncol* 14.11 (11/2012), S. 1316–1324. DOI: 10 . 1093/ neuonc/nos208.
- [39] INTERPHONE Study Group. „Brain tumour risk in relation to mobile telephone use: results of the INTERPHONE international case-control study.“ *Int J Epidemiol* 39.3 (6/2010), S. 675–694. DOI: 10 . 1093/ije/dyq079.

- [40] L. Hardell, M. Carlberg und K. Hansson Mild. „Re-analysis of risk for glioma in relation to mobile telephone use: comparison with the results of the Interphone international case-control study.“ *Int J Epidemiol* 40.4 (8/2011), S. 1126–1128. DOI: 10.1093/ije/dyq246.
- [41] L. Hardell, M. Carlberg und K. Hansson Mild. „Pooled analysis of case-control studies on malignant brain tumours and the use of mobile and cordless phones including living and deceased subjects.“ *Int J Oncol* 38.5 (5/2011), S. 1465–1474. DOI: 10.3892/ijo.2011.947.
- [42] M. Carlberg und L. Hardell. „On the association between glioma, wireless phones, heredity and ionising radiation.“ *Pathophysiology* 19.4 (9/2012), S. 243–252. DOI: 10.1016/j.pathophys.2012.07.001.
- [43] M. P. Little, P. Rajaraman, R. E. Curtis, S. S. Devesa, P. D. Inskip, D. P. Check und M. S. Linet. „Mobile phone use and glioma risk: comparison of epidemiological study results with incidence trends in the United States.“ *BMJ* 344 (2012), e1147.
- [44] I. A. on Cancer Research. *IARC Classifies Radiofrequency Electromagnetic Fields as Possibly Carcinogenic to Humans*. Press Release. 208. 5/2011. URL: [http://www.iarc.fr/en/media-centre/pr/2011/pdfs/pr208\\_E.pdf](http://www.iarc.fr/en/media-centre/pr/2011/pdfs/pr208_E.pdf).
- [45] A. J. Moore. *Neurosurgery. Principles and Practice*. Hrsg. von D. W. Newell. @Springer Specialist Surgery Series. London: Springer-Verlag London Ltd, 2005. Online-Ressource. ISBN: 9781846280511. URL: <http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10228864>.
- [46] S. R. VandenBerg. „Current diagnostic concepts of astrocytic tumors.“ *J Neuropathol Exp Neurol* 51.6 (11/1992), S. 644–657.
- [47] J. M. Kros. „From expert opinion to evidence-based: changes in the gold standard of primary brain tumour diagnosis.“ *The Journal of pathology* 213.1 (09/2007), S. 1–3. DOI: 10.1002/path.2201.
- [48] J. A. Schwartzbaum, J. L. Fisher, K. D. Aldape und M. Wrensch. „Epidemiology and molecular pathology of glioma.“ *Nat Clin Pract Neurol* 2.9 (9/2006), 494–503, quiz 1 p following 516. DOI: 10.1038/ncpneuro0289.
- [49] N. F. Marko, J. Quackenbush und R. J. Weil. „Why is there a lack of consensus on molecular subgroups of glioblastoma? Understanding the nature of biological and statistical variability in glioblastoma expression data.“ *PLoS ONE* 6.7 (2011), e20826. DOI: 10.1371/journal.pone.0020826.
- [50] M. K. Nicholas. „Glioblastoma multiforme: evidence-based approach to therapy.“ *Expert Rev Anticancer Ther* 7.12 Suppl (12/2007), S23–S27. DOI: 10.1586/14737140.7.12s.S23.
- [51] K. A. Carson, S. A. Grossman, J. D. Fisher und E. G. Shaw. „Prognostic factors for survival in adult patients with recurrent glioma enrolled onto the new approaches to brain tumor therapy CNS consortium phase I and II clinical trials.“ *Journal of clinical oncology: official journal of the American Society of Clinical Oncology* 25.18 (6/2007), S. 2601–2606. DOI: 10.1200/JCO.2006.08.1661.
- [52] C. S. Hwang, M. H. Marymont und K. Muro. „Photon radiotherapy for the treatment of high-grade gliomas.“ *Expert Rev Anticancer Ther* 7.12 Suppl (12/2007), S37–S43. DOI: 10.1586/14737140.7.12s.S37.

- [53] R. V. Lukas, A. Boire und M. K. Nicholas. „Emerging therapies for malignant glioma.“ *Expert Rev Anticancer Ther* 7.12 Suppl (12/2007), S29–S36. DOI: 10.1586/14737140.7.12s.S29.
- [54] T. R. Coté, A. Manns, C. R. Hardy, F. J. Yellin und P. Hartge. „Epidemiology of brain lymphoma among people with or without acquired immunodeficiency syndrome. AIDS/Cancer Study Group.“ *J Natl Cancer Inst* 88.10 (5/1996), S. 675–679.
- [55] P. Berlit, Hrsg. *Klinische Neurologie*. 2. Aufl. Springer, 2006. ISBN: 3-540-01982-0.
- [56] F. DeMonte, M. Gilbert, A. Mahajan und I. McCutcheon, Hrsg. *Tumors of the Brain and Spine*. M. D. Anderson Cancer Care Series. Springer, 2006. ISBN: 978-0387-29201-4.
- [57] T. Batchelor und J. S. Loeffler. „Primary CNS lymphoma.“ *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 24.8 (3/2006), S. 1281–1288. DOI: 10.1200/JCO.2005.04.8819.
- [58] E. Gerstner und T. Batchelor. „Primary CNS lymphoma.“ *Expert Rev Anticancer Ther* 7.5 (5/2007), S. 689–700. DOI: 10.1586/14737140.7.5.689.
- [59] C. Diamond, T. H. Taylor, T. Aboumrad und H. Anton-Culver. „Changes in acquired immunodeficiency syndrome-related non-Hodgkin lymphoma in the era of highly active antiretroviral therapy: incidence, presentation, treatment, and survival.“ *Cancer* 106.1 (01/2006), S. 128–135. DOI: 10.1002/cncr.21562.
- [60] M. Hensel, A. Goetzenich, T. Lutz, A. Stoehr, A. Moll, J. Rockstroh, N. Hanhoff, H. Jäger und F. Mosthaf. „HIV und Krebs in Deutschland“. *Deutsches Ärzteblatt* 108.8 (2011), S. 117–122. DOI: 10.3238/arztebl.2010.0117.
- [61] A. Juergens, H. Pels, S. Rogowski, K. Fliessbach, A. Glasmacher, A. Engert, M. Reiser, V. Diehl, M. Vogt-Schaden, G. Egerer, G. Schackert, H. Reichmann, F. Kroschinsky, U. Bode, U. Herrlinger, M. Linnebank, M. Deckert, R. Fimmers, I. G. H. Schmidt-Wolf und U. Schlegel. „Long-term survival with favorable cognitive outcome after chemotherapy in primary central nervous system lymphoma.“ *Ann Neurol* 67.2 (02/2010), S. 182–189. DOI: 10.1002/ana.21824.
- [62] C. Nimsky, O. Ganslandt, S. Cerny, P. Hastreiter, G. Greiner und R. Fahlbusch. „Quantification of, visualization of, and compensation for brain shift using intraoperative magnetic resonance imaging.“ *Neurosurgery* 47.5 (11/2000), S. 1070–9, 1070–9.
- [63] D. A. Orringer, A. Golby und F. Jolesz. „Neuronavigation in the surgical management of brain tumors: current and future trends.“ *Expert Rev Med Devices* 9.5 (09/2012), S. 491–500. DOI: 10.1586/erd.12.42.
- [64] M. Kirsch. persönliche Mitteilung.
- [65] S. N. Reske. „Positronen-Emissions-Tomographie in der Onkologie“. *Deutsches Ärzteblatt* 95.30 (1998), pages.
- [66] M. R. Fetell. „Merrit’s Textbook of Neurology“. In: Hrsg. von L. P. Rowland. 9. Aufl. Williams & Wilkins, 1995. Kap. 51. Gliomas, S. 336–351.
- [67] A. Fallert-Müller, Hrsg. *Lexikon der Biochemie. in zwei Bänden*. Heidelberg: Elsevier, Spektrum Akad. Verl., 2004. ISBN: 3827415802.
- [68] L. Stryer. *Biochemie*. Heidelberg: Spektrum Akademischer Verl., 1991.
- [69] S. J. Madsen, C.-H. Sun, B. J. Tromberg, V. Cristini, N. D. Magalhães und H. Hirschberg. „Multicell tumor spheroids in photodynamic therapy.“ *Lasers Surg Med* 38.5 (6/2006), S. 555–564. DOI: 10.1002/lsm.20350.

- [70] W. Stummer, U. Pichlmeier, T. Meinel, O. D. Wiestler, F. Zanella, H.-J. Reulen und A. L. A.-G. S. Group. „Fluorescence-guided surgery with 5-aminolevulinic acid for resection of malignant glioma: a randomised controlled multicentre phase III trial.“ *The lancet oncology* 7.5 (5/2006), S. 392–401. DOI: 10.1016/S1470-2045(06)7066569.
- [71] S. B. Sobottka, K. D. Geiger, R. Salzer, G. Schackert und C. Krafft. „Suitability of infrared spectroscopic imaging as an intraoperative tool in cerebral glioma surgery.“ *Anal Bioanal Chem* 393.1 (1/2009), S. 187–195. DOI: 10.1007/s00216-008-2443-8.
- [72] M. Makary, E. A. Chiocca, N. Erminy, M. Antor, S. D. Bergese, M. Abdel-Rasoul, S. Fernandez und R. Dzwonczyk. „Clinical and economic outcomes of low-field intraoperative MRI-guided tumor resection neurosurgery.“ *J Magn Reson Imaging* 34.5 (11/2011), S. 1022–1030. DOI: 10.1002/jmri.22739.
- [73] P. L. Kubben, K. J. ter Meulen, O. E. M. G. Schijns, M. P. ter Laak-Poort, J. J. van Overbeeke und H. van Santbrink. „Intraoperative MRI-guided resection of glioblastoma multiforme: a systematic review.“ *Lancet Oncol* 12.11 (10/2011), S. 1062–1070. DOI: 10.1016/S1470-2045(11)70130-9.
- [74] P. L. Kubben und H. van Santbrink. „Intraoperative magnetic resonance imaging for high grade glioma resection: Evidence-based or wishful thinking?“ *Surg Neurol Int* 4 (2013), S. 1. DOI: 10.4103/2152-7806.106114.
- [75] D. Liang und M. Schulder. „The role of intraoperative magnetic resonance imaging in glioma surgery.“ *Surg Neurol Int* 3.Suppl 4 (2012), S320–S327. DOI: 10.4103/2152-7806.103029.
- [76] C. Senft, A. Bink, K. Franz, H. Vatter, T. Gasser und V. Seifert. „Intraoperative MRI guidance and extent of resection in glioma surgery: a randomised, controlled trial.“ *Lancet Oncol* 12.11 (10/2011), S. 997–1003. DOI: 10.1016/S1470-2045(11)70196-6.
- [77] C. Powell. *Akron General installs magnet for planned intraoperative MRI*. 12/2011. URL: <http://www.ohio.com/business/akron-general-installs-magnet-for-planned-intraoperative-mri-1.356912>.
- [78] *Royal Children's Hospital first to offer state-of-the-art intraoperative MRI*. RCH News. 10/2010. URL: <http://blogs.rch.org.au/news/2010/10/29/royal-childrens-hospital-first-to-offer-state-of-the-art-intraoperative-mri/>.
- [79] C. Werner. *Einzigartige Gehirn-OP in der Heidelberg-Klinik*. 08/2011. URL: <http://www.abendblatt.de/ratgeber/wissen/article1978160/Einzigartige-Gehirn-OP-in-der-Heidelberg-Klinik.html>.
- [80] D. P. Archer, R. A. McTaggart Cowan, R. J. Falkenstein und G. R. Sutherland. „Intraoperative mobile magnetic resonance imaging for craniotomy lengthens the procedure but does not increase morbidity.“ *Can J Anaesth* 49.4 (04/2002), S. 420–426.
- [81] M. F. Kircher, A. de la Zerda, J. V. Jokerst, C. L. Zavaleta, P. J. Kempen, E. Mittra, K. Pitter, R. Huang, C. Campos, F. Habte, R. Sinclair, C. W. Brennan, I. K. Mellinshoff, E. C. Holland und S. S. Gambhir. „A brain tumor molecular imaging strategy using a new triple-modality MRI-photoacoustic-Raman nanoparticle.“ *Nat Med* 18.5 (5/2012), S. 829–834. DOI: 10.1038/nm.2721.
- [82] R. L. McCreery. *Raman Spectroscopy for Chemical Analysis*. Wiley-Interscience, 2000. ISBN: 0471252875. DOI: 10.1002/0471721646. URL: <http://onlinelibrary.wiley.com/book/10.1002/0471721646>.



- [83] T. Renner. *Quantities, Units and Symbols in Physical Chemistry*. Hrsg. von E. R. Cohen, T. Cvitas, J. G. Frey, B. Holstrom, K. Kuchitsu, R. Marquardt, I. Mills, F. Pavese, M. Quack, J. Stohner, H. L. Strauss, M. Takami und A. J. Thor. The Royal Society of Chemistry, 2007. ISBN: 978-0-85404-433-7. DOI: 10.1039/9781847557889. URL: <http://dx.doi.org/10.1039/9781847557889>.
- [84] P. R. Griffiths und J. A. de Haseth. *Fourier transform infrared spectrometry*. 2. ed. Chemical analysis 171. Hoboken, NJ: Wiley-Interscience, 2007. XVII, 529. ISBN: 9780471194040.
- [85] J. M. Chalmers und P. R. Griffiths, Hrsg. *Handbook of Vibrational Spectroscopy*. Wiley, 2006. ISBN: 0470027320.
- [86] G. Sutherland. persönliche Mitteilung.
- [87] A. P. Shreve, N. J. Cherepy und R. A. Mathies. „Effective Rejection of Fluorescence Interference in Raman Spectroscopy Using a Shifted Excitation Difference Technique“. *Appl. Spectrosc.* 46.4 (04/1992), S. 707–711.
- [88] S. Dochow, N. Bergner, C. Krafft, J. Clement, M. Malizu, B. Balagopal, R. Marchington, K. Dholakia und J. Popp. „Wavelength Modulated Raman Spectroscopy for Biomedical Applications.“ *Biomed Tech (Berl)* (08/2012). DOI: 10.1515/bmt-2012-4280.
- [89] F. Bonnier, A. Mehmood, P. Knief, A. Meade, W. Hornebeck, H. Lambkin, K. Flynn, V. McDonagh, C. Healy, T. C. Lee, F. M. Lyng und H. J. Byrne. „In vitro analysis of immersed human tissues by Raman microspectroscopy“. *J Raman Spectrosc* 41 (2010). DOI: 10.1002/jrs.2825.
- [90] W. Gellert, Hrsg. *Kleine Enzyklopädie Mathematik*. 9., gekürzte Aufl. Leipzig: Bibliogr. Inst., 1974. 739 S.
- [91] X. Shao, W. Zheng und Z. Huang. „In vivo diagnosis of colonic precancer and cancer using near-infrared autofluorescence spectroscopy and biochemical modeling.“ *J Biomed Opt* 16.6 (06/2011), S. 067005. DOI: 10.1117/1.3589099.
- [92] V. Hollis. „Non-invasive monitoring of brain tissue temperature by near-infrared spectroscopy“. Doktorarbeit. University College London, 2002. URL: [http://www.medphys.ucl.ac.uk/research/borg/homepages/veronica/veronica\\_2003.html](http://www.medphys.ucl.ac.uk/research/borg/homepages/veronica/veronica_2003.html).
- [93] I. Latka, S. Dochow, C. Krafft, B. Dietzek und J. Popp. „Fiber optic probes for linear and nonlinear Raman applications – Current trends and future development“. *Laser & Photonics Reviews* (2013), n/a–n/a. DOI: 10.1002/lpor.201200049.
- [94] U. Utzinger und R. R. Richards-Kortum. „Fiber optic probes for biomedical optical spectroscopy“. *J Biomed Opt* 8.1 (1/2003), S. 121–147. DOI: 10.1117/1.1528207.
- [95] InPhotonics, Inc. *Background Filtering in Fiber Optic Raman Sampling Probes*. 111 Downey Street, Norwood, MA 02062, 1999. URL: <http://www.inphotonics.com/technote13.pdf>.
- [96] S. Koljenović, T. C. B. Schut, R. Wolthuis, B. de Jong, L. Santos, P. J. Caspers, J. M. Kros und G. J. Puppels. „Tissue characterization using high wave number Raman spectroscopy“. *J Biomed Opt* 10.3 (5/2005), pages. DOI: 10.1117/1.1922307.
- [97] C. Krafft, L. Neudert, T. Simat und R. Salzer. „Near infrared Raman spectra of human brain lipids.“ *Spectrochim Acta A Mol Biomol Spectrosc* 61.7 (5/2005), S. 1529–1535. DOI: 10.1016/j.saa.2004.11.017.
- [98] C. Krafft und V. Sergo. „Biomedical applications of Raman and infrared spectroscopy to diagnose tissues“. *Spectroscopy* 20 (2006), S. 195–218.

- [99] C. Kendall, M. Isabelle, F. Bazant-Hegemark, J. Hutchings, L. Orr, J. Babrah, R. Baker und N. Stone. „Vibrational spectroscopy: a clinical tool for cancer diagnostics.“ *The Analyst* 134.6 (2009), S. 1029–1045.
- [100] C. Krafft, K. Thümmeler, S. B. Sobottka, G. Schackert und R. Salzer. „Classification of malignant gliomas by infrared spectroscopy and linear discriminant analysis“. *Biopolymers* 82.4 (2006), S. 301–305.
- [101] C. Krafft, S. B. Sobottka, G. Schackert und R. Salzer. „Analysis of human brain tissue, brain tumors and tumor cells by infrared spectroscopic mapping.“ *The Analyst* 129.10 (10/2004), S. 921–925. DOI: 10.1039/b408934k.
- [102] K. Gajjar, L. D. Heppenstall, W. Pang, K. M. Ashton, J. Trevisan, I. I. Patel, V. Llabjani, H. F. Stringfellow, P. L. Martin-Hirsch, T. Dawson und F. L. Martin. „Diagnostic segregation of human brain tumours using Fourier-transform infrared and/or Raman spectroscopy coupled with discriminant analysis“. *Analytical Methods* 5 (2013), S. 89–102. DOI: 10.1039/C2AY25544H.
- [103] A. Stelling, R. Salzer, M. Kirsch, S. B. Sobottka, K. Geiger, E. Koch, G. Schackert und G. Steiner. „Intra-operative optical diagnostics with vibrational spectroscopy.“ *Anal Bioanal Chem* 400.9 (7/2011), S. 2745–2753. DOI: 10.1007/s00216-011-5022-3.
- [104] A. L. Stelling, D. Toher, O. Uckermann, J. Tavkin, E. Leipnitz, J. Schweizer, H. Cramm, G. Steiner, K. D. Geiger und M. Kirsch. „Infrared spectroscopic studies of cells and tissues: triple helix proteins as a potential biomarker for tumors.“ *PLoS One* 8.3 (2013), e58332. DOI: 10.1371/journal.pone.0058332.
- [105] M. Köhler, S. Machill, R. Salzer und C. Krafft. „Characterization of lipid extracts from brain tissue and tumors using Raman spectroscopy and mass spectrometry.“ *Anal Bioanal Chem* 393.5 (3/2009), S. 1513–1520. DOI: 10.1007/s00216-008-2592-9.
- [106] A. Beljebbar, N. Amharref, A. Lévêques, S. Dukic, L. Venteo, L. Schneider, M. Pluot und M. Manfait. „Modeling and quantifying biochemical changes in C6 tumor gliomas by Fourier transform infrared imaging.“ *Anal Chem* 80.22 (11/2008), S. 8406–8415. DOI: 10.1021/ac800990y.
- [107] I. Dreissig, S. Machill, R. Salzer und C. Krafft. „Quantification of brain lipids by FTIR spectroscopy and partial least squares regression.“ *Spectrochim Acta A Mol Biomol Spectrosc* 71.5 (1/2009), S. 2069–2075. DOI: 10.1016/j.saa.2008.08.008.
- [108] S. Koljenović, L.-P. Choo-Smith, T. C. B. Schut, J. M. Kros, H. J. van den Berge und G. J. Puppels. „Discriminating vital tumor from necrotic tissue in human glioblastoma tissue samples by Raman spectroscopy.“ *Laboratory investigation; a journal of technical methods and pathology* 82.10 (10/2002), S. 1265–1277. DOI: 10.1097/01.LAB.0000032545.96931.B8.
- [109] N. Bergner. „Schwingungsspektroskopische Bildgebung zur charakterisierung, Klassifizierung und Identifizierung von humanem Hirngewebe auf zellulärer Ebene“. Doktorarbeit. Friedrich-Schiller-Universität Jena, 2013.
- [110] N. Bergner, B. F. M. Romeike, R. Reichart, R. Kalff, C. Krafft und J. Popp. „Tumor margin identification and prediction of the primary tumor from brain metastases using FTIR imaging and support vector machines.“ *Analyst* 138.14 (07/2013), S. 3983–3990. DOI: 10.1039/c3an00326d.

- [111] N. Bergner, C. Krafft, K. D. Geiger, M. Kirsch, G. Schackert und J. Popp. „Unsupervised unmixing of Raman microspectroscopic images for morphochemical analysis of non-dried brain tumor specimens.“ *Anal Bioanal Chem* 403.3 (5/2012), S. 719–725. DOI: 10.1007/s00216-012-5858-1.
- [112] N. Bergner, T. Bocklitz, B. F. Romeike, R. Reichart, R. Kalff, C. Krafft und J. Popp. „Identification of primary tumors of brain metastases by Raman imaging and support vector machines“. *Chemom Intell Lab Syst* (2012).
- [113] D. G. Leslie, R. E. Kast, J. M. Poulik, R. Rabah, S. Sood, G. W. Auner und M. D. Klein. „Identification of Pediatric Brain Neoplasms Using Raman Spectroscopy.“ *Pediatr Neurosurg* (11/2012), S. 109–117. DOI: 10.1159/000343285.
- [114] R. Marx. „Raman-Mapping von Hirntumor-Dünnschnitten“. Diplomarbeit. Technische Universität Dresden, 2009.
- [115] C. L. Evans, X. Xu, S. Kesari, X. S. Xie, S. T. C. Wong und G. S. Young. „Chemically-selective imaging of brain structures with CARS microscopy.“ *Opt Express* 15.19 (9/2007), S. 12076–12087. DOI: 10.1364/OE.15.012076.
- [116] T. Meyer, N. Bergner, A. Medyukhina, B. Dietzek, C. Krafft, B. F. M. Romeike, R. Reichart, R. Kalff und J. Popp. „Interpreting CARS images of tissue within the C-H-stretching region.“ *J Biophotonics* 5.10 (10/2012), S. 729–733. DOI: 10.1002/jbio.201200104.
- [117] T. Meyer, N. Bergner, C. Bielecki, C. Krafft, D. Akimov, B. F. M. Romeike, R. Reichart, R. Kalff, B. Dietzek und J. Popp. „Nonlinear microscopy, infrared, and Raman microspectroscopy for brain tumor analysis.“ *J Biomed Opt* 16.2 (02/2011), S. 021113. DOI: 10.1117/1.3533268.
- [118] R. Noreen, R. Pineau, C.-C. Chien, M. Cestelli-Guidi, Y. Hwu, A. Marcelli, M. Moenner und C. Petibois. „Functional histology of glioma vasculature by FTIR imaging.“ *Anal Bioanal Chem* 401.3 (8/2011), S. 795–801. DOI: 10.1007/s00216-011-5069-1.
- [119] R. Noreen, M. Moenner, Y. Hwu und C. Petibois. „FTIR spectro-imaging of collagens for characterization and grading of gliomas.“ *Biotechnol Adv* 30.6 (11/2012), S. 1432–1446. DOI: 10.1016/j.biotechadv.2012.03.009.
- [120] H. N. Banerjee und L. Zhang. „Deciphering the finger prints of brain cancer astrocytoma in comparison to astrocytes by using near infrared Raman spectroscopy.“ *Mol Cell Biochem* 295.1-2 (1/2007), S. 237–240. DOI: 10.1007/s11010-006-9278-4.
- [121] L. J. Kaufman, C. P. Brangwynne, K. E. Kasza, E. Filippidi, V. D. Gordon, T. S. Deisboeck und D. A. Weitz. „Glioma expansion in collagen I matrices: analyzing collagen concentration-dependent growth and motility patterns.“ *Biophys J* 89.1 (7/2005), S. 635–650. DOI: 10.1529/biophysj.105.061994.
- [122] N. Amharref, A. Beljebbar, S. Dukic, L. Venteo, L. Schneider, M. Pluot, R. Vistelle und M. Manfait. „Brain tissue characterisation by infrared imaging in a rat glioma model.“ *Biochim Biophys Acta* 1758.7 (7/2006), S. 892–899. DOI: 10.1016/j.bbamem.2006.05.003.
- [123] A. Beljebbar, S. Dukic, N. Amharref, S. Bellefqih und M. Manfait. „Monitoring of biochemical changes through the c6 gliomas progression and invasion by fourier transform infrared (FTIR) imaging.“ *Anal Chem* 81.22 (11/2009), S. 9247–9256. DOI: 10.1021/ac901464v.

- [124] A. Beljebbar, S. Dukic, N. Amharref und M. Manfait. „Screening of biochemical/histological changes associated to C6 glioma tumor development by FTIR/PCA imaging.“ *The Analyst* 135.5 (5/2010), S. 1090–1097. DOI: 10.1039/b922184k.
- [125] N. Amharref, A. Beljebbar, S. Dukic, L. Venteo, L. Schneider, M. Pluot und M. Manfait. „Discriminating healthy from tumor and necrosis tissue in rat brain tissue samples by Raman spectral imaging.“ *Biochim Biophys Acta* 1768.10 (10/2007), S. 2605–2615. DOI: 10.1016/j.bbame.2007.06.032.
- [126] A. Beljebbar, S. Dukic, N. Amharref und M. Manfait. „Ex vivo and in vivo diagnosis of C6 glioblastoma development by Raman spectroscopy coupled to a microprobe.“ *Anal Bioanal Chem* 398.1 (9/2010), S. 477–487. DOI: 10.1007/s00216-010-3910-6.
- [127] K. Wehbe, R. Pineau, S. Eimer, A. Vital, H. Loiseau und G. Déléris. „Differentiation between normal and tumor vasculature of animal and human glioma by FTIR imaging.“ *Analyst* 135.12 (12/2010), S. 3052–3059. DOI: 10.1039/c0an00513d.
- [128] M. J. Clark, N. Homer, B. D. O’Connor, Z. Chen, A. Eskin, H. Lee, B. Merriman und S. F. Nelson. „U87MG decoded: the genomic sequence of a cytogenetically aberrant human cancer cell line.“ *PLoS Genet* 6.1 (01/2010), e1000832. DOI: 10.1371/journal.pgen.1000832.
- [129] R. A. Wolfe, G. H. Sato und D. B. McClure. „Continuous culture of rat C6 glioma in serum-free medium.“ *J Cell Biol* 87.2 Pt 1 (11/1980), S. 434–441.
- [130] B. Grobbs, P. P. De Deyn und H. Slegers. „Rat C6 glioma as experimental model system for the study of glioblastoma growth and invasion.“ *Cell Tissue Res* 310.3 (12/2002), S. 257–270. DOI: 10.1007/s00441-002-0651-7.
- [131] M. G. Shim und B. C. Wilson. „The effects of ex vivo handling procedures on the near-infrared Raman spectra of normal mammalian tissues.“ *Photochem Photobiol* 63.5 (5/1996), S. 662–671. DOI: 10.1111/j.1751-1097.1996.tb05671.x.
- [132] C. Krafft, S. B. Sobottka, G. Schackert und R. Salzer. „Near infrared Raman spectroscopic mapping of native brain tissue and intracranial tumors.“ *The Analyst* 130.7 (7/2005), S. 1070–1077. DOI: 10.1039/b419232j.
- [133] C. Matthäus, S. Dochow, G. Bergner, A. Lattermann, B. F. M. Romeike, E. T. Marple, C. Krafft, B. Dietzek, B. R. Brehm und J. Popp. „In vivo characterization of atherosclerotic plaque depositions by Raman-probe spectroscopy and in vitro coherent anti-stokes Raman scattering microscopic imaging on a rabbit model.“ *Anal Chem* 84.18 (09/2012), S. 7845–7851. DOI: 10.1021/ac301522d.
- [134] M. Kirsch, G. Schackert, R. Salzer und C. Krafft. „Raman spectroscopic imaging for in vivo detection of cerebral brain metastases.“ *Anal Bioanal Chem* 398.4 (10/2010), S. 1707–1713. DOI: 10.1007/s00216-010-4116-7.
- [135] M. Kammer. „Datenanalyse von FTIR-Maps“. Diplomarbeit. Technische Universität Dresden, 2007.
- [136] P. G. Andrus und R. D. Strickland. „Cancer grading by Fourier transform infrared spectroscopy.“ *Biospectroscopy* 4.1 (1998), S. 37–46. DOI: 10.1007/s00216-010-4116-7.
- [137] J. Ramesh, J. Kapelushnik, J. Mordehai, A. Moser, M. Huleihel, V. Erukhimovitch, C. Levi und S. Mordechai. „Novel methodology for the follow-up of acute lymphoblastic leukemia using FTIR microspectroscopy.“ *J Biochem Biophys Methods* 51.3 (5/2002), S. 251–261. DOI: 10.1016/S0165-022X(02)00004-0.

- [138] J. Ramesh, M. Huleihel, J. Mordehai, A. Moser, V. Erukhimovich, C. Levi, J. Kapelushnik und S. Mordechai. „Preliminary results of evaluation of progress in chemotherapy for childhood leukemia patients employing Fourier-transform infrared microspectroscopy and cluster analysis.“ *The Journal of laboratory and clinical medicine* 141.6 (6/2003), S. 385–394. DOI: 10.1016/S0022-2143(03)00025-8.
- [139] S. Wold, M. Sjöström und L. Eriksson. „PLS-regression: a basic tool of chemometrics“. *Chemom Intell Lab Syst* 58.2 (10/2001), S. 109–130.
- [140] S. Wold, J. Trygg, A. Berglund und H. Antti. „Some recent developments in PLS modeling“. *Chemom Intell Lab Syst* 58.2 (10/2001), S. 131–150.
- [141] A. Höskuldsson. „PLS regression methods“. *J Chemom* 2.3 (1988), S. 211–228.
- [142] P. Geladi und B. R. Kowalski. „Partial least-squares regression: a tutorial“. *Anal Chim Acta* 185 (1986), S. 1–17. DOI: 10.1016/0003-2670(86)80028-9.
- [143] M. Barker und W. Rayens. „Partial least squares for discrimination“. *J Chemom* 17.3 (2003), S. 166–173. DOI: 10.1002/cem.785.
- [144] G. C. Cawley und N. L. C. Talbot. „On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation“. *Journal of Machine Learning Research* 11 (2010), S. 2079–2107.
- [145] T. Hastie, R. Tibshirani und J. Friedman. *The Elements of Statistical Learning; Data mining, Inference and Prediction*. 2. Aufl. Springer Verlag, New York, 2009.
- [146] R. Wolthuis, M. van Aken, K. Fountas, J. S. Robinson, H. A. Bruining und G. J. Puppels. „Determination of water concentration in brain tissue by Raman spectroscopy.“ *Anal Chem* 73.16 (8/2001), S. 3915–3920.
- [147] M. Sattlecker, N. Stone, J. Smith und C. Bessant. „Assessment of robustness and transferability of classification models built for cancer diagnostics using Raman spectroscopy“. *J Raman Spectrosc* (2010), S. 897–903.
- [148] P. C. Mahalanobis. „On the generalized distance in statistics“. In: *Proceedings of the National Institute of Sciences of India*. Bd. 2. 1. New Delhi. 1936, S. 49–55.
- [149] R. A. Fisher. „The Use of Multiple Measurements in Taxonomic Problems“. *Annals of Eugenics* 7.7 (1936), S. 179–188.
- [150] W. N. Venables und B. D. Ripley. *Modern Applied Statistics with S*. 4th. New York: Springer, 9/2002. ISBN: 9780387954578. URL: <http://www.stats.ox.ac.uk/pub/MASS4>.
- [151] C. Krafft, L. Shapoval, S. B. Sobottka, G. Schackert und R. Salzer. „Identification of primary tumors of brain metastases by infrared spectroscopic imaging and linear discriminant analysis.“ *Technology in cancer research & treatment* 5.3 (6/2006), S. 291–298.
- [152] C. Krafft, S. Sobottka, K. Geiger, G. Schackert und R. Salzer. „Classification of malignant gliomas by infrared spectroscopic imaging and linear discriminant analysis“. *Anal Bioanal Chem* 387.5 (3/2007), S. 1669–1677.
- [153] G. Steiner, S. Küchler, A. Hermann, E. Koch, R. Salzer, G. Schackert und M. Kirsch. „Rapid and label-free classification of human glioma cells by infrared spectroscopic imaging.“ *Cytometry. Part A: the journal of the International Society for Analytical Cytology* 73A.12 (12/2008), S. 1158–1164. DOI: 10.1002/cyto.a.20639.
- [154] B. Appiah, V. Nammalvar und R. Drezek. „Statistical analysis of FTIR spectra of cervical tissues and diagnostic algorithms for cervical cancer“. In: *Proc. SPIE 6863, Optical Diagnostics and Sensing VIII*. 2008, S. 68630V. DOI: 10.1117/12.764271.

- [155] S. Pieters, Y. Vander Heyden, J.-M. Roger, M. D'Hondt, L. Hansen, B. Palagos, B. De Spiegeleer, J.-P. Remon, C. Vervaeet und T. De Beer. „Raman spectroscopy and multivariate analysis for the rapid discrimination between native-like and non-native states in freeze-dried protein formulations.“ *Eur J Pharm Biopharm* 85.2 (10/2013), S. 263–271. DOI: 10.1016/j.ejpb.2013.03.035.
- [156] E. Fernández-Ahumada, J. M. Roger, B. Palagos, J. E. Guerrero, D. Pérez-Marín und A. Garrido-Varo. „Multivariate near-infrared reflection spectroscopy strategies for ensuring correct labeling at feed bagging in the animal feed industry.“ *Appl Spectrosc* 64.1 (01/2010), S. 83–91.
- [157] H. S. Tapp, M. Defernez und E. K. Kemsley. „FTIR spectroscopy and multivariate analysis can distinguish the geographic origin of extra virgin olive oils.“ *J Agric Food Chem* 51.21 (10/2003), S. 6110–6115. DOI: 10.1021/jf030232s.
- [158] B. Efron. „The efficiency of logistic regression compared to normal discriminant analysis.“ *J Am Stat Assoc* 70.352 (1975), S. 892–898.
- [159] R. N. Forthofer, E. S. Lee und M. Hernandez. *Biostatistics - A Guide to Design, Analysis, and Discovery*. 2. Aufl. Academic Press, 2007.
- [160] F. E. Harrell und K. L. Lee. „A comparison of the discrimination of discriminant analysis and logistic regression under multivariate normality“. *Biostatistics: Statistics in Biomedical, Public Health and Environmental Sciences*, (1985), S. 333–343.
- [161] H. da Silva Martinho, C. M. de Oliveira Monteiro da Silva, M. C. B. M. Yassoyama, P. de Oliveira Andrade, R. A. Bitar, A. M. do Espírito Santo, E. A. L. Arisawa und A. A. Martin. „Role of cervicitis in the Raman-based optical diagnosis of cervical intraepithelial neoplasia.“ *J Biomed Opt* 13.5 (2008), S. 054029. DOI: 10.1117/1.2976114.
- [162] A. Robichaux-Viehoever, E. Kanter, H. Shappell, D. Billheimer, H. Jones und A. Mahadevan-Jansen. „Characterization of Raman spectra measured in vivo for the detection of cervical dysplasia.“ *Appl Spectrosc* 61.9 (9/2007), S. 986–993.
- [163] S. C. G. Penteado, B. P. Fogazza, C. da Silva Carvalho, E. A. L. Arisawa, M. A. Martins, A. A. Martin und H. da Silva Martinho. „Diagnosis of degenerative lesions of supraspinatus rotator cuff tendons by Fourier transform-Raman spectroscopy.“ *J Biomed Opt* 13.1 (2008), S. 014018. DOI: 10.1117/1.2841017.
- [164] A. Nijssen, T. C. B. Schut, F. Heule, P. J. Caspers, D. P. Hayes, M. H. A. Neumann und G. J. Puppels. „Discriminating basal cell carcinoma from its surrounding tissue by Raman spectroscopy.“ *The Journal of investigative dermatology* 119.1 (7/2002), S. 64–69. DOI: 10.1046/j.1523-1747.2002.01807.x.
- [165] S. K. Majumder, M. D. Keller, F. I. Boulos, M. C. Kelley und A. Mahadevan-Jansen. „Comparison of autofluorescence, diffuse reflectance, and Raman spectroscopy for breast tissue discrimination.“ *J Biomed Opt* 13.5 (2008), S. 054009. DOI: 10.1117/1.2975962.
- [166] C. A. Lieber, S. K. Majumder, D. Billheimer, D. L. Ellis und A. Mahadevan-Jansen. „Raman microspectroscopy for skin cancer detection in vitro.“ *J Biomed Opt* 13.2 (2008), S. 024013. DOI: 10.1117/1.2899155.
- [167] R. Brereton. *Chemometrics for pattern recognition*. Chichester, U.K: Wiley, 2009. ISBN: 9780470987254.
- [168] R. G. Brereton und G. R. Lloyd. „Support vector machines for classification and regression.“ *Analyst* 135.2 (2/2010), S. 230–267. DOI: 10.1039/b918972f.

- [169] T. Czekaj, W. Wu und B. Walczak. „About kernel latent variable approaches and SVM“. *Journal of Chemometrics* 19.5–7 (2006), S. 342–354. DOI: 10.1002/cem.937.
- [170] S. Mika, G. Ratsch, J. Weston, B. Schölkopf und K. Müller. „Fisher discriminant analysis with kernels“. In: *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop*. 1999, S. 41–48. DOI: 10.1109/NNSP.1999.788121. URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=788121>.
- [171] J. Zhu und T. Hastie. „Kernel Logistic Regression and the Import Vector Machine“. *Journal of Computational and Graphical Statistics* 14.1 (2005), pages.
- [172] M. Sattlecker, R. Baker, N. Stone und C. Bessant. „Support vector machine ensembles for breast cancer type prediction from mid-FTIR micro-calcification spectra“. *Chemom Intell Lab Syst* 107.2 (2011), S. 363–370. DOI: 10.1016/j.chemolab.2011.05.007.
- [173] C.-C. Chang und C.-J. Lin. „LIBSVM: A library for support vector machines“. *ACM Transactions on Intelligent Systems and Technology* 2 (3 2011). Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>, 27:1–27:27.
- [174] W. S. Sarle. „AI-FAQ“. comp.ai.neural-nets FAQ. 1997–2002. URL: <ftp://ftp.sas.com/pub/neural/FAQ.html>.
- [175] K. K. Dobbin und R. M. Simon. „Sample size planning for developing classifiers using high-dimensional DNA microarray data.“ *Biostatistics* 8.1 (1/2007), S. 101–117. DOI: 10.1093/biostatistics/kxj036.
- [176] K. K. Dobbin, Y. Zhao und R. M. Simon. „How large a training set is needed to develop a classifier for microarray data?“ *Clin Cancer Res* 14.1 (1/2008), S. 108–114. DOI: 10.1158/1078-0432.CCR-07-0443.
- [177] S. Mukherjee, P. Tamayo, S. Rogers, R. Rifkin, A. Engle, C. Campbell, T. R. Golub und J. P. Mesirov. „Estimating dataset size requirements for classifying DNA microarray data.“ *J Comput Biol* 10.2 (2003), S. 119–142. DOI: 10.1089/106652703321825928.
- [178] R. L. Figueroa, Q. Zeng-Treitler, S. Kandula und L. H. Ngo. „Predicting sample size required for classification performance.“ *BMC Med Inform Decis Mak* 12.1 (2/2012), S. 8. DOI: 10.1186/1472-6947-12-8.
- [179] D. M. Tax. „One-class classification – Concept-learning in the absence of counter-examples“. Doktorarbeit. Technische Universiteit Delft, 2001.
- [180] R. G. Brereton. „One-class classifiers“. *J Chemom* 25 (2011), S. 225–246. DOI: 10.1002/cem.1397.
- [181] S. Wold und M. Sjöström. „SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy“. In: *Chemometrics: Theory and Application*. 1977. Kap. 12, S. 243–282. DOI: 10.1021/bk-1977-0052.ch012. URL: <http://pubs.acs.org/doi/abs/10.1021/bk-1977-0052.ch012>.
- [182] C. Krafft, L. Shapoval, S. B. Sobottka, K. D. Geiger, G. Schackert und R. Salzer. „Identification of primary tumors of brain metastases by SIMCA classification of IR spectroscopic images.“ *Biochim Biophys Acta* 1758.7 (7/2006), S. 883–891. DOI: 10.1016/j.bbamem.2006.05.001.
- [183] L. Zadeh. „Fuzzy sets“. *Information and Control* 8.3 (6/1965), S. 338–353.
- [184] P. Lasch, W. Haensch, D. Naumann und M. Diem. „Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis.“ *Biochim Biophys Acta* 1688.2 (3/2004), S. 176–186. DOI: 10.1016/j.bbadis.2003.12.006.

- [185] W. Steller, J. Einkenkel, L.-C. Horn, U.-D. Braumann, H. Binder, R. Salzer und C. Krafft. „Delimitation of squamous cell cervical carcinoma using infrared microspectroscopic imaging.“ *Anal Bioanal Chem* 384.1 (1/2006), S. 145–154. DOI: 10.1007/s00216-005-0124-4.
- [186] M. Kuske, R. Rubio, A. C. Romain, J. Nicolas und S. Marco. „Fuzzy k-NN applied to moulds detection“. *Sensors and Actuators B: Chemical* 106.1 (4/2005), S. 52–60.
- [187] C. Kendall, N. Stone, N. Shepherd, K. Geboes, B. Warren, R. Bennett und H. Barr. „Raman spectroscopy, a potential tool for the objective identification and classification of neoplasia in Barrett’s oesophagus.“ *The Journal of pathology* 200.5 (8/2003), S. 602–609. DOI: 10.1002/path.1376.
- [188] G. R. Lloyd, M. Almond, N. Stone, N. Shepherd, S. Sanders, J. Hutchings, H. Barr und C. Kendall. „Utilising non-consensus pathology measurements to improve the diagnosis of oesophageal cancer using a Raman spectroscopic probe“. *Analyst* (2013), pages. DOI: 10.1039/C3AN01163A.
- [189] A. de Juan, M. Mäder, M. Martínez und R. Tauler. „Combining hard- and soft-modelling to solve kinetic problems“. *Chemom Intell Lab Syst* 54 (2000), S. 123–141. DOI: 10.1016/S0169-7439(00)00112-X.
- [190] K. Varmuza und P. Filzmoser. *Introduction to multivariate statistical analysis in chemometrics*. Boca Raton: CRC Press, 2009. ISBN: 9781420059472.
- [191] E. Binaghi, P. A. Brivio, P. Ghezzi und A. Rampini. „A fuzzy set-based accuracy assessment of soft classification“. *Pattern Recognition Letters* 20.9 (9/1999), S. 935–948.
- [192] R. G. Pontius und M. L. Cheuk. „A generalized cross-tabulation matrix to compare soft-classified maps at multiple resolutions“. *International Journal of Geographical Information Science* 20.1 (2006), S. 1–30.
- [193] J. Pontius Robert Gilmore und J. Connors. „Expanding the conceptual, mathematical and practical methods for map comparison“. In: *7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*. Hrsg. von M. Caetano und M. Painho. Proceedings of Accuracy 2006. plenary lecture. 2006. URL: [http://www.clarku.edu/~rpontius/pontius\\_connors\\_2006\\_spatialaccuracy.pdf](http://www.clarku.edu/~rpontius/pontius_connors_2006_spatialaccuracy.pdf) <http://spatial-accuracy.org/system/files/Pontius2006accuracy.pdf>.
- [194] J. Silván-Cárdenas und L. Wang. „Sub-pixel confusion-uncertainty matrix for assessing soft classifications“. *Remote Sensing of Environment* 112.3 (3/2008), S. 1081–1095. DOI: 10.1016/j.rse.2007.07.017.
- [195] V. Fedorov, F. Mannino und R. Zhang. „Consequences of dichotomization“. *Pharm Stat* 8 (2009), S. 50–61. DOI: 10.1002/pst.331.
- [196] P. Royston, D. G. Altman und W. Sauerbrei. „Dichotomizing continuous predictors in multiple regression: a bad idea.“ *Stat Med* 25.1 (1/2006), S. 127–141. DOI: 10.1002/sim.2331.
- [197] R. C. MacCallum, S. Zhang, K. J. Preacher und D. D. Rucker. „On the practice of dichotomization of quantitative variables.“ *Psychol Methods* 7.1 (3/2002), S. 19–40. DOI: 10.1037/1082-989X.7.1.19.



- [198] S. J. Dixon, N. Heinrich, M. Holmboe, M. L. Schaefer, R. R. Reed, J. Trevejo und R. G. Brereton. „Application of classification methods when group sizes are unequal by incorporation of prior probabilities to three common approaches: Application to simulations and mouse urinary chemosignals“. *Chemom Intell Lab Syst* 99.2 (2009), S. 111–120. DOI: 10.1016/j.chemolab.2009.07.016.
- [199] L. Breiman. „Bagging Predictors“. *Machine Learning* 24 (1996), S. 123–140.
- [200] B. H. Menze, W. Petrich und F. A. Hamprecht. „Multivariate feature selection and hierarchical classification for infrared spectroscopy: serum-based detection of bovine spongiform encephalopathy.“ *Anal Bioanal Chem* 387.5 (3/2007), S. 1801–1807. DOI: 10.1007/s00216-006-1070-5.
- [201] R. K. Harrop Galvão, M. C. U. Araújo, M. do Nascimento Martins, G. E. José, M. J. Coelho Pontes, E. C. Silva und T. C. Bezerra Saldanha. „An application of subagging for the improvement of prediction accuracy of multivariate calibration models“. *Chemom Intell Lab Syst* 81.1 (3/2006), S. 60–67.
- [202] R. W. Johnson. „An Introduction to the Bootstrap“. *Teaching Statistics* 23.2 (2001), S. 49–54. DOI: doi :10.1111/1467-9639.00050.
- [203] R. Wehrens, H. Putter und L. M. C. Buydens. „The bootstrap: a tutorial“. *Chemom Intell Lab Syst* 54.1 (12/2000), S. 35–52.
- [204] B. Efron und R. Tibshirani. *An introduction to the bootstrap*. Chapman & Hall, New York, 1993.
- [205] R. Kohavi. „A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection“. In: *Artificial Intelligence Proceedings 14<sup>th</sup> International Joint Conference, 20 – 25. August 1995, Montréal, Québec, Canada*. Hrsg. von C. S. Mellish. Morgan Kaufmann, USA, 1995, S. 1137–1145.
- [206] A. Blum, A. Kalai und J. Langford. „Beating the Hold-Out: Bounds for K-fold and Progressive Cross-Validation“. In: *COLT*. 1999, S. 203–208. DOI: <http://doi.acm.org/10.1145/307400.307439>.
- [207] S. Arlot und A. Celisse. „A survey of cross-validation procedures for model selection“. *Statist. Surv.* 4 (2010), S. 40–79. DOI: 10.1214/09-SS054.
- [208] K. H. Esbensen und P. Geladi. „Principles of Proper Validation: use and abuse of re-sampling for validation“. *J Chemom* 24.3-4 (2010), S. 168–187.
- [209] B. Hanczar, J. Hua und E. R. Dougherty. „Decorrelation of the true and estimated classifier errors in high-dimensional settings.“ *EURASIP J Bioinform Syst Biol* (2007), S. 38473. DOI: 10.1155/2007/38473.
- [210] U. M. Braga-Neto und E. R. Dougherty. „Is cross-validation valid for small-sample microarray classification?“ *Bioinformatics* 20.3 (2/2004), S. 374–380. DOI: 10.1093/bioinformatics/btg419.
- [211] J.-H. Kim. „Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap“. *Computational Statistics & Data Analysis* 53.11 (2009), S. 3735–3745. DOI: 10.1016/j.csda.2009.04.009.
- [212] L. Breiman. *Out-Of-Bag Estimation*. Techn. Ber. 1996.
- [213] P. Filzmoser, B. Liebmann und K. Varmuza. „Repeated double cross validation“. *J Chemom* 23.4 (2009), S. 160–171. DOI: 10.1002/cem.1225.

- [214] F. Provost, T. Fawcett und R. Kohavi. „The Case Against Accuracy Estimation for Comparing Induction Algorithms“. In: *In Proceedings of the Fifteenth International Conference on Machine Learning*. 1998.
- [215] A.-L. Boulesteix und C. Strobl. „Optimal classifier selection and negative bias in error rate estimation: an empirical study on high-dimensional prediction.“ *BMC Med Res Methodol* 9 (2009), S. 85. DOI: 10.1186/1471-2288-9-85.
- [216] M. R. Yousefi, J. Hua und E. R. Dougherty. „Multiple-rule bias in the comparison of classification rules.“ *Bioinformatics* 27.12 (6/2011), S. 1675–1683. DOI: 10.1093/bioinformatics/btr262.
- [217] M. M. Leeftang, K. G. Moons, J. B. Reitsma und A. H. Zwinderman. „Bias in Sensitivity and Specificity Caused by Data-Driven Selection of Optimal Cutoff Values: Mechanisms, Magnitude, and Solutions“. *Clin Chem* 54.4 (2008), S. 729–737.
- [218] S. Salzberg. „On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach“. *Data Mining and Knowledge Discovery* 1.3 (9/1997), S. 317–328.
- [219] P. Filzmoser, S. Serneels, R. Maronna und P. J. V. Espen. „Robust multivariate methods in chemometrics“. In: *Comprehensive Chemometrics*. Hrsg. von B. Walczak, R. Ferre und S. Brown. 2009.
- [220] S. Ellison und T. Fearn. „Characterising the performance of qualitative analytical methods: Statistics and terminology“. *TrAC Trends in Analytical Chemistry* 24.6 (6/2005), S. 468–476.
- [221] T. Fawcett. „An introduction to ROC analysis“. *Pattern Recognition Letters* 27.8 (6/2006), S. 861–874.
- [222] D. Simon und I. John R. Boring. „Clinical Methods: The History, Physical and Laboratory Examinations“. In: Hrsg. von J. W. Hurst. 3. Aufl. Butterworth-Heinemann, 1990. Kap. Sensitivity, Specificity, and Predictive Value, S. 49–54. URL: <http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=cm&part=A347>.
- [223] J. L. Fleiss, B. Levin und M. C. Paik. *Statistical Methods for Rates and Proportions*. 3rd Edition. New Jersey: Wiley-Interscience, 2003, S. xxvii + 760.
- [224] K. Danzer. *Analytical Chemistry — Theoretical and Metrological Fundamentals*. Springer, 2007. ISBN: 978-3-540-35990-6. DOI: 10.1007/978-3-540-35990-6.
- [225] A. D. McNaught und A. Wilkinson, Hrsg. *Compendium of Chemical Terminology (the "Gold Book")*. 2. Aufl. Blackwell Scientific, 1997. ISBN: 0-9678550-9-8. DOI: doi:10.1351/goldbook. URL: <http://goldbook.iupac.org/>.
- [226] J. Hilden. „Prevalence-free utility-respecting summary indices of diagnostic power do not exist.“ *Stat Med* 19.4 (2/2000), S. 431–440. DOI: 10.1002/(SICI)1097-0258(20000229)19:4<431::AID-SIM348>3.0.CO;2-R.
- [227] D. L. Simel, G. P. Samsa und D. B. Matchar. „Likelihood ratios with confidence: sample size estimation for diagnostic test studies.“ *J Clin Epidemiol* 44.8 (1991), S. 763–770. DOI: doi:10.1016/0895-4356(91)90128-V.
- [228] N. J. Perkins und E. F. Schisterman. „The inconsistency of optimal cutpoints obtained using two criteria based on the receiver operating characteristic curve.“ *Am J Epidemiol* 163.7 (4/2006), S. 670–675. DOI: 10.1093/aje/kwj063.

- [229] T. D. Ross. „Accurate confidence intervals for binomial proportion and Poisson rate estimation“. *Comput Biol Med* 33 (2003), S. 509–531. DOI: 10 . 1016 / S0010 - 4825 (03 ) 00019-2.
- [230] L. Brown, T. Cai und A. DasGupta. „Interval Estimation for a Binomial Proportion“. *Statistical Science* 16 (2001), S. 101–133.
- [231] A. M. Pires und C. Amado. „Interval Estimators for a Binomial Proportion: Comparison of Twenty Methods“. *Revstat – Statistical Journal* 6.2 (2008), S. 165–197.
- [232] Y. Bengio und Y. Grandvalet. „No Unbiased Estimator of the Variance of K-Fold Cross-Validation“. *Journal of Machine Learning Research* 5 (2004), S. 1089–1105.
- [233] S. A. Julious. *Sample sizes for clinical trials*. Boca Raton, Fla. [u.a.]: CRC Press, 2010. XXVII, 299. ISBN: 9781584887393.
- [234] G. M. Foody. „Status of land cover classification accuracy assessment“. *Remote Sensing of Environment* 80.1 (4/2002), S. 185–201.
- [235] H. G. Lewis und M. Brown. „A generalized confusion matrix for assessing area estimates from remotely sensed data“. *Int J Remote Sens* 22.16 (2001), S. 3223–3235.
- [236] S. Gottwald. „Many-Valued Logic“. In: *The Stanford Encyclopedia of Philosophy*. Hrsg. von E. N. Zalta. Spring 2010. 2010.
- [237] P. Hajek. „Fuzzy Logic“. In: *The Stanford Encyclopedia of Philosophy*. Hrsg. von E. N. Zalta. Fall 2010. 2010.
- [238] D. Dubois und H. Prade. „A review of fuzzy set aggregation connectives“. *Information sciences* 36.1-2 (7/1985), S. 85–121.
- [239] J. Łukasiewicz und A. Tarski. „Untersuchungen über den Aussagenkalkül“. *Sprawozdania z posiedzeń Towarzystwa Naukowego Warszawskiego* 3.23 (1930), S. 1–21.
- [240] H. Reichenbach. „Wahrscheinlichkeitslogik“. *Erkenntnis* 5 (1935), S. 37–43.
- [241] D. Gómez, G. Biging und J. Montero. „Accuracy statistics for judging soft classification“. *Int J Remote Sens* 29.3 (2008), S. 693–709.
- [242] J. Chen, X. Zhu, H. Imura und X. Chen. „Consistency of accuracy assessment indices for soft classification: Simulation analysis“. *ISPRS Journal of Photogrammetry and Remote Sensing* 65.2 (3/2010), S. 156–164. DOI: 10 . 1016 / j . isprsjprs . 2009 . 10 . 003.
- [243] A. Comber, P. Fisher, C. Brunsdon und A. Khmag. „Spatial analysis of remote sensing image classification accuracy“. *Remote Sensing of Environment* 127 (2012), S. 237–246. DOI: 10 . 1016 / j . rse . 2012 . 09 . 005.
- [244] M. Diem, P. R. Griffiths und J. M. Chalmers, Hrsg. *Vibrational Spectroscopy for Medical Diagnosis*. Wiley, 2008. ISBN: 978-0-470-01214-7.
- [245] U. Neugebauer, T. Bocklitz, J. H. Clement, C. Krafft und J. Popp. „Towards detection and identification of circulating tumour cells using Raman spectroscopy.“ *Analyst* 135.12 (12/2010), S. 3178–3182. DOI: 10 . 1039 / c0an00608d.
- [246] U. Neugebauer, J. H. Clement, T. Bocklitz, C. Krafft und J. Popp. „Identification and differentiation of single cells from peripheral blood by Raman spectroscopic imaging.“ *J Biophotonics* 3.8-9 (8/2010), S. 579–587. DOI: 10 . 1002 / jbio . 201000020.
- [247] G. W. Brier. „Verification of Forecasts Expressed in Terms of Probability“. *Mon Wea Rev* 78.1 (1/1950), S. 1–3.

## Referenzen

- [248] T. Gneiting und A. E. Raftery. *Strictly Proper Scoring Rules, Prediction, and Estimation*. Techn. Ber. 463R. Department of Statistics, University of Washington, 2005. URL: <http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA454828>.
- [249] A. Nikulin, B. Dolenko, T. Bezabeh und R. Somorjai. „Near-optimal region selection for feature space reduction: novel preprocessing methods for classifying MR spectra“. *NMR Biomed* 11 (1998), S. 209–216.
- [250] S. Droste, T. Jansen und I. Wegener. „Optimization with randomized search heuristics—the (A)NFL theorem, realistic scenarios, and difficult functions“. *Theoretical Computer Science* 287.1 (2002). <ce:title>Natural Computing</ce:title>, S. 131–144. DOI: 10.1016/S0304-3975(02)00094-4.
- [251] J. C. Culberson. *On the Futility of Blind Search*. Technical Report TR 96-18. Department of Computing Science, The University of Alberta, Edmonton, AB, Canada, 9/1996.
- [252] U. Braga-Neto und E. Dougherty. „Bolstered error estimation“. *Pattern recognition* 37 (2004), S. 1267–1281.
- [253] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2013. URL: <http://www.R-project.org>.
- [254] E. Check. „Cancer atlas maps out sample worries“. *Nature* 447.7148 (6/2007), S. 1036–1037.
- [255] K. D. Geiger. persönliche Mitteilung. 20. 11. 2006.
- [256] Z. Xianyi, W. Qian und Z. Yunquan. „Model-driven Level 3 BLAS Performance Optimization on Loongson 3A Processor“. In: *IEEE 18th International Conference on Parallel and Distributed Systems (ICPADS) 17–19 Dec. 2012*. 2012.
- [257] D. Sarkar. *Lattice*. R package version 0.18-8. Springer New York, 2008. ISBN: 978-0-387-75968-5. URL: <http://www.springerlink.com/content/978-0-387-75968-5>.
- [258] H. Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009. ISBN: 978-0-387-98140-6. URL: <http://had.co.nz/ggplot2/book>.
- [259] H. Wickham. „The Split-Apply-Combine Strategy for Data Analysis“. *Journal of Statistical Software* 40.1 (2011), S. 1–29.
- [260] H. Wickham. „Reshaping Data with the reshape Package“. *Journal of Statistical Software* 21.12 (2007), S. 1–20.
- [261] Y. Xie. *knitr: A general-purpose package for dynamic report generation in R*. R package version 1.2. 2013. URL: <http://yihui.name/knitr/>.
- [262] Y. Xie. *Dynamic Documents with R and knitr*. ISBN 978-1482203530. Chapman und Hall/CRC, 2013. URL: <http://yihui.name/knitr/>.
- [263] Y. Xie. „knitr: A Comprehensive Tool for Reproducible Research in R“. In: *Implementing Reproducible Computational Research*. Hrsg. von V. Stodden, F. Leisch und R. D. Peng. ISBN 978-1466561595. Chapman und Hall/CRC, 2013. URL: <http://www.crcpress.com/product/isbn/9781466561595>.
- [264] H. Bengtsson. *R.matlab: Read and write of MAT files together with R-to-Matlab connectivity*. R package version 1.7.0. 2013. URL: <http://CRAN.R-project.org/package=R.matlab>.
- [265] B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 18. 1. 1996. ISBN: 0521460867. URL: <http://www.worldcat.org/isbn/0521460867>.

- [266] B. Mevik und R. Wehrens. „The pls Package: Principal Component and Partial Least Squares Regression in R“. *Journal of Statistical Software* 18.2 (10. 1. 2007), S. 1–24.
- [267] L. Tierney, A. J. Rossini, N. Li und H. Sevcikova. *snow: Simple Network of Workstations*. R package version 0.3-10. 2012. URL: <http://CRAN.R-project.org/package=snow>.
- [268] J. L. Koenig und A. C. Angood. „Raman spectra of poly(ethylene glycols) in solution“. *Journal of Polymer Science Part A-2: Polymer Physics* 8.10 (1970), S. 1787–1796. DOI: 10.1002/pol.1970.160081013.
- [269] G. Schwedt. *Analytische Chemie: Grundlagen, Methoden Und Praxis*. 2. Aufl. Wiley-VCH, 2008. ISBN: 9783527312061.
- [270] J. M. Kros, T. Gorlia, M. C. Kouwenhoven, P-P. Zheng, V. P. Collins, D. Figarella-Branger, F. Giangaspero, C. Giannini, K. Mokhtari, S. J. Mørk, A. Paetau, G. Reifenberger und M. J. van den Bent. „Panel review of anaplastic oligodendroglioma from European Organization For Research and Treatment of Cancer Trial 26951: assessment of consensus in diagnosis, influence of 1p/19q loss, and correlations with outcome.“ *J Neuropathol Exp Neurol* 66.6 (6/2007), S. 545–551. DOI: 10.1097/01.jnen.0000263869.84188.72.
- [271] R. A. Prayson, D. P. Agamanolis, M. L. Cohen, M. L. Estes, B. K. Kleinschmidt-DeMasters, F. Abdul-Karim, S. P. McClure, B. A. Sebek und R. Vinay. „Interobserver reproducibility among neuropathologists and surgical pathologists in fibrillary astrocytoma grading.“ *J Neurol Sci* 175.1 (4/2000), S. 33–39. DOI: [http://dx.doi.org/10.1016/S0022-510X\(00\)00274-4](http://dx.doi.org/10.1016/S0022-510X(00)00274-4).
- [272] D. Petrowsky. „Zusammensetzung der grauen und der weissen Substanz des Gehirns“. *Archiv für die gesamte Physiologie des Menschen und der Tiere* 7.1 (1873), S. 367–370. DOI: 10.1007/BF01613333.
- [273] D. Carr, ported by Nicholas Lewin-Koh und M. Maechler. *hexbin: Hexagonal Binning Routines*. R package version 1.26.2. 2013. URL: <http://CRAN.R-project.org/package=hexbin>.
- [274] S. K. Majumder, S. Gebhart, M. D. Johnson, R. Thompson, W.-C. Lin und A. Mahadevan-Jansen. „A probability-based spectroscopic diagnostic algorithm for simultaneous discrimination of brain tumor and tumor margins from normal brain tissue.“ *Appl Spectrosc* 61.5 (5/2007), S. 548–557. DOI: 10.1366/000370207780807704.
- [275] G. Socrates. *Infrared and Raman Characteristic Group Frequencies*. 3. Aufl. Wiley, 2001. ISBN: 978-0-470-09307-8.
- [276] I. Berget und T. Næs. „Using unclassified observations for improving classifiers“. *J Chemom* 18.2 (2004), S. 103–111.
- [277] „WHO Classification of tumors of the central nervous system“. In: Hrsg. von D. N. Louis, W. K. Cavenee, H. Ohgaki und O. D. Wiestler. World Health Organization, 2007. Kap. Astrocytic tumors. ISBN: 9789283224303.
- [278] J. De Gelder, K. De Gussem, P. Vandenabeele und L. Moens. „Reference database of Raman spectra of biological molecules“. *J Raman Spectrosc* 38.9 (2007), S. 1133–1147.
- [279] M. S. Bergholt, W. Zheng, K. Lin, K. Y. Ho, M. Teh, K. G. Yeoh, J. B. So und Z. Huang. „In vivo diagnosis of esophageal cancer using image-guided Raman endoscopy and biomolecular modeling.“ *Technol Cancer Res Treat* 10.2 (04/2011), S. 103–112.

- [280] O. Chapelle, B. Schölkopf und A. Zien. „Semi-Supervised Learning — Adaptive Computation and Machine Learning“. In: MIT Press, 2006. Kap. Introduction to Semi-Supervised Learning, S. 1–12. URL: <http://mitpress.mit.edu/books/semi-supervised-learning>.
- [281] S. Ben-David, T. Lu und D. Pál. „Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning“. In: *21st Annual Conference on Learning Theory*. 2008.
- [282] F. G. Cozman und I. Cohen. *Unlabeled Data Can Degrade Classification Performance of Generative Classifiers*. Techn. Ber. HPL-2001-234. HP Laboratories Palo Alto, 9/2001. URL: <http://www.hpl.hp.com/techreports/2001/HPL-2001-234.html>.
- [283] I. Cohen, F. G. Cozman und A. Bronstein. *The effect of unlabeled data on generative classifiers, with application to model selection*. Techn. Ber. HPL-2002-140. HP Labs, 2002. URL: <http://www.hpl.hp.com/techreports/2002/HPL-2002-140.html>.
- [284] L. Xu, M. White und D. Schuurmans. „Optimal reverse prediction: a unified perspective on supervised, unsupervised and semi-supervised learning“. In: *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, S. 1137–1144. ISBN: 978-1-60558-516-1. DOI: <http://doi.acm.org/10.1145/1553374.1553519>.
- [285] T. Dörfer, T. Bocklitz, N. Tarcea, M. Schmitt und J. Popp. „Checking and Improving Calibration of Raman Spectra using Chemometric Approaches“. *Zeitschrift für Physikalische Chemie* 225.6-7 (2011), S. 753–764. DOI: 10.1524/zpch.2011.0077.
- [286] H. Mark und J. Workman. *Chemometrics in spectroscopy*. Amsterdam Oxford: Academic, 2007. ISBN: 012374024X.
- [287] C. M. Gryniewicz-Ruzicka, S. Arzhantsev, L. N. Pelster, B. J. Westenberger, L. F. Buhse und J. F. Kauffman. „Multivariate calibration and instrument standardization for the rapid detection of diethylene glycol in glycerin by Raman spectroscopy.“ *Appl Spectrosc* 65.3 (03/2011), S. 334–341. DOI: 10.1366/10-05976.
- [288] J. D. Rodriguez, B. J. Westenberger, L. F. Buhse und J. F. Kauffman. „Standardization of Raman spectra for transfer of spectral libraries across different instruments.“ *Analyst* 136.20 (10/2011), S. 4232–4240. DOI: 10.1039/c1an15636e.
- [289] M. Kompany-Zareh und F. van den Berg. „Multi-way based calibration transfer between two Raman spectrometers.“ *Analyst* 135.6 (06/2010), S. 1382–1388. DOI: 10.1039/b927501k.
- [290] J. Engel, J. Gerretzen, E. Szymańska, J. J. Jansen, G. Downey, L. Blanchet und L. M. Buydens. „Breaking with trends in pre-processing?“ *TrAC Trends in Analytical Chemistry* 50 (2013), S. 96–106. DOI: <http://dx.doi.org/10.1016/j.trac.2013.04.015>.
- [291] T. Bocklitz, A. Walter, K. Hartmann, P. Rösch und J. Popp. „How to pre-process Raman spectra for reliable and stable models?“ *Anal Chim Acta* 704.1-2 (10/2011), S. 47–56. DOI: 10.1016/j.aca.2011.06.043.
- [292] M. A. Bezerra, R. E. Santelli, E. P. Oliveira, L. S. Villar und L. A. Escalera. „Response surface methodology (RSM) as a tool for optimization in analytical chemistry.“ *Talanta* 76.5 (09/2008), S. 965–977. DOI: 10.1016/j.talanta.2008.05.019.
- [293] R. Leardi. „Experimental design in chemistry: A tutorial“. *Analytica Chimica Acta* 652.1–2 (2009), S. 161–172. DOI: <http://dx.doi.org/10.1016/j.aca.2009.06.015>.

- [294] L. Buchen. „Cancer: Missing the mark.“ *Nature* 471.7339 (3/2011), S. 428–432. DOI: 10.1038/471428a.
- [295] G. Gigerenzer. *Das Einmaleins der Skepsis. über den richtigen Umgang mit Zahlen und Risiken*. Hrsg. von M. Zillgitt. 6. Aufl. BvT 41. Berlin: Berliner Taschenbuch-Verl., 2009. 406 S. ISBN: 9783833300417.
- [296] T. W. Bocklitz, A. C. Crecelius, C. Matthäus, N. Tarcea, F. von Eggeling, M. Schmitt, U. S. Schubert und J. Popp. „Deeper understanding of biological tissue: quantitative correlation of MALDI-TOF and Raman imaging.“ *Anal Chem* 85.22 (11/2013), S. 10829–10834. DOI: 10.1021/ac402175c.
- [297] E. Merck. *Klinisches Labor*. 12. E. Merck, Darmstadt, 1974.
- [298] *Pschyrembel Klinisches Wörterbuch*. 258., neu bearbeitete Auflage. Walter de Gruyter & Co., 1998.
- [299] *Digital Library of Mathematical Functions*. National Institute of Standards und Technology. 05/2010. URL: <http://dlmf.nist.gov/>.
- [300] *NIST/SEMATECH e-Handbook of Statistical Methods*. NIST/SEMATECH. URL: <http://www.itl.nist.gov/div898/handbook/>.

# Glossar

***a-posteriori*-Wahrscheinlichkeit** abhängige Wahrscheinlichkeit. Wahrscheinlichkeit, dass ein Ereignis eintritt, die die konkrete Beobachtung (Ausprägung der unabhängigen Variaten) berücksichtigt.

***a-priori*-Wahrscheinlichkeit** unabhängige Wahrscheinlichkeit. Eine Wahrscheinlichkeit, dass ein Ereignis eintritt, unabhängig von etwaigen Beobachtungen (Ausprägungen unabhängiger Variate). Siehe auch *a-posteriori*-Wahrscheinlichkeit.

**benigne** gutartiger Tumor. Oberbegriff: Dignität [298].

**Beta-Verteilung** Die (Standard) Beta-Verteilung ist definiert als

$$f(x) = \frac{1}{B(p, q)} x^{p-1} (1-x)^{q-1}$$

mit  $p$  und  $q > 0$  und  $x \in [0, 1]$ .

Der Normierungsfaktor ist Eulers Beta:

$$B(p, q) = \int_0^1 t^{p-1} (1-t)^{q-1} dt$$

Die kumulierte Wahrscheinlichkeitsverteilung ist die regularisierte unvollständige Betafunktion  $I_x$

$$I_x(p, q) = B_x(p, q) / B(p, q)$$

[299, 300].

**Clusteranalyse** Gruppe von unüberwachten Verfahren, die Gruppen ähnlicher Objekte finden. Meist werden Gruppen ähnlicher Proben (Spektren) gebildet, aber auch Variate können gruppiert werden. Siehe [145].

**deskriptiv** Beschreibende Modelle sollen Zusammenhänge in den Daten aufzeigen, aber nicht (zwingend) die Vorhersage neuer Daten ermöglichen. Daher wird bei deskriptiven Modellen größerer Wert auf die Interpretierbarkeit der Modelle gelegt, und weniger Wert auf eine Validierung mit unabhängigen Testdaten.

**Dignität** Biologische Wertigkeit eines Tumors [298]. Siehe maligne und benigne.

**Drift** Langsame, gerichtete Veränderung (zum Beispiel der Modellparameter).

**Endothel** Einschichtige Auskleidung der Blut- und Lymphgefäße. Siehe Epithel [298].

**Epithel** Geschlossener Zellverband, der innere oder äußere Körperoberflächen bedeckt. Seine Funktionen sind Schutz, Stoffaustausch und Reizaufnahme. Epithelien werden nach ihrer Histologie eingeteilt [298].



**ex vivo** Analyse: eine Probe wird entnommen und dann untersucht. Siehe auch *in vivo*, *in situ*.

**Fehler I. Art,  $\alpha$**  Gibt es bei einem statistischen Modellvergleich (allgemein: Hypothesentest) tatsächlich keinen Unterschied zwischen den Modellen (ist die Nullhypothese also tatsächlich wahr), so wird bei Wiederholungen des Experiments ein Anteil von  $\alpha$  falsch positiven Ergebnissen erwartet. Das heißt: bei ungefähr  $\alpha n$  Experimenten wird zufällig ein Unterschied zwischen den Modellen beobachtet, der mindestens so groß wie der angenommene Unterschied zwischen den Modellen (Hypothesen) ist.

$1 - \alpha$  heißt Konfidenz- oder Signifikanzniveau. Siehe auch Fehler II. Art.

**Fehler II. Art,  $\beta$**  Siehe Fehler I. Art: Sind die Modelle tatsächlich unterschiedlich (ist die Nullhypothese tatsächlich falsch), so wird das zufällig bei einem Anteil  $\beta$  der Wiederholungsexperimente nicht erkannt (falsch negativ).  $1 - \beta =$  Teststärke. Oft wird  $\beta = 4\alpha$  angestrebt..

**Gliose** Vermehrung des Glia im Hirngewebe, zum Beispiel Vernarbungen nach Verletzungen oder Operationen [298].

**Grading** Histopathologische Differenzierung maligner Tumore in vier Grade. Einteilung: °I (gut differenziert) bis °IV (undifferenziert). Je höher die Gradzahl, d. h. je undifferenzierter der Tumor ist, desto maligner ist er [298].

**Hyperparameter** Als *Parameter* eines chemometrischen Modells bezeichnet man alle Komponenten, die während der Modellbildung geschätzt werden, also zum Beispiel die Koeffizienten einer LDA oder logistischen Regression, oder die latenten Variablen und Gewichte bei einer PLS"-Regression. Viele Modelle benutzen weitere, sogenannte *Hyperparameter*, die zum Beispiel bestimmte Eigenschaften des Modelltrainingsalgorithmus oder die gewünschte Modellkomplexität steuern. Beispiele sind die Kostenfunktion für SVM mit weichem Rand (engl. *soft margin*), der Radius eines Gauß"-Kerns, Grenzwerte eines Klassifikationsmodells zum Umrechnen von *Scores* in harte Klassen oder auch die Anzahl der verwendeten Komponenten bei PCA und PLS [144].

Diese Hyperparameter werden in der Regel entweder direkt vom Nutzer festgelegt, oder einer Optimierung unterzogen.

**in situ** während der Operation im OP"-Bereich. Siehe auch *in vivo*, *ex vivo*.

**in vivo** am Lebenden. Analyse: Die Messung findet direkt am Patienten statt, eine Probennahme ist nicht erforderlich. Siehe auch *ex vivo*, *in situ*.

**Interquartilsabstand** Differenz zwischen drittem und erstem Quartil.

**Inzidenz** Anteil *Neuerkrankungen*, bezogen auf eine Population und einen Zeitraum (meist 100 000 Personenjahre).

$\kappa$  Die Kappastatistik ist eine um die erwartete Anzahl an zufälligen Übereinstimmungen korrigierte Version der Trefferrate [223].

**Kalibrierung** Eine Kalibrierung überprüft ein Messverfahren auf Richtigkeit und Präzision oder bestätigt, dass das Verfahren vorgegebene Toleranzen einhält. In diesem Sinne sind auch die Aussagen eines gut kalibrierten chemometrischen Modells richtig und präzise. In der chemischen Analytik bezeichnet Kalibrierung auch die Umrechnungsfunktion zwischen unabhängiger Größe, meist Konzentration, und abhängigen Messgrößen.

**Konfidenzniveau** auch Signifikanzniveau. Siehe Fehler I. Art.

**Liquor** Gehirn-/Rückenmarksflüssigkeit.

**Mahalanobis-Distanz** Die Mahalanobis-Distanz beschreibt den Abstand zweier Punkte (Spektren), die zu einer Gruppe (von Spektren) gehören. Dabei wird die Kovarianzstruktur dieser Gruppe berücksichtigt:

$$d_M(\mathbf{x}_a, \mathbf{x}_b) = \sqrt{(\mathbf{x}_a - \mathbf{x}_b)^T \mathbf{S}^{-1} (\mathbf{x}_a - \mathbf{x}_b)} \quad (1)$$

mit  $\mathbf{S} = \text{COV}(\mathbf{X})$  (2)

Der Abstand wird sozusagen in „Standardabweichungs-Einheiten“ angegeben.

Das „Berücksichtigen der Kovarianzstruktur“ ist eine Koordinatentransformation. Dabei wird aus der elliptischen Kovarianzstruktur eine sphärische Kovarianzstruktur. Diese Koordinatentransformation erfolgt mit Hilfe der Cholesky-Zerlegung von  $\mathbf{S}$ :  $\mathbf{G}\mathbf{G}^T = \mathbf{S}$ .

Die Mahalanobis-Distanz ist dann

$$d_M(\mathbf{x}_a, \mathbf{x}_b) = \sqrt{(\mathbf{G}^{-1}\mathbf{x})^T (\mathbf{G}^{-1}\mathbf{x})} \quad \text{mit } \mathbf{x} = \mathbf{x}_a - \mathbf{x}_b \quad (3)$$

$$= \sqrt{\mathbf{x}'^T \mathbf{x}'} \quad \text{mit } \mathbf{x}' = \mathbf{G}^{-1} (\mathbf{x}_a - \mathbf{x}_b) \quad (4)$$

Im neuen Koordinatensystem wird die euklidische Distanz gemessen.

**maligne** bösartiger Tumor. Oberbegriff: Dignität [298].

**Matrix** siehe Probenmatrix.

**Median** Die Hälfte der Werte einer Verteilung ist kleiner als ihr Median. Damit ist der Median das 50. Perzentil und das 2. Quartil.

**metrisch** Eine metrische Größe kann Zahlen und physikalische Größen abbilden. Die Werte sind geordnet, Abstände können berechnet werden und Werte können zueinander ins Verhältnis gesetzt werden. Siehe nominalskaliert, ordinalskaliert.

**nominalskaliert** Nominalskalierte Größen können nur bestimmte Werte annehmen, diese Werte sind nicht geordnet. Beispiel: Gewebearten. Siehe ordinalskaliert metrisch.

**Ödem** Bei einem Ödem wird zusätzliches Wasser ins Gewebe eingelagert und das Gewebe schwillt.

**überwachtes Modell** Modell, das unter Verwendung von abhängigen Variaten gebildet wird. Siehe unüberwachtes Modell.

**ordinalskaliert** Ordinalskalierte Größen können wie nominalskalierte Größen nur bestimmte Werte annehmen. Zwischen diesen möglichen Werten besteht aber eine Ordnung. Beispiel: Schulnoten. Siehe nominalskaliert metrisch.

**Perzentil**  $p$  % der Werte einer Verteilung sind kleiner als ihr  $p$ . Perzentil. Siehe Quantil, Quartil.

**Polymorphie** Vielgestaltigkeit.

**Probenmatrix** Alle Bestandteile einer Probe außer dem betrachteten Analyten [225].

**prädiktiv** Vorhersagende Modelle sollen in erster Linie auf neue Daten angewandt werden und daher für unbekannte Daten gute Vorhersagen liefern. Es wird gegebenenfalls weniger Wert auf die Interpretierbarkeit des Modells gelegt. Vorhersagende Modelle sind immer auch überwachtes Modelle Modelle.

**Prävalenz** Anteil Erkrankungen, bezogen auf eine Population (meist 100 000 Einwohner) zu einem gegebenen Zeitpunkt.

**Quantil** Der Anteil  $p$  einer Anzahl von Werten ist kleiner als das  $p$ -Quantil. Siehe Quartil, Perzentil, Median.

**Quartil**  $\frac{1}{4}$  der Werte ist kleiner als das 1. Quartil,  $\frac{2}{4}$  kleiner als das 2. Quartil (Median) und  $\frac{3}{4}$  kleiner als das 3. Quartil. Das 4. Quartil ist das Maximum der Werte. Siehe Quartil, Perzentil.

**Regularisierung** Regularisierung schränkt die Modellkomplexität ein. Dadurch lässt sich die Varianz der Modelle reduzieren. Andererseits wird ein Bias in Kauf genommen. Die Stärke der Einschränkung wird in der Regel durch einen Hyperparameter eingestellt. Unterschiedliche Regularisierungsmethoden erlauben es, unterschiedliches Vorwissen oder Erwartungen in das Modell einfließen zu lassen. Bei der PLS"-Regression ist das die Erwartung, dass sich sowohl die Spektren als auch die Abhängige durch wenige Komponenten beschreiben lassen, weil nur eine eingeschränkte Anzahl an chemischen Spezies eine Rolle spielt. Andere Regularisierungsmethoden wie der LASSO beschreiben die Erwartung, dass nicht alle oder sogar nur wenige Variate überhaupt Information zum gegebenen Problem bereitstellen, und daher die Koeffizienten für die anderen Variate Null sein sollten [145].

**Rezidiv** Wiederkehren einer Krankheit [298].

**Selektivität** Grad der Abwesenheit von Kreuzempfindlichkeiten [225].

**Sensitivität (engl. *sensitivity*)** *sensitivity* kann im Englischen sowohl für die Empfindlichkeit im Sinne der analytischen Chemie (also die Steigung einer Kalibrierkurve) stehen [225], als auch für die Sensitivität eines diagnostischen Tests oder eines Klassifikationsmodells. Siehe Kapitel 4.8.5 (S. 48), Gleichung 4.17 (S. 53) und Kapitel 8.2 (S. 78).

**Signifikanzniveau** auch Konfidenzniveau. Siehe Fehler I. Art.

**Spezies** Eine chemische Spezies besteht aus einer Menge an chemisch identischen molekularen Bestandteilen, die (auf der Zeitskala des jeweiligen Experiments) nicht unterschieden werden können. Insbesondere teilen sie dieselben Energielevels, können also auch spektroskopisch nicht unterschieden werden. Dabei ist nicht notwendig, dass ein einzelnes Teilchen der Spezies isoliert existieren kann, zum Beispiel kann ein Oberflächenoxid eine eigene chemische Spezies bilden, obwohl es nur auf der Oberfläche des Bulkmaterials tatsächlich vorkommen kann [225].

**Spezifität (engl. *specificity*)** In der analytischen Chemie die Abwesenheit von Kreuzempfindlichkeiten (qualitativ) [225]. In der medizinischen Diagnostik die Wahrscheinlichkeit, mit der ein Test Negative Proben richtig erkennt. Siehe Kapitel 4.8.5 (S. 48) Gleichung 4.18 und Kapitel 8.2 (S. 78).

**Stratifizierung** Randbedingung beim Ziehen von Stichproben. In der Grundgesamtheit gibt es verschiedene Gruppen. Stratifizierung stellt sicher, dass der Anteil der einzelnen Gruppen in der Stichprobe (annähernd) derselbe ist wie in der Grundgesamtheit.

**Surrogatmarker** Eine einfach(er) messbare Ersatzgröße, die anstelle einer nur schwer oder nicht zugänglichen Größe gemessen wird. Zum Beispiel messen viele klinische Studien über die Wirksamkeit von chirurgischen Hirntumorbehandlungen nicht direkt die Überlebenszeit oder Lebensqualität der Patienten, sondern ziehen als Ersatz die Größe des im Patienten verbliebenen Tumorstes heran, die bereits kurz nach der Operation gemessen wird.

**Teststärke** Die Teststärke (engl. *power*) ist die Wahrscheinlichkeit, bei einem statistischen Test die Nullhypothese richtig abzulehnen.

Die Teststärke für Hypothesentests entspricht der Sensitivität eines medizinischen Tests.

**Trefferquote** Auch Richtig-positiv-Quote bzw. Trefferrate (engl. *true positive rate, TPR*). Die Trefferquote ist das Verhältnis der richtig erkannten (positiv getesteten) Proben zu allen wirklich zu der Klasse gehörenden Proben. Siehe Kapitel 4.8.5 (S. 48).

**unüberwachtes Modell** Modell, das ohne Verwendung von abhängigen Variaten gebildet wird. Siehe überwachtes Modell.

**Variate** Ein- oder Ausgangsgröße eines statistischen/chemometrischen Modells.

**Violinen-Diagramm** Ein Violinen-Diagramm besteht aus spiegelsymmetrisch angeordneten Diagramm der Dichteverteilung einer Größe. Diese übernimmt die Funktion von Box und Whiskers beim Boxplot. Im Inneren können Median, Mittelwert oder andere Werte eingezeichnet werden.

# Index

- .632+bootstrap, 45, 73
- .632bootstrap, 45, 73
- 5-Aminolävulinsäure, *siehe* ALA
  
- a posteriori Wahrscheinlichkeit, 27
- Aggregation
  - als Filter, 46
  - Wiederholungsmessungen, 75
- AIDS, 7, 8
- ALA, 9
- Aminolävulinsäure, *siehe* ALA
- ANN, 30, 37
- Anregungswellenlänge, 14, 15
- Anti-Stokes-Strahlung, 14
- arrayhelpers, 97
- Astrozyt, 6
- Astrozytom, 2, 3, 6–7
- Auflösung
  - räumlich, 36
- Ausreißer, 30
- Ausschluss
  - Grenzfälle, 3, 38
  - Proben, 3, 38, 101
- Autofluoreszenz, 16, 127
  
- bagging, 40
- Bayes-Fehler, 32
- benigne, 216
- Beta-Verteilung, 216
- Bias, 32
  - optimistisch, 37, 45, 69, 72, 73, 77, 79, 95, 153
  - pessimistisch, 44, 45, 77–79
- bias-variance-tradeoff, 39
- bilineare Modelle, 23–26
- Binomialverteilung, 56
- Biopsie, 2, 3
- Blindwert, 23
- Blut-Hirn-Schranke, 6, 9, 11
- bolstered error estimation, 75
- Bootstrap, 41
- brain shift, 8
  
- Bulkprobe, 19, 20, 33, 89–91, 97, 101, 139, 156
- CARS, 19
- CCD-Kamera, 15
- Chance, 29
- Chancen, 51
- charakteristische Banden, 18
- chemischer Rang, 24
- closed world, *siehe* geschlossenes System
- Cluster-Annahme, 150
- Clusteranalyse, 216
  - fuzzy, 35
- Computertomographie, 8
- Cosmic Ray Filter, 93, 98
  
- Dateiserver, 190
- Datenbank, 189
  - PostgreSQL, 191
- Datenmatrix, 21
- Datenpunkt
  - $p$ -dimensional, 21
- Datenstruktur
  - hierarchisch, 47
- Datenvorbehandlung, 22–24, 97–101
  - als Teil des Modells, 22
  - Basislinien-Korrektur, 99
  - datengesteuert, 23, 33, 47
  - Dimensionsreduktion, 22
  - Dunkelstromkorrektur, 98
  - Intensitätskalibrierung, 98
  - normieren, 100
  - physikalische Effekte, 22
  - skalieren, 23, 101, 110
  - Verbesserung numerischer Eigenschaften,  
22, 23
  - zentrieren, 23, 100, 110
- deskriptiv, 216
- Diagnostik
  - intraoperativ, 2
- Dichotomisierung, *siehe* härten
- Differentialdiagnostik, 3, 9, 35, 38, 52, 53, 88,  
101, 105, 108, 109, 133–146, 150, 151,  
154

## Index

- Dignität, 5, 216
  - benigne, 216
  - maligne, 5, 6, 8, 9, 218
- Doppelblindstudie, 96
- Drift, 38, 43, 216
  
- Einbettmedium, 88, 112
- Einklassen-Klassifikation, 34, 35
- Endothel, 6, 216
- Ensemble-Modell, 31, 39, 40, 45–47, 68, 75, 109, 139
  - Validierung, 45
- Entdifferenzierung, 3, 6, 7, 36, 146
- Epithel, 6, 216
- ex vivo, 149, 217
- Extrapolation, 40
  - Lernkurve, 65
  - räumlich, 36, 116
  - zeitlich, 43, 152
  
- Fehler
  - Bayes, *siehe* Bayes-Fehler
  - I. Art, 217
  - II. Art, 217
  - Probennahme, 94, 115
  - systematisch, *siehe* Unsicherheit
  - zufällig, *siehe* Unsicherheit
- Fehlerquote, 45, 50, 51, 58
- Fehlerrate, *siehe* Fehlerquote
- Fernerkundung, 35
- Feuchte
  - Probe, 130
- fluorescence-guided surgery, 9
- Fluoreszenz, 15
- Fluoreszenz-geleitete Chirurgie, 9
- Freiheitsgrade, 33
  
- Gefrierartefakt, 89, 90
- Gefriermedium, *siehe* Einbettmedium
- Gefrierschnitt, 88
- ggplot2, 97
- Glioblastom, *siehe* Astrozytom, 2, 6–7, 33
- Gliom, 6–7
  - genetische Variation, 7
  - hochgradig, 6
  - Infiltration, 7
  - niedriggradig, 6
  - Polymorphie, 7
- Gliose, 6, 145, 217
- Größe
  - ordinalskaliert, 218
- Grading (Tumor), 3, 5, 17, 20, 32, 33, 36, 38, 74, 88, 91, 94, 95, 105, 108, 109, 114–116, 119–132, 136, 146, 148, 217
- Grenzfälle, 3
  - Ausschluss, 37
- Grenzfaelle, 4
- Grenzwert, 37
- grid search, *siehe* Rastersuche
- Größe
  - metrisch, 26, 27, 29, 37, 39, 40, 54, 218
  - nominalskaliert, 27, 39, 40, 218
  - ordinalskaliert, 27
  
- härten, 27, 37, 46, 83, 116, 131
- Halbüberwachte Modellbildung, 150
- Heterogenität, 3
  - molekularbiologisch, 7
  - räumlich, 7
- Hirntumor
  - Diagnostik, 8
- Hirntumordiagnostik, 2, 36
- Hirntumore, 5–11
- Histogramm
  - 2d, 120
- Histologie, 2, 3, 5, 7–9, 36, 38, 88, 89, 94–97, 103, 105, 114–116, 134, 138, 140–142, 144, 146, 147, 149, 151, 154
- HIV, 8
- hold-out-Validierung, 42, 45, 68, 72, 76
- HoloGRAM, 91, 188, 189
- Hyperparameter, 25, 32, 47, 48, 59, 68–70, 109, 153, 154, 217
- hyperSpec, 97
  
- in situ, 217
- in vivo, 2, 3, 14, 16, 18, 20, 31, 88, 131, 146, 149, 150, 217
- Infiltration, 3, 7, 36
- Interquartilsabstand, 217
  - 2d, 120
- Inzidenz, 5, 7, 8, 27, 51, 145, 217
- Iterationen
  - Validierung, *siehe* Validierung
  
- Jackknife, 41
  
- Kalibrierung, 217
  - von Vorhersagen, 144
- $\kappa$ , 217

- Kenngrößen, 4, 42, 48–51, 53–55, 58, 61, 68,  
 69, 75, 77, 79, 81–85, 134, 138, 140–  
 142, 147  
 weich, 61, 77–84, 122, 125, 144, 147, 152,  
 156
- Kenngrößen  
 weich, 122
- Kernel-Trick, 32
- Klassenzugehörigkeit  
 weich, 36, 37
- Klassenzugehörigkeitswahrscheinlichkeit, 27  
 a posteriori, *siehe* a posteriori Wahrscheinlichkeit
- Klassifikation, 26–32  
 closed world, 28  
 Einklassen-, *siehe* Einklassen-Klassifikation  
 geschlossenes System, 28, 30, 35, 50, 52,  
 53, 61, 78, 79, 82  
 offenes System, 35, 52, 78, 79  
 weich, *siehe* weiche Klassifikation, 35, 59,  
 116
- knitr, 97
- kohärente anti-Stokes Raman-Spektroskopie,  
*siehe* CARS
- Komplexität  
 Modell, 34, 39, 67, 69, 70, 153
- Konfidenzintervall, 55, 56
- Konfidenzniveau, 56, 217
- Konvergenz, 23
- Korrektheit, 85
- Kreuzvalidierung, 41, 44, 69, 74, 148  
*k*-fach, 41  
 iteriert, 44, 57, 64, 72–75, 109, 110, 121,  
 122, 126, 132, 139  
*k*-fach, 39, 40, 45, 57, 65, 67, 73–75, 110,  
 121, 122, 126, 132, 139  
 Leave-one-out, 44  
 leave-one-out, 41
- künstliches neuronales Netz, 30
- künstliches neuronales Netz, 37
- lattice, 97
- LDA, 24–32, 54, 69, 70, 100, 101, 108, 109, 117,  
 120, 127, 134, 145  
 Annahmen, 29  
 Projektion, 120
- Lebenserwartung, 2
- Leerwert, 23
- Lernkurve, 33, 34, 42, 64–67, 72
- Lipid, 12
- Liquor, 218
- logistische Funktion, 29
- logistische Regression, 29–31  
 Projektion, 120
- Logit-Funktion, 29
- Lymphom, 2, 3, 7–8
- MAE, 81, 122
- Magnetresonanztomographie (MRT), 7, 8  
 intraoperativ, 9–10
- Mahalanobis-Distanz, 218
- maligne, *siehe* Dignität, 218
- MASS, 108
- Matlab, 91, 92, 97, 188, 190, 191
- Matrix, *siehe* Probenmatrix, *siehe* Probenma-  
 trix
- Median, 218  
 2d, 120
- Mehrfachbestimmung, 40, 57, 75, 76, 98
- Messprogramm, 188
- metrisch, 218
- mittlerer absoluter Fehler, MAE, 81
- model update, 43
- Modell  
 überwacht, 218  
 Aggregation, 139  
 Stabilität, *siehe* Stabilität  
 triviales, 53  
 unüberwacht, 220
- Modelle  
 weich, 36
- Modellkomplexität, 33
- Modelloptimierung  
 Testprobenzahl, 58
- Modellqualität, 23, 30, 32–34, 37, 42, 43, 48,  
 58, 60, 64, 67–69, 77, 78, 84, 105, 107,  
 122, 125, 139, 148, 152, 153  
 Trainingsprobenzahl, 32–34
- Modellselektion  
 datengesteuert, 48, 55
- Modellvergleich, 68  
 Testprobenzahl, 58
- molekulare Veränderungen, 6, 7
- Morphologie, 3, 6, 7, 94, 95, 103, 114, 115, 126,  
 132, 146, 154
- MSE, 122
- multiple Sklerose, 9
- multiplicative signal correction, 98

## Index

- Nachweisgrenze, 26, 27
- nnet, 108
- No free lunch Theoreme, 70
- nominalskaliert, 218
- normales Gewebe, 2
- Normalmode, 12
- Normalschwingung, 12
- numerische Apertur, 17, 91
  
- Ödem, 218
- OpenBlasThreads, 97
- Optimierung, 32, 48, 70, 153, 154
  - datengesteuert, 48, 55, 59, 68–70, 105, 153
- ordinalskaliert, 218
- out-of-bootstrap-Validierung, 45, 72, 73
- overfitting, *siehe* Ueberanpassung
  
- partial least squares regression, *siehe* PLS
- Patientenzahl, 3
  - Ausschluss, 38
  - Literatur, 33
- PBS, 20
- PCA, 24–26, 29, 31–33, 48, 127
- Perzentil, 218
  - 2d, 120
- PLS, 24–26, 29, 31–33, 48, 109, 110, 112, 113,
  - 122, 127, 139, 152
- pls, 109
- PLS-DA, 26
- PLSR, *siehe* PLS
- plyr, 97
- Polymorphie, 7, 218
- PostgreSQL, 97, 189, 191
- power, 220
- prädiktiv, 218
- Prävalenz, 27, 51–53, 145, 219
- pre-processing, *siehe* Datenvorbehandlung
- Probe
  - Feuchte, 130
- Probenalter, 117
- Probendichte, 33
- Probenlagerung, 88, 117–118
- Probenmatrix, 12, 23, 218
- Probenpräparation, 88
- Probenzahl
  - Empfehlung, 33
  - Modelltraining, 30, 32, 33
  - Validierung, 57
- producer's accuracy, *siehe* Sensitivität
  
- Protein, 12
  - Faltung, 12
  
- Quantil, 219
  - 2d, 120
- Quartil, 219
  
- R, 97, 108, 110, 191
- R.matlab, 97
- Raman-Messungen, 91
- RamanGUI, 91, 92, 188–190
- random forest, 29, 31, 37, 40
- Rastersuche, 32, 59, 153
- Raumdiagonale, 22, 33
- Rayleigh-Streuung, 13
- Receiver-Operating-Curve, 54
- Referenzdaten, 26, 46, 49, 59–61, 66, 77–82,
  - 95, 105, 116, 121–123, 134, 135, 138,
  - 140–142, 144, 147, 149, 156
- Referenzdiagnose, 94–96, 114–116
- Regularisierung, 25, 219
- Resampling, 40–41
  - Interpretation der Datensätze, 74
- Resektion
  - vollständig, 7, 10, 11
- reshape2, 97
- Resubstitution, 43, 72, 73
- Rezidiv, 219
- RMSE, 122
- Robustheit, 38
- ROC, 54
- Rohdaten, 22
- ruggedness, *siehe* Robustheit
  
- Schwingungsspektroskopie, 12–18
- Selektivität, 219
- Sensitivität, 4, 47, 50–55, 57, 58, 61, 64–69, 82–
  - 84, 110, 121, 125, 126, 131, 134, 138,
  - 140–142, 144, 147, 148, 153, 219
  - weich, 61, 62, 77, 80, 82, 83, 122, 123, 131,
  - 134, 135, 140, 141, 145
- SERS, 11
- Set-Validierung, 45
- shot noise, 16
- Sicherheitsabstand
  - Probe, 92, 100
  - Tumor, 2
- Signal-Rausch-Verhältnis, 14, 16, 17, 22, 28,
  - 70, 90, 97, 100–102, 117, 119, 125, 138,
  - 154



- Signifikanzniveau, 219
- SIMCA, 37
- snow, 110
- Soft Independent Modeling of Class Analogies, *siehe* SIMCA
- softclassval, 84, 85, 97, 110
- softmax, 29, 30, 109
- Software
  - HoloGRAM, 91, 188, 189
  - Matlab, 91, 92, 97, 188, 190, 191
  - PostgreSQL, 97, 191
  - R, 97, 108, 110, 191
    - arrayhelpers, 97
    - ggplot2, 97
    - hyperSpec, 97
    - knitr, 97
    - lattice, 97
    - MASS, 108
    - nnet, 108
    - OpenBlasThreads, 97
    - pls, 109
    - plyr, 97
    - R.matlab, 97
    - reshape2, 97
    - snow, 110
    - softclassval, 84, 85, 97, 110
  - RamanGUI, 91, 92, 188–190
- Sonde
  - faseroptisch, 2, 14
- Sonden
  - faseroptisch, 16
- Spektroskopie
  - Infrarot, 12, 14
    - Literatur, 18–20
  - Infrarot, 14
  - Raman, 2, 3, 11, 13–18
    - Anregungswellenlänge, 15
    - biologische Proben, 16
    - CARS, *siehe* CARS
    - Literatur, 18–20
    - räumliche Auflösung, 18
    - Rauschen, 16
    - Sonden), 16
- Spektrum
  - als Datenpunkt, 21
  - weich, 96
- Spezies
  - chemische, 12–14, 18, 25, 219
- Spezifität, 4, 50–55, 58, 61, 69, 77, 79, 81, 82, 110, 121–123, 125, 131, 134, 135, 138, 140–142, 144, 145, 147, 148, 153, 219
  - weich, 61, 62
- Spikes, 98
- Stabilisierung
  - Modell, 31
- Stabilität, 39
  - Modell, 26, 31, 38–40, 44, 57, 67–69, 73, 74, 109, 127, 139, 142, 148
  - Vorhersagen, 39, 74, 75, 125, 132, 134, 139
- statistische Unabhängigkeit, 46
- Sternzelle, 6
- Stichprobenumfang, 44, 55, 56, 67
  - effektiv, 33, 57, 148
  - Empfehlung, 33
  - notwendiger, 57
- Stokes-Strahlung, 14
- Stratifizierung, 219
- Summenparameter, 12
- Supportvektor-Maschine, *siehe* SVM
- Surface Enhanced Raman Spectroscopy, *siehe* SERS
- Surrogatmarker, 11, 219
- Surrogatmodell, 44–47, 49, 57, 68, 69, 72–76, 79, 109, 110, 124, 138–142, 148
- SVM, 31, 32, 37, 48
- Testproben, 42, 43, 49, 74, 131, 151, 152
  - Anzahl, 42, 44, 57–59, 66, 68, 75, 79, 148
  - statistische Unabhängigkeit, 44, 45, 48, 58, 59, 64, 68, 72, 109
- Teststärke, 220
- Toxoplasmose, 9
- Trefferquote, 46, 47, 50, 51, 82, 220
- Trefferrate, *siehe* Trefferquote
- Tumorproben, 88
  - PostgreSQL-Datenbank, 189
- Tumorrand, 3
- Überanpassung, 32, 43, 45, 69, 70, 72, 73, 153
- Unit-Tests, 85
- Unsicherheit
  - Bayes, *siehe* Bayes-Fehler
  - systematisch
    - Datenvorbehandlung, 23
    - Modell, 32, 39
  - zufällig, 76
    - Datenvorbehandlung, 23

## Index

- Modell, 32, 38, 39, 46, 54, 57, 75
- Validierung, 42, 44, 45, 57, 58, 62, 65–70, 72, 73, 75, 77, 81, 83, 84, 147, 152, 153
- Vorhersage, 46, 75, 83, 125
- Validierung
  - Resampling, *siehe* Resampling
- Validierung, 3, 42–62
  - Ensemble-Modell, 45
  - geschachtelt, 48, 69
  - hold out, *siehe* hold-out-Validierung
  - Iterationen, 44–46, 57, 66, 73–75, 121–125, 132, 134, 135, 138–142
  - Kreuz-, *siehe* Kreuzvalidierung
  - nested, 48
  - Resampling, *siehe* Resampling
  - Schemata, 42–45, 72–73
- Variate, 220
  - Anzahl, 33
  - uninformativ, 24
- Vergleich
  - Modelle, *siehe* Modellvergleich
- Violinen-Diagramm, 220
- vollständige chirurgische Entfernung, *siehe* Resektion
- vollständige chirurgische Entfernung, *siehe* Resektion
- Vorbehandlung, *siehe* Datenvorbehandlung
- Vorhersagewert
  - negativer, 50, 52, 53, 61, 62, 79, 145, 147, 153
  - positiver, 50, 52, 53, 61, 62, 145, 147, 153
- Wahrscheinlichkeit
  - a posteriori, 216
  - a priori, 216
- Wassergehalt, 117, 126, 130, 131
- weich
  - Chemometrie, 36, 37
  - Kenngröße, 82, 134
  - Klassenzugehörigkeit, 36, 37, 59
  - Klassifikation, 34, 35, 59, 116
  - Referenz, 37, 95, 140, 141
  - Spektrum, 103, 108, 120–122, 131
  - Vorhersage, 37, 81, 84
  - Zuordnungsmatrix, 61
- Wellenzahlgenauigkeit, 14
- WHO, 5
- Wiederholungsmessung, *siehe* Mehrfachbestimmung
- wMAE, 81
- Zellpopulation, 3
- Ziehen
  - mit Zurücklegen, 41
  - ohne Zurücklegen, 41
- ZNS, 5
- Zuordnungsmatrix, 55